



© M. Romanică

Exerciții de învățare automată

Liviu Ciortuz, Alina Munteanu, Elena Bădărău

Draft, 17 mai 2024

Pentru corecții și sugestii de îmbunătățire, vă rugăm să folosiți adresa
ciortuz@info.uaic.ro

Motto:

*„Tot ce găsește mâna ta să facă, fă cu
toată puterea ta, căci în locuința morților,
în care mergi, nu mai este nici lucrare, nici
chibzuială, nici știință, nici înțelepciune.“*

Cartea Eclesiastului 9:10

Cuprins

Multumiri	7
Cuvânt înainte	9
0 Fundamente matematice ale învățării automate	16
0.1 Probleme rezolvate	23
0.1.1 Evenimente aleatoare și formula lui Bayes	23
0.1.2 Variabile aleatoare	30
0.1.3 Distribuții probabiliste uzuale	56
0.1.4 Estimarea parametrilor unor distribuții probabiliste	93
0.1.5 Elemente de teoria informației	129
0.1.6 Funcții-nucleu	157
0.1.7 Metode de optimizare în învățarea automată	169
0.2 Probleme propuse	214
0.2.1 Evenimente aleatoare și formula lui Bayes	214
0.2.2 Variabile aleatoare	216
0.2.3 Distribuții probabiliste uzuale	221
0.2.4 Estimarea parametrilor unor distribuții probabiliste	228
0.2.5 Elemente de teoria informației	238
0.2.6 Funcții-nucleu	247
0.2.7 Metode de optimizare în învățarea automată	253
1 Metode de regresie	268
1.1 Probleme rezolvate	271
1.1.1 Regresia liniară	271
1.1.2 Regresia logistică	300
1.2 Probleme propuse	317
1.2.1 Regresia liniară	317
1.2.2 Regresia logistică	329
1.2.3 Modele liniare generalizate	338
2 Clasificare bayesiană	344
2.1 Probleme rezolvate	348
2.1.1 Ipoteze de probabilitate maximă a posteriori (MAP)	348
2.1.2 Algoritmii Bayes Naiv și Bayes Optimal	358
2.1.3 Clasificare bayesiană gaussiană	384
2.2 Probleme propuse	404
2.2.1 Ipoteze de probabilitate maximă a posteriori (MAP)	404
2.2.2 Algoritmii Bayes Naiv și Bayes Optimal	406
2.2.3 Clasificare bayesiană gaussiană	418

3 Învățare bazată pe memorare — Algoritmul k-NN	428
3.1 Probleme rezolvate	431
3.2 Probleme propuse	462
4 Arbori de decizie	472
4.1 Probleme rezolvate	478
4.1.1 Algoritmul ID3	478
4.1.2 Algoritmul AdaBoost	527
4.2 Probleme propuse	575
4.2.1 Algoritmul ID3	575
4.2.2 Algoritmul AdaBoost	597
5 Mașini cu vectori-suport	620
5.1 Probleme rezolvate	624
5.1.1 SVM cu margine “hard”	624
5.1.2 SVM cu margine “soft”	650
5.1.3 Alte probleme de optimizare de tip SVM	690
5.2 Probleme propuse	709
5.2.1 SVM cu margine “hard”	709
5.2.2 SVM cu margine “soft”	717
5.2.3 Alte probleme de optimizare de tip SVM	726
6 Rețele neuronale artificiale	736
6.1 Probleme rezolvate	739
6.1.1 Chestiuni introductive	739
6.1.2 Unități neuronale — algoritmi de antrenare	755
6.1.3 Rețele “feed-forward” — algoritmul de retropropagare	777
6.2 Probleme propuse	795
6.2.1 Chestiuni introductive	795
6.2.2 Unități neuronale — algoritmi de antrenare	799
6.2.3 Rețele “feed-forward” — algoritmul de retropropagare	808
7 Clusterizare	820
7.1 Probleme rezolvate	827
7.1.1 Clusterizare ierarhică	827
7.1.2 Algoritmul K -means	841
7.1.3 Algoritmul EM pentru modele de mixturi gaussiene	861
7.2 Probleme propuse	907
7.2.1 Clusterizare ierarhică	907
7.2.2 Algoritmul K -means	917
7.2.3 Algoritmul EM pentru modele de mixturi gaussiene	930
8 Schema algoritică EM	952
8.1 Probleme rezolvate	955
8.1.1 Fundamente teoretice	955
8.1.2 Mixturi de distribuții Bernoulli / categoriale	965
8.1.3 Distribuții binomiale / multinomiale	1014
8.1.4 Sume de variabile aleatoare	1021
8.1.5 Alte instanțe ale schemei algoritmice EM	1032
8.2 Probleme propuse	1037
8.2.1 Fundamente teoretice	1037

8.2.2	Mixturi de distribuții Bernoulli / categoriale	1041
8.2.3	Distribuții binomiale / multinomiale	1051
8.2.4	Alte câteva instanțe ale schemei algoritmice EM	1058
9	Modele Markov ascunse	1066
9.1	Probleme rezolvate	1067
9.2	Probleme propuse	1090

Mulțumiri

O mare parte dintre exercițiile și problemele din această culegere au fost date la examenele sau temele pentru acasă de la cursurile de învățare automată ținute în anii 1994-2022 la Carnegie Mellon University (CMU) din Pittsburgh, Statele Unite ale Americii, de către profesorii Tom Mitchell, Andrew Moore, Carlos Guestrin, Geoff Gordon, Eric Xing, William Cohen, Ziv Bar-Joseph, Aarti Singh, Roni Rosenfeld, Alex Smola, Barnabás Póczos, Seyoung Kim, Nina Balcan și Matt Gormley. Alte exerciții și probleme provin de la cursul de învățare automată al profesorului Andrew Ng de la Universitatea Stanford, precum și cel al profesorilor Tommi Jaakkola, Regina Barzilai și Leslie Kaelbling de la MIT, S.U.A.

Facem mențiunea că, în general, rezolvările care apar în prezenta culegere au un nivel de detaliere semnificativ mai elaborat decât rezolvările originale, prezentate adeseori doar în mod esențializat de către profesorii mai sus menționați și/sau asistenții dânsilor.

Tuturor acestora le aducem pe această cale cele mai sincere mulțumiri pentru generozitatea cu care au pus la dispoziție pe internet aceste materiale didactice.

Mai mulți studenți de la Facultatea de Informatică a Universității „Alexandru Ioan Cuza“ din Iași ne-au ajutat de-a lungul ultimilor ani la redactarea textelor sau a imaginilor pentru unele probleme. Îi menționăm în ordinea alfabetică a numelui de familie: Luisa Apachiței, Gheorghe Balan, Ciprian Băetu, Cristian Budăianu, Giorgiana Caltais, Petru Cehan, Ivona Chili, Sebastian Ciobanu, Lidia Corciova, Oana Cotoman, Denis Crusos, Sergiu Dinu, Ahmad Cezar El-Nazli, Oana Florean, Eugen Goriac, Irina Grosu, Andrei-Constantin Iacob, Sergiu Iacob, Anca Luca, Diana Lucaci, Ștefan Matcovici, Dorin Miron, Alexandru Mitan, Andi Munteanu, Georgiana Ojoc, Oriana Oniciuc, Mihaela Potlog, Andreea Prodan, Mădălina Racoviță, Elena-Irina Radu, Ciprian Recianu, Cristian Rotaru, Irina Roznovăț, Alexandra Sbiera, Marius Spănu, Andreea Stanciu, Cristina Șerban, Octavian Tamaș, Călin-Raș Turliuc și Liana-Ștefania Tucăr. (Ne cerem sincer scuze pentru evenualele omisiuni.)

De asemenea, alături de Ștefan Panțiru, Anca Ignat, Mihaela Breabă și Adrian Zălinescu, care au ținut seminarii, mai mulți studenți — dintre care îi menționăm pe Ramona Albert, Luisa Apachiței, Vlad Aioanei, Mihai Baboi, Ștefan Bălăucă, Diana Carcea, Ștefan Catîru, Nicolae Căpățînă, Sebastian Ciobanu, Andrei Cioromilă, Cristian Cojocaru, Laura Cornei, Rareș Dima, Alexandru Dobranici, Alina Duca, Marius Georgică, Mihai Grigoriță, Sorin Grițco, Mircea Irimescu, Diana Lucaci, Cecilia Mariciuc Mălină Marin, Ștefan Matcovici, Alexandru Mărtinaș, Mihaela Mengheres, Alexandra Minghel, Ionuț-Vlad Modoranu, Petru Munteanu, Lucian Neștian, Lucian Nevoe, Axenia Niță, Alexandru Oloieri, Alexandra Pal, Andreea Pădurariu, Iulian Pichiu, Codrina Prisecaru, Emanuel Pușcașu, Mihaela Radu, Carol Rameder, Ștefan-Vladimir Sbârcea, Raluca Scortanu, Cristian Simionescu, Nicolae Șoitu, Ionel Ștefanucă, Silviu Șutea-Drăgan, Cezar Todirișcă, Iulian Vâscu, Silviu Vițel și Ina Vivdici —, precum și domnul Liviu Olaru și conf. dr. Elena Nenciu, au identificat anumite erori în versiuni preliminare ale prezentei culegeri.

Mulțumim domnilor Ștefan Panțiru, Daniel Munteanu și Dorin Miron pentru instalarea programelor necesare pentru tehnoredactare. Mulțumim redactorilor de carte Eduard Dulman și Cerasela Cirimpei, precum și doamnelor Dana Lungu și Luminița Răducanu de la Editura Universității „Alexandru Ioan Cuza“ din Iași pentru asistența acordată în vederea publicării primei ediții a acestei cărți. Mul-

țumiri speciale se cuvin Mihaelei Romanică din Antwerpen, Belgia, care ne-a pus la dispoziție desenele care apar în deschiderea fiecărui capitol, precum și desenul de pe copertă.

Liviu Ciortuz îi mulțumește pe această cale profesorului Cristian Preda de la Universitatea din Lille și INRIA – Nord Europe pentru posibilitatea de a lucra acolo la finalizarea primei ediții a acestei culegeri, în luna iulie 2015.

Cuvânt înainte — de Liviu Ciortuz

Așa cum se cuvine, vom preciza aici următoarele chestiuni: de ce anume a fost scrisă această culegere, cum a fost ea alcătuită în timp, care au fost criteriile de selecție atât la nivel tematic — adică, de ce au fost alese respectivele capitole de învățare automată și nu altele — cât și la nivelul tipului / tipurilor de exerciții incluse, cum se situează această culegere în raport cu alte culegeri din aria învățării automate, care este suportul bibliografic,¹ documentele on-line care o însoțesc — site-ul asociat, slide-uri, un index tematic și un „companion“ de tip probleme de implementare —, precum și ideile pe care le avem acum cu privire la o posibilă extindere ulterioară a culegerii.

Cartea aceasta s-a născut adunând mai întâi exercițiile care au fost date la examenele și teste „pe parcurs“ din cadrul cursului de *Învățare automată* de la Facultatea de Informatică a Universității „Alexandru Ioan Cuza“ din Iași. Cursul acesta s-a ținut începând din anul 2003.² Atunci când s-au acumulat un număr semnificativ de exerciții traduse în limba română, a fost impede că ar fi de mare ajutor ca rezolvările acestor exerciții să fie puse la dispoziția studenților din seriile noi, sub forma unei culegeri. Pentru aceasta, am socotit că este benefic să apelez la ajutorul unora dintre cei mai buni studenți care au absolvit deja cursul. Astfel, mai întâi Elena Bădărău (studentă la master, seria 2008–2010) și apoi, într-o măsură considerabil sporită, Alina Munteanu (studentă la master, seria 2009–2011, iar după aceea doctorand) au redactat soluții (în general, semnificativ mai extinse față de original), la multe dintre problemele pe care le culesesem anterior.

De atunci și până acum am cules, am tradus și chiar am formulat multe alte exerciții, cu scopul acoperirii unei arii suficient de largi pentru fiecare dintre capituloare predate la curs. Fiecare capitol din culegere a fost împărțit într-o secțiune de probleme în probleme rezolvate și o secțiune de probleme propuse. În ultimii ani, după ce cele două co-autoare au plecat din facultate,³ am redactat multe soluții și am realizat șlefuirea întregului material. De-a lungul acestor ani, am prezentat la cursurile de învățare automată următoarele capituloare (selecția și ordinea lor de succesiune variind ușor de la an la an): fundamente matematice — și anume, probabilități și statistică, estimarea parametrilor unor distribuții probabiliste, teoria informației, funcții-nucleu și metode de optimizare (toate aceste secțiuni având rol de suport pentru capituloarele următoare) —, modele de regresie, clasificare bayesiană, învățare bazată pe memorare, arbori de decizie, mașini cu vectori-suport,

¹ Este vorba despre cartea sau cărțile care sunt mai potrivite pentru a fi folosite în tandem cu această culegere și, foarte posibil, pentru cursul aferent.

² În perioada 2003–2008 el a fost curs optional la ciclul de licență, anul IV. Apoi, odată cu trecerea la sistemul Bologna, el a devenit curs obligatoriu la ciclul de master. Ulterior, două dintre cele cinci specializări / programe de master de la facultatea noastră l-au pus la dispoziția studenților drept curs obligatoriu, iar celelalte trei l-au păstrat drept curs optional.

Începând din anul 2015, cursul acesta s-a scindat într-un curs de bază (obligatoriu) în cadrul ciclului de licență, la anul III, și un curs avansat (sub denumirea de *Capitole speciale de Învățare automată*) la master.

În anii universitari 2018–2019 și 2019–2020 cursul de bază de *Învățare automată* a fost ținut sub formă de curs optional și la masterul de *Matematici aplicate* de la Facultatea de Matematică a Universității „Alexandru Ioan Cuza“ din Iași. De asemenea, în semestrul întâi al anului universitar 2012–2013 am ținut cursul de *Învățare automată* în regim modular (timp de 4 săptămâni) și la studenții de la master, specializarea informatică, de la Facultatea de Matematică-Informatică a Universității de Vest din Timișoara.

³ După absolvirea masterului, Elena Bădărău a lucrat pentru un timp la compania Amazon din Iași. În perioada 2015–2017, dânsa a avut amabilitatea de a ține seminarii la cursul de *Învățare automată* de la ciclul de licență din facultatea noastră. Alina Munteanu a început doctoratul la Iași și apoi l-a continuat la *Max Delbrück Center for Molecular Medicine* din Berlin, Germania, folosind tehnici de învățare automată în rezolvarea unor probleme de cercetare în bioinformatică. Dânsa și-a susținut teza de doctorat la *Universitatea Humboldt* din Berlin în octombrie 2017.

rețele neuronale artificiale, clusterizare, schema algoritmică EM, modele Markov ascunse, și, în sfârșit, teoria învățării computaționale.

În actuala versiune a culegerii apar exerciții pentru primele zece capitole dintre cele unsprezece enumerate mai sus. Dintre acestea, capitolul 0 (anumite secțiuni) și capitolele 2-4 și 7 corespund acum programei cursului de *Învățare automată* de la licență, iar restul capitolelor corespund programei cursului de *Capitole speciale de Învățare automată* de la programul de master de la facultatea noastră.⁴ Pe lângă adăugarea de noi capitole, ne propunem ca pe viitor să extindem culegerea prin introducerea de noi exerciții care să acopere „segmente“ care nu sunt prezente în capitolele deja elaborate. Ne gândim, de exemplu, la unele exerciții legate de analiza componentelor principale (engl., *Principal Component Analysis*), clusterizare spectrală (engl., *spectral clustering*), precum și analiza prin factori (engl., *factor analysis*).

Majoritatea exercițiilor din această culegere sunt destinate a fi rezolvate „cu creionul pe hârtie“. Ele solicită în general fie *aplicarea* principaliilor algoritmi din capitolele / domeniile de învățare automată care au fost menționate mai sus, fie *demonstrarea* unor proprietăți (teoretice) ale acestor algoritmi sau ale conceptelor utilizate de ei.

Remarcăm modul în care au fost și sunt în general concepute exercițiile cu caracter teoretic de la cursurile de *Machine Learning* de la CMU. Și anume, ele călăuzesc pas cu pas studentul în aşa fel încât să ajungă să demonstreze el însuși (cu încredere crescândă în forțele proprii) rezultate importante. În acest fel se micșorează sau chiar se elimină efortul necesar pentru memorarea integrală a acestor demonstrații, care devin astfel mai degrabă obiect de lucru acasă sau în grup decât de prezentare abstractă și aridă la curs. Pe această cale se realizează o conlucrare foarte bine mediată între profesor (și / sau asistent) pe de o parte și student pe de altă parte, pe baza acestor texte de tip problemă / exercițiu, bine concepute și elaborate având în minte acest scop didactic.

În alcătuirea acestei culegeri am lăsat în mod intenționat de o parte problemele cu specific de implementare. Am procedat aşa din două motive. În primul rând, există suficient de multe cărți de specialitate care se ocupă de astfel de chestiuni. În al doilea rând, am început deja să creăm pentru studentii noștri un repertoir (engl., *repository*) de seturi de date, implementări și scripturi pentru cele mai instructive probleme practice date la lucrările pentru acasă (engl., *homeworks*) de la cursurile de *Machine Learning* de la CMU și eventual de la alte universități. Corespunzător, este în curs avansat de elaborare (în limba engleză) o fasciculă care constituie un *Companion practic* pentru prezenta culegere.

Culegerea are un site asociat: <http://profs.info.uaic.ro/~ciortuz/ML.ex-book/>. Acolo se găsesc — pe lângă unele *seturi de date* folosite la exerciții pentru care au fost create grafice ce pun în evidență comportamentul unor algoritmi de învățare automată pe care îi studiem — *slide-uri* pentru cele mai utile exerciții din culegere. Menționăm că multe dintre aceste slide-uri au fost folosite nu doar la seminarii ci și la curs, pentru exemplificarea aplicării algoritmilor predăți, demonstrarea unor proprietăți importante ale lor etc. Sperăm că acest set de slide-uri se va mări în timp și că ele vor fi folosite și de cadre didactice de la alte facultăți. Majoritatea slide-urilor au fost redactate *în limba engleză* (spre deosebire de culegere!), pentru ca de ele să beneficieze un număr cât mai mare de studenți.⁵

⁴Capitolul de *Teoria învățării computaționale* (engl., Computational learning theory) se află acum într-un stadiu incipient de elaborare. Sperăm că vom avea răstimpul necesar pentru finalizarea lui și atunci acest capitol va fi adăugat la culegere într-o ediție viitoare.

⁵La fiecare dintre exercițiile pentru care am făcut deja slide-uri, am pus în antet — mai precis, acolo unde

Ca *suport bibliografic* la *curs* am folosit cu preponderență în primii ani de predare cartea *Machine Learning* a profesorului Tom Mitchell de la CMU (ed. McGraw-Hill, 1997). Deși nu mai este de dată recentă, această carte are un stil ușor accesibil studenților noștri. Menționăm că, dintre cele unsprezece capitole enumerate mai înainte, cartea profesorului Tom Mitchell nu conține capitolele de fundamente, clusterizare, mașini cu vectori-suport și modele Markov ascunse (dar conține alte capitole, precum învățare bazată pe ranforsare și algoritmi genetici). Pentru trei dintre cele patru capitole pe care tocmai le-am menționat,⁶ am folosit cartea *Foundations of Statistical Natural Language Processing* (Christopher Manning, Hinrich Schütze, MIT Press, 2002).⁷ Pe parcursul acestor ani de predare, am făcut referire din ce în ce mai mult și la notițele de curs (engl., *Lecture notes*) ale renumitului profesor Andrew Ng de la Stanford University, în care se folosesc adeseori formalizări matematice mai sofisticate.⁸

La începutul fiecărui capitol din carte am alcătuit un *sumar*, sau o *privire de ansamblu*, un conspect (engl., overview) asupra capitolului respectiv. El poate fi văzut și ca un *index de probleme*. Ansamblul acestor conspective constituie un rezumat al cursului / cursurilor de *Învățare automată* de la facultatea noastră. Pentru aproape fiecare chestiune care se pretează a fi predată la curs, acest *sumar* / *index* indică unul sau mai multe exerciții care se referă la (sau folosesc, exemplifică) chestiunea respectivă.

Sintetizând, putem spune că această culegere este o încercare de a pune împreună în mod unitar exerciții cu caracter preponderent conceptual și teoretic pentru principalele capitole de învățare automată. Ea folosește un aparat matematic care este bazat în special pe noțiunile de matematică din liceu și de la cursurile de matematică și de probabilități din primul an de facultate.

Remarcăm că, pe de o parte, există cărți conținând exerciții specializate pe [câte] un capitol de învățare automată, de exemplu *Regression Analysis by Example* (Samprit Chatterjee, Ali S. Hadi, Wiley-Interscience, 2006) și *Reinforcement Learning: An Introduction* (Richard S. Sutton, Andrew G. Barto, MIT Press, 1998). Însă, pe de altă parte, nu știm să existe o culegere similară culegerii noastre, având o arie suficient de vastă, cu exerciții situate pe două palieri, unul exemplificativ, iar celălalt conceptual și teoretic, în aşa fel încât (foarte important!) să fie adecvate pentru nivelul studenților de la facultatea noastră. Trebuie spus că exercițiile din cărțile consacrate în domeniul învățării automate, care acoperă un număr mare de capitole — de exemplu *Pattern Classification* (Richard O. Duda, Peter E. Hart, David G. Stork, Wiley-Interscience, 2003), *Pattern Recognition and Machine Learning* (Christopher M. Bishop, Springer, 2006), *The Elements of Statistical Learning. Data Mining, Inference and Prediction* (Trevor Hastie, Robert Tibshirani, Jerome Friedman, Springer, 2009), precum și *Pattern Recognition* (Sergios Theodoridis, Konstantinos Koutroumbas, Academic Press, 2009) —, folosesc adeseori un aparat matematic destul de complex. Rolul culegerii noastre este de

se indică *subiectul* și *sursa* respectivului exercițiu — semnul ■.

⁶Excepție face capitolul despre mașini cu vectori-suport (SVM). O foarte bună prezentare a acestui domeniu este făcută în cartea *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods* de Nello Cristianini și John Shawe-Taylor, publicată la editura Cambridge University Press, în anul 2000.

⁷Alegerea acestei cărți se datorează experienței pe care am acumulat-o mai înainte în acest domeniu (NLP), asupra căruia învățarea automată a avut un foarte puternic impact în ultimele două decenii, iar acest fapt este bine reflectat în cartea menționată.

⁸Pentru studenții cei mai sărguinčioși, care vor să se inițieze în cercetarea din domeniul învățării automate de tip „shallow“, recomandăm acum cartea *Machine Learning — A Probabilistic Perspective* de Kevin P. Murphy, publicată la editura MIT Press în anul 2012.

a fi o „punte“ înspre nivelul cărților pe care tocmai le-am menționat,⁹ astfel încât studenții noștri cei mai buni să poată să acceadă la acel nivel.¹⁰

Experiența pe care am acumulat-o în ultimii ani, în care am folosit deja culegerea, dotând studenții de la cursurile noastre de *Învățare automată* cu drafturi preliminare ale ei, ne arată că exercițiile din această carte sunt suficiente nu doar pentru două cursuri de câte un semestru (având săptămânal două ore de predare propriu-zisă și două ore de seminar), ci și pentru seminarii suplimentare / avansate, cu un grup de studenți de “top”.

Înainte de a încheia, trebuie precizate câteva *detalii tehnice*:

1. La folosirea ghilimelelor, am utilizat “ ” pentru expresii în limba engleză, respectiv „ „ pentru expresii românești, conform regulilor de ortografie corespunzătoare.
2. Pentru numere zecimale, din pură conveniență, am folosit în general notația de tip anglo-saxon (adică, de exemplu, 1.5 în loc de 1,5, cum ar fi trebuit), datorită faptului că am tehnoredactat în paralel două versiuni (engleză și română) ale culegerii, iar formulele matematice sunt (deocamdată) comune pentru cele două versiuni.
3. Menționăm că, tot din motive de conveniență la tehnoredactare, figurile (dar și tabelele) din carte nu au fost numerotate și, în consecință, fiecare figură a fost plasată în imediata vecinătate a textului care se referă la ea.
4. Precizăm că atunci când, în cadrul unei probleme (fie al enunțului, fie al soluției), ne vom referi la un anumit punct al problemei respective, îl vom desemna printr-unul din caracterele italice *a*, *b*, *c* etc, deși la începutul punctului respectiv apare litera scrisă sub forma uzuală: *a*, *b*, *c* etc. De asemenea, atunci când, într-o anumită problemă, ne vom referi la un punct al unei alte probleme, vom scrie pe scurt, de exemplu, „vedeți problema 14.b“; se va înțelege că facem trimitere la punctul *b* al problemei 14.

Atât eu cât și ceilalți doi coautori ai acestei culegeri ne dorim ca vastul efort depus pentru alcătuirea ei — în primul rând, de către autorii exercițiilor originale dar, cu modestia de rigoare, și de către noi — să permită generațiilor actuale și viitoare de studenți și doctoranți să-și însușească cunoștințelele de bază din domeniul învățării automate, care a ajuns azi la un nivel impresionant de aplicabilitate, atât în industria IT cât și în cercetarea științifică.

Liviu Ciortuz
Iași, septembrie 2023

⁹Pentru soluții la exercițiile din cărțile enumerate, vedeți: *Solution Manual to accompany Pattern Classification (2nd ed.) by R. O. Duda, P. E. Hart and D. G. Stork* (David G. Stork, 2001), *Pattern Recognition and Machine Learning Solutions to the Exercises: Web-Edition* (Markus Svensén and Christopher M. Bishop, 2009), *A Solution Manual and Notes for: The Elements of Statistical Learning by Jerome Friedman, Trevor Hastie, and Robert Tibshirani* (John L. Weatherwax, David Epstein, http://waxworksmath.com/Authors/G_M/Hastie/hastie.html, 21 June 2013), *Notes and Solutions for: Pattern Recognition by Sergios Theodoridis and Konstantinos Koutroumbas* (John L. Weatherwax, http://waxworksmath.com/Authors/N_Z/Theodoridis/theodoridis.html, October 17, 2015).

¹⁰O culegere de exerciții de tip implementare în domeniul învățării automate este următoarea: *Introduction to Pattern Recognition: A Matlab Approach* (Sergios Theodoridis, Aggelos Pikrakis, Konstantinos Koutroumbas, Dionisis Cavouras, Academic Press, 2010). Pentru exemplificări pe aceeași „linie“, vedeți și *Computational Statistics Handbook with Matlab* (Wendy Martinez, Angel Martinez, CRC Press, 2007, 2015).

“Puerile as such an exercise may seem, it sharpens the faculties of observation, and teaches one where to look and what to look for.”

Sherlock Holmes in *Study in Scarlet*,

cited by

Wolfgang Karl Härdle and Zdeněk Hlávka in
Multivariate Statistics — Exercises and Solutions,
second edition, Springer, 2015



© M. Romanică

- 0 Fundamente matematice:**
 probabilități, variabile aleatoare,
 estimarea parametrilor unor distribuții probabiliste,
 teoria informației, funcții-nucleu
 și metode de optimizare

Sumar

Evenimente aleatoare și formula lui Bayes

- *funcția de probabilitate* – câteva proprietăți care derivă din definiția ei: ex. 1, ex. 93;
- calcularea unor *probabilități elementare*: ex. 2.a, ex. 89;
- calcularea unor *probabilități condiționate*: ex. 2.b, ex. 3, ex. 91.a;
- *regula de înmulțire*: ex. 93.b;
- *formula probabilității totale*: ex. 90, ex. 93.e;
 formula probabilității totale – varianta condițională: ex. 96.cd;
- *independența* evenimentelor aleatoare: ex. 4, ex. 5, ex. 91.bc;
- *independența condițională* a evenimentelor aleatoare – legătura dintre forma „tare“ și forma „slabă“ a definiției pentru această noțiune: ex. 92;
- *formula lui Bayes* – aplicare: ex. 6, ex. 7, ex. 94, ex. 95;
 formula lui Bayes – varianta condițională: ex. 96.b;
- recapitulare: probabilități elementare și probabilități condiționate – câteva proprietăți (A/F): ex. 8, ex. 96.

Variabile aleatoare

- funcție de *distribuție cumulativă* (engl., cumulative distribution function, c.d.f.), un exemplu: ex. 33 (vedeți și ex. 9 de la capitolul *Învățare bazată pe memorare*);
- proprietatea de *liniaritate a mediei*: ex. 9.a;
- *varianța și covarianța*: proprietăți de tip *caracterizare*: ex. 9.bc;
- covarianța oricărora două variabile aleatoare independente este 0: ex. 10; reciproca acestei afirmații nu este adevărată în general: ex. 11.a, ex. 100; totuși ea are loc dacă variabilele sunt de tip binar (adică iau doar valorile 0 și 1): ex. 11.b;
- o *condiție suficientă* pentru ca $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$: independența variabilelor X și Y : ex. 23.c
- pentru *variabile aleatoare discrete*:
- calcularea mediilor și a varianțelor – exemplificare: ex. 12, ex. 97, ex. 99.a;
- definirea unei *variabile-indicator* cu ajutorul unui eveniment aleator; calcularea mediei acestei variabile: ex. 98;

- regula de *înlățuire* (pt. var. aleat.) – aplicare: ex. 13;
- regula de *înmulțire* (pt. var. aleat.), varianta condițională – demonstrație: ex. 14;
- *independență*: ex. 99.b, ex. 101.a, ex. 16.a;
independență condițională: ex. 15, ex. 101.b, ex. 16.b, ex. 102.b;
o condiție suficientă pentru independentă condițională: ex. 103;
- pentru *variabile aleatoare continue*:
- dată fiind o funcție care depinde de un parametru real, să se calculeze valoarea respectivului parametru astfel încât funcția respectivă să fie o *funcție de densitate de probabilitate* (engl., probability density function, p.d.f.): ex. 17.a, ex. 18, ex. 104;
- dat fiind un p.d.f., să se calculeze o anumită probabilitate: ex. 17.c, ex. 105;
- variabile aleatoare discrete vs. variabile aleatoare continue; p.m.f. vs. p.d.f.: ex. 107;
- variabile aleatoare discrete și variabile aleatoare continue; independentă și calcul de medii: ex. 108;
- *coeficientul de corelație* pentru două variabile aleatoare: ex. 19;
- *vector de variabile aleatoare*:
 - o proprietate: matricea de covarianță este simetrică și pozitiv definită: ex. 20; calculul matricei de covarianță când asupra vectorului de variabile aleatoare operăm transformări liniare: ex. 109;
 - câteva inegalități de bază în teoria probabilităților: margini superioare pentru probabilități de forma $P(Z \geq t)$ și $P(Z - E[Z] \geq t)$: ex. 21, ex. 22;
- recapitulare: ex. 23, ex. 110.

Distribuții probabiliste uzuale

- **distribuții discrete**
- **distribuția Bernoulli**:
 - suma de variabile identic distribuite; media sumei: ex. 24;
 - mixturi* de distribuții Bernoulli: ex. 113, ex. 114;
- **distribuția binomială**:
 - verificarea condițiilor de definiție pentru p.m.f.: ex. 25.a;
 - calculul mediei și al varianței: ex. 25.b, ex. 26.bf;
 - calcularea unor probabilități (simple și respectiv condiționale): ex. 111, ex. 26.c;
- **distribuția categorială**:
 - calcularea unor probabilități și a unor medii: ex. 112;
 - mixtură* de distribuții categoriale: calculul mediei și al varianței: ex. 29;
- **distribuția geometrică**:
 - calcularea numărului „așteptat“ / mediu de „observații“ necesare pentru ca un anumit eveniment să se producă: ex. 28;
- **distribuția Poisson**:
 - verificarea condițiilor de definiție pentru p.m.f., calculul mediei și al varianței: ex. 27;

- **distribuții continue**

- distribuția *continuă uniformă*:
exemplu de distribuție continuă uniformă unidimensională; calcularea mediei și a varianței: ex. 115;
calculul unei p.d.f. comune, pornind de la două variabile [urmând distribuții continue uniforme unidimensionale] independente; calculul unei anumite probabilități, folosind p.d.f. comună: ex. 30;
exemplu de distribuție continuă uniformă bidimensională; calcularea unei p.d.f. condiționale; verificarea independenței celor două distribuții marginale; calculul mediilor unor p.d.f.-uri condiționale: ex. 106, ex. 108;
- distribuția *exponențială*:
verificarea condițiilor de definiție pentru p.d.f.,
calculul mediei și al varianței: ex. 31.a;
- distribuția *Gamma*:
verificarea condițiilor de definiție pentru p.d.f., calculul mediei și al varianței: ex. 31.b;
- distribuția *gaussiană unidimensională*:
verificarea condițiilor de definiție pentru p.d.f., calculul mediei și al varianței: ex. 32;
„standardizare“ (i.e., reducerea cazului nestandard la cazul standard): ex. 33;
- distribuția *gaussiană bidimensională*: exemplificare; calcularea explicită a p.d.f.-ului, dat fiind vectorul de medii și matricea de covarianță: ex. 35;
o proprietate pentru distribuția *gaussiană bidimensională*: distribuția condițională a unei componente în raport cu cealaltă componentă este tot de tip gaussian; calculul parametrilor acestei distribuții condiționale: ex. 38;
- distribuția *gaussiană multidimensională*:
matrice simetrice și pozitiv definite¹¹ o proprietate de tip *factorizare* folosind *vectori proprii*: ex. 36;
densitatea distribuției gaussiene multidimensionale este într-adevăr p.d.f.¹²: ex. 37;
o proprietate importantă, în cazul în care matricea de covarianță este diagonală: p.d.f.-ul comun este produsul p.d.f.-urilor marginale (care sunt independente): ex. 34;
- În ce privește specificul datelor generate de distribuții gaussiene multidimensionale:¹³
 - în cazul cel mai general (deci când matricea Σ nu este neapărat diagonală), datele generate de acest tip de distribuție tind să se grupeze în elipse (corpuri elipsoidale) cu *axe de simetrie* [desigur, perpendiculare, dar] în general ne-paralele cu *axe de coordonate*;
 - dacă matricea de covarianță Σ este diagonală, atunci datele generate tind să se grupeze în elipse (dacă se lucrează în \mathbb{R}^2) sau, mai general, corpuri elipsoidale având axe de simetrie paralele cu axele sistemului de coordonate;
 - dacă matricea Σ este de forma $\sigma^2 I$, unde I este matricea identitate, datele generate de respectiva distribuție tind să se grupeze în sfere;

¹¹Așa sunt matricele de covarianță ale variabilelor gaussiene multidimensionale.

¹²Adică satisfac condițiile din definiția noțiunii de p.d.f.

¹³Vedeți corespondența imediată cu alura curbelor de izocontur / nivel determinate de aceste distribuții.

- **mixturi de distribuții gaussiene multidimensionale:**
exprimarea vectorului de medii și a matricei de covarianță în funcție de mediile și matricele de covarianță ale distribuțiilor componente: ex. 118;
- **mixturi de distribuții oarecare:**
calculul mediilor și al varianțelor (în funcție de mediile și matricele de covarianță ale distribuțiilor componente): ex. 119;
în cazul unei distribuții multidimensionale de tip mixtură $p(x) = \sum_{k=1}^K \pi_k p(x|z_k)$, dacă partaționă variabila $x \in \mathbb{R}^d$ în două, sub forma $x \stackrel{\text{not.}}{=} (x_1, x_2)$, atunci variabila condiționată $x_2|x_1$ urmează tot o distribuție de tip mixtură: ex. 39;
- **distribuția Bernoulli și distribuția normală standard:**
intervale de încredere, legea numerelor mari, teorema limită centrală; aplicație la calculul erorii reale a unui clasificator: ex. 40;
- **funcția generatoare de momente** pentru o variabilă aleatoare: ex. 120;
- **familia de distribuții exponentiale:** ex. 41 și ex. 122 (precum și ex. 40.a și ex. 41.a de la capitolul *Metode de regresie*);
- **chestiuni recapitulative** (corespondența dintre nume de distribuții și expresiile unor p.d.f.-uri date): ex. 121.

Estimarea parametrilor unor distribuții probabiliste uzuale

- **distribuția Bernoulli:** ex. 43 (+MAP, folosind distr. Beta), ex. 42, ex. 125 (un caz particular), ex. 126.a, ex. 124.bcd (bias-ul și varianța estimatorului MLE);
- **distribuția binomială:** ex. 126.bc și ex. 127 (ultimul, folosind și metoda gradientului și metoda lui Newton);
- **distribuția categorială:** ex. 44, ex. 128 (+MAP, folosind distr. Dirichlet);
- **distribuția geometrică:** ex. 129 (+MAP, folosind distr. Beta);
- **distribuția Poisson:** ex. 46 (+MAP, folosind distr. Gamma);
- **distribuția uniformă continuă:** calcul de probabilități și MLE:
în \mathbb{R} : ex. 47, ex. 130, ex. 131; în \mathbb{R}^2 : ex. 48, ex. 132;
- **distribuția gaussiană unidimensională:**
MLE pt. μ , considerând σ^2 cunoscut: ex. 50 (+MAP, folosind distribuția gaussiană);
MLE pt. σ^2 , atunci când nu se impun restricții asupra lui μ : ex. 51 (+deplasare);
MLE pt. σ^2 , atunci când $\mu = 0$: ex. 134 (+nedeleplasare);
- **distribuția gaussiană multidimensională:** ex. 53, ex. 135 (+MAP, folosind distr. Gauss-Wishart);
- **distribuția exponentială:** ex. 49, ex. 133 (+MAP, folosind distr. Gamma);
- **distribuția Gamma:** ex. 52 și ex. 136 (ultimul, folosind metoda gradientului și metoda lui Newton);
- **existența și unicitatea MLE:** ex. 54;
- **MLE și parametrizarea alternativă:** ex. 137.

Elemente de teoria informației

- definiții și proprietăți imediate pentru entropie, entropie comună, entropie condițională specifică, entropie condițională medie, câștig de informație: ex. 55; exemplificarea acestor noțiuni (varianta discretă): ex. 56, ex. 57, ex. 138, ex. 139, ex. 140;
- exemple de calculare a entropiei unor variabile aleatoare continue: distribuția continuă uniformă (ex. 60.a), distribuția gaussiană unidimensională (ex. 60.b) și distribuția gaussiană multidimensională (ex. 60.c), distribuția exponentială (ex. 61.a), distribuția Gamma (ex. 61.b);
- o proprietate a entropiei: nenegativitatea: ex. 55.a, ex. 149.a;
- o margine superioară pentru valorea entropiei unei variabile aleatoare discrete: ex. 141;
- două proprietăți ale câștigului de informație: nenegativitatea (ex. 63.c, ex. 144) și $IG(X; Y) = 0 \Leftrightarrow$ variabila X este independentă de Y (ex. 63.c); pentru o demonstrație imediată a implicației directe, vedeți ex. 139.b;
- re-descoperirea definiției entropiei, pornind de la un set de proprietăți dezirabile: ex. 62;
- o aplicație pentru câștigul de informație: selecția de trăsături: ex. 58;
- entropia comună: exemplu de calculare: ex. 140.d
forma particulară a relației de „înlănțuire“ în cazul variabilelor aleatoare independente: ex. 59, ex. 142, ex. 149.c;
entropie comună și condiționată: formula de „înlănțuire“ condițională: ex. 143;
- entropia relativă (divergența Kullback-Leibler / KL):
definiție și proprietăți elementare: ex. 63.a;
exprimarea câștigului de informație cu ajutorul entropiei relative: ex. 63.b;
o proprietate de tip „relație de înlănțuire“: ex. 146;
echivalența dintre minimizarea entropiei relative și maximizarea funcției de log-verosimilitate: ex. 147;
- cross-entropie: definiție, o proprietate (nenegativitatea) și un exemplu simplu de calculare a valorii cross-entropiei: ex. 64;
un exemplu de aplicație pentru cross-entropie: selecția modelelor probabiliste: ex. 145;
- inegalitatea lui Gibbs: un caz particular; comparație între valorile entropiei și ale cross-entropiei: ex. 65;
- entropia văzută ca o funcțională în raport cu p ; calculul derivatei funcționale a entropiei în raport cu p : ex. 66;
- determinarea distribuțiilor probabiliste unidimensionale care — în anumite condiții — au entropii maxime, folosind metoda dualității Lagrange: ex. 148.

Funcții-nucleu

- aflarea funcției de „mapare“ a trăsăturilor care corespunde unei funcții-nucleu date: ex. 67, ex. 151, ex. 152.a;
comparații asupra numărului de operații efectuate la calcularea valorii unor funcții-nucleu (în spațiul inițial vs. spațiul nou de „trăsături“): ex. 152.b;
calculul distanțelor euclidiene în spații de „trăsături“ folosind doar funcții-nucleu: ex. 71;

- *teorema lui Mercer* (1909): condiții necesare și suficiente pentru ca o funcție să fie funcție-nucleu: ex. 68.ab;
- rezultate de tip „constructiv“ pentru [obținerea de noi] funcții-nucleu: ex. 68.c, 69, ex. 70, ex. 154, ex. 77.b; contraexemple: ex. 77.ac, ex. 155; „normalizarea“ funcțiilor-nucleu: ex. 153;
- o inegalitate [derivată din inegalitatea Cauchy-Buniakovski-Schwarz], care furnizează o margine superioară pentru $K(x, x')$, valoarea absolută a unei funcții-nucleu oarecare: ex. 156;
- un exemplu de funcție-nucleu care servește la a măsura similaritatea dintre două imagini oarecare: ex. 72;
- exemple de [funcții de] *mapare a atributelor* care asigură separabilitate liniară [în spațiul de trăsături] pentru orice set de instanțe de antrenament: ex. 73, ex. 157;
- funcția-nucleu gaussiană / *funcția cu baza radială* (engl., Radial Basis Function, RBF):
 - demonstrația faptului că RBF este într-adevăr funcție-nucleu: ex. 74;
 - funcția de „mapare“ corespunzătoare funcției-nucleu RBF ia valori într-un spațiu [de „trăsături“] de dimensiune infinită: ex. 75;
 - 2 proprietăți simple ale nucleului RBF: ex. 76, ex. 158;
 - RBF menține monotonia distanței euclidiene dintre perechi de instanțe:¹⁴ ex. 15 de la capitolul *Învățare bazată pe memorare*;
 - orice mulțime de instanțe distințe, având orice etichetare posibilă, este separabilă liniar în spațiul de „trăsături“ dacă se folosește nucleul RBF: ex. 26.a de la capitolul *Mașini cu vectori-suport* (aici parametrul nucleului RBF este ales în mod convenabil);¹⁵
- trecerea de la ne-separabilitate liniară [a unor exemple de antrenament] în *spațiul de intrare* la separabilitate liniară în *spațiul de trăsături*: ex. 150, ex. 159;
- recapitulare (A/F): ex. 161.

Metode de optimizare în învățarea automată

- (P0) definiții, caracterizări și câteva proprietăți pentru funcții convexe: ex. 78;
- analiza convexității unor funcții de cost folosite în învățarea automată [profundă]: ex. 163;
- inegalitatea lui Jensen: ex. 79;
- *metoda gradientului*, exemplificare: ex. 80.c, ex. 162, ex. 166, ex. 127, ex. 136.a; implementare: ex. 165.
subderivată, subdiferențială și subgradient: ex. 81, ex. 167;

¹⁴Relativ la k -NN vecinătăți, această proprietate se enunță astfel: pentru orice x, y și z din \mathbb{R}^d , are loc relația echivalență

$$\|x - y\| \leq \|x - z\| \Leftrightarrow \|\phi(x) - \phi(y)\| \leq \|\phi(x) - \phi(z)\|,$$

unde ϕ este funcția de „mapare“ corespunzătoare nucleului RBF.

¹⁵Problema 5.A de la capitolul *Mașini cu vectori-suport* arată că această proprietate este valabilă pentru orice valoare a parametrului nucleului RBF.

algoritmul Perceptron¹⁶ poate fi văzut ca o instanță a algoritmului subgradientului descendent ciclic (varianta stochastică): ex. 168;

metoda lui Newton, exemplificare: ex. 80.d, ex.127, ex.136.b;

(P1) condiții suficiente pentru convergența metodei gradientului: ex. 164;

(P2) o proprietate interesantă a metodei lui Newton: în cazul oricărei funcții de gradul al doilea (de una sau mai multe variabile), aplicarea acestei metode de optimizare implică / necesită execuția unei singure iterații: ex. 169;

(P3) *reparametrizarea liniară* a atributelor nu afectează [rezultatele obținute cu] metoda lui Newton, însă afectează metoda gradientului: ex. 170;

– *metoda dualității Lagrange*:

(P4) demonstrarea proprietății de *dualitate slabă*: ex. 82;

(P5) demonstrarea unei părți din *teorema Karush-Kuhn-Tucker*: ex. 83;

- exemple de aplicare: ex. 84, ex. 85, ex. 86, ex. 141, ex. 171, ex. 172, ex. 173, ex. 174;

- un exemplu de problemă de optimizare convexă pentru care condițiile Karush-Kuhn-Tucker nu sunt satisfăcute: ex. 175;

- folosirea metodei multiplicatorilor lui Lagrange la estimarea în sens MLE a parametrilor distribuției categoriale: ex. 44;

- folosirea metodei multiplicatorilor lui Lagrange pentru determinarea marginii superioare pentru valorile entropiei unei variabile aleatoare discrete: ex. 141;

- aplicarea metodei dualității Lagrange la determinarea distribuțiilor probabiliste unidimensionale care — în anumite condiții — au entropii maxime: ex. 148;

– două variante a algoritmului Perceptron,¹⁷ pentru care relația de actualizare a ponderilor se obține rezolvând [câte] o problemă de optimizare [convexă] cu restricții: ex. 87, ex. 176;

– *metoda descreșterii pe coordonate*, în contextul regresiei liniare cu regularizare L_1 : ex. 11 de la capitolul *Metode de regresie*;

metoda subgradientului, tot în contextul regresiei liniare cu regularizare L_1 : ex. 27 de la capitolul *Metode de regresie*;

– teorema de reprezentare: ex. 88, ex. 177.

¹⁶Vedeți pr. 16 de la capitolul *Rețele neuronale artificiale*.

¹⁷Vedeți problema 16 de la capitolul *Rețele neuronale artificiale*.

0.1 Fundamente — Probleme rezolvate

0.1.1 Evenimente aleatoare și formula lui Bayes

1. (Proprietăți derive din definiția funcției de probabilitate)
 • CMU, 2009 spring, Tom Mitchell, HW2, pr. 1.1

Fie două evenimente A și B .

a. Folosind doar proprietățile din definiția funcției de probabilitate, arătați că $P(A \cap \bar{B}) = P(A) - P(A \cap B)$.

b. Demonstrați *inegalitatea lui Bonferroni*: $P(A \cap B) \geq P(A) + P(B) - 1$.

c. Spunem că evenimentele A și B sunt *incompatibile* dacă $P(A \cap B) = 0$.

În ipoteza în care $P(A) = 1/3$ și $P(B) = 5/6$, este posibil ca evenimentele A și B să fie incompatibile? Justificați răspunsul.

Răspuns:

a. $A = A \cap \Omega = A \cap (B \cup \bar{B}) = (A \cap B) \cup (A \cap \bar{B})$.

Cum evenimentele $(A \cap B)$ și $(A \cap \bar{B})$, văzute ca multimi, sunt disjuncte, conform proprietății de „aditivitate numărabilă“ din definiția funcției de probabilitate putem scrie $P(A) = P(A \cap B) + P(A \cap \bar{B})$, deci $P(A \cap \bar{B}) = P(A) - P(A \cap B)$.

b. Întrucât $A \cup B = (A \cap \bar{B}) \cup (A \cap B) \cup (\bar{A} \cap B)$, este imediat că $P(A \cup B) = P(A \cap \bar{B}) + P(A \cap B) + P(\bar{A} \cap B)$. Deci,

$$\begin{aligned} P(A \cap B) &= P(A \cup B) - P(A \cap \bar{B}) - P(\bar{A} \cap B) \\ &= [P(A \cup B) - P(A \cap \bar{B})] + [P(A \cup B) - P(\bar{A} \cap B)] - P(A \cup B) \\ &= P(A) + P(B) - P(A \cup B). \end{aligned}$$

Cum $P(A \cup B) \leq 1$ (fiind o probabilitate), putem scrie că: $P(A \cap B) = P(A) + P(B) - P(A \cup B) \geq P(A) + P(B) - 1$, deci $P(A \cap B) \geq P(A) + P(B) - 1$.

Observație: Din egalitatea $P(A \cap B) = P(A) + P(B) - P(A \cup B)$ dedusă mai sus rezultă

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Această egalitate se numește *regula de adunare / însumare a probabilităților* (engl., additivity rule).

c. Pentru a studia posibilitatea ca evenimentele A și B să fie incompatibile vom aplica inegalitatea lui Bonferroni:

$$P(A) + P(B) = \frac{1}{3} + \frac{5}{6} = \frac{7}{6} \Rightarrow P(A \cap B) \geq \frac{1}{6} > 0$$

Deci $P(A \cap B)$ nu poate fi 0 și, în consecință, evenimentele A și B nu sunt incompatibile.

2. (Calcularea de probabilități elementare și probabilități condiționate)

• CMU, 2009 spring, Tom Mitchell, HW2, pr. 1.4

În acest exercițiu vom arăta că probabilitatea unui eveniment se poate schimba (într-un anumit sens) dacă știm probabilitatea unui alt eveniment, legat de cel dintâi.

Se aruncă simultan două zaruri. Notăm cu S variabila aleatoare care desemnează suma numerelor rezultate din aruncarea celor două zaruri.

- Calculați $P(S = 11)$.
- Dacă știm că S este număr prim, cât devine probabilitatea de mai sus?

Răspuns:

- Probabilitatea unui eveniment precum cel din enunț este dată de raportul dintre numărul de cazuri favorabile și numărul cazurilor posibile.

La aruncarea simultană a două zaruri, fiecare zar poate cădea pe una dintre cele 6 fețe, independent de celălalt zar. Prin urmare, există $6 \cdot 6 = 36$ posibilități (cazuri posibile). Pentru a se obține $S = 11$ cele două zaruri trebuie să aibă fie valorile (5, 6), fie (6, 5), deci există 2 posibilități (cazuri favorabile). Prin urmare,

$$P(S = 11) = \frac{2}{36} = \frac{1}{18}$$

- Probabilitatea căutată este:

$$P(S = 11 | S = \text{prim}) = \frac{P(S = 11 \cap S = \text{prim})}{P(S = \text{prim})} = \frac{P(S = 11)}{P(S = \text{prim})}$$

Trebuie calculată probabilitatea ca S să fie număr prim. Pentru aceasta este necesar numărul de cazuri favorabile, adică numărul cazurilor pentru care $S \in \{2, 3, 5, 7, 11\}$. Există următoarele 15 posibilități:

- $S = 2 : (1, 1)$
- $S = 3 : (1, 2), (2, 1)$
- $S = 5 : (1, 4), (4, 1), (2, 3), (3, 2)$
- $S = 7 : (1, 6), (6, 1), (2, 5), (5, 2), (3, 4), (4, 3)$
- $S = 11 : (5, 6), (6, 5)$

Deci $P(S = \text{prim}) = 15/36$. Prin urmare,

$$P(S = 11 | S = \text{prim}) = \frac{2/36}{15/36} = \frac{2}{15}$$

Întrucât $2/15 > 1/18$, putem spune că probabilitatea evenimentului $S = 11$ a crescut după ce am aflat un fapt suplimentar, și anume că suma rezultată la aruncarea celor două zaruri este un număr prim ($S = \text{prim}$).

Ca o observație cu caracter general: dacă $A \subseteq B$ (așa cum a fost cazul în această problemă), atunci rezultă imediat că $P(A) \leq P(A|B)$.

3.

(Spațiu de eşantionare – exemplificare; calcul de probabilități condiționate)

- CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW1, pr. 1.4

O cutie conține trei carduri. Un card este roșu pe ambele părți, un alt card este verde pe ambele părți, iar cel care a rămas este roșu pe o parte și verde pe partea cealaltă. Selectăm în mod aleatoriu un card din această cutie; presupunem că nu-i vedem decât culoarea de pe fața superioară. Dacă această față este verde, care este probabilitatea ca și cealaltă față să fie verde?

Răspuns:

Pentru a rezolva această problemă este foarte util să stabilim mai întâi care sunt elementele ce compun *spațiul de eşantionare* (engl., sample space), Ω .¹⁸ Contra intuitiei primare, Ω nu va fi constituit din cele trei carduri, ci din fețele lor, fiindcă ceea ce observăm după o extragere este doar o față a unui card, nu ambele fețe ale cardului extras.

Din punct de vedere formal, vom folosi următoarea *notăție* pentru carduri:

$$C_1 = (R1, R2), C_2 = (R3, V4), C_3 = (V5, V6),$$

unde $R1, R2, R3, V4, V5$ și $V6$ desemnează cele 6 fețe ale cardurilor. Așadar, $\Omega = \{R1, R2, R3, V4, V5, V6\}$.

Observație: Am fi putut nota fețele cardurilor folosind pur și simplu numerele $1, \dots, 6$, însă am preferat să însotim fiecare dintre aceste numere cu litera R sau V care desemnează culoarea feței respective.

După ce am făcut această pregătire, probabilitatea cerută în enunțul problemei se calculează simplu, folosind regula clasice: $p = m/n$, unde m este numărul de cazuri favorabile, iar n este numărul de cazuri posibile.

Cazurile posibile sunt $V4, V5, V6$, iar cazurile favorabile sunt $V5$ și $V6$ deoarece doar pentru ele fața cealaltă a cardului este verde (este vorba de $V6$ și respectiv $V5$). Așadar, probabilitatea cerută este $\frac{2}{3}$.

4.

(Evenimente aleatoare independente)

- CMU, 2009 spring, Tom Mitchell, HW2, pr. 1.2.1
- CMU, 2009 spring, Ziv Bar-Joseph, HW1, pr. 1.1

Două evenimente A și B sunt independente statistic dacă $P(A \cap B) = P(A) \cdot P(B)$.

a. Arătați că dacă A și B sunt evenimente independente, atunci:

- A și \bar{B} sunt independente;
- \bar{A} și \bar{B} sunt independente.

b. Dacă evenimentul A este independent în raport cu el însuși, ce puteți spune despre $P(A)$?

Răspuns:

¹⁸A se vedea noțiunea prezentată la curs.

a. $P(A \cap \bar{B}) = P(A) - P(A \cap B) = P(A) - P(A) \cdot P(B) = P(A) \cdot (1 - P(B)) = P(A) \cdot P(\bar{B})$, deci A și \bar{B} sunt independente. (S-a folosit independenta evenimentelor A și B .)

Pentru independenta evenimentelor \bar{A} și \bar{B} se procedează analog: $P(\bar{A} \cap \bar{B}) = P(\bar{B}) - P(A \cap \bar{B}) = P(\bar{B}) - P(A) \cdot P(\bar{B}) = P(\bar{B}) \cdot (1 - P(A)) = P(\bar{A}) \cdot P(\bar{B})$, deci \bar{A} și \bar{B} sunt independente. (La cea de-a doua egalitate s-a folosit independenta evenimentelor A și \bar{B} demonstrată mai sus.)

b. Condiția de independentă a evenimentului A față de el însuși se scrie astfel:

$$P(A \cap A) = P(A) \cdot P(A) \Rightarrow P(A) = [P(A)]^2 \Rightarrow P(A)[P(A) - 1] = 0$$

Tinând cont și de restricția $P(A) \in [0, 1]$, rezultă că $P(A) = 0$ sau $P(A) = 1$. (Atenție: nu rezultă neapărat că $A = \emptyset$ respectiv $A = \Omega$!)

5. (Evenimente aleatoare independente; aplicarea proprietăților din definiția funcției de probabilitate)

• CMU, 2009 spring, Tom Mitchell, HW2, pr. 1.2.3

Robert și Alina dau cu banul alternativ. Cel dintâi dintre ei care va obține stema (în engleză: head) câștigă jocul. Alina este prima care va da cu banul.

a. Dacă $P(stemă) = 1/2$, care este probabilitatea ca Alina să câștige jocul?

Indicație (1): Încercați să enumerați toate situațiile în care Alina poate câștiga.

Indicație (2): Pentru orice $a \in [0, 1]$, avem $\sum_{i=0}^{i=\infty} a^i = 1 + a + a^2 + \dots + a^n + \dots = \lim_{n \rightarrow +\infty} \frac{1 - a^n}{1 - a} = \frac{1}{1 - a}$.

b. Dacă $P(stemă) = p \in (0, 1]$, care este probabilitatea ca Alina să câștige jocul?

c. Tinând cont de expresia obținută la punctul b, dacă ar fi ca tu să joci acest joc, cum ai decide să intri în joc: primul ori al doilea (presupunând, bineînteles, că ai avea posibilitatea să alegi)?

Răspuns:

a. Alina aruncă moneda în „pașii“ impari. Ea câștigă jocul dacă la pasul $2n+1$ obține stema și până la pasul respectiv nimeni nu a obținut stema.

Notăm cu A evenimentul ca Alina să câștige jocul și cu A_i evenimentul ca Alina să câștige jocul la a i -a aruncare a banului. Întrucât evenimentele A_i , văzute ca mulțimi, sunt mutual disjuncte ($A_i \cap A_j = \emptyset$ pentru orice $i \neq j$) și $A = A_1 \cup A_3 \cup \dots$, conform proprietății de aditivitate numărabilă din definiția funcției de probabilitate rezultă că

$$P(A) = P(A_1) + P(A_3) + P(A_5) + \dots$$

Întrucât $p = \frac{1}{2}$, ținând cont [și] de faptul că toate aruncările sunt independente, probabilitățile corespunzătoare evenimentelor A_i se calculează astfel:

$$\begin{aligned} P(A_1) &= P(\text{stemă}) = \frac{1}{2} \\ P(A_3) &= (1 - P(\text{stemă})) \cdot (1 - P(\text{stemă})) \cdot P(\text{stemă}) = \left(\frac{1}{2}\right)^3 \\ P(A_5) &= (1 - \frac{1}{2}) \cdot (1 - \frac{1}{2}) \cdot (1 - \frac{1}{2}) \cdot (1 - \frac{1}{2}) \cdot P(\text{stemă}) = \left(\frac{1}{2}\right)^5 \\ &\dots \\ P(A_{2i+1}) &= (1 - P(\text{stemă}))^{2i} \cdot P(\text{stemă}) = \left(\frac{1}{2}\right)^{2i+1} \end{aligned}$$

Prin urmare,

$$\begin{aligned} P(A) &= \sum_{i=0}^{\infty} P(A_{2i+1}) = \sum_{i=0}^{\infty} \left(\frac{1}{2}\right)^{2i+1} = \sum_{i=0}^{\infty} \left(\frac{1}{2}\right)^{2i} \cdot \frac{1}{2} = \frac{1}{2} \cdot \sum_{i=0}^{\infty} \left(\frac{1}{4}\right)^i = \frac{1}{2} \cdot \frac{1}{1 - 1/4} \\ &= \frac{1}{2} \cdot \frac{4}{3} = \frac{2}{3}. \end{aligned}$$

Așadar, pentru $p = 1/2$ probabilitatea ca Alina să câștige jocul este $2/3$ (deci de două ori mai mare decât probabilitatea ca Robert să câștige jocul), pur și simplu datorită faptului că ea este prima care dă cu banul. (Avantajul primului jucător!)

b. Dacă $P(\text{stemă}) = p \in (0, 1]$, se urmează același raționament ca mai sus, cu modificarea valorilor $P(A_i)$ astfel:

$$\begin{aligned} P(A_1) &= P(\text{stemă}) = p \\ P(A_3) &= (1 - P(\text{stemă})) \cdot (1 - P(\text{stemă})) \cdot P(\text{stemă}) = (1 - p)^2 \cdot p \\ P(A_5) &= (1 - p) \cdot (1 - p) \cdot (1 - p) \cdot (1 - p) \cdot P(\text{stemă}) = (1 - p)^4 \cdot p \\ &\dots \\ P(A_{2i+1}) &= (1 - P(\text{stemă}))^{2i} \cdot P(\text{stemă}) = (1 - p)^{2i} \cdot p \end{aligned}$$

Prin urmare, probabilitatea ca Alina să câștige devine:

$$P(A) = \sum_{i=0}^{\infty} P(A_{2i+1}) = \sum_{i=0}^{\infty} (1-p)^{2i} \cdot p = p \cdot \sum_{i=0}^{\infty} (1-p)^{2i} \stackrel{p \neq 0}{=} p \cdot \frac{1}{1 - (1-p)^2} = \frac{p}{2p - p^2} = \frac{1}{2 - p}$$

c. Jucătorul care aruncă primul banul câștigă jocul cu o probabilitate de $1/(2-p)$ (calculată mai sus). Cum $p \geq 0$, rezultă că $\frac{1}{2-p} \geq \frac{1}{2}$. Așadar, în cazul în care există posibilitatea de a alege, este de preferat să arunci primul.

6.

(Formula lui Bayes)

• CMU, 2001 fall, Andrew Moore, midterm exam, pr. 5

Se consideră două variabile aleatoare A și B despre care știm următoarele informații:

- a. $P(A | B) = 2/3$
- b. $P(A | B) = 2/3$ și $P(A | \bar{B}) = 1/3$
- c. $P(A | B) = 2/3$, $P(A | \bar{B}) = 1/3$ și $P(B) = 1/3$
- d. $P(A | B) = 2/3$, $P(A | \bar{B}) = 1/3$, $P(B) = 1/3$ și $P(A) = 4/9$.

În care din cele patru cazuri informațiile date sunt suficiente pentru a calcula $P(B | A)$? Există vreun caz în care apar informații superflue (i.e., informații care pot fi deduse din celelalte informații furnizate în cazul respectiv)?

Răspuns:

Conform teoremei lui Bayes, vom avea:

$$\begin{aligned} P(B | A) &= \frac{P(A | B) \cdot P(B)}{P(A)} = \frac{P(A | B) \cdot P(B)}{P(A | B) \cdot P(B) + P(A | \bar{B}) \cdot P(\bar{B})} \\ &= \frac{P(A | B) \cdot P(B)}{P(A | B) \cdot P(B) + P(A | \bar{B}) \cdot (1 - P(B))} \end{aligned}$$

Devine astfel evident că în cazurile c și d informațiile deținute sunt suficiente, pe când în celelalte două cazuri nu. În cazul d, informația $P(A) = 4/9$ este superfluă.

7.

(Formula lui Bayes)

• o CMU, 2008 spring, T. Mitchell, W. Cohen, midterm, pr. 1.5

Presupunem că, răspunzând la o întrebare cu răspuns de genul adevărat / fals, un student fie cunoaște răspunsul, fie ghicește răspunsul. Probabilitatea de a cunoaște răspunsul este p , iar probabilitatea de a ghici răspunsul este $1 - p$.

Presupunem că probabilitatea de a răspunde corect la întrebare este

- 1 în cazul în care studentul cunoaște răspunsul
- și 0.5 dacă studentul ghicește răspunsul.

Exprimăți în funcție de p care este probabilitatea ca studentul examinat să cunoască răspunsul la întrebare, în ipoteza că el a răspuns corect (notăție: $P(knew | correct)$).

Răspuns:

Evenimentele de interes în problema dată sunt:

correct = studentul a răspuns corect la întrebare,

knew = studentul cunoștea răspunsul corect

și complementarul acestuia din urmă:

guess $\stackrel{\text{not.}}{=} \neg knew$ = studentul ghicește răspunsul.

Aplicând formula lui Bayes obținem:

$$\begin{aligned} P(knew | correct) &= \frac{P(correct | knew) \cdot P(knew)}{P(correct | knew) \cdot P(knew) + P(correct | guess) \cdot P(guess)} \\ &= \frac{1 \cdot p}{1 \cdot p + 0.5 \cdot (1 - p)} = \frac{p}{0.5p + 0.5} = \frac{p}{0.5(p + 1)} = \frac{2p}{p + 1} \end{aligned}$$

8.

(Probabilități, chestiuni elementare:
Adevărat sau Fals?) • CMU, 2016 fall, N. Balcan, M. Gormley, HW1, pr. 6.1.1-4

În cele de mai jos vom nota cu Ω spațiul de eșantionare, iar cu \bar{A} complementul evenimentului A .

Marcați cu *adevărat* sau *fals* fiecare dintre afirmațiile următoare:

- a. Pentru orice $A, B \subseteq \Omega$ astfel încât $P(A) > 0$ și $P(B) > 0$, are loc egalitatea $P(A|B)P(B) = P(B|A)P(A)$.
- b. Pentru orice $A, B \subseteq \Omega$ astfel încât $P(B) > 0$, are loc egalitatea $P(A \cup B) = P(A) + P(B) - P(A|B)$.
- c. Pentru orice $A, B, C \subseteq \Omega$ astfel încât $P(B \cup C) > 0$, urmează că $\frac{P(A \cup B \cup C)}{P(B \cup C)} \geq P(A|B \cup C)P(B \cup C)$
- d. Pentru orice $A, B \subseteq \Omega$ astfel încât $P(A) > 0$ și $P(\bar{A}) > 0$, urmează că $P(B|A) + P(B|\bar{A}) = 1$.

Indicație:

Pentru fiecare afirmație adevărată, faceți demonstrația proprietății respective. Pentru fiecare afirmație falsă, dați fie un contraexemplu fie o justificare riguroasă.

Răspuns:

a. Adevărat. Demonstrația este imediată.

b. Fals.

Stim că pentru orice $A, B \subseteq \Omega$, are loc egalitatea $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Așadar, egalitatea din enunț ar fi adevărată dacă pentru orice $A, B \subseteq \Omega$ am avea $P(A \cap B) = P(A|B)$. Însă,

$$P(A \cap B) = P(A|B) \Leftrightarrow P(A \cap B) = \frac{P(A \cap B)}{P(B)} \Leftrightarrow P(B) = 1 \text{ sau } P(A \cap B) = 0.$$

Deci egalitatea nu este adevărată pentru orice A și B .

c. Adevărat.

Observăm că dacă se notează $D = B \cup C$, atunci inegalitatea din enunț devine mai ușor de manipulat: $\frac{P(A \cup D)}{P(D)} \geq P(A|D) \cdot P(D)$. Este imediat că ea este echivalentă cu $\frac{P(A \cup D)}{P(D)} \geq P(A \cap D)$, ceea ce implică $P(A \cup D) \geq P(A \cap D) \cdot P(D)$.

Ultima inegalitate este adevărată fiindcă pe de o parte $P(A \cup D) \geq P(A \cap D)$ și pe de altă parte $P(D) \in [0, 1]$.

d. Fals.

Stim că pentru orice $A, B \subseteq \Omega$ cu $P(A) > 0$, are loc egalitatea $P(B|A) + P(\bar{B}|A) = 1$ (o puteți demonstra imediat). Așadar, egalitatea din enunț ar fi adevărată dacă pentru orice $A, B \subseteq \Omega$ astfel încât $P(A) > 0$ și $P(\bar{A}) > 0$ am avea $P(B|\bar{A}) = P(\bar{B}|A)$.

Vom construi următorul contraexemplu: considerăm că la aruncarea unui zar perfect, evenimentul A este apariția unei fețe pare, iar evenimentul B este apariția feței 1. Urmează că

$$P(B|\bar{A}) = \frac{1}{3} \text{ și } P(\bar{B}|A) = 1.$$

Deci egalitatea nu este adevărată pentru orice evenimente A (cu $P(A) > 0$ și $P(\bar{A}) > 0$) și B .

0.1.2 Variabile aleatoare

9. (Variabile aleatoare: proprietăți de bază pentru medii, varianță, covarianță)

Fie variabila aleatoare $X : \Omega \rightarrow \mathbb{R}$, cu funcția de probabilitate P .

Dacă X este variabilă aleatoare *discretă*, atunci prin definiție $P(x) \stackrel{\text{not.}}{=} P(X = x) \stackrel{\text{not.}}{=} P(\{\omega \mid X(\omega) = x\}) \geq 0$ pentru orice $x \in \mathbb{R}$, și $\sum_{x_i \in \text{Val}(X)} P(x_i) = 1$, unde $\text{Val}(X)$ este mulțimea valorilor variabilei aleatoare X .

Dacă X este variabilă aleatoare *continuă*, având funcția densitate de probabilitate p , atunci prin definiție $p(X = x) \geq 0$ pentru orice $x \in \mathbb{R}$, și $\int_{-\infty}^{\infty} p(X = x)dx = 1$ (sau, scris mai simplu: $\int_{-\infty}^{+\infty} p(x)dx = 1$).

- a. ■ CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh, HW1, pr. 1.1

Dacă X este variabilă aleatoare discretă, *media* sa se definește ca fiind numărul real $E[X] = \sum_{x_i \in \text{Val}(X)} x_i \cdot P(X = x_i)$. Dacă X este variabilă aleatoare continuă, media sa este $E[X] = \int_{-\infty}^{\infty} x \cdot p(X = x)dx$.

Arătați că pentru orice două variabile aleatoare W și Z de același tip (adică fie ambele discrete fie ambele continue), având același domeniu de definiție (Ω), avem

$$E[W + Z] = E[W] + E[Z].$$

De asemenea, demonstrați că pentru orice constantă $a \in \mathbb{R}$, are loc egalitatea

$$E[aX] = aE[X]. \quad (1)$$

Notați că aX este o variabilă aleatoare definită pe același domeniu (Ω) ca și variabila X , cu proprietatea că $(aX)(\omega) \stackrel{\text{def.}}{=} aX(\omega)$ pentru orice $\omega \in \Omega$.

Observație: Cele două egalități de mai sus se pot combina sub o formă mai generală: pentru orice variabile aleatoare (fie toate discrete fie toate continue) X_1, \dots, X_n și pentru orice constante $a_1, \dots, a_n \in \mathbb{R}$, cu $n \geq 1$, are loc egalitatea

$$E[a_1X_1 + \dots + a_nX_n] = a_1E[X_1] + \dots + a_nE[X_n]. \quad (2)$$

Această egalitate este cunoscută sub numele de *proprietatea de liniaritate a mediei*.

b. *CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh, HW1, pr. 1.3*

Fie X o variabilă aleatoare. Notăm $\bar{X} = E[X]$. Varianța lui X se definește ca fiind $Var(X) = E[(X - \bar{X})^2]$. Arătați că:

$$Var(X) = E[X^2] - (E[X])^2.$$

Observație importantă: Veți vedea că această proprietate este adeseori folosită (în locul definiției varianței) în diverse demonstrații care vor urma.

De asemenea, demonstrați că pentru orice constantă $a \in \mathbb{R}$, are loc egalitatea

$$Var(aX) = a^2 Var(X). \quad (3)$$

Prin urmare, în cazul varianței nu avem o proprietate de liniaritate similară cu cea din cazul mediei.

Indicație: La acest punct nu este necesar să faceți demonstrațiile separat pentru cele două cazuri, discret și respectiv continuu.

c. *CMU, 2009 spring, Tom Mitchell, HW2, pr. 1.3.1*

Covarianța a două variabile aleatoare X și Y care au același domeniu de definiție (Ω) se definește astfel: $Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$, unde $E[X]$ este media lui X . (Este imediat faptul că noțiunea de covarianță generalizează noțiunea de varianță.)

Demonstrați egalitatea:

$$Cov(X, Y) = E[XY] - E[X] \cdot E[Y]. \quad (4)$$

Consecință imediată (din relațiile (1) și (4)):

$$Cov(aX, bY) = ab Cov(X, Y) \quad \forall a, b \in \mathbb{R}. \quad (5)$$

Observații:

1. Este imediat faptul că proprietatea (5) generalizează proprietatea (3).
2. Spre deosebire de varianță, care poate lua doar valori mai mari sau egale cu 0 (ceea ce decurge imediat din definiția de la punctul b), covarianța poate lua și valori negative. Mai mult, la problema 19 se va demonstra că atunci când $Var(X) \neq 0$ și $Var(Y) \neq 0$, avem următoarele margini (una inferioară și cealaltă superioară) pentru $Cov(X, Y)$:¹⁹

$$-\sqrt{Var(X) \cdot Var(Y)} \leq Cov(X, Y) \leq +\sqrt{Var(X) \cdot Var(Y)}. \quad (6)$$

Răspuns:

- a. Pentru cazul discret²⁰ vom folosi o formă echivalentă a formulei pentru media unei variabile aleatoare și anume $E[X] = \sum_{\omega \in \Omega} X(\omega) \cdot P(\omega)$. Prin urmare, putem scrie:

¹⁹ Atunci când $Var(X) = 0$ sau $Var(Y) = 0$, rezultă că variabila X (respectiv Y) este funcție constantă (fie peste tot domeniul de definiție, fie cu excepția unei mulțimi de probabilitate 0), ceea ce implică imediat $Cov(X, Y) = 0$. Așadar, dubla inegalitate (6) este satisfăcută și în acest caz.

²⁰ *Observație importantă:* În demonstrația care urmează — și, de asemenea, în toate problemele din acest capitol în care se fac / cer demonstrații pentru variabile aleatoare discrete — se presupune că mulțimile de valori ale acestor variabile sunt finite. Așadar, tratăm doar cazul discret finit.

$$\begin{aligned}
E[W + Z] &= \sum_{u \in Val(W+Z)} u \cdot P(W + Z = u) \\
&= \sum_{w \in Val(W), z \in Val(Z)} (w + z) \cdot P(W + Z = w + z) \\
&= \sum_{w \in Val(W)} \sum_{z \in Val(Z)} (w + z) \cdot P(\{\omega \in \Omega \mid (W + Z)(\omega) = w + z\}) \\
&= \sum_{w \in Val(W)} \sum_{z \in Val(Z)} w \cdot P(\{\omega \in \Omega \mid (W + Z)(\omega) = w + z\}) + \\
&\quad \sum_{w \in Val(W)} \sum_{z \in Val(Z)} z \cdot P(\{\omega \in \Omega \mid (W + Z)(\omega) = w + z\}) \\
&= \sum_{w \in Val(W)} w \sum_{z \in Val(Z)} P(\{\omega \in \Omega \mid W(\omega) = w, Z(\omega) = z\}) + \\
&\quad \sum_{z \in Val(Z)} z \sum_{w \in Val(W)} P(\{\omega \in \Omega \mid W(\omega) = w, Z(\omega) = z\}) \\
&= \sum_{w \in Val(W)} w P(\{\omega \in \Omega \mid W(\omega) = w\}) + \sum_{z \in Val(Z)} z P(\{\omega \in \Omega \mid Z(\omega) = z\}) \\
&= E[W] + E[Z].
\end{aligned}$$

Pentru cazul continuu:

$$\begin{aligned}
E[W + Z] &= \int_w \int_z (w + z) p_{WZ}(w, z) dz dw \\
&= \int_w \int_z w p_{WZ}(w, z) dz dw + \int_w \int_z z p_{WZ}(w, z) dz dw \\
&= \int_w w \int_z p_{WZ}(w, z) dz dw + \int_z z \int_w p_{WZ}(w, z) dw dz \\
&= \int_w w p_W(w) dw + \int_z z p_Z(z) dz = E[W] + E[Z].
\end{aligned}$$

În cazul în care variabila aleatoare X este discretă, egalitatea $E[aX] = aE[X]$ se poate demonstra imediat, aplicând definiția mediei. Vom considera mai întâi subcazul $a \neq 0$:

$$\begin{aligned}
E[aX] &= \sum_{u \in Val(aX)} u P(aX = u) = \sum_{x \in Val(X)} ax P(X = x), \quad \text{unde } x = \frac{1}{a}u \\
&= a \sum_{x \in Val(X)} x P(X = x) = aE[X].
\end{aligned}$$

Pentru subcazul $a = 0$, egalitatea $E[aX] = aE[X]$ este trivială.

Similar se face demonstrația egalității $E[aX] = aE[X]$ în cazul în care variabila aleatoare X este continuă.

b. Pentru a demonstra această proprietate vom ține cont de liniaritatea mediei (vedeți *Observația de la punctul a*):

$$\begin{aligned}
Var(X) &= E[(X - \bar{X})^2] = E[X^2 - 2X\bar{X} + \bar{X}^2] \\
&= E[X^2] - E[2X\bar{X}] + E[\bar{X}^2] = E[X^2] - 2\bar{X}E[X] + \bar{X}^2 \\
&= E[X^2] - 2\bar{X}^2 + \bar{X}^2 = E[X^2] - \bar{X}^2 = E[X^2] - (E[X])^2.
\end{aligned}$$

Egalitatea $\text{Var}(aX) = a^2 \text{Var}(X)$ rezultă imediat, aplicând definiția varianței și ținând cont de proprietatea $E[aX] = aE[X]$ pe care am demonstrat-o la punctul a.

$$\begin{aligned}\text{Var}(aX) &\stackrel{\text{def.}}{=} E[(aX - \underbrace{E[aX]}_{aE[X]})^2] = E[(a(X - E[X]))^2] \\ &= E[a^2(X - E[X])^2] = a^2E[(X - E[X])^2] \stackrel{\text{def.}}{=} a^2 \text{Var}(X).\end{aligned}$$

Alternativ, putem folosi proprietatea pe care am demonstrat-o mai sus:

$$\begin{aligned}\text{Var}(aX) &= E[(aX)^2] - (E[aX])^2 = E[a^2X^2] - (aE[X])^2 = a^2E[X^2] - a^2(E[X])^2 \\ &= a^2(E[X^2] - (E[X])^2) = a^2 \text{Var}(X).\end{aligned}$$

c. Se folosește același gen de raționament ca la punctul precedent (prima parte):

$$\begin{aligned}\text{Cov}(X, Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY - XE[Y] - E[X]Y + E[X]E[Y]] \\ &= E[XY] - E[XE[Y]] - E[E[X]Y] + E[E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] = E[XY] - E[X]E[Y].\end{aligned}$$

10.

(Rezultat teoretic:
covarianța oricărora 2 variabile aleatoare independente este 0)

■ CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh, HW1, pr. 1.2

În mod intuitiv, două variabile aleatoare X și Y sunt *independente* atunci când cunoașterea valorii uneia dintre ele (de exemplu X) nu furnizează niciun indiciu despre valoarea celeilalte variabile (Y , în acest caz).

Formal, dacă X și Y sunt variabile aleatoare discrete, independența lor revine la egalitatea $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$ pentru orice $x \in \text{Val}(X)$ și orice $y \in \text{Val}(Y)$.

Similar, dacă X și Y sunt variabile aleatoare continue, atunci $p(X = x, Y = y) = p(X = x) \cdot p(Y = y)$ pentru orice valori x și y posibile.

Arătați că dacă X și Y sunt variabile aleatoare independente de același tip (adică fie discret fie continuu), atunci

$$E[XY] = E[X] \cdot E[Y].$$

Echivalent: X, Y independente $\Rightarrow \text{Cov}(X, Y) = 0$. (A se vedea problema 9 punctul c.)

Observație: Reciproca implicației de mai sus nu este în general adevărată. A se vedea problema 11.

Răspuns:

Pentru cazul în care variabilele aleatoare independente X și Y sunt discrete, putem scrie:

$$\begin{aligned}
E[XY] &= \sum_{x \in Val(X)} \sum_{y \in Val(Y)} xyP(X = x, Y = y) \\
&\stackrel{indep.}{=} \sum_{x \in Val(X)} \sum_{y \in Val(Y)} xyP(X = x) \cdot P(Y = y) \\
&= \sum_{x \in Val(X)} \left(xP(X = x) \sum_{y \in Val(Y)} yP(Y = y) \right) \\
&= \sum_{x \in Val(X)} xP(X = x)E[Y] = \left(\sum_{x \in Val(X)} xP(X = x) \right) E[Y] = E[X] \cdot E[Y]
\end{aligned}$$

Pentru cazul continuu demonstrația este similară:

$$\begin{aligned}
E[XY] &= \int_x \int_y xy p(X = x, Y = y) dy dx \\
&\stackrel{indep.}{=} \int_x \int_y xy p(X = x) \cdot p(Y = y) dy dx \\
&= \int_x x p(X = x) \int_y y p(Y = y) dy dx \\
&= \int_x x p(X = x) E[Y] dx = E[Y] \cdot \int_x x p(X = x) dx \\
&= E[Y] \cdot E[X] = E[X] \cdot E[Y]
\end{aligned}$$

11. (Covarianța nulă nu implică în mod neapărat independența variabilelor aleatoare)

*CMU, 2009 spring, Tom Mitchell, HW2, pr. 1.3.2
CMU, 2009 fall, Geoff Gordon, HW1, pr. 3.1*

a. Reciproca afirmației din problema 10 nu este în general adevărată, deci $Cov(X, Y) = 0 \not\Rightarrow X$ și Y sunt independente. Arătați aceasta folosind ca exemplu variabilele aleatoare ale căror distribuții sunt date în tabelul alăturat.

b. Totuși, dacă X și Y sunt variabile aleatoare binare luând valori în mulțimea $\{0, 1\}$, iar $E[XY] = E[X] \cdot E[Y]$, atunci X și Y sunt independente. Justificați.

(Așadar, două variabile aleatoare binare cu valori în mulțimea $\{0, 1\}$ sunt independente atunci și numai atunci când au covarianță nulă.)

X	Y	$P(X, Y)$
0	0	1/3
1	0	0
2	0	1/3
0	1	0
1	1	1/3
2	1	0

Răspuns:

a. Din datele furnizate în exemplul a rezultă următoarele probabilități:

$$\begin{aligned}
P(X = 0) &= 1/3 & P(Y = 0) &= 2/3 & P(XY = 0) &= 2/3 \\
P(X = 1) &= 1/3 & P(Y = 1) &= 1/3 & P(XY = 1) &= 1/3 \\
P(X = 2) &= 1/3 & & & P(XY = 2) &= 0.
\end{aligned}$$

Putem calcula mediile acestor variabile aleatoare folosind formula de definiție pentru variabile discrete $E[X] = \sum_x x \cdot P(X = x)$:

$$\begin{aligned} E[X] &= 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + 2 \cdot \frac{1}{3} = 1 \\ E[Y] &= 0 \cdot \frac{2}{3} + 1 \cdot \frac{1}{3} = \frac{1}{3} \\ E[XY] &= 0 \cdot \frac{2}{3} + 1 \cdot \frac{1}{3} + 2 \cdot 0 = \frac{1}{3}. \end{aligned}$$

Conform formulei care a fost demonstrată la problema 10, covarianța variabilelor X și Y este: $Cov(X, Y) = E[XY] - E[X] \cdot E[Y] = \frac{1}{3} - 1 \cdot \frac{1}{3} = 0$. Cu toate acestea, variabilele X și Y nu sunt independente. Într-adevăr, pentru $X = 0$ și $Y = 0$ se observă că $P(X = 0, Y = 0) = 1/3$ dar $P(X = 0)P(Y = 0) = 1/3 \cdot 2/3 = 2/9 \neq \frac{1}{3}$.

Prin acest contraexemplu am arătat că implicația „ $Cov(X, Y) = 0 \Rightarrow X$ și Y sunt independente“ nu este în general adevărată.

b. În continuare vom demonstra că în cazul în care X și Y iau valori în mulțimea $\{0, 1\}$ implicația de mai sus este adevărată.

Dacă X și Y sunt variabile aleatoare binare, atunci:

$$\begin{aligned} E[X] &= 0 \cdot P(X = 0) + 1 \cdot P(X = 1) = P(X = 1) \\ E[Y] &= P(Y = 1) \\ E[XY] &= P(X = 1, Y = 1). \end{aligned}$$

Covarianța nulă înseamnă — conform problemei 9.c — că $E[XY] = E[X] \cdot E[Y]$, adică:

$$P(X = 1, Y = 1) = P(X = 1)P(Y = 1).$$

Însă, a demonstra independența variabilelor aleatoare X și Y revine la a arăta că $P(X, Y) = P(X)P(Y)$ pentru toate combinațiile posibile de valori ale variabilelor. Unul din cazuri ($X = 1, Y = 1$) este deja demonstrat, deci mai există încă 3 cazuri. Pentru acestea vom utiliza formulele: $P(A \cap B) = P(A) - P(A \cap \bar{B})$ și $P(\bar{A}) = 1 - P(A)$.

Cazul $X = 1, Y = 0$:

$$\begin{aligned} P(X = 1, Y = 0) &= P(X = 1) - P(X = 1, Y = 1) \\ &= P(X = 1) - P(X = 1)P(Y = 1) \\ &= P(X = 1)(1 - P(Y = 1)) \\ &= P(X = 1)P(Y = 0). \end{aligned}$$

Cazul $X = 0, Y = 1$ se tratează similar cu cazul anterior:

$$\begin{aligned} P(X = 0, Y = 1) &= P(Y = 1) - P(X = 1, Y = 1) \\ &= P(Y = 1) - P(X = 1)P(Y = 1) \\ &= P(Y = 1)(1 - P(X = 1)) \\ &= P(Y = 1)P(X = 0). \end{aligned}$$

Cazul $X = 0, Y = 0$:

$$\begin{aligned} P(X = 0, Y = 0) &= P(X = 0) - P(X = 0, Y = 1) \\ &= P(X = 0) - P(X = 0)P(Y = 1) \\ &= P(X = 0)(1 - P(Y = 1)) \\ &= P(X = 0)P(Y = 0). \end{aligned}$$

Prin urmare, egalitatea $P(X, Y) = P(X)P(Y)$ este adevărată pentru toate căzurile, deci variabilele X și Y sunt independente.

12.

(Variabile aleatoare: calcul de medii)

Fie X o variabilă aleatoare pentru care $E(X) = \mu$ și $Var(X) = \sigma^2$.

a. *CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW1, pr. 2.4*

Cât este $E[X(X - 1)]$ în funcție de μ și σ ?

b. *CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, midterm, pr. 1.c*

Fie $c \in R$. Care din următoarele variante sunt adevărate?

- | | |
|--|---|
| A. $E[(X - c)^2] = (\mu - c)^2 + \sigma^2$ | D. $E[(X - c)^2] = (\mu - c)^2 + 2\sigma^2$ |
| B. $E[(X - c)^2] = (\mu - c)^2$ | E. $E[(X - c)^2] = \mu^2 + c^2 + 2\sigma^2$ |
| C. $E[(X - c)^2] = (\mu - c)^2 - \sigma^2$ | F. $E[(X - c)^2] = \mu^2 + c^2 - 2\sigma^2$. |

Răspuns:

a. De la problema 9.a știm că media sumei a două variabile aleatoare este suma mediilor variabilelor respective. De asemenea, este imediat demonstrabil că $E[c \cdot X] = c \cdot E[X]$, unde X, Y sunt variabile aleatoare, iar c este o constantă. Tinând cont că $Var(X) = E[X^2] - (E[X])^2$ (vedeți problema 9.b), putem scrie:

$$\begin{aligned} E[X(X - 1)] &= E[X^2 - X] = E[X^2] - E[X] \\ &= E[X^2] - (E[X])^2 + (E[X])^2 - E[X] \\ &= Var(X) + (E[X])^2 - E[X] \\ &= \sigma^2 + \mu^2 - \mu \end{aligned}$$

b. Pentru a găsi varianta adevărată vom calcula $E[(X - c)^2]$:

$$\begin{aligned} E[(X - c)^2] &= E[X^2 - 2cX + c^2] = E[X^2] - 2cE[X] + c^2 \\ &= E[X^2] - (E[X])^2 + (E[X])^2 - 2cE[X] + c^2 \\ &= \sigma^2 + \mu^2 - 2c\mu + c^2 \\ &= \sigma^2 + (\mu - c)^2 \end{aligned}$$

Deci varianta A este adevărată. (Toate celelalte variante sunt, în general, false.)

13.

(Variabile aleatoare discrete:
distribuții de probabilitate comune, marginale, condiționate;
regula de înlățuire)

CMU, 2002 fall, Andrew Moore, final exam, pr. 4.a

Considerăm un set de date definit cu ajutorul a 3 variabile aleatoare cu valori booleene X, Y și Z . Care dintre seturile de informații de mai jos sunt suficiente pentru a specifica distribuția comună $P(x, y, z)$?

A.	B.	C.	D.
$P(\neg X Z)$	$P(\neg X \neg Z)$	$P(X Z)$	$P(X Z)$
$P(\neg X \neg Z)$	$P(X \neg Z)$	$P(X \neg Z)$	$P(X \neg Z)$
$P(\neg Y X, Z)$	$P(Y X, Z)$	$P(Y X, Z)$	$P(Y X, Z)$
$P(\neg Y X, \neg Z)$	$P(Y X, \neg Z)$	$P(Y X, \neg Z)$	$P(Y X, \neg Z)$
$P(\neg Y \neg X, Z)$	$P(Y \neg X, Z)$	$P(Y \neg X, Z)$	$P(\neg Y \neg X, \neg Z)$
$P(\neg Y \neg X, \neg Z)$	$P(Y \neg X, \neg Z)$	$P(\neg Y \neg X, \neg Z)$	$P(Y \neg X, \neg Z)$
$P(Z)$	$P(Z)$	$P(\neg Z)$	$P(Z)$

Răspuns:

Pentru a calcula distribuția comună a mai multor variabile aleatoare se poate aplica regula de înlățuire.²¹ În cazul nostru, putem scrie:

$$P(X, Y, Z) = P(Z) \cdot P(X | Z) \cdot P(Y | X, Z)$$

Deoarece variabilele aleatoare X , Y și Z au valori booleene, pentru a specifica distribuția comună $P(X, Y, Z)$ este nevoie să se calculeze valoarea acesteia în fiecare din cele 8 cazuri posibile:

$$\begin{array}{cccc} P(X, Y, Z) & P(X, Y, \neg Z) & P(X, \neg Y, Z) & P(X, \neg Y, \neg Z) \\ P(\neg X, Y, Z) & P(\neg X, Y, \neg Z) & P(\neg X, \neg Y, Z) & P(\neg X, \neg Y, \neg Z) \end{array}$$

Cunoaștem de asemenea relații de calcul de forma:

$$\begin{aligned} P(\neg X) &= 1 - P(X) \\ P(\neg X | Y) &= 1 - P(X | Y) \end{aligned}$$

Așadar, pentru a aplica regula de înlățuire de mai sus este nevoie să cunoaștem

$$\begin{array}{ll} P(Z) \text{ sau } P(\neg Z); & P(Y | X, Z) \text{ sau } P(\neg Y | X, Z); \\ P(X | Z) \text{ sau } P(\neg X | Z); & P(Y | \neg X, Z) \text{ sau } P(\neg Y | \neg X, Z); \\ P(X | \neg Z) \text{ sau } P(\neg X | \neg Z); & P(Y | X, \neg Z) \text{ sau } P(\neg Y | X, \neg Z); \\ & P(Y | \neg X, \neg Z) \text{ sau } P(\neg Y | \neg X, \neg Z). \end{array}$$

Cu aceste precizări, putem specifica pentru fiecare dintre seturile de informații din enunț dacă sunt suficiente pentru a calcula distribuția comună $P(X, Y, Z)$.

Cazul A. Da. Se observă că putem calcula $P(X | Z)$ și $P(X | \neg Z)$ din primele două probabilități din enunț. De asemenea, utilizând următoarele 4 probabilități se pot deduce: $P(Y | X, Z)$, $P(Y | X, \neg Z)$, $P(Y | \neg X, Z)$ și $P(Y | \neg X, \neg Z)$. Iar din $P(Z)$ se obține $P(\neg Z)$. Prin urmare, există toate informațiile necesare distribuției comune $P(X, Y, Z)$.

Cazul B. Nu, informațiile din enunț nu sunt suficiente. Nu putem deduce valoarea pentru $P(X | Z)$.

²¹Pentru variabile aleatoare, regula de înlățuire

$$P(A_1, A_2, \dots, A_n) = P(A_1) \cdot P(A_2 | A_1) \cdot P(A_n | A_1, A_2, \dots, A_{n-1})$$

se demonstrează imediat pornind de la regula de înlățuire pentru (probabilități de) evenimente aleatoare. A se vedea problema 14.

Cazul C. Da, informațiile din enunț sunt suficiente. Din $P(X | Z)$ se obține $P(\neg X | Z)$, iar din $P(X | \neg Z)$ se obține $P(\neg X | \neg Z)$. Din următoarele 4 probabilități se obțin celelalte 4 necesare pentru $P(Y | X, Z)$, și anume: $P(\neg Y | X, Z)$, $P(\neg Y | X, \neg Z)$, $P(\neg Y | \neg X, Z)$ și respectiv $P(Y | \neg X, \neg Z)$.

Cazul D. Nu, informațiile din enunț nu sunt suficiente. Nu putem deduce valoarea pentru $P(Y | \neg X, Z)$.

14.

(Variabile aleatoare discrete:
regula de înmulțire, varianta condițională)
CMU, 2009 fall, Geoff Gordon, HW2, pr. 2.1

Arătați că pentru orice valori x, y și z ale variabilelor aleatoare X, Y și Z respectiv, avem:

$$P(X = x, Y = y | Z = z) = P(X = x | Y = y, Z = z) \cdot P(Y = y | Z = z).$$

În notație simplificată: $P(X, Y | Z) = P(X | Y, Z) \cdot P(Y | Z)$.

Indicație: Folosiți regula de înlățuire pentru evenimente aleatoare.

Răspuns:

Folosind definiția probabilității condiționate și regula de înlățuire (cu termeni ordonați în mod convenabil), egalitatea cerută se obține astfel:

$$\begin{aligned} P(X, Y | Z) &\stackrel{\text{def.}}{=} \frac{P(X, Y, Z)}{P(Z)} \stackrel{\text{not.}}{=} \frac{P(Z \cap Y \cap X)}{P(Z)} \\ &= \frac{P(Z)P(Y | Z)P(X | Y, Z)}{P(Z)} = P(Y | Z) \cdot P(X | Y, Z) \end{aligned}$$

Observație: O altă metodă de rezolvare constă în a aplica pentru fiecare membru al egalității din enunț definiția probabilității condiționate, după simplificări obținându-se pentru ambii membri aceeași valoare:

$$\begin{aligned} P(X, Y | Z) &= \frac{P(X, Y, Z)}{P(Z)} \\ P(X | Y, Z) \cdot P(Y | Z) &= \frac{P(X, Y, Z)}{P(Y, Z)} \cdot \frac{P(Y, Z)}{P(Z)} = \frac{P(X, Y, Z)}{P(Z)} \end{aligned}$$

15.

(Variabile aleatoare discrete: independentă condițională)
CMU, 2005 fall, T. Mitchell, A. Moore, midterm, pr. 4

Fie variabilele aleatoare discrete A, B și C având distribuția comună conform tabelului de mai jos.

- Este variabila A independentă condițional de B în raport cu variabila C ?
- Dacă ați răspuns afirmativ la întrebarea a, faceți o schimbare în primele două linii ale tabelului de mai sus pentru a obține o distribuție pentru care răspunsul la aceeași întrebare să devină negativ. Invers, dacă ați răspuns negativ la întrebarea a, faceți o schimbare în primele două linii ale tabelului încât răspunsul să devină afirmativ.

A	B	C	$P(A, B, C)$
0	0	0	1/8
0	0	1	1/8
0	1	0	1/8
0	1	1	1/8
1	0	0	1/8
1	0	1	1/8
1	1	0	1/8
1	1	1	1/8

Răspuns:

a. Faptul că variabila A este independentă condițional de B în raport cu variabila C se mai notează prin $A \perp B | C$ și poate fi demonstrat prin una din următoarele două relații:

$$P(A = a, B = b | C = c) = P(A = a | C = c) \cdot P(B = b | C = c) \text{ sau}$$

$$P(A = a | B = b, C = c) = P(A = a | C = c), \text{ dacă } P(B = b, C = c) \neq 0.$$

pentru orice $a \in Val(A), b \in Val(B), c \in Val(C)$. Deși în general se folosește prima relație, pentru acest exercițiu este mai ușor să utilizăm cea de-a doua relație. Conform tabelului dat în enunț, rezultă imediat că $P(B = b, C = c) \neq 0$ pentru orice $b, c \in \{0, 1\}$. Vom demonstra că pentru orice $a, b, c \in \{0, 1\}$ este adevărat că $P(A = a | B = b, C = c) = P(A = a | C = c)$. Cele două probabilități condiționate vor fi calculate folosind datele din tabel:

$$\text{Cazul } (0, 0, 0): P(A = 0 | B = 0, C = 0) = \frac{1 \cdot \frac{1}{8}}{\frac{1}{2} \cdot \frac{1}{8}} = \frac{1}{2} \text{ și } P(A = 0 | C = 0) = \frac{2 \cdot \frac{1}{8}}{4 \cdot \frac{1}{8}} = \frac{1}{2}.$$

$$\text{Cazul } (0, 0, 1): P(A = 0 | B = 0, C = 1) = \frac{1 \cdot \frac{1}{8}}{\frac{1}{2} \cdot \frac{1}{8}} = \frac{1}{2} \text{ și } P(A = 0 | C = 1) = \frac{2 \cdot \frac{1}{8}}{4 \cdot \frac{1}{8}} = \frac{1}{2}.$$

Se observă că pentru toate celelalte cazuri se obțin aceleași valori, deci

$$P(A = a | B = b, C = c) = P(A = a | C = c), \forall a, b, c \in \{0, 1\}.$$

Așadar, variabila A este independentă condițional de B în raport cu variabila C .

b. Schimbarea trebuie făcută în aşa fel încât să se păstreze relația $\sum P(A, B, C) = 1$. O variantă posibilă este:

A	B	C	$P(A, B, C)$
0	0	0	1/4
0	0	1	0
...			...

Pentru aceste noi valori se observă că: $P(A = 0 | B = 0, C = 0) = \frac{1 \cdot 1/4}{1 \cdot 1/4 + 1 \cdot 1/8} = \frac{1 \cdot 1/4}{1 \cdot 1/4 + 3 \cdot 1/8} = \frac{2}{3}$ și $P(A = 0 | C = 0) = \frac{1 \cdot 1/4 + 1 \cdot 1/8}{1 \cdot 1/4 + 3 \cdot 1/8} = \frac{3}{5}$. Este suficient un singur caz în care probabilitățile respective nu sunt egale, prin urmare variabila A nu este independentă condițional de variabila B în raport cu a treia variabilă, C .

16.

(Variabile aleatoare discrete:
distribuții comune, distribuții marginale, distribuții condiționale;
independență, independență condițională)

• CMU, 2016 fall, N. Balcan, M. Gormley, HW2, pr. 1.4

Fie trei variabile aleatoare X, Y și Z care iau valori în mulțimea $\{0, 1\}$. În tabelul următor este dată distribuția probabilistă comună a acestor trei variabile, $P(X, Y, Z)$.

		$Z = 0$		$Z = 1$	
		$X = 0$	$X = 1$	$X = 0$	$X = 1$
$Y = 0$	1/24	1/12	1/12	5/24	
	1/12	p	q	7/24	

a. Considerând că X și Y sunt independente, găsiți valorile lui p și q .

Indicație: Este util (deși nu obligatoriu) să calculați mai întâi $P(X, Y)$, distribuția comună a variabilelor X și Y , completând tabelul alăturat, după care veți calcula și distribuțiile (marginale) pentru X și Y , de preferință ca o linie și respectiv o coloană suplimentară la acest tabel.

	$X = 0$	$X = 1$
$Y = 0$		
$Y = 1$		

b. Considerând valorile lui p și q determinate la punctul a, sunt X și Y independente condițional în raport cu Z ? De ce?

Răspuns:

a. Conform definiției, variabilele aleatoare X și Y sunt independente dacă și numai dacă

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y), \forall x \in Val(X), \forall y \in Val(Y). \quad (7)$$

Observație: Evident, pentru a determina p și q , ar fi de dorit să scriem un sistem de două ecuații cu aceste două necunoscute. Cele două ecuații pot fi obținute prin particularizarea relației (7) pentru două perechi distincte de valori $(x, y) \in Val(X) \times Val(Y)$. Alternativ, va fi suficient să reținem doar o astfel de ecuație, pentru că în locul celeilalte putem folosi una din proprietățile din definiția funcției de probabilitate, și anume

$$\sum_{x,y,z} P(X = x, Y = y, Z = z) = 1. \quad (8)$$

Aceasta revine practic la a însuma elementele din tabelul dat în enunț pentru distribuția $P(X, Y, Z)$.²²

Pentru a calcula probabilitățile marginale $P(X = x, Y = y)$, vom proceda conform definiției: $P(X = x, Y = y) = \sum_{z \in Val(Z)} P(X = x, Y = y, Z = z)$. Si anume:

$$\begin{aligned} P(X = 0, Y = 0) &= P(X = 0, Y = 0, Z = 0) + P(X = 0, Y = 0, Z = 1) = \frac{1}{24} + \frac{1}{12} = \frac{1}{8} \\ P(X = 0, Y = 1) &= P(X = 0, Y = 1, Z = 0) + P(X = 0, Y = 1, Z = 1) = \frac{1}{12} + q \\ P(X = 1, Y = 0) &= P(X = 1, Y = 0, Z = 0) + P(X = 1, Y = 0, Z = 1) = \frac{1}{12} + \frac{5}{24} = \frac{7}{24} \\ P(X = 1, Y = 1) &= P(X = 1, Y = 1, Z = 0) + P(X = 1, Y = 1, Z = 1) = \frac{7}{24} + p. \end{aligned}$$

În mod similar,

$$P(X = 0) = \sum_{y \in Val(Y)} P(X = 0, Y = y) = \frac{1}{8} + \frac{1}{12} + q = \frac{5}{24} + q$$

²²Remarcăm că, dacă vom urma *Indicația* din enunț, după ce vom scrie probabilitățile marginale $P(X, Y)$, $P(X)$ și $P(Y)$, va fi ușor să scriem încă alte trei proprietăți similare cu (8). Deci, o vom putea alege atunci pe cea mai „directă”/simplă dintre ele.

$$P(X = 1) = \frac{7}{24} + p + \frac{7}{24} = \frac{7}{12} + p$$

și

$$\begin{aligned} P(Y = 0) &= \sum_{x \in Val(X)} P(X = x, Y = 0) = \frac{1}{8} + \frac{7}{24} = \frac{5}{12} \\ P(Y = 1) &= 1 - P(Y = 0) = \frac{7}{12}. \end{aligned}$$

Punând acum toate aceste rezultate împreună, obținem:

	$X = 0$	$X = 1$	
$Y = 0$	$1/8$	$7/24$	$5/12$
$Y = 1$	$1/12 + q$	$7/24 + p$	$7/12$
	$5/24 + q$	$7/12 + p$	

Acum este simplu de văzut că $5/24 + q + p + 14/24 = 1 \Leftrightarrow p + q = 5/24$. (Aceasta constituie prima noastră ecuație în p și q .) În consecință, am putea chiar să eliminăm q din tabelul de mai sus, scriind $P(X = 0, Y = 1) = 1/12 + q = 1/12 + 5/24 - p = 7/24 - p$ și $P(X = 0) = 5/24 + q = 5/24 + 5/24 - p = 5/12 - p$.

Aplicând definiția independenței variabilelor X și Y pentru $x = 0$ și $y = 0$, vom avea:

$$\begin{aligned} P(X = 0, Y = 0) &= P(X = 0) \cdot P(Y = 0) \Leftrightarrow \frac{1}{8} = \left(\frac{5}{12} - p \right) \cdot \frac{5}{12} \Leftrightarrow \frac{5}{12} - p = \frac{3}{2} \cdot \frac{1}{5} \\ &\Leftrightarrow p = \frac{5}{12} - \frac{3}{10} = \frac{25 - 18}{60} = \frac{7}{60}, \end{aligned}$$

de unde rezultă imediat

$$q = \frac{5}{24} - p = \frac{5}{24} - \frac{7}{60} = \frac{25 - 14}{120} = \frac{11}{120}.$$

De asemenea, vom avea $P(X = 0) = 5/12 - p = 5/12 - 7/60 = 3/10$ și $P(X = 1) = 7/10$.

În final, va trebui să verificăm și celelalte trei egalități din definiția independenței lui X și Y , pentru că, în general, este posibil ca sistemul de ecuații corespunzător relației (7) să fie supra-restricționat (deci, incompatibil):²³

$$P(X = 0, Y = 1) = P(X = 0) \cdot P(Y = 1) :$$

$$\frac{7}{24} - p = \frac{3}{10} \cdot \frac{7}{12} \Leftrightarrow \frac{7}{24} - \frac{7}{60} = \frac{7}{40} \Leftrightarrow 7 \cdot \left(\frac{1}{24} - \frac{1}{60} \right) = \frac{7}{40} \Leftrightarrow 7 \cdot \frac{5 - 2}{120} = \frac{7}{40} \quad (A)$$

$$P(X = 1, Y = 0) = P(X = 1) \cdot P(Y = 0) :$$

$$\frac{7}{24} = \frac{7}{10} \cdot \frac{5}{12} \quad (A)$$

$$P(X = 1, Y = 1) = P(X = 1) \cdot P(Y = 1) :$$

$$\frac{7}{24} + p = \frac{7}{10} \cdot \frac{7}{12} \Leftrightarrow \frac{7}{24} + \frac{7}{60} = 7^2 \cdot \frac{1}{10} \cdot \frac{1}{12} \Leftrightarrow \frac{1}{24} + \frac{1}{60} = \frac{7}{120} \Leftrightarrow \frac{5 + 2}{120} = \frac{7}{120} \quad (A)$$

²³Ar fi suficient chiar să verificăm doar două dintre cele trei egalități, fiindcă ultima decurge automat din celelalte, ținând cont de proprietățile $\sum_{x,y} P(X = x, Y = y) = 1$, $\sum_x P(X = x) = 1$ și $\sum_y P(Y = y) = 1$.

Așadar, date fiind cele două valori ale lui p și q , variabilele X și Y sunt independente.

b. Conform definiției, variabilele aleatoare X și Y sunt independente condițional în raport cu variabila Z dacă și numai dacă

$$P(X = x, Y = y|Z = z) = P(X = x|Z = z) \cdot P(Y = y|Z = z), \quad (9)$$

$$\forall x \in \text{Val}(X), \forall y \in \text{Val}(Y), \forall z \in \text{Val}(Z).$$

Dacă X și Y nu sunt independente condițional în raport cu Z , atunci $\exists x \in \text{Val}(X), \exists y \in \text{Val}(Y), \exists z \in \text{Val}(Z)$ astfel încât $P(X = x, Y = y|Z = z) \neq P(X = x|Z = z) \cdot P(Y = y|Z = z)$.

Pentru a putea exprima probabilitățile condiționate care apar în formulele de mai sus, vom calcula în prealabil distribuția marginală a lui Z pornind de la tabelul din enunț:

$$\begin{aligned} P(Z = 0) &= \sum_{x \in \text{Val}(X)} \sum_{y \in \text{Val}(Y)} P(X = x, Y = y, Z = 0) \\ &= 5/24 + p = 5/24 + 7/60 = 39/120 = 13/40 \\ P(Z = 1) &= 1 - P(Z = 0) = 27/40. \end{aligned}$$

Vom verifica relația (9) pentru $x = y = z = 0$:

$$\begin{aligned} P(X = 0, Y = 0|Z = 0) &= P(X = 0|Z = 0) \cdot P(Y = 0|Z = 0) \\ \Leftrightarrow \frac{P(X = 0, Y = 0, Z = 0)}{P(Z = 0)} &= \frac{P(X = 0, Z = 0)}{P(Z = 0)} \cdot \frac{P(Y = 0, Z = 0)}{P(Z = 0)} \\ \Leftrightarrow P(X = 0, Y = 0, Z = 0) \cdot P(Z = 0) &= P(X = 0, Z = 0) \cdot P(Y = 0, Z = 0) \\ \Leftrightarrow \frac{1}{24} \cdot \frac{13}{40} &= \left(\frac{1}{24} + \frac{1}{12}\right) \cdot \left(\frac{1}{24} + \frac{1}{12}\right) \\ \Leftrightarrow \frac{13}{24 \cdot 40} &= \frac{1}{8} \cdot \frac{1}{8} \Leftrightarrow \frac{13}{3 \cdot 5} = 1 \quad (F) \end{aligned}$$

Prin urmare, variabilele X și Y nu sunt independente condițional în raport cu variabila Z .

Observație: Această problemă pune în evidență următoarea proprietate, care merită să fie reținută: independența a două variabile aleatoare nu implică (în general) independența lor condițională în raport cu o a treia variabilă aleatoare.

17.

(Variabile aleatoare continue:
funcția densitate de probabilitate)

CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW1, pr. 1.5

Fie variabila aleatoare continuă X a cărei funcție densitate de probabilitate (în limba engleză: “probability density functions”; scris, sub formă abreviată, p.d.f.) este:

$$p(x) = \begin{cases} cx^2 & \text{pentru } 1 \leq x \leq 2 \\ 0 & \text{în caz contrar.} \end{cases}$$

- a. Tinând cont de proprietățile funcției densitate de probabilitate (a se vedea notițele de la curs), cât trebuie să fie valoarea constantei c ?
- b. Desenați graficul funcției de mai sus.
- c. Calculați $P(X > 3/2)$.

Răspuns:

- a. Faptul că $p(x)$ este funcție densitate de probabilitate pentru variabila aleatoare continuă X înseamnă că $p(x) \geq 0, \forall x$ și că $\int_{-\infty}^{+\infty} p(x)dx = 1$. Pentru a afla constanta c vom calcula valoarea integralei:

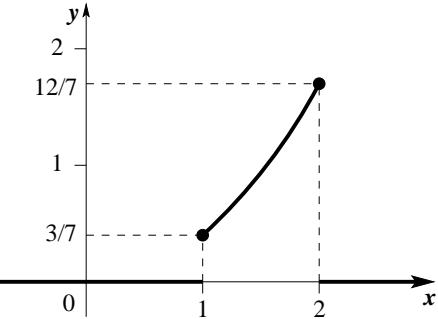
$$\begin{aligned}\int_{-\infty}^{+\infty} p(x)dx &= \underbrace{\int_{-\infty}^1 p(x)dx}_{=0} + \int_1^2 p(x)dx + \underbrace{\int_2^{+\infty} p(x)dx}_{=0} \\ &= \int_1^2 cx^2 dx = c \cdot \int_1^2 x^2 dx = c \cdot \frac{x^3}{3} \Big|_1^2 = c \left(\frac{2^3}{3} - \frac{1^3}{3} \right) = c \cdot \frac{7}{3}\end{aligned}$$

Prin urmare, $c \cdot \frac{7}{3} = 1 \Rightarrow c = \frac{3}{7}$.

- b. Trebuie să reprezentăm grafic funcția:

$$p(x) = \begin{cases} \frac{3}{7}x^2 & \text{pentru } 1 \leq x \leq 2 \\ 0 & \text{în caz contrar.} \end{cases}$$

În intervalul $[1, 2]$, această funcție este un fragment din parabola corespunzătoare funcției de gradul al doilea: $\frac{3}{7}x^2$, parabolă care are vârful în punctul $(0, 0)$. Putem calcula $p(1) = 3/7 \approx 0.42$ și $p(2) = 12/7 \approx 1.71$.



- c. Valoarea probabilității cerute se poate calcula astfel:

$$\begin{aligned}P(X > 3/2) &\stackrel{\text{not.}}{=} P(\{\omega \mid X(\omega) > \frac{3}{2}\}) \stackrel{\text{def.}}{=} \int_{X(\omega)=x>3/2} p(x)dx = \int_{3/2}^{+\infty} p(x)dx \\ &= \int_{3/2}^2 p(x)dx + \underbrace{\int_2^{+\infty} p(x)dx}_{=0} = \int_{3/2}^2 \frac{3}{7}x^2 dx = \frac{3}{7} \cdot \int_{3/2}^2 x^2 dx \\ &= \frac{3}{7} \cdot \frac{x^3}{3} \Big|_{3/2}^2 = \frac{1}{7} \cdot x^3 \Big|_{3/2}^2 = \frac{1}{7} \cdot \left(8 - \frac{27}{8} \right) = \frac{1}{7} \cdot \frac{64 - 27}{8} = \frac{37}{56}.\end{aligned}$$

O altă variantă de rezolvare este bazată pe folosirea *funcției cumulative de distribuție*, care, după cum știm, se definește prin relația $F(x) \stackrel{\text{def.}}{=} P(X \leq x)$, pentru orice $x \in \mathbb{R}$:

$$\begin{aligned}P(X > 3/2) &= 1 - P(X \leq 3/2) = 1 - F\left(\frac{3}{2}\right) = 1 - \int_{-\infty}^{3/2} p(x)dx \\ &= 1 - \left(\underbrace{\int_{-\infty}^1 p(x)dx}_0 + \int_1^{3/2} p(x)dx \right) = 1 - \left(0 + \int_1^{3/2} p(x)dx \right)\end{aligned}$$

$$\begin{aligned}
&= 1 - \int_1^{3/2} p(x)dx = 1 - \int_1^{3/2} \frac{3}{7}x^2 dx = 1 - \frac{1}{7} \cdot x^3 \Big|_1^{3/2} \\
&= 1 - \frac{1}{7} \cdot \left(\frac{27}{8} - 1 \right) = 1 - \frac{1}{7} \cdot \frac{19}{8} \\
&= 1 - \frac{19}{56} = \frac{37}{56}.
\end{aligned}$$

18.

(Variabile aleatoare continue:
funcția densitate de probabilitate)

CMU, 2008 spring, Eric Xing, HW1, pr. 1.1.b

Fie funcția

$$p(x) = \begin{cases} cx^{-d} & \text{pentru } x > 1 \\ 0 & \text{în caz contrar.} \end{cases}$$

Care sunt valorile posibile pentru c și d în aşa fel ca p să poată reprezenta o funcție densitate de probabilitate?Răspuns:O funcție p poate reprezenta o funcție densitate de probabilitate dacă $p(x) \geq 0$ pentru $\forall x$, și $\int_{-\infty}^{\infty} p(x)dx = 1$. Vom folosi aceste două condiții pentru a calcula valorile posibile pentru c și d .Prima condiție $p(x) \geq 0$ implică faptul că $c \geq 0$, asupra lui d neimpunând nicio restricție. Mai mult, ținând cont de forma lui p și de cea de-a doua condiție, vom avea chiar $c > 0$, fiindcă $c = 0$ ar implica $\int_{-\infty}^{\infty} p(x)dx = 0 \neq 1$.

Pentru a aplica cea de-a doua condiție, calculăm integrala:

$$\int_{-\infty}^{\infty} p(x)dx = \int_1^{\infty} cx^{-d} dx = c \int_1^{\infty} x^{-d} dx.$$

Vom avea de tratat două cazuri:

$$\text{Cazul 1: } d = 1 \Rightarrow c \int_1^{\infty} x^{-d} dx = c \int_1^{\infty} \frac{1}{x} dx = c \cdot \ln x \Big|_1^{\infty} = \infty$$

$$\text{Cazul 2: } d \neq 1 \Rightarrow c \int_1^{\infty} x^{-d} dx = c \cdot \frac{x^{-d+1}}{-d+1} \Big|_1^{\infty} = \frac{c}{1-d} \cdot x^{1-d} \Big|_1^{\infty}$$

Subcazul $d > 1$: $x^{1-d} \Big|_1^{\infty} = 0 - 1 = -1$. Deci în acest caz $c \int_1^{\infty} x^{-d} dx = \frac{c}{d-1}$.Subcazul $d < 1$: $x^{1-d} \Big|_1^{\infty} = +\infty$. Deci în acest caz $c \int_1^{\infty} x^{-d} dx = +\infty$.Prin urmare, $\int_{-\infty}^{\infty} p(x)dx = 1 \Rightarrow d > 1$ și $\frac{c}{d-1} = 1$.În concluzie, p poate reprezenta o funcție densitate de probabilitate dacă

$$c > 0, d > 1 \text{ și } c = d - 1.$$

Referitor la aceste trei restricții, se observă că ultimele două o implică pe cea dintâi, deci o fac superfluă.

19. (Coeficientul de corelație pentru două variabile aleatoare: două proprietăți)

*Liviu Ciortuz, 2019, după
■ Sheldon Ross, A First Course in Probability, 5th ed.,
Prentice Hall, 1997, pag. 332*

Pentru două variabile aleatoare oarecare X și Y având $\text{Var}(X) \neq 0$ și $\text{Var}(Y) \neq 0$, coeficientul de corelație se definește astfel:

$$\rho(X, Y) \stackrel{\text{def.}}{=} \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y},$$

unde $\sigma_X \stackrel{\text{not.}}{=} \sqrt{\text{Var}(X)}$ și $\sigma_Y \stackrel{\text{not.}}{=} \sqrt{\text{Var}(Y)}$ sunt deviațiile standard ale celor două variabile aleatoare.

- a. Să se demonstreze că $-1 \leq \rho(X, Y) \leq 1$.

Consecință: $\text{Cov}(X, Y) \in [-\sigma_X \sigma_Y, +\sigma_X \sigma_Y]$.

- b. Să se arate că dacă $\rho(X, Y) = 1$ (deci $\text{Cov}(X, Y) = \sigma_X \sigma_Y$), atunci $Y = aX + b$, cu $a = \sigma_Y/\sigma_X > 0$. Similar, dacă $\rho(X, Y) = -1$ (deci $\text{Cov}(X, Y) = -\sigma_X \sigma_Y$, atunci $Y = aX + b$, cu $a = -\sigma_Y/\sigma_X < 0$).²⁴

Indicații:

- La punctul a, pentru a demonstra inegalitatea $\rho(X, Y) \geq -1$ vă sugerăm să dezvoltați expresia $\text{Var}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right)$ folosind următoarele două proprietăți, valabile pentru orice variabile aleatoare X și Y : $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$,²⁵ și $\text{Cov}(aX, bY) = ab\text{Cov}(X, Y)$, pentru orice $a, b \in \mathbb{R}$.²⁶ Apoi veți proceda similar, dezvoltând expresia $\text{Var}\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right)$, ca să demonstrați inegalitatea $\rho(X, Y) \leq 1$.
- La punctul b, veți ține cont de faptul că pentru o variabilă aleatoare oarecare X , avem $\text{Var}(X) = 0$ dacă și numai dacă variabila X este constantă.²⁷

Răspuns:

- a. Pentru a demonstra inegalitatea $\rho(X, Y) \geq -1$, procedăm conform *Indicației 1*:

$$\begin{aligned} \text{Var}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) &= \text{Var}\left(\frac{X}{\sigma_X}\right) + \text{Var}\left(\frac{Y}{\sigma_Y}\right) + 2\text{Cov}\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right) \\ &= \frac{1}{\sigma_X^2} \text{Var}(X) + \frac{1}{\sigma_Y^2} \text{Var}(Y) + 2 \frac{1}{\sigma_X \sigma_Y} \text{Cov}(X, Y) \\ &= 1 + 1 + 2\rho(X, Y) = 2[1 + \rho(X, Y)]. \end{aligned}$$

²⁴ Așadar, coeficientul de corelație reprezintă o „măsură“ a gradului de „dependență liniară“ dintre X și Y . LC: Coeficientul a din relația $Y = aX + b$ nu poate lua valori în afara intervalului $[-\sigma_Y/\sigma_X, +\sigma_Y/\sigma_X]$ — așa cum ne-am așteptat dacă facem legătura cu ecuația unei drepte oarecare din planul euclidian — din cauza simetriei $\text{Cov}(X, Y) = \text{Cov}(Y, X)$.

²⁵ Vedeți rezolvarea problemei 23.c.

²⁶ Vedeți *Consecința* de la problema 9.c.

²⁷ Mai precis, există $c \in \mathbb{R}$ astfel încât $P(X = c) = 1$, unde P este distribuția de probabilitate considerată la definirea variabilelor din enunțul problemei.

Întrucât $\text{Var}(X) \geq 0$ pentru orice variabilă aleatoare X , rezultă că $\text{Var}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) \geq 0$, deci $1 + \rho(X, Y) \geq 0$, adică $\rho(X, Y) \geq -1$.

În mod similar, putem să arătăm că

$$\text{Var}\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right) = 2[1 - \rho(X, Y)] \geq 0,$$

deci $\rho(X, Y) \leq 1$.

b. Dacă $\rho(X, Y) = -1$, atunci din primul calcul de la punctul a va rezulta că $\text{Var}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) = 0$. Stim că aceasta se întâmplă dacă $\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}$ este o variabilă aleatoare constantă, adică, mai precis, există $a' \in \mathbb{R}$ astfel încât $\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y} = a'$ cu probabilitate 1. Prin urmare, putem scrie $Y = a'\sigma_Y - \frac{\sigma_Y}{\sigma_X}X$. Rezultă că $Y = aX + b$, unde $a = -\frac{\sigma_Y}{\sigma_X} < 0$ și $b = a'\sigma_Y$.

În mod similar, dacă $\rho(X, Y) = 1$, atunci din al doilea calcul de la punctul a va rezulta că $\text{Var}\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right) = 0$, deci există $a'' \in \mathbb{R}$ astfel încât $\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y} = a''$ cu probabilitate 1. Așadar, $Y = -a''\sigma_Y + \frac{\sigma_Y}{\sigma_X}X$. Renotând, obținem $Y = aX + b$, cu $a = \frac{\sigma_Y}{\sigma_X} > 0$ și $b = -a''\sigma_Y$.

20.

(O proprietate: matricea de covarianță a oricărui vector de variabile aleatoare este simetrică și pozitiv semidefinită)

■ □ prelucrare de Liviu Ciortuz, după
“The Multivariate Gaussian Distribution”, Chuong B. Do, 2008

Fie variabilele aleatoare X_1, \dots, X_n , cu $X_i : \Omega \rightarrow \mathbb{R}$ pentru $i = 1, \dots, n$. Matricea de covarianță a vectorului de variabile aleatoare $X = (X_1, \dots, X_n)$ este o matrice pătratică de dimensiune $n \times n$, ale cărei elemente se definesc astfel: $[\text{Cov}(X)]_{ij} \stackrel{\text{def.}}{=} \text{Cov}(X_i, X_j)$, pentru orice $i, j \in \{1, \dots, n\}$.

Arătați că $\Sigma \stackrel{\text{not.}}{=} \text{Cov}(X)$ este matrice simetrică și pozitiv semidefinită, cea de-a doua proprietate însemnând că pentru orice vector $z \in \mathbb{R}^n$ are loc inegalitatea $z^\top \Sigma z \geq 0$. (Vectorii $z \in \mathbb{R}^n$ sunt considerați vectori-coloană, iar simbolul \top reprezintă operația de transpunere a matricelor.)

Răspuns:

Faptul că matricea Σ este simetrică decurge imediat din definiția ei: dacă $X = (X_1, \dots, X_n)$, atunci $[\text{Cov}(X)]_{i,j} \stackrel{\text{def.}}{=} \text{Cov}(X_i, X_j) = E[(X_i - E[X_i])(X_j - E[X_j])] = E[(X_j - E[X_j])(X_i - E[X_i])] = \text{Cov}(X_j, X_i) = [\text{Cov}(X)]_{j,i}$, pentru orice $i, j \in \{1, \dots, n\}$.

Apoi, pentru orice vector $z \in \mathbb{R}^n$ de forma $z = (z_1, \dots, z_n)^\top$, avem:

$$\begin{aligned}
z^\top \Sigma z &= \sum_{i=1}^n z_i \left(\sum_{j=1}^n \Sigma_{ij} z_j \right) = \sum_{i=1}^n \sum_{j=1}^n (z_i \Sigma_{ij} z_j) = \sum_{i=1}^n \sum_{j=1}^n (z_i \operatorname{Cov}[X_i, X_j] z_j) \\
&= \sum_{i=1}^n \sum_{j=1}^n (z_i E[(X_i - E[X_i])(X_j - E[X_j])] z_j) \\
&= E \left[\sum_{i=1}^n \sum_{j=1}^n z_i (X_i - E[X_i])(X_j - E[X_j]) z_j \right]
\end{aligned}$$

Ultima dintre egalitățile de mai sus derivă din proprietatea de liniaritate a mediilor (vedeți ex. 9.a). Mai departe,

$$\begin{aligned}
z^\top \Sigma z &= E \left[\left(\sum_{i=1}^n z_i (X_i - E[X_i]) \right) \left(\sum_{j=1}^n (X_j - E[X_j]) z_j \right) \right] \\
&= E \left[\left(\sum_{i=1}^n (X_i - E[X_i]) z_i \right) \left(\sum_{j=1}^n (X_j - E[X_j]) z_j \right) \right] \\
&= E \left[\left(\sum_{i=1}^n (X_i - E[X_i]) z_i \right)^2 \right],
\end{aligned}$$

ceea ce evident, reprezintă o cantitate nenegativă. Așadar, $z^\top \Sigma z \geq 0$.

Să mai observăm că ultima expresie obținută mai sus se scrie mai simplu / compact sub formă vectorială astfel:

$$E[((X - E[X])^\top \cdot z)^2].$$

21.

(Câteva inegalități de bază în teoria probabilităților:
margini superioare
pentru probabilități de forma $P(Z \geq t)$ și $P(Z - E[Z] \geq t)$
și unele consecințe ale lor)

□ • · Liviu Ciortuz, Andi Munteanu, pornind de la
A first course in Probability, Sheldon Ross, 8th edition, 2010,
Stanford, Machine Learning course, John Duchi,
Supplemental Lecture Notes – Hoeffding's inequality

a. Inegalitatea lui [Andrey] Markov.²⁸

Fie $Z \geq 0$ o variabilă aleatoare cu valori nenegative. Demonstrați că pentru orice $t > 0$, are loc inegalitatea

$$P(Z \geq t) \leq \frac{E[Z]}{t}.$$

²⁸ Andrey Markov, 1856 - 1922, matematician rus, care l-a avut ca profesor pe Pafnutiy Chebyshev. Inegalitatea aceasta a apărut în lucrările lui Chebyshev (și ale altor matematicieni), însă ea a primit mai târziu numele lui Markov, în onoarea lui. Markov a avut contribuții în domeniul proceselor stochastice. Cercetările sale au stat la baza creării lanțurilor Markov și a proceselor Markov. Cf. https://en.wikipedia.org/wiki/Andrey_Markov, accesat la 22 martie 2022.

b. Inegalitatea lui Chebyshev²⁹ – consecință a inegalității Markov.

Fie Z o variabilă aleatoare având atât media ($E[Z]$) cât și varianța ($Var[Z]$) finite. Demonstrați că pentru orice $t > 0$,

$$P(Z - E[Z] \geq t \text{ sau } Z - E[Z] \leq -t) = P(|Z - E[Z]| \geq t) \leq \frac{Var[Z]}{t^2}.$$

c. O consecință imediată a inegalității Chebyshev.

Fie Z o variabilă aleatoare a cărei medie ($E[Z]$) este finită. Arătați că în cazul în care $Var[Z] = 0$, rezultă

$$P(Z = E[Z]) = 1.$$

d. O altă inegalitate de tip Chebyshev — consecință a inegalității Markov — cunoscută și sub numele de inegalitatea lui Cantelli.³⁰

Fie Z o variabilă aleatoare pentru care media ($E[Z]$) și varianța ($Var[Z]$) sunt finite. Demonstrați că pentru orice $t > 0$, are loc inegalitatea

$$P(Z - E[Z] \geq t) \leq \frac{Var[Z]}{Var[Z] + t^2}.$$

e. Inegalitățile lui Chernoff³¹ – consecințe ale inegalității Markov.

Fie Z o variabilă aleatoare a cărei medie ($E[Z]$) este finită. Demonstrați că pentru orice $t \geq 0$,

$$P(Z \geq E[Z] + t) \leq \min_{\lambda \geq 0} \{E[\exp(\lambda(Z - E[Z]))] e^{-\lambda t}\} \stackrel{\text{not.}}{=} \min_{\lambda \geq 0} \{M_{Z-E[Z]}(\lambda) e^{-\lambda t}\}$$

și

$$P(Z \leq E[Z] - t) \leq \min_{\lambda \geq 0} \{E[\exp(\lambda(E[Z] - Z))] e^{-\lambda t}\} \stackrel{\text{not.}}{=} \min_{\lambda \geq 0} \{M_{E[Z]-Z}(\lambda) e^{-\lambda t}\},$$

unde prin $M(\lambda)$ am desemnat funcția generatoare de momente (vedeți pr. 120).

f. Legea numerelor mari, varianta slabă.³²

Folosiți inegalitatea Chebyshev pentru a demonstra următoarea proprietate: dacă Z_1, Z_2, \dots sunt un sir de variabile aleatoare i.i.d., toate având media finită $E[Z_i] = \mu$, atunci urmează că pentru orice $\varepsilon > 0$

$$P\left(\left|\frac{Z_1 + \dots + Z_n}{n} - \mu\right| \geq \varepsilon\right) \rightarrow 0 \text{ atunci când } n \rightarrow \infty.$$

²⁹Pafnutij Chebyshev, 1821 - 1894, matematician rus; este considerat părintelele fondator al matematicii rusești. A avut contribuții în domeniile probabilităților, statisticii, mecanicii, precum și în teoria numerelor. Cf. https://en.wikipedia.org/wiki/Pafnutij_Chebyshev, accesat la 22 martie 2022.

³⁰Din inegalitatea lui Chebyshev, deducem că $P(Z - E[Z] \geq t) \leq \frac{Var[Z]}{t^2}$. Vom arăta aici că putem stabili pentru această probabilitate o margine superioară mai bună, și anume $\frac{Var[Z]}{Var[Z] + t^2}$.

³¹Herman Chernoff, 1923- , matematician american, ai căruia părinți au fost imigranți evrei proveniți din Rusia. Herman Chernoff a predat la University of Illinois Urbana-Champaign, Stanford, MIT și Harvard University. Cf. https://en.wikipedia.org/wiki/Herman_Chernoff, accesat la 22 martie 2022.

³²Alături de teoremele limită centrală, teoremele cunoscute sub numele de legi ale numerelor mari constituie cele mai importante rezultate din teoria probabilităților. Legile numerelor mari stabilesc condițiile în care media aritmetică (sau, „empirică“) a unei secvențe de variabile aleatoare aproximează (într-un anumit sens) media probabilistică a variabilelor respective. Teoremele limită centrală studiază condițiile în care suma unui mare număr de variabile aleatoare urmează o distribuție aproximativ normală / gaussiană. Cf. *A first course in Probability*, Sheldon Ross, 8th edition, 2010, Prentice Hall, pag. 395.

Sugestie: În demonstrație puteți folosi o *presupozitie suplimentară*, și anume că toate variabilele au varianță finită, $\text{Var}[Z_i] \stackrel{\text{not.}}{=} \sigma^2$.

Răspuns:

a. Inegalitatea lui Markov.

Definim variabila-indicator $1_{\{Z \geq t\}}$, care ia valoarea 1 atunci când $Z \geq t$ și 0 în cazul contrar. Întrucât Z ia doar valori nenegative, are loc inegalitatea $1_{\{Z \geq t\}} \leq \frac{Z}{t}$. Aplicând operatorul E (media) ambilor membrii ai acestei inegalități, obținem $E[1_{\{Z \geq t\}}] \leq E\left[\frac{Z}{t}\right] \Leftrightarrow E[1_{\{Z \geq t\}}] \leq \frac{E[Z]}{t}$, fiindcă t este constant. Din modul în care este definită variabila aleatoare $1_{\{Z \geq t\}}$, rezultă că $E[1_{\{Z \geq t\}}] = P(Z \geq t)$ (vedeți pr. 98). Înlocuind în ultima inegalitate, obținem că $P(Z \geq t) \leq \frac{E[Z]}{t}$, ceea ce trebuia demonstrat.

b. Inegalitatea lui Chebyshev.

Variabila aleatoare $(Z - E[Z])^2$ ia doar valori nenegative, deci putem să-i aplicăm inegalitatea lui Markov cu $t = t'^2$:

$$P((Z - E[Z])^2 \geq t'^2) \leq \frac{E[(Z - E[Z])^2]}{t'^2}.$$

Inegalitatea $(Z - E[Z])^2 \leq t'^2$ este satisfăcută dacă și numai dacă $|Z - E[Z]| \geq t'$. Ajungem deci la următoarea inegalitate:

$$P(|Z - E[Z]| \geq t') \leq \frac{E[(Z - E[Z])^2]}{t'^2} = \frac{\text{Var}[Z]}{t'^2}.$$

c. O consecință imediată a inegalității Chebyshev.

Conform inegalității Chebyshev, pentru variabila aleatoare Z are loc inegalitatea $P(|Z - E[Z]| \geq t) \leq \frac{\text{Var}[Z]}{t^2}$, pentru orice $t > 0$. Din ipoteză stim că $\text{Var}[Z] = 0$, deci $P(|Z - E[Z]| \geq t) \leq 0$. Tinând cont de faptul că probabilitățile pot lua valori doar între 0 și 1, deducem că $P(|Z - E[Z]| \geq t) = 0$, pentru orice $t > 0$.

Întrucât expresia $|Z - E[Z]|$ ia ca valori doar numere reale nenegative, iar evenimentele $|Z - E[Z]| = t$ sunt imposibile pentru orice $t > 0$, rezultă că $P(|Z - E[Z]| = 0) = 1 \Leftrightarrow P(Z = E[Z]) = 1$, adică Z este variabilă aleatoare constantă.

d. O altă inegalitate de tip Chebyshev.

Fie $a > 0$ un termen ce va fi adăugat de o parte și alta a inegalității $Z - E[Z] \geq t$. Folosindu-ne de inegalitatea evidentă $P(X \geq x) \leq P(X^2 \geq x^2)$ pentru orice $x > 0$, obținem:

$$\begin{aligned} P(Z - E[Z] \geq t) &= P(\underbrace{Z - E[Z] + a}_{\text{not.: } X} \geq \underbrace{t + a}_{\text{not.: } x}) \leq P((Z - E[Z] + a)^2 \geq (t + a)^2) \\ &\stackrel{\text{Markov}}{\leq} \frac{E[(Z - E[Z] + a)^2]}{(t + a)^2}. \end{aligned}$$

În continuare, vom simplifica numărătorul fracției:

$$E[(Z - E[Z] + a)^2] = E[(Z - E[Z])^2 + 2a(Z - E[Z]) + a^2] \stackrel{\text{lin. med.}}{=} a^2 + 2a(E[Z] - t).$$

$$= E[(Z - E[Z])^2] + 2aE[Z] - 2aE[Z] + a^2 = Var[Z] + a^2.$$

Înlocuind în inegalitatea de mai sus, obținem $P(Z - E[Z] \geq t) \leq \frac{Var[Z] + a^2}{(t + a)^2}$.

Această inegalitate este valabilă pentru orice $a > 0$, deci pentru îmbunătățirea marginii superioare, vom minimiza membrul drept.

Fie $f : (0, \infty) \rightarrow \mathbb{R}$, $f(x) = \frac{x^2 + \sigma^2}{(x + t)^2}$, unde $\sigma \stackrel{\text{not.}}{=} \sqrt{Var[Z]}$ și t sunt constante strict pozitive. Funcția f este derivabilă. Calculând derivata ei de ordinul unu, obținem:

$$\begin{aligned} f'(x) &= \frac{(x^2 + \sigma^2)'(x + t)^2 - (x^2 + \sigma^2)((x + t)^2)'}{(x + t)^4} = \frac{2x(x + t)^2 - (x^2 + \sigma^2)2(x + t)}{(x + t)^4} \\ &= \frac{2(xt - \sigma^2)}{(x + t)^3}. \end{aligned}$$

Egalându-l pe $f'(x)$ cu 0, vom obține $\frac{2(xt - \sigma^2)}{(x + t)^3} = 0 \Leftrightarrow xt - \sigma^2 = 0 \Leftrightarrow x_0 = \frac{\sigma^2}{t}$. Știind că $x + t > 0, \forall x$, remarcăm faptul că $f'(x) < 0, \forall x < x_0$, respectiv $f'(x) > 0, \forall x > x_0$. Deducem astfel că punctul $x_0 = \frac{\sigma^2}{t}$ minimizează funcția f . Revenind la îmbunătățirea marginii superioare, obținem:

$$P(Z - E[Z] \geq t) \leq \min_{a>0} \frac{\sigma^2 + a^2}{(t + a)^2} = \frac{\sigma^2 + \frac{\sigma^4}{t^2}}{\left(t + \frac{\sigma^2}{t}\right)^2} = \frac{t^2\sigma^2 + \sigma^4}{(t^2 + \sigma^2)^2} = \frac{\sigma^2(t^2 + \sigma^2)}{(t^2 + \sigma^2)^2} = \frac{\sigma^2}{\sigma^2 + t^2}.$$

e. Inegalitățile lui Chernoff.

Prima inegalitate: Tinând cont de faptul că funcția exponentială este strict crescătoare (păstrând deci sensul inegalității), inegalitatea $Z - E[Z] \geq t$ este echivalentă cu $\exp(\lambda(Z - E[Z])) \geq e^{\lambda t}$, pentru orice $\lambda > 0$, deci $P(Z \geq E[Z] + t) = P(\exp(\lambda(Z - E[Z])) \geq e^{\lambda t})$. Conform inegalității lui Markov, această nouă probabilitate va fi majorată astfel:

$$P(\exp(\lambda(Z - E[Z])) \geq e^{\lambda t}) \leq \frac{E[\exp(\lambda(Z - E[Z]))]}{\exp(\lambda t)} \stackrel{\text{not.}}{=} M_{Z-E[Z]}(\lambda) e^{-\lambda t}.$$

Această inegalitate fiind valabilă pentru orice $\lambda > 0$, putem îmbunătăți marginea superioară a probabilității prin aplicarea operatorului de minimizare peste parametrul $\lambda > 0$ (și chiar pentru $\lambda \geq 0$, fiindcă $M_{Z-E[Z]}(0) e^{-0 \cdot t} = 1$). Obținem, astfel:

$$\begin{aligned} P(Z \geq E[Z] + t) &= P(Z - E[Z] \geq t) = P(\exp(\lambda(Z - E[Z])) \geq e^{\lambda t}) \\ &\leq \min_{\lambda \geq 0} \{M_{Z-E[Z]}(\lambda) e^{-\lambda t}\}. \end{aligned}$$

Cea de-a doua inegalitate: Inversăm semnul inegalității $Z \leq E[Z] - t$, ajungând astfel la $-Z \geq t - E[Z]$. Tinând cont de faptul că funcția exponentială este strict crescătoare, inegalitatea $E[Z] - Z \geq t$ este echivalentă cu $\exp(\lambda(E[Z] - Z)) \geq e^{\lambda t}$, pentru orice $\lambda > 0$, deci $P(Z \leq E[Z] - t) = P(E[Z] - Z \geq t) = P(\exp(\lambda(E[Z] -$

$Z)) \geq e^{\lambda t}$). Conform inegalității lui Markov, această nouă probabilitate va fi majorată astfel:

$$P(\exp(\lambda(E[Z] - Z)) \geq e^{\lambda t}) \leq \frac{E[\exp(\lambda(E[Z] - Z))]}{\exp(\lambda t)} \stackrel{\text{not.}}{=} M_{E[Z]-Z}(\lambda) e^{-\lambda t}.$$

Această inegalitate fiind valabilă pentru orice $\lambda > 0$, putem îmbunătăți marginea superioară a probabilității prin aplicarea operatorului de minimizare peste parametrul $\lambda > 0$ (și chiar pentru $\lambda \geq 0$, fiindcă $M_{E[Z]-Z}(0) e^{-0 \cdot t} = 1$). Obținem, astfel:

$$\begin{aligned} P(Z \leq E[Z] - t) &= P(E[Z] - Z \geq t) = P(\exp(\lambda(E[Z] - Z)) \geq e^{\lambda t}) \\ &\leq \min_{\lambda \geq 0} \{M_{E[Z]-Z}(\lambda) e^{-\lambda t}\}. \end{aligned}$$

f. Legea numerelor mari, varianta slabă.

Vom nota $Z = \frac{Z_1 + \dots + Z_n}{n}$. Știm din enunț că variabilele aleatoare Z_i sunt i.i.d. și au varianță (σ^2) finită, pentru $i \in \{1, \dots, n\}$. Tinând cont de proprietățile de liniaritate ale mediei $E[aX + bY] = aE[X] + bE[Y]$ și ale varianței $Var[aX + bY] \stackrel{\text{indep.}}{=} a^2Var[X] + b^2Var[Y]$, obținem că:

$$E[Z] = \frac{1}{n} \sum_{i=1}^n E[Z_i] = \frac{n\mu}{n} = \mu; \quad Var[Z] = \frac{1}{n^2} \sum_{i=1}^n Var[Z_i] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Probabilitatea $P\left(\left|\frac{Z_1 + \dots + Z_n}{n} - \mu\right| \geq \varepsilon\right)$ se poate scrie ca $P(|Z - E[Z]| \geq \varepsilon)$. Aplicând inegalitatea lui Chebyshev, rezultă:

$$P(|Z - E[Z]| \geq \varepsilon) \leq \frac{Var[Z]}{\varepsilon^2} = \frac{1}{n^2} \cdot n\sigma^2 \cdot \frac{1}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}.$$

Tinând cont de faptul că $\varepsilon > 0$ și σ nu depind de n , deducem că $\lim_{n \rightarrow \infty} \frac{\sigma^2}{n\varepsilon^2} = 0$. Întrucât probabilitatea nu poate fi negativă, rezultă prin trecere la limită că $0 \leq \lim_{n \rightarrow \infty} P(|Z - E[Z]| \geq \varepsilon) \leq 0$. Așadar, conform teoremei cleștelui obținem rezultatul

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{Z_1 + \dots + Z_n}{n} - \mu\right| \geq \varepsilon\right) \stackrel{\text{not.}}{=} \lim_{n \rightarrow \infty} P(|Z - E[Z]| \geq \varepsilon) = 0.$$

22.

(Inegalitățile lui Hoeffding)

□ Andi Munteanu, Liviu Ciortuz, pornind de la Stanford, Machine Learning course, John Duchi, Supplemental Lecture Notes – Hoeffding's inequality, University of Massachusetts at Amherst, Statistical ML course, 2010 spring, Justin Domke, Learning theory (Lecture Notes 10)

a. Demonstrați următorul rezultat, cunoscut sub numele de *lema lui Hoeffding*:

Dacă Z este o variabilă aleatoare mărginită, cu $Z \in [a', b']$, atunci are loc inegalitatea:

$$M_{Z-E[Z]}(\lambda) \stackrel{\text{def.}}{=} E[\exp(\lambda(Z - E[Z]))] \leq \exp\left(\frac{\lambda^2(b' - a')^2}{8}\right) \text{ pentru orice } \lambda \in \mathbb{R},$$

unde prin $M(\lambda)$ am desemnat funcția generatoare de momente (vedeți pr. 120).

b. Folosind inegalitățile lui Chernoff (vedeți problema 21.e), precum și lema lui Hoeffding, demonstrați *inegalitatea lui Hoeffding*,³³ care sunt probabil inegalitățile cel mai des folosite în *teoria învățării computaționale*:

Dacă Z_1, \dots, Z_n sunt variabile aleatoare independente și mărginite, cu $Z_i \in [a_i, b_i]$ pentru $i = 1, \dots, n$ și notăm $\bar{Z} = \frac{1}{n} \sum_{i=1}^n Z_i$, atunci pentru orice $t \geq 0$ au loc inegalitățile următoare:

$$\begin{aligned} P(\bar{Z} - E[\bar{Z}] \geq t) &\leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \\ P(E[\bar{Z}] - \bar{Z} \geq t) &\leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right), \end{aligned}$$

de unde rezultă imediat că

$$P(|\bar{Z} - E[\bar{Z}]| \geq t) \leq 2 \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \quad (10)$$

c. Arătați că dacă se consideră un număr oarecare $\delta \in (0, 1]$ — care poate fi asociat cu un anumit „nivel de încredere“ (engl., confidence) —, atunci pentru orice

$$n \geq \frac{1}{t} \sqrt{\frac{1}{2} \left(\sum_{i=1}^n (b_i - a_i)^2 \right) \ln \frac{2}{\delta}}$$

este satisfăcută inegalitatea $P(|\bar{Z} - E[\bar{Z}]| \geq t) \leq \delta$. În mod alternativ, putem afirma cu un grad de încredere de cel puțin $1 - \delta$ că, atunci când n satisfacă inegalitatea de mai sus, diferența (în modul) dintre \bar{Z} și $E[\bar{Z}]$ este de cel mult t .

Răspuns:

a. Notăm $Z - E[Z]$ cu \tilde{Z} . Rezultă că $E[\tilde{Z}] = E[Z - E[Z]] = 0$. De asemenea, notăm $a = a' - E[Z]$ și $b = b' - E[Z]$. Evident, $Z \in [a', b'] \Rightarrow E[Z] \in [a', b']$, deci $a \leq 0 \leq b$, iar $\tilde{Z} = Z - E[Z] \in [a, b]$. Pe lângă relațiile acestea, mai avem și egalitatea $b - a = b' - a'$, care se poate verifica imediat.

Stim că $e^{\lambda x}$ este o funcție convexă pentru orice $\lambda \in \mathbb{R}$. Considerând $x \in [a, b]$, vom scrie λx sub forma combinației convexe $\frac{b-x}{b-a}\lambda a + \frac{x-a}{b-a}\lambda b$ și apoi vom aplica inegalitatea lui Jensen:³⁴

³³Wassily Hoeffding (1914 - 1991), statistician, născut în Finlanda (la vremea respectivă Finlanda era ducat în imperiul țarist). Bunicii lui pe linie paternă au fost de origine daneză. Familia lui a emigrat în 1920 în Danemarca. Wassily Hoeffding a obținut doctoratul în 1940 la universitatea din Berlin. A emigrat în 1946 în USA și a activat la universitatea Carolinei de Nord la Chapel Hill. (Cf. https://en.wikipedia.org/wiki/Wassily_Hoeffding, accesat la 22 martie 2022.)

³⁴Vedeți problema 141 de la acest capitol.

$$e^{\lambda x} \leq \frac{b-x}{b-a} e^{\lambda a} + \frac{x-a}{b-a} e^{\lambda b}.$$

Inegalitatea are loc pentru orice $x \in [a, b]$, deci în particular pentru $x = \tilde{Z}$. Aplicând apoi operatorul E (medie) ambelor părți ale inegalității, vom obține:

$$M_{\tilde{Z}}(\lambda) \stackrel{\text{def.}}{=} E[\exp(\lambda \tilde{Z})] \leq \frac{b - E[\tilde{Z}]}{b-a} e^{\lambda a} + \frac{E[\tilde{Z}] - a}{b-a} e^{\lambda b} \stackrel{E[\tilde{Z}] = 0}{=} \frac{b}{b-a} e^{\lambda a} - \frac{a}{b-a} e^{\lambda b}.$$

Notăm $-\frac{a}{b-a}$ cu θ . Tinând cont de *Observația* de mai sus legată de semnul lui a și semnul lui b , rezultă că $\theta \geq 0$. Inegalitatea de mai sus poate fi rescrisă astfel:

$$M_{\tilde{Z}}(\lambda) \leq (1 - \theta) e^{\lambda a} + \theta e^{\lambda b} = e^{\lambda a} (1 - \theta + \theta e^{\lambda(b-a)}) \stackrel{\text{def. } \theta}{=} e^{-\lambda\theta(b-a)} (1 - \theta + \theta e^{\lambda(b-a)}).$$

Notând $u = \lambda(b-a)$, expresia $e^{-\lambda\theta(b-a)} (1 - \theta + \theta e^{\lambda(b-a)})$ poate fi rescrisă sub formă $e^{-\theta u} (1 - \theta + \theta e^u)$. Pentru a studia această expresie într-o manieră convenabilă, vom defini funcția $f : \mathbb{R} \rightarrow \mathbb{R}$, $f(u) = -\theta u + \ln(1 - \theta + \theta e^u)$. Verificăm dacă argumentul logaritmului este pozitiv:

$$1 - \theta + \theta e^u = \theta \left(\frac{1}{\theta} - 1 + e^u \right) \stackrel{\text{def. } \theta}{=} \theta \left(-\frac{b}{a} + e^u \right) \stackrel{a < 0, b \geq 0}{>} 0.$$

Studiem derivatele de ordinul unu și doi ale funcției f :

$$\begin{aligned} f'(u) &= -\theta + \frac{\theta e^u}{1 - \theta + \theta e^u} \\ f''(u) &= \frac{\theta e^u (1 - \theta + \theta e^u) - (\theta e^u)^2}{(1 - \theta + \theta e^u)^2} = \frac{\theta e^u}{1 - \theta + \theta e^u} \left(1 - \frac{\theta e^u}{1 - \theta + \theta e^u} \right) \stackrel{\text{not.}}{=} v(1 - v), \\ \text{unde } v &\stackrel{\text{not.}}{=} \frac{\theta e^u}{1 - \theta + \theta e^u}. \end{aligned}$$

Observăm faptul că $v - v^2$ este o funcție de gradul al doilea, concavă, având coeficienții $c_2 = -1$, $c_1 = 1$ și $c_0 = 0$. Funcția își atinge maximul în $v_0 = -\frac{c_1}{2c_2} = \frac{1}{2}$.

În consecință, $f''(u) \leq \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$, pentru orice u .

Aplicând Teorema lui Taylor, rezultă că pentru orice $u \in \mathbb{R}^*$, există un w situat între 0 și u , astfel încât

$$f(u) = f(0) + (u-0)f'(0) + (u-0)^2 \frac{f''(w)}{2!} = (0 + \ln 1) + u \left(-\theta + \frac{\theta}{1} \right) + \frac{1}{2} u^2 f''(w) = \frac{u^2 f''(w)}{2}.$$

Știind că $f''(u) \leq \frac{1}{4}$ pentru orice u , rezultă că $f(u) \leq \frac{u^2}{2} \cdot \frac{1}{4} = \frac{u^2}{8}$.

Revenind acum la inegalitatea relativă la $M_{\tilde{Z}}(\lambda)$, putem scrie:

$$\begin{aligned} M_{\tilde{Z}}(\lambda) &\leq e^{-\lambda\theta(b-a)} (1 - \theta + \theta e^{\lambda(b-a)}) = e^{f(u)} \\ &\leq \exp \left(\frac{1}{8} u^2 \right) \stackrel{\text{def. } u}{=} \exp \left(\frac{1}{8} \lambda^2 (b-a)^2 \right) = \exp \left(\frac{1}{8} \lambda^2 (b'-a')^2 \right). \end{aligned}$$

b. **Prima inegalitate:** Aplicăm prima inegalitate a lui Chernoff (varianta fără $\min_{\lambda \geq 0}$) pentru a mărgini superior probabilitatea $P(\bar{Z} - E[\bar{Z}] \geq t)$:

$$\begin{aligned} P(\bar{Z} - E[\bar{Z}] \geq t) &= P\left(\frac{1}{n} \sum_{i=1}^n (Z_i - E[Z_i]) \geq t\right) = P\left(\sum_{i=1}^n (Z_i - E[Z_i]) \geq nt\right) \\ &\stackrel{\text{Chernoff}}{\leq} E\left[\exp\left(\lambda \sum_{i=1}^n (Z_i - E[Z_i])\right)\right] e^{-n\lambda t} = E\left[\prod_{i=1}^n \exp(\lambda(Z_i - E[Z_i]))\right] e^{-n\lambda t} \\ &\stackrel{\text{pr. 120.a}}{=} e^{-n\lambda t} \prod_{i=1}^n E[\exp(\lambda(Z_i - E[Z_i]))] \stackrel{\text{not.}}{=} e^{-n\lambda t} \prod_{i=1}^n M_{Z_i - E[Z_i]}(\lambda). \end{aligned}$$

Ne vom folosi în continuare de lema lui Hoeffding, pe care o vom aplica fiecărui factor al produsului și, prin urmare, vom obține:

$$\begin{aligned} P(\bar{Z} - E[\bar{Z}] \geq t) &\leq e^{-n\lambda t} \prod_{i=1}^n \exp\left(\frac{\lambda^2(b_i - a_i)^2}{8}\right) = \exp\left(-n\lambda t + \sum_{i=1}^n \frac{\lambda^2(b_i - a_i)^2}{8}\right) \\ &= \exp\left(-n\lambda t + \frac{\lambda^2}{8} \sum_{i=1}^n (b_i - a_i)^2\right). \end{aligned} \quad (11)$$

Inegalitatea aceasta are loc pentru orice $\lambda \geq 0$ (vedeți pr. 21.e), deci putem să îmbunătățim marginea superioară prin aplicarea operatorului de minimizare peste parametrul $\lambda \geq 0$. Observăm că argumentul funcției \exp este un polinom de gradul al doilea în raport cu parametrul λ , deci de forma $c_2\lambda^2 + c_1\lambda + c_0$, având coeficienții

$$c_2 = \frac{1}{8} \sum_{i=1}^n (b_i - a_i)^2, \quad c_1 = -nt, \quad c_0 = 0.$$

Coefficientul c_2 este pozitiv, deci funcția aceasta își atinge minimul în $\lambda' = -\frac{c_1}{2c_2} = \frac{4nt}{\sum_{i=1}^n (b_i - a_i)^2}$. Prin urmare, putem deduce că marginea superioară a probabilității $P(\bar{Z} - E[\bar{Z}] \geq t)$ este următoarea:

$$\begin{aligned} P(\bar{Z} - E[\bar{Z}] \geq t) &\leq \min_{\lambda \geq 0} \exp\left(-n\lambda t + \frac{\lambda^2}{8} \sum_{i=1}^n (b_i - a_i)^2\right) \\ &= \exp\left(-n\lambda' t + \frac{\lambda'^2}{8} \sum_{i=1}^n (b_i - a_i)^2\right) = \exp\left(-\frac{4n^2 t^2}{\sum_{i=1}^n (b_i - a_i)^2} + \frac{16n^2 t^2}{8 \sum_{i=1}^n (b_i - a_i)^2}\right) \\ &= \exp\left(-\frac{2n^2 t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \end{aligned} \quad (12)$$

A doua inegalitate: Aplicăm inegalitatea două a lui Chernoff (varianta fără $\min_{\lambda \geq 0}$) pentru a mărgini superior probabilitatea $P(E[\bar{Z}] - \bar{Z} \geq t)$:

$$\begin{aligned} P(E[\bar{Z}] - \bar{Z} \geq t) &= P\left(\frac{1}{n} \sum_{i=1}^n (E[Z_i] - Z_i) \geq t\right) = P\left(\sum_{i=1}^n (E[Z_i] - Z_i) \geq nt\right) \\ &\stackrel{\text{Chernoff}}{\leq} E\left[\exp\left(\lambda \sum_{i=1}^n (E[Z_i] - Z_i)\right)\right] e^{-n\lambda t} = E\left[\prod_{i=1}^n \exp(\lambda(E[Z_i] - Z_i))\right] e^{-n\lambda t} \\ &\stackrel{\text{pr. 120.a}}{=} e^{-n\lambda t} \prod_{i=1}^n E[\exp(\lambda(E[Z_i] - Z_i))] \stackrel{\text{not.}}{=} e^{-n\lambda t} \prod_{i=1}^n M_{E[Z_i] - Z_i}(\lambda). \end{aligned}$$

Vom folosi în din nou de lema lui Hoeffding (ținem cont de faptul că lema este valabilă pentru orice $\lambda \in \mathbb{R}$ și de faptul că $M_{E[Z_i] - Z_i}(\lambda)$, cu un λ pozitiv, este echivalent cu $M_{Z_i - E[Z_i]}(\lambda)$, cu un λ negativ), pe care o vom aplica pentru fiecare factor al produsului. Vom obține:

$$\begin{aligned} P(E[\bar{Z}] - \bar{Z} \geq t) &\leq e^{-n\lambda t} \prod_{i=1}^n \exp\left(\frac{\lambda^2(b_i - a_i)^2}{8}\right) = \exp\left(-n\lambda t + \sum_{i=1}^n \frac{\lambda^2(b_i - a_i)^2}{8}\right) \\ &= \exp\left(-n\lambda t + \frac{\lambda^2}{8} \sum_{i=1}^n (b_i - a_i)^2\right). \end{aligned} \quad (13)$$

Că și mai sus, inegalitatea aceasta are loc pentru orice $\lambda \geq 0$, deci putem să îmbunătățim marginea superioară prin aplicarea operatorului de minimizare peste parametrul $\lambda \geq 0$. Observăm că expresiile (13) și (11) sunt identice, deci au aceeași valoare minimă, și anume (12). Prin urmare,

$$P(E[\bar{Z}] - \bar{Z} \geq t) \leq \min_{\lambda \geq 0} \exp\left(-n\lambda t + \frac{\lambda^2}{8} \sum_{i=1}^n (b_i - a_i)^2\right) = \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

c. Impunând condiția ca membrul drept al inegalității (10) să fie mai mic sau egal cu δ , vom obține:

$$\begin{aligned} 2 \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \leq \delta &\Leftrightarrow \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \leq \frac{\delta}{2} \Leftrightarrow \\ -\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2} &\leq \ln \frac{\delta}{2} \Leftrightarrow \frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2} \geq \ln \frac{2}{\delta} \Leftrightarrow \\ n^2 \geq \frac{\sum_{i=1}^n (b_i - a_i)^2}{2t^2} \ln \frac{2}{\delta} &\Leftrightarrow n \geq \frac{1}{t} \sqrt{\frac{1}{2} \sum_{i=1}^n (b_i - a_i)^2 \ln \frac{2}{\delta}}. \end{aligned}$$

23.

(Variabile aleatoare: Adevărat sau Fals?)

*CMU, 2006 fall, E. Xing, T. Mitchell, final exam, pr. 1.b
CMU, 2008 fall, Eric Xing, midterm exam, pr. 1.1, 1.2, 1.3*

- a. Dacă o variabilă aleatoare continuă X are funcția densitate de probabilitate p diferită de zero pe tot domeniul de definiție, atunci probabilitatea ca X să ia o valoare oarecare x (notație: $P(X = x)$) este egală cu $p(x)$.
- b. $E[X + Y] = E[X] + E[Y]$ pentru orice două variabile aleatoare X și Y .
- c. $Var[X + Y] = Var[X] + Var[Y]$ pentru orice două variabile aleatoare X și Y .
- d. $E[XY] = E[X] \cdot E[Y]$ pentru orice două variabile aleatoare X și Y .

Răspuns:

- a. Fals (în general). Dacă variabila aleatoare continuă X are funcția densitate de probabilitate p , atunci $P(a \leq X \leq b) = \int_a^b p(x)dx$. Prin urmare, $P(X = x) = 0$ pentru orice $x \in \mathbb{R}$. Enunțul afirmă pe de o parte că $p(x) = P(X = x) = 0$ pentru

un anume $x \in \mathbb{R}$, iar pe de altă parte că p este diferită de zero pe tot domeniul de definiție, ceea ce este absurd.

b. Adevărat. Demonstrația este făcută în problema 9 punctul a.

c. Fals (în general). Se poate considera situația $Y = -X$, caz în care $\text{Var}[X + Y] = 0$, dar $\text{Var}[X] + \text{Var}[Y] = E[X^2] - (E[X])^2 + E[(-X)^2] - (E[-X])^2 = 2\text{Var}[X]$. Așadar, pentru orice variabilă aleatoare X cu $\text{Var}[X] \neq 0$, luând $Y = -X$, rezultă că $\text{Var}[X + Y] = 0$, iar $\text{Var}[X] + \text{Var}[Y] \neq 0$. Un exemplu de variabilă aleatoare cu varianță nenulă este distribuția gaussiană.

Mai general, se poate demonstra că $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y]$ astfel:

$$\begin{aligned}\text{Var}[X + Y] &= E[(X + Y)^2] - (E[X + Y])^2 \\ &= E[X^2 + 2XY + Y^2] - (E[X] + E[Y])^2 \\ &= E[X^2] + 2E[XY] + E[Y^2] - (E[X])^2 - 2E[X] \cdot E[Y] - (E[Y])^2 \\ &= (E[X^2] - (E[X])^2) + (E[Y^2] - (E[Y])^2) + (2E[XY] - 2E[X] \cdot E[Y]) \\ &= \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y]\end{aligned}$$

Prin urmare, egalitatea din enunț este adevărată pentru orice două variabile aleatoare X și Y pentru care $\text{Cov}[X, Y] = 0$ (vedeți problema 9.c), dar este falsă în rest. Egalitatea $\text{Cov}[X, Y] = 0$ este adevărată de exemplu atunci când X și Y sunt variabile independente (vedeți problema 10).

d. Fals, în general. Afirmația din enunț este echivalentă cu $\text{Cov}[X, Y] = 0$. Așa cum am menționat la punctul c, ea este adevărată dacă, de exemplu X și Y sunt variabile aleatoare independente. Dacă, în schimb, vom considera de pildă cazul $Y = X$, cu X variabilă aleatoare binară care ia valoarea 1 cu probabilitatea p și valoarea 0 cu probabilitatea $1 - p$, se poate deduce imediat că $E[X^2] \neq (E[X])^2$, fiindcă $p(1 - p) \neq 0$ pentru orice $p \in (0, 1)$.

0.1.3 Distribuții probabiliste uzuale

24.

(Distribuția Bernoulli;
variabile aleatoare identic distribuite:
calcul de valori medii folosind proprietatea de liniaritate a mediilor)

• CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW1, pr. 2.2

Un iepuraș se joacă „de-a sărilea“. Poziția lui inițială coincide cu originea axei reale. După aceea, iepurașul face câte un salt de-a lungul axei, fie la stânga fie la dreapta. Pentru a determina în ce direcție să sară, iepurașul dă cu banul. Dacă obține stema, va sări spre dreapta, iar dacă obține banul, va sări spre stânga. Probabilitatea de a obține stema este p . Se presupune că toate salturile iepurașului au aceeași lungime, și anume 1.

Care va fi poziția (medie) la care ne aşteptăm să fie iepurașul după ce face n salturi?

Răspuns:

Fiecare săritură a iepurașului este modelată de o variabilă aleatoare. Un salt spre dreapta înseamnă o deplasare cu $+1$ pe axă, iar un salt spre stânga -1 . Să notăm cu X_i variabila aleatoare corespunzătoare săriturii i . Aceasta este:

$$X_i : \begin{pmatrix} -1 & 1 \\ 1-p & p \end{pmatrix}$$

Media acestei variabile aleatoare este $E[X_i] = -1 \cdot (1-p) + 1 \cdot p = 2p - 1$.

Tinând cont de proprietatea de liniaritate a mediilor (vedeți pr. 9.a), poziția iepurașului după n salturi este de așteptat să fie:

$$E[X_1 + X_2 + \dots + X_n] = E[X_1] + E[X_2] + \dots + E[X_n] = n(2p - 1).$$

De exemplu, pentru $p = 1/2$, se va obține poziția 0 pe axa reală (așa cum este de așteptat dacă n este număr par), în vreme ce pentru $p = 2/3$ va rezulta poziția $n/3$ (dacă $n/3 \in \mathbb{N}$), iar pentru $p = 1/3$ poziția $-n/3$ (similar).

25.

(Distribuția binomială:
verificarea condițiilor de definiție pentru p.m.f.;
calculul mediei și al varianței)

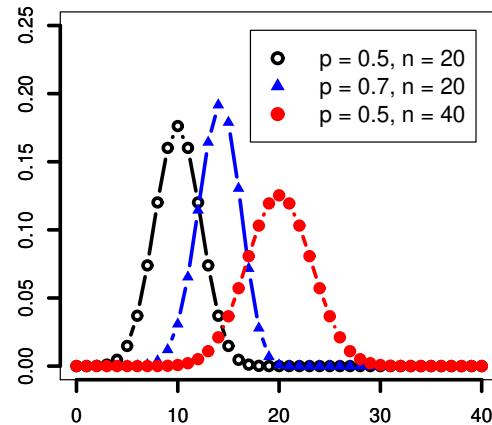
■ □ Liviu Ciortuz, 2015

Distribuția binomială de parametri n și p are funcția masă de probabilitate (engl., probability mass function, p.m.f.) definită astfel:

$$b(r; n, p) = C_n^r p^r (1-p)^{n-r} \quad \forall r \in \{0, \dots, n\}.$$

Distribuția binomială: p.m.f.

Vă reamintim că $b(r; n, p)$ este probabilitatea care corespunde numărului (r) de apariții ale feței cu stema (corespondent engl., *head*) obținute la efectuarea a n aruncări independente ale unei monede, atunci când se presupune că probabilitatea de apariție a stemei la o aruncare oarecare a acestei monede este p .



- a. Verificați că funcția $b(r; n, p)$, așa cum a fost definită mai sus, reprezintă într-adevăr o funcție masă de probabilitate. Aceasta revine la a arăta că $b(r; n, p) \geq 0$ pentru orice $p \in [0, 1]$, $n \in \mathbb{N}$ și $r \in \{0, 1, \dots, n\}$, iar $\sum_{r=0}^n b(r; n, p) = 1$ pentru orice astfel de n și p , fixați.

- b. Calculați media și varianța distribuției binomiale.

Răspuns:

a. Evident, $b(r; n, p) \stackrel{\text{def.}}{=} C_n^r p^r (1-p)^{n-r} \geq 0$ pentru orice $p \in [0, 1]$, $n \in \mathbb{N}$ și $r \in \{0, 1, \dots, n\}$, iar

$$\begin{aligned} \sum_{r=0}^n b(r; n, p) &= (1-p)^n + C_n^1 p(1-p)^{n-1} + \dots + C_n^{n-1} p^{n-1}(1-p) + p^n \\ &= [p + (1-p)]^n = 1 \end{aligned}$$

b. Calculul mediei se poate face pornind de la definiție:

$$\begin{aligned} E[b(r; n, p)] &\stackrel{\text{def.}}{=} \sum_{r=0}^n r \cdot b(r; n, p) = \\ &= 1 \cdot C_n^1 p(1-p)^{n-1} + 2 \cdot C_n^2 p^2(1-p)^{n-2} + \dots + (n-1) \cdot C_n^{n-1} p^{n-1}(1-p) + n \cdot p^n \\ &= p [C_n^1 (1-p)^{n-1} + 2 \cdot C_n^2 p(1-p)^{n-2} + \dots + (n-1) \cdot C_n^{n-1} p^{n-2}(1-p) + n \cdot p^{n-1}] \\ &= np [(1-p)^{n-1} + C_{n-1}^1 p(1-p)^{n-2} + \dots + C_{n-1}^{n-2} p^{n-2}(1-p) + C_{n-1}^{n-1} p^{n-1}] \quad (14) \\ &= np[p + (1-p)]^{n-1} = np = nE[Bernoulli(p)]. \quad (15) \end{aligned}$$

Pentru egalitatea (14) am folosit faptul că

$$\begin{aligned} k C_n^k &= k \frac{n!}{k!(n-k)!} = \frac{n!}{(k-1)!(n-k)!} = \frac{n(n-1)!}{(k-1)!(n-1-(k-1))!} \\ &= n C_{n-1}^{k-1}, \forall k = 1, \dots, n. \end{aligned}$$

Pentru calculul varianței, vom folosi formula $Var[X] = E[X^2] - E^2[X]$, care a fost demonstrată la problema 9.b. Întrucât am calculat deja $E[b(r; n, p)]$, rămâne să calculăm $E[b^2(r; n, p)]$.³⁵ Notând $q = 1 - p$, vom avea:

$$\begin{aligned} E[b^2(r; n, p)] &\stackrel{\text{def.}}{=} \sum_{r=0}^n r^2 C_n^r p^r q^{n-r} = \sum_{r=0}^n r^2 \frac{n(n-1)\dots(n-r+1)}{r!} p^r q^{n-r} \\ &= \sum_{r=1}^n rn \frac{(n-1)\dots(n-r+1)}{(r-1)!} p^r q^{n-r} = \sum_{r=1}^n rn C_{n-1}^{r-1} p^r q^{n-r} \\ &= np \sum_{r=1}^n r C_{n-1}^{r-1} p^{r-1} q^{(n-1)-(r-1)}. \end{aligned}$$

Mai departe, notând pentru conveniență $r - 1$ cu j , urmează:

$$\begin{aligned} E[b^2(r; n, p)] &= np \sum_{j=0}^{n-1} (j+1) C_{n-1}^j p^j q^{(n-1)-j} \\ &= np \left[\sum_{j=0}^{n-1} j C_{n-1}^j p^j q^{(n-1)-j} + \sum_{j=0}^{n-1} C_{n-1}^j p^j q^{(n-1)-j} \right]. \end{aligned}$$

Prima sumă din paranteza patrată de mai sus este chiar $E[b(r; n-1, p)]$, conform relației (15), iar cea de-a doua sumă este egală cu 1, conform unui calcul absolut similar cu cel de la punctul a. Prin urmare,

$$E[b^2(r; n, p)] = np[(n-1)p + 1] = n^2 p^2 - np^2 + np.$$

³⁵Rezolvarea de mai jos urmează îndeaproape linia demonstrației găsite pe site-ul www.proofwiki.org/wiki/Variance_of_Binomial_Distribution, accesat la data de 5 octombrie 2015. La rândul său, acest site menționează ca sursă "Probability: An Introduction" de Geoffrey Grimmett și Dominic Welsh, Oxford Science Publications, 1986.

Așadar, $Var[X] = E[b^2(r; n, p)] - (E[b(r; n, p)])^2 = n^2p^2 - np^2 + np - n^2p^2 = np(1 - p)$.

Observație: O altă cale de a calcula varianța distribuției binomiale este următoarea:

- se demonstrează relativ ușor că orice variabilă aleatoare urmând distribuția binomială $b(r; n, p)$ poate fi văzută ca o sumă de n variabile independente care urmează distribuția Bernoulli de parametru p ;³⁶
- se știe (sau, se poate dovedi imediat) că varianța distribuției Bernoulli de parametru p este $p(1 - p)$;
- ținând cont de proprietatea $Var[X_1 + X_2 + \dots + X_n] = Var[X_1] + Var[X_2] + \dots + Var[X_n]$ atunci când X_1, X_2, \dots, X_n sunt variabile independente, conform demonstrației de la problema 23.c, rezultă că $Var[X] = np(1 - p)$.

26.

(Distribuția categorială; distribuția binomială:
calcularea unor probabilități, medii, numere „medii“ etc.)

* UAIC, 2020 fall, Stefan Bălăucă, student

Mickey vrea să își cumpere un nou zar, dar se teme că și acesta ar putea fi măsluit. Vânzătorul îi spune că din noul lot, o fracțiune f dintre zaruri sunt măsluite având distribuția de probabilitate $P = [0.5, 0.1, 0.1, 0.1, 0.1, 0.1]$.

Pentru a testa dacă zarul pe care l-a ales este sau nu măsluit, Mickey face următorul experiment: aruncă zarul de n ori și numără de câte ori cade față cu numărul 1. Notează acest număr cu a . Pentru a-l ajuta pe Mickey să decidă dacă să cumpere acest zar, ne propunem să răspundem la următoarele întrebări:

- a. Considerăm variabila aleatoare $S =$ numărul de aruncări la care se obține față / valoarea 1. Care este distribuția de probabilitate a variabilei S dacă zarul este măsluit? Dar dacă zarul este corect?
- b. Care este media variabilei aleatoare S în fiecare din cele două cazuri?
- c. Care este probabilitatea *a posteriori* ca zarul să fie măsluit, dacă Mickey obține în exact a din cele n aruncări valoarea 1?
- d. Câte dintre cele n aruncări trebuie să aibă valoarea 1 pentru ca Mickey să poată afirma că zarul este măsluit, cu probabilitate de cel puțin p ?
- e. Presupunând că zarul este măsluit, care este numărul *minim* de aruncări (n) necesare pentru a putea afirma că acesta este măsluit, cu probabilitate de cel puțin p ? (Veți da răspunsul în funcție de f și p .)
- f. Presupunând că zarul este măsluit, care este numărul *mediu* de aruncări necesare pentru a putea afirma că acesta este măsluit, cu probabilitate de cel puțin p ? (Veți da răspunsul în funcție de f și p .)
- g. Aplicație numerică pentru punctele e și f : $f = 1\%$ și 5% , iar $p = 50\%, 95\%$ și 99% .

³⁶Vedeți www.proofwiki.org/wiki/Bernoulli_Process_as_Binomial_Distribution, care citează ca sursă “Probability: An Introduction” de Geoffrey Grimmett și Dominic Welsh, Oxford Science Publications, 1986.

Răspuns:

a. Considerăm variabilele aleatoare $X_k =$ valoarea zarului la aruncarea cu numărul k și evenimentele aleatoare $Y_k = (X_k \text{ este } 1)$ și $M =$ zarul este măsluit. Atunci, $P(Y_k|\neg M) = P(X_k = 1|\neg M) = 1/6$ și $P(Y_k|M) = P(X_k = 1|M) = 1/2$.

Valoarea variabilei aleatoare S este a (adică, $S = a$) dacă și numai dacă mulțimea $A = \{k | X_k = 1\}$ are cardinalul $|A| = a$. Cum fiecare dintre configurațiile posibile ale mulțimii A sunt disjuncte (ca evenimente aleatoare), putem scrie

$$P(S = a|\neg M) = \sum_{A, |A|=a} \left(\prod_{k \in A} P(Y_k|\neg M) \prod_{k \notin A} P(\neg Y_k|\neg M) \right) = \sum_{A, |A|=a} \left((1/6)^a (5/6)^{n-a} \right),$$

deci

$$P(S = a|\neg M) = C_n^a (1/6)^a (5/6)^{n-a}.$$

Similar, putem scrie

$$P(S = a|M) = \sum_{A, |A|=a} \left(\prod_{k \in A} P(Y_k|M) \prod_{k \notin A} P(\neg Y_k|M) \right) = \sum_{A, |A|=a} \left((1/2)^a (1/2)^{n-a} \right),$$

deci

$$P(S = a|M) = \sum_{A, |A|=a} (1/2)^n = C_n^a (1/2)^n.$$

Putem observa că datorită construcției lui S , distribuțiile de probabilitate condiționale $S = a|\neg M$ și $S = a|M$ sunt de tip binomial.

b. Folosind formula mediei pentru distribuțiile binomiale,³⁷ vom avea:

$$E[S|\neg M] = n \cdot P(Y_k|\neg M) = n/6, \text{ iar } E[S|M] = n \cdot P(Y_k|M) = n/2.$$

c. Aplicând formula lui Bayes și ținând cont de informația din enunț că probabilitatea ca un zar oarecare să fie măsluit este f , găsim probabilitatea cerută:

$$P(M|S = a) = \frac{P(S = a|M) \cdot P(M)}{P(S = a)} = \frac{P(S = a|M) \cdot f}{P(S = a|M) \cdot f + P(S = a|\neg M) \cdot (1-f)}.$$

Pe baza relațiilor deduse la punctul a, putem scrie:

$$\begin{aligned} P(M|S = a) &= \frac{C_n^a (1/2)^n \cdot f}{C_n^a (1/2)^n \cdot f + C_n^a (1/6)^a (5/6)^{n-a} \cdot (1-f)} \\ &= \frac{f}{f + \frac{5^{n-a}}{3^n} (1-f)} = \frac{3^n f}{3^n f + 5^{n-a} (1-f)}. \end{aligned}$$

d. Pentru a putea afirma că zarul este măsluit cu probabilitate de cel puțin p trebuie să avem $P(M|S = a) \geq p$. Pe baza relației deduse la punctul c, avem:

$$\begin{aligned} \frac{3^n f}{3^n f + 5^{n-a} (1-f)} \geq p &\Leftrightarrow 3^n f \geq 3^n f p + 5^{n-a} (1-f)p \Leftrightarrow 3^n f (1-p) \geq 5^{n-a} (1-f)p \\ &\Leftrightarrow 5^a \geq \left(\frac{5}{3}\right)^n \frac{1-f}{f} \frac{p}{1-p} \Leftrightarrow a \ln 5 \geq n(\ln 5 - \ln 3) + \ln \frac{1-f}{f} + \ln \frac{p}{1-p} \\ &\Leftrightarrow a \geq n \left(1 - \frac{\ln 3}{\ln 5}\right) + \frac{1}{\ln 5} \left(\ln \frac{1-f}{f} + \ln \frac{p}{1-p}\right). \end{aligned}$$

³⁷Vedeți ex. 25.b.

Observăm că din punct de vedere *asimptotic*, [adică] atunci când $n \rightarrow \infty$, raportul a/n trebuie să fie cel puțin egal cu $1 - \frac{\ln 3}{\ln 5} \approx 0.317$. Din ultima inegalitate de mai sus putem trage concluzia că indiferent de valoarea lui f , alegând n suficient de mare, vom putea efectua teste pentru a verifica dacă zarul este sau nu măsluit, cu orice precizie p .

e. Numărul minim de aruncări se va obține dacă toate aruncările au valoarea 1 (dacă $a = n$). În acest caz trebuie ca $P(M|S = n) \geq p$. Pe baza relațiilor de la punctul anterior, putem scrie:

$$\begin{aligned} 3^n f(1-p) &\geq 5^{n-n}(1-f)p \Leftrightarrow 3^n \geq \frac{1-f}{f} \frac{p}{1-p} \Leftrightarrow \\ n &\geq \frac{1}{\ln 3} \left(\ln \frac{1-f}{f} + \ln \frac{p}{1-p} \right). \end{aligned}$$

f. La punctul b am arătat că $E[S|M] = n/2$, deci $a \approx n/2$ în acest caz. Înlocuind în relațiile de la punctul d, putem scrie:

$$\begin{aligned} 3^n f(1-p) &\geq 5^{n-n/2}(1-f)p \Leftrightarrow 3^n f(1-p) \geq 5^{n/2}(1-f)p \Leftrightarrow \\ 3^n f(1-p) &\geq (\sqrt{5})^n(1-f)p \Leftrightarrow \left(\frac{3}{\sqrt{5}} \right)^n \geq \frac{1-f}{f} \frac{p}{1-p} \Leftrightarrow \\ n &\geq \frac{1}{\ln \frac{3}{\sqrt{5}}} \left(\ln \frac{1-f}{f} + \ln \frac{p}{1-p} \right). \end{aligned}$$

g. Începem prin a calcula valorile logaritmilor:

$$\begin{aligned} f = 5\% &\rightarrow \ln \frac{1-f}{f} = \ln \frac{95}{5} = \ln 19 \approx 2.94 \\ f = 1\% &\rightarrow \ln \frac{1-f}{f} = \ln \frac{99}{1} = \ln 99 \approx 4.60 \\ p = 50\% &\rightarrow \ln \frac{p}{1-p} = \ln \frac{50}{50} = \ln 1 = 0 \\ p = 95\% &\rightarrow \ln \frac{p}{1-p} = \ln \frac{95}{5} = \ln 19 \approx 2.94 \\ p = 99\% &\rightarrow \ln \frac{p}{1-p} = \ln \frac{99}{1} = \ln 99 \approx 4.60. \end{aligned}$$

Valorile minime pentru n se calculează folosind formula de la punctul e, iar valorile medii pentru n se calculează folosind formula de la punctul f:

n_{min}	$f = 5\%$	$f = 1\%$
$p = 50\%$	2.68	4.18
$p = 95\%$	5.36	6.86
$p = 99\%$	6.86	8.36

n_{med}	$f = 5\%$	$f = 1\%$
$p = 50\%$	10.02	15.63
$p = 95\%$	20.03	25.65
$p = 99\%$	25.65	31.27

27.

(Distribuția Poisson:
verificarea condițiilor de definiție pentru p.m.f.;
calculul mediei și a varianței)

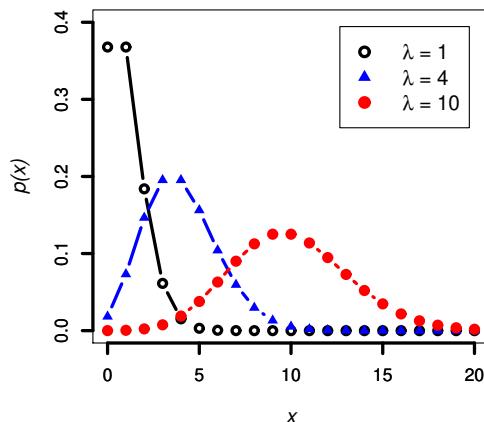
Liviu Ciortuz, 2017

Distribuția Poisson: p.m.f.

Distribuția Poisson este o distribuție discretă de parametru $\lambda > 0$, a cărei funcție masă de probabilitate este dată de expresia

$$p(x | \lambda) = \frac{1}{e^\lambda} \cdot \frac{\lambda^x}{x!}, \text{ pentru orice } x \in \mathbb{N}.$$

Prin convenție, se consideră că $0! = 1$. Factorul $\frac{1}{e^\lambda}$, care nu depinde de x , este aşa-numita *constantă de normalizare*.



Demonstrați mai întâi că funcția $p(\)$ este într-adevăr funcție masă de probabilitate (engl., probability mass function, p.m.f.) și apoi că media acestei distribuții este λ , iar varianța ei este tot λ .

Sugestie: Țineți cont că $\sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^\lambda$ (limită fundamentală).

Răspuns:

Vom arăta mai întâi că $p(\)$ este într-adevăr o funcție masă de probabilitate (p.m.f.). Evident, $p(x|\lambda) > 0$ pentru orice $x \in \mathbb{N}$, fiindcă $\lambda > 0$. Apoi,

$$\sum_{x \in \mathbb{N}} \frac{1}{e^\lambda} \frac{\lambda^x}{x!} = \frac{1}{e^\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = \frac{1}{e^\lambda} e^\lambda = 1,$$

ținând cont de limita fundamentală care apare în *Sugestia* din enunț.

Pentru calcularea mediei acestei distribuții, aplicăm definiția:

$$\sum_{x=0}^{\infty} x \frac{1}{e^\lambda} \frac{\lambda^x}{x!} = \frac{1}{e^\lambda} \sum_{x=0}^{\infty} x \frac{\lambda^x}{x!} = \frac{1}{e^\lambda} \sum_{x=1}^{\infty} \frac{\lambda^x}{(x-1)!} = \frac{\lambda}{e^\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} = \underbrace{\frac{\lambda}{e^\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}}_{e^\lambda} = \frac{\lambda}{e^\lambda} e^\lambda = \lambda.$$

În ce privește calculul varianței pentru distribuția Poisson, știm că $Var[X] = E[X^2] - E^2[X]$ (proprietate demonstrată la problema 9.b) pentru orice distribuție aleatoare X și, prin urmare, pentru că am calculat deja media acestei distribuții, trebuie să mai calculăm $\sum_{x=0}^{\infty} x^2 p(x|\lambda)$.

$$\begin{aligned} \sum_{x=0}^{\infty} x^2 p(x|\lambda) &= \sum_{x=0}^{\infty} x^2 \frac{1}{e^\lambda} \frac{\lambda^x}{x!} = \frac{1}{e^\lambda} \sum_{x=1}^{\infty} x^2 \frac{\lambda^x}{x!} = \frac{1}{e^\lambda} \left[\sum_{x=1}^{\infty} [x(x-1) + x] \frac{\lambda^x}{x!} \right] \\ &= \frac{1}{e^\lambda} \left[\sum_{x=1}^{\infty} x(x-1) \frac{\lambda^x}{x!} + \sum_{x=1}^{\infty} x \frac{\lambda^x}{x!} \right] = \frac{1}{e^\lambda} \left[\sum_{x=2}^{\infty} \frac{\lambda^x}{(x-2)!} + \sum_{x=1}^{\infty} \frac{\lambda^x}{(x-1)!} \right] \\ &= \frac{1}{e^\lambda} \left[\lambda^2 \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} + \lambda \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \right] = \frac{1}{e^\lambda} \left[\lambda^2 \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} + \lambda \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \right] \end{aligned}$$

$$= \frac{1}{e^\lambda} [\lambda^2 e^\lambda + \lambda e^\lambda] = \frac{1}{e^\lambda} e^\lambda (\lambda^2 + \lambda) = \lambda^2 + \lambda.$$

Prin urmare, varianța distribuției Poisson este $\lambda^2 + \lambda - \lambda^2 = \lambda$.

28.

(Distribuția geometrică:
numărul „așteptat“ / mediu de „observații“ necesare
pentru ca un anumit eveniment să se producă)

• CMU, 2012 spring, Ziv Bar-Joseph, HW1, pr. 1.4

În cazul unui zar perfect cu șase fețe, probabilitățile de apariție pentru fiecare dintre fețele zarului sunt egale.

Mickey se duce la un cazar și dorește ca, folosindu-și cunoștințele din domeniul probabilităților, să-și evalueze șansa pe care o are de a obține la aruncarea unui astfel de zar față 6.

Mai precis, Mickey se întrebă care este numărul mediu (sau, numărul „așteptat“; engl., expected number) de aruncări ale zarului pe care ar trebui să le efectueze până să obțină față 6.

Justificați răspunsul în detaliu.

Răspuns:

Experimentul lui Mickey poate fi modelat cu ajutorul *distribuției geometrice*.³⁸ Stîm că distribuția geometrică poate fi reprezentată de o tabelă de „repartiție“ de forma

$$\begin{pmatrix} 1 & 2 & 3 & \dots & n & \dots \\ q & pq & p^2q & \dots & p^{n-1}q & \dots \end{pmatrix},$$

unde $p, q \geq 0$ și $p + q = 1$. Se verifică imediat că $\sum_{i=1}^{\infty} p^{i-1}q = 1$ atunci când $p \in [0, 1)$. În particular, pentru $p = \frac{5}{6}$ și $q = \frac{1}{6}$, avem

$$\sum_{i=1}^{\infty} \frac{1}{6} \cdot \left(\frac{5}{6}\right)^{i-1} = 1. \quad (16)$$

Vom nota cu E numărul mediu de aruncări ale zarului pe care ar trebui să le efectueze Mickey până să obțină față 6.

Conform definiției mediei, E se poate exprima astfel:

$$E = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} \cdot \frac{5}{6} + 3 \cdot \frac{1}{6} \cdot \left(\frac{5}{6}\right)^2 + \dots + n \cdot \frac{1}{6} \cdot \left(\frac{5}{6}\right)^{n-1} + \dots$$

Punând termenul generic $n \cdot \frac{1}{6} \cdot \left(\frac{5}{6}\right)^{n-1}$ sub forma $(1 + (n - 1)) \cdot \frac{1}{6} \cdot \left(\frac{5}{6}\right)^{n-1}$ și aplicând apoi distributivitatea înmulțirii față de adunare, putem scrie E în mod echivalent astfel:

$$E = \frac{1}{6} + \frac{5}{6} \cdot \left[\frac{1}{6} + \frac{1}{6} \cdot \frac{5}{6} + \dots + \frac{1}{6} \cdot \left(\frac{5}{6}\right)^{n-1} + \dots \right]$$

³⁸ Simplu spus, distribuția geometrică poate fi gândită ca modelând următorul experiment aleatoriu: Fie o monedă a cărei probabilitate de apariție a feței-stemă este p . Aruncăm moneda o dată sau de mai multe ori, până când apare stema. Notăm numărul de aruncări care au precedat apariția stemei cu k . Acest număr $k \in \{0, 1, \dots\}$ va fi [asociat cu] valoarea unei variabile aleatoare X , despre care spunem că urmează distribuția geometrică. Evident, $P(X = k) = (1 - p)^k p$.

$$+ \frac{5}{6} \cdot \left[1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} \cdot \frac{5}{6} + \dots + (n-1) \cdot \frac{1}{6} \cdot \left(\frac{5}{6}\right)^{n-1} + \dots \right].$$

Tinând cont de relația (16), egalitatea precedentă se scrie astfel:

$$E = \frac{1}{6} + \frac{5}{6} \cdot 1 + \frac{5}{6} \cdot E.$$

Rezultă imediat că $\frac{1}{6}E = \frac{1}{6} + \frac{5}{6}$, deci $E = 6$.

29.

(O mixtură de distribuții categoriale: calculul mediei și al varianței)

■ □ • ○ CMU, 2010 fall, Aarti Singh, HW1, pr. 2.2.1-2

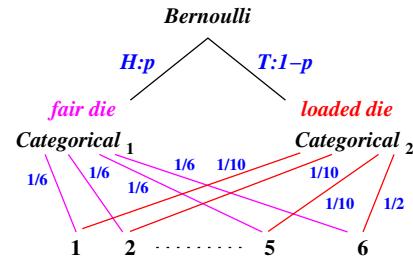
Presupunem că avem două zaruri cu șase fețe. Un zar este perfect — deci vom considera funcția sa masă de probabilitate (p.m.f.) definită prin $P_1(x) = 1/6$ pentru $x = 1, \dots, 6$ —, iar celălalt zar este măsluit și are următoarea funcție masă de probabilitate:

$$P_2(x) = \begin{cases} \frac{1}{2} & \text{pentru } x = 6; \\ \frac{1}{10} & \text{pentru } x \in \{1, 2, 3, 4, 5\}. \end{cases}$$

Pentru a decide ce zar să aruncăm, vom folosi o monedă; considerăm că probabilitatea să obținem stema la aruncarea monedei este $p \in (0, 1)$. Dacă obținem stema (engl., head), vom arunca apoi zarul perfect; în caz contrar vom arunca zarul măsluit.

Putem reprezenta grafic „mixtura“ formată din cele două distribuții categoriale ca în figura alăturată.

Notând cu X variabila aleatoare corespunzătoare acestei mixturi și cu P funcția ei masă de probabilitate (p.m.f.), putem scrie:



$$P(i) = P(i|fair) \cdot p + P(i|loaded) \cdot (1 - p) = P_1(i) \cdot p + P_2(i) \cdot (1 - p) \text{ pentru } i = 1, \dots, 6.$$

- Calculați în funcție de p media variabilei aleatoare X .
- Calculați în funcție de p varianța variabilei aleatoare X .

Răspuns:

a. Tinând cont de modul în care a fost definită în enunț mixtura celor două distribuții categoriale, putem calcula media variabilei X , care reprezintă această mixtură, în felul următor:

$$\begin{aligned} E[X] &= \sum_{i=1}^6 i \cdot P(i) = \sum_{i=1}^6 i \cdot [P_1(i) \cdot p + P_2(i) \cdot (1 - p)] \\ &= \left[\sum_{i=1}^6 i \cdot P_1(i) \right] p + \left[\sum_{i=1}^6 i \cdot P_2(i) \right] (1 - p) = \frac{7}{2} \cdot p + \frac{9}{2} \cdot (1 - p) = \frac{9}{2} - p. \end{aligned}$$

Am folosit formula $\sum_{i=1}^n i = \frac{n(n+1)}{2}$. La punctul următor vom avea nevoie de o altă formulă: $\sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6}$.

b. Conform formulei $Var(X) = E[X^2] - (E[X])^2$, care a fost demonstrată la problema 9.b, va trebui să calculăm $E[X^2]$. Procedăm astfel:

$$\begin{aligned} E[X^2] &= \sum_{i=1}^6 i^2 \cdot P(i) = \sum_{i=1}^6 i^2 \cdot [P_1(i) \cdot p + P_2(i) \cdot (1-p)] \\ &= \left[\sum_{i=1}^6 i^2 \cdot P_1(i) \right] p + \left[\sum_{i=1}^6 i^2 \cdot P_2(i) \right] (1-p) = \frac{91}{6} \cdot p + \left(\frac{55}{10} + \frac{36}{2} \right) \cdot (1-p) \\ &= \frac{47}{2} - \frac{25}{3} \cdot p. \end{aligned}$$

Combinând acest rezultat cu cel pe care l-am obținut la punctul precedent, vom obține:

$$\begin{aligned} Var(X) &= E[X^2] - (E[X])^2 = \frac{47}{2} - \frac{25}{3}p - \left(\frac{9}{2} - p \right)^2 = \frac{47}{2} - \frac{25}{3}p - \left(\frac{81}{4} - 9p + p^2 \right) \\ &= \left(\frac{47}{2} - \frac{81}{4} \right) - \left(\frac{25}{3} - 9 \right) \cdot p - p^2 = \frac{13}{4} + \frac{2}{3} \cdot p - p^2. \end{aligned}$$

30.

(Variabile aleatoare uniforme continue, independente; funcții densitate de probabilitate)

CMU, 2008 spring, Eric Xing, HW1, pr. 1.5

O persoană pleacă la serviciu între orele 8:00 și 8:30, iar timpul necesar deplasării este între 40 și 50 de minute. Considerăm X variabila aleatoare care reprezintă timpul de plecare exprimat în minute scurse după ora 8:00 și Y variabila aleatoare care reprezintă durata deplasării. Presupunând că aceste două variabile sunt independente și uniform distribuite, calculați:

- funcțiile densitate de probabilitate $p(x) \stackrel{\text{not.}}{=} P_X(X = x)$, $p(y) \stackrel{\text{not.}}{=} P_Y(Y = y)$ și $p(x, y) \stackrel{\text{not.}}{=} P_{X,Y}(X = x, Y = y)$.
- probabilitatea ca acea persoană să ajungă la serviciu înainte de ora 9.

Răspuns:

- Conform enunțului $Val(X) = [0, 30]$ și $Val(Y) = [40, 50]$. Pentru a determina funcția densitate de probabilitate pentru X și respectiv Y vom impune restricția ca integrala valorilor pe domeniul de definiție să fie 1:

$$\int_{-\infty}^{\infty} p(x)dx = 1 \Leftrightarrow \int_0^{30} p(x)dx = 1 \stackrel{p\text{-unif.}}{\Leftrightarrow} p(x) = \begin{cases} \frac{1}{30} & \text{pentru } 0 \leq x \leq 30 \\ 0 & \text{în caz contrar.} \end{cases}$$

$$\int_{-\infty}^{\infty} p(y)dy = 1 \Leftrightarrow \int_{40}^{50} p(y)dy = 1 \stackrel{p\text{-unif.}}{\Leftrightarrow} p(y) = \begin{cases} \frac{1}{10} & \text{pentru } 40 \leq y \leq 50 \\ 0 & \text{în caz contrar.} \end{cases}$$

Deoarece variabilele X și Y sunt independente, funcția densitate de probabilitate comună este:

$$p(x, y) = p(x) \cdot p(y) = \begin{cases} \frac{1}{300} & \text{pentru } 0 \leq x \leq 30 \text{ și } 40 \leq y \leq 50 \\ 0 & \text{în caz contrar.} \end{cases}$$

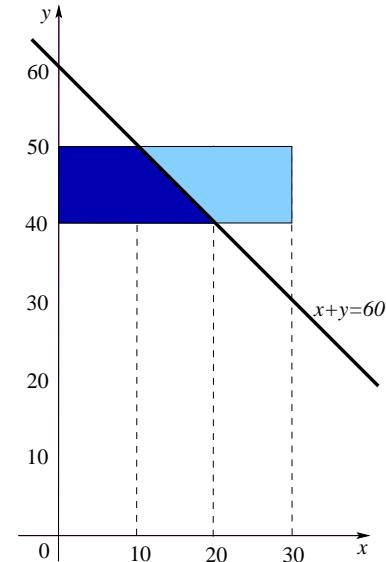
b. Probabilitatea ca persoana respectivă să ajungă la serviciu înainte de ora 9 este $P(X + Y \leq 60)$.

Grafic, aşa cum se observă din figura alăturată, această probabilitate este foarte simplu de găsit: $p(x + y \leq 60) = \frac{1}{2}$. Într-adevăr, stim că aria întregii zone dreptunghiulare pe care $p(x, y) \neq 0$ este

$$\int_{x=0}^{30} \int_{y=40}^{50} p(x, y) dy dx = 1.$$

Probabilitatea urmărită este aria zonei din acest dreptunghi care se găsește „sub“ dreapta de ecuație $x + y = 60$. Această zonă corespunde perechilor de valori (x, y) care satisfac condiția ca persoana să ajungă la serviciu înainte de ora 9.

Alternativ, această arie se poate obține prin calcul direct:



$$\begin{aligned} P(X + Y \leq 60) &= \int_{x=0}^{10} \int_{y=40}^{50} \frac{1}{300} dy dx + \int_{x=10}^{20} \int_{y=40}^{60-x} \frac{1}{300} dy dx \\ &= \int_{x=0}^{10} \frac{1}{300} \left(\int_{y=40}^{50} dy \right) dx + \int_{x=10}^{20} \frac{1}{300} \left(\int_{y=40}^{60-x} dy \right) dx \\ &= \int_{x=0}^{10} \frac{1}{300} \cdot y|_{40}^{50} dx + \int_{x=10}^{20} \frac{1}{300} \cdot y|_{40}^{60-x} dx = \int_{x=0}^{10} \frac{1}{300} \cdot 10 dx + \int_{x=10}^{20} \frac{1}{300} \cdot (20-x) dx \\ &= \frac{1}{30} \cdot x|_0^{10} + \frac{1}{300} \cdot \left(20x - \frac{1}{2}x^2 \right) \Big|_{10}^{20} = \frac{10}{30} + \frac{50}{300} = \frac{1}{2}. \end{aligned}$$

31.

(Distribuția exponențială și distribuția Gamma:
verificarea condițiilor de definiție pentru p.d.f.,
calculul mediilor și al varianțelor)

Liviu Ciortuz, 2017

- a. Distribuția *exponențială* este o distribuție continuă, care are funcția densitate de probabilitate

$$p(x | \theta) = \begin{cases} \theta e^{-\theta x} & \text{pentru } x \geq 0 \\ 0 & \text{pentru } x < 0. \end{cases}$$

unde $\theta > 0$ este un parametru real.

Arătați mai întâi că funcția $p(\cdot)$ este într-adevăr funcție densitate de probabilitate (engl., probability density function, p.d.f.) și apoi că media și respectiv varianța distribuției exponentiale sunt $\frac{1}{\theta}$ și respectiv $\frac{1}{\theta^2}$.

- b. Funcția Γ a lui Euler (1707-1783) este o generalizare în \mathbb{R}^+ a definiției numerelor factoriale din \mathbb{N} (și anume, $\Gamma(r) = (r-1)!$ pentru orice $r \in \mathbb{N}^*$).

Formula de definiție a acestei funcții este următoarea:

$$\Gamma(r) \stackrel{\text{def.}}{=} \int_0^{+\infty} t^{r-1} e^{-t} dt \text{ pentru orice } r > 0.$$

Demonstrați următoarele proprietăți ale funcției Γ :

i. $\Gamma(r+1) = r \Gamma(r)$ pentru orice $r > 0$.

ii. $\Gamma(1) = 1$ și $\Gamma(1/2) = \sqrt{\pi}$.³⁹

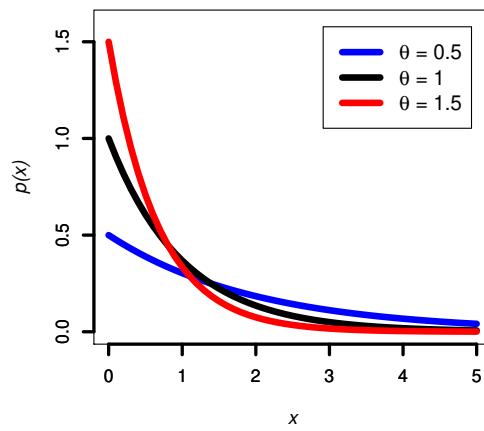
- c. Distribuția *Gamma* este o distribuție continuă, de parametri $r > 0$ (care dă *forma* distribuției, engl., *shape*) și $\alpha > 0$ (numit *rata*, engl., *rate*), cu funcția densitate de probabilitate definită pe \mathbb{R}^+ prin expresia următoare:

$$\begin{aligned} p(x) &\stackrel{\text{not.}}{=} \text{Gamma}(x|r, \alpha) \\ &\stackrel{\text{def.}}{=} \frac{\alpha^r}{\Gamma(r)} x^{r-1} e^{-\alpha x} \text{ pentru } x \geq 0, \end{aligned}$$

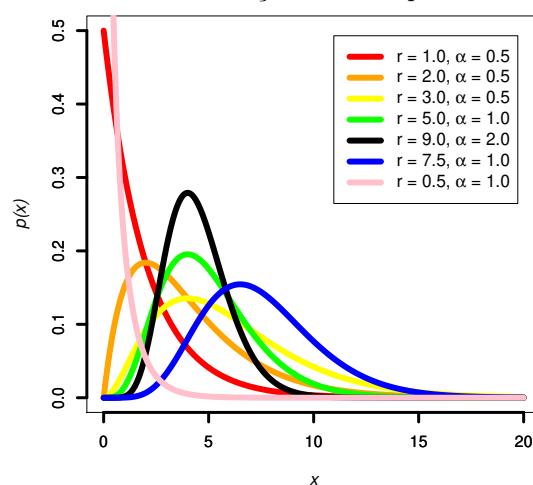
unde *constanta de normalizare* este $\frac{\alpha^r}{\Gamma(r)}$.

Demonstrați mai întâi că funcția $p(\cdot)$ este într-adevăr p.d.f. și apoi că media distribuției Gamma este $\frac{r}{\alpha}$, iar varianța ei este $\frac{r}{\alpha^2}$.

Distribuția exponențială: p.d.f.



Distribuția Gamma: p.d.f.



³⁹Din relația de recurență i. și din relația $\Gamma(1) = 1$, rezultă $\Gamma(r+1) = r \cdot \Gamma(r) = r \cdot (r-1) \cdot \Gamma(r-1) = \dots = r \cdot (r-1) \cdot \dots \cdot 2 \cdot 1 = r!$. Așadar, într-adevăr funcția Γ a lui Euler generalizează noțiunea de *produs factorial*.

Sugestie: Puteți folosi următoarea proprietate, care este demonstrată la problema 32:

$$\int_{-\infty}^{\infty} v^2 e^{-\frac{v^2}{2}} dv = 2 \int_0^{\infty} v^2 e^{-\frac{v^2}{2}} dv = \sqrt{2\pi}. \quad (17)$$

Observație: Se poate vedea imediat că $\text{Gamma}(x|1, \alpha) = \alpha e^{-\alpha x}$ pentru orice $x \geq 0$, ceea ce corespunde [definiției] funcției de densitate a[l] distribuției exponentiale (vedeți punctul a). Așadar, se poate spune că distribuția exponențială este membru al familiei de distribuții Gamma.

Răspuns:

a. Verificăm mai întâi că funcția $p(x|\theta)$ dată în enunț este într-adevăr o funcție densitate de probabilitate. Evident, $p(x|\theta) \geq 0$ pentru orice x , fiindcă $\theta > 0$ și $e^{-\theta x} > 0$ pentru orice x . Apoi,

$$\int_0^{\infty} \theta e^{-\theta x} dx = - \int_0^{\infty} (e^{-\theta x})' dx = e^{-\theta x} \Big|_0^{\infty} = 1 - 0 = 1.$$

La calculul mediei distribuției exponentiale vom folosi formula de integrare prin părți:

$$\begin{aligned} \int_0^{\infty} x \theta e^{-\theta x} dx &= - \int_0^{\infty} x (e^{-\theta x})' dx = -x(e^{-\theta x}) \Big|_0^{\infty} + \int_0^{\infty} x' e^{-\theta x} dx \\ &= 0 + \int_0^{\infty} e^{-\theta x} dx = -\frac{1}{\theta} e^{-\theta x} \Big|_0^{\infty} = -\frac{1}{\theta} (0 - 1) = \frac{1}{\theta}. \end{aligned}$$

Facem mențiunea că pentru a deduce $x(e^{-\theta x}) \Big|_0^{\infty} = 0$, am ținut cont că $\lim_{x \rightarrow \infty} \frac{x}{e^{\theta x}} = 0$, conform regulii lui l'Hôpital.

Pentru calculul varianței distribuției exponentiale, vom apela la formula $\text{Var}[X] = E[X^2] - E^2[X]$ (pe care am demonstrat-o la problema 9.b), unde X este o variabilă aleatoare oarecare. Întrucât în cazul nostru cunoaștem deja $E[X]$, calculăm $E[X^2]$:

$$\begin{aligned} \int_0^{\infty} x^2 \theta e^{-\theta x} dx &= - \int_0^{\infty} x^2 (e^{-\theta x})' dx = -x^2(e^{-\theta x}) \Big|_0^{\infty} + \int_0^{\infty} (x^2)' e^{-\theta x} dx \\ &= 0 + 2 \int_0^{\infty} x e^{-\theta x} dx = 2 \int_0^{\infty} x e^{-\theta x} dx = 2 \frac{1}{\theta} \int_0^{\infty} x \theta e^{-\theta x} dx = 2 \frac{1}{\theta} E[X] = \frac{2}{\theta^2}. \end{aligned}$$

Ca și mai înainte, folosind regula lui l'Hôpital, am dedus $x^2(e^{-\theta x}) \Big|_0^{\infty} = 0$. Așadar, $\text{Var}[X] = \frac{2}{\theta^2} - \left(\frac{1}{\theta}\right)^2 = \frac{1}{\theta^2}$.

b. Pentru demonstrarea relației i., $\Gamma(r+1) = r\Gamma(r)$, vom folosi formula de integrare prin părți:

$$\Gamma(r+1) = \int_0^{\infty} t^r e^{-t} dt = \int_0^{\infty} t^r (-e^{-t})' dt = -e^{-t} t^r \Big|_0^{\infty} + r \int_0^{\infty} t^{r-1} e^{-t} dt = r\Gamma(r).$$

Am ținut cont că $\lim_{t \rightarrow \infty} \frac{t^r}{e^t} = 0$ (pentru $r \in \mathbb{N}^*$ se poate folosi teorema lui l'Hôpital, însă pentru $r \in \mathbb{R}$, $r > 0$ se folosește teorema cleștelui).

ii. Valoarea funcției Γ în 1 se calculează astfel:

$$\Gamma(1) = \int_0^\infty e^{-t} dt = \int_0^\infty (-e^{-t})' dt = -e^{-t} \Big|_0^\infty = 1 - 0 = 1.$$

În sfârșit, calculăm valoarea funcției Γ în 1/2:

$$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty t^{-\frac{1}{2}} e^{-t} dt = \int_0^\infty \frac{1}{\sqrt{t}} e^{-t} dt = \int_0^\infty (\sqrt{t})' \cdot 2 \cdot e^{-t} dt = 2 \underbrace{\sqrt{t} e^{-t} \Big|_0^\infty}_{0} + 2 \int_0^\infty \sqrt{t} e^{-t} dt.$$

Făcând schimbarea de variabilă $\sqrt{t} = \frac{v}{\sqrt{2}} \Leftrightarrow t = \frac{v^2}{2}$ cu $v \geq 0$, rezultă $dt = v dv$ și

$$\Gamma\left(\frac{1}{2}\right) = 2 \int_0^\infty \sqrt{t} e^{-t} dt = 2 \int_0^\infty \frac{v}{\sqrt{2}} e^{-\frac{v^2}{2}} v dv = \sqrt{2} \int_0^\infty v^2 e^{-\frac{v^2}{2}} dv \stackrel{(17)}{=} \sqrt{2} \cdot \frac{1}{2} \cdot \sqrt{2\pi} = \sqrt{\pi}.$$

c. Folosind notația din enunț, unde $p(\)$ desemnează funcția densitate de probabilitate a distribuției Gamma, vom arăta mai întâi că $p(\)$ este într-adevăr funcție densitate de probabilitate. Evident, $p(x) > 0$ pentru orice $x > 0$, pentru că α și r sunt strict pozitive. Apoi, a verifică condiția $\int_0^{+\infty} p(x) dx = 1$ revine la a arăta că $\int_0^{+\infty} x^{r-1} e^{-\alpha x} dx = \frac{\Gamma(r)}{\alpha^r}$. Într-adevăr, făcând schimbarea de variabilă $y = \alpha x$, care implică $dy = \alpha dx$, urmează

$$\int_0^{+\infty} x^{r-1} e^{-\alpha x} dx = \int_0^{+\infty} \left(\frac{y}{\alpha}\right)^{r-1} e^{-y} \cdot \frac{1}{\alpha} dy = \frac{1}{\alpha^r} \int_0^{+\infty} y^{r-1} e^{-y} dy = \frac{\Gamma(r)}{\alpha^r}. \quad (18)$$

Acum arătăm că media distribuției Gamma este $\frac{r}{\alpha}$:

$$\int_0^{+\infty} xp(x) dx = \frac{\alpha^r}{\Gamma(r)} \int_0^{+\infty} x^r e^{-\alpha x} dx \stackrel{(18)}{=} \frac{\alpha^r}{\Gamma(r)} \cdot \frac{\Gamma(r+1)}{\alpha^{r+1}} \stackrel{b.i.}{=} \frac{\alpha^r}{\Gamma(r)} \cdot \frac{r \Gamma(r)}{\alpha^{r+1}} = \frac{r}{\alpha}.$$

În sfârșit, arătăm că varianța distribuției Gamma este $\frac{r}{\alpha^2}$. Apelăm încă o dată la formula $Var[X] = E[X^2] - E^2[X]$. Pentru că am calculat deja media distribuției Gamma, vom calcula acum $\int_0^{+\infty} x^2 p(x) dx$.

$$\begin{aligned} \int_0^{+\infty} x^2 p(x) dx &= \frac{\alpha^r}{\Gamma(r)} \int_0^{+\infty} x^2 x^{r-1} e^{-\alpha x} dx = \frac{\alpha^r}{\Gamma(r)} \int_0^{+\infty} x^{r+1} e^{-\alpha x} dx \\ &\stackrel{(18)}{=} \frac{\alpha^r}{\Gamma(r)} \cdot \frac{\Gamma(r+2)}{\alpha^{r+2}} \stackrel{b.i.}{=} \frac{(r+1)r \Gamma(r)}{\Gamma(r) \alpha^2} = \frac{r(r+1)}{\alpha^2}. \end{aligned}$$

În consecință, varianța distribuției Gamma este

$$\frac{r(r+1)}{\alpha^2} - \left(\frac{r}{\alpha}\right)^2 = \frac{r}{\alpha^2}.$$

Menționăm faptul că dând lui r valoarea 1 în formulele pe care tocmai le-am obținut pentru media și varianța distribuției Gamma, obținem media și respectiv varianța distribuției exponențiale, ceea ce era de altfel de așteptat, conform *Observației* din enunț.

32.

(Distribuția gaussiană unidimensională: verificarea condițiilor de definiție pentru p.d.f., calculul mediei și al varianței)

prelucrare de Liviu Ciortuz, după

■ CMU, 2010 spring, T. Mitchel, E. Xing, A. Singh, HW1, pr. 1.3.2

Considerăm o variabilă aleatoare X care urmează distribuția normală (gaussiană):⁴⁰

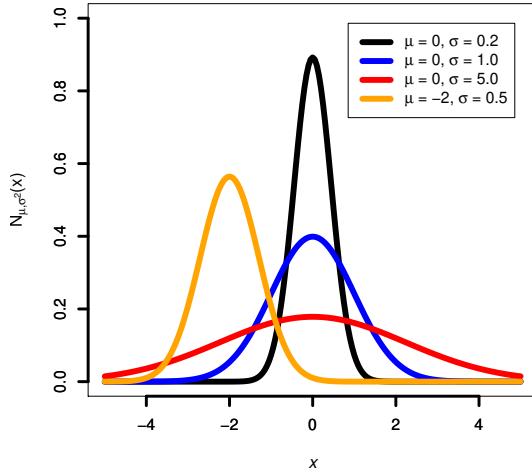
$$\mathcal{N}(X = x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x - \mu)^2}{2\sigma^2}}$$

unde σ poate fi orice număr real pozitiv, iar μ orice număr real.

Arătați că:

- a. \mathcal{N} este într-adevăr o funcție de densitate de probabilitate (p.d.f.), adică $\int_{-\infty}^{+\infty} \mathcal{N}(x) dx = 1$.
- b. $E[X] = \mu$.
- c. $Var[X] = \sigma^2$.

Distribuția gaussiană: p.d.f.



Sugestie (pentru punctul a): Pentru cazul distribuției gaussiene *standard* — la care veți face „reducere“, de la cazul *general*, printr-o schimbare liniară de variabilă —, proprietatea de demonstrat devine

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 1 \text{ sau, echivalent } \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx = \sqrt{2\pi}.$$

Pentru demonstrarea ultimei egalități vă recomandăm să arătați că

$$\left(\int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx \right)^2 = 2\pi \text{ sau, echivalent } \left(\int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx \right) \cdot \left(\int_{-\infty}^{+\infty} e^{-\frac{y^2}{2}} dy \right) = 2\pi,$$

deci

$$\int_{x=-\infty}^{+\infty} \int_{y=-\infty}^{+\infty} e^{-\frac{x^2+y^2}{2}} dx dy = 2\pi.$$

⁴⁰După numele lui Karl Friedrich Gauss (1777 – 1855), matematician, fizician și astronom german.

Ultima egalitate poate fi demonstrată prin trecerea din sistemul de coordinate cartezian în sistemul de coordonate polare. În mod concret, $(x, y) \mapsto (r, \theta)$, unde $r \in [0, +\infty)$ și $\theta \in [0, 2\pi)$, cu $x = r \cos \theta$ și $y = r \sin \theta$, ceea ce constituie o corespondență bijectivă. Vă reamintim că *regula de schimbare de variabilă* pentru cazul vectorial, formulată (aici) pentru cazul probabilist, este următoarea:⁴¹

Presupunem că $V = [V_1 \dots V_n]^\top \in \mathbb{R}^n$ este un vector de variabile aleatoare având funcția de densitate de probabilitate comună $f_V : \mathbb{R}^n \rightarrow \mathbb{R}$. Dacă definim un alt vector de variabile aleatoare, Z , obținut prin compunerea $Z = H(V)$, unde $H : \mathbb{R}^n \rightarrow \mathbb{R}^n$ este o funcție bijectivă și derivabilă [pe componente] în raport cu fiecare dintre argumentele sale, atunci Z va avea funcția de densitate de probabilitate comună $f_Z : \mathbb{R}^n \rightarrow \mathbb{R}$, unde

$$f_Z(z) = f_V(v) \cdot \left| \det \left(\begin{bmatrix} \frac{\partial v_1}{\partial z_1} & \dots & \frac{\partial v_1}{\partial z_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial v_n}{\partial z_1} & \dots & \frac{\partial v_n}{\partial z_n} \end{bmatrix} \right) \right|. \quad (19)$$

Matricea al cărei determinant este calculat în expresia de mai sus se numește matrice *jacobiană*, iar determinantul respectiv se numește *determinant jacobian*.⁴²

Răspuns:

a. Folosind schimbarea de variabilă *sugerată* în enunț, integrala dublă

$$\int_{x=-\infty}^{+\infty} \int_{y=-\infty}^{+\infty} e^{-\frac{x^2+y^2}{2}} dx dy$$

devine

$$\int_{r=0}^{+\infty} \int_{\theta=0}^{2\pi} e^{-\frac{r^2}{2}} r dr d\theta,$$

întrucât avem determinantul jacobian⁴³

$$\begin{aligned} \left| \frac{\partial(x, y)}{\partial(r, \theta)} \right| &= \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \frac{\partial r \cos \theta}{\partial r} & \frac{\partial r \cos \theta}{\partial \theta} \\ \frac{\partial r \sin \theta}{\partial r} & \frac{\partial r \sin \theta}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{vmatrix} \\ &= r \cos^2 \theta + r \sin^2 \theta = r \geq 0, \end{aligned}$$

de unde a rezultat $dx dy = r dr d\theta$.

În continuare, vom putea scrie:

$$\int_{r=0}^{\infty} \int_{\theta=0}^{2\pi} e^{-\frac{r^2}{2}} (r dr d\theta) = \int_{r=0}^{\infty} r e^{-\frac{r^2}{2}} \left(\int_{\theta=0}^{2\pi} d\theta \right) dr = \int_{r=0}^{\infty} r e^{-\frac{r^2}{2}} \theta \Big|_0^{2\pi} dr$$

⁴¹Cf. *The Multivariate Gaussian Distribution*, Chuong Do, Stanford University, 2008.

⁴²După numele matematicianului german evreu Carl Gustav Jacob Jacobi (1804 – 1851).

⁴³În relația (19) vom face următoarele înlocuiri: $z_1 = r$, $z_2 = \theta$, $v_1 = x = r \cos \theta$, $v_2 = y = r \sin \theta$.

$$= 2\pi \int_{r=0}^{\infty} r e^{-\frac{r^2}{2}} dr = 2\pi \left(-e^{-\frac{r^2}{2}} \right) \Big|_0^{\infty} = 2\pi e^{-\frac{r^2}{2}} \Big|_0^{\infty} = 2\pi(1 - 0) = 2\pi,$$

ceea ce practic finalizează demonstrația pentru *cazul distribuției gaussiene unidimensionale standard*.

Pentru *cazul nestandard*, folosim schimbarea de variabilă $v = \frac{x-\mu}{\sigma} \Rightarrow x = \sigma v + \mu$ cu $dx = \sigma dv$ și obținem:

$$\begin{aligned} \int_{-\infty}^{\infty} \mathcal{N}(x) dx &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-\frac{v^2}{2}} \sigma dv = \frac{1}{\sqrt{2\pi}\sigma} \sigma \int_{-\infty}^{\infty} e^{-\frac{v^2}{2}} dv = \frac{1}{\sqrt{2\pi}} \sqrt{2\pi} = 1, \end{aligned}$$

ceea ce era de demonstrat.

b. Vom calcula media variabilei aleatoare X folosind formula de definiție:

$$E[X] \stackrel{\text{def.}}{=} \int_{-\infty}^{\infty} xp(x) dx = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx.$$

Facând (din nou) schimbarea de variabilă $v = \frac{x-\mu}{\sigma} \Rightarrow x = \sigma v + \mu$ și $dx = \sigma dv$, vom obține:

$$\begin{aligned} E[X] &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (\sigma v + \mu) e^{-\frac{v^2}{2}} (\sigma dv) = \frac{\sigma}{\sqrt{2\pi}\sigma} \left(\sigma \int_{-\infty}^{\infty} ve^{-\frac{v^2}{2}} dv + \mu \int_{-\infty}^{\infty} e^{-\frac{v^2}{2}} dv \right) \\ &= \frac{1}{\sqrt{2\pi}} \left(-\sigma \int_{-\infty}^{\infty} (-v)e^{-\frac{v^2}{2}} dv + \mu \int_{-\infty}^{\infty} e^{-\frac{v^2}{2}} dv \right) \\ &= \frac{1}{\sqrt{2\pi}} \left(\underbrace{-\sigma e^{-\frac{v^2}{2}}}_{=0} \Big|_{-\infty}^{\infty} + \mu \int_{-\infty}^{\infty} e^{-\frac{v^2}{2}} dv \right) = \frac{\mu}{\sqrt{2\pi}} \underbrace{\int_{-\infty}^{\infty} e^{-\frac{v^2}{2}} dv}_{=\sqrt{2\pi}} = \mu. \end{aligned}$$

La ultima egalitate am folosit rezultatul obținut la punctul a (cazul standard).

c. Trebuie să arătăm că $\text{Var}[X] = \sigma^2$. Vom calcula varianța lui X utilizând formula $\text{Var}[X] = E[X^2] - (E[X])^2$. Întrucât am calculat $E[X]$, trebuie să mai calculăm valoarea medie a lui X^2 .

$$E[X^2] = \int_{-\infty}^{\infty} x^2 p(x) dx = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x^2 \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx.$$

Făcând aceeași schimbare de variabilă $v = \frac{x-\mu}{\sigma} \Rightarrow x = \sigma v + \mu$, cu $dx = \sigma dv$, obținem:

$$\begin{aligned}
E[X^2] &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (\sigma v + \mu)^2 e^{-\frac{v^2}{2}} (\sigma dv) = \frac{\sigma}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (\sigma^2 v^2 + 2\sigma\mu v + \mu^2) e^{-\frac{v^2}{2}} dv \\
&= \frac{1}{\sqrt{2\pi}} \left(\sigma^2 \int_{-\infty}^{\infty} v^2 e^{-\frac{v^2}{2}} dv + 2\sigma\mu \int_{-\infty}^{\infty} v e^{-\frac{v^2}{2}} dv + \mu^2 \int_{-\infty}^{\infty} e^{-\frac{v^2}{2}} dv \right). \quad (20)
\end{aligned}$$

Stim de la punctul a că $\int_{-\infty}^{\infty} e^{-\frac{v^2}{2}} dv = \sqrt{2\pi}$, iar de la punctul b că $\int_{-\infty}^{\infty} ve^{-\frac{v^2}{2}} dv = 0$.

Mai rămâne să calculăm prima integrală din expresia (20). Pentru aceasta vom utiliza metoda de integrare prin părți:

$$\begin{aligned}
\int_{-\infty}^{\infty} v^2 e^{-\frac{v^2}{2}} dv &= \int_{-\infty}^{\infty} (-v) \left(-ve^{-\frac{v^2}{2}} \right) dv = \int_{-\infty}^{\infty} (-v) \left(e^{-\frac{v^2}{2}} \right)' dv \\
&= (-v) e^{-\frac{v^2}{2}} \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} (-1)e^{-\frac{v^2}{2}} dv = 0 + \int_{-\infty}^{\infty} e^{-\frac{v^2}{2}} dv = \sqrt{2\pi}.
\end{aligned}$$

Pentru a calcula valoarea integralei definite am folosit următoarea proprietate:

$$\lim_{v \rightarrow \infty} v e^{-\frac{v^2}{2}} = \lim_{v \rightarrow \infty} \frac{v}{e^{\frac{v^2}{2}}} \stackrel{l'Hôpital}{=} \lim_{v \rightarrow \infty} \frac{1}{v e^{\frac{v^2}{2}}} = 0 = \lim_{v \rightarrow -\infty} v e^{-\frac{v^2}{2}}.$$

Așadar,

$$E[X^2] = \frac{1}{\sqrt{2\pi}} \left(\sigma^2 \sqrt{2\pi} + 2\sigma\mu \cdot 0 + \mu^2 \sqrt{2\pi} \right) = \sigma^2 + \mu^2.$$

Deci varianța variabilei aleatoare $X \sim \mathcal{N}(\mu, \sigma^2)$ este:

$$Var[X] = E[X^2] - (E[X])^2 = (\sigma^2 + \mu^2) - \mu^2 = \sigma^2.$$

33.

(Distribuția gaussiană unidimensională:
reducerea cazului nestandard la cazul standard,
folosind funcția cumulativă de distribuție)

CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW1, pr. 3

Considerăm variabila aleatoare X cu distribuția normală de medie $\mu = 1$ și varianță $\sigma^2 = 4$. Calculați următoarele probabilități:

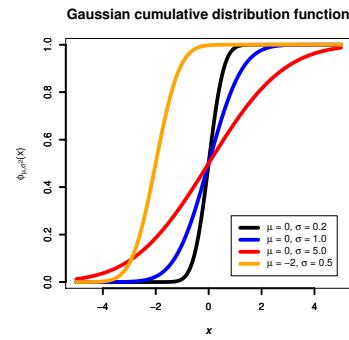
a. $P(X \leq 3)$.

b. $P(|X| \leq 2)$.

Indicație: Pentru a calcula aceste probabilități, soluția este să transformați variabila aleatoare X în distribuția normală standard Z după formula: $Z =$

$\frac{X - \mu}{\sigma}$. Pentru distribuția probabilistă normală standard ($\mu = 0$ și $\sigma = 1$), valoările funcției de distribuție cumulativă (engl., cumulative distribution function), notată Φ și definită prin relația $\Phi(x) \stackrel{\text{not.}}{=} P(Z \leq x)$, se consideră că sunt deja calculate. Puteți folosi următorul tabel de valori:

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
-3.4	0.0003	-1.7	0.0446	0.0	0.5000	1.7	0.9554
-3.3	0.0005	-1.6	0.0548	0.1	0.5398	1.8	0.9641
-3.2	0.0007	-1.5	0.0668	0.2	0.5793	1.9	0.9713
-3.1	0.0010	-1.4	0.0808	0.3	0.6179	2.0	0.9772
-3.0	0.0013	-1.3	0.0968	0.4	0.6554	2.1	0.9821
-2.9	0.0019	-1.2	0.1151	0.5	0.6915	2.2	0.9861
-2.8	0.0026	-1.1	0.1357	0.6	0.7257	2.3	0.9893
-2.7	0.0035	-1.0	0.1587	0.7	0.7580	2.4	0.9918
-2.6	0.0062	-0.9	0.1841	0.8	0.7881	2.5	0.9938
-2.5	0.0062	-0.8	0.2119	0.9	0.8159	2.6	0.9953
-2.4	0.0082	-0.7	0.2420	1.0	0.8413	2.7	0.9965
-2.3	0.0107	-0.6	0.2743	1.1	0.8643	2.8	0.9974
-2.2	0.0139	-0.5	0.3085	1.2	0.8849	2.9	0.9981
-2.1	0.0179	-0.4	0.3446	1.3	0.9032	3.0	0.9987
-2.0	0.0228	-0.3	0.3821	1.4	0.9192	3.1	0.9990
-1.9	0.0287	-0.2	0.4207	1.5	0.9332	3.2	0.9993
-1.8	0.0359	-0.1	0.4602	1.6	0.9452	3.3	0.9995



Notă: Pentru distribuția normală standard (Φ , în textul problemei), vedeti curba de culoare albastră.

Răspuns:

a.

$$\begin{aligned} P(X \leq 3) &\stackrel{\text{not.}}{=} P(\{\omega | X(\omega) \leq 3\}) = P\left(\left\{\omega \middle| \frac{X(\omega) - \mu}{\sigma} \leq \frac{3 - \mu}{\sigma}\right\}\right) \\ &\stackrel{\text{not.}}{=} P\left(\frac{X - \mu}{\sigma} \leq \frac{3 - \mu}{\sigma}\right) = P\left(Z \leq \frac{3 - \mu}{\sqrt{4}}\right) = P(Z \leq 1) = \Phi(1) = 0.8413. \end{aligned}$$

b. Vom proceda analog, descompunând probabilitatea $P(|X| \leq 2) = P(-2 \leq X \leq 2)$ într-o diferență de două probabilități:

$$\begin{aligned} P(|X| \leq 2) &= P(-2 \leq X \leq 2) = P(X \leq 2) - P(X \leq -2) \\ &= P(X - 1 \leq 2 - 1) - P(X - 1 \leq -2 - 1) \\ &= P\left(\frac{X - 1}{\sqrt{4}} \leq \frac{2 - 1}{\sqrt{4}}\right) - P\left(\frac{X - 1}{\sqrt{4}} \leq \frac{-2 - 1}{\sqrt{4}}\right) \\ &= P\left(Z \leq \frac{2 - 1}{2}\right) - P\left(Z \leq \frac{-2 - 1}{2}\right) = \Phi\left(\frac{1}{2}\right) - \Phi\left(-\frac{3}{2}\right) \\ &= 0.6915 - 0.0668 = 0.6247 \end{aligned}$$

34.

(Distribuții gaussiene multidimensionale: o proprietate importantă, în cazul în care matricea de covarianță este diagonală)

■ □ • prelucrare de Liviu Ciortuz, după “The Multivariate Gaussian Distribution”, Chuong B. Do, 2008

Fie o variabilă aleatoare $X : \Omega \rightarrow \mathbb{R}^d$. În cele ce urmează elementele din \mathbb{R}^d vor fi considerate vectori-coloană. Vom nota cu \mathbb{S}_+^d spațiul matricelor simetrice pozitiv definite⁴⁴ de dimensiune $d \times d$.

⁴⁴Prin definiție, o matrice $A \in \mathbb{R}^{d \times d}$ este pozitiv definită dacă $x^\top A x > 0$ pentru orice $x \neq 0$ din \mathbb{R}^d , unde simbolul \top reprezintă operația de transpunere a matricelor.

Vom spune că variabila X , reprezentată sub forma $X = [X_1 \dots X_d]^\top$, urmează o distribuție gaussiană (sau, *normală*) multidimensională, având media $\mu \in \mathbb{R}^d$ și matricea de covarianță $\Sigma \in \mathbb{S}_+^d$, dacă funcția ei de densitate [de probabilitate] are forma analitică următoare:⁴⁵

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1} (x - \mu) \right), \quad (21)$$

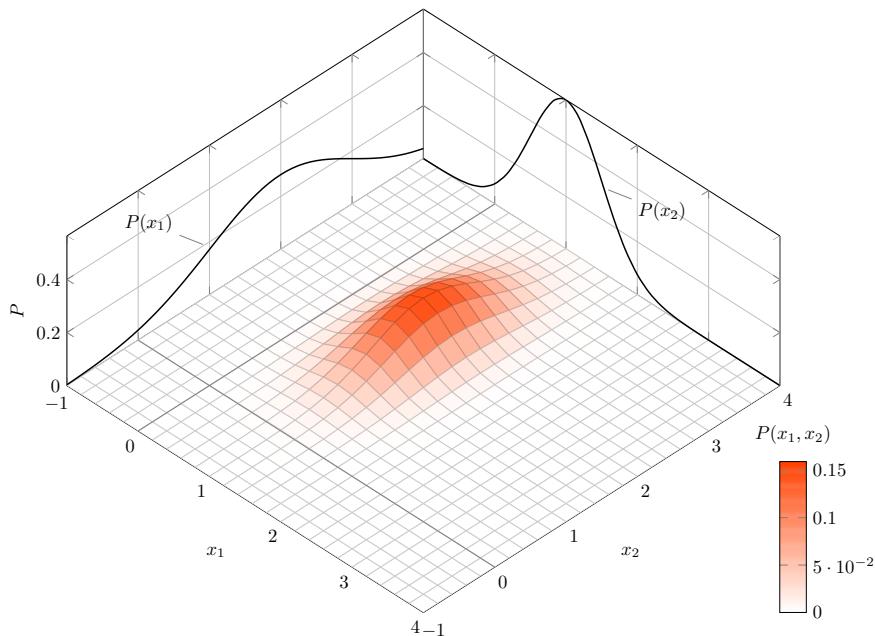
unde notația $|\Sigma|$ desemnează determinantul matricei Σ , iar $\exp(\cdot)$ desemnează funcția exponentială având baza e .⁴⁶ Pe scurt, vom nota această proprietate de definiție a lui X sub forma $X \sim \mathcal{N}(\mu, \Sigma)$.

Observații:

- Se poate demonstra că orice matrice pozitiv definită Σ este inversabilă (deci $|\Sigma| \neq 0$), iar Σ^{-1} este de asemenea matrice pozitiv definită. Așadar pentru orice vector nenul z , vom avea $z^\top \Sigma^{-1} z > 0$. Aceasta implică faptul că pentru orice vector $x \neq \mu$, vom avea:

$$(x - \mu)^\top \Sigma^{-1} (x - \mu) > 0, \text{ deci } -\frac{1}{2}(x - \mu)^\top \Sigma^{-1} (x - \mu) < 0.$$

Similar cazului unidimensional, graficul acestei funcții exponentiale va avea forma unui clopot cu vârful având abscisa $x = \mu$ și cu deschiderea cuadratică⁴⁷ îndreptată în jos.



[Credit: [http://www.pgflots.net/tikz/examples/all/](http://www/pgfplots.net/tikz/examples/all/)]

⁴⁵Remarcați faptul că expresia din dreapta egalității (21) implică un ușor abuz (sau, mai degrabă, o convenție) de notație: într-un astfel de context, o matrice reală de dimensiune 1×1 — în acest caz matricea $(x - \mu)^\top \Sigma^{-1} (x - \mu)$ — este identificată cu un număr real, care este chiar singurul ei element.

⁴⁶În cazul general al unui vector X format din d variabile aleatoare, elementul generic (i, j) al matricei de covarianță asociate lui X este prin definiție $Cov(X_i, X_j) = E[(X_i - E[X_i])(X_j - E[X_j])]$ pentru $i, j \in \{1, \dots, d\}$. Pentru demonstrația faptului că matricea de covarianță Σ este întotdeauna simetrică și pozitiv semidefinită vedeti problema 20.

⁴⁷Este vorba de o elipsă având axele principale determinate — ca vectori — de vectorii proprii ai matricei de covarianță Σ . Mărimile axelor principale sunt respectiv $\sqrt{\lambda_i}$, unde prin λ_i am notat vectorii proprii ai matricei Σ . (Cf. *Pattern Classification*, Duda, Hart and Stork, Appendix A.5.2, 2001.)

2. Coeficientul funcției exponențiale, adică $\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}$, nu depinde de x ; el este *constanta de normalizare* care ne asigură că

$$\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp((x - \mu)^\top \Sigma^{-1}(x - \mu)) dx_1 dx_2 \cdots dx_d = 1.$$

Considerăm cazul simplu, în care $d = 2$ și matricea de covarianță Σ este diagonală,⁴⁸ deci

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

cu $\sigma_1 > 0$ și $\sigma_2 > 0$.

Arătați că într-un astfel de caz, expresia funcției de densitate de probabilitate gaussiană bidimensională este identică cu produsul a două funcții de densitate de tip gaussian unidimensionale independente, prima funcție având media μ_1 și varianța σ_1^2 , iar cea de-a doua având media μ_2 și varianța σ_2^2 .

Răspuns:

În condițiile de mai sus, funcția de densitate de probabilitate gaussiană multidimensională are forma

$$p(x; \mu, \Sigma) = \frac{1}{2\pi \begin{vmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{vmatrix}^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^\top \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}\right)$$

Aplicând mai întâi formula pentru determinantul de ordin 2,⁴⁹ și calculând apoi inversa matricei Σ , obținem:

$$\begin{aligned} p(x; \mu, \Sigma) &= \frac{1}{2\pi \sigma_1 \sigma_2} \exp\left(-\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^\top \left(\frac{1}{\sigma_1^2 \sigma_2^2} \begin{bmatrix} \sigma_2^2 & 0 \\ 0 & \sigma_1^2 \end{bmatrix}\right) \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}\right) \\ &= \frac{1}{2\pi \sigma_1 \sigma_2} \exp\left(-\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^\top \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}\right) \\ &= \frac{1}{2\pi \sigma_1 \sigma_2} \exp\left(-\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^\top \begin{bmatrix} \frac{1}{\sigma_1^2} (x_1 - \mu_1) \\ \frac{1}{\sigma_2^2} (x_2 - \mu_2) \end{bmatrix}\right) \\ &= \frac{1}{2\pi \sigma_1 \sigma_2} \exp\left(-\frac{1}{2\sigma_1^2} (x_1 - \mu_1)^2 - \frac{1}{2\sigma_2^2} (x_2 - \mu_2)^2\right) \\ &= \frac{1}{\sqrt{2\pi} \sigma_1} \exp\left(-\frac{1}{2\sigma_1^2} (x_1 - \mu_1)^2\right) \frac{1}{\sqrt{2\pi} \sigma_2} \exp\left(-\frac{1}{2\sigma_2^2} (x_2 - \mu_2)^2\right) \\ &= p(x_1; \mu_1, \sigma_1^2) p(x_2; \mu_2, \sigma_2^2). \end{aligned}$$

⁴⁸În aceste condiții, se poate demonstra ușor că elementele de pe diagonala matricei Σ sunt strict pozitive. Într-adevăr, este suficient să se particularizeze vectorul z din formalizarea proprietății de *pozitiv-definire* a matricei Σ la valoarea $z = (1, 0)^\top$ și respectiv $z = (0, 1)^\top$.

⁴⁹Anume, $\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$.

Generalizând, este imediat că orice variabilă gaussiană d -dimensională de medie $\mu \in \mathbb{R}^d$ și matrice de covarianță diagonală $\Sigma \stackrel{\text{not.}}{=} \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_d^2)$, se comportă ca o colecție de d variabile aleatoare gaussiene unidimensionale independente, având respectiv mediile μ_i și varianțele σ_i^2 .

Observație: Știm de la exercițiul 10 că dacă X_1 și X_2 sunt variabile aleatoare independente, atunci $\text{Cov}(X_1, X_2) = 0$. Deși reciproca acestei afirmații nu este în general valabilă, la exercițiul 11.b am arătat că ea este adevărată în cazul în care valorile acestor două variabile aleatoare sunt binare. La exercițiul de față am arătat că implicația $\text{Cov}(X_1, X_2) = 0 \Rightarrow X_1 \perp X_2$ are loc și în cazul în care perechea (X_1, X_2) urmează o distribuție gaussiană bivariată (sau, mai general, atunci când X_1 și X_2 sunt componente ale unei distribuții gaussiene multivariate a cărei matrice de covarianță este diagonală).

35.

(Distribuția gaussiană, cazul bidimensional:
explicitarea p.d.f. într-un caz particular)

*prelucrare de Liviu Ciortuz, 2019, după
 • MIT, 2006 fall, Tommi Jaakkola, HW1, pr. A.5.a*

Fie X un vector de variabile aleatoare de tip gaussian, cu

$$E[X] = \begin{pmatrix} 10 \\ 5 \end{pmatrix} \quad \text{și} \quad \text{Cov}(X) = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}.$$

a. Verificați că matricea $\text{Cov}(X)$, care este simetrică, este și pozitiv definită.⁵⁰

b. Pornind de la notația matriceală pentru funcția densitate de probabilitate (p.d.f.) pentru o distribuție gaussiană multidimensională oarecare,

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(\sqrt{2\pi})^d \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right)$$

și considerând $x = (x_1, x_2) \in \mathbb{R}^2$, scrieți expresia funcției de densitate de probabilitate comună a variabilei X de mai sus. Veți nota această funcție cu $p(x_1, x_2)$.

Răspuns:

a. *Soluția I-a:* folosind definiția pentru matrice pozitiv definite.

Matricea $\text{Cov}(X)$ este pozitiv definită dacă și numai dacă pentru orice vector $z \stackrel{\text{not.}}{=} (z_1, z_2)^\top \in \mathbb{R}^2$ care diferă de vectorul $0 \in \mathbb{R}^2$ are loc inegalitatea $z^\top \text{Cov}(X) z > 0$.

Fie acum un vector oarecare $z = (z_1, z_2)^\top \in \mathbb{R}^2$ diferit de vectorul 0.

$$(z_1, z_2) \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = (z_1, z_2) \begin{pmatrix} 2z_1 + z_2 \\ z_1 + z_2 \end{pmatrix} = 2z_1^2 + 2z_1z_2 + z_2^2 = z_1^2 + (z_1 + z_2)^2.$$

Evident, $z_1^2 + (z_1 + z_2)^2 \geq 0$ pentru orice z_1 și $z_2 \in \mathbb{R}$. Vom arăta că $z_1^2 + (z_1 + z_2)^2 > 0$ pentru orice $(z_1, z_2)^\top \neq 0 \in \mathbb{R}^2$. Vom trata două cazuri:

⁵⁰Vedeți problema 20.

Cazul 1: Dacă $z_1 \neq 0$, atunci $z_1^2 \neq 0$ și, în consecință, $z_1^2 + (z_1 + z_2)^2 > 0$, indiferent de valoarea lui $z_2 \in \mathbb{R}$.

Cazul 2: Dacă $z_1 = 0$, atunci din condiția $(z_1, z_2) \neq 0 \in \mathbb{R}^2$ rezultă că $z_2 \neq 0$. Prin urmare, $z_1^2 + (z_1 + z_2)^2 = z_2^2 > 0$.

Soluția a II-a: Vom folosi următorul rezultat teoretic (de tip *caracterizare*): Fie A o matrice de numere reale, simetrică. Matricea A este pozitiv definită dacă și numai dacă toate valorile ei proprii (engl., eigenvalues) sunt pozitive.⁵¹

Mai întâi vom afla valorile proprii ale matricei $\Sigma \stackrel{\text{not.}}{=} \text{Cov}(X)$. Pentru aceasta, vom rezolva ecuația $(\Sigma - \lambda I)X = 0$, impunând *restricția ca soluția $X \in \mathbb{R}^2$ să fie diferită de vectorul 0*. Notând $X = (x_1, x_2)^\top$, ecuația $(\Sigma - \lambda I)X = 0$ va fi rescrisă astfel:

$$\begin{pmatrix} 2 - \lambda & 1 \\ 1 & 1 - \lambda \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Faptul că soluția $X \in \mathbb{R}^2$ trebuie să fie diferită de vectorul 0 implică în mod necesar

$$\begin{vmatrix} 2 - \lambda & 1 \\ 1 & 1 - \lambda \end{vmatrix} = 0 \Leftrightarrow 2 - 3\lambda + \lambda^2 - 1 = 0 \Leftrightarrow \lambda^2 - 3\lambda + 1 = 0 \Leftrightarrow \lambda_{1,2} = \frac{3 \pm \sqrt{5}}{2}.$$

Se observă imediat că λ_1 și λ_2 (vectorii proprii ai matricei Σ) sunt pozitivi. Așadar, ținând cont de *Indicația* din enunț, rezultă că Σ este matrice pozitiv definită.

b. Avem: $|\text{Cov}(X)| = \begin{vmatrix} 2 & 1 \\ 1 & 1 \end{vmatrix} = 1$, deci $(\text{Cov}(X))^{-1} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$.

Așadar, funcția de densitate de probabilitate comună a variabilei X este următoarea:

$$\begin{aligned} p(x_1, x_2) &= \frac{1}{(\sqrt{2\pi})^2} \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right) \\ &= \frac{1}{2\pi} \exp\left(-\frac{1}{2}(x_1 - 10, x_2 - 5) \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 - 10 \\ x_2 - 5 \end{pmatrix}\right) \\ &= \frac{1}{2\pi} \exp\left(-\frac{1}{2}(x_1 - x_2 - 5, -x_1 + 2x_2) \begin{pmatrix} x_1 - 10 \\ x_2 - 5 \end{pmatrix}\right) \\ &= \frac{1}{2\pi} \exp\left(-\frac{1}{2}(x_1^2 - 10x_1 - x_1x_2 + 10x_2 - 5x_1 + 50 - x_1x_2 + 5x_1 + 2x_2^2 - 10x_2)\right) \\ &= \frac{1}{2\pi} \exp\left(-\frac{1}{2}(x_1^2 - 10x_1 - 2x_1x_2 + 2x_2^2 + 50)\right). \end{aligned}$$

⁵¹Cf. <http://mathworld.wolfram.com/PositiveDefiniteMatrix.html>.

36.

(Matrice simetrice și pozitiv definite:
 [o proprietate de tip] factorizare folosind vectori proprii)

□ UAIC Iași, 2018 spring, Sebastian Ciobanu, după
 Appendix A.2, in The Multivariate Gaussian Distribution,
 by Chuong B. Do, Stanford, 2008

Fie o variabilă aleatoare $X : \Omega \rightarrow \mathbb{R}^d$. În cele ce urmează, elementele din \mathbb{R}^d vor fi considerate vectori-coloană. Vom nota cu \mathbb{S}_+^d mulțimea matricelor simetrice pozitiv definite de dimensiune $d \times d$.⁵²

Definiție: Fie $A \in \mathbb{R}^{d \times d}$. Se numește *valoare proprie* a matricei A un număr complex $\lambda \in \mathbb{C}$ pentru care există un vector nenul $x \in \mathbb{C}^d$ astfel încât să aibă loc egalitatea $Ax = \lambda x$. Acest vector x se numește *vector propriu* asociat valorii proprii λ .

a. Demonstrați că pentru orice matrice $\Sigma \in \mathbb{S}_+^d$ există o matrice $B \in \mathbb{R}^{d \times d}$ astfel încât Σ să poată fi factorizată sub forma următoare: $\Sigma = BB^\top$.

Observație: Factorizarea aceasta nu este unică, adică există mai multe posibilități de a scrie matricea Σ sub forma $\Sigma = BB^\top$.

Indicație: Vă puteți folosi de următoarele proprietăți:

i. Orice matrice $A \in \mathbb{R}^{d \times d}$ care este simetrică poate fi scrisă astfel: $A = U\Lambda U^\top$, unde $U \in \mathbb{R}^{d \times d}$ este o matrice ortonormală conținând vectorii proprii (pentru care impunem să aibă norma 1) ai lui A drept coloane, iar Λ este matricea diagonală conținând valorile proprii ale lui A în ordinea corespunzătoare coloanelor (adică, a vectorilor proprii) din matricea U .⁵³

ii. Fie $A \in \mathbb{R}^{d \times d}$ o matrice simetrică. A este pozitiv definită dacă și numai dacă toate valorile sale proprii sunt reale și pozitive.⁵⁴

iii. $(AB)^\top = B^\top A^\top$, pentru orice matrice $A, B \in \mathbb{R}^{d \times d}$.⁵⁵

b. La acest punct vom face o *exemplificare* pentru chestiunile prezentate la punctul a.

Considerând matricea

$$\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix},$$

(deci, $d = 2$) determinați o matrice $B \in \mathbb{R}^{2 \times 2}$ astfel încât $\Sigma = BB^\top$.

Indicație: Folosind notațiile din **Definiția** dată pentru valorile proprii și vectorii proprii, vom avea:

$$Ax = \lambda x \Leftrightarrow (\lambda I_d - A)x = \mathbf{0}, x \neq \mathbf{0} \Leftrightarrow \det(\lambda I_d - A) = 0,$$

⁵²Pentru definiția noțiunii de matrice pozitiv definită, vedeți problema 20.

⁵³Această proprietate este enunțată în secțiunea 0.8 din documentul *Matrix Identities* de Sam Roweis, 1999. Faptul că U este matrice ortonormală se poate scrie în mod formal astfel: $U^\top U = U U^\top = I$.

⁵⁴ https://en.wikipedia.org/wiki/Positive-definite_matrix.

Observație: Proprietății a.ii îl corespund alte trei proprietăți similare: Fie $A \in \mathbb{R}^{d \times d}$ o matrice simetrică. A este pozitiv semidefinită dacă și numai dacă toate valorile sale proprii sunt reale și neneegative. A este negativ definită dacă și numai dacă toate valorile sale proprii sunt reale și negative. A este negativ semidefinită dacă și numai dacă toate valorile sale proprii sunt reale și nepozitive. (Vedeți https://en.wikipedia.org/wiki/Definiteness_of_a_matrix#Eigenvalues.)

⁵⁵Formula (1c) din același document (*Matrix Identities*) de Sam Roweis.

unde $\mathbf{0}$ este vectorul coloană nul d -dimensional, iar I_d este matricea identitate d -dimensională.

c. Demonstrați că orice matrice B care satisface proprietatea de la punctul a este inversabilă.

Indicație: Vă puteți folosi de următoarele proprietăți:

i. Matricea $A \in \mathbb{R}^{d \times d}$ este inversabilă dacă și numai dacă $\det(A) \neq 0$.

ii. $\det(AB) = \det(A)\det(B)$, pentru orice matrice $A, B \in \mathbb{R}^{d \times d}$.⁵⁶

iii. $\det(A) = \det(A^\top)$, unde $A \in \mathbb{R}^{d \times d}$.

iv. Orice matrice pozitiv definită este inversabilă (iar inversa ei este de asemenea matrice pozitiv definită).⁵⁷

Răspuns:

a. Stim, prin ipoteză, că matricea $\Sigma \in \mathbb{S}_+^d$, deci este simetrică. Conform proprietății (a.i), putem scrie următoarea factorizare pentru Σ :

$$\Sigma = U\Lambda U^\top,$$

unde U este matricea ortonormală conținând vectorii proprii (cu normă 1) ai lui Σ drept coloane, iar $\Lambda \in \mathbb{R}^{d \times d}$ este matricea diagonală conținând valorile proprii ale lui Σ , în ordinea corespunzătoare coloanelor matricei U .

Întrucât $\Sigma \in \mathbb{S}_+^d$ este matrice pozitiv definită, conform proprietății (a.ii) rezultă că toate valorile proprii ale lui Σ sunt pozitive. Prin urmare, există matricea

$$\Lambda^{1/2} = \begin{bmatrix} \sqrt{\lambda_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sqrt{\lambda_d} \end{bmatrix} \in \mathbb{R}^{d \times d}.$$

Observații:

(1) Este imediat faptul că putem factoriza / descompune matricea Λ în felul următor:

$$\Lambda^{1/2} \cdot \Lambda^{1/2} = \Lambda. \quad (22)$$

(2) $\Lambda^{1/2}$ este matrice diagonală, deci simetrică. Așadar,

$$(\Lambda^{1/2})^\top = \Lambda^{1/2}. \quad (23)$$

Acum putem relua factorizarea matricei $\Sigma = U\Lambda U^\top$. Elaborând — adică, factorizând matricea Λ — în partea dreaptă a acestei egalități, vom putea obține o nouă factorizare pentru matricea Σ în felul următor:

$$\begin{aligned} \Sigma &= U\Lambda U^\top \\ &\stackrel{(22)}{=} U\Lambda^{1/2}\Lambda^{1/2}U^\top \stackrel{(23)}{=} U\Lambda^{1/2}(\Lambda^{1/2})^\top U^\top \\ &\stackrel{asoc.}{=} U\Lambda^{1/2}((\Lambda^{1/2})^\top U^\top) \stackrel{(a.iii)}{=} U\Lambda^{1/2}(U\Lambda^{1/2})^\top \\ &= BB^\top, \text{ unde } B \stackrel{not.}{=} U\Lambda^{1/2}. \end{aligned} \quad (24)$$

⁵⁶Formula (2a) din același document (*Matrix Identities*) de Sam Roweis.

⁵⁷https://en.wikipedia.org/wiki/Positive-definite_matrix.

Observații:

(3) O altă modalitate de a factoriza / descompune matricea Σ sub forma $\Sigma = BB^\top$ este dată de *factorizarea Cholesky*: dacă A este o matrice simetrică și pozitiv definită, atunci există o unică matrice $L \in \mathbb{R}^{d \times d}$, inferior triunghiu-lară, astfel încât $A = LL^\top$.⁵⁸

(4) Factorizarea (24) este de fapt valabilă și pentru matrice pozitiv semidefinite. Justificarea se bazează pe proprietatea menționată la nota de subsol 54. Însă proprietatea de inversabilitate (vedeți punctul c) nu este valabilă decât pentru matrice pozitiv definite.

b. Mai întâi vom calcula valorile proprii ale matricei Σ din enunț. Fie deci $\lambda \in \mathbb{R}$ și $x \in \mathbb{R}^d$, $x \neq 0$. Trebuie să rezolvăm ecuația $\Sigma x = \lambda x$ și, conform *Indicației* din enunț, aceasta revine la a rezolva ecuația $\det(\lambda I_d - \Sigma) = 0$.

$$\begin{aligned}\det(\lambda I_d - \Sigma) = 0 &\Leftrightarrow \det\left(\begin{bmatrix} \lambda - 1 & -0.5 \\ -0.5 & \lambda - 1 \end{bmatrix}\right) = 0 \Leftrightarrow (\lambda - 1)^2 - 0.5^2 = 0 \\ &\Leftrightarrow (\lambda - 1 - 0.5)(\lambda - 1 + 0.5) = 0 \Leftrightarrow (\lambda - 1.5)(\lambda - 0.5) = 0 \\ &\Leftrightarrow \lambda_1 = 0.5, \lambda_2 = 1.5.\end{aligned}$$

Deci valorile proprii ale matricei Σ sunt $\lambda_1 = 0.5$ și $\lambda_2 = 1.5$.

Se observă că $\lambda_1, \lambda_2 > 0$. Așadar, conform proprietății (a, ii),⁵⁹ matricea Σ dată în enunț este pozitiv definită. Conform punctului a, știm acum că există o matrice B astfel încât matricea Σ să poată fi factorizată sub forma $\Sigma = BB^\top$. Pentru a determina matricea B , trebuie să calculăm vectorii proprii ai matricei Σ . Îi vom obține rezolvând următorul sistem, care este compatibil nedeterminat (pentru că $\det(\lambda I_d - \Sigma) = 0$):⁶⁰

$$\begin{cases} (\lambda - 1)v - 0.5u = 0 \Rightarrow v = \frac{0.5u}{\lambda - 1} \\ -0.5v + (\lambda - 1)u = 0 \end{cases}.$$

Deci, vectorii proprii sunt de forma:

$$x = (v, u)^\top = \left(\frac{0.5}{\lambda - 1}u, u \right)^\top, u \in \mathbb{R}.$$

În mod concret,

$$\lambda_1 = 0.5 \Rightarrow x_1 = (-u, u)^\top, u \in \mathbb{R}$$

și

$$\lambda_2 = 1.5 \Rightarrow x_2 = (u, u)^\top, u \in \mathbb{R}.$$

Deoarece vrem ca vectorii x_1 și x_2 să aibă normă 1, îi vom „normaliza“. Mai întâi,

$$\frac{x_1}{\|x_1\|} = \frac{1}{\sqrt{2u^2}}(-u, u)^\top = \frac{1}{\sqrt{2}|u|}(-u, u)^\top \quad (25)$$

și, considerând $u > 0$, deci $|u| = u$, rezultă că

$$\frac{x_1}{\|x_1\|} = \frac{1}{\sqrt{2}u}(-u, u)^\top = \left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right)^\top.$$

⁵⁸https://en.wikipedia.org/wiki/Positive-definite_matrix.

⁵⁹Mai exact, ne referim la partea „numai atunci“ a acestei proprietăți, care reprezintă o caracterizare de tip echivalentă pentru noțiunea de matrice pozitiv definită. (Observați că matricea Σ din enunț este simetrică.)

⁶⁰Vedeți *Indicația* din enunț.

Apoi, în mod similar obținem

$$\frac{x_2}{\|x_2\|} = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right)^\top.$$

De la punctul a știm că putem alege $B = U\Lambda^{1/2}$, deci

$$\begin{aligned} U &= \begin{bmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}, \quad \Lambda^{1/2} = \begin{bmatrix} \sqrt{0.5} & 0 \\ 0 & \sqrt{1.5} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{\sqrt{3}}{\sqrt{2}} \end{bmatrix} \\ \Rightarrow B &= \begin{bmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{\sqrt{3}}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{bmatrix}. \end{aligned}$$

Verificăm că într-adevăr $\Sigma = BB^\top$:

$$BB^\top = \begin{bmatrix} -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{bmatrix} \begin{bmatrix} -\frac{1}{2} & \frac{1}{2} \\ \frac{\sqrt{3}}{2} & \frac{\sqrt{3}}{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{4} + \frac{3}{4} & -\frac{1}{4} + \frac{3}{4} \\ -\frac{1}{4} + \frac{3}{4} & \frac{1}{4} + \frac{3}{4} \end{bmatrix} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} = \Sigma.$$

În ton cu *Observația* de la pagina 79, dacă în relația (25) vom considera $u < 0$, vom obține încă o soluție pentru factorizarea matricei Σ :

$$\begin{aligned} \frac{x'_1}{\|x'_1\|} &= \frac{1}{\sqrt{2}|u|}(-u, u)^\top = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}} \right)^\top = -\frac{x_1}{\|x_1\|} \\ \frac{x'_2}{\|x'_2\|} &= \frac{1}{\sqrt{2}|u|}(u, u)^\top = \left(-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}} \right)^\top = -\frac{x_2}{\|x_2\|}, \end{aligned}$$

deci

$$U' = -U, B' = U'\Lambda^{1/2} = -U\Lambda^{1/2} = -B$$

și, în consecință,

$$B'(B')^\top = (-B)(-B)^\top = BB^\top = \Sigma.$$

c. De la punctul a rezultă că pentru orice matrice $\Sigma \in \mathbb{S}_+^d$ există o descompunere / factorizare de forma $\Sigma = BB^\top$, deci

$$\det(\Sigma) = \det(BB^\top) \stackrel{(c.ii)}{=} \det(B)\det(B^\top) \stackrel{(c.iii)}{=} \det(B)^2. \quad (26)$$

Prin urmare,

$$\Rightarrow \det(B) = \pm \sqrt{\det(\Sigma)}. \quad (27)$$

Pe de altă parte, din faptul că Σ este matrice pozitiv definită, rezultă conform proprietății (c.iv) că Σ este matrice inversabilă. Mai departe, conform proprietății (c.i), vom avea $\det(\Sigma) \neq 0$. Coroborând aceasta cu relația (27), rezultă că $\det(B) \neq 0$ și în consecință (din nou, conform proprietății (c.i)) că matricea B este inversabilă (ceea ce era de demonstrat).

37.

(Distribuția gaussiană multidimensională:
funcția ei de densitate de probabilitate
satisfacă într-adevăr proprietățile de definiție)

□ UAIC Iași, 2018 spring, Sebastian Ciobanu, după
Appendix A.2, in The Multivariate Gaussian Distribution,
by Chuong B. Do, Stanford, 2008

Definiție: Notăm cu \mathbb{S}_+^d mulțimea matricelor simetrice pozitiv definite de dimensiune $d \times d$.⁶¹ Spunem că variabila aleatoare vectorială $X : \Omega \rightarrow \mathbb{R}^d$, reprezentată sub forma $X = (X_1 \dots X_d)^\top$, urmează o distribuție gaussiană multidimensională, având media $\mu \in \mathbb{R}^d$ și matricea de covarianță $\Sigma \in \mathbb{S}_+^d$, dacă funcția ei de densitate [de probabilitate] are forma analitică următoare:

$$p_X(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right), \quad (28)$$

unde notația $\det(\Sigma)$ desemnează determinantul matricei Σ , iar $\exp(\)$ desemnează funcția exponentială având baza e . Pe scurt, vom nota această proprietate de definiție a lui X sub forma $X \sim \mathcal{N}(\mu, \Sigma)$.

Observație (1): La problema 36 am demonstrat că pentru orice matrice $\Sigma \in \mathbb{R}^d$ simetrică și pozitiv definită există (însă nu neapărat în mod unic) o matrice $B \in \mathbb{R}^{d \times d}$ cu proprietatea că Σ se poate „factoriza“ sub forma $\Sigma = BB^\top$. Mai mult, am demonstrat că orice matrice B care satisfac această proprietate este în mod necesar inversabilă. De asemenea, din relația (26) și din faptul că matricea Σ este pozitiv definită (ceea ce implică $\det(\Sigma) \neq 0$; vedeti proprietatea c.iv de la problema 36) rezultă că $\det(\Sigma) > 0$, ceea ce justifică faptul că $\det(\Sigma)^{1/2} = \sqrt{\det(\Sigma)}$ din relația (28) este bine definit.

a. Demonstrați că dacă definim $Z = B^{-1}(X - \mu)$, atunci $Z \sim \mathcal{N}(0, I_d)$, unde 0 este vectorul coloană nul d -dimensional, iar I_d este matricea identitate d -dimensională.

Observație (2): Proprietatea aceasta este o generalizare a metodei de „standardizare“ pe care am întâlnit-o deja la problema 33, aplicată în cazul distribuțiilor gaussiene unidimensionale: $Z = \frac{X - \mu}{\sigma}$. În esență, acolo ne refeream la schimbarea de variabilă care realizează punerea în corespondență a unei distribuții gaussiene unidimensionale oarecare cu distribuția gaussiană / normală standard (aceasta din urmă având media 0 și varianța 1).

Indicație (1): Vă puteți folosi de următoarele proprietăți:

i. Fie $X = (X_1 \dots X_d)^\top \in \mathbb{R}^d$ o variabilă aleatoare de tip vector, cu distribuția comună dată de funcția de densitate $p_X : \mathbb{R}^d \rightarrow \mathbb{R}$. Dacă $Z = H(X) \in \mathbb{R}^d$, unde H este funcție bijectivă și derivabilă [pe componente] în raport cu fiecare dintre argumentele sale, atunci Z are distribuția comună dată de funcția de densitate $p_Z : \mathbb{R}^d \rightarrow \mathbb{R}$, unde

$$p_Z(z) = p_X(x) \cdot \left| \det \begin{pmatrix} \frac{\partial x_1}{\partial z_1} & \cdots & \frac{\partial x_1}{\partial z_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_d}{\partial z_1} & \cdots & \frac{\partial x_d}{\partial z_d} \end{pmatrix} \right|.$$

⁶¹Pentru definiția noțiunii de matrice pozitiv definită, vedeti problema 20.

Matricea $\begin{bmatrix} \frac{\partial x_1}{\partial z_1} & \dots & \frac{\partial x_1}{\partial z_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_d}{\partial z_1} & \dots & \frac{\partial x_d}{\partial z_d} \end{bmatrix}$ se numește *matricea jacobiană* a lui x în raport cu z și se notează, în acest caz, cu $\frac{\partial x}{\partial z}$.

ii. $\frac{\partial Ax + b}{\partial x} = A$, unde $x, b \in \mathbb{R}^d$, $A \in \mathbb{R}^{d \times d}$, iar A și b nu depind de x .⁶²

iii. $(AB)^{-1} = B^{-1}A^{-1}$, unde $A, B \in \mathbb{R}^{d \times d}$.⁶³

iv. $(AB)^\top = B^\top A^\top$, unde $A, B \in \mathbb{R}^{d \times d}$.⁶⁴

v. $\det(AB) = \det(A)\det(B)$, unde $A, B \in \mathbb{R}^{d \times d}$.⁶⁵

vi. $\det(A) = \det(A^\top)$, unde $A \in \mathbb{R}^{d \times d}$.

b. Arătați că funcția $p_X(x; \mu, \Sigma)$ care a fost dată în enunț este într-adevăr funcție densitate de probabilitate (p.d.f.).

Indicație (2): Vă puteți folosi de următoarele proprietăți:

i. Pentru cazul $d = 1$, funcția $p_X(x; \mu, \sigma^2)$ este funcție densitate de probabilitate. (Demonstrația a fost făcută la problema 32.a.)

ii. În cazul în care matricea Σ este diagonală,

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_d^2 \end{bmatrix}, \text{ iar } \mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_d \end{bmatrix},$$

expresia funcției de densitate gaussiană multidimensională este identică cu produsul a d funcții de densitate de tip gaussian, unidimensionale și independente, prima funcție având media μ_1 și varianța σ_1^2 , a doua funcție având media μ_2 și varianța σ_2^2 , ..., a d -a funcție având media μ_d și varianța σ_d^2 . (Demonstrația acestei proprietăți a fost făcută la problema 34.)

Răspuns:

a. Sunt imediate următoarele relații:

$$z = B^{-1}(x - \mu) \xrightarrow{B \cdot} Bz = x - \mu \Rightarrow x = Bz + \mu \quad (29)$$

$$\frac{\partial x}{\partial z} = \frac{\partial(Bz + \mu)}{\partial z} \stackrel{(a.ii)}{=} B \quad (30)$$

$$(\det(\Sigma))^{1/2} = \sqrt{\det(BB^\top)} \stackrel{(a.v)}{=} \sqrt{\det(B)\det(B^\top)} \stackrel{(a.vi)}{=} \sqrt{(\det(B))^2} = |\det(B)|. \quad (31)$$

⁶²Aceasta este o consecință care decurge imediat din formula (6b) din documentul *Matrix Identities* de Sam Roweis, 1999.

⁶³Formula (1d) din același document (*Matrix Identities*) de Sam Roweis.

⁶⁴Formula (1c) din același document (*Matrix Identities*) de Sam Roweis.

⁶⁵Formula (2a) din același document (*Matrix Identities*) de Sam Roweis.

Vom rescrie acum expresia care constituie argumentul funcției $\exp()$ din definiția p.d.f.-ului distribuției gaussiene multidimensionale (vedeți relația (28)), în funcție de vectorul z .

$$\begin{aligned}
 & (x - \mu)^\top \Sigma^{-1} (x - \mu) \\
 & \stackrel{(29)}{=} (Bz + \mu - \mu)^\top (BB^\top)^{-1} (Bz + \mu - \mu) = (Bz)^\top (BB^\top)^{-1} (Bz) \\
 & \stackrel{(a.iii)}{=} (Bz)^\top ((B^\top)^{-1} B^{-1}) (Bz) \\
 & \stackrel{(a.iv)}{=} (z^\top B^\top) ((B^\top)^{-1} B^{-1}) (Bz) \\
 & \stackrel{\text{asoc.}}{=} z^\top (B^\top (B^\top)^{-1}) (B^{-1} B) z = z^\top I_d I_d z = z^\top z. \tag{32}
 \end{aligned}$$

Aplicând acum proprietatea (a.i), vom putea calcula p.d.f.-ul distribuției gaussiene asociate variabilei Z :

$$\begin{aligned}
 p_Z(z) &= p_X(x) \cdot \left| \det \left(\frac{\partial x}{\partial z} \right) \right| \\
 &= \frac{1}{(2\pi)^{d/2} \det(\Sigma)^{1/2}} \exp \left(-\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right) \left| \det \left(\frac{\partial x}{\partial z} \right) \right| \\
 &\stackrel{(30)(31)(32)}{=} \frac{1}{(2\pi)^{d/2} |\det(B)|} \exp \left(-\frac{1}{2} z^\top z \right) \frac{1}{|\det(B)|} \\
 &= \frac{1}{(2\pi)^{d/2}} \exp \left(-\frac{1}{2} z^\top z \right) \stackrel{z^\top = z^\top I_d}{=} \frac{1}{(2\pi)^{d/2} (\det(I_d))^{1/2}} \exp \left(-\frac{1}{2} z^\top I_d z \right) \\
 &= \mathcal{N}(z; \mathbf{0}, I_d).
 \end{aligned}$$

b. $p_X(x; \mu, \Sigma)$ este funcție densitate de probabilitate (p.d.f.) dacă:

- $p_X(x; \mu, \Sigma) \geq 0, \forall x \in \mathbb{R}^d$
- $I \stackrel{\text{not.}}{=} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} p_X(x; \mu, \Sigma) dx_d \dots dx_2 dx_1 = 1$.

Prima condiție este satisfăcută, pentru că numitorul fractiei [care este *constanta de normalizare*] din definiția funcției de densitate de probabilitate $p_X(x; \mu, \Sigma)$ este pozitiv, iar $\exp(y) > 0$ pentru orice $y \in \mathbb{R}$.⁶⁶

În continuare vom verifica a doua condiție.

În integrala I facem substituția (schimbarea de variabilă): $z = B^{-1}(x - \mu)$, unde $\Sigma = BB^\top, B \in \mathbb{R}^{d \times d}$. Matricea B există și este inversabilă după cum s-a precizat în enunț (vedeți *Observația* (1)).

De la punctul a rezultă imediat că:

$$\begin{aligned}
 I &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} p_Z(z; \mathbf{0}, I_d) dz_d \dots dz_2 dz_1 \\
 &\stackrel{(b.ii)}{=} \left(\int_{-\infty}^{\infty} p_{Z_1}(z_1; 0, 1) dz_1 \right) \left(\int_{-\infty}^{\infty} p_{Z_2}(z_2; 0, 1) dz_2 \right) \dots \left(\int_{-\infty}^{\infty} p_{Z_d}(z_d; 0, 1) dz_d \right) \\
 &\stackrel{(b.i)}{=} 1 \cdot 1 \cdot \dots \cdot 1 = 1.
 \end{aligned}$$

⁶⁶ *Observație importantă*: Spre deosebire de problema 32 — unde, la calcularea varianței pentru distribuția gaussiană unidimensională, făceam trecerea din sistemul de coordonate carteziene în sistemul de coordonate polare —, aici nu am schimbat sistemul de coordonate, ci practic am translat punctele, conform unei transformări liniare bijective. Am trecut astfel de la un “clopot” cu deschidere de formă eliptică la un “clopot” cu deschidere de formă eliptică circulară, cu centrul în originea sistemului de coordonate.

Deci, și a doua condiție este satisfăcută, ceea ce înseamnă că funcția $p_X(x; \mu, \Sigma)$ este într-adevăr funcție densitate de probabilitate (p.d.f.).

38.

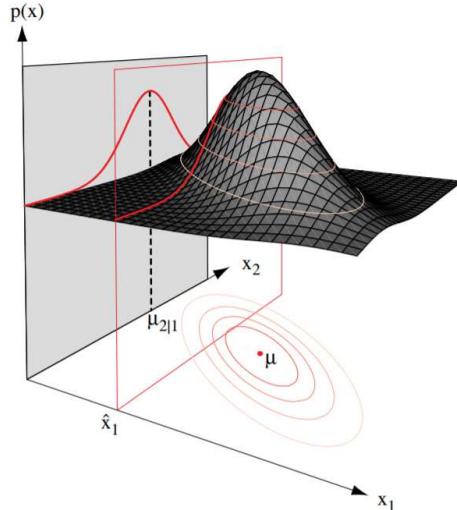
(Distribuția gaussiană bidimensională; o proprietate: distribuția condițională a unei componente în raport cu celalaltă componentă este tot de tip gaussian; identificarea parametrilor acestei distribuții condiționale)

■ □ *formulare de Liviu Ciortuz, după "Pattern Classification" (2nd ed.), [Appendix A. Mathematical foundations]*
R. Duda, P. Hart, D. Stork. John Wiley & Sons Inc., 2001

Fie X o variabilă aleatoare care urmează o distribuție gaussiană bidimensională de parametri μ (vectorul de medii) și Σ (matricea de covarianță). Așadar, $\mu = (\mu_1, \mu_2) \in \mathbb{R}^2$, iar $\Sigma \in \mathcal{M}_{2 \times 2}(\mathbb{R})$ este matrice simetrică și pozitiv definită.

Prin definiție, $\Sigma = Cov(X, X)$, unde $X \stackrel{\text{not.}}{=} (X_1, X_2)$, așadar $\Sigma_{ij} = Cov(X_i, X_j)$ pentru $i, j \in \{1, 2\}$. De asemenea, $Cov(X_i, X_i) = Var[X_i] \stackrel{\text{not.}}{=} \sigma_i^2 \geq 0$ pentru $i \in \{1, 2\}$, în vreme ce pentru $i \neq j$ avem $Cov(X_i, X_j) = Cov(X_j, X_i) \stackrel{\text{not.}}{=} \sigma_{ij}$. În sfârșit, dacă folosim *coeficientul de corelație* $\rho \stackrel{\text{def.}}{=} \frac{\sigma_{12}}{\sigma_1 \sigma_2}$, rezultă că putem scrie matricea de covarianță Σ sub forma următoare:⁶⁷

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{bmatrix}. \quad (33)$$



Demonstrați că ipoteza $X \sim \mathcal{N}(\mu, \Sigma)$ implică faptul că distribuția condițională $X_2|X_1$ este de tip gaussian, și anume $X_2|X_1 = x_1 \sim \mathcal{N}(\mu_{2|1}, \sigma_{2|1}^2)$, cu $\mu_{2|1} = \mu_2 + \rho \frac{\sigma_2}{\sigma_1} (x_1 - \mu_1)$ și $\sigma_{2|1}^2 = \sigma_2^2 (1 - \rho^2)$.

Observație: Pentru $X_1|X_2$, rezultatul este similar: $X_1|X_2 = x_2 \sim \mathcal{N}(\mu_{1|2}, \sigma_{1|2}^2)$, cu $\mu_{1|2} = \mu_1 + \rho \frac{\sigma_1}{\sigma_2} (x_2 - \mu_2)$ și $\sigma_{1|2}^2 = \sigma_1^2 (1 - \rho^2)$.

Credit: *Pattern Classification*, 2nd ed., R. Duda, P. Hart and D. Stork, 2001

Răspuns:

Conform definiției distribuției condiționale, $X_2|X_1 = x_1$ are funcția de densitate de probabilitate

$$p(x_2|x_1) = \frac{p_{X_1, X_2}(x_1, x_2)}{p_{X_1}(x_1)}, \quad (34)$$

⁶⁷Pentru definiția coeficientului de corelație, vedeți problema 19.

unde p_{X_1, X_2} este densitatea distribuției gaussiene bidimensionale de parametri μ și Σ , iar p_{X_1} este distribuția marginală a variabilei X_1 . Așadar,⁶⁸

$$p_{X_1, X_2}(x_1, x_2) = \frac{1}{(\sqrt{2\pi})^2 \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right) \text{ și} \quad (35)$$

$$p_{X_1}(x_1) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{2\sigma_1^2}(x_1 - \mu_1)^2\right). \quad (36)$$

Se constată imediat din relația (33) că *determinantul* matricei Σ este $|\Sigma| = \sigma_1^2 \sigma_2^2 (1 - \rho^2)$. Întrucât din expresia lui p_{X_1, X_2} de mai sus rezultă că Σ trebuie să fie matrice inversabilă, vom avea $\rho \in (-1, 1)$. De asemenea, pentru că se consideră $\sigma_1, \sigma_2 > 0$, rezultă că $\sqrt{|\Sigma|} = \sigma_1 \sigma_2 \sqrt{1 - \rho^2}$.

Inversa matricei Σ se calculează astfel:

$$\begin{aligned} \Sigma^{-1} &= \frac{1}{\sigma_1^2 \sigma_2^2 (1 - \rho^2)} \Sigma^* = \frac{1}{\sigma_1^2 \sigma_2^2 (1 - \rho^2)} \begin{bmatrix} \sigma_2^2 & -\rho \sigma_1 \sigma_2 \\ -\rho \sigma_1 \sigma_2 & \sigma_1^2 \end{bmatrix} \\ &= \frac{1}{(1 - \rho^2)} \begin{bmatrix} \frac{1}{\sigma_1^2} & -\frac{\rho}{\sigma_1 \sigma_2} \\ -\frac{\rho}{\sigma_1 \sigma_2} & \frac{1}{\sigma_2^2} \end{bmatrix} \end{aligned}$$

Prin urmare,

$$\begin{aligned} p_{X_1, X_2}(x_1, x_2) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \\ &\exp\left(-\frac{1}{2(1-\rho^2)}(x_1 - \mu_1, x_2 - \mu_2) \begin{bmatrix} \frac{1}{\sigma_1^2} & -\frac{\rho}{\sigma_1 \sigma_2} \\ -\frac{\rho}{\sigma_1 \sigma_2} & \frac{1}{\sigma_2^2} \end{bmatrix} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}\right) \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \\ &\exp\left(-\frac{1}{2(1-\rho^2)} \left[\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1}\right) \left(\frac{x_2 - \mu_2}{\sigma_2}\right) + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2 \right]\right) \end{aligned} \quad (37)$$

Observație: Din expresia pe care tocmai am obținut-o se observă ușor că $p_{X_1, X_2}(x_1, x_2)$ își atinge maximul atunci când $(x_1, x_2) = (\mu_1, \mu_2)$. Curbele de ecuație $p_{X_1, X_2}(x_1, x_2) = c$, pentru diverse valori ale lui $c \in \mathbb{R}^+$, se numesc *curbe de izocontur*. Ele au formă elipsoidală.

Înlocuind expresiile (36) și (37) în definiția (34), vom obține:

$$\begin{aligned} p(x_2|x_1) &= \frac{p_{X_1, X_2}(x_1, x_2)}{p_{X_1}(x_1)} \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \\ &\exp\left(-\frac{1}{2(1-\rho^2)} \left[\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1}\right) \left(\frac{x_2 - \mu_2}{\sigma_2}\right) + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2 \right]\right). \end{aligned}$$

⁶⁸Se poate demonstra că relația (36) rezultă din relația (35), prin integrare în raport cu x_2 . Demonstrația depășește cadrul de față și este lăsată ca exercițiu.

$$\begin{aligned}
& \left(\sqrt{2\pi}\sigma_1 \exp\left(\frac{1}{2} \left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2\right) \right)^{-1} \\
&= \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)} \left(\frac{x_2 - \mu_2}{\sigma_2} - \rho \frac{x_1 - \mu_1}{\sigma_1}\right)^2\right] \\
&= \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2} \left(\frac{x_2 - [\mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x_1 - \mu_1)]}{\sigma_2\sqrt{1-\rho^2}}\right)^2\right].
\end{aligned}$$

Din această expresie se observă că $X_2|X_1 = x_1$ urmează o distribuție gaussiană de parametri $\mu_{2|1} \stackrel{\text{not.}}{=} \mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x_1 - \mu_1)$ și $\sigma_{2|1}^2 \stackrel{\text{not.}}{=} \sigma_2^2(1 - \rho^2)$.

Observație: Rezultatul obținut la acest exercițiu se generalizează la [condiționarea pentru] distribuții marginale pentru distribuția gaussiană multidimensională, astfel: dacă $x \sim \mathcal{N}(\mu, \Sigma)$ și $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^{r+s}$, $x_1 \in \mathbb{R}^r$, $x_2 \in \mathbb{R}^s$, $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$, $\mu_1 \in \mathbb{R}^r$, $\mu_2 \in \mathbb{R}^s$, $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$, $\Sigma_{11} \in \mathbb{R}^{r \times r}$, $\Sigma_{12} = \Sigma_{21}^\top \in \mathbb{R}^{r \times s}$, $\Sigma_{22} \in \mathbb{R}^{s \times s}$, atunci $[x_1 \sim \mathcal{N}(\mu_1, \Sigma_{11}), x_2 \sim \mathcal{N}(\mu_2, \Sigma_{22})]$ și $x_1|x_2 \sim \mathcal{N}(\mu_{1|2}, \Sigma_{1|2})$, unde

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \quad \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \quad (38)$$

Similar, $x_{2|1} \sim \mathcal{N}(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$. Pentru demonstrație, vedeti documentul *More on Multivariate Gaussians* de Chuong B. Do (21 November 2008), sect. 3.2 și 3.3.

39. (Mixturi de distribuții multidimensionale: o proprietate)

• CMU, 2015 fall, Z. Bar-Joseph, E. Xing, midterm exam, pr. 7

Considerăm mixtura de distribuții

$$p(x) = \sum_{k=1}^K \pi_k p(x|z_k),$$

unde $x \in \mathbb{R}^d$, cu $d \in \mathbb{N}^*$ și, ca de obicei, z_k sunt valorile unei variabile de tip categorial, $\pi_k \in [0, 1]$ cu $\sum_{k=1}^K \pi_k = 1$, iar $p(x|z_k)$ sunt distribuții probabiliste, bineînțeles cu $k = 1, \dots, K$.

Presupunem că împărțim vectorul x în două părți, astfel: $x = (x_1, x_2)$.

Arătați că distribuția condiționată $p(x_2|x_1)$ este de asemenea o mixtură de distribuții,

$$p(x_2|x_1) = \sum_{k=1}^K \varepsilon_k p(x_2|x_1, z_k) \quad (39)$$

și găsiți expresia probabilității de selecție ε_k în funcție de π_k , $p(x_1|z_k)$ și $p(x_1)$.

Răspuns:

Vom exprima în mod convenabil distribuția condiționată $p(x_2|x_1)$ — care constituie membrul stâng al egalității din concluzia problemei — în aşa fel încât să punem în evidență relația dintre această distribuție și distribuțiile condiționate $p(x_2|x_1, z_k)$, care apar în membrul drept al aceleiași egalități:

$$\begin{aligned} p(x_2|x_1) &= \sum_{k=1}^K p(x_2, z_k|x_1) \stackrel{\text{form. mult.}}{=} \sum_{k=1}^K p(x_2|z_k, x_1) p(z_k|x_1) \\ &\stackrel{F. Bayes}{=} \sum_{k=1}^K p(x_2|z_k, x_1) p(x_1|z_k) \underbrace{p(z_k)}_{\pi_k} / p(x_1) \end{aligned}$$

Prima egalitate este justificată de proprietatea de aditivitate numărabilă din definiția funcției de probabilitate. Pentru obținerea celei de-a doua egalități am folosit regula de înmulțire (în varianta conditională), iar pentru cea de-a treia am aplicat regula lui Bayes.

Din ultima egalitate rezultă:⁶⁹

$$\varepsilon_k = \pi_k p(x_1|z_k) / p(x_1).$$

40. (Distribuția Bernoulli, distribuția gaussiană standard; intervale de încredere, teorema limită centrală, legea numerelor mari: aplicație la calculul erorii reale a unui clasificator)

■ • CMU, 2008 fall, Eric Xing, HW3, pr. 3.3

Nu demult, Chris s-a decis să folosească un nou clasificator (binar) care filtrează emailurile spam. Ulterior, el a vrut să evalueze cât de bun este acest clasificator. În acest scop, el a testat clasificatorul pe un mic set de date constituit din 100 de emailuri alese în mod aleatoriu dintre toate emailurile sale. Rezultatul pe care l-a obținut a fost următorul: 83 de emailuri au fost clasificate corect. Așadar rata erorii produse (sau: eroarea medie produsă) de clasificator pe acest mic set de date este de 17%. Este evident însă că eroarea aceasta este mai mică sau mai mare decât *eroarea reală*, pur și simplu datorită alegerii aleatoare a celor 100 de emailuri.

Dacă se consideră un nivel de încredere de 95%, în ce interval se va situa eroarea reală (ținând cont de acest experiment)?

Răspuns:

Vom nota cu X_i variabila aleatoare care reprezintă producerea unui mesaj email ($i = 1, \dots, n = 100$) și vom considera că $X_i = 1$ dacă emailul respectiv este clasificat eronat și 0 în cazul contrar.

Notăm cu $\mu \stackrel{\text{not.}}{=} e_{real}$ media variabilei X_i și cu σ^2 varianța variabilei X_i . (Valorile lui μ și σ nu depind de i).

Conform Teoremei Limită Centrală, eroarea la eșantionare,

$$e_{sample} \stackrel{\text{not.}}{=} \frac{X_1 + \dots + X_n}{n},$$

⁶⁹Se va observa că ε_k nu depinde de x_2 , însă depinde de x_1 .

văzută ca variabilă aleatoare, este aproximată de $\mathcal{N}(\mu, \sigma^2)$, distribuția normală de medie μ și variantă σ^2 . În consecință, notând

$$\begin{aligned} Z_n &= \frac{X_1 + \dots + X_n - n\mu}{\sqrt{n}\sigma} = \frac{\frac{X_1 + \dots + X_n}{n} - \mu}{\frac{\sigma}{\sqrt{n}}} \\ \text{fiindcă } \operatorname{Var} \left[\frac{X_1 + \dots + X_n}{n} \right] &= \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}, \end{aligned}$$

rezultă că $Z_n \sim \mathcal{N}(0, 1)$, deci pentru orice $a \in \mathbb{R}$ vom avea $P(Z_n \leq a) \rightarrow \Phi(a)$, unde $\Phi(x) \stackrel{\text{def}}{=} P(Z \leq x)$ desemnează funcția de distribuție cumulativă (engl., cumulative distribution function, c.d.f.) pentru distribuția normală standard (vedeți problema 33).

Folosind notațiile de mai sus, vom scrie următoarele echivalențe, care au loc pentru orice $a \geq 0$:

$$\begin{aligned} |Z_n| \leq a &\Leftrightarrow \left| \frac{X_1 + \dots + X_n - n\mu}{\sqrt{n}\sigma} \right| \leq a \Leftrightarrow \left| \frac{X_1 + \dots + X_n - n\mu}{n\sigma} \right| \leq \frac{a}{\sqrt{n}} \\ &\Leftrightarrow \left| \frac{X_1 + \dots + X_n - n\mu}{n} \right| \leq \frac{a\sigma}{\sqrt{n}} \Leftrightarrow \left| \frac{X_1 + \dots + X_n}{n} - \mu \right| \leq \frac{a\sigma}{\sqrt{n}} \\ &\Leftrightarrow |e_{sample} - e_{real}| \leq \frac{a\sigma}{\sqrt{n}} \Leftrightarrow |e_{real} - e_{sample}| \leq \frac{a\sigma}{\sqrt{n}} \\ &\Leftrightarrow -\frac{a\sigma}{\sqrt{n}} \leq e_{real} - e_{sample} \leq \frac{a\sigma}{\sqrt{n}} \Leftrightarrow e_{sample} - \frac{a\sigma}{\sqrt{n}} \leq e_{real} \leq e_{sample} + \frac{a\sigma}{\sqrt{n}} \\ &\Leftrightarrow e_{real} \in [e_{sample} - \frac{a\sigma}{\sqrt{n}}, e_{sample} + \frac{a\sigma}{\sqrt{n}}] \end{aligned} \tag{40}$$

Pentru a determina efectiv intervalul de încredere cerut, trebuie ca mai întâi să mai aflăm a și σ și apoi să le înlocuim în relația (40).

Constanta a se determină ușor ținând cont de faptul că în enunț se precizează *nivelul de încredere* (95%) pentru apartenența erorii reale la intervalul specificat. Această restricție corespunde relației $P(|Z_n| \leq a) = 0.95$. Pe de altă parte, întrucât $\Phi(-a) + \Phi(a) = 1$, avem:

$$P(|Z_n| \leq a) = \Phi(a) - \Phi(-a) = 2\Phi(a) - 1,$$

deci

$$P(|Z_n| \leq a) = 0.95 \Leftrightarrow 2\Phi(a) - 1 = 0.95 \Leftrightarrow \Phi(a) = 0.975 \Leftrightarrow a \cong 1.97 \text{ (vedeți pr. 33).}$$

Ne-a mai rămas de calculat valoarea lui σ . Vom ține cont de faptul că

$$\sigma^2 \stackrel{\text{not.}}{=} \operatorname{Var}_{real} = e_{real}(1 - e_{real}), \tag{41}$$

ultima egalitate având loc fiindcă toate variabilele X_i , identic distribuite, sunt de tip Bernoulli. În relația (41) vom putea approxima e_{real} cu e_{sample} , conform legii numerelor mari (varianta „slabă“ a acestei legi este prezentată la problema 21.f).

În sfârșit, putem scrie:

$$\frac{a\sigma}{\sqrt{n}} = \frac{1.97\sqrt{0.17(1-0.17)}}{\sqrt{100}} \cong 0.07,$$

deci relația (40) devine $e_{real} \in [0.10, 0.24]$.

Sumarizând, putem afirma cu încredere de 95% că eroarea reală a clasificatorului folosit de Chris este mai mare sau egală cu 0.1 și mai mică sau egală cu 0.24.

41. (Familia de distribuții exponențiale: definiție, exemplificare)

• CMU, 2015 fall, Z. Bar-Joseph, E. Xing, HW1, pr. 1.1.a

Multe dintre distribuțiile probabiliste des folosite în statistică și învățarea automată fac parte din *familia exponențială*. Prin *definiție*, familia exponențială este formată din acele distribuții ale căror funcții masă de probabilitate (în cazul distribuțiilor discrete), respectiv funcții densitate de probabilitate (în cazul distribuțiilor continue) pot fi exprimate sub forma următoare:

$$f_X(x) = b(x) e^{\eta(\theta) \cdot T(x) - a(\theta)},$$

unde $T(x)$, $b(x)$, $\eta(\theta)$ și $a(\theta)$ sunt cunoscute.

Demonstrați că următoarele trei distribuții

- a. **distribuția multinomială**, având parametrii $\theta_i > 0$ pentru $i = 1, \dots, k$, astfel încât $\sum_{i=1}^k \theta_i = 1$ și fiind definită prin

$$p(x|\theta) = \frac{(\sum_{i=1}^k x_i)!}{\prod_{i=1}^k x_i!} \prod_{i=1}^k \theta_i^{x_i}, \quad (42)$$

unde $x = (x_1, \dots, x_k)$, cu $x_i \in \mathbb{N}$;

- b. **distribuția Dirichlet**,⁷⁰ având parametrii $\theta_i > 0$ pentru $i = 1, \dots, k$, astfel încât $\sum_{i=1}^k \theta_i = 1$ și fiind definită prin

$$p(x|\theta) = \frac{\Gamma(\sum_{i=1}^k \theta_i)}{\prod_{i=1}^k \Gamma(\theta_i)} \prod_{i=1}^k x_i^{\theta_i-1}; \quad (43)$$

unde Γ este funcția Gamma a lui Euler,⁷¹ iar $x = (x_1, \dots, x_k)$, cu $x_i > 0$;

- c. **distribuția gaussiană multidimensională**, având parametrii $\mu \in \mathbb{R}^d$ și $\Sigma \in \mathbb{R}^{d \times d}$, matrice simetrică și pozitiv definită, și fiind definită prin

$$p(x|\mu, \Sigma) = \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} \exp \left(-\frac{1}{2} (x - \mu)^\top \Sigma (x - \mu) \right).$$

⁷⁰Această distribuție poartă numele lui Johann Peter Gustav Lejeune Dirichlet, matematician prusac (1805-1859). Dirichlet a studiat la universitatea din Paris și a predat la universitățile din Breslau, Berlin și Göttinghen. Soția sa a fost Rebecka Mendelssohn Bartholdy, sora mai mică a compozitorului Felix Mendelssohn Bartholdy.

⁷¹Vedeți problema 31.b.

fac parte din familia exponențială.

Răspuns:

a. Cazul distribuției multinomiale:⁷²

$$\begin{aligned} p(x|\theta) &= \frac{(\sum_{i=1}^k x_i)!}{\prod_{i=1}^k x_i!} \prod_{i=1}^k \theta_i^{x_i} = \frac{(\sum_{i=1}^k x_i)!}{\prod_{i=1}^k x_i!} \exp \left(\ln \prod_{i=1}^k \theta_i^{x_i} \right) \\ &= \frac{(\sum_{i=1}^k x_i)!}{\prod_{i=1}^k x_i!} \exp \left(\sum_{i=1}^k \ln \theta_i^{x_i} \right). \end{aligned}$$

Așadar, luând $b(x) = \frac{(\sum_{i=1}^k x_i)!}{\prod_{i=1}^k x_i!}$, $\eta(\theta) = (\ln \theta_1, \dots, \ln \theta_k)$, $T(x) = x$ și $a(\theta) = 0$ rezultă că distribuția multinomială aparține familiei exponențiale.

b. Cazul distribuției Dirichlet:⁷³

$$\begin{aligned} p(x|\theta) &= \frac{\Gamma(\sum_{i=1}^k \theta_i)}{\prod_{i=1}^k \Gamma(\theta_i)} \prod_{i=1}^k x_i^{\theta_i-1} = \exp \left(\ln \left(\frac{\Gamma(\sum_{i=1}^k \theta_i)}{\prod_{i=1}^k \Gamma(\theta_i)} \prod_{i=1}^k x_i^{\theta_i-1} \right) \right) \\ &= \exp \left(\ln \frac{\Gamma(\sum_{i=1}^k \theta_i)}{\prod_{i=1}^k \Gamma(\theta_i)} + \sum_{i=1}^k (\theta_i - 1) \ln x_i \right). \end{aligned}$$

Considerând $b(x) = 1$, $\eta(\theta) = (\theta_1 - 1, \dots, \theta_k - 1)$, $T(x) = (\ln x_1, \dots, \ln x_k)$ și $a(\theta) = -\ln \frac{\Gamma(\sum_{i=1}^k \theta_i)}{\prod_{i=1}^k \Gamma(\theta_i)}$, rezultă că distribuția Dirichlet este (și ea) membră a familiei exponențiale.

c. Cazul distribuției gaussiene multidimensionale:

$$\begin{aligned} p(x|\mu, \Sigma) &= \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} \exp \left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1} (x - \mu) \right) \\ &= \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} \exp \left(-\frac{1}{2}x^\top \Sigma^{-1} x + \mu^\top \Sigma^{-1} x - \frac{1}{2}\mu^\top \Sigma^{-1} \mu \right) \\ &= \frac{1}{(2\pi)^{k/2}} \exp \left(-\frac{1}{2}tr(\Sigma^{-1}xx^\top) + \mu^\top \Sigma^{-1} x - \frac{1}{2}\mu^\top \Sigma^{-1} \mu - \frac{1}{2}\ln |\Sigma| \right). \end{aligned}$$

Și această distribuție aparține familiei exponențiale, fiindcă putem scrie:

$$\begin{aligned} b(x) &= (2\pi)^{-k/2} \\ \eta(\theta) &= (\Sigma^{-1}\mu; -\frac{1}{2}vec(\Sigma^{-1})) \\ T(x) &= (x; vec(xx^\top)) \\ a(\theta) &= \frac{1}{2}\mu^\top \Sigma^{-1} \mu + \frac{1}{2}\ln |\Sigma|. \end{aligned}$$

⁷²Distribuția multinomială constituie o generalizare a distribuției binomiale (care la rândul ei generalizează distribuția Bernoulli). Pe de altă parte, distribuția multinomială poate fi văzută ca o generalizare a distribuției categoriale (care este, și ea, o generalizare a distribuției Bernoulli).

⁷³Constanta de normalizare din expresia distribuției Dirichlet este $1/B(\theta_1, \dots, \theta_k)$, unde B este funcția Beta multidimensională. Distribuția Dirichlet — o generalizare a distribuției Beta; vedeti exercițiile 43, 128 și 129 — este distribuție conjugată (a priori) pentru distribuția categorială și pentru distribuția multinomială. (Pentru definitia noțiunii de *distribuții conjugate* vedeti enunțul problemei 43.B.) Înțând cont de faptul că $\Gamma(x) = (x-1)!$ pentru orice $x \in \mathbb{N}^*$, veți putea observa că există o mare similaritate între expresiile (42) și (43).

Precizăm că $x^\top \Sigma^{-1} x = \text{tr}(x^\top \Sigma^{-1} x)$ fiindcă $x^\top \Sigma^{-1} x \in \mathbb{R}$, apoi $\text{tr}(x^\top \Sigma^{-1} x) = \text{tr}(\Sigma^{-1} x x^\top)$ fiindcă $\text{tr}(ABC) = \text{tr}(BCA)$ pentru orice matrice A, B și C astfel încât ABC este matrice pătratică⁷⁴ și, în sfârșit, se poate verifica ușor că $\text{tr}(\Sigma^{-1} x x^\top) = \text{vec}(\Sigma^{-1}) \cdot \text{vec}(x x^\top)$ fiindcă matricea $x x^\top$ este simetrică ($(x x^\top)^\top = x x^\top$). Prin $\text{vec}(A)$ am desemnat vectorul obținut prin liniarizarea matricei A . Similar, am notat cu $\text{tr}(A)$ suma elementelor de pe diagonala principală a unei matrice pătratice oarecare A .

0.1.4 Estimarea parametrilor unor distribuții probabiliste

42.

(Distribuții de tip Bernoulli;
calculul verosimilității datelor;
estimarea parametrilor în sensul MLE)

CMU, 2005 fall, T. Mitchell, A. Moore, midterm, pr. 1.3

Avem două monede. Probabilitatea de apariție a stemei este θ în cazul primei monede și 2θ în cazul celei de-a două monede.

Presupunem că aruncăm aceste două monede de mai multe ori, în mod independent una de cealaltă, și obținem rezultatele din tabelul alăturat.

Moneda	Rezultat
1	stema
2	ban
2	ban
2	ban
2	stema

- a. Care este log-verosimilitatea acestor date în funcție de θ ?
- b. Cât este estimarea / valoarea de verosimilitate maximă (engl., maximum likelihood, ML) a lui θ ?

Răspuns:

a. Dacă notăm pe de o parte cu $stemă_1$ și ban_1 fețele primei monede și cu $stemă_2$ și ban_2 fețele celei de-a două monede, iar pe de altă parte cu x_1, x_2, \dots, x_5 cele cinci evenimente aleatoare din enunțul problemei, atunci verosimilitatea acestor date în raport cu parametrul θ este:

$$\begin{aligned} L(\theta) &\stackrel{\text{def.}}{=} P(date \mid \theta) = P(x_1 = stemă_1, x_2 = ban_2, x_3 = ban_2, x_4 = ban_2, x_5 = stemă_2 \mid \theta) \\ &\stackrel{i.i.d.}{=} P(stemă_1 \mid \theta) \cdot [P(ban_2 \mid \theta)]^3 \cdot P(stemă_2 \mid \theta) \\ &= \theta \cdot (1 - 2\theta)^3 \cdot 2\theta = 2\theta^2 \cdot (1 - 2\theta)^3. \end{aligned}$$

Notăm log-verosimilitatea acestor date cu $\ell(\theta)$ și o calculăm astfel:⁷⁵

$$\ell(\theta) \stackrel{\text{def.}}{=} \ln L(\theta) = \ln P(date \mid \theta) = \ln(2\theta^2 \cdot (1 - 2\theta)^3) = \ln 2 + 2 \ln \theta + 3 \ln(1 - 2\theta).$$

⁷⁴Vedeți documentul *Linear algebra review and reference* (29.09.2012) de Zico Kolter de la Stanford University, pag. 9.

⁷⁵Am ales ca bază a logaritmului numărul e . De fapt, pentru a asigura consecvența proprietăților de monotonie a funcției de verosimilitate, ar fi suficient să lucrăm cu o bază oarecare, supraunitară, a . Într-un astfel de caz, la derivare, vom avea în plus (față de cazul logaritmului natural, \ln) un factor constant, pozitiv.

b. Estimarea de verosimilitate maximă a lui θ este prin definiție $\hat{\theta}_{MLE} \stackrel{not.}{=} \operatorname{argmax}_\theta \ell(\theta)$ — mai exact, ținând cont de enunț, $\operatorname{argmax}_{\theta \in (0,1/2)} \ell(\theta)$ — și se calculează cu ajutorul derivatei de ordinul întâi a funcției $\ell(\theta)$:

$$\frac{\partial \ell(\theta)}{\partial \theta} = \frac{\partial}{\partial \theta} (\ln 2 + 2 \ln \theta + 3 \ln(1 - 2\theta)) = 0 + \frac{2}{\theta} + \frac{3(-2)}{1 - 2\theta}$$

$$\frac{\partial \ell(\theta)}{\partial \theta} = 0 \Leftrightarrow \frac{2}{\theta} - \frac{6}{1 - 2\theta} = 0 \Leftrightarrow 2(1 - 2\theta) - 6\theta = 0 \Leftrightarrow 2 - 10\theta = 0 \Leftrightarrow \theta_{MLE} = \frac{1}{5} \in (0, \frac{1}{2}).$$

Se poate verifica ușor că $\theta = \frac{1}{5}$ este punct de maxim pentru funcția ℓ . Într-adevăr, ținând cont de faptul că din enunț se poate deduce că $\theta \in (0, 1/2)$, rezultă că $\frac{\partial \ell(\theta)}{\partial \theta} \stackrel{not.}{=} \ell'(\theta) > 0$ pentru $\theta < \frac{1}{5}$ și $\ell'(\theta) < 0$ pentru $\theta > \frac{1}{5}$, deci funcția ℓ este strict crescătoare pe intervalul $(0, 1/5]$ și strict descrescătoare pe intervalul $[1/5, 1/2)$.

43.

(Distribuția Bernoulli: estimarea parametrului în sensul MLE și, respectiv, MAP)

■ □ • ○ CMU, 2015 spring, T. Mitchell, N. Balcan, HW2, pr. 2

Presupunem că „observăm“ valorile variabilelor aleatoare X_1, \dots, X_n care sunt distribuite în mod identic și independent (engl., independent and identically distributed, i.i.d.), conform unei singure distribuții Bernoulli având parametrul θ . Cu alte cuvinte, pentru fiecare variabilă X_i , știm că

$$P(X_i = 1) = \theta, \quad \text{iar} \quad P(X_i = 0) = 1 - \theta.$$

Scopul nostru în cele ce urmează este să estimăm valoarea parametrului θ pornind de la valorile „observate“ ale variabilelor X_1, \dots, X_n .

A. Estimarea în sensul verosimilității maxime (engl., Maximum Likelihood Estimation, MLE)

Introducere: Pentru orice valoare $\hat{\theta}$ a parametrului θ , fixată în mod arbitrar, putem să calculăm probabilitatea producerii „observațiilor“ [adică a valorilor variabilelor aleatoare] X_1, \dots, X_n . Această probabilitate a [producerii] datelor observabile este adeseori numită *verosimilitatea datelor*, iar funcția $L(\hat{\theta})$ care asociază fiecarei valori $\hat{\theta}$ verosimilitatea respectivă a datelor se numește *funcția de verosimilitate*. În mod natural, a „estima“ valoarea necunoscută a parametrului θ revine la a identifica acea valoare $\hat{\theta}$ [a lui θ] care maximizează funcția de verosimilitate. În mod formal, putem scrie acest fapt astfel:

$$\hat{\theta}_{MLE} \stackrel{def.}{=} \operatorname{argmax}_{\hat{\theta}} L(\hat{\theta}).$$

Această notație înseamnă: i. $\hat{\theta}_{MLE}$ aparține domeniului de valori posibile ale parametrului θ , și ii. $L(\hat{\theta}) \leq L(\hat{\theta}_{MLE})$ pentru orice $\hat{\theta}$ care aparține respectivului domeniu de valori.

a. Scrieți expresia [pentru calculul valorilor] funcției de verosimilitate, $L(\hat{\theta})$. Această expresie ar trebui să depindă de valorile variabilelor aleatoare X_1, \dots, X_n , precum și de $\hat{\theta}$, valoarea ipotetică a parametrului θ . Depinde oare valoarea acestei expresii de ordinea în care apar variabilele aleatoare?

- b. Presupunem că $n = 10$ și că setul de date conține șase de 1 și patru de 0. Scrieți un mic program de calculator care trasează graficul funcției de verosimilitate pentru acest set de date pentru fiecare valoare a lui $\hat{\theta}$ din multimea $\{0, 0.01, 0.02, \dots, 1.0\}$.⁷⁶ În acest grafic, axa x -ilor va trebui să-i corespundă lui $\hat{\theta}$, iar axa y -ilor lui $L(\hat{\theta})$. Pe axa y -ilor veți scala valorile astfel încât să se poată vedea variația respectivelor valori. Stabiliți cât este $\hat{\theta}_{MLE}$, identificând pe axa x -ilor acea valoare a lui $\hat{\theta}$ pentru care se atinge maximul funcției de verosimilitate.
- c. Găsiți expresia analitică (engl., closed-form formula) pentru $\hat{\theta}_{MLE}$, estimarea de verosimilitate maximă a lui $\hat{\theta}$. Pentru datele de la punctul b, concordă oare rezultatul dat de expresia analitică cu rezultatul obținut folosind graficul?
- d. Generați alte trei grafice pentru funcția de verosimilitate:
 unul pentru $n = 5$, setul de date conținând trei de 1 și doi de 0;
 unul pentru $n = 100$, setul de date conținând șaizeci de 1 și patruzeci de 0;
 unul pentru $n = 10$, cu cinci de 1 și cinci de 0.
- e. Pentru diferențele seturi de date de mai sus (la punctele b și d), comparați funcțiile de verosimilitate și estimările în sens MLE.

B. Estimarea în sensul probabilității maxime a posteriori (engl., Maximum A posteriori (MAP) Probability Estimation)

Introducere: La estimarea în sens MLE (vedeți partea A), am tratat valoarea „adevărată“ a parametrului θ ca pe un număr fixat (nealeator). Însă alteori — de exemplu, în cazurile în care dispunem de anumite cunoștințe a priori în legătură cu θ —, este util să-l tratăm pe θ ca fiind o variabilă aleatoare și să exprimăm aceste cunoștințe a priori sub forma unei distribuții de probabilitate a priori peste θ . Să presupunem, de exemplu, că valorile [variabilelor] X_1, \dots, X_n sunt generate în modul următor:

- Mai întâi, valoarea lui θ este generată folosind o anumită distribuție de probabilitate a priori.
- Apoi, valorile [variabilelor] X_1, \dots, X_n sunt generate în mod independent cu ajutorul unei distribuții Bernoulli care folosește pentru parametrul ei (θ) valoarea generată mai sus.

Întrucât atât θ cât și X_1, \dots, X_n sunt văzute ca [fiind] variabile aleatoare, lor li se poate asocia o distribuție de probabilitate comună (engl., joint probability). În acest context / cadru, o modalitate naturală de a estima valoarea lui θ constă pur și simplu în a alege valoarea sa cea mai probabilă ținând cont de distribuția a priori aleasă și de datele observabile X_1, \dots, X_n :⁷⁷

$$\hat{\theta}_{MAP} \stackrel{\text{def.}}{=} \underset{\hat{\theta}}{\operatorname{argmax}} P(\theta = \hat{\theta} | X_1, \dots, X_n).$$

Aceasta se numește estimarea de probabilitate maximă a posteriori (engl., maximum a posteriori probability, MAP) a lui θ . Folosind formula lui Bayes, putem scrie probabilitatea a posteriori a lui θ astfel:

$$P(\theta = \hat{\theta} | X_1, \dots, X_n) = \frac{P(X_1, \dots, X_n | \theta = \hat{\theta}) P(\theta = \hat{\theta})}{P(X_1, \dots, X_n)}.$$

⁷⁶LC: Alternativ, puteți folosi cunoștințele de analiză matematică din liceu pentru a trasa graficul cerut aici (ori pentru a stabili, în principiu, alura lui). Similar pentru punctul d.

⁷⁷Din punct de vedere matematic, expresia ținând cont se va traduce prin folosirea unei probabilități condiționate.

Întrucât probabilitatea de la numitor nu depinde de $\hat{\theta}$, estimarea în sens MAP a lui θ poate fi scrisă astfel:

$$\hat{\theta}_{MAP} = \operatorname{argmax}_{\hat{\theta}} P(X_1, \dots, X_n | \theta = \hat{\theta}) P(\theta = \hat{\theta}) = \operatorname{argmax}_{\hat{\theta}} L(\hat{\theta}) P(\theta = \hat{\theta}).$$

Cu alte cuvinte, estimarea în sens MAP a lui θ este valoarea $\hat{\theta}$ care maximizează produsul dintre funcția de verosimilitate și distribuția a priori a lui θ . În cazul în care această distribuție a priori este continuă și funcția ei de densitate de probabilitate (engl., probability density function, p.d.f.) este p , estimarea în sens MAP a lui θ este dată de formula

$$\hat{\theta}_{MAP} = \operatorname{argmax}_{\hat{\theta}} L(\hat{\theta}) p(\hat{\theta}).$$

Definiție: Dacă, la estimarea în sens MAP, distribuția *a posteriori* a parametrului λ în raport cu datele X , adică $P(\lambda|X)$, face parte din aceeași familie ca și distribuția *a priori* a parametrului, adică $P(\lambda)$, spunem că *i.* distribuția a priori $P(\lambda)$ și distribuția a posteriori $P(\lambda|X)$ sunt *distribuții conjugate*, și *ii.* distribuția a priori $P(\lambda)$ este o *conjugată* [a priori] pentru funcția de verosimilitate $P(X|\lambda)$.⁷⁸

În cele ce urmează vom folosi ca distribuție a priori pentru parametrul θ distribuția Beta(3, 3), care are funcția densitate de probabilitate următoare:

$$p(\hat{\theta}) = \frac{\hat{\theta}^2(1-\hat{\theta})^2}{B(3,3)},$$

unde $B(\alpha, \beta)$ desemnează funcția Beta, iar $B(3,3) = \frac{1}{30}$.⁷⁹

f. Să presupunem, ca la punctul b , că $n = 10$ și că „observăm“ șase de 1 și patru de 0. Scrieți un mic program care trasează graficul funcției $\hat{\theta} \mapsto L(\hat{\theta}) p(\hat{\theta})$ pentru aceleasi valori ale lui $\hat{\theta}$ ca la punctul b . Estimați $\hat{\theta}_{MAP}$, identificând pe axa x -ilor acea valoare a lui $\hat{\theta}$ pentru care se atinge maximul funcției de mai sus.

g. Determinați formula analitică pentru $\hat{\theta}_{MAP}$, estimarea în sens MAP a lui $\hat{\theta}$. Pentru datele de la punctul f, concordă oare rezultatul dat de expresia analitică cu rezultatul obținut folosind graficul?

h. Comparați estimările în sens MAP cu cele în sens MLE pentru datele de la punctul b. Explicați în mod succint diferențele semnificative.

i. Comentați modul cum evoluează relația dintre estimările MAP și MLE pe măsură ce n (numărul de „observări“) tinde la infinit, presupunând că în procesul de trecere la limită raportul $\#\{X_i = 1\}/\#\{X_i = 0\}$ rămâne constant.

Răspuns:

⁷⁸Pentru un exemplu, vedeți rezolvarea problemei 46.b.ii, unde se arată că distribuția Gamma este distribuție conjugată (a priori) pentru distribuția Poisson.

⁷⁹P.d.f.-ul distribuției Beta este definit astfel:

$$f(x; \alpha, \beta) = \text{constant} \cdot x^{\alpha-1} (1-x)^{\beta-1} = \frac{1}{B(\alpha, \beta)} \cdot x^{\alpha-1} (1-x)^{\beta-1} = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot x^{\alpha-1} (1-x)^{\beta-1},$$

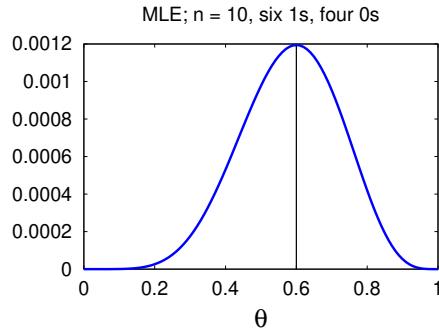
unde Γ desemnează funcția Gamma a lui Euler. Vă reamintim — vedeți problema 31.b — că $\Gamma(x) = (x-1)!$ pentru orice $x \in \mathbb{N}^*$. Pentru câteva combinații de valori ale parametrilor α și β , graficele p.d.f.-urilor corespunzătoare sunt date la problema 129.

a. Întrucât variabilele X_i sunt independente, putem scrie:

$$\begin{aligned} L(\hat{\theta}) &\stackrel{\text{def.}}{=} P_{\hat{\theta}}(X_1, \dots, X_n) \stackrel{i.i.d.}{=} \prod_{i=1}^n P_{\hat{\theta}}(X_i) = \prod_{i=1}^n (\hat{\theta}^{X_i} \cdot (1 - \hat{\theta})^{1-X_i}) \\ &= \hat{\theta}^{\#\{X_i=1\}} \cdot (1 - \hat{\theta})^{\#\{X_i=0\}}, \end{aligned}$$

unde notația $\#\{\cdot\}$ desemnează numărul de variabile X_i pentru care este satisfăcută condiția înscrisă între paranteze. La ultima egalitate am utilizat faptul că $Val(X_i) = \{0, 1\}$.⁸⁰ Evident, funcția de verosimilitate nu depinde de ordinea prezentării datelor.

b. Prezentăm în figura alăturată graficul cerut, pe care l-am obținut cu ajutorul unui program Matlab relativ simplu [pe care-l puteți găsi pe site-ul acestei culegeri].



c. Din cauza produselor care apar în funcția de verosimilitate $L(\theta)$, este mai convenabil să folosim funcția de log-verosimilitate: $l(\theta) \stackrel{\text{def.}}{=} \ln(L(\theta))$. Întrucât funcția \ln este crescătoare, acea valoare $\hat{\theta}$ [a argumentului θ] pentru care se atinge maximul funcției de log-verosimilitate este identică cu acel $\hat{\theta}$ pentru care se atinge maximul funcției de verosimilitate. Făcând uz de proprietățile funcției \ln , atunci când $\hat{\theta} \in (0, 1)$ putem scrie $l(\hat{\theta})$ astfel:

$$l(\hat{\theta}) = \ln(\hat{\theta}^{n_1} \cdot (1 - \hat{\theta})^{n_0}) = n_1 \ln(\hat{\theta}) + n_0 \ln(1 - \hat{\theta}),$$

unde $n_1 \stackrel{\text{not.}}{=} \#\{X_i = 1\}$, iar $n_0 \stackrel{\text{not.}}{=} \#\{X_i = 0\}$.

Derivatele de ordinul întâi și al doilea ale funcției de log-verosimilitate l pentru $\hat{\theta} \in (0, 1)$ se scriu astfel:

$$l'(\hat{\theta}) = \frac{n_1}{\hat{\theta}} - \frac{n_0}{1 - \hat{\theta}} \quad \text{și} \quad l''(\hat{\theta}) = -\left(\frac{n_1}{\hat{\theta}^2} + \frac{n_0}{(1 - \hat{\theta})^2}\right) \text{ pentru } \hat{\theta} \in (0, 1).$$

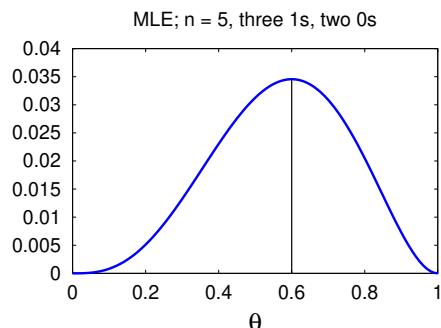
Întrucât derivata de ordin secund a funcției l este negativă pe tot domeniul ei de definiție (adică, intervalul $(0, 1)$), rezultă că l este funcție concavă și vom putea găsi valoarea argumentului pentru care ea își atinge maximul rezolvând ecuația $l'(\theta) = 0$. Soluția acestei ecuații se obține printr-un calcul algebraic simplu și este

$$\hat{\theta}_{MLE} = \frac{n_1}{n_1 + n_0}.$$

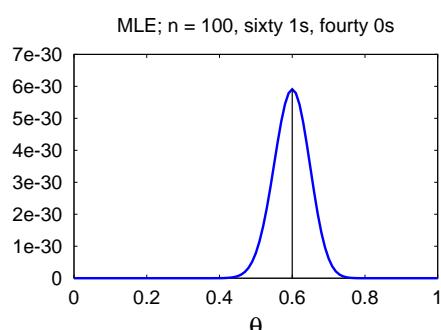
Se constată ușor că această formulă / expresie este în concordanță cu punctul de maxim din graficul care a fost obținut la punctul f .

⁸⁰Acesta este așa-numitul *artificiu al ridicării la putere* (engl., the exponentiation trick). El va fi folosit la unele probleme de la capitolul de *Clasificare Bayesiană*, la secțiunea *Algoritmul EM/GMM* de la capitolul de *Clusterizare*, precum și la capitolul *Schema algoritmică EM*.

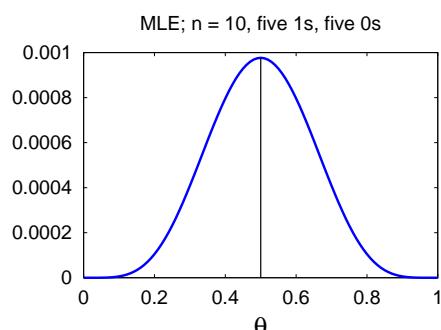
d. Pentru $n = 5$, cu trei de 1 și doi de 0, modificând ușor programul scris pentru punctul b, se va obține graficul alăturat.



În mod similar, pentru $n = 100$, cu șaizeci de 1 și patruzeci de 0 obținem graficul alăturat.



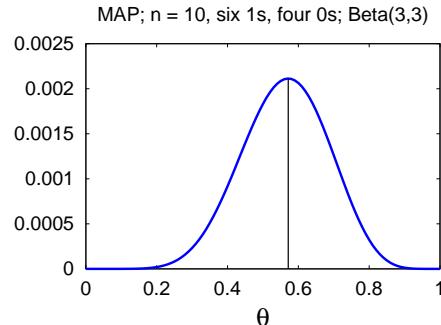
În sfârșit, pentru $n = 10$, cu cinci de 1 și cinci de 0 obținem:



e. Conform expresiei analitice obținute la punctul c, estimarea de verosimilitate maximă (MLE) a parametrului θ al distribuției Bernoulli este egală cu proporția „observațiilor“ 1 în ansamblul datelor. Așadar, în cazul primelor trei grafice, estimarea în sens MLE a lui θ este 0.6, iar în cazul ultimului grafic este 0.5. Pe măsură ce numărul de instanțe (n) crește, atunci când proporția instanțelor 1 se păstrează funcția de verosimilitate va avea vârful „adunat“ din ce în ce mai mult (sub forma unei *fleșe*⁸¹) în jurul valorii maxime, iar aceste valori maxime vor fi din ce în ce mai mici.

⁸¹Cf. DEX, *fleșă* (din fr. flèche) este un acoperiș foarte înalt, sub formă de piramidă sau de con, folosit mai ales în evul mediu la clădirile monumentale ale bisericilor.

- f. Un alt program Matlab relativ simplu — aflat de asemenea pe site-ul acestei culegeri — a produs graficul alăturat. (Este util să-l comparați cu graficul de la punctul b.) Abscisa punctului de maxim este $\theta \approx 0.571$.



- g. La fel ca în cazul estimării în sensul verosimilității maxime (MLE), vom aplica și aici funcția \ln înainte de a găsi valoarea lui $\hat{\theta}$ care maximizează funcția de probabilitate a posteriori. Așadar, urmărim să maximizăm funcția

$$l(\hat{\theta}) = \ln(L(\hat{\theta}) \cdot p(\hat{\theta})) = \ln(\hat{\theta}^{n_1+2} \cdot (1 - \hat{\theta})^{n_0+2}) - \ln(B(3, 3)).$$

Constanta de normalizare pentru distribuția a priori corespunde aici unei constante adiționale. Prin urmare, derivatele de ordinul întâi și al doilea [ale acestei noi funcții] devin identice cu cele din cazul estimării în sensul MLE, cu singura diferență că $n_1 + 2$ și $n_0 + 2$ îi vor înlocui pe n_1 și respectiv n_0 . Rezultă că forma analitică (engl., the closed form formula) pentru estimarea în sensul MAP a parametrului θ este următoarea:

$$\hat{\theta}_{MAP} = \frac{n_1 + 2}{n_1 + n_0 + 4}$$

Se constată ușor că această formulă / expresie este în concordanță cu punctul de maxim din graficul care a fost obținut la punctul f : $8/14 \approx 0.571$.

- h. Estimarea în sens MAP [a parametrului θ] este identică / egală cu estimarea în sens MLE dacă se mai adaugă patru variabile aleatoare *virtuale*, dintre care două iau valoarea 1, iar două iau valoarea 0. Acestea fac ca valoarea estimatorului MAP să fie împinsă mai aproape de valoarea 0.5 (și anume, la $8/14 \approx 0.571$); din această cauză $\hat{\theta}_{MAP}$ este mai mic decât $\hat{\theta}_{MLE}$ (care este $6/10 = 0.6$).

- i. Este evident că pe măsură ce n crește, tinzând la infinit, influența celor patru variabile aleatoare virtuale dispare, iar cei doi estimatori devin egali:

$$\hat{\theta}_{MAP} = \frac{n_1 + 2}{\underbrace{n_1 + n_0 + 4}_n} = \frac{\frac{n_1}{n} + \frac{2}{n}}{1 + \frac{4}{n}} \rightarrow \frac{n_1}{n} = \text{const} = \hat{\theta}_{MLE}.$$

44.

(Distribuția categorială: estimarea parametrilor, în sensul verosimilității maxime (MLE))

■ □ • ○ CMU, 2009 spring, Ziv Bar-Joseph, HW1, pr. 2.3

În acest exercițiu veți deriva estimările în sens MLE pentru parametrii unei distribuții categoriale. O variabilă aleatoare X care urmează o astfel de distribuție poate lua k valori (nu doar 2, cum este cazul distribuției Bernoulli, vedeti

problema 124), și anume a_1, a_2, \dots, a_k , probabilitatea de a vedea / „observă“ un eveniment de tip j fiind θ_j , pentru $j = 1, \dots, k$.

a. Fie D o mulțime formată din n „observații“ ale lui X , independente și identic distribuite, și anume $\{d_1, \dots, d_n\}$, fiecare d_i fiind una dintre cele k valori ale lui X . Vom nota cu n_i numărul care desemnează de câte ori variabila X ia / produce valoarea a_i în mulțimea D . Exprimăți verosimilitatea lui D ca o funcție de $k - 1$ parametri (din totalul de k parametri) pentru distribuția lui X , și anume $\theta_1, \theta_2, \dots, \theta_{k-1}$.

Observație: Calculul verosimilității în funcție de $k - 1$ parametri (în loc de k) se impune datorită restricției $\sum_{i=1}^k \theta_i = 1$. Maximizările care vor fi cerute la punctul următor trebuie să țină cont de această restricție.

b. Găsiți $\hat{\theta}_j$, estimarea de verosimilitate maximă (MLE) pentru θ_j , un parametru (oarecare) dintre cei $k - 1$ de la punctul a. Pentru aceasta, veți calcula derivata parțială a funcției de verosimilitate în raport cu θ_j , apoi o veți egala cu 0 și în final veți calcula $\hat{\theta}_j$, rădăcina acestei ecuații.

Sugestie: Este recomandabil ca înainte de calculul derivatei parțiale să înlocuiți funcția de verosimilitate (de la punctul a) cu funcția de log-verosimilitate.

c. Arătați că pentru fiecare $j = 1, \dots, k$, estimatorul în sens MLE pentru parametrul θ_j este egal cu $\frac{n_j}{n}$ (așa cum era de așteptat).

Sugestie: Pornind de la expresiile deduse la punctul b pentru $\hat{\theta}_j$ pentru $j = 1, \dots, k - 1$ și notând $\hat{\theta}_k \stackrel{not.}{=} 1 - \sum_{i=1}^{k-1} \hat{\theta}_j$, arătați că sunt satisfăcute egalitățile $\hat{\theta}_j n_k = n_j \hat{\theta}_k$ pentru $j = 1, \dots, k - 1$, iar aceste egalități la rândul lor implică $\hat{\theta}_k = \frac{n_k}{n}$ și, în consecință, $\hat{\theta}_j = \frac{n_j}{n}$ pentru $j = 1, \dots, k - 1$.

Răspuns:

a. Verosimilitatea datelor se calculează astfel:

$$\begin{aligned} P(d_1, \dots, d_n | \theta) &= \prod_{j=1}^n \left(\sum_{i=1}^k \theta_i \cdot 1_{\{d_j=a_i\}} \right), \text{ unde } 1_{\{\cdot\}} \text{ este funcția-indicator} \\ &\quad \text{și } \theta \stackrel{not.}{=} (\theta_1, \dots, \theta_k) \\ \Rightarrow L(\theta) &= \prod_{i=1}^k \theta_i^{n_i}, \text{ unde } n_i \stackrel{not.}{=} \sum_{j=1}^n 1_{\{d_j=a_i\}} \\ \Rightarrow L(\theta) &= \left(1 - \sum_{i=1}^{k-1} \theta_i \right)^{n_k} \prod_{i=1}^{k-1} \theta_i^{n_i}, \text{ fiindcă } \sum_{i=1}^k \theta_i = 1. \end{aligned}$$

b. Vom proceda conform cerințelor din enunț, și anume:

$$\begin{aligned} \ln L(\theta) &= n_k \ln \left(1 - \sum_{i=1}^{k-1} \theta_i \right) + \sum_{i=1}^{k-1} n_i \ln \theta_i \\ \Rightarrow \frac{\partial \ln L(\theta)}{\partial \theta_j} &= -\frac{n_k}{1 - \sum_{i=1}^{k-1} \theta_i} + \frac{n_j}{\theta_j} \\ \frac{\partial \ln L(\theta)}{\partial \theta_j} = 0 &\Leftrightarrow -\frac{n_k}{1 - \hat{\theta}_j - \sum_{i \neq j, k} \theta_i} + \frac{n_j}{\hat{\theta}_j} = 0 \end{aligned}$$

$$\begin{aligned} &\Leftrightarrow \frac{n_j}{\hat{\theta}_j} = \frac{n_k}{1 - \hat{\theta}_j - \sum_{i \neq j, k}^{k-1} \theta_i} \\ &\Rightarrow n_j \left(1 - \sum_{i \neq j, k}^{k-1} \theta_i \right) = (n_k + n_j) \hat{\theta}_j \\ &\Rightarrow \hat{\theta}_j = \frac{n_j}{n_j + n_k} \left(1 - \sum_{i \neq j, k}^{k-1} \theta_i \right) \text{ pentru orice } j \in \{1, \dots, k-1\}. \end{aligned}$$

Observații:

1. Am presupus că $\theta_i > 0$, pentru orice $i = 1, \dots, k$ ca să existe $\ln \theta_i$ în scrierea funcției de log-verosimilitate $\ln L(\theta)$. De asemenea, din scrierea lui $\hat{\theta}_j$ de pe ultima linie de mai sus, coroborat cu condiția $\hat{\theta}_j > 0$, rezultă că este necesar să presupunem $n_j > 0$, pentru orice $j = 1, \dots, k$.
2. Se poate demonstra relativ ușor că funcția de log-verosimilitate $\ln L(\theta)$ este concavă. Într-adevăr, vă puteți convinge singuri că matricea hessiană a acestei funcții se scrie sub forma $-cI - \text{diag}\left(\frac{n_1}{\theta_1^2}, \dots, \frac{n_1}{\theta_1^2}\right)$, unde c este constantă pozitivă $\frac{n_k}{1 - \sum_{i=1}^{k-1} \theta_i}$, iar notația $\text{diag}\left(\frac{n_1}{\theta_1^2}, \dots, \frac{n_1}{\theta_1^2}\right)$ desemnează matricea pătratică având pe diagonala principală elementele $\frac{n_1}{\theta_1^2}, \dots, \frac{n_1}{\theta_1^2}$ și 0 în rest. Este evident că matricea hessiană menționată mai sus este negativ definită, ceea ce implică faptul că $\ln L(\theta)$ este funcție [strict] concavă, iar maximul ei se află egalând cu 0 derivatele ei parțiale de ordinul întâi.

c. Preluând rezultatul de la punctul b, putem scrie mai departe:

$$\hat{\theta}_j = \frac{n_j}{n_j + n_k} \left(1 - \sum_{i \neq j, k}^{k-1} \hat{\theta}_i \right) = \frac{n_j}{n_j + n_k} (\hat{\theta}_j + \hat{\theta}_k), \text{ fiindcă } \hat{\theta}_k = 1 - \sum_{i=1}^{k-1} \hat{\theta}_i.$$

Din primul și ultimul termen al acestei duble egalități rezultă

$$\begin{aligned} &\hat{\theta}_j \left(1 - \frac{n_j}{n_j + n_k} \right) = \frac{n_j}{n_j + n_k} \hat{\theta}_k \\ &\Rightarrow \hat{\theta}_j \frac{n_k}{n_j + n_k} = \frac{n_j}{n_j + n_k} \hat{\theta}_k \Rightarrow \hat{\theta}_j n_k = n_j \hat{\theta}_k. \end{aligned}$$

Prin urmare, $\hat{\theta}_j = \frac{n_j}{n_k} \hat{\theta}_k$ pentru orice $j \in \{1, \dots, k-1\}$. Mai departe, substituind aceste egalități în relația $\hat{\theta}_k = 1 - \hat{\theta}_1 - \dots - \hat{\theta}_{k-1}$, rezultă

$$\begin{aligned} &\hat{\theta}_k = 1 - \frac{n_1}{n_k} \hat{\theta}_k - \dots - \frac{n_{k-1}}{n_k} \hat{\theta}_k \\ &\Rightarrow n_k \hat{\theta}_k = n_k - (n_1 + \dots + n_{k-1}) \hat{\theta}_k \\ &\Rightarrow \hat{\theta}_k \underbrace{(n_1 + \dots + n_{k-1} + n_k)}_n = n_k \\ &\Rightarrow \hat{\theta}_k = \frac{n_k}{n} \\ &\Rightarrow \hat{\theta}_j = \frac{n_j}{n_k} \cdot \frac{n_k}{n} = \frac{n_j}{n} \in [0, 1], \text{ pentru orice } j \in \{1, \dots, k-1\}. \end{aligned}$$

Observație: Rezolvarea unei astfel de probleme de estimare a verosimilității maxime (MLE) în care intervin și restricții — precum este, în cazul nostru, restricția $\sum_{i=1}^k \theta_i = 1$ — se poate face (și de obicei, se face) folosind *metoda multiplicatorilor lui Lagrange*. În cazul nostru, aplicarea acestei metode revine la a rezolva problema de optimizare $\max_{\theta_1, \dots, \theta_k} \ln P(d_1, \dots, d_n | \theta)$, cu restricția $\sum_{i=1}^k \theta_i = 1$. Conform rezolvării de la punctul *a*, aceasta *problemă de optimizare* este echivalentă cu următoarea:

$$\max_{\theta} (n_1 \ln \theta_1 + \dots + n_k \ln \theta_k) \text{ a.i. } \sum_{i=1}^k \theta_i = 1.$$

Lagrangeanul generalizat corespunzător acestei probleme este:⁸²

$$\ell(\theta, \lambda) = n_1 \ln \theta_1 + \dots + n_k \ln \theta_k - \lambda(\theta_1 + \dots + \theta_k - 1).$$

Aplicând *condiția de staționaritate / optimalitate*,⁸³ adică derivând $\ell(\theta)$ în raport cu fiecare θ_j pentru $j = 1, \dots, k$ și egalând aceste deriveate parțiale cu 0, vom obține:

$$\frac{n_j}{\theta_j} - \lambda = 0 \Rightarrow \frac{n_j}{\theta_j} = \lambda \Rightarrow \hat{\theta}_j = \frac{n_j}{\lambda}.$$

Întrucât $\sum_{j=1}^k \hat{\theta}_j = 1$, rezultă

$$\frac{1}{\lambda} \underbrace{\sum_{j=1}^k n_j}_{n} = 1.$$

Prin urmare, $\lambda = n$, ceea ce implică

$$\hat{\theta}_j = \frac{n_j}{n} \text{ pentru } j = 1, \dots, k.$$

45.

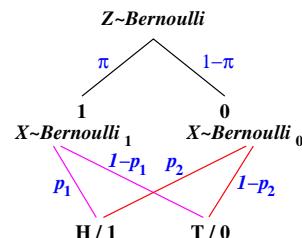
(O mixtură de distribuții Bernoulli; estimarea unui parametru prin diferite metode, folosind distribuția binomială⁸⁴)

■ □ • CMU, 2014 fall, W. Cohen, Z. Bar-Joseph, HW1, pr. 2.b

Presupunem că avem două monede: una este perfectă, având deci probabilitatea de apariție a feței cu stema $p_1 = 1/2$, iar cealaltă monedă este măsluită / imperfectă, având probabilitatea de apariție a stemei $p_2 = 1/3$.

Facem 100 de aruncări după cum urmează. De fiecare dată alegem una dintre cele două monede. Cu o probabilitate necunoscută π , alegem moneda perfectă, iar cu probabilitatea $1 - \pi$ alegem moneda măsluită. Considerăm că am obținut de 40 de ori stema din totalul celor 100 de aruncări.

a. Calculați în manieră directă / analitică estimarea de verosimilitate maximă (MLE) a parametrului π .



⁸²Vedeți problema 82.a.

⁸³Vedeți problema 83.

⁸⁴Pentru estimarea acelaiași parametru folosind algoritmul EM, vedeți ex. 5 de la capitolul *Schema algoritmică EM*.

b. În locul metodei analitice (de la punctul *a*), pentru estimarea în sensul verosimilității maxime (MLE) a parametrului π se poate folosi medoda gradientului sau metoda lui Newton. Deducreți regulile de actualizare pentru parametrul π în cazul medodei gradientului și respectiv în cazul metodei lui Newton. (Vedeți ex. 80.cd.)

Observație: Pentru o comparație între numărul de iterații [și timpul necesar] pentru a se ajunge la convergență pentru metoda gradientului și metoda lui Newton, vedeți de exemplu pr. 5 de la capitolul *Schema algoritmică EM*.

Răspuns:

a. Fie X o variabilă aleatoare având distribuția probabilistă descrisă de mixtura de distribuții Bernoulli din enunț. Vom desemna cu $Z = 1$ (respectiv $Z = 0$) moneda perfectă (respectiv, cea imperfectă), vom nota cu H fața stemă (engl., head) și apoi vom considera următoarea probabilitate:

$$\begin{aligned} q &\stackrel{\text{not.}}{=} P(X = H) \stackrel{\text{F.P.T.}}{=} P(X = H, Z = 1) + P(X = H, Z = 0) \\ &= P(X = H|Z = 1) \cdot P(Z = 1) + P(X = H|Z = 0) \cdot P(Z = 0) \\ &= \frac{1}{2} \cdot \pi + \frac{1}{3} \cdot (1 - \pi) = \frac{1}{3} + \frac{\pi}{6}, \end{aligned}$$

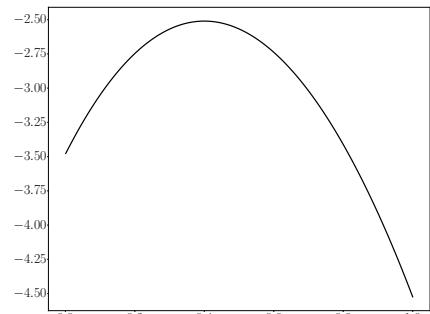
ceea ce implică $1 - q = \frac{2}{3} - \frac{\pi}{6}$.

Dacă X' este o variabilă aleatoare Bernoulli echivalentă cu mixtura dată în enunț, putem scrie $X' \sim \text{Bernoulli}(q)$. În sfârșit, întrucât nu ni s-a precizat ordinea în care au fost obținute cele 100 de rezultate / „observații“ (de tip stemă sau ban), vom lucra cu o [a treia] variabilă aleatoare, $X'' \sim \text{binomial}(r; n, q)$, unde $n = 100$. În cazul nostru, avem $r = 40$, deci funcția de verosimilitate a datelor este:

$$L(\pi) = C_{100}^{40} q^{40} (1 - q)^{60} = C_{100}^{40} \left(\frac{1}{3} + \frac{\pi}{6} \right)^{40} \cdot \left(\frac{2}{3} - \frac{\pi}{6} \right)^{60}.$$

Putem scrie acum funcția de log-verosimilitate a datelor:

$$\begin{aligned} \ell(\pi) &\stackrel{\text{def.}}{=} \ln(L(\pi)) \\ &= \ln C_{100}^{40} + 40 \ln \left(\frac{1}{3} + \frac{\pi}{6} \right) + 60 \ln \left(\frac{2}{3} - \frac{\pi}{6} \right). \end{aligned}$$



Derivata întâi și derivata a doua pentru funcția de log-verosimilitate au respectiv următoarele expresii:

$$\begin{aligned} l'(\pi) &= \frac{40}{\frac{1}{3} + \frac{\pi}{6}} \cdot \frac{1}{6} - \frac{60}{\frac{2}{3} - \frac{\pi}{6}} \cdot \frac{1}{6} = \frac{40}{2 + \pi} - \frac{60}{4 - \pi} = \frac{40 - 100\pi}{(2 + \pi)(4 - \pi)} \\ l''(\pi) &= -\frac{40}{(2 + \pi)^2} - \frac{60}{(4 - \pi)^2} < 0, \forall \pi \in [0, 1], \end{aligned}$$

ceea ce implică faptul că funcția ℓ (ca și L) are un singur punct de maxim, iar acesta este obținut egalând prima derivată cu 0: $\ell'(\pi) = 0 \Leftrightarrow \pi = \frac{40}{100} = 0.4 \in [0, 1]$. Așadar, $\pi_{MLE} = 0.4$.

b. În cazul metodei gradientului ascendent, regula de „actualizare“ este următoarea:

$$\pi_{t+1} = \pi_t + \eta \ell'(\pi_t) \stackrel{a.}{=} \pi_t + \eta \frac{40 - 100\pi}{(2 + \pi)(4 - \pi)},$$

unde η este *rata de învățare*.

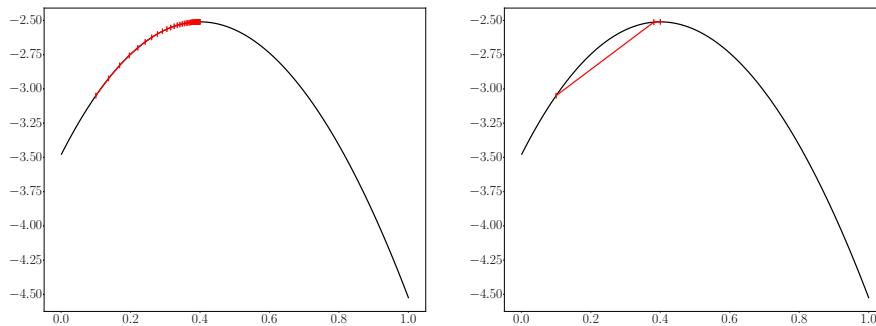
În cazul metodei lui Newton, regula de „actualizare“ este următoarea:

$$\begin{aligned} \pi_{t+1} &= \pi_t - \frac{\ell'(\pi_t)}{\ell''(\pi_t)} \\ \frac{\ell'(\pi_t)}{\ell''(\pi_t)} &\stackrel{a.}{=} -\frac{20(2 - 5\pi)}{(2 + \pi)(4 - \pi)} \cdot \frac{(2 + \pi)^2(4 - \pi)^2}{20[2(4 - \pi)^2 + 3(2 + \pi)^2]} = -\frac{(2 - 5\pi)(2 + \pi)(4 - \pi)}{5\pi^2 - 4\pi + 44}. \end{aligned}$$

Așadar,

$$\pi_{t+1} = \pi_t + \frac{(2 - 5\pi_t)(2 + \pi_t)(4 - \pi_t)}{5\pi_t^2 - 4\pi_t + 44}.$$

Grafcile următoare (primul fiind pentru metoda gradientului ascendent, al doilea pentru metoda lui Newton), ca și graficul precedent, au fost făcute de către Andi Munteanu. Ele servesc pentru a compara evoluțiile celor două metode. Pe axa Ox a fost reprezentată valoarea probabilității π , iar pe axa Oy funcția de verosimilitate. În ambele cazuri, punctul de plecare a fost $\pi^{(0)} = 0.1$. Numărul de iterații necesare pentru a ajunge la convergență (pentru $\epsilon = 10^{-4}$): pentru metoda lui Newton a fost 3, iar în cazul metodei gradientului ascendent a fost 49 dacă rata de învățare a fost setată la valoarea $\eta = 0.01$. Timpul necesar pentru executarea acestor iterații a fost de 0.00041 secunde în cazul metodei lui Newton și de 0.00054 secunde în cazul metodei gradientului.



46.

(Distribuția Poisson: estimarea parametrului în sens MLE și în sens MAP, folosind distribuția Gamma)

prelucrare de Liviu Ciortuz, după

• CMU, 2005 fall, T. Mitchel, A. Moore, HW1, pr. 5

La o fabrică de ciocolată, o persoană responsabilă cu controlul calității trebuie să estimeze numărul de fragmente de insecte care se găsesc în batoanele de ciocolată. Persoana respectivă procedează astfel:

- Mai întâi, sunt alese și analizate în mod independent 15 batoane de ciocolată. Să presupunem că numărul de fragmente de insecte găsite în fiecare din aceste batoane este înregistrat în manieră vectorială astfel: $(x_1, \dots, x_{15}) \stackrel{\text{not.}}{=} (2, 1, 0, 1, 0, 0, 1, 0, 2, 0, 1, 1, 0, 2, 1)$.
- Apoi, pentru a estima numărul de fragmente de insecte din (toate) batoanele de ciocolată, este folosită variabila aleatoare X care urmează distribuția Poisson de parametru real $\lambda > 0$.

Notă (1): Distribuția Poisson de parametru $\lambda > 0$ are funcția masă de probabilitate

$$p(x | \lambda) = \frac{1}{e^\lambda} \cdot \frac{\lambda^x}{x!}, \text{ pentru orice } x \in \mathbb{N}.$$

Prin convenție, se consideră că $0! = 1$.

Factorul $\frac{1}{e^\lambda}$, care nu depinde de x , este aşa-numita *constantă de normalizare*.⁸⁵ Se poate demonstra că media acestei distribuții este λ , iar varianța tot λ .⁸⁶

a. Calculați log-verosimilitatea datelor în funcție de λ și arătați că estimarea de verosimilitate maximă a lui λ este:

$$\lambda_{MLE} = \frac{1}{n} \sum_{i=1}^n x_i$$

unde n este numărul de instanțe considerate. Calculați apoi valoarea lui λ_{MLE} corespunzătoare datelor furnizate mai sus.

b. Presupunem că valorile parametrului λ urmează în realitate distribuția Gamma de parametri $r = 2$ și $\alpha = 8$.

Notă (2): Vă reamintim că distribuția Gamma de parametri $r > 0$ (care dă *forma* distribuției, engl., shape) și $\alpha > 0$ (numit *rata*, engl., rate) are funcția densitate de probabilitate definită pe \mathbb{R}^+ astfel:

$$p(x) \stackrel{\text{not.}}{=} \text{Gamma}(x|r, \alpha) = \frac{\alpha^r}{\Gamma(r)} x^{r-1} e^{-\alpha x}, \quad (44)$$

unde Γ desemnează funcția [Gamma a] lui Euler, care are proprietatea $\Gamma(r) = (r-1)!$ pentru orice $r \in \mathbb{N}^*$.⁸⁷ Aici *constantă de normalizare* este $\frac{\alpha^r}{\Gamma(r)}$.⁸⁸ Se poate demonstra că media distribuției Gamma este $\frac{r}{\alpha}$,⁸⁹ iar dacă $r > 1$ atunci *modul* ei (engl., mode) este $\frac{r-1}{\alpha}$. (*Modul* sau *valoarea dominantă* este, prin definiție, valoarea cu frecvența cea mai mare de apariție.)

În continuare veți răspunde la următoarele întrebări:

⁸⁵Semnificație: $\sum_{x \in \mathbb{N}} \frac{\lambda^x}{x!} = e^\lambda$.

⁸⁶A se vedea ex. 27.

⁸⁷Vedeți problema 31.b.

⁸⁸Semnificație: $\int_{x=-\infty}^{+\infty} x^{r-1} e^{-\alpha x} dx = \frac{\Gamma(r)}{\alpha^r}$.

⁸⁹A se vedea ex. 31.c.

i. Câte fragmente de insecte estimăm că vom găsi, în medie, într-un baton de ciocolată înainte de a face colectarea datelor?

Indicație: Veți lucra cu distribuția de probabilitate comună $p(x, \lambda)$, unde $x \in \mathbb{N}$ (discret), iar $\lambda \in (0, +\infty)$. Cele două argumente ale funcției p nu sunt independente, fiindcă x depinde de λ prin intermediul legii de distribuție Poisson:

$$p(x, \lambda) = p(x|\lambda) \cdot p(\lambda).$$

Media cerută în enunț se calculează însumând toate produsele de forma $x \cdot p(X = x)$, unde prin X am notat variabila aleatoare care modeleză numărul de fragmente de insecte găsite în batoanele de ciocolată. Cum $p(x)$ este distribuție marginală în raport cu distribuția comună $p(x, \lambda)$, ea se va calcula astfel:

$$p(x) = \int_0^\infty p(x, \lambda) d\lambda = \int_0^\infty p(x|\lambda)p(\lambda)d\lambda.$$

Prin urmare, va trebui să calculați (în funcție de datele din enunț) valoarea expresiei:

$$E[X] \stackrel{\text{def.}}{=} \sum_x x \cdot p(X = x) = \sum_x x \left(\int_0^\infty p(x|\lambda)p(\lambda)d\lambda \right).$$

ii. Folosind formula lui Bayes, calculați expresia funcției de densitate de probabilitate a posteriori a lui λ , notată cu $p(\lambda | x_1, \dots, x_n)$.

Indicație: Va fi util să observați că expresia $\lambda^{\sum x_i + r - 1} e^{-\lambda(n+\alpha)}$ seamănă foarte mult cu funcția densitate de probabilitate a distribuției Gamma [de mai sus] dacă se renunță la constanta de normalizare și se înlocuiește argumentul x cu λ .

iii. Calculați estimarea de probabilitate maximă a posteriori (MAP) a lui λ în raport cu datele.

c. Dacă presupunem că valorile lui λ urmează în realitate distribuția Gamma de parametri $r = 4$ și $\alpha = 16$, care este estimarea de probabilitate maximă a posteriori a lui λ ?

Răspuns:

a. Vom nota log-verosimilitatea datelor cu $\ell(\lambda)$ și o vom calcula astfel:

$$\begin{aligned} \ell(\lambda) &\stackrel{\text{def.}}{=} \ln p(x_1, \dots, x_n | \lambda) \stackrel{i.i.d.}{=} \ln \prod_{i=1}^n p(x_i | \lambda) = \sum_{i=1}^n \ln(p(x_i | \lambda)) = \sum_{i=1}^n \ln \left(e^{-\lambda} \cdot \frac{\lambda^{x_i}}{x_i!} \right) \\ &= \sum_{i=1}^n (\ln(e^{-\lambda}) + \ln(\lambda^{x_i}) - \ln(x_i!)) = \sum_{i=1}^n (-\lambda + x_i \ln \lambda - \ln(x_i!)) \\ &= -n\lambda + \left(\sum_{i=1}^n x_i \right) \cdot \ln \lambda - \sum_{i=1}^n \ln(x_i!). \end{aligned}$$

În cazul datelor din enunț, avem $\sum_{i=1}^{15} x_i = 12$ și

$$\sum_{i=1}^{15} \ln(x_i!) = 3 \cdot \ln(2!) + 6 \cdot \ln(1!) + 6 \cdot \ln(0!) = 3 \cdot \ln 2 + 6 \cdot \underbrace{\ln 1}_{=0} + 6 \cdot \underbrace{\ln 1}_{=0} = 3 \cdot \ln 2.$$

Prin urmare, log-verosimilitatea datelor este:

$$\ell(\lambda) = -15\lambda + 12 \ln \lambda - 3 \cdot \ln 2.$$

Această funcție este concavă (fapt care se poate verifica imediat cu ajutorul derivatei de ordinul al doilea), deci λ_{MLE} este rădăcina derivatei ei de ordinul întâi. În cazul general,

$$\frac{\partial l}{\partial \lambda} \left(-n\lambda + \left(\sum_i x_i \right) \ln \lambda - \sum_{i=1}^n (\ln x_i !) \right) = 0 \Leftrightarrow -n + \frac{\sum_i x_i}{\lambda} = 0 \Leftrightarrow \lambda = \frac{\sum_i x_i}{n},$$

cu condiția ca $\sum_i x_i > 0$, întrucât λ , parametrul distribuției Poisson trebuie să fie strict pozitiv. Așadar, estimarea verosimilității maximă a lui λ corespunzătoare datelor din enunț este:

$$\lambda_{MLE} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{12}{15} = \frac{4}{5} = 0.8$$

b.i. Conform *Indicației* din enunț,

$$\begin{aligned} E[X] &= \sum_x x p(x) = \sum_x x \int_{\lambda} p(x, \lambda) d\lambda = \sum_x x \left(\int_{\lambda} p(x | \lambda) p(\lambda) d\lambda \right) \\ &= \int_{\lambda} p(\lambda) \cdot \left(\sum_x x p(x | \lambda) \right) d\lambda = E_{\lambda}[E_X[X | \lambda]], \end{aligned}$$

unde $E_X[X | \lambda]$ este media variabilei X considerând parametrul λ cunoscut / fixat, iar $E_{\lambda}[\cdot]$ reprezintă media obținută atunci când parametrul λ este lăsat să varieze.

Stim că pentru orice variabilă X care urmează distribuția Poisson de parametru λ avem $E[X] = Var(X) = \lambda$ (vedeți *Nota (1)* din enunț sau problema 27).

Așadar, $E_X[X | \lambda] = \lambda$. Apoi, conform distribuției Gamma, $E_{\lambda}[\lambda] = \frac{r}{\alpha}$ (vedeți *Nota (2)*). Conform enunțului, $\lambda \sim Gamma(r = 2, \alpha = 8)$. Prin urmare,

$$E[X] = E_{\lambda}[E_X[\underbrace{X}_{\sim Poisson(\lambda)} | \lambda]] = E_{\lambda}[\underbrace{\lambda}_{\sim Gamma(r=2, \alpha=8)}] = \frac{r}{\alpha} = \frac{2}{8} = 0.25$$

b.ii. Folosind formula lui Bayes, vom scrie expresia funcției de probabilitate a posteriori astfel:

$$p(\lambda | x_1, \dots, x_n) = \frac{p(x_1, \dots, x_n | \lambda) \cdot p(\lambda)}{p(x_1, \dots, x_n)} \quad (45)$$

Vom calcula pe rând expresiile care intervin în membrul drept al acestei relații:

$$p(x_1, \dots, x_n | \lambda) \stackrel{i.i.d.}{=} \prod_{i=1}^n p(x_i | \lambda) = \prod_{i=1}^n \left(\frac{e^{-\lambda} \lambda^{x_i}}{x_i !} \right) = \frac{e^{-n\lambda} \lambda^{\sum_i x_i}}{\prod_i x_i !}$$

Așadar,

$$\begin{aligned} p(x_1, \dots, x_n | \lambda) \cdot p(\lambda) &= \frac{e^{-n\lambda} \lambda^{\sum_i x_i}}{\prod_i x_i !} \cdot \frac{\alpha^r}{\Gamma(r)} \lambda^{r-1} e^{-\alpha\lambda} \\ &= \frac{\alpha^r}{\Gamma(r) \cdot \prod_i x_i !} \cdot \lambda^{r-1 + \sum_i x_i} \cdot e^{-\lambda(\alpha+r)} \end{aligned} \quad (46)$$

Totodată,

$$\begin{aligned} p(x_1, \dots, x_n) &= \int_0^\infty p(x_1, \dots, x_n | \lambda) p(\lambda) d\lambda \\ &= \int_0^\infty \frac{\alpha^r}{\Gamma(r) \cdot \prod_i x_i!} \cdot \lambda^{r-1+\sum_i x_i} \cdot e^{-\lambda(n+\alpha)} d\lambda \\ &= \frac{\alpha^r}{\Gamma(r) \cdot \prod_i x_i!} \int_0^\infty \lambda^{r-1+\sum_i x_i} \cdot e^{-\lambda(n+\alpha)} d\lambda \end{aligned}$$

Așa cum se precizează și în *Indicația* din enunț, se observă că expresia de sub integrală arată asemănător cu funcția densitate de probabilitate $\text{Gamma}(\lambda | r + \sum_i x_i, n + \alpha)$, însă *constanta de normalizare* lipsește.⁹⁰ Prin urmare, vom continua calculele astfel:

$$\begin{aligned} p(x_1, \dots, x_n) &= \\ &= \frac{\alpha^r}{\Gamma(r) \cdot \prod_i x_i!} \left(\frac{(n + \alpha)^{r + \sum_i x_i}}{\Gamma(r + \sum_i x_i)} \right)^{-1} \underbrace{\int_0^\infty \frac{(n + \alpha)^{r + \sum_i x_i}}{\Gamma(r + \sum_i x_i)} \lambda^{r-1+\sum_i x_i} \cdot e^{-\lambda(n+\alpha)} d\lambda}_{= 1 \text{ (p.d.f.)}} \\ &= \frac{\alpha^r}{\Gamma(r) \cdot \prod_i x_i!} \cdot \frac{\Gamma(r + \sum_i x_i)}{(n + \alpha)^{r + \sum_i x_i}} \end{aligned} \quad (47)$$

Înlocuind cantitățile (46) și (47) în expresia funcției de probabilitate a posteriori (45), vom obține:

$$\begin{aligned} p(\lambda | x_1, \dots, x_n) &= \\ &= \frac{\alpha^r}{\Gamma(r) \cdot \prod_i x_i!} \lambda^{r-1+\sum_i x_i} \cdot e^{-\lambda(\alpha+n)} \left(\frac{\alpha^r}{\Gamma(r) \cdot \prod_i x_i!} \cdot \frac{\Gamma(r + \sum_i x_i)}{(n + \alpha)^{r + \sum_i x_i}} \right)^{-1} \\ &= \frac{(n + \alpha)^{r + \sum_i x_i}}{\Gamma(r + \sum_i x_i)} \lambda^{r-1+\sum_i x_i} \cdot e^{-\lambda(\alpha+n)} \end{aligned} \quad (48)$$

Așadar, funcția de probabilitate a posteriori a parametrului λ are distribuția $\text{Gamma}(\lambda | r + \sum_i x_i, n + \alpha)$.

Observație: Acest fapt înseamnă că distribuția Poisson are ca *distribuție a priori conjugată* distribuția Gamma.⁹¹

iii. Estimarea de probabilitate maximă a posteriori a lui λ în raport cu datele se calculează după formula:

$$\lambda_{MAP} = \underset{\lambda}{\operatorname{argmax}} p(\lambda | x_1, \dots, x_n)$$

De la punctul *ii* știm că parametrul λ este distribuit a posteriori conform distribuției $\text{Gamma}(\lambda | r + \sum_i x_i, n + \alpha)$, deci valoarea căutată este *modul* acestei distribuții (vedeți *Nota (2)* din enunț):

$$\lambda_{MAP} = \frac{(r + \sum_i x_i) - 1}{n + \alpha} = \frac{2 + 12 - 1}{15 + 8} = \frac{13}{23} \approx 0.56$$

⁹⁰ Altfel spus, în expresia funcției de densitate a distribuției Gamma din enunț (44) facem substituțiile $x \rightarrow \lambda$, $\alpha \rightarrow n + \alpha$ și $r \rightarrow r + \sum_i x_i$.

⁹¹ Pentru definiția noțiunii de *distribuții conjugate* vedeți enunțul problemei 43.B.

Prin urmare, în condițiile de la punctul b se observă că estimarea de probabilitate maximă a posteriori pentru parametrul λ este $\lambda_{MAP} = 0.56$, care este semnificativ mai mică decât $\lambda_{MLE} = 0.8$, estimarea de verosimilitate maximă a lui λ (vedeți punctul a).

c. Dacă $\lambda \sim \text{Gamma}(r = 4, \alpha = 16)$, atunci estimarea de probabilitate maximă a posteriori a lui λ este:

$$\lambda_{MAP} = \frac{(r + \sum_i x_i) - 1}{n + \alpha} = \frac{4 + 12 - 1}{15 + 16} = \frac{15}{31} \approx 0.48$$

Așadar, în noile condiții se obține o valoare și mai mică pentru λ_{MAP} .

47. (O distribuție uniformă (continuă) definită pe \mathbb{R} : estimarea parametrului, în sensul verosimilității maxime (MLE))

*prelucrare de Liviu Ciortuz, după
□ • ○ CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, midterm, pr. 5*

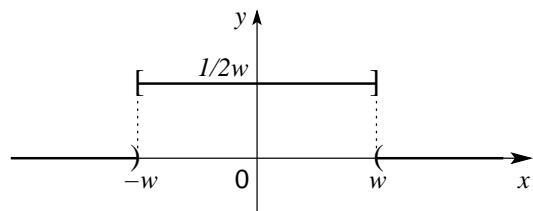
Presupunem că instanțele x_1, \dots, x_n au fost generate în mod independent de către o distribuție uniformă continuă $U(-w, w)$, având parametrul $w > 0$. Funcția densitate de probabilitate (p.d.f.) a acestei distribuții este:

$$p(x) = \begin{cases} 0, & \text{dacă } x < -w; \\ \frac{1}{2w}, & \text{dacă } -w \leq x \leq w; \\ 0, & \text{dacă } x > w. \end{cases}$$

- a. Desenați graficul funcției p și apoi demonstrați că într-adevăr p reprezintă o funcție de densitate de probabilitate.
- b. Găsiți formula de calcul pentru estimarea de verosimilitate maximă (MLE) a lui w .

Răspuns:

- a. Graficul funcției p este prezentat în figura alăturată.



Pentru ca funcția p să reprezinte o p.d.f., ea trebuie să satisfacă două cerințe: i. $p(x) \geq 0$ pentru orice x din domeniul de definiție și ii. $\int_{-\infty}^{\infty} p(x) dx = 1$.

Prima dintre aceste două cerințe este imediat satisfăcută, fiindcă $w > 0$. Pentru a verifica îndeplinirea celei de-a doua cerințe, vom calcula integrala definită:

$$\begin{aligned} \int_{-\infty}^{\infty} p(x) dx &= \int_{-\infty}^{-w} \underbrace{p(x)}_0 dx + \int_{-w}^w p(x) dx + \int_w^{\infty} \underbrace{p(x)}_0 dx \\ &= \int_{-w}^w \frac{1}{2w} dx = \frac{1}{2w} \int_{-w}^w dx \\ &= \frac{1}{2w} \cdot x \Big|_{-w}^w = \frac{1}{2w} \cdot (w - (-w)) = \frac{1}{2w} \cdot 2w = 1. \end{aligned}$$

b. Vom nota cu \hat{w} estimarea de verosimilitate maximă a lui w . Conform definiției, $\hat{w} = \arg \max_{w>0} p(x_1, \dots, x_n | w)$. Întrucât instanțele x_1, \dots, x_n au fost generate de către p în mod independent unele de altele, rezultă că $\hat{w} = \arg \max_{w>0} \prod_{i=1}^n p(x_i | w)$.

În cele ce urmează vom nota cu x_M maximul dintre $|x_1|, |x_2|, \dots, |x_n|$, deci

$$x_M = \max\{|x_1|, |x_2|, \dots, |x_n|\}.$$

În cazul în care $w < x_M$, conform definiției funcției p rezultă că $\prod_{i=1}^n p(x_i | w) = 0$. În caz contrar, adică atunci când $w \geq x_M$, conform aceleiași definiții a lui p vom avea $\prod_{i=1}^n p(x_i | w) = \frac{1}{(2w)^n}$, care este o cantitate pozitivă. Prin urmare,

$$\begin{aligned}\hat{w} &= \arg \max_{w \geq x_M} \frac{1}{(2w)^n} = \arg \max_{w \geq x_M} \ln \frac{1}{(2w)^n} = \arg \max_{w \geq x_M} -n \ln(2w) \\ &= \arg \min_{w \geq x_M} n \ln(2w) = \arg \min_{w \geq x_M} \ln w = \arg \min_{w \geq x_M} w = x_M.\end{aligned}$$

Observație: La cea de-a doua egalitate din sirul egalităților de mai sus, am utilizat faptul că funcția \ln este strict crescătoare; deci, aplicând-o unei expresii oarecare, ea conservă monotonia.

Așadar, am obținut rezultatul $\hat{w} = x_M = \max\{|x_1|, |x_2|, \dots, |x_n|\}$.

48.

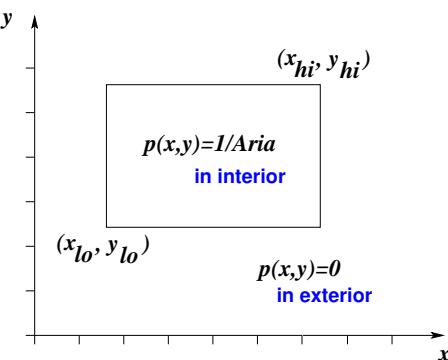
(Variabile aleatoare uniforme definite pe \mathbb{R}^2 :
funcția densitate de probabilitate;
distribuții comune, marginale, condiționale;
formula lui Bayes)

CMU, 2002 fall, Andrew Moore, midterm, pr. 2

Figura alăturată ilustrează o clasă simplă de funcții densitate de probabilitate (p.d.f.) definite peste perechi de variabile continue reale x, y .

Numim această clasă de funcții *Rectangle-PDF*. O funcție din această clasă este desemnată în mod generic prin $\text{Rectangle}(x_{lo}, y_{lo}, x_{hi}, y_{hi})$.

Așadar, parametrii clasei *Rectangle-PDF* sunt cele 4 coordonate: x_{lo}, y_{lo}, x_{hi} și y_{hi} .



Definiția funcției densitate de probabilitate $\text{Rectangle}(x_{lo}, y_{lo}, x_{hi}, y_{hi})$ este:

$$p(x, y) = \begin{cases} \frac{1}{(x_{hi} - x_{lo})(y_{hi} - y_{lo})} & \text{dacă } x_{lo} \leq x \leq x_{hi} \text{ și } y_{lo} \leq y \leq y_{hi} \\ 0 & \text{în caz contrar.} \end{cases}$$

Observație: Se poate arăta imediat că $\int_x \int_y p(x, y) dx dy = 1$.

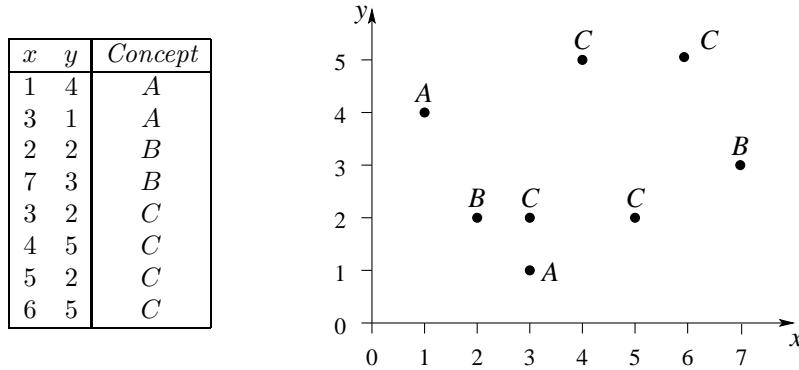
a. Pentru $\text{Rectangle}(0, 0, 1/2, 2)$, calculați: $p(x = 1/4, y = 1/4)$, $p(y = 1/4)$, $p(x = 1/4)$, $p(x = 1/4 | y = 1/4)$.

b. Fie D o mulțime de R puncte, $(x_1, y_1), (x_2, y_2), \dots, (x_R, y_R)$, selectate din $\text{Rectangle}(x_{lo}, y_{lo}, x_{hi}, y_{hi})$ în mod independent unele de altele.⁹² Pentru simplitate, vom nota $\theta = (x_{lo}, y_{lo}, x_{hi}, y_{hi})$. Prin definiție, verosimilitatea datelor D este $P(D | \theta)$.

Dacă vrem să găsim $\theta^{\text{MLE}} \stackrel{\text{not.}}{=} (x_{lo}^{\text{MLE}}, y_{lo}^{\text{MLE}}, x_{hi}^{\text{MLE}}, y_{hi}^{\text{MLE}})$, valorile parametrilor care maximizează verosimilitatea datelor D , atunci este evident că vom lua⁹³

$$x_{lo}^{\text{MLE}} = \min_k \{x_k\}, \quad y_{lo}^{\text{MLE}} = \min_k \{y_k\}, \quad x_{hi}^{\text{MLE}} = \max_k \{x_k\}, \quad y_{hi}^{\text{MLE}} = \max_k \{y_k\}.$$

În continuare, ca date concrete, vom folosi punctele $(x_i, y_i), i = 1, \dots, 6$, clasificate în trei „concepte“ A, B, C care fac parte din clasa *Rectangle-PDF*:



Vom avea, de exemplu, $p(2.5, 2.5 | A) \stackrel{\text{not.}}{=} p_{X,Y}(2.5, 2.5 | A) = 1/6$, fiindcă A este estimat în sensul verosimilității maxime (MLE) ca fiind $\text{Rectangle}(1, 1, 3, 4)$, deci cele două laturi ale dreptunghiului A sunt de mărime 2 și respectiv 3.

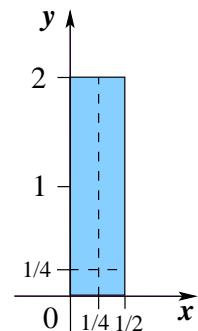
Folosind formula lui Bayes, calculați:

- $P(\text{Concept} = A | x = 1.5, y = 3)$
- $P(\text{Concept} = A | x = 2.5, y = 2.5)$
- $P(\text{Concept} = A | y = 5)$.

Răspuns:

a. În figura alăturată avem reprezentarea grafică a funcției $\text{Rectangle}(0, 0, 1/2, 2)$. Cum punctul $(1/4, 1/4)$ se găsește în interiorul dreptunghiului $(0 < \frac{1}{4} < \frac{1}{2} \text{ și } 0 < \frac{1}{4} < 2)$, rezultă că valoarea p.d.f. comune cerute este:

$$p\left(x = \frac{1}{4}, y = \frac{1}{4}\right) = \frac{1}{\left(\frac{1}{2} - 0\right)(2 - 0)} = \frac{1}{\frac{1}{2} \cdot 2} = 1$$



⁹²În mod riguros, în acest context conceptual $\text{Rectangle}(x_{lo}, y_{lo}, x_{hi}, y_{hi})$ desemnează mulțimea punctelor din plan pentru care funcția omonimă $\text{Rectangle}(x_{lo}, y_{lo}, x_{hi}, y_{hi})$ care a fost definită mai sus ia valori nenule. Această mulțime este exact dreptunghiul reprezentat în imaginea de mai sus și punctele din interiorul lui. Altfel spus, cu suprapunere de notație, funcția $\text{Rectangle}(x_{lo}, y_{lo}, x_{hi}, y_{hi})$ desemnează conceptul geometric $\text{Rectangle}(x_{lo}, y_{lo}, x_{hi}, y_{hi})$.

⁹³Pentru justificare riguroasă, vedeti cazul (mai simplu) al problemei 47.

Valorile p.d.f. marginale cerute sunt:

$$\begin{aligned}
 p\left(y = \frac{1}{4}\right) &= \int_{-\infty}^{\infty} p(x, y = \frac{1}{4}) dx \\
 &= \int_{-\infty}^0 p(x, y = \frac{1}{4}) dx + \int_0^{1/2} p(x, y = \frac{1}{4}) dx + \int_{1/2}^{\infty} p(x, y = \frac{1}{4}) dx \\
 &= \int_{-\infty}^0 0 dx + \int_0^{1/2} \frac{1}{\left(\frac{1}{2} - 0\right)(2 - 0)} dx + \int_{1/2}^{\infty} 0 dx \\
 &= 0 + \int_0^{1/2} 1 dx + 0 = x|_0^{1/2} = \frac{1}{2} - 0 = \frac{1}{2} \\
 p\left(x = \frac{1}{4}\right) &= \int_{-\infty}^{\infty} p(x = \frac{1}{4}, y) dy \\
 &= \int_{-\infty}^0 p(x = \frac{1}{4}, y) dy + \int_0^2 p(x = \frac{1}{4}, y) dy + \int_2^{\infty} p(x = \frac{1}{4}, y) dy \\
 &= \int_{-\infty}^0 0 dy + \int_0^2 \frac{1}{\left(\frac{1}{2} - 0\right)(2 - 0)} dy + \int_2^{\infty} 0 dy = 0 + \int_0^2 1 dy + 0 = y|_0^2 \\
 &= 2 - 0 = 2
 \end{aligned}$$

În sfârșit, valoarea p.d.f. condiționale cerute este:

$$p\left(x = \frac{1}{4} \mid y = \frac{1}{4}\right) = \frac{p(x = \frac{1}{4}, y = \frac{1}{4})}{p(y = \frac{1}{4})} = \frac{\frac{1}{2}}{\frac{1}{2}} = 2$$

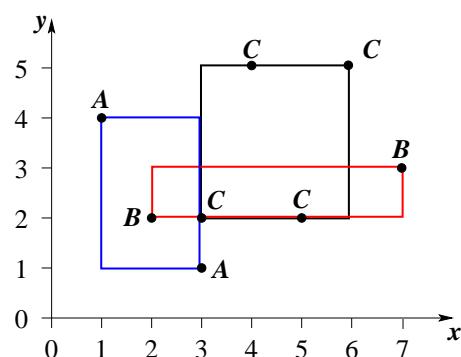
Observație: Aceeași valoare pentru $p(x = \frac{1}{4} \mid y = \frac{1}{4})$ se obține și dacă observăm că variabilele x și y sunt independente, deci $p(x = \frac{1}{4} \mid y = \frac{1}{4}) = p(x = \frac{1}{4}) = 2$.

b. Tinând cont de date (vedeți tabelul din enunț), cele trei concepte A, B, C vor fi estimate în sensul verosimilității maxime ca:

$$\text{Rectangle}_A(1, 1, 3, 4)$$

$$\text{Rectangle}_B(2, 2, 7, 3)$$

$$\text{Rectangle}_C(3, 2, 6, 5)$$



Așadar, vom avea:

$$p(x, y \mid A) = \begin{cases} \frac{1}{(3-1)(4-1)} = \frac{1}{6}, & \text{dacă } 1 \leq x \leq 3 \text{ și } 1 \leq y \leq 4 \\ 0, & \text{altfel;} \end{cases}$$

$$p(x, y \mid B) = \begin{cases} \frac{1}{(7-2)(3-2)} = \frac{1}{5}, & \text{dacă } 2 \leq x \leq 7 \text{ și } 2 \leq y \leq 3 \\ 0, & \text{altfel;} \end{cases}$$

$$p(x, y | C) = \begin{cases} \frac{1}{(6-3)(5-2)} = \frac{1}{9}, & \text{dacă } 3 \leq x \leq 6 \text{ și } 2 \leq y \leq 5 \\ 0, & \text{altfel.} \end{cases}$$

Acum putem calcula cele trei probabilități condiționate cerute, folosind formula lui Bayes. Probabilitățile a priori $P(A) = \frac{2}{8}$, $P(B) = \frac{2}{8}$ și $P(C) = \frac{4}{8}$ care intervin în aplicarea acestei formule au fost estimate din setul de date D din enunț, în sensul verosimilității maxime. (A se vedea *Observația* de mai jos.)

$$\begin{aligned} P(\text{Concept} = A | x = 1.5, y = 3) &= \frac{p(x = 1.5, y = 3 | A) \cdot P(A)}{p(1.5, 3 | A) \cdot P(A) + p(1.5, 3 | B) \cdot P(B) + p(1.5, 3 | C) \cdot P(C)} \\ &= \frac{\frac{1}{6} \cdot \frac{2}{8}}{\frac{1}{6} \cdot \frac{2}{8} + 0 \cdot \frac{2}{8} + 0 \cdot \frac{4}{8}} = 1 \end{aligned}$$

$$\begin{aligned} P(\text{Concept} = A | x = 2.5, y = 2.5) &= \frac{p(x = 2.5, y = 2.5 | A) \cdot P(A)}{p(2.5, 2.5 | A) \cdot P(A) + p(2.5, 2.5 | B) \cdot P(B) + p(2.5, 2.5 | C) \cdot P(C)} \\ &= \frac{\frac{1}{6} \cdot \frac{2}{8}}{\frac{1}{6} \cdot \frac{2}{8} + \frac{1}{5} \cdot \frac{2}{8} + 0 \cdot \frac{4}{8}} = \frac{\frac{1}{6}}{\frac{1}{6} + \frac{1}{5}} = \frac{5}{5+6} = \frac{5}{11} \end{aligned}$$

$$\begin{aligned} P(\text{Concept} = A | y = 5) &= \frac{p(y = 5 | A) \cdot P(A)}{p(y = 5)} \\ &= \frac{[\int_{-\infty}^{\infty} p(x = t, y = 5 | A) dt] \cdot P(A)}{p(y = 5)} = \frac{\int_{-\infty}^{\infty} 0 dt \cdot P(A)}{p(y = 5)} = \frac{0 \cdot P(A)}{p(y = 5)} = 0 \end{aligned}$$

Observație: Așa cum am precizat deja, valorile probabilităților a priori $P(A) = \frac{2}{8}$, $P(B) = \frac{2}{8}$ și $P(C) = \frac{4}{8}$ au fost estimate (în sensul verosimilității maxime) din tabelul de date din enunț. Pentru a înțelege mai bine de ce s-a procedat așa, ar fi fost mai explicit dacă făceam condiționarea (și) în funcție de D , setul de date din enunț:

$$\begin{aligned} P(\text{Concept} = A | x = 1.5, y = 3, D) &= \frac{p(x = 1.5, y = 3 | A, D) \cdot P(A | D)}{p(1.5, 3 | A, D) \cdot P(A | D) + p(1.5, 3 | B, D) \cdot P(B | D) + p(1.5, 3 | C, D) \cdot P(C | D)} \end{aligned}$$

Totuși, din motive legate de simplitate, am optat pentru notația folosită mai sus.

49.

(Distribuția exponentială: estimarea parametrului
în sensul verosimilității maxime (MLE))

□ • CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, final exam, pr. 1.a

Presupunem că valorile reale x_1, x_2, \dots, x_n au fost obținute prin eșantionare stochastică folosind o distribuție $p(x)$ de forma

$$p(x | \lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{pentru } x \geq 0 \\ 0 & \text{pentru } x < 0. \end{cases}$$

unde $\lambda > 0$ este un parametru necunoscut. Aceasta este *distribuția exponențială*.⁹⁴

Care dintre următoarele expresii reprezintă valoarea estimată pentru λ , obținută prin aplicarea metodei verosimilității maxime pe aceste date? (Se va presupune că în multimea dată toate elementele x_i sunt mai mari decât 1.)

$$\begin{array}{lllll} i. & \frac{\sum_{i=1}^n \ln(x_i)}{n} & ii. & \frac{\max_{i=1}^n \ln(x_i)}{n} & iii. & \frac{n}{\sum_{i=1}^n \ln(x_i)} \\ v. & \frac{\sum_{i=1}^n x_i}{n} & vi. & \frac{\max_{i=1}^n x_i}{n} & vii. & \frac{n}{\sum_{i=1}^n x_i} \\ ix. & \frac{\sum_{i=1}^n x_i^2}{n} & x. & \frac{\max_{i=1}^n x_i^2}{n} & xi. & \frac{n}{\sum_{i=1}^n x_i^2} \\ xiii. & \frac{\sum_{i=1}^n e^{x_i}}{n} & xiv. & \frac{\max_{i=1}^n e^{x_i}}{n} & xv. & \frac{n}{\sum_{i=1}^n e^{x_i}} \\ & & & & xvi. & \frac{n}{\max_{i=1}^n e^{x_i}} \end{array}$$

Răspuns:

Verosimilitatea datelor x_1, \dots, x_n în raport cu parametrul λ se calculează astfel:

$$P(x_1, \dots, x_n | \lambda) \stackrel{i.i.d.}{=} \prod_{i=1}^n p(x_i | \lambda) = \prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n \cdot \prod_{i=1}^n e^{-\lambda x_i} = \lambda^n \cdot e^{-\lambda \sum_{i=1}^n x_i}$$

Așadar, funcția de log-verosimilitate se va scrie astfel:

$$l(\lambda) \stackrel{\text{def.}}{=} \ln P(x_1, \dots, x_n | \lambda) = \ln \left(\lambda^n \cdot e^{-\lambda \sum_{i=1}^n x_i} \right) = n \ln \lambda - \lambda \sum_{i=1}^n x_i.$$

Derivata acestei funcții este

$$l'(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^n x_i, \quad \text{pentru orice } \lambda > 0.$$

Rădăcina lui l' se obține ușor:

$$\frac{n}{\lambda} - \sum_{i=1}^n x_i = 0 \Leftrightarrow \lambda = \frac{n}{\sum_{i=1}^n x_i},$$

iar această valoare este strict pozitivă fiindcă $x_i > 1$ pentru $i = 1, \dots, n$ (vedeți enunțul). Este imediat că $l'(\lambda) > 0$ pentru $\lambda < \frac{n}{\sum_{i=1}^n x_i}$ și $l'(\lambda) < 0$ pentru $\lambda > \frac{n}{\sum_{i=1}^n x_i}$. Prin urmare, funcția l are un punct de maxim. Abscisa acestui punct de maxim este rădăcina primei derivate. Așadar, valoarea de verosimilitate maximă a lui λ este

$$\lambda_{MLE} = \frac{n}{\sum_{i=1}^n x_i}$$

În concluzie, răspunsul corect este *vii.*

⁹⁴Se poate demonstra că media și varianța distribuției exponențiale sunt λ^{-1} și respectiv λ^{-2} . A se vedea ex. 31.a.

50. (Distribuția gaussiană unidimensională: estimarea mediei în sensul MLE și respectiv MAP, atunci când varianța este cunoscută)

■ □ • CMU, 2011 fall, T. Mitchell, A. Singh, HW2, pr. 1

În această problemă ne propunem să calculăm *estimatorul de verosimilitate maximă* (engl., maximum likelihood estimator, MLE) și *estimatorul de probabilitate maximă a posteriori* (engl., maximum a posteriori probability (MAP) estimator) pentru media unei distribuții gaussiene unidimensionale. Concret, presupunem că avem n instanțe, x_1, \dots, x_n generate în mod independent de către o distribuție normală cu varianță *cunoscută*, σ^2 , și media *necunoscută*, μ .

a. Calculați estimatorul MLE pentru media μ . Elaborați calculele în mod detaliat.

b. Arătați că $E[\mu_{MLE}] = \mu$.

Observație: Această egalitate înseamnă că estimatorul μ_{MLE} este *fără deplasare* sau *nedeplasat* (engl., unbiased).

c. Calculați $Var[\mu_{MLE}]$. Ce tendință au valorile acestei varianțe atunci când numărul de instanțe (n) tinde la infinit?

d. Presupunem că media urmează la rândul ei o distribuție normală de medie ν și varianță β^2 . Calculați estimatorul MAP pentru media μ . Elaborați calculele în mod detaliat.

e. Ce se întâmplă cu estimatorii MLE and MAP atunci când numărul de instanțe (n) tinde la infinit?

Răspuns:

a. Funcția de verosimilitate a datelor se exprimă astfel:

$$L(\mu) \stackrel{\text{def.}}{=} p(x_1, \dots, x_n | \mu) = \prod_{i=1}^n p(x_i | \mu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$$

Pentru a doua egalitate am ținut cont de proprietatea de independentă a datelor, iar ulterior am folosit definiția funcției de densitate a distribuției gaussiene de medie μ și varianță σ^2 . Funcția de log-verosimilitate este:

$$l(\mu) \stackrel{\text{def.}}{=} \ln p(x_1, \dots, x_n | \mu) = \sum_{i=1}^n \left(\ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{(x_i - \mu)^2}{2\sigma^2} \right)$$

În scrierea expresiei funcției de log-verosimilitate am aplicat proprietățile logaritmului. Maximul funcției de (log-)verosimilitate se află cu ajutorul derivatei:⁹⁵

$$\frac{\partial}{\partial \mu} l(\mu) = \sum_{i=1}^n \frac{x_i - \mu}{\sigma^2}$$

⁹⁵Se va observa imediat că derivata de ordinul al doilea este strict negativă, deci funcția de (log-)verosimilitate este strict concavă.

Rădăcina derivatei de ordinul întâi a funcției de (log-)verosimilitate se calculează ușor:

$$\begin{aligned}\frac{\partial}{\partial \mu} l(\mu) = 0 &\Leftrightarrow \sum_{i=1}^n \frac{x_i - \mu}{\sigma^2} = 0 \Leftrightarrow \sum_{i=1}^n (x_i - \mu) = 0 \Leftrightarrow \sum_{i=1}^n x_i = n\mu \\ &\Leftrightarrow \mu = \frac{\sum_{i=1}^n x_i}{n}\end{aligned}$$

Prin urmare, estimarea de verosimilitate maximă a parametrului μ (media) pentru distribuția gaussiană este $\mu_{MLE} = \frac{\sum_{i=1}^n x_i}{n}$.

b. Considerând instanțele x_1, \dots, x_n ca fiind generate respectiv de variabilele aleatoare X_1, \dots, X_n , toate având aceeași distribuție gaussiană (cu media μ și varianță σ^2), rezultă $\mu_{MLE} = \frac{\sum_{i=1}^n X_i}{n}$. Este natural să considerăm μ_{MLE} ca fiind o variabilă aleatoare. Aplicând proprietatea de liniaritate a mediilor variabilelor aleatoare, vom obține:

$$E[\mu_{MLE}] = E\left[\frac{X_1 + \dots + X_n}{n}\right] = \frac{E[X_1] + \dots + E[X_n]}{n} = \frac{n\mu}{n} = \mu$$

c. Tinând cont de faptul că $Var[aX] = a^2 Var[X]$ pentru orice variabilă aleatoare X și orice constantă $a \in \mathbb{R}$, precum și de faptul că varianța unei sume de variabile independente independente este suma varianțelor lor (vedeți pr. 23.c), rezultă:

$$Var[\mu_{MLE}] = Var\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n^2} \sum_{i=1}^n Var[X_i] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

În consecință, $Var[\mu_{MLE}] \rightarrow 0$ atunci când $n \rightarrow \infty$.

d. Funcția care ne interesează de data aceasta este $p(\mu|x_1, \dots, x_n)$, probabilitatea (de fapt, p.d.f.) a posteriori a parametrului μ , date fiind instanțele x_1, \dots, x_n generate cu ajutorul unei distribuții gaussiane de medie μ și varianță σ^2 .⁹⁶ Tinând cont mai întâi de teorema lui Bayes, iar apoi de expresia funcției de verosimilitate $L(\mu)$ obținută la punctul a precum și de informația din enunț conform căreia parametrul μ urmează o distribuție gaussiană de medie ν și varianță β^2 , vom putea scrie:

$$p(\mu|x_1, \dots, x_n) \stackrel{T. Bayes}{=} \frac{p(x_1, \dots, x_n|\mu) p(\mu)}{p(x_1, \dots, x_n)} \quad (49)$$

$$= \frac{\left(\prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \right) \cdot \frac{1}{\sqrt{2\pi}\beta} \exp\left(-\frac{(\mu - \nu)^2}{2\beta^2}\right)}{C} \quad (50)$$

unde probabilitatea $p(x_1, \dots, x_n)$ a fost înlocuită cu simbolul C (o constantă), fiindcă această probabilitate nu depinde de parametrul μ .

⁹⁶ De fapt, mai corect ar fi ca și aici — și peste tot la acest punct — în loc de $p(\mu|x_1, \dots, x_n)$ să scriem $p(\mu|x_1, \dots, x_n, \nu, \beta)$. Similar, în loc de $p(\mu)$ ar trebui să scriem $p(\mu|\nu, \beta)$. Pentru a simplifica redactarea calculelor, vom presupune că se subînțelege de fiecare dată „condiționarea“ în raport cu parametrii ν și β , acolo unde este necesară.

Ca și la punctul precedent, pentru studiul maximului acestei funcții este convenabil ca în prealabil să o logaritmăm:

$$\ln p(\mu|x_1, \dots, x_n) = -\sum_{i=1}^n \left(\ln \sqrt{2\pi}\sigma + \frac{(x_i - \mu)^2}{2\sigma^2} \right) - \ln \sqrt{2\pi}\beta - \frac{(\mu - \nu)^2}{2\beta^2} - \ln C$$

Maximul acestei funcții se obține tot cu ajutorul derivatei:⁹⁷

$$\begin{aligned} \frac{\partial}{\partial \mu} \ln p(\mu|x_1, \dots, x_n) &= \sum_{i=1}^n \frac{x_i - \mu}{\sigma^2} - \frac{\mu - \nu}{\beta^2} \\ \frac{\partial}{\partial \mu} \ln p(\mu|x_1, \dots, x_n) = 0 &\Leftrightarrow \sum_{i=1}^n \frac{x_i - \mu}{\sigma^2} = \frac{\mu - \nu}{\beta^2} \Leftrightarrow \mu \left(\frac{1}{\beta^2} + \frac{n}{\sigma^2} \right) = \frac{\sum_{i=1}^n x_i}{\sigma^2} + \frac{\nu}{\beta^2} \end{aligned}$$

Așadar, estimarea de probabilitate maximă a posteriori a parametrului μ (media) pentru distribuția gaussiană este $\mu_{MAP} = \frac{\sigma^2\nu + \beta^2 \sum_{i=1}^n x_i}{\sigma^2 + n\beta^2}$.

O altă soluție:

În loc să studiem probabilitatea a posteriori $p(\mu|x_1, \dots, x_n)$ cu ajutorul derivelor, vom arăta mai întâi că partea dreaptă a expresiei (50) este ea însăși o gaussiană, iar apoi vom ține cont de faptul că valoarea maximă a unei gausiene este atinsă exact în dreptul mediei.⁹⁸

$$\begin{aligned} p(\mu|x_1, \dots, x_n) &= \\ &= \frac{1}{C} \left(\prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \right) \cdot \frac{1}{\sqrt{2\pi}\beta} e^{-\frac{(\mu - \nu)^2}{2\beta^2}} \\ &= const \cdot e^{-\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2} - \frac{(\mu - \nu)^2}{2\beta^2}} \\ &= const \cdot e^{-\frac{\beta^2 \sum_{i=1}^n (x_i - \mu)^2 + \sigma^2(\mu - \nu)^2}{2\sigma^2\beta^2}} \\ &= const \cdot \exp \left(-\frac{n\beta^2 + \sigma^2}{2\sigma^2\beta^2} \mu^2 + \frac{\beta^2 \sum_{i=1}^n x_i + \nu\sigma^2}{\sigma^2\beta^2} \mu - \frac{\beta^2 \sum_{i=1}^n x_i^2 + \nu^2\sigma^2}{2\sigma^2\beta^2} \right) \\ &= const \cdot \exp \left(-\frac{\mu^2 - 2\mu \frac{\beta^2 \sum_{i=1}^n x_i + \nu\sigma^2}{n\beta^2 + \sigma^2} + \frac{\beta^2 \sum_{i=1}^n x_i^2 + \nu^2\sigma^2}{n\beta^2 + \sigma^2}}{\frac{2\sigma^2\beta^2}{n\beta^2 + \sigma^2}} \right) \\ &= const \cdot \exp \left(-\frac{\left(\mu - \frac{\beta^2 \sum_{i=1}^n x_i + \nu\sigma^2}{n\beta^2 + \sigma^2} \right)^2 - \left(\frac{\beta^2 \sum_{i=1}^n x_i + \nu\sigma^2}{n\beta^2 + \sigma^2} \right)^2 + \frac{\beta^2 \sum_{i=1}^n x_i^2 + \nu^2\sigma^2}{n\beta^2 + \sigma^2}}{2 \frac{\sigma^2\beta^2}{n\beta^2 + \sigma^2}} \right) \end{aligned}$$

⁹⁷Veți observa că derivata de ordinul al doilea este negativă, ceea ce convine.

⁹⁸Pentru a asigura o bună lizibilitate, în calculul care urmează am înlocuit — la un moment dat — expresiile de tipul e^x cu $\exp(x)$.

$$\begin{aligned}
&= \text{const} \cdot \exp \left(-\frac{\left(\mu - \frac{\beta^2 \sum_{i=1}^n x_i + \nu \sigma^2}{n \beta^2 + \sigma^2} \right)^2}{2 \frac{\sigma^2 \beta^2}{n \beta^2 + \sigma^2}} \right) \\
&\quad \cdot \exp \left(\frac{\left(\frac{\beta^2 \sum_{i=1}^n x_i + \nu \sigma^2}{n \beta^2 + \sigma^2} \right)^2 - \frac{\beta^2 \sum_{i=1}^n x_i^2 + \nu^2 \sigma^2}{n \beta^2 + \sigma^2}}{2 \frac{\sigma^2 \beta^2}{n \beta^2 + \sigma^2}} \right) \\
&= \text{const}' \cdot \exp \left(-\frac{\left(\mu - \frac{\beta^2 \sum_{i=1}^n x_i + \nu \sigma^2}{n \beta^2 + \sigma^2} \right)^2}{2 \frac{\sigma^2 \beta^2}{n \beta^2 + \sigma^2}} \right)
\end{aligned}$$

Se observă că factorul neconstant din expresia de mai sus reprezintă chiar o distribuție gaussiană de medie $\frac{\beta^2 \sum_{i=1}^n x_i + \nu \sigma^2}{n \beta^2 + \sigma^2}$ și varianță $\frac{\sigma^2 \beta^2}{n \beta^2 + \sigma^2}$. Așadar, valoarea ei maximă, ca și cea pentru expresia (50), se realizează pentru $\mu = \frac{\beta^2 \sum_{i=1}^n x_i + \nu \sigma^2}{n \beta^2 + \sigma^2} = \mu_{MAP}$.

e. Startul la acest punct îl constituie rezultatele obținute mai sus: $\mu_{MLE} = \frac{\sum_{i=1}^n x_i}{n}$ și $\mu_{MAP} = \frac{\sigma^2 \nu + \beta^2 \sum_{i=1}^n x_i}{\sigma^2 + n \beta^2}$. Punând sub o formă convenabilă ultima dintre cele două expresii, vom arăta că μ_{MAP} se poate scrie în funcție de μ_{MLE} :

$$\begin{aligned}
\mu_{MAP} &= \frac{\sigma^2 \nu + \beta^2 \sum_{i=1}^n x_i}{\sigma^2 + n \beta^2} = \frac{\sigma^2 \nu}{\sigma^2 + n \beta^2} + \frac{\beta^2 \sum_{i=1}^n x_i}{\sigma^2 + n \beta^2} \\
&= \frac{\sigma^2 \nu}{\sigma^2 + n \beta^2} + \frac{\frac{1}{n} \sum_{i=1}^n x_i}{1 + \frac{\sigma^2}{n \beta^2}} = \frac{\sigma^2 \nu}{\sigma^2 + n \beta^2} + \frac{\mu_{MLE}}{1 + \frac{\sigma^2}{n \beta^2}}
\end{aligned}$$

Evident, pentru $n \rightarrow \infty$ va rezulta că $\frac{\sigma^2 \nu}{\sigma^2 + n \beta^2} \rightarrow 0$ și $\frac{\sigma^2}{n \beta^2} \rightarrow 0$, deci $\mu_{MAP} \rightarrow \mu_{MLE}$. Așadar, pentru valori suficient de mari ale lui n , cele două estimări ale lui μ vor avea valori foarte apropiate.

51.

(Distribuția gaussiană unidimensională:
estimarea varianței în sensul MLE
[în cazul când nu se impun restricții asupra mediei μ])

□ • ○ prelucrare de Liviu Ciortuz, după
■ CMU, 2010 fall, Ziv Bar-Joseph, HW1, pr. 2.1.1-2
CMU, 2007 fall, Carlos Guestrin, HW1, pr. 3.2.1

Fie $x_1, \dots, x_n \in \mathbb{R}$ instanțe generate în mod independent și identic distribuite conform gausienei $N(x|\mu, \sigma^2)$.

Comentarii:

1. Vă reamintim că funcția densitate de probabilitate (p.d.f.) a distribuției gausiene

unidimensionale este definită de expresia:

$$p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

2. La problema 50.a am calculat estimatorul MLE al mediei μ pentru această distribuție. (Am notat acest estimator cu μ_{MLE} .)

a. Calculați estimatorul MLE al varianței σ^2 (notat cu σ_{MLE}^2), presupunând că nu se impune nicio restricție asupra lui μ .⁹⁹

b. Arătați că $E[\sigma_{MLE}^2] = \frac{n-1}{n}\sigma^2$.

c. Rezultatul de la punctul precedent ne arată că σ_{MLE}^2 este un estimator *cu deplasare* (sau, *deplasat*; engl., biased) în raport cu σ^2 .¹⁰⁰ Totuși, același rezultat ne permite să găsim ușor un estimator *nedeplasat* (engl., unbiased) pentru parametrul σ^2 . Care este acest estimator?

Răspuns:

a. Scriem mai întâi funcția de log-verosimilitate a datelor $x \stackrel{\text{not.}}{=} (x_1, \dots, x_n)$:

$$\begin{aligned} \ell(\mu, \sigma^2) &\stackrel{\text{def.}}{=} \ln P(x \mid \mu, \sigma^2) \stackrel{i.i.d.}{=} \ln \prod_{i=1}^n p(x_i) = \sum_{i=1}^n \left(-\frac{1}{2} \ln(2\pi\sigma^2) - \frac{(x_i - \mu)^2}{2\sigma^2} \right) \\ &= -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2. \end{aligned}$$

Derivata parțială a funcției ℓ în raport cu σ^2 este:¹⁰¹

$$\frac{\partial \ell(\mu, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2.$$

Tinând cont de semnele acestei derivate parțiale și egalând-o cu 0, vom obține:

$$\sigma_{MLE}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_{MLE})^2.$$

Observație: Am ținut cont de faptul că maximul funcției $\ell(\mu, \sigma^2)$ se atinge pentru $\mu = \mu_{MLE}$ și $\sigma^2 = \sigma_{MLE}^2$.¹⁰² De aceea, la rezolvarea ecuației de mai sus ($\frac{\partial \ell(\mu, \sigma^2)}{\partial \sigma^2} = 0$), am înlocuit μ cu μ_{MLE} .

⁹⁹Este util de știut de asemenea că la problema 134 se cere ca (tot pentru distribuția gaussiană unidimensională) să se estimateze varianța σ^2 în sens MLE, presupunând însă că media μ este 0.

¹⁰⁰Vă readucem aminte că estimatorul σ_{MLE} ar fi fost fără deplasare (sau, nedeplasat; engl., unbiased) dacă ar fi fost adevărată egalitatea $E[\sigma_{MLE}^2] = \sigma^2$.

¹⁰¹Puteți constata singuri că este mult mai convenabil să facem derivarea în raport cu σ^2 , decât în raport cu σ .

¹⁰²Într-adevăr, la problema 50.a s-a demonstrat de fapt că

$$l(\mu, \sigma^2) \leq l(\mu_{MLE}, \sigma^2) \text{ pentru orice } \sigma > 0 \text{ fixat,}$$

iar la exercițiul de față demonstrația de mai sus poate fi rescrisă / adaptată imediat pentru a demonstra inegalitatea

$$l(\mu_{MLE}, \sigma^2) \leq l(\mu_{MLE}, \sigma_{MLE}^2).$$

b. Vom folosi acum rezultatul de la punctul precedent pentru a calcula media estimatorului σ_{MLE}^2 . Într-adevăr, are sens să vorbim despre media lui σ_{MLE}^2 atunci când acesta este văzut ca o variabilă aleatoare (ceea ce este o consecință a faptului că instanțele x_i sunt generate în mod aleator). Făcând uz de proprietatea de liniaritate a mediei (a cărei folosire este desemnată mai jos prin simbolul (*)),¹⁰³ vom putea scrie:

$$\begin{aligned}
 E[\sigma_{MLE}^2] &= E\left[\frac{1}{n} \sum_{i=1}^n (x_i - \mu_{MLE})^2\right] \stackrel{(*)}{=} E[(x_1 - \mu_{MLE})^2] \\
 &= E\left[(x_1 - \frac{1}{n} \sum_{i=1}^n x_i)^2\right] = E\left[x_1^2 - \frac{2}{n} x_1 \sum_{i=1}^n x_i + \frac{1}{n^2} (\sum_{i=1}^n x_i)^2\right] \\
 &= E\left[x_1^2 - \frac{2}{n} x_1 \sum_{i=1}^n x_i + \frac{1}{n^2} \sum_{i=1}^n x_i^2 + \frac{2}{n^2} \sum_{i < j} x_i x_j\right] \\
 &\stackrel{(*)}{=} E[x_1^2] + \frac{1}{n^2} \sum_{i=1}^n E[x_i^2] - \frac{2}{n} \sum_{i=1}^n E[x_1 x_i] + \frac{2}{n^2} \sum_{i < j} E[x_i x_j] \\
 &= E[x_1^2] + \frac{1}{n^2} n E[x_1^2] - \frac{2}{n} E[x_1^2] - \frac{2}{n} (n-1) E[x_1 x_2] + \frac{2}{n^2} \frac{n(n-1)}{2} E[x_1 x_2] \\
 &= \frac{n-1}{n} E[x_1^2] - \frac{n-1}{n} E[x_1 x_2]. \tag{51}
 \end{aligned}$$

Ca să finalizăm calculul lui $E[\sigma_{MLE}^2]$, va trebui să calculăm separat mediile $E[x_1^2]$ și $E[x_1 x_2]$. Pentru prima dintre ele, vom folosi o cunoscută proprietate a varianței, pe care am demonstrat-o la problema 9.b:

$$\sigma^2 = Var(x_1) = E[x_1^2] - (E[x_1])^2 = E[x_1^2] - \mu^2 \Rightarrow E[x_1^2] = \sigma^2 + \mu^2.$$

Pentru calculul lui $E[x_1 x_2]$, ținând cont de faptul că x_1 și x_2 sunt independente, rezultă $Cov(x_1, x_2) = 0$, conform problemei 10. Deci vom putea scrie:

$$\begin{aligned}
 0 &= Cov(x_1, x_2) = E[(x_1 - E[x_1])(x_2 - E[x_2])] = E[(x_1 - \mu)(x_2 - \mu)] \\
 &\stackrel{(*)}{=} E[x_1 x_2] - \mu E[x_1 + x_2] + \mu^2 \stackrel{(*)}{=} E[x_1 x_2] - \mu(E[x_1] + E[x_2]) + \mu^2 \\
 &= E[x_1 x_2] - \mu(2\mu) + \mu^2 = E[x_1 x_2] - \mu^2.
 \end{aligned}$$

Prin urmare, $E[x_1 x_2] = \mu^2$.

Substituind cele două rezultate intermediare — și anume, $E[x_1^2] = \sigma^2 + \mu^2$ și $E[x_1 x_2] = \mu^2$ — în relația (51), vom obține:

$$E[\sigma_{MLE}^2] = \frac{n-1}{n}(\sigma^2 + \mu^2) - \frac{n-1}{n}\mu^2 = \frac{n-1}{n}\sigma^2.$$

c. Din relația obținută la punctul b, ținând din nou cont de liniaritatea mediei, rezultă că

$$E\left[\frac{n}{n-1}\sigma_{MLE}^2\right] = \sigma^2.$$

Prin urmare, $\frac{1}{n-1} \sum_{i=1}^n (x_i - \mu_{MLE})^2$ este un estimator nedeplasat pentru parametrul σ^2 .

¹⁰³Vedeți problema 9.a.

52.

(Distribuția Gamma:
estimarea parametrilor în sens MLE)■ □ • ○ *Liviu Ciortuz, 2017*

Fie distribuția Gamma de parametri $r > 0$ și $\beta > 0$, a cărei funcție [de] densitate de probabilitate (p.d.f.) este dată de formula următoare:¹⁰⁴

$$\text{Gamma}\left(x|r, \frac{1}{\beta}\right) = \frac{1}{\beta^r \Gamma(r)} x^{r-1} e^{-\frac{x}{\beta}}, \text{ pentru orice } x > 0.$$

În această formulă, am notat cu Γ funcția lui Euler, care constituie o generalizare pentru [funcția care definește] factorialul unui număr natural ($\Gamma(r) = (r-1)!$ pentru orice $r \in \mathbb{N}^*$).¹⁰⁵

Presupunând că instanțele $x_1, \dots, x_n \in \mathbb{R}^+$ au fost generate de către o astfel de distribuție Gamma, cu parametrii r și β fixați, calculați estimările de verosimilitate maximă (MLE) pentru cei doi parametri, r și β .

Sugestie: Scrieți mai întâi $\ell(r, \beta)$, funcția de log-verosimilitate a datelor x_1, x_2, \dots, x_n . Apoi, calculați $\hat{\beta}$, rădăcina derivatei parțiale a lui ℓ în raport cu parametrul β . În expresia pe care ați obținut-o anterior pentru funcția de log-verosimilitate ℓ , înlocuiți β cu expresia lui $\hat{\beta}$. Calculați derivata parțială a lui $\ell(r, \hat{\beta})$ în raport cu parametrul r . Egalând cu 0 expresia acestei derive parțiale veți obține o ecuație de formă

$$\ln r - \frac{\Gamma'(r)}{\Gamma(r)} = \dots . \quad (52)$$

Raportul $\frac{\Gamma'(r)}{\Gamma(r)}$, notat în general cu $\psi(r)$, se numește funcția digamma.

Observație importantă: Nu se cunosc soluții analitice pentru rezolvarea ecuațiilor de tipul (52). Însă găsirea maximului funcției $\ell(r, \hat{\beta})$ în raport cu parametrul r se poate face folosind metode de optimizare precum metoda gradientului ascendent sau metoda lui Newton. (Vedeți problema 136.)

Răspuns:

Vom scrie mai întâi expresia funcției de verosimilitate:

$$\begin{aligned} L(r, \beta) &\stackrel{\text{def.}}{=} P(x_1, \dots, x_n | r, \beta) \stackrel{i.i.d.}{=} \prod_{i=1}^n P(x_i | r, \beta) \\ &= \beta^{-rn} (\Gamma(r))^{-n} \left(\prod_{i=1}^n x_i \right)^{r-1} e^{-\frac{1}{\beta} \sum_{i=1}^n x_i} \end{aligned}$$

Logaritmând, vom obține funcția de log-verosimilitate:

$$\ell(r, \beta) \stackrel{\text{def.}}{=} \ln L(r, \beta) = -rn \ln \beta - n \ln \Gamma(r) + (r-1) \sum_{i=1}^n \ln x_i - \frac{1}{\beta} \sum_{i=1}^n x_i.$$

¹⁰⁴ Atenție! Spre deosebire de problema 46, precum și problema 31.c, unde am lucrat cu parametrii r și α , aici lucrăm cu r și $\beta = 1/\alpha$.

¹⁰⁵ Vedeți problema 31.b.

Vom calcula acum derivata parțială a funcției de log-verosimilitate $\ell(r, \beta)$ în raport cu β , iar apoi vom egala această derivată cu 0 pentru a obține rădăcina ei:

$$\begin{aligned}\frac{\partial}{\partial \beta} \ell(r, \beta) &= -\frac{rn}{\beta} + \frac{1}{\beta^2} \sum_{i=1}^n x_i = \frac{1}{\beta^2} \left[\sum_{i=1}^n x_i - rn\beta \right] \\ \frac{\partial}{\partial \beta} \ell(r, \beta) = 0 &\Leftrightarrow \hat{\beta} = \frac{1}{rn} \sum_{i=1}^n x_i > 0.\end{aligned}$$

Dacă în expresia derivatei parțiale $\frac{\partial}{\partial \beta} \ell(r, \beta)$ fixăm valoarea lui r , din analiza semnelor derivatei se constată imediat că $\hat{\beta}$ este punct de maxim și, mai mult, acest punct de maxim este unic (repetăm, pentru fiecare valoare a lui r , fixată). Pentru a calcula acum maximul funcției de log-verosimilitate $\ell(r, \beta)$ în raport cu r , mai întâi vom înlocui β cu $\hat{\beta}$ în expresia lui $\ell(r, \beta)$ și vom obține:

$$\begin{aligned}\ell(r, \hat{\beta}) &= -rn \ln \hat{\beta} - n \ln \Gamma(r) + (r-1) \sum_{i=1}^n \ln x_i - \frac{1}{\hat{\beta}} \sum_{i=1}^n x_i \\ &= rn \ln(rn) - rn \ln \sum_{i=1}^n x_i - n \ln \Gamma(r) + (r-1) \sum_{i=1}^n \ln x_i - \frac{rn}{\sum_{i=1}^n x_i} \cdot \sum_{i=1}^n x_i \\ &= rn \ln(rn) - rn \left(\ln \sum_{i=1}^n x_i + 1 \right) - n \ln \Gamma(r) + (r-1) \sum_{i=1}^n \ln x_i.\end{aligned}$$

Apoi, încercând să calculăm rădăcina derivatei parțiale a funcției $\ell(r, \hat{\beta})$ (pe care tocmai am calculat-o mai sus) în raport cu r , scriem următoarele echivalențe:

$$\begin{aligned}\frac{\partial}{\partial r} \ell(r, \hat{\beta}) = 0 &\Leftrightarrow n \ln(nr) + n - n \left(\ln \sum_{i=1}^n x_i + 1 \right) - n \cdot \frac{\Gamma'(r)}{\Gamma(r)} + \sum_{i=1}^n \ln x_i = 0 \Leftrightarrow \\ n(\ln r - \psi(r)) &= -n \ln n - \sum_{i=1}^n \ln x_i + n \ln \sum_{i=1}^n x_i \Leftrightarrow \\ \ln r - \psi(r) &= -\ln n - \frac{1}{n} \sum_{i=1}^n \ln x_i + \ln \sum_{i=1}^n x_i.\end{aligned}$$

Soluția ultimei ecuații este \hat{r} , estimarea de verosimilitate maximă a parametrului r al distribuției $Gamma\left(x|r, \frac{1}{\beta}\right)$ și, conform precizării din enunț, ea poate fi obținută nu în mod direct (ca în cazul altor distribuții probabiliste), ci prin anumite metode / programe de analiză numerică. (Vedeți problema 136.)

53. (Distribuția gaussiană multidimensională: estimarea în sens MLE¹⁰⁶ a vectorului de medii μ și a matricei de precizie $\Lambda = \Sigma^{-1}$)

*prelucrare de Liviu Ciortuz, după
■ □ • ○ CMU, 2010 fall, Aarti Singh, HW1, pr. 3.2.1*

Funcția de densitate de probabilitate (p.d.f.) pentru o distribuție gaussiană d -dimensională este definită prin expresia următoare:¹⁰⁷

$$\mathcal{N}(x | \mu, \Lambda^{-1}) \stackrel{\text{def.}}{=} \frac{\exp\left(-\frac{1}{2}(x - \mu)^\top \Lambda(x - \mu)\right)}{(2\pi)^{d/2} \sqrt{|\Lambda^{-1}|}}, \quad (53)$$

unde $\mu \in \mathbb{R}^d$ este media distribuției, iar $\Lambda \in \mathbb{R}^{d \times d}$ este așa-numita *matrice de precizie* (inversa matricei de covarianță, pe care am notat-o cu Σ).¹⁰⁸

Fie $\{x_1, x_2, \dots, x_n\} \subseteq \mathbb{R}^d$ un eşantion, adică un set de instanțe generate în mod independent și distribuite identic, conform unei gaussiene d -dimensionale.

a. Presupunem că $n \gg d$ (adică n este mult mai mare decât d).

Calculați estimările în sensul verosimilității maxime (MLE) pentru parametrii μ și Λ . (Veți nota aceste estimări cu $\hat{\mu}$ și respectiv $\hat{\Lambda}$.)

b. Particularizați rezultatele pentru cazul $d = 2$.

Sugestie: Întrucât se lucrează în \mathbb{R}^d , vă recomandăm să folosiți derive [partiale] vectoriale. Pentru acest tip de derive, variabila în raport cu care se derivează este din \mathbb{R}^d , nu din \mathbb{R} cum suntem obișnuiți. Unele din următoarele formule (preluate din documentul *Matrix Identities*, de Sam Roweis, 1999) vă pot fi de folos:

$$(2b) \quad |A^{-1}| = \frac{1}{|A|}$$

$$(2e) \quad \text{Tr}(AB) = \text{Tr}(BA);^{109}$$

mai general, $\text{Tr}(ABC\dots) = \text{Tr}(BC\dots A) = \text{Tr}(C\dots AB) = \dots$

$$(3b) \quad \frac{\partial}{\partial X} \text{Tr}(XA) = \frac{\partial}{\partial X} \text{Tr}(AX) = A^\top$$

$$(4b) \quad \frac{\partial}{\partial X} \ln |X| = (X^{-1})^\top = (X^\top)^{-1}$$

$$(5c) \quad \frac{\partial}{\partial X} a^\top X b = ab^\top$$

$$(5g) \quad \frac{\partial}{\partial X} (Xa + b)^\top C(Xa + b) = (C + C^\top)(Xa + b)a^\top$$

¹⁰⁶Pentru estimarea în sens MAP a parametrilor acestei distribuții, vedeți problema 135.

¹⁰⁷Comparativ cu expresia (28) de la problema 37, aici s-a notat Σ^{-1} cu Λ , iar $\frac{1}{|\Sigma|^{1/2}}$ a devenit $\frac{1}{\sqrt{|\Lambda^{-1}|}}$, care ulterior va fi folosit sub forma $\sqrt{|\Lambda|}$, fiindcă $|\Lambda^{-1}| = |\Lambda|^{-1}$ (vedeți regula (2b) de mai jos).

Justificarea pentru această schimbare de notație este că, din perspectiva calculelor ulterioare (găsirea optimului funcției de log-verosimilitate) este convenabil ca în p.d.f.-ul funcției \mathcal{N} să nu avem atât Σ cât și Σ^{-1} . Un pic surprinzător, calculele vor arăta (în mod indirect) că este mai convenabil ca din cele două să păstrăm Σ^{-1} *not.* Λ . Aceasta este un „artificiu“ standard în astfel de contexte. (Vedeți de exemplu problema 24 de la capitolul *Clusterizare*.)

¹⁰⁸Operatorul \top desemnează transpunerea vectorilor / matricelor. Considerăm că vectorii x și μ din \mathbb{R}^d sunt vectori-coloană.

¹⁰⁹Vedeți teorema 1.3.d din *Matrix Analysis for Statistics*, 2017, James R. Schott.

În formulele (2e) și (3b), notația $\text{Tr}(A)$ desemnează *urma* (engl., trace) unei matrice pătratice A , de dimensiune $n \times n$; ea se definește ca fiind suma elementelor de pe diagonala principală a matricei,¹¹⁰ adică $a_{11} + \dots + a_{nn}$.

Răspuns:

a. Vom scrie mai întâi log-verosimilitatea datelor x_1, \dots, x_n ca funcție de cei doi parametri, μ și Λ :

$$\begin{aligned} l(\mu, \Lambda) &\stackrel{i.i.d.}{=} \ln \prod_{i=1}^n \mathcal{N}(x_i | \mu, \Lambda^{-1}) = \sum_{i=1}^n \ln \mathcal{N}(x_i | \mu, \Lambda^{-1}) \\ &= -\frac{nd}{2} \ln(2\pi) - \frac{n}{2} \ln |\Lambda^{-1}| - \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^\top \Lambda (x_i - \mu) \\ &\stackrel{(2b)}{=} -\frac{nd}{2} \ln(2\pi) + \frac{n}{2} \ln |\Lambda| - \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^\top \Lambda (x_i - \mu). \end{aligned} \quad (54)$$

Fixând matricea de precizie Λ la o valoare oarecare (matrice pozitiv definită), se poate demonstra relativ ușor (vedeți mai jos) că matricea hessiană a acestei funcții de log-verosimilitate în raport cu variabila μ este negativ definită, deci avem de-a face cu o funcție strict concavă în raport cu μ . Pentru a identifica maximul acestei funcții, vom urmări să rezolvăm ecuația

$$\begin{aligned} \nabla_\mu l(\mu, \Lambda) = 0 &\stackrel{(5g)}{\iff} -\frac{1}{2} (\Lambda + \Lambda^\top) \sum_{i=1}^n (x_i - \mu) (-1) = 0 \iff \\ &\Lambda \sum_{i=1}^n (x_i - \mu) = 0 \iff n\Lambda\mu = \Lambda \sum_{i=1}^n x_i, \end{aligned} \quad (55)$$

unde notația $\nabla_\mu l(\mu, \Lambda)$ desemnează vectorul gradient (adică vectorul de derive parțiale) al funcției $l(\mu, \Lambda)$ în raport cu variabila vectorială μ .

Menționăm faptul că atunci când, la elaborarea șirului de echivalențe (55), am aplicat formula (5g), am luat $C = \Lambda$, $X = \mu$, $a = -1$ și $b = x_i$. În plus, am folosit faptul că Λ este matrice simetrică (adică $\Lambda = \Lambda^\top$), întrucât Λ este inversa matricei Σ , care este matricea de covarianță a distribuției gaussiene multidimensionale considerate și, conform problemei 20, Σ este matrice simetrică.

Tinând cont că Λ , ca matrice pozitiv definită, este inversabilă, din relația (55) vom obține următoarea valoare (de fapt, exact estimarea în sens MLE) pentru μ :

$$\hat{\mu} = \frac{\sum_{i=1}^n x_i}{n}, \quad (56)$$

iar această expresie coincide cu *media* [aritmetică a] *eșantionului* considerat (engl., sample mean), medie notată îndeobște cu \bar{x} , și care este constantă în raport cu matricea de precizie Λ .

Ceea ce am arătat până acum este

$$l(\mu, \Lambda) \leq l(\hat{\mu}, \Lambda) \text{ pentru orice } \mu \in \mathbb{R}^d,$$

¹¹⁰Diagonala principală a unei matrice este diagonala care pornește din colțul stânga-sus al matricei și merge până în colțul din dreapta-jos.

unde Λ este o matrice simetrică și pozitiv definită arbitrară (dar fixată). Observați că în inegalitatea aceasta, ambii membri (atât cel stâng cât și cel drept) au ca al doilea argument aceeași matrice Λ .

Urmărind să ne ocupăm acum de Λ — mai precis, să demonstrăm că are loc încă o inegalitate, $l(\hat{\mu}, \Lambda) \leq l(\hat{\mu}, \hat{\Lambda})$, pentru orice matrice pozitiv definită Λ —, vom încerca mai întâi μ cu $\hat{\mu}$ (adică, cu \bar{x}) în expresia pe care am obținut-o mai sus pentru funcția de log-verosimilitate, (54):

$$l(\hat{\mu}, \Lambda) = -\frac{nd}{2} \ln(2\pi) + \frac{n}{2} \ln |\Lambda| - \frac{1}{2} \sum_{i=1}^n (x_i - \bar{x})^\top \Lambda (x_i - \bar{x}) \quad (57)$$

$$\stackrel{(2e)}{=} -\frac{nd}{2} \ln(2\pi) + \frac{n}{2} (\ln |\Lambda| - \text{Tr}(S\Lambda)), \quad (58)$$

unde S este *matricea de covariantă la eşantionare* (engl., sample covariance matrix): $S \stackrel{\text{not.}}{=} \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^\top$. Oferim mai jos câteva detalii de calcul pentru a arăta modul în care am procedat pentru a obține expresia lui $l(\hat{\mu}, \Lambda)$.

Explicație: $(x_i - \bar{x})^\top \Lambda (x_i - \bar{x})$ este o matrice de dimensiune 1×1 , deci $(x_i - \bar{x})^\top \Lambda (x_i - \bar{x}) = \text{Tr}((x_i - \bar{x})^\top \Lambda (x_i - \bar{x}))$. Folosind regula (2e), această expresie poate fi scrisă mai departe ca $\text{Tr}((x_i - \bar{x})(x_i - \bar{x})^\top \Lambda)$.

Tinând cont de o altă regulă simplă, $\text{Tr}(A + B) = \text{Tr}(A) + \text{Tr}(B)$ (care poate fi demonstrată ușor), rezultă că

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})^\top \Lambda (x_i - \bar{x}) &= \sum_{i=1}^n \text{Tr}((x_i - \bar{x})^\top \Lambda (x_i - \bar{x})) = \sum_{i=1}^n \text{Tr}((x_i - \bar{x})(x_i - \bar{x})^\top \Lambda) \\ &= \text{Tr}\left(\sum_{i=1}^n ((x_i - \bar{x})(x_i - \bar{x})^\top \Lambda)\right) = \text{Tr}\left(\left(\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^\top\right) \Lambda\right) = \text{Tr}((nS)\Lambda) \\ &= n\text{Tr}(S\Lambda). \end{aligned}$$

Conform relației (3b), avem $\frac{\partial}{\partial \Lambda} \text{Tr}(S\Lambda) = S^\top$. Folosind în continuare relația (4b), rezultă că $\nabla_\Lambda l(\hat{\mu}, \Lambda) = \frac{n}{2}(\Lambda^\top)^{-1} - \frac{n}{2}S^\top$. Prin urmare,¹¹¹ matricea hessiană a funcției $l(\hat{\mu}, \Lambda)$ este $-\frac{n}{2}(\Lambda^{-1})^2 = -\frac{n}{2}\Sigma^2$, despre care se poate arăta imediat că este negativ definită.

Așadar, vom putea să căutăm maximul expresiei (58) rezolvând ecuația

$$\nabla_\Lambda l(\hat{\mu}, \Lambda) = 0,$$

despre care se poate arăta¹¹² că este echivalentă cu

$$\Lambda^{-1} - S = 0. \quad (\text{Deci, } \Sigma = \Lambda^{-1} = S.)$$

În general, pentru $n \gg d$ matricea S este inversabilă, așa că obținem următoarea estimare în sens MLE pentru Λ :

$$\hat{\Lambda} = S^{-1}.$$

¹¹¹Vedeți relația (61) din documentul *The Matrix Cookbook*, K.B. Petersen, M.S. Pedersen, 2012: $\frac{\partial}{\partial X} a^\top X^{-1} b = -(X^{-1})^\top ab^\top (X^{-1})^\top$.

¹¹² $\frac{n}{2}(\Lambda^\top)^{-1} - \frac{n}{2}S^\top = 0 \Leftrightarrow (\Lambda^\top)^{-1} - S^\top = 0 \Leftrightarrow (\Lambda^{-1})^\top = S^\top \Leftrightarrow \Lambda^{-1} = S$.

Observații:

1. În calculele de mai sus, am urmărit să ne asigurăm că estimările $\hat{\mu}$ și $\hat{\Lambda}$ satisfac dubla inegalitate

$$l(\mu, \Lambda) \leq l(\hat{\mu}, \Lambda) \leq l(\hat{\mu}, \hat{\Lambda})$$

pentru orice $\mu \in \mathbb{R}^d$ și orice Λ matrice simetrică și pozitiv definită. Mai mult, se poate demonstra ușor că S (deci și inversa sa) este matrice simetrică și pozitiv definită. Așadar, rezultă că $\hat{\mu}$ și $\hat{\Lambda}$ sunt estimări în sensul MLE.

2. În loc să fi folosit relația (58), adică să lucrăm cu operatorul Tr , am fi putut să calculăm derivata parțială a funcției $l(\hat{\mu}, \Lambda)$ în raport cu matricea Λ direct din relația (57):

$$\begin{aligned} \nabla_{\Lambda} l(\hat{\mu}, \Lambda) &\stackrel{(4b),(5c)}{=} \frac{n}{2} (\Lambda^{\top})^{-1} - \frac{1}{2} \sum_{i=1}^n (x_i - \hat{\mu})(x_i - \hat{\mu})^{\top} \\ &= \frac{n}{2} \Lambda^{-1} - \frac{1}{2} \sum_{i=1}^n (x_i - \hat{\mu})(x_i - \hat{\mu})^{\top} = \frac{n}{2} \Lambda^{-1} - \frac{n}{2} S. \end{aligned} \quad (59)$$

Așadar,

$$\nabla_{\Lambda} l(\hat{\mu}, \Lambda) = 0 \Leftrightarrow \hat{\Lambda}^{-1} \stackrel{\text{not.}}{=} \hat{\Sigma} = S \stackrel{\text{not.}}{=} \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})(x_i - \hat{\mu})^{\top}. \quad (60)$$

Observație importantă:¹¹³

În enunț a fost inclusă condiția $n \gg d$. În cazul în care n (numărul de instanțe) nu este semnificativ mai mare decât d (numărul de dimensiuni / attribute), estimările în sens MLE pentru parametrii μ și Σ pot fi nesatisfăcătoare (engl., poor). Dacă $d \gg n$ și încercăm să estimăm media și matricea de covarianță folosind relațiile (56) și respectiv (60), se poate să obținem $|\Sigma| = 0$,¹¹⁴ întrucât cele n instanțe generează (engl., *span*)¹¹⁵ doar un subspațiu de dimensiune redusă al lui \mathbb{R}^d . În acest caz, Σ^{-1} nu există, deci nu vom putea scrie funcția de densitate a distribuției gaussiene multidimensionale (53). Totuși, în unele cazuri particulare este posibil ca matricea Σ să fie nesingulară chiar pentru $n \geq 2$, de exemplu atunci când Σ este matrice diagonală sau, și mai restrictiv, când $\Sigma = \sigma^2 I$. Dezavantajul în cazul în care se lucrează cu Σ matrice diagonală este că se presupune în mod implicit că attributele sunt independente (cf. ex. 34).¹¹⁶

- b. Pentru cazul distribuției gaussiene bidimensionale (adică, pentru $d = 2$), pur și simplu ca să ne re-familiarizăm cu matricea de covarianță $\Sigma \stackrel{\text{not.}}{=} \Lambda^{-1}$, mai întâi ne aducem aminte că aceasta poate fi scrisă astfel:¹¹⁷

¹¹³Cf. Andrew Ng, *Factor Analysis* (ML Lecture Notes, Part X), pag. 1-3.

¹¹⁴Într-un astfel de caz spunem că Σ este matrice singulară.

¹¹⁵Cf. documentului *Linear Algebra Review and Reference* de Zico Kolter (și Chuong Do), 2012, pag. 12, *spațiu liniar generat* de vectorii $\{x_1, x_2, \dots, x_n\}$ se definește ca fiind mulțimea formată din toți acei vectori care pot fi scriși ca o combinație liniară de x_1, x_2, \dots, x_n :

$$\text{span}(x_1, x_2, \dots, x_n) = \left\{ v \mid v = \alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_n x_n, \text{ cu } \alpha_i \in \mathbb{R} \right\}.$$

¹¹⁶Din această cauză, pentru a putea identifica eventualele dependențe dintre attribute, atunci când $d \gg n$ se recomandă să se aplique metoda *analizei prin factori* (engl., Factor Analysis), care face estimarea parametrilor folosind *algoritmul EM*. Vedeti documentul *Factor Analysis* de Andrew Ng (ML Lecture Notes, Part X), pag. 4-5.

¹¹⁷Vedeți problema 38.

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

cu $\sigma_1 > 0$, $\sigma_2 > 0$ și $\rho \in (-1, 1)$.

Urmează apoi că matricea de precizie Λ are forma următoare:

$$\Lambda \stackrel{\text{not.}}{=} \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)} \begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix} = \frac{1}{(1-\rho^2)} \begin{pmatrix} \frac{1}{\sigma_1^2} & -\frac{\rho}{\sigma_1\sigma_2} \\ -\frac{\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{pmatrix}.$$

Rezultatele pe care le-am obținut la punctul a se particularizează în felul următor pentru cazul distribuției gaussiene bidimensionale ($d = 2$):

$$\begin{aligned} \hat{\mu} &= \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \begin{pmatrix} \sum_{i=1}^n x_{i,1} \\ \sum_{i=1}^n x_{i,2} \end{pmatrix} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \end{pmatrix}, \text{ unde} \\ x_i &\stackrel{\text{not.}}{=} \begin{pmatrix} x_{i,1} \\ x_{i,2} \end{pmatrix} \text{ pentru } i = 1, \dots, n, \\ \bar{x}_1 &\stackrel{\text{not.}}{=} \frac{1}{n} \sum_{i=1}^n x_{i,1} \text{ și } \bar{x}_2 \stackrel{\text{not.}}{=} \frac{1}{n} \sum_{i=1}^n x_{i,2}; \\ \hat{\Sigma} &\stackrel{\text{not.}}{=} \hat{\Lambda}^{-1} = \frac{1}{n} \sum_{i=1}^n \left(\begin{pmatrix} x_{i,1} - \bar{x}_1 \\ x_{i,2} - \bar{x}_2 \end{pmatrix} (x_{i,1} - \bar{x}_1, x_{i,2} - \bar{x}_2) \right) \\ &= \frac{1}{n} \sum_{i=1}^n \begin{pmatrix} (x_{i,1} - \bar{x}_1)^2 & (x_{i,1} - \bar{x}_1)(x_{i,2} - \bar{x}_2) \\ (x_{i,1} - \bar{x}_1)(x_{i,2} - \bar{x}_2) & (x_{i,2} - \bar{x}_2)^2 \end{pmatrix} \\ &= \frac{1}{n} \begin{pmatrix} \sum_{i=1}^n (x_{i,1} - \bar{x}_1)^2 & \sum_{i=1}^n (x_{i,1} - \bar{x}_1)(x_{i,2} - \bar{x}_2) \\ \sum_{i=1}^n (x_{i,1} - \bar{x}_1)(x_{i,2} - \bar{x}_2) & \sum_{i=1}^n (x_{i,2} - \bar{x}_2)^2 \end{pmatrix}. \end{aligned}$$

Aceasta este exact matricea de covariantă la eșantionare,

$$\begin{pmatrix} Var(X_1) & Cov(X_1, X_2) \\ Cov(X_1, X_2) & Var(X_2) \end{pmatrix},$$

unde am presupus că variabilele X_1 și X_2 reprezintă cele două componente pentru X , variabila gaussiană bidimensională care a generat instanțele x_1, \dots, x_n .

54.

(Estimatori MLE: existența și unicitatea)

$\square \bullet \circ$ CMU, 2010 fall, Aarti Singh, HW1, pr. 3.3

În această problemă vom arăta că estimările de verosimilitate maximă (MLE) nu există în mod neapărat. Își, chiar atunci când există, se poate ca ele să nu fie unice.

a. Puneți în evidență un caz în care nu există estimarea / estimările de verosimilitate maximă (MLE). Vă cerem să specificați familia de distribuții pe care ați ales-o, precum și tipul de eșantioane (engl., samples) pentru care MLE nu este [bine] definită.

- b. Dați un exemplu de caz în care MLE există, însă nu în mod unic. Specificați familia de distribuții pe care ati ales-o, precum și tipul de eșantioane pentru care pot exista estimări multiple de verosimilitate maximă.
- c. Identificând cele două exemple descrise mai sus, sperăm că v-ați format o anumită intuiție despre proprietățile funcției de [log]-verosimilitate care sunt cruciale pentru existența și unicitatea MLE. Formulați aceste proprietăți.

Răspuns:

- a. La rezolvarea problemei 46.a am demonstrat pentru distribuția Poisson de parametru $\lambda > 0$ că estimarea lui în sensul MLE este $\lambda_{MLE} = \frac{1}{n} \sum_{i=1}^n x_i$, pentru orice set de instanțe $x_1, \dots, x_n \geq 0$. Evident, această estimare MLE există doar dacă $\sum_{i=1}^n x_i > 0$. Atunci când $\sum_{i=1}^n x_i = 0$, funcția de log-verosimilitate

$$l(\lambda) \stackrel{\text{def.}}{=} \ln P(x_1, \dots, x_n | \lambda) = -n\lambda + \left(\sum_{i=1}^n x_i \right) \cdot \ln \lambda - \sum_{i=1}^n \ln(x_i!)$$

nu-și atinge maximul în interiorul domeniului de existență pentru λ , și anume $(0, +\infty)$.

- b. Considerăm familia mixturilor de două distribuții gaussiene unidimensionale:

$$f(x|\theta, \mu_1, \sigma_1, \mu_2, \sigma_2) = \theta \mathcal{N}(x|\mu_1, \sigma_1) + (1-\theta) \mathcal{N}(x|\mu_2, \sigma_2),$$

unde $0 < \theta < 1$, $0 < \sigma_1$, $0 < \sigma_2$, iar μ_1, μ_2 sunt numere reale oarecare, distincte. Este imediat că pentru orice estimări bine-definite $\hat{\theta}, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1$ și $\hat{\sigma}_2$ obținute pe un set oarecare de instanțe x_1, \dots, x_n , următorul set de estimări

$$\hat{\theta}' = 1 - \hat{\theta}, \quad \hat{\mu}'_1 = \hat{\mu}_2, \quad \hat{\mu}'_2 = \hat{\mu}_1, \quad \hat{\sigma}'_1 = \hat{\sigma}_2, \quad \hat{\sigma}'_2 = \hat{\sigma}_1$$

vor produce întotdeauna aceeași verosimilitate ca și estimările inițiale. Și totuși $(\hat{\theta}', \hat{\mu}'_1, \hat{\mu}'_2, \hat{\sigma}'_1, \hat{\sigma}'_2) \neq (\hat{\theta}, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2)$. Așadar, estimările în sensul MLE nu sunt în mod necesar unice.

- c. În exemplul de la punctul a, domeniul de existență pentru parametrul distribuției pe care am ales-o este, ca și în multe alte cazuri, un interval deschis. Însă valoarea acestui parametru pentru care se atinge maximul funcției de log-verosimilitate este situată la „marginea“ [inferioară a] intervalului, nu în interiorul intervalului respectiv. Această problemă apare adeseori atunci când mărimea eșantionului (adică numărul instanțelor x_i) este mică în comparație cu numărul parametrilor care trebuie estimati.

În exemplul de la punctul b, funcția de log-verosimilitate nu este concavă — pentru exemplificare, vedeti problema 17 de la capitolul de *Clusterizare*, în special graficele care reprezintă curbele de izocontur ale funcției de log-verosimilitate ale mixturilor de gaussiene, la pag. 874 și pag. 875 —, ceea ce implică faptul că pot exista mai multe puncte de maxim local.

0.1.5 Elemente de teoria informației¹¹⁸

55.

(Entropie, entropie comună, entropie condițională, câștig de informație: definiții și proprietăți imediate)

■ • Livi Ciortuz, pornind de la CMU, 2005 fall, T. Mitchell, A. Moore, HW1, pr. 2

Fie X și Y variabile aleatoare discrete. Dăm pe scurt următoarele *definiții*:

- **Entropia variabilei X :**

$$H(X) \stackrel{\text{def.}}{=} -\sum_i P(X = x_i) \log P(X = x_i) \stackrel{\text{not.}}{=} E_X[-\log P(X)].$$

Prin convenție, dacă $p(x) = 0$ atunci vom considera $p(x) \log p(x) = 0$.

- **Entropia condițională specifică a variabilei Y în raport cu valoarea x_k a variabilei X :**

$$\begin{aligned} H(Y | X = x_k) &\stackrel{\text{def.}}{=} -\sum_j P(Y = y_j | X = x_k) \log P(Y = y_j | X = x_k) \\ &\stackrel{\text{not.}}{=} E_{Y|X=x_k}[-\log P(Y | X = x_k)]. \end{aligned}$$

- **Entropia condițională medie a variabilei Y în raport cu variabila X :**

$$H(Y | X) \stackrel{\text{def.}}{=} \sum_k P(X = x_k) H(Y | X = x_k) \stackrel{\text{not.}}{=} E_X[H(Y | X)].$$

- **Entropia comună a variabilelor X și Y :**

$$\begin{aligned} H(X, Y) &\stackrel{\text{def.}}{=} -\sum_i \sum_j P(X = x_i, Y = y_j) \log P(X = x_i, Y = y_j) \\ &\stackrel{\text{not.}}{=} E_{X,Y}[-\log P(X, Y)]. \end{aligned}$$

- **Câștigul de informație** al variabilei X în raport cu variabila Y (sau invers), numit de asemenea *informația mutuală* (engl., mutual information) a variabilelor X și Y :

$$IG(X, Y) \stackrel{\text{not.}}{=} MI(X, Y) \stackrel{\text{def.}}{=} H(X) - H(X | Y) = H(Y) - H(Y | X).$$

(Observație: ultima egalitate de mai sus are loc datorită rezultatului de la punctul c de mai jos.)

Arătați că:

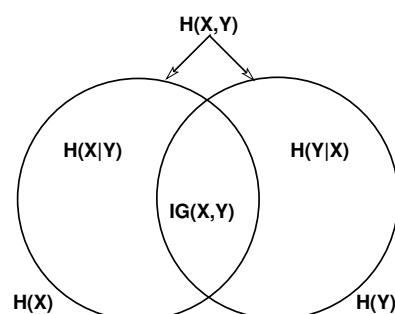
a. $H(X) \geq 0$. În particular, $H(X) = 0$ dacă și numai dacă variabila X este constantă.

b. $H(Y | X) = -\sum_i \sum_j P(X = x_i, Y = y_j) \log P(Y = y_j | X = x_i)$.

c. $H(X, Y) = H(X) + H(Y | X) = H(Y) + H(X | Y)$.

Mai general: $H(X_1, \dots, X_n) = H(X_1) + H(X_2 | X_1) + \dots + H(X_n | X_1, \dots, X_{n-1})$ (regula de înlățuire a entropiilor).

Observație: Relația precedentă, precum și relația de definiție pentru câștigul de informație sunt ilustrate în figura alăturată.



¹¹⁸Observație importantă: În toate problemele care urmează, referitor la entropie / teoria informației se va considera în mod implicit că notația ‘log’ desemnează logaritmul în baza 2. De asemenea, prin convenție, se va considera $p \cdot \log p = 0$ pentru $p = 0$.

Răspuns:

a. Este ușor să arătăm că $H(X) = -\sum_i P(X = x_i) \log P(X = x_i) \geq 0$.

Stim că $\log x \leq 0$ pentru $\forall x \leq 1$ și $\log x \geq 0$ pentru $\forall x \geq 1$. De asemenea stim că $P(X = x_i) \in [0, 1]$ (fiind o probabilitate). Așadar,

$$H(X) = -\sum_i P(X = x_i) \log P(X = x_i) = \sum_i \underbrace{P(X = x_i)}_{\geq 0} \underbrace{\log \frac{1}{P(X = x_i)}}_{\geq 0} \geq 0.$$

Pentru a arăta că $H(X) = 0$ dacă și numai dacă X este constantă vom demonstra că ambele implicații au loc:

„ \Rightarrow “ Presupunem că $H(X) = 0$, adică $\sum_i P(X = x_i) \log \frac{1}{P(X = x_i)} = 0$. Datorită faptului că fiecare termen din această sumă este mai mare sau egal cu 0, rezultă că $H(X) = 0$ doar dacă pentru $\forall i$, $P(X = x_i) = 0$ sau $\log \frac{1}{P(X = x_i)} = 0$, adică dacă pentru $\forall i$, $P(X = x_i) = 0$ sau $P(X = x_i) = 1$. Cum însă $\sum_i P(X = x_i) = 1$ rezultă că există o singură valoare x_1 pentru X astfel încât $P(X = x_1) = 1$, iar $P(X = x) = 0$ pentru orice $x \neq x_1$. Altfel spus, variabila aleatoare discretă X este constantă.¹¹⁹

„ \Leftarrow “ Presupunem că variabila X este constantă, ceea ce înseamnă că X ia o singură valoare x_1 , cu probabilitatea $P(X = x_1) = 1$. Prin urmare, $H(X) = -1 \cdot \log 1 = 0$.

b. Pentru a demonstra egalitatea cerută vom porni de la definiția lui $H(Y | X)$ și apoi vom efectua câteva transformări elementare:

$$\begin{aligned} H(Y | X) &= \sum_i P(X = x_i) H(Y | X = x_i) \\ &= \sum_i P(X = x_i) \left[-\sum_j P(Y = y_j | X = x_i) \log P(Y = y_j | X = x_i) \right] \\ &= -\sum_i \sum_j \underbrace{P(X = x_i) P(Y = y_j | X = x_i)}_{=P(X=x_i, Y=y_j)} \log P(Y = y_j | X = x_i) \\ &= -\sum_i \sum_j P(X = x_i, Y = y_j) \log P(Y = y_j | X = x_i). \end{aligned}$$

c. În primul rând, trebuie să demonstrează că

$$H(X, Y) = H(X) + H(Y | X) = H(Y) + H(X | Y).$$

Din definiția entropiei comune stim că $H(X, Y) = -\sum_i \sum_j P(X = x_i, Y = y_j) \log P(X = x_i, Y = y_j)$. Vom aplica mai întâi regula de înmulțire, $P(X, Y) = P(X) \cdot P(Y | X)$, după care vom transforma logaritmul produsului în sumă de logaritmi. Pentru claritatea demonstrației vom nota prescurtat $p(x_i) = P(X = x_i)$, $p(x_i, y_j) = P(X = x_i, Y = y_j)$ etc.

$$H(X, Y) = -\sum_i \sum_j p(x_i, y_j) \log p(x_i, y_j)$$

¹¹⁹Mai corect spus, X este constantă pe tot domeniul de definiție, eventual cu excepția unei multimi de probabilitate 0.

$$\begin{aligned}
&= - \sum_i \sum_j p(x_i) \cdot p(y_j | x_i) \log[p(x_i) \cdot p(y_j | x_i)] \\
&= - \sum_i \sum_j p(x_i) \cdot p(y_j | x_i) [\log p(x_i) + \log p(y_j | x_i)] \\
&= - \sum_i \sum_j p(x_i) \cdot p(y_j | x_i) \log p(x_i) - \sum_i \sum_j p(x_i) \cdot p(y_j | x_i) \log p(y_j | x_i) \\
&= - \sum_i p(x_i) \log p(x_i) \cdot \underbrace{\sum_j p(y_j | x_i)}_{=1} - \sum_i p(x_i) \sum_j p(y_j | x_i) \log p(y_j | x_i) \\
&= H(X) + \sum_i p(x_i) H(Y | X = x_i) = H(X) + H(Y | X).
\end{aligned}$$

Egalitatea $\sum_j p(y_j | x_i) = 1$ se justifică ușor ținând cont de proprietatea de aditivitate numărabilă din definiția funcției / distribuției de probabilitate.

Pentru a demonstra egalitatea $H(X, Y) = H(Y) + H(X | Y)$, se procedează analog, înlocuind $p(x_i, y_j)$ nu cu $p(x_i) \cdot p(y_j | x_i)$, ci cu $p(y_i) \cdot p(x_j | y_i)$.

Pentru cazul general

$$H(X_1, \dots, X_n) = H(X_1) + H(X_2 | X_1) + \dots + H(X_n | X_1, \dots, X_{n-1}),$$

vom folosi regula de înlățuire de la variabile aleatoare

$$P(X_1, \dots, X_n) = P(X_1) \cdot P(X_2 | X_1) \cdot P(X_3 | X_1, X_2) \cdot \dots \cdot P(X_n | X_1, \dots, X_{n-1}),$$

precum și scrierea entropiei sub formă de medie, $H(X) = E \left[\log \frac{1}{P(X)} \right]$:

$$\begin{aligned}
H(X_1, \dots, X_n) &= E \left[\log \frac{1}{p(x_1, \dots, x_n)} \right] \\
&= - E_{p(x_1, \dots, x_n)} \left[\log \underbrace{\frac{p(x_1, \dots, x_n)}{p(x_1) \cdot p(x_2 | x_1) \cdots p(x_n | x_1, \dots, x_{n-1})}} \right] \\
&= - E_{p(x_1, \dots, x_n)} [\log p(x_1) + \log p(x_2 | x_1) + \dots + \log p(x_n | x_1, \dots, x_{n-1})] \\
&= - E_{p(x_1)} [\log p(x_1)] - E_{p(x_1, x_2)} [\log p(x_2 | x_1)] - \dots \\
&\quad - E_{p(x_1, \dots, x_n)} [\log p(x_n | x_1, \dots, x_{n-1})] \\
&\stackrel{(b)}{=} H(X_1) + H(X_2 | X_1) + \dots + H(X_n | X_1, \dots, X_{n-1}).
\end{aligned}$$

La penultima egalitate am ținut cont de definiția distribuției marginale pornind de la distribuția comună, iar la ultima egalitate am folosit rezultatul de la punctul b.

56. (Entropie, entropie condițională specifică, câștig de informație: exemplificare)

■ □ • ○ CMU, 2012 spring, Roni Rosenfeld, HW2, pr. 2

Problema aceasta se referă la aruncarea a două zaruri perfecte, cu 6 fețe.

- a. Calculează distribuția probabilistă a sumei numerelor de pe cele două fețe care au fost obținute / „observate“ în urma aruncării zarurilor.

În continuare, suma aceasta va fi asimilată cu o variabilă aleatoare, notată cu S .

b. Cantitatea de *informație* obținută (sau: *surpriza pe care o resimțim*) la „observarea“ producerii valorii x a unei variabile aleatoare X oarecare este prin *definiție*

$$\text{Information}(P(X = x)) = \text{Surprise}(P(X = x)) = \log_2 \frac{1}{P(X = x)} = -\log_2 P(X = x).$$

Această cantitate este exprimată (numeric) în *biți de informație*.

Cât de surprins vei fi atunci când vei „observa“ $S = 2$, respectiv $S = 11$, $S = 5$ și $S = 7$? (Vei exprima de fiecare dată rezultatul în biți. Puteți folosi $\log_2 3 = 1.584962501$.)

c. Calculează entropia variabilei S .

d. Să presupunem acum că vei arunca aceste două zaruri pe rând, iar la aruncarea primului zar se obține numărul 4. Cât este entropia lui S în urma acestei „observații“? S-a pierdut, ori s-a câștigat informație în acest proces? Calculează cât de multă informație (exprimată în biți) s-a pierdut ori s-a câștigat.

Răspuns:

a. Redăm distribuția lui S (ușor de calculat) în următorul tabel:

S	2	3	4	5	6	7	8	9	10	11	12
$P(S)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

b. Conform definiției date, vom avea:

$$\begin{aligned} \text{Information}(S = 2) &= -\log_2(1/36) = \log_2 36 = 2 \log_2 6 = 2(1 + \log_2 3) \\ &= 5.169925001 \text{ biți} \end{aligned}$$

$$\text{Information}(S = 11) = -\log_2 2/36 = \log_2 18 = 1 + 2 \log_2 3 = 4.169925001 \text{ biți}$$

$$\text{Information}(S = 5) = -\log_2 4/36 = \log_2 9 = 2 \log_2 3 = 3.169925001 \text{ biți}$$

$$\text{Information}(S = 7) = -\log_2 6/36 = \log_2 6 = 1 + \log_2 3 = 2.584962501 \text{ biți}$$

c. Conform definiției pentru entropie (vedeți problema 55), $H(S)$ este media ponderată (cu ajutorul probabilităților) a „surprizelor“ / cantităților de informație produse la „observarea“ tuturor valorilor variabilei S . Făcând calculele, vom obține:

$$\begin{aligned} H(S) &= -\sum_{i=1}^n p_i \log_2 p_i \\ &= -\left(2 \cdot \frac{1}{36} \log_2 \frac{1}{36} + 2 \cdot \frac{2}{36} \log_2 \frac{2}{36} + 2 \cdot \frac{3}{36} \log_2 \frac{3}{36} + 2 \cdot \frac{4}{36} \log_2 \frac{4}{36} + \right. \\ &\quad \left. 2 \cdot \frac{5}{36} \log_2 \frac{5}{36} + \frac{6}{36} \log_2 \frac{6}{36}\right) \\ &= \frac{1}{36} \left(2 \log_2 36 + 4 \log_2 18 + 6 \log_2 12 + 8 \log_2 9 + 10 \log_2 \frac{36}{5} + 6 \log_2 6\right) \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{36} \left(2 \log_2 6^2 + 4 \log_2 6 \cdot 3 + 6 \log_2 6 \cdot 2 + 8 \log_2 3^2 + 10 \log_2 \frac{6^2}{5} + 6 \log_2 6 \right) \\
&= \frac{1}{36} (40 \log_2 6 + 20 \log_2 3 + 6 - 10 \log_2 5) \\
&= \frac{1}{36} (60 \log_2 3 + 46 - 10 \log_2 5) = 3.274401919 \text{ biți.}
\end{aligned}$$

d. Distribuția variabilei S condiționată de observarea feței 4 la prima aruncare este:

S	2	3	4	5	6	7	8	9	10	11	12
$P(S ...)$	0	0	0	1/6	1/6	1/6	1/6	1/6	1/6	0	0

În consecință, folosind definiția entropiei condiționale specifice (vedeți de asemenea problema 55), vom avea:

$$H(S|First-die-shows-4) = -6 \cdot \frac{1}{6} \log_2 \frac{1}{6} = \log_2 6 = 2.58 \text{ biți,}$$

ceea ce înseamnă că se obține următorul câștig de informație:

$$IG(S; First-die-shows-4) = H(S) - H(S|First-die-shows-4) = 3.27 - 2.58 = 0.69 \text{ biți.}$$

Altfel spus, atunci când ni se comunică faptul că la aruncarea celor două zaruri primul dintre ele produce fața 4, această informație va reduce ulterior entropia variabilei S (sau, am putea spune, „surpriza“ medie provocată de valorile ei) cu 0.69 biți.

57.

(Probabilități marginale, entropii, entropii condiționale medii)

■ □ • CMU, 2012 spring, Roni Rosenfeld, HW2, pr. 3

Un doctor trebuie să pună un diagnostic unui pacient care are simptome de răceală (C , de la engl. cold). Factorul principal pe care doctorul îl ia în considerare pentru a elabora diagnosticul este timpul, adică starea vremii de afară (T). Variabila aleatoare C ia două valori, *yes* și *no*, iar variabila aleatoare T ia 3 valori: *sunny* (însorit), *rainy* (ploios) și *snowy* (foarte rece, să zicem). Distribuția comună a celor două variabile este dată în tabelul următor:

	$T = sunny$	$T = rainy$	$T = snowy$
$C = no$	0.30	0.20	0.10
$C = yes$	0.05	0.15	0.20

a. Calculați probabilitățile marginale $P(C)$ și $P(T)$.

Sugestie: Folosiți formula $P(X = x) = \sum_y P(X = x; Y = y)$. De exemplu,

$$P(C = no) = P(C = no, T = sunny) + P(C = no, T = rainy) + P(C = no, T = snowy).$$

b. Calculați entropiile $H(C)$ și $H(T)$.

c. Calculați entropiile condiționale medii $H(C|T)$ și $H(T|C)$.

Răspuns:

a. Folosind formula dată, vom obține: $P_C = (0.6, 0.4)$ și $P_T = (0.35, 0.35, 0.30)$.

b. Aplicând definiția pentru entropie (vedeți problema 55), rezultă:

$$\begin{aligned} H(C) &= 0.6 \log_2 \frac{5}{3} + 0.4 \log_2 \frac{5}{2} = \log_2 5 - 0.6 \log_2 3 - 0.4 = 0.971 \text{ biți} \\ H(T) &= 2 \cdot 0.35 \log_2 \frac{20}{7} + 0.3 \log_2 \frac{10}{3} \\ &= 0.7(2 + \log_2 5 - \log_2 7) + 0.3(1 + \log_2 5 - \log_2 3) \\ &= 1.7 + \log_2 5 - 0.7 \log_2 7 - 0.3 \log_2 3 = 1.581 \text{ biți}. \end{aligned}$$

c. Aplicând definiția pentru entropie condițională medie (vedeți de asemenea problema 55), vom avea:

$$\begin{aligned} H(C|T) &\stackrel{\text{def.}}{=} \sum_{t \in \text{Val}(T)} P(T=t) \cdot H(C|T=t) \\ &= P(T=sunny) \cdot H(C|T=sunny) + P(T=rainy) \cdot H(C|T=rainy) + \\ &\quad P(T=snowy) \cdot H(C|T=snowy) \\ &= 0.35 \cdot H\left(\frac{0.30}{0.30+0.05}, \frac{0.05}{0.30+0.05}\right) + 0.35 \cdot H\left(\frac{0.20}{0.20+0.15}, \frac{0.15}{0.20+0.15}\right) + \\ &\quad 0.30 \cdot H\left(\frac{0.10}{0.10+0.20}, \frac{0.20}{0.20+0.10}\right) \\ &= \frac{7}{20} \cdot H\left(\frac{6}{7}, \frac{1}{7}\right) + \frac{7}{20} \cdot H\left(\frac{4}{7}, \frac{3}{7}\right) + \frac{3}{10} \cdot H\left(\frac{1}{3}, \frac{2}{3}\right) \\ &= \frac{7}{20} \cdot \left(\frac{6}{7} \log_2 \frac{7}{6} + \frac{1}{7} \log_2 7\right) + \frac{7}{20} \cdot \left(\frac{4}{7} \log_2 \frac{7}{4} + \frac{3}{7} \log_2 \frac{7}{3}\right) + \frac{3}{10} \cdot \left(\frac{1}{3} \log_2 3 + \frac{2}{3} \log_2 \frac{3}{2}\right) \\ &= \frac{7}{20} \cdot \left(\log_2 7 - \frac{6}{7} - \frac{6}{7} \log_2 3\right) + \frac{7}{20} \cdot \left(\log_2 7 - \frac{8}{7} - \frac{3}{7} \log_2 3\right) + \frac{3}{10} \cdot \left(\log_2 3 - \frac{2}{3}\right) \\ &= \frac{7}{10} \log_2 7 - \left(\frac{3}{10} + \frac{4}{10} + \frac{2}{10}\right) - \left(\frac{6}{20} + \frac{3}{20} - \frac{3}{10}\right) \cdot \log_2 3 \\ &= \frac{7}{10} \log_2 7 - \frac{3}{20} \log_2 3 - \frac{9}{10} = 0.82715 \text{ biți}. \end{aligned}$$

Similar,

$$\begin{aligned} H(T|C) &\stackrel{\text{def.}}{=} \sum_{c \in \text{Val}(C)} P(C=c) \cdot H(T|C=c) \\ &= P(C=no) \cdot H(T|C=no) + P(C=yes) \cdot H(T|C=yes) \\ &= 0.60 \cdot H\left(\frac{0.30}{0.30+0.20+0.10}, \frac{0.20}{0.30+0.20+0.10}, \frac{0.10}{0.30+0.20+0.10}\right) \\ &\quad + 0.40 \cdot H\left(\frac{0.05}{0.05+0.15+0.20}, \frac{0.15}{0.05+0.15+0.20}, \frac{0.20}{0.05+0.15+0.20}\right) \\ &= \frac{3}{5} \cdot H\left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right) + \frac{2}{5} \cdot H\left(\frac{1}{8}, \frac{3}{8}, \frac{1}{2}\right) \\ &= \frac{3}{5} \left(\frac{1}{2} + \frac{1}{3} \log_2 3 + \frac{1}{6}(1 + \log_2 3)\right) + \frac{2}{5} \left(\frac{1}{8} \cdot 3 + \frac{3}{8}(3 - \log_2 3) + \frac{1}{2}\right) \\ &= \frac{3}{5} \left(\frac{2}{3} + \frac{1}{2} \log_2 3\right) + \frac{2}{5} \left(2 - \frac{3}{8} \log_2 3\right) = \frac{6}{5} + \frac{3}{20} \log_2 3 = 1.43774 \text{ biți}. \end{aligned}$$

58.

(Câștigul de informație / informația mutuală,
o aplicație: selecția de trăsături)

■ □ • ○ CMU, 2009 spring, Ziv Bar-Joseph, HW5, pr. 6

În tabelul următor se dă un set de opt observații / instanțe, reprezentate ca tupluri de valori ale variabilelor aleatoare binare de „intrare“ X_1, X_2, X_3, X_4, X_5 și ale variabilei aleatoare binare de „ieșire“ Y .

Am dori să reducem spațiul de trăsături $\{X_1, X_2, X_3, X_4, X_5\}$ folosind o metodă de selecție de tip *filtru*.

a. Calculați câștigul de informație / informația mutuală $MI(X_i, Y)$ pentru fiecare i .

b. Înținând cont de rezultatul de la punctul precedent, alegeți cel mai mic subset de trăsături în aşa fel încât cel mai bun clasificator antrenat pe acest spațiu (redus) de trăsături să fie cel puțin la fel de bun ca și cel mai bun clasificator antrenat pe întreg spațiul de trăsături. Justificați alegerea pe care ați făcut-o.

X_1	X_2	X_3	X_4	X_5	Y
0	1	1	0	1	0
1	0	0	0	1	0
0	1	0	1	0	1
1	1	1	1	0	1
0	1	1	0	0	1
0	0	0	1	1	1
1	0	0	1	0	1
1	1	1	0	1	1

Răspuns:

a. Pentru calculul informației mutuale putem folosi formula din problema 139 sau problema 63.b:

$$MI(X, Y) = \sum_x \sum_y p_{X,Y}(x, y) \log \left(\frac{p_{X,Y}(x, y)}{p_X(x)p_Y(y)} \right)$$

Probabilitățile marginale, estimate în sensul verosimilității maxime (MLE), sunt:

	P_{X_1}	P_{X_2}	P_{X_3}	P_{X_4}	P_{X_5}	P_Y
0	1/2	3/8	1/2	1/2	1/2	1/4
1	1/2	5/8	1/2	1/2	1/2	3/4

iar probabilitățile comune sunt:

X_i	Y	$P_{X_1,Y}$	$P_{X_2,Y}$	$P_{X_3,Y}$	$P_{X_4,Y}$	$P_{X_5,Y}$
0	0	1/8	1/8	1/8	1/4	0
0	1	3/8	1/4	3/8	1/4	1/2
1	0	1/8	1/8	1/8	0	1/4
1	1	3/8	1/2	3/8	1/2	1/4

Se poate observa că X_1 și Y sunt independente, deci $MI(X_1, Y) = 0$, conform proprietății care este demonstrată la problema 139 (sau la problemele 63.c și 144.b). Similar, $MI(X_3, Y) = 0$. În rest, efectuând calculele obținem $MI(X_2, Y) = 0.01571$, $MI(X_4, Y) = 0.3113$ și $MI(X_5, Y) = 0.3113$.

b. La selecția de trăsături vom alege acele trăsături X_i care au informație mutuală nenulă în raport cu Y . Acestea sunt X_2, X_4 și X_5 . Celelalte două trăsături, X_1 și X_3 sunt independente în raport cu Y .

Totuși, inspectând datele, observăm că dacă vom selecta doar trăsăturile X_2, X_4 și X_5 vom avea două instanțe (vedeți prima și ultima linie din tabel) care au aceleași trăsături ($X_2 = 1, X_4 = 0, X_5 = 1$) dar au etichete / ieșiri diferite: $Y = 0$, respectiv $Y = 1$. Așadar, vom adăuga la setul de trăsături selectate anterior și variabila X_1 , care va permite dezambiguizarea în cazul acestor două instanțe, menținând astfel „consistența“ setului de date.

Observație: Deși $MI(X_1, Y) = 0$ — sau, echivalent spus, X_1 este independent de Y —, nu rezultă că variabila X_1 combinată cu una sau mai multe variabile X_j , cu $j \in \{2, 4, 5\}$, și formând astfel o nouă variabilă aleatoare, rămâne independentă de Y . Noua variabilă poate avea câștig de informație nenul (în unele cazuri chiar maxim!) în raport cu Y .

59. (Entropia comună: forma particulară a relației de „înlănțuire“ în cazul variabilelor aleatoare independente)

prelucrare de Liviu Ciortuz, după CMU, 2012 spring, Roni Rosenfeld, HW2, pr. 7.b

Conform problemei 55.c, formula de înlănțuire a entropiilor pentru cazul general (adică, indiferent dacă X și Y sunt sau nu independente) este:

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y). \quad (61)$$

Demonstrați că dacă X și Y sunt variabile aleatoare discrete independente, atunci $H(X, Y) = H(X) + H(Y)$, și reciproc: atunci când are loc egalitatea $H(X, Y) = H(X) + H(Y)$ rezultă că variabilele X și Y sunt independente.

Răspuns:

Conform definiției câștigului de informație (vedeți problema 55),

$$IG(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \quad (62)$$

De asemenea, conform problemei 144.b (sau, conform problemei 63.c),

$$IG(X, Y) = 0 \Leftrightarrow X \text{ și } Y \text{ sunt independente.} \quad (63)$$

Din relațiile (62) și (63) rezultă că

$$H(Y) = H(Y|X) \Leftrightarrow X \text{ și } Y \text{ sunt independente.} \quad (64)$$

Așadar, dacă X și Y sunt independente, coroborând relațiile (64) și (61) vom avea $H(X, Y) = H(Y) + H(X)$.

Invers, dacă $H(X, Y) = H(X) + H(Y)$, din relația (61) rezultă că $H(Y) = H(Y|X)$, ceea ce implică faptul că X și Y sunt independente, conform relației (64).

Observație: Din egalitățile (61) și (62), rezultă

$$H(X, Y) = H(X) + H(Y) - IG(X, Y).$$

Conform proprietății de nenegativitate a câștigului de informație ($IG(X, Y) \geq 0$; vedeți problema 144.a sau problema 63.c), rezultă

$$H(X, Y) \leq H(X) + H(Y).$$

Această ultimă relație este ilustrată în figura de la finalul enunțului problemei 55.

60.

(Entropia distribuției continue uniforme și a distribuțiilor gaussiene unidimensionale și multidimensionale)

 Liviu Ciortuz, 2019

Pentru o variabilă aleatoare X care urmează o distribuție continuă având funcția densitate de probabilitate (p.d.f.) p , entropia se definește astfel:

$$H(X) = \int_{-\infty}^{+\infty} p(x) \log_2 \frac{1}{p(x)} dx \stackrel{p(x) \neq 0}{=} - \int_{-\infty}^{+\infty} p(x) \log_2 p(x) dx.$$

Indicație: Dacă $p(x) = 0$, veți presupune că $-p(x) \log_2 p(x) = 0$.

- a. Calculați entropia distribuției uniforme definite pe intervalul $[a, b]$.
- b. Calculați entropia distribuției gaussiene unidimensionale de parametri $\mu \in \mathbb{R}$ și $\sigma^2 \in \mathbb{R}_+$, pentru care funcția de densitate de probabilitate (p.d.f.) este:

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ pentru } x \in \mathbb{R}.$$

- c. Calculați entropia distribuției gaussiene multidimensionale de parametri $\mu \in \mathbb{R}^d$ și $\Sigma \in \mathbb{R}^{d \times d}$ (matrice simetrică și pozitiv definită), pentru care funcția de densitate de probabilitate (p.d.f.) este:

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1} (x-\mu)\right) \text{ pentru } x \in \mathbb{R}^d.$$

Sugestie: Următoarea formulă, preluată din documentul *Matrix Identities*, de Sam Roweis, 1999, vă poate fi de folos:

$$(2e) \text{ Tr}(AB) = \text{Tr}(BA).$$

Răspuns:

- a. Dacă f este funcția densitate de probabilitate (p.d.f.) a distribuției uniforme continue pe intervalul $[a, b]$, atunci prin definiție există o constantă $c \in \mathbb{R}_+$ cu proprietatea că $f(x) = c$ pentru orice $x \in [a, b]$. Din condiția [care ține tot de definiție] $\int_a^b f(x) dx = 1$ rezultă imediat că valoarea constantei c este $\frac{1}{b-a}$. Așadar,

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{pentru } \forall x \in [a, b]; \\ 0 & \text{altfel.} \end{cases}$$

Entropia unei variabile aleatoare X care urmează această distribuție uniformă continuă se calculează astfel:

$$\begin{aligned} H(X) &= - \int_a^b f(x) \log_2 f(x) dx = \int_a^b \frac{1}{b-a} \log_2 (b-a) dx = \frac{1}{b-a} \log_2(b-a) \cdot x \Big|_a^b \\ &= \frac{1}{b-a} \cdot \log_2(b-a) \cdot (b-a) = \log_2(b-a). \end{aligned}$$

b.¹²⁰ Pentru a calcula entropia distribuției gaussiene unidimensionale, ne vom folosi de următoarele rezultate:

$$\int_{-\infty}^{\infty} \mathcal{N}(x) dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 1 \quad (\text{deoarece } \mathcal{N} \text{ este p.d.f.}) \quad (65)$$

$$\int_{-\infty}^{\infty} v^2 e^{-\frac{v^2}{2}} dv = \sqrt{2\pi} \quad (\text{din rezolvarea ex. 32.c}) \quad (66)$$

Fie $X \sim \mathcal{N}(x|\mu, \sigma^2)$. Entropia lui X se calculează astfel:

$$\begin{aligned} H(X) &\stackrel{\text{def.}}{=} \int_{-\infty}^{\infty} \mathcal{N}(x) \log_2 \frac{1}{\mathcal{N}(x)} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \log_2 \left(\sqrt{2\pi}\sigma \cdot e^{+\frac{(x-\mu)^2}{2\sigma^2}} \right) dx \\ &= \log_2(\sqrt{2\pi}\sigma) \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx}_1 + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \log_2 e^{+\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &\stackrel{(65)}{=} \frac{1}{\ln 2} \left(\ln \sqrt{2\pi}\sigma + \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{2\sigma^2} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \right). \end{aligned}$$

Pentru a calcula integrala, vom face schimbarea de variabilă $v = \frac{x-\mu}{\sigma}$, din care rezultă $x = \sigma v + \mu$ și $dx = \sigma dv$. Mai departe, reluând calculul entropiei $H(X)$, vom avea:

$$\begin{aligned} H(X) &= \frac{1}{\ln 2} \left(\ln \sqrt{2\pi}\sigma + \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \frac{v^2}{2} e^{-\frac{v^2}{2}} \sigma dv \right) \stackrel{(66)}{=} \frac{1}{\ln 2} \left(\frac{1}{2} \ln(2\pi\sigma^2) + \frac{1}{2\sqrt{2\pi}} \sqrt{2\pi} \right) \\ &= \frac{1}{2} \frac{\ln(2\pi\sigma^2 e)}{\ln 2} = \frac{1}{2} \log_2(2\pi\sigma^2 e) = \log_2 \sqrt{2\pi\sigma^2 e}. \end{aligned}$$

c.¹²¹ Fie $X \sim \mathcal{N}(x|\mu, \Sigma)$ și $x \in \mathbb{R}^d$. Vom folosi următoarele proprietăți:

$$E[\ln p(x)] \stackrel{\text{def.}}{=} \int_{-\infty}^{+\infty} p(x) \ln p(x) dx \quad (67)$$

$$(x - \mu)^\top \in \mathbb{R}^{1 \times d}, \Sigma^{-1} \in \mathbb{R}^{d \times d}, (x - \mu) \in \mathbb{R}^{d \times 1} \Rightarrow (x - \mu)^\top \Sigma^{-1} (x - \mu) \in \mathbb{R}$$

$$\Rightarrow (x - \mu)^\top \Sigma^{-1} (x - \mu) = \text{Tr}((x - \mu)^\top \Sigma^{-1} (x - \mu)) \quad (68)$$

$$\Sigma \mu = E[X] \quad (69)$$

$$\Sigma = \text{Var}(X) \stackrel{\text{def.}}{=} \text{Cov}(X, X) \stackrel{\text{def.}}{=} E[(X - E[X])(X - E[X])^\top] \stackrel{(69)}{=} E[(X - \mu)(X - \mu)^\top]. \quad (70)$$

Notă: Pentru demonstrarea egalităților $\mu = E[X]$ și $\Sigma = \text{Var}(X)$, vedeti www.statlect.com/probability-distributions/multivariate-normal-distribution.

¹²⁰Soluție redactată inițial de Andi Munteanu (std., FII, 2019f).

¹²¹Soluție redactată inițial de Georgiana Ojoc (std., FII, 2020f).

Entropia lui X se calculează astfel:

$$\begin{aligned}
 H(X) &\stackrel{\text{def.}}{=} - \int_{-\infty}^{+\infty} p(x) \log_2 p(x) dx = -\frac{1}{\ln 2} \int_{-\infty}^{+\infty} p(x) \ln p(x) dx \\
 &\stackrel{(67)}{=} -\frac{1}{\ln 2} E[\ln p(x)] dx = -\frac{1}{\ln 2} E\left[\ln\left(\frac{1}{(\sqrt{2\pi})^d \sqrt{|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1} (x-\mu)}\right)\right] dx \\
 &= -\frac{1}{\ln 2} E\left[-\frac{d}{2} \ln(2\pi) - \frac{1}{2} \ln(|\Sigma|) - \frac{1}{2}(x-\mu)^\top \Sigma^{-1} (x-\mu)\right] \\
 &= \frac{1}{\ln 2} \left(\frac{d}{2} \ln(2\pi) + \frac{1}{2} \ln(|\Sigma|) + \frac{1}{2} E[(x-\mu)^\top \Sigma^{-1} (x-\mu)]\right) \\
 &\stackrel{(68)}{=} \frac{1}{\ln 2} \left(\frac{d}{2} \ln(2\pi) + \frac{1}{2} \ln(|\Sigma|) + \frac{1}{2} E[\text{Tr}((x-\mu)^\top \Sigma^{-1} (x-\mu))]\right) \\
 &\stackrel{(2e)}{=} \frac{1}{\ln 2} \left(\frac{d}{2} \ln(2\pi) + \frac{1}{2} \ln(|\Sigma|) + \frac{1}{2} E[\text{Tr}(\Sigma^{-1}(x-\mu)(x-\mu)^\top)]\right) \\
 &\stackrel{\text{lin. med.}}{=} \frac{1}{\ln 2} \left(\frac{d}{2} \ln(2\pi) + \frac{1}{2} \ln(|\Sigma|) + \frac{1}{2} \text{Tr}(\Sigma^{-1} E[(x-\mu)(x-\mu)^\top])\right) \\
 &\stackrel{(70)}{=} \frac{1}{\ln 2} \left(\frac{d}{2} \ln(2\pi) + \frac{1}{2} \ln(|\Sigma|) + \frac{1}{2} \text{Tr}(\Sigma^{-1} \Sigma)\right) \\
 &= \frac{1}{\ln 2} \left(\frac{d}{2} \ln(2\pi) + \frac{1}{2} \ln(|\Sigma|) + \frac{1}{2} \text{Tr}(I_d)\right) \\
 &= \frac{1}{\ln 2} \left(\frac{d}{2} \ln(2\pi) + \frac{1}{2} \ln(|\Sigma|) + \frac{d}{2}\right).
 \end{aligned}$$

61.

(Calcularea entropiei unor variabile aleatoare continue:
cazul distribuției exponențiale și al distribuției Gamma)

a.

■ CMU, 2011 spring, Roni Rosenfeld, HW2, pr. 2.c

Calculați entropia *distribuției* continue *exponențiale* de parametru $\lambda > 0$. Vă reamintim că definiția p.d.f.-ului acestei distribuții este următoarea:¹²²

$$p(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{dacă } x \geq 0; \\ 0, & \text{dacă } x < 0. \end{cases}$$

b.

UAIC, 2020, Georgiana Ojoc, Liviu Ciortuz

Calculați entropia distribuției Gamma, pentru care funcția de densitate de probabilitate este:¹²³

$$p(x) \stackrel{\text{not.}}{=} \text{Gamma}(x | r, \alpha) \stackrel{\text{def.}}{=} \frac{\alpha^r}{\Gamma(r)} x^{r-1} e^{-\alpha x} \text{ pentru } x \geq 0,$$

cu $r > 0$ (forma), $\alpha > 0$ (rata) și $\Gamma(r) \stackrel{\text{def.}}{=} \int_0^{+\infty} x^{r-1} e^{-x} dx$ pentru orice $r > 0$ (funcția lui Euler). (Vedeți ex. 31.b.)

Indicație:

Este posibil să fie nevoie să folosiți o proprietate specială, numită *regula de derivare sub semnul de integrală* (engl., differentiation under the integral sign),

¹²²La ex. 31.a puteți vedea graficul acestei funcții de densitate pentru câteva valori ale parametrului λ .

¹²³La ex. 31.c puteți vedea graficul acestei funcții de densitate pentru câteva valori ale parametrilor r și α .

care constituie obiectul pentru următoarea

Teoremă: Dacă $f(r, t)$ este o funcție cu valori reale, continuă și derivabilă în raport cu r pe intervalul (a, b) , iar derivata parțială $\frac{\partial}{\partial r} f(r, t)$ este de asemenea continuă pe intervalul (a, b) , atunci¹²⁴

$$\int_a^b \frac{\partial}{\partial r} f(r, t) dt = \frac{\partial}{\partial r} \int_a^b f(r, t) dt.$$

Menționăm că regula aceasta, datorată lui Gottfried Leibnitz (1646-1716), se generalizează — sub o formă semnificativ mai elaborată — la cazul când limitele de integrare a și b depind de r , adică sunt de forma $a(r)$ și respectiv $b(r)$, iar aceste două funcții sunt continue și au derivatele continue.

Răspuns:

a. Dat fiind faptul că funcția p se anulează pe intervalul $(-\infty, 0)$, este natural ca mai întâi să „rupem“ intervalul de integrare pentru $\int_{-\infty}^{\infty} p(x) \log_2 \frac{1}{p(x)} dx$ în două: $(-\infty, 0)$ și $[0, \infty)$. Așadar,

$$\begin{aligned} H(X) &= \int_{-\infty}^0 p(x) \log_2 \frac{1}{p(x)} dx + \int_0^{\infty} p(x) \log_2 \frac{1}{p(x)} dx \\ &\stackrel{\text{def. } p}{=} - \int_{-\infty}^0 0 \log_2 0 dx + \int_0^{\infty} \lambda e^{-\lambda x} \log_2 \frac{1}{\lambda e^{-\lambda x}} dx. \end{aligned}$$

Prima dintre aceste două ultime integrale este 0, conform *indicării* din enunț. Pentru a putea calcula mai ușor cea de-a două integrală (în expresia căreia apare numărul e), vom schimba baza logaritmului, și anume vom trece din baza 2 în baza e (baza logaritmului natural, \ln).¹²⁵

Prin urmare,

$$\begin{aligned} H(X) &= \frac{1}{\ln 2} \int_0^{\infty} \lambda e^{-\lambda x} \ln \frac{1}{\lambda e^{-\lambda x}} dx \\ &= \frac{1}{\ln 2} \int_0^{\infty} \lambda e^{-\lambda x} \left(\ln \frac{1}{\lambda} + \ln \frac{1}{e^{-\lambda x}} \right) dx \\ &= \frac{1}{\ln 2} \int_0^{\infty} \lambda e^{-\lambda x} (-\ln \lambda + \ln e^{\lambda x}) dx \\ &= \frac{1}{\ln 2} \int_0^{\infty} \lambda e^{-\lambda x} (-\ln \lambda + \lambda x) dx \\ &= \frac{1}{\ln 2} \int_0^{\infty} \lambda e^{-\lambda x} (-\ln \lambda) dx + \frac{1}{\ln 2} \int_0^{\infty} \lambda e^{-\lambda x} \lambda x dx \\ &= \frac{-\ln \lambda}{\ln 2} \int_0^{\infty} \lambda e^{-\lambda x} dx + \frac{\lambda}{\ln 2} \int_0^{\infty} \lambda e^{-\lambda x} x dx. \end{aligned}$$

Prima integrală are valoarea 1, întrucât $p()$ este p.d.f.-ul distribuției exponentiale (vedeți ex. 31.a). Cea de-a două integrală are valoarea $1/\lambda$, întrucât

¹²⁴Pentru demonstrație, vedeți „Proof of basic form“ pe site-ul https://en.wikipedia.org/wiki/Leibniz_integral_rule#Alternative_Proof_of_General_Form_with_Variable_Limits,_using_the_Chain_Rule (accesat la data de 15.12.2020).

¹²⁵Pentru aceasta, vom folosi formula $\log_a b = \frac{\log_c b}{\log_c a}$, valabilă pentru orice $a > 0$, $b > 0$ și $c > 0$, cu $a \neq 1$ și $c \neq 1$. În calculele care urmează vom folosi și alte formule specifice logaritmilor.

reprezintă media distribuției exponențiale (vedeți enunțul același ex. 31.a). Prin urmare,

$$H(X) = -\frac{\ln \lambda}{\ln 2} + \frac{\lambda}{\ln 2} \cdot \frac{1}{\lambda} = -\frac{\ln \lambda}{\ln 2} + \frac{1}{\ln 2} = \frac{1 - \ln \lambda}{\ln 2}.$$

b. Fie $X \sim \text{Gamma}(x | r, \alpha)$. Entropia lui X se calculează astfel:

$$\begin{aligned} H(X) &\stackrel{\text{def.}}{=} - \int_0^{+\infty} p(x) \log_2 p(x) dx = -\frac{1}{\ln 2} \int_0^{+\infty} p(x) \ln p(x) dx \\ &= -\frac{1}{\ln 2} \int_0^{+\infty} p(x) \ln \left(\frac{\alpha^r}{\Gamma(r)} x^{r-1} e^{-\alpha x} \right) dx \\ &= -\frac{1}{\ln 2} \int_0^{+\infty} p(x) \left(\ln \frac{\alpha^r}{\Gamma(r)} + \ln x^{r-1} + \ln e^{-\alpha x} \right) dx \\ &= -\frac{1}{\ln 2} \left(\ln \frac{\alpha^r}{\Gamma(r)} \underbrace{\int_0^{+\infty} p(x) dx}_{I_1} + (r-1) \underbrace{\int_0^{+\infty} p(x) \ln x dx}_{I_2} - \alpha \underbrace{\int_0^{+\infty} x p(x) dx}_{I_3} \right). \end{aligned}$$

Integrala I_1 are valoarea 1, deoarece $p(\cdot)$ este p.d.f. Integrala I_3 , care reprezintă media distribuției Gamma, are valoarea $\frac{r}{\alpha}$ (vedeți exercițiul 31.c). Pentru a calcula integrala I_2 , vom folosi următoarele egalități:

$$\frac{\partial}{\partial r} (e^{-\alpha x} x^{r-1}) = e^{-\alpha x} x^{r-1} \ln x \quad (71)$$

$$\frac{\partial}{\partial r} \ln \frac{\Gamma(r)}{\alpha^r} = \frac{\alpha^r}{\Gamma(r)} \cdot \frac{\partial}{\partial r} \left(\frac{\Gamma(r)}{\alpha^r} \right) \quad (72)$$

$$\psi(x) \stackrel{\text{def.}}{=} \frac{\Gamma'(r)}{\Gamma(r)} = \frac{\partial}{\partial r} \ln \Gamma(r) \text{ pentru orice } r > 0 \text{ (funcția digamma)} \quad (73)$$

$$\frac{\partial}{\partial r} \Gamma(r) = \frac{\partial}{\partial r} \int_0^{+\infty} t^{r-1} e^{-t} dt = \int_0^{+\infty} \frac{\partial}{\partial r} (t^{r-1} e^{-t}) dt = \int_0^{+\infty} t^{r-1} e^{-t} \ln t dt. \quad (74)$$

Interschimbarea simbolilor de derivare și respectiv de integrare de pe ultima linie este posibilă întrucât sunt îndeplinite condițiile de aplicare pentru regula de „derivare sub semnul de integrală“ a lui Leibnitz.

Acum,

$$\begin{aligned} I_2 &= \int_0^{+\infty} \frac{\alpha^r}{\Gamma(r)} x^{r-1} e^{-\alpha x} \ln x dx = \frac{\alpha^r}{\Gamma(r)} \int_0^{+\infty} x^{r-1} e^{-\alpha x} \ln x dx \\ &\stackrel{(71)}{=} \frac{\alpha^r}{\Gamma(r)} \int_0^{+\infty} \frac{\partial}{\partial r} (x^{r-1} e^{-\alpha x}) dx \\ &= \frac{\alpha^r}{\Gamma(r)} \int_0^{+\infty} \frac{\partial}{\partial r} \left(\frac{1}{\alpha^{r-1}} \cdot (\alpha x)^{r-1} e^{-\alpha x} \right) dx = \frac{\alpha^r}{\Gamma(r)} \frac{\partial}{\partial r} \int_0^{+\infty} \frac{1}{\alpha^{r-1}} \cdot (\alpha x)^{r-1} e^{-\alpha x} dx \\ &= \frac{\alpha^r}{\Gamma(r)} \cdot \frac{\partial}{\partial r} \left(\frac{1}{\alpha^{r-1}} \int_0^{+\infty} (\alpha x)^{r-1} e^{-\alpha x} dx \right). \end{aligned}$$

Egalitatea de pe penultima linie are loc întrucât putem aplica tot regula de „derivare sub semnul de integrală“ a lui Leibnitz (așa cum s-a procedat și la calcularea lui $\Gamma'(r) \stackrel{\text{not.}}{=} \frac{\partial}{\partial r} \Gamma(r)$, la egalitatea (74)).

Folosind schimbarea de variabilă $v = \alpha x$ care implică $dx = \frac{dv}{\alpha}$, obținem:

$$\begin{aligned} I_2 &= \frac{\alpha^r}{\Gamma(r)} \cdot \frac{\partial}{\partial r} \left(\frac{1}{\alpha^r} \int_0^{+\infty} v^{r-1} e^{-v} dv \right) = \frac{\alpha^r}{\Gamma(r)} \cdot \frac{\partial}{\partial r} \left(\frac{\Gamma(r)}{\alpha^r} \right) \stackrel{(72)}{=} \frac{\partial}{\partial r} \ln \frac{\Gamma(r)}{\alpha^r} \\ &= \frac{\partial}{\partial r} (\ln \Gamma(r) - \ln \alpha^r) = \frac{\partial}{\partial r} \ln \Gamma(r) - \frac{\partial}{\partial r} (r \ln \alpha) \stackrel{(73)}{=} \psi(r) - \ln \alpha. \end{aligned}$$

Folosind toate aceste rezultate intermediiare, în final vom obține:

$$\begin{aligned} H(X) &= -\frac{1}{\ln 2} \left(\ln \frac{\alpha^r}{\Gamma(r)} + (r-1)(\psi(r) - \ln \alpha) - \alpha \frac{r}{\alpha} \right) \\ &= -\frac{1}{\ln 2} (r \ln \alpha - \ln \Gamma(r) + (r-1)\psi(r) - r \ln \alpha + \ln \alpha - r) \\ &= \frac{1}{\ln 2} (r - \ln \alpha + \ln \Gamma(r) + (1-r)\psi(r)). \end{aligned}$$

62. (Redescoperirea definiției entropiei
pornind de la un set de proprietăți dezirabile ale ei)
■ □ • CMU, 2005 fall, T. Mitchell, A. Moore, HW1, pr. 2.2

Prin definiție, *entropia* (în sens Shannon) a unei variabile aleatoare discrete X ale cărei valori sunt luate cu probabilitățile p_1, p_2, \dots, p_n este $H(X) = -\sum_i p_i \log p_i$. Însă legătura dintre această definiție formală și obiectivul avut în vedere — și anume, acela de a exprima gradul de *incertitudine* cu care se produc valorile unei astfel de variabile aleatoare — nu este foarte intuitivă.

Scopul acestui exercițiu este de a arăta că orice funcție $\psi_n(p_1, \dots, p_n)$ care satisface trei proprietăți dezirabile (sau, axiome) pentru entropie este în mod necesar de forma $-K \sum_i p_i \log p_i$ unde K este o constantă reală pozitivă. Iată care sunt aceste proprietăți:¹²⁶

A1. Funcția $\psi_n(p_1, \dots, p_n)$ este continuă în fiecare din argumentele ei și simetrică.

Din punct de vedere formal, în acest caz simetria se traduce prin egalitatea $\psi_n(p_1, \dots, p_i, \dots, p_j, \dots, p_n) = \psi_n(p_1, \dots, p_j, \dots, p_i, \dots, p_n)$ pentru orice $i \neq j$. Informal spus, dacă două dintre valorile care sunt luate de variabila aleatoare X (și anume x_i și x_j) își schimbă între ele probabilitățile (p_i și respectiv p_j), valoarea entropiei lui X nu se schimbă.

A2. Funcția $\psi_n(1/n, \dots, 1/n)$ este strict crescătoare în raport cu n .

Altfel spus, dacă toate evenimentele sunt echiprobabile, atunci entropia crește odată cu numărul de evenimente posibile.

A3. Dacă faptul de a alege între mai multe evenimente posibile poate fi realizat prin mai multe alegeri succesive, atunci $\psi_n(p_1, \dots, p_n)$ trebuie să se poată scrie ca o sumă ponderată a entropiilor calculate la fiecare stadiu / alegere.¹²⁷

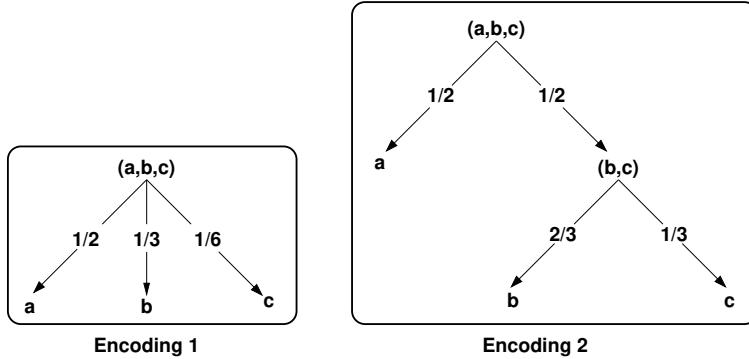
De exemplu, dacă evenimentele (a, b, c) se produc respectiv cu probabilitățile $(1/2, 1/3, 1/6)$, atunci acest fapt poate fi echivalat cu

¹²⁶LC: Deși nu se specifică în enunțul original al problemei, este necesar / natural să considerăm și proprietatea următoare: [A0.] $\psi_1(1) = 0$ pentru că ψ_n este văzută că măsură a *dezordinității* / *incertitudinii*, iar în cazul particular $n = 1$ nu există niciun fel de dezordine / incertitudine.

¹²⁷Această proprietate poate fi folosită pentru a elabora o [variantă de] rezolvare pentru problema 50 de la capitolul *Arbore de decizie*.

- a alege mai întâi cu probabilitate de $1/2$ între a și (b, c) ,
- urmat de a alege între b și c cu probabilitățile $2/3$ și $1/3$ respectiv.
(A se vedea imaginile de mai jos, Encoding 1 și Encoding 2.)

Din punct de vedere formal, proprietatea A3 impune ca, pe acest exemplu, $\psi_3(1/2, 1/3, 1/6)$ să fie egal cu $\psi_2(1/2, 1/2) + 1/2 \cdot \psi_2(2/3, 1/3)$.



Așadar, în acest exercițiu vi se cere să arătați că dacă o funcție de n variabile $\psi_n(p_1, \dots, p_n)$ satisfac proprietățile A1, A2 și A3 de mai sus, atunci există $K \in \mathbb{R}^+$ astfel încât $\psi_n(p_1, \dots, p_n) = -K \sum_i p_i \log p_i$ unde $K \in \mathbb{R}_+$ este o constantă.¹²⁸

Indicație:

Veți face rezolvarea acestei probleme în mod gradual, parcurgând următoarele puncte (dintre care primele două puncte au rolul de a vă acomoda cu noțiunile din enunț):

- Arătați că $H(1/2, 1/3, 1/6) = H(1/2, 1/2) + \frac{1}{2}H(2/3, 1/3)$. Altfel spus, verificați faptul că definiția clasică a entropiei, $H(X) = \sum_i p_i \log 1/p_i$, satisfac proprietatea A3 pe exemplul care a fost dat mai sus.
- Calculați entropia în cazul distribuției / „codificării“ din figura alăturată, folosind din nou proprietatea A3.

Următoarele întrebări tratează cazul particular $A(n) \stackrel{\text{not.}}{=} \psi_n(1/n, 1/n, \dots, 1/n)$.

- Arătați că

$$A(s^m) = m A(s) \text{ pentru orice } s, m \in \mathbb{N}^*. \quad (75)$$

La punctele $d - g$ de mai jos, pentru orice număr $t \in \mathbb{N}^*$ (fixat), vom considera — pe lângă $n \in \mathbb{N}^*$, care a fost de fapt introdus atunci când am spus că aici ne ocupăm de $A(n)$ — numerele $s, m \in \mathbb{N}$, cu $s > 1$ astfel încât¹²⁹

¹²⁸În demonstrație, vom vedea, va rezulta $K = \frac{1}{\log s} \psi_s \left(\frac{1}{s}, \dots, \frac{1}{s} \right)$ pentru un număr oarecare $s \in \mathbb{N}^* \setminus \{1\}$, fixat.

¹²⁹LC: În idee, am putea să fixăm $s = 2$ și apoi să alegem $m \in \mathbb{N}$ (în funcție de t și n) astfel încât relația (76) să fie satisfăcută. Însă raționamentul următor nu depinde (în esență) de valoarea aleasă (și fixată) pentru s .

$$s^m \leq t^n \leq s^{m+1}. \quad (76)$$

Precizare: Odată ce am fixat t , s și n , numărul m se alege astfel încât să aibă loc dubla inegalitate (76).

d. Verificați că, prin logaritmarea¹³⁰ dublei inegalități (76) și apoi prin reaaranjare, obținem $\frac{m}{n} \leq \frac{\log t}{\log s} \leq \frac{m}{n} + \frac{1}{n}$ pentru $s \neq 1$, și deci

$$\left| \frac{m}{n} - \frac{\log t}{\log s} \right| \leq \frac{1}{n}. \quad (77)$$

e. Explicați de ce $A(s^m) \leq A(t^n) \leq A(s^{m+1})$.

f. Combinând ultima inegalitate de mai sus cu egalitatea (75), avem $A(s^m) \leq A(t^n) \leq A(s^{m+1}) \Rightarrow m A(s) \leq n A(t) \leq (m+1) A(s)$. Verificați că

$$\left| \frac{m}{n} - \frac{A(t)}{A(s)} \right| \leq \frac{1}{n} \text{ pentru } s \neq 1. \quad (78)$$

g. Combinând inegalitățile (77) și (78), arătați că

$$\left| \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \right| \leq \frac{2}{n} \text{ pentru } s \neq 1 \quad (79)$$

și, în consecință

$$A(t) = K \log t \text{ cu } K > 0 \text{ (din cauza proprietății A2).} \quad (80)$$

h. Arătați că rezultatul de mai sus ($A(t) = K \log t \Leftrightarrow \psi_t(1/t, \dots, 1/t) = Kt \frac{1}{t} \log t$) se generalizează ușor la cazul $\psi_k(p_1, \dots, p_k)$ cu $p_i \in \mathbb{Q}^+$ pentru $i = 1, \dots, k$ și $\sum_{i=1}^k p_i = 1$:

$$\psi_k(p_1, \dots, p_k) = -K \sum_i p_i \log p_i.$$

i. În sfârșit, tratați și cazul $p_i \in \mathbb{R}^+$, $i = 1, \dots, k$, cu $\sum_i p_i = 1$.

Răspuns:

a. Facem calculele:

$$\begin{aligned} H\left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right) &= \frac{1}{2} \log 2 + \frac{1}{3} \log 3 + \frac{1}{6} \log 6 = \left(\frac{1}{2} + \frac{1}{6}\right) \log 2 + \left(\frac{1}{3} + \frac{1}{6}\right) \log 3 = \frac{2}{3} + \frac{1}{2} \log 3 \\ H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2} H\left(\frac{2}{3}, \frac{1}{3}\right) &= 1 + \frac{1}{2} \left(\frac{2}{3} \log \frac{3}{2} + \frac{1}{3} \log 3\right) = 1 + \frac{1}{2} \left(\log 3 - \frac{2}{3}\right) = \frac{2}{3} + \frac{1}{2} \log 3 \end{aligned}$$

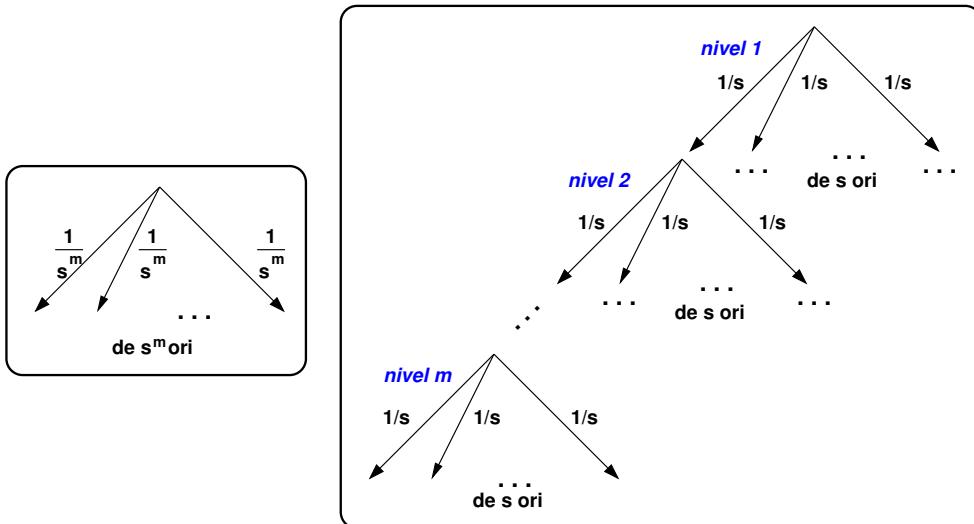
$$\text{și rezultă că } H\left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right) = H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2} H\left(\frac{2}{3}, \frac{1}{3}\right).$$

¹³⁰Puteți folosi ca bază a logaritmului orice număr supra-unitar, arbitrar ales, dar fixat.

b. Folosind proprietatea A3, entropia „codificării“ din figura dată în enunț este:

$$\begin{aligned}
 & H\left(\frac{1}{6}, \frac{1}{3}, \frac{1}{2}\right) + \frac{1}{6}H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2}H\left(\frac{3}{4}, \frac{1}{8}, \frac{1}{8}\right) \\
 & = \frac{1}{6}\log 6 + \frac{1}{3}\log 3 + \frac{1}{2}\log 2 + \frac{1}{6} + \frac{1}{2}\left(\frac{3}{4}\log \frac{4}{3} + \frac{2}{8}\log 8\right) \\
 & = \frac{1}{6}\log 2 + \frac{1}{6}\log 3 + \frac{1}{3}\log 3 + \frac{1}{2} + \frac{1}{6} + \frac{3}{8}(2 - \log 3) + \frac{3}{8} \\
 & = \frac{1}{6} + \frac{1}{2} + \frac{1}{6} + \frac{3}{4} + \frac{3}{8} + \left(\frac{1}{6} + \frac{1}{3} - \frac{3}{8}\right)\log 3 \\
 & = \frac{47}{24} + \frac{1}{8}\log 3 = 1.958 + 0.125\log 3 = 2.156
 \end{aligned}$$

c. Pentru calculul lui $A(s^m)$ se poate folosi atât o „codificare“ imediată cât și una (des)compusă, ca în figura următoare:



Aplicând proprietatea A3 pe „codificarea“ din figura de mai sus, partea dreaptă, avem:

$$\begin{aligned}
 A(s^m) &= A(s) + s \cdot \frac{1}{s}A(s) + s^2 \cdot \frac{1}{s^2}A(s) + \dots + s^{m-1} \cdot \frac{1}{s^{m-1}}A(s) \\
 &= \underbrace{A(s) + A(s) + A(s) + \dots + A(s)}_{\text{de } m \text{ ori}} = mA(s)
 \end{aligned}$$

d. Aplicând funcția \log fiecărui termen al inegalității $s^m \leq t^n \leq s^{m+1}$ obținem $m \log s \leq n \log t \leq (m+1) \log s$. Apoi, pentru $s \neq 1$, împărțind prin $n \log s$, rezultă:

$$\frac{m}{n} \leq \frac{\log t}{\log s} \leq \frac{m}{n} + \frac{1}{n} \Rightarrow 0 \leq \frac{\log t}{\log s} - \frac{m}{n} \leq \frac{1}{n} \Rightarrow \left| \frac{\log t}{\log s} - \frac{m}{n} \right| \leq \frac{1}{n}$$

e. Datorită proprietății A2 din enunț, inegalitatea $s^m \leq t^n \leq s^{m+1}$ implică

$$\psi_{s^m}\left(\frac{1}{s^m}, \dots, \frac{1}{s^m}\right) \leq \psi_{t^n}\left(\frac{1}{t^n}, \dots, \frac{1}{t^n}\right) \leq \psi_{s^{m+1}}\left(\frac{1}{s^{m+1}}, \dots, \frac{1}{s^{m+1}}\right)$$

ceea ce reprezintă exact $A(s^m) \leq A(t^n) \leq A(s^{m+1})$.

f. Datorită proprietății (75), dubla inegalitate $A(s^m) \leq A(t^n) \leq A(s^{m+1})$ devine $m A(s) \leq n A(t) \leq (m+1) A(s)$. Împărțind această inegalitate prin $n A(s)$, despre care se poate spune că este nenul pentru orice $s \neq 1$,¹³¹ rezultă:

$$\frac{m}{n} \leq \frac{A(t)}{A(s)} \leq \frac{m}{n} + \frac{1}{n} \Rightarrow 0 \leq \frac{A(t)}{A(s)} - \frac{m}{n} \leq \frac{1}{n} \Rightarrow \left| \frac{A(t)}{A(s)} - \frac{m}{n} \right| \leq \frac{1}{n}$$

g. Inegalitățile duble de mai jos rescriu convenabil proprietățile (77) și (78):

$$-\frac{1}{n} \leq \frac{m}{n} - \frac{\log t}{\log s} \leq \frac{1}{n} \quad \text{și} \quad -\frac{1}{n} \leq \frac{A(t)}{A(s)} - \frac{m}{n} \leq \frac{1}{n}$$

Însumându-le membru cu membru, rezultă

$$-\frac{2}{n} \leq \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \leq \frac{2}{n} \Rightarrow \left| \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \right| \leq \frac{2}{n}$$

Dacă trecem la limită inegalitatea $\left| \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \right| \leq \frac{2}{n}$ pentru $n \rightarrow \infty$, rezultă $\left| \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \right| \rightarrow 0$, de unde avem $\left| \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \right| = 0$ și deci $\frac{A(t)}{A(s)} = \frac{\log t}{\log s}$. Rezultă că $A(t) = \frac{A(s)}{\log s} \log t = K \log t$. Evident, constanta $K = \frac{A(s)}{\log s} = \frac{1}{\log s} \psi_s \left(\frac{1}{s}, \dots, \frac{1}{s} \right)$ nu depinde de t . Variind valorile lui t , rezultă că $A(t) = K \log t$ pentru orice $t \in \mathbb{N}^*$ astfel încât relația (76) are loc.

Observație: Pentru $t \in \mathbb{N}^*$ și $p_1 = \dots = p_t = \frac{1}{t}$, este imediat că $-\sum_i p_i \log p_i = \log t$. Așadar, rezultatul $A(t) = K \log t$ pe care tocmai l-am obținut mai sus implică faptul că egalitatea $\psi_t(p_1, \dots, p_t) = -K \sum_i p_i \log p_i$ are loc pentru cazul $p_1 = \dots = p_t = \frac{1}{t}$.

h. Considerăm o mulțime de N evenimente echiprobabile. Fie $\mathcal{P} = (S_1, S_2, \dots, S_k)$ o partitioare a acestei mulțimi de evenimente. Notăm $p_i = |S_i|/N$.

Propunem *codificarea* din figura alăturată. Vom alege mai întâi S_i , una dintre submulțimile din partitia \mathcal{P} , în funcție de probabilitățile p_1, \dots, p_k . Extragem apoi unul dintre elementele mulțimii S_i , cu probabilitate uniformă.

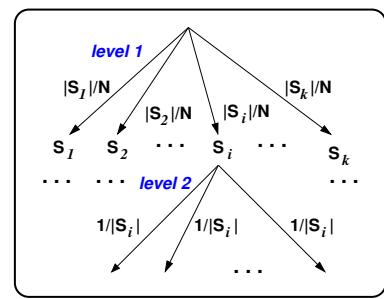
Conform egalității (80), avem $A(N) = K \log N$. Folosind proprietatea A3 și *codificarea* în doi pași propusă mai sus, rezultă că

$$A(N) = \psi_k(p_1, \dots, p_k) + \sum_i p_i A(|S_i|).$$

Așadar,

$$K \log N = \psi_k(p_1, \dots, p_k) + K \sum_i p_i \log |S_i|.$$

¹³¹Putem considera în mod natural $A(1) = 0$ (vedeți proprietatea A0). Conform proprietății A2, urmează că $A(s) > A(1) = 0$ pentru orice $s > 1$.



Prin urmare,

$$\begin{aligned}\psi_k(p_1, \dots, p_k) &= K[\log N - \sum_i p_i \log |S_i|] = K[(\log N) \sum_i p_i - \sum_i p_i \log |S_i|] \\ &= -K \sum_i p_i \log \frac{|S_i|}{N} = -K \sum_i p_i \log p_i.\end{aligned}$$

i.¹³² Considerăm pentru fiecare p_i , $i = 1, \dots, k-1$ un sir de aproximări succesive $\{q_i^{(n)}\}_{n \geq 1} \subset \mathbb{Q}^+$ astfel încât $\lim_{n \rightarrow \infty} q_i^{(n)} = p_i$. Definim de asemenea sirul $\{q_k^{(n)}\}_{n \geq 1} \subset \mathbb{Q}$ astfel: $q_k^{(n)} = 1 - \sum_{i=1}^{k-1} q_i^{(n)}$. Se pot alege aproximările $\{q_i^{(n)}\}$ astfel încât cel puțin de la un loc încolo $q_k^{(n)} \in \mathbb{Q}^+$.¹³³

Trecând la limită în relația de definiție $q_k^{(n)} = 1 - \sum_{i=1}^{k-1} q_i^{(n)}$, obținem:

$$\lim q_k^{(n)} = 1 - \sum_{i=1}^{k-1} \lim q_i^{(n)} = 1 - \sum_{i=1}^{k-1} p_i = p_k.$$

Așadar, avem $\lim_{n \rightarrow \infty} q_i^{(n)} = p_i$, $\forall i = 1, \dots, k$. Deoarece $q_i^{(n)} \in \mathbb{Q}^+$, pentru $i = 1, \dots, k$, cu $k \geq 1$, este valabilă relația

$$\psi_k(q_1^{(n)}, \dots, q_k^{(n)}) = -K \sum_i q_i^{(n)} \log q_i^{(n)}, \quad \forall n \geq 1.$$

Trecând la limită în această relație și ținând cont de continuitatea funcției \log , vom avea:

$$\begin{aligned}\lim_{n \rightarrow \infty} \psi_k(q_1^{(n)}, \dots, q_k^{(n)}) &= \lim_{n \rightarrow \infty} \left(-K \sum_i q_i^{(n)} \log q_i^{(n)} \right) \\ &= -K \sum_i \lim_{n \rightarrow \infty} q_i^{(n)} \log q_i^{(n)} = -K \sum_i p_i \log p_i.\end{aligned}\quad (81)$$

Datorită proprietății A1, funcția ψ_k este continuă în fiecare dintre argumentele ei, astfel că

$$\lim_{n \rightarrow \infty} \psi_k(q_1^{(n)}, \dots, q_k^{(n)}) = \psi_k(\lim_{n \rightarrow \infty} q_1^{(n)}, \dots, \lim_{n \rightarrow \infty} q_k^{(n)}) = \psi_k(p_1, \dots, p_k). \quad (82)$$

Din relațiile (81) și (82) putem concluziona că

$$\psi_k(p_1, \dots, p_k) = -K \sum_i p_i \log p_i, \quad \text{pentru } \forall p_i \in [0, 1], \text{ a.i. } \sum_i p_i = 1.$$

Astfel se atinge obiectivul propus, acela de a arăta că $\psi_k(p_1, \dots, p_k)$ este de forma $-K \sum_i p_i \log p_i$, pentru situațiile în care probabilitățile p_i sunt numere reale.

O observație finală: Egalitatea $A(t) = K \log t$ implică $A(1) = K \log 1 = 0$, ceea ce este în concordanță cu proprietatea A0. În baza proprietății A2 va rezulta $\psi_n(p_1, \dots, p_n) \geq 0$ pentru orice $n \in \mathbb{N}^*$ și orice $p_1, \dots, p_n \in [0, 1]$ astfel încât $\sum_i p_i = 1$.

¹³²Rezolvare redactată de Ștefan Bălăucă, student, FII, 2020.

¹³³De exemplu, putem defini $q_i^{(n)} =$ numărul format luând în considerare doar primele n zecimale ale lui p_i .

63.

(Entropia relativă: definiție și proprietăți elementare; exprimarea câștigului de informație cu ajutorul entropiei relative)

■ □ • prelucrare de Liviu Ciortuz, după CMU, 2007 fall, Carlos Guestrin, HW1, pr. 1.2

Entropia relativă sau *divergența Kullback-Leibler (KL)* a unei distribuții p în raport cu o altă distribuție q — ambele distribuții fiind discrete — se definește astfel:¹³⁴

$$KL(p||q) \stackrel{\text{def.}}{=} - \sum_{x \in X} p(x) \log \frac{q(x)}{p(x)}$$

Din perspectiva teoriei informației, divergența KL specifică numărul de *biti additionali* care sunt necesari în medie pentru a transmite valorile variabilei X atunci când presupunem că aceste valori sunt distribuite conform distribuției („model“) q , dar în realitate ele urmează o altă distribuție, p .¹³⁵

a. Demonstrați inegalitatea $KL(p||q) \geq 0$ și apoi arătați că egalitatea are loc dacă și numai dacă $p = q$.¹³⁶

Indicație:

Pentru a demonstra punctul acesta puteți folosi *inegalitatea lui Jensen*:¹³⁷

Dacă $f : \mathbb{R} \rightarrow \mathbb{R}$ este o *funcție convexă*, atunci pentru orice $a_i \geq 0$, $i = 1, \dots, n$ cu $\sum_i a_i = 1$ și orice $x_i \in \mathbb{R}$, $i = 1, \dots, n$, avem $f(\sum_i a_i x_i) \leq \sum_i a_i f(x_i)$. Dacă f este strict convexă, atunci egalitatea are loc doar dacă $x_1 = \dots = x_n$. Pentru funcții concave, semnul inegalității este \geq .

b. Câștigul de informație poate fi (re)definit ca fiind entropia relativă dintre distribuția comună observată a lui X și Y pe de o parte, și produsul distribuțiilor marginale p_X și p_Y de cealaltă parte:

$$\begin{aligned} IG(X, Y) &\stackrel{\text{def.}}{=} KL(p_{X,Y} || (p_X p_Y)) = - \sum_x \sum_y p_{X,Y}(x, y) \log \left(\frac{p_X(x)p_Y(y)}{p_{X,Y}(x, y)} \right) \\ &\stackrel{\text{not.}}{=} - \sum_x \sum_y p(x, y) \log \left(\frac{p(x)p(y)}{p(x, y)} \right) \end{aligned}$$

Arătați că această nouă definiție a câștigului de informație este echivalentă cu definiția dată anterior (vedeți problema 55). Cu alte cuvinte, demonstrați egalitatea

¹³⁴În învățarea automată, divergența KL este folosită de exemplu la fundamentarea schemei algoritmice EM. Vedeți problemele 1.b și 2 de la capitolul *Schema algoritmică EM*. Mai general, maximizarea verosimilității datelor — în vederea estimării parametrilor distribuțiilor probabiliste — poate fi văzută ca fiind echivalentă cu minimizarea divergenței KL (vedeți problema 147).

¹³⁵Atenție: Divergența KL nu este o măsură de *distanță* între două distribuții probabiliste, fiindcă în general ea nu este simetrică ($KL(p||q) \neq KL(q||p)$) și nici nu satisface inegalitatea triunghiului. Pentru „simetrizare“, se consideră $M(p, q) = \frac{1}{2}(p + q)$, apoi se definește funcția $JSD(p||q) = \frac{1}{2}KL(p||M) + \frac{1}{2}KL(q||M)$, care se numește *divergență Jensen-Shannon*. În sfârșit, se poate arăta că $\sqrt{JSD(p||q)}$ definește o măsură de distanță (metrică), adică este nenegativă, simetrică, implică identitatea indiscernabililor și satisface inegalitatea triunghiului; ea este numită *distanță Jensen-Shannon*.

Variația informației, definită prin

$$VI(X, Y) \stackrel{\text{def.}}{=} H(X, Y) - IG(X, Y) = H(X) + H(Y) - 2IG(X, Y) = H(X | Y) + H(Y | X),$$

este de asemenea o măsură de distanță.

¹³⁶Mai general, $KL(p||q)$ este cu atât mai mică cu cât „asemănarea“ dintre distribuțiile p și q este mai mare.

¹³⁷Vedeți problema 79.

$$KL(p_{X,Y} \parallel (p_X p_Y)) = H[X] - H[X | Y].$$

Observație: Din noua definiție introdusă mai sus pentru câștigul de informație, rezultă imediat că

$$\begin{aligned} IG(X, Y) &= \sum_y p(y) \sum_x p(x | y) \log \frac{p(x | y)}{p(x)} = \sum_y p(y) KL(p_{X|Y} \parallel p_X) \\ &= E_Y [KL(p_{X|Y} \parallel p_X)] \end{aligned}$$

ceea ce înseamnă că $IG(X, Y)$ poate fi văzută ca o medie (în raport cu distribuția lui Y) a divergenței KL dintre distribuția condițională a lui X în raport cu Y pe de o parte, și distribuția lui X pe de altă parte.

c. O consecință imediată a punctelor a și b este faptul că $IG(X, Y) \geq 0$ (deci $H(X) \geq H(X|Y)$ și $H(Y) \geq H(Y|X)$) pentru orice variabile aleatoare discrete X și Y . Folosind din nou rezultatele de la punctele a și b , arătați că $IG(X, Y) = 0$ dacă și numai dacă X și Y sunt independente.

Răspuns:

a. Vom dovedi inegalitatea $KL(p||q) \geq 0$ folosind inegalitatea lui Jensen, în expresia căreia vom înlocui f cu funcția convexă $-\log_2$, pe a_i cu $p(x_i)$ și pe x_i cu $\frac{q(x_i)}{p(x_i)}$. (Pentru conveniență, în cele ce urmează vom renunța la indicele variabilei x .) Vom avea:

$$\begin{aligned} KL(p \parallel q) &\stackrel{\text{def.}}{=} -\sum_x p(x) \log \frac{q(x)}{p(x)} \\ &\stackrel{\text{Jensen}}{\geq} -\log \left(\sum_x p(x) \frac{q(x)}{p(x)} \right) = -\log \underbrace{\left(\sum_x q(x) \right)}_1 = -\log 1 = 0. \end{aligned}$$

Așadar, $KL(p \parallel q) \geq 0$, oricare ar fi distribuțiile (discrete) p și q .

Vom demonstra acum că $KL(p||q) = 0 \Leftrightarrow p = q$.

Egalitatea $p(x) = q(x)$ implică $\frac{q(x)}{p(x)} = 1$, deci $\log \frac{q(x)}{p(x)} = 0$ pentru orice x , de unde rezultă imediat $KL(p||q) = 0$.

Pentru a demonstra implicația inversă, se ține cont că în inegalitatea lui Jensen, în cazul funcțiilor strict convexe (cum este $-\log_2$) are loc egalitatea doar în cazul în care $x_i = x_j$ pentru orice i și j . În cazul de față, această condiție se traduce prin faptul că raportul $\frac{q(x)}{p(x)}$ este același (α) pentru orice valoare

a lui x . Tinând cont că $\sum_x p(x) = 1$ și $\sum_x q(x) = \sum_x p(x) \frac{q(x)}{p(x)} = 1$, rezultă că $\alpha = \frac{q(x)}{p(x)} = 1$, deci $p(x) = q(x)$ pentru orice x , ceea ce înseamnă că distribuțiile p și q sunt identice.

b. Vom folosi regula de înmulțire, conform căreia $p(x, y) = p(x | y)p(y)$:

$$\begin{aligned}
 KL(p_{X,Y} || (p_X p_Y)) &\stackrel{\text{def. } KL}{=} -\sum_x \sum_y p(x, y) \log \left(\frac{p(x)p(y)}{p(x, y)} \right) \\
 &= -\sum_x \sum_y p(x, y) \log \left(\frac{p(x)}{p(x | y)} \right) = -\sum_x \sum_y p(x, y) [\log p(x) - \log p(x | y)] \\
 &= -\sum_x \sum_y p(x, y) \log p(x) - \left(-\sum_x \sum_y p(x, y) \log p(x | y) \right) \\
 &\stackrel{pr. 55.b}{=} -\sum_x \log p(x) \underbrace{\sum_y p(x, y)}_{=p(x)} - H[X | Y] = \sum_x p(x) \log p(x) - H[X | Y] \\
 &= H[X] - H[X | Y] = IG(X, Y)
 \end{aligned}$$

c. Conform punctului b, egalitatea $IG(X, Y) = 0$ este echivalentă cu egalitatea $KL(p_{X,Y} || p_X p_Y) = 0$. Conform punctului a, această a doua relație este adevărată dacă și numai dacă distribuțiile $p_{X,Y}$ și $p_X p_Y$ sunt identice, or aceasta este exact definiția independenței variabilelor X și Y .

64. (Cross-entropie: definiție, o proprietate (nenegativitatea) și un exemplu simplu de calculare a valorii cross-entropiei)
■ □ ● ○ CMU, 2011 spring, Roni Rosenfeld, HW2, pr. 3.c

Cross-entropia a două distribuții p și q , desemnată prin $CH(p, q)$, reprezintă numărul mediu de biți necesari pentru a codifica un eveniment dintr-o mulțime oarecare de posibilități, atunci când schema de cod[ific]are folosită se bazează pe o distribuție de probabilitate dată q , în loc să se bazeze pe distribuția „adevărată“ p . În cazul în care distribuțiile p și q sunt discrete, această noțiune se definește formal astfel:¹³⁸

$$CH(p, q) = -\sum_x p(x) \log q(x).$$

În cazul distribuțiilor continue, definiția se obține / construiește prin analogie:

$$CH(p, q) = -\int_X p(x) \log q(x) dx.$$

Observație: Tinând cont de definiția entropiei relative (cunoscută și sub numele de divergență Kullback-Leibler), vedeti pr. 63, putem scrie:

$$KL(p||q) = CH(p, q) - H(p).$$

Cross-entropia — ca și entropia relativă; vedeti problema 63 —, spre deosebire de entropia comună, nu este simetrică în raport cu cele două distribuții / argumente: în general, $CH(p, q) \neq CH(q, p)$.

¹³⁸Pentru exemple de folosire a cross-entropiei în învățarea automată, vedeti problema 15 de la capitolul *Rețele neuronale artificiale*, precum și problema 2 de la capitolul *Schema algoritmică EM*.

a. Poate oare cross-entropia să ia valori negative? Faceți o demonstrație sau dați un contraexemplu.

b. În multe experimente, pentru a stabili calitatea diferitelor ipoteze / modele, se procedează la evaluarea / compararea lor pe un set de date. Să presupunem că, urmărind să faci predicția *funcției de probabilitate* asociate unei anumite *variabile aleatoare* care are 7 valori posibile, ai obținut (printr-un procedeu oarecare) două *modele* diferite, iar *distribuțiile de probabilitate* prezise de către aceste două modele sunt respectiv:

$$q_1 = \left(\frac{1}{10}, \frac{1}{10}, \frac{1}{5}, \frac{3}{10}, \frac{1}{5}, \frac{1}{20}, \frac{1}{20} \right) \text{ și } q_2 = \left(\frac{1}{20}, \frac{1}{10}, \frac{3}{20}, \frac{7}{20}, \frac{1}{5}, \frac{1}{10}, \frac{1}{20} \right).$$

Să zicem că pentru evaluare folosești un set de date caracterizat de următoarea distribuție *empirică*:

$$p_{\text{empiric}} = \left(\frac{1}{20}, \frac{1}{10}, \frac{1}{5}, \frac{3}{10}, \frac{1}{5}, \frac{1}{10}, \frac{1}{20} \right).$$

Calculează cross-entropiile $CH(p_{\text{empiric}}, q_1)$ și $CH(p_{\text{empiric}}, q_2)$.

Care dintre aceste două modele va conduce la o cross-entropie mai mică? Putem oare garanta că acest model este într-adevăr [cel] mai bun? Explică / justifică răspunsul [pe care l-ați dat].

Răspuns:

a. Nu, cross-entropia nu poate lua valori negative. Iată cum demonstrăm:

Stim că pentru orice funcții de probabilitate p și q și pentru orice x (care aparține domeniului de valori al unei variabile aleatoare care are o astfel de distribuție de probabilitate), valorile $p(x)$ și $q(x)$ satisfac inegalitățile $0 \leq p(x) \leq 1$ and $0 \leq q(x) \leq 1$. Inegalitatea $q(x) \leq 1$ implică faptul că $\log q(x) \leq 0$. Din $0 \leq p(x)$ și $-\log q(x) \geq 0$, rezultă că $0 \leq -p(x) \log q(x)$. În consecință, suma tuturor acestor termeni va fi de asemenea mai mare sau egală cu 0, deci cross-entropia nu poate fi niciodată negativă.

Observație importantă: Spre deosebire de entropie (vedeți problema 141), cross-entropia nu este mărginită superior. Ea poate crește la infinit; vedeți cazul când pentru o anumită valoare x sunt adevărate simultan relațiile $p(x) \neq 0$ și $q(x) = 0$.¹³⁹

b. Facem calculele, folosind formula cross-entropiei:

$$\begin{aligned} CH(p_{\text{empiric}}, q_1) &= \\ &- \left(\frac{1}{20} \log_2 \frac{1}{10} + \frac{1}{10} \log_2 \frac{1}{10} + \frac{1}{5} \log_2 \frac{1}{5} + \frac{3}{10} \log_2 \frac{3}{10} + \frac{1}{5} \log_2 \frac{1}{5} + \frac{1}{10} \log_2 \frac{1}{20} \right. \\ &\quad \left. + \frac{1}{20} \log_2 \frac{1}{20} \right) = \frac{3}{20} \log_2 10 + \frac{2}{5} \log_2 5 + \frac{3}{10} \log_2 \frac{10}{3} + \frac{3}{20} \log_2 20 = \\ &= \frac{3}{20} \log_2 2 \cdot 5 + \frac{2}{5} \log_2 5 + \frac{3}{10} \log_2 \frac{2 \cdot 5}{3} + \frac{3}{20} \log_2 2^2 \cdot 5 \\ &= \left(\frac{3}{20} + \frac{3}{10} + 2 \cdot \frac{3}{20} \right) + \left(\frac{3}{20} + \frac{2}{5} + \frac{3}{10} + \frac{3}{20} \right) \log_2 5 - \frac{3}{10} \log_2 3 \\ &= \frac{3}{4} + \log_2 5 - \frac{3}{10} \log_2 3 = 2.596439345 \text{ biți} \end{aligned}$$

¹³⁹Mai precis, $\lim_{q(x) \rightarrow +0} (-p(x) \cdot \log_2 q(x)) = -p(x)(-\infty) = +\infty$.

$$\begin{aligned}
CH(p_{empiric}, q_2) &= \\
&- \left(\frac{1}{20} \log_2 \frac{1}{20} + \frac{1}{10} \log_2 \frac{1}{10} + \frac{1}{5} \log_2 \frac{3}{20} + \frac{3}{10} \log_2 \frac{7}{20} + \frac{1}{5} \log_2 \frac{1}{5} + \frac{1}{10} \log_2 \frac{1}{10} \right. \\
&\quad \left. + \frac{1}{20} \log_2 \frac{1}{20} \right) = \\
&= \frac{1}{10} \log_2 20 + \frac{1}{5} \log_2 10 + \frac{1}{5} \log_2 \frac{20}{3} + \frac{3}{10} \log_2 \frac{20}{7} + \frac{1}{5} \log_2 5 \\
&= \frac{1}{10} \log_2 2^2 \cdot 5 + \frac{1}{5} \log_2 2 \cdot 5 + \frac{1}{5} \log_2 \frac{2^2 \cdot 5}{3} + \frac{3}{10} \log_2 \frac{2^2 \cdot 5}{7} + \frac{1}{5} \log_2 5 \\
&= \left(2 \cdot \frac{1}{10} + \frac{1}{5} + 2 \cdot \frac{1}{5} + 2 \cdot \frac{3}{10} \right) + \left(\frac{1}{10} + 3 \cdot \frac{1}{5} + \frac{3}{10} \right) \log_2 5 - \frac{1}{5} \log_2 3 - \frac{3}{10} \log_2 7 \\
&= \frac{7}{5} + \log_2 5 - \frac{1}{5} \log_2 3 - \frac{3}{10} \log_2 7 = 2.562729118 \text{ biti}.
\end{aligned}$$

Se observă că distribuția $p_{empiric}$ are o cross-entropie mai mică în [raport cu] modelul q_2 . Este deci rezonabil să afirmăm că alegerea modelului q_2 este mai bună.

Totuși, nu putem garanta că acest model este întotdeauna cel mai bun, fiindcă aici lucrăm cu o distribuție „empirică“, iar distribuția „adevărată“ nu neapărat se reflectă în mod complet / perfect în această distribuție empirică.

De obicei, *bias-ul de eşantionare* (engl., sampling bias), precum și *insuficiența datelor de antrenament* vor contribui la lărgirea „spațiului“ care diferențiază distribuția adevărată de distribuția empirică. Prin urmare, în practică, atunci când concepem un [astfel de] experiment de evaluare a mai multor distribuții probabiliste, trebuie să avem permanent în minte faptul acesta și, dacă este posibil, să folosim tehnici care reduc / minimizează aceste riscuri.

65.

(Inegalitatea lui Gibbs:¹⁴⁰ un caz particular; comparație între valorile entropiei și ale cross-entropiei)

Liviu Ciortuz, 2012, după www.en.wikipedia.org și CMU, T. Mitchell, A. Moore, 2003 fall, HW1, pr. 1.2

Fie $P = \{p_1, \dots, p_n\}$ o distribuție de probabilitate discretă.

a. Arătați că pentru orice distribuție de probabilitate $Q = \{q_1, \dots, q_n\}$ are loc inegalitatea:

$$-\sum_{i=1}^n p_i \log_2 p_i \leq -\sum_{i=1}^n p_i \log_2 q_i$$

Altfel spus, $H(P) \leq CH(P, Q)$, unde $H(P)$ este entropia distribuției P , iar $CH(P, Q)$ este *cross-entropia* lui P în raport cu Q .

Observație: Pentru un exemplu de utilizare a acestei inegalități, cunoscută sub numele de *inegalitatea lui Gibbs*, în învățarea automată, puteți vedea problema 2 de la capitolul *Schema algoritmică EM*.

¹⁴⁰Josiah Willard Gibbs (1839-1903), om de știință american, a avut contribuții teoretice majore în fizică, chimie și matematică. A fost unul dintre părinții fondatori ai domeniului mecanicii statistice. A inventat teoria modernă a calculului vectorial. A predat la universitatea Yale în perioada 1871-1903.

b. Arătați că în formula de mai sus egalitatea are loc dacă și numai dacă $p_i = q_i$ pentru $i = 1, \dots, n$.

Observație: În formula din enunț, în locul bazei 2 pentru logaritm poate fi folosită orice bază supraunitară.

Indicații:

1. Dacă în inegalitatea dată se trece termenul din partea stângă în partea dreaptă, obținem $0 \leq \sum_{i=1}^n p_i \log_2 p_i - \sum_{i=1}^n p_i \log_2 q_i \Leftrightarrow 0 \leq -\sum_{i=1}^n p_i \log_2 \frac{q_i}{p_i}$. Puteți face legătura dintre expresia din partea dreaptă a acestei ultime inegalități și definiția *entropiei relative* (numită de asemenea *divergența Kullback-Leibler*; vedeti problema 63) și apoi să folosiți proprietățile entropiei relative.
2. Pentru a demonstra într-o manieră directă inegalitatea lui Gibbs, puteți folosi inegalitatea lui Jensen (vedeti problema 79).

Răspuns:

Expresia $-\sum_{i=1}^n p_i \log_2 \frac{q_i}{p_i}$ la care s-a ajuns în *Indicație* este exact divergența Kullback-Leibler dintre distribuțiile P și Q . Formal, scriem acest lucru astfel: $KL(P||Q) = -\sum_{i=1}^n p_i \log_2 \frac{q_i}{p_i} = CH(P, Q) - H(P)$.

a. La problema 63.a am demonstrat inegalitatea $KL(P||Q) \geq 0$, care are loc pentru orice distribuții probabiliste discrete P și Q . Aceasta este exact proprietatea de care avem nevoie pentru a justifica inegalitatea dată în enunț la acest punct ($H(P) \leq CH(P, Q)$).

Inegalitatea lui Gibbs se poate demonstra în mod direct folosind inegalitatea lui Jensen în versiune probabilistă (cf. problemei 79.c): dacă φ este o funcție convexă, iar X este variabilă aleatoare de distribuție p , atunci $E[\varphi(X)] \geq \varphi(E[X])$. Particularizând această inegalitate pentru $\varphi = -\log_a$, cu $a > 1$, și înlocuind X cu $\frac{q(x)}{p(x)}$, vom avea:

$$\begin{aligned} E\left[-\log_a \frac{q(x)}{p(x)}\right] &\geq -\log_a E\left[\frac{q(x)}{p(x)}\right] \Leftrightarrow -\sum_i p(x_i) \log_a \frac{q(x_i)}{p(x_i)} \geq -\log_a \underbrace{\left(\sum_i p(x_i) \cdot \frac{q(x_i)}{p(x_i)}\right)}_1 \\ &\Leftrightarrow -\sum_i p(x_i) \log_a \frac{q(x_i)}{p(x_i)} \geq 0 \Leftrightarrow -\sum_i p(x_i) \log_a q(x_i) \geq -\sum_i p(x_i) \log_a p(x_i) \\ &\stackrel{not.}{\Leftrightarrow} -\sum_i p_i \log_a q_i \geq -\sum_i p_i \log_a p_i. \end{aligned}$$

b. Tot la problema 63.a s-a demonstrat că $KL(P||Q)$ are valoarea 0 dacă și numai dacă distribuțiile P și Q sunt identice. În contextul nostru, această proprietate se transpune imediat sub forma $H(P) = CH(P, Q) \Leftrightarrow p_i = q_i$ pentru $i = 1, \dots, n$.

Alternativ, vom folosi *Observația 1* de la problema 79, care afirmă următoarele: în cazul unei funcții strict convexe inegalitatea lui Jensen se realizează egalitatea dacă și numai dacă $x_1 = \dots = x_n$. În cazul nostru, rezultă că în inegalitatea lui Gibbs se realizează egalitatea dacă și numai dacă $\frac{p_i}{q_i} = \alpha$ (constant), pentru orice i . Cum $\sum_i p_i = \sum_i q_i = 1$, rezultă imediat că $\alpha = 1$, deci distribuțiile probabiliste P și Q sunt identice.

66. (Entropia văzută ca o *funcțională* în raport cu p (p.d.f. sau p.m.f); calculul *derivatei funcționale* a entropiei în raport cu p)

□ *Liviu Ciortuz, 2021, pornind de la CMU, 2013 spring, A. Smola, B. Poczos, HW2, pr. 1 și https://en.wikipedia.org/wiki/Functional_derivative*

Comentariu:

Stim că entropia unei variabile aleatoare X continuă se definește după cum urmează: dacă funcția de densitate de probabilitate (p.d.f.) a lui X este p , atunci $H(X) = \int p(x) \log_2 p(x)$.¹⁴¹ Abstractizând, noțiunea de entropie H poate fi interpretată [și] ca o funcție care ia ca argument funcția p . În matematică, astfel de „funcții de funcții“ — acestea din urmă îndeplinind anumite proprietăți, lăsate nespecificate aici — se numesc *funcționale*.

Ca și pentru funcțiile de variabilă reală, și în cazul funcțiilor se poate defini o noțiune de tip derivată, numită *derivată funcțională*.

Fie F o funcțională definită peste mulțimea de funcții M . Considerăm f și h din M , iar $t \in \mathbb{R}$. Noțiunea de *derivată funcțională* se definește pornind de la calculul următoarei expresii, care reprezintă valoarea unei *derivate numerice* — adică, definită în sensul clasic — în punctul $t = 0$:

$$\left[\frac{d}{dt} F(f + th) \right]_{t=0} \stackrel{\text{def.}}{=} \left[\lim_{t \rightarrow 0} \frac{F(f + th) - F(f)}{t} \right]_{t=0} \quad (83)$$

Dacă această expresie se poate calcula ca integrala unui produs de două funcții, dintre care una este h , adică

$$\int \left[\frac{\delta F(f)}{\delta f}(x) \right] h(x) dx, \quad (84)$$

atunci funcția $\frac{\delta F(f)}{\delta f}(x)$ este considerată *derivata funcțională* a lui F în raport cu funcția f .

Derivata funcțională $\frac{\delta F(f)}{\delta f}$ poate fi interpretată ca fiind gradientul lui F în „punctul“ f . Integrala $\int \frac{\delta F(f)}{\delta f}(x) h(x) dx$ poate fi interpretată ca fiind *derivata direcțională* a lui F în „punctul“ f pe direcția lui h (care poate fi văzut ca un vector infinit, ale cărui componente sunt valorile $h(x)$, pentru $x \in \mathbb{R}$).

*Exemplu:*¹⁴² Considerăm $F(f) \stackrel{\text{def.}}{=} \int f^2(x) dx$. Atunci,

$$\begin{aligned} F(f + th) &= \int (f + th)^2(x) dx = \int f^2(x) dx + 2t \int f(x) h(x) dx + t^2 \int h^2(x) dx \\ \Rightarrow \frac{dF(f + th)}{dt} &= \lim_{t \rightarrow 0} \frac{F(f + th) - F(f)}{t} = 2 \int f(x) h(x) dx + 2t \int h^2(x) dx \\ \Rightarrow \frac{dF(f + th)}{dt} \Big|_{t=0} &= 2 \int f(x) h(x) dx = \int 2f(x) h(x) dx \end{aligned}$$

¹⁴¹Dacă X este variabilă aleatoare discretă, iar funcția sa masă de probabilitate (p.m.f.) este p , atunci $H(X) = -\sum_{x \in Val(X)} p(x) \log_2 p(x)$.

¹⁴²Acest exemplu a fost preluat din cartea *A modern approach to functional integration*, 2010, de J. Klauder, pag. 38-39.

$$\Rightarrow \frac{\delta F(f)}{\delta f}(x) = 2f(x).$$

Arătați că derivata funcțională a entropiei (H) unei variabile aleatoare continue X având p.d.f. p este

$$\frac{\delta H(p)}{\delta p}(x) = -\frac{1}{\ln 2}[1 + \ln p(x)].$$

Răspuns:

Pentru simplitate, vom considera funcționala $F(p) \stackrel{\text{def.}}{=} \int p(x) \ln p(x) dx$. Așadar, $H(p) = -\frac{1}{\ln 2}F(p)$, de unde va rezulta $\frac{\delta H(p)}{\delta p}(x) = -\frac{1}{\ln 2}\frac{\delta F(p)}{\delta p}(x)$. Vom arăta că $\frac{\delta F(p)}{\delta p}(x) = 1 + \ln p(x)$. Pentru aceasta, vom porni de la definiția (83), mai precis de la expresia din dreapta egalității respective și vom explicita mai întâi $F(p + th)$:

$$\begin{aligned} F(p + th) &= \int (p + th)(x) \cdot \ln(p + th)(x) dx = \int (p(x) + th(x)) \cdot \ln(p(x) + th(x)) dx \\ &= \int p(x) \ln(p(x) + th(x)) dx + t \int h(x) \ln(p(x) + th(x)) dx. \end{aligned}$$

Prin urmare,

$$\begin{aligned} \frac{dF(p + th)}{dt} &\stackrel{\text{def.}}{=} \lim_{t \rightarrow 0} \frac{F(p + th) - F(p)}{t} \\ &= \lim_{t \rightarrow 0} \left[\frac{1}{t} \left\{ \int p(x) \ln(p(x) + th(x)) dx - \int p(x) \ln p(x) dx \right\} \right. \\ &\quad \left. + \int h(x) \ln(p(x) + th(x)) dx \right] \\ &= \lim_{t \rightarrow 0} \left[\frac{1}{t} \left\{ \int p(x) (\ln(p(x) + th(x)) - \ln p(x)) dx \right\} \right] + \\ &\quad \lim_{t \rightarrow 0} \int h(x) \ln(p(x) + th(x)) dx. \end{aligned} \tag{85}$$

Vom calcula acum aceste (ultime) două limite. Vom începe cu cea de-a două limită, fiindcă este mai simplă.

$$\lim_{t \rightarrow 0} \int h(x) \ln(p(x) + th(x)) dx = \int \lim_{t \rightarrow 0} h(x) \ln(p(x) + th(x)) dx \tag{86}$$

$$= \int h(x) \ln(p(x)) dx. \tag{87}$$

Observați că am putut interschimba simbolii $\lim_{t \rightarrow 0}$ și \int , pentru că integrala aceasta este definită, deci reprezintă de fapt tot o limită. (Aceste două limite se calculează folosind variabile diferite, t și x .)

Revenim acum la prima limită de mai sus:

$$\begin{aligned} &\lim_{t \rightarrow 0} \left[\frac{1}{t} \left\{ \int p(x) (\ln(p(x) + th(x)) - \ln p(x)) dx \right\} \right] \\ &= \lim_{t \rightarrow 0} \int \frac{1}{t} p(x) (\ln(p(x) + th(x)) - \ln p(x)) dx \end{aligned}$$

$$= \int \lim_{t \rightarrow 0} \frac{1}{t} p(x)(\ln(p(x) + th(x)) - \ln p(x)) dx \quad (88)$$

$$= \int \left[\frac{d}{dt} p(x) \ln(p(x) + th(x)) \right]_{t=0} dx \quad (89)$$

$$= \int \left[p(x) \cdot \frac{1}{p(x) + th(x)} \cdot h(x) \right]_{t=0} dx = \int h(x) dx. \quad (90)$$

Observați că la egalitatea (88) am folosit aceeași justificare ca și pentru egalitatea (86). La egalitatea (89) am folosit definiția derivatei [numerice] pentru funcția $p(x) \ln(p(x) + th(x))$ în raport cu t , calculată în punctul $t = 0$.

Acum, înlocuind rezultatele (87) și (90) în relația (85), obținem:

$$\frac{dF(p + th)}{dt} = \int h(x) dx + \int h(x) \ln(p(x)) dx = \int (1 + \ln p(x)) h(x) dx, \quad (91)$$

rezultat care nu depinde de t (deci nu mai este nevoie să aplicăm operatorul $[]_{t=0}$). Așadar, din relația (91) rezultă că $\frac{\delta F(p)}{\delta p}(x) = 1 + \ln p(x)$ și, în consecință, $\frac{\delta H(p)}{\delta p}(x) = -\frac{1}{\ln 2}(1 + \ln p(x))$.

Observație: Vom demonstra că formula din enunț este valabilă și pentru cazul discret.¹⁴³ Așadar, vom folosi definiția $H[p(x)] = -\sum_x p(x) \log_2 p(x)$, unde x ia valori într-o mulțime discretă. De data aceasta, este mai convenabil să pornim de la membrul stâng al egalității (83), care a fost dată pentru a putea introduce noțiunea de derivată funcțională.

$$\begin{aligned} \left[\frac{d}{dt} H[p(x) + th(x)] \right]_{t=0} &= - \left[\frac{d}{dt} \sum_x (p(x) + th(x)) \log_2(p(x) + th(x)) \right]_{t=0} \\ &= -\frac{1}{\ln 2} \left[\frac{d}{dt} \sum_x (p(x) + th(x)) \ln(p(x) + th(x)) \right]_{t=0} \\ &= -\frac{1}{\ln 2} \left[\sum_x \frac{d}{dt} (p(x) + th(x)) \ln(p(x) + th(x)) \right]_{t=0} \\ &= -\frac{1}{\ln 2} \left[\sum_x [h(x) \ln(p(x) + th(x)) + \cancel{(p(x) + th(x))} \frac{h(x)}{\cancel{p(x) + th(x)}}] \right]_{t=0} \\ &= -\frac{1}{\ln 2} \sum_x [h(x) \ln p(x) + h(x)] = -\frac{1}{\ln 2} \sum_x [1 + \ln p(x)] h(x) \\ &= \sum_x -\frac{1}{\ln 2} [1 + \ln p(x)] h(x). \end{aligned} \quad (92)$$

În concluzie, $\frac{\delta H(p)}{\delta p}(x) = -\frac{1}{\ln 2}(1 + \ln p(x))$.¹⁴⁴ Rezultatul acesta este identic (d.p.v. sintactic) cu cel din cazul continuu.

¹⁴³Prin urmare, vom arăta [și] cum se poate defini — și calculă — derivata funcțională în raport cu o funcție care nu este continuă.

¹⁴⁴Observați că semnului de integrare $\int dx$ din relația de definiție (84) îi corespunde în relația (92) simbolul de sumare \sum_x , așa cum este natural la trecerea de la cazul continuu la cazul discret. (Într-o versiune viitoare a culegerii, vom face probabil legătura în sens invers, adică vom prezenta definiția noțiunii de derivată funcțională mai întâi pentru cazul discret și apoi pentru cazul continuu.)

0.1.6 Funcții-nucleu

67.

(Găsirea mapării care corespunde unei funcții-nucleu polinomiale particulare)

University of Utah, 2008 fall, Hal Daumé III, HW5, pr. 1.2

Considerăm $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ o funcție polinomială de gradul al doilea: $K(x, z) = (1 + x \cdot z)^2$. Scrieți forma detaliată a acestei funcții pentru cazul 3-dimensional (adică $x = (x_1, x_2, x_3)$ și $z = (z_1, z_2, z_3)$).

Arătați că funcția K satisface condițiile din *definiția funcției-nucleu*, adică există o funcție $\phi : \mathbb{R}^3 \rightarrow \mathbb{R}^n$ cu n ales convenabil, astfel încât $K(x, z) = \phi(x) \cdot \phi(z)$.

Răspuns:

Produsul scalar $x \cdot z$ se scrie desfășurat $x_1 z_1 + x_2 z_2 + x_3 z_3$, deci

$$\begin{aligned} K(x, z) = & 1 + x_1^2 z_1^2 + x_2^2 z_2^2 + x_3^2 z_3^2 + 2x_1 z_1 + 2x_2 z_2 + 2x_3 z_3 + \\ & + 2x_1 x_2 z_1 z_2 + 2x_1 x_3 z_1 z_3 + 2x_2 x_3 z_2 z_3. \end{aligned}$$

Conform *definiției* funcției-nucleu, va trebui să obținem forma analitică a mapării $\phi : \mathbb{R}^3 \rightarrow \mathbb{R}^n$ (unde n va fi stabilit îndată) astfel încât $K(x, z) = \phi(x) \cdot \phi(z)$. Din expresia lui $K(x, z)$ de mai sus este evident că dacă definim

$$\phi(x) = (1, x_1^2, x_2^2, x_3^2, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_3, \sqrt{2}x_1 x_2, \sqrt{2}x_1 x_3, \sqrt{2}x_2 x_3)$$

atunci rezultă că $K(x, z) = \phi(x) \cdot \phi(z)$. Din forma analitică a lui ϕ rezultă că $n = 10$.

68.

(Matrice-nucleu: teorema lui Mercer — condiții necesare [și suficiente] pentru ca o funcție de două variabile din \mathbb{R}^d să fie funcție-nucleu; o proprietate de tip „construcție“ de noi funcții-nucleu)

• o CMU, 2008 fall, Eric Xing, final exam, pr. 2

a. Demonstrați că orice funcție-nucleu K este simetrică, adică pentru orice elemente x_1 și x_2 din domeniul de definiție al lui K , are loc egalitatea $K(x_1, x_2) = K(x_2, x_1)$.

b. Considerăm un set de instanțe $x_1, \dots, x_m \in \mathbb{R}^d$ și o funcție-nucleu $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$. Fie A matricea de tip $m \times m$ având elementele $A(i, j) \stackrel{\text{def}}{=} K(x_i, x_j)$, pentru $i, j = 1, \dots, m$. Matricea A astfel definită se numește *matricea-nucleu* asociată funcției-nucleu K .¹⁴⁵

Prin *definiție*, se spune că o matrice M de tip $m \times m$ este *pozitiv semidefinită* (sau: *nenegativ definită*) dacă pentru orice vector m -dimensional f (văzut ca vector-colonă), are loc inegalitatea $f^\top M f \geq 0$.¹⁴⁶

¹⁴⁵ Matricea-nucleu mai este numită și *matrice Gram*. Notiunea de matrice Gram este definită de unii autori independent de notiunea de funcție-nucleu: date fiind elementele $z_1, \dots, z_m \in \mathbb{R}^m$, matricea Gram este matricea constituită din produsele scalare $z_i \cdot z_j$, cu $i, j = 1, m$.

¹⁴⁶ M este *matrice pozitiv definită* dacă $f^\top M f > 0$ pentru orice f . Similar se definesc și notiunile de matrice negativ semidefinită și matrice negativ definită.

Demonstrați că orice matrice-nucleu este pozitiv semidefinită.

Indicație: Dacă socotiți că este mai ușor pentru dumneavoastră, demonstrați această afirmație în cazul particular al funcției-nucleu $K(x_i, x_j) = (1 + x_i \cdot x_j)^2$, cu x_i și x_j din \mathbb{R}^2 , văzuți ca vectori-coloană.

Observație: Punctele a și b de mai sus arată că *orice matrice-nucleu este simetrică și pozitiv semidefinită*. Se poate arăta — dar nu demonstrăm aici efectiv — că este adevărată și reciproca acestei afirmații, și anume:

Data fiind funcția $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$, dacă pentru orice număr $m \in \mathbb{N}$ și pentru orice elemente $x_1, \dots, x_m \in \mathbb{R}^d$, matricea A de tip $m \times m$, având elementele $A(i, j) \stackrel{\text{def.}}{=} K(x_i, x_j)$ pentru $i, j \in \{1, \dots, m\}$ este simetrică și pozitiv semidefinită, atunci K este funcție-nucleu.¹⁴⁷

c. Folosind *observația* de mai sus, demonstrați că *suma a două funcții-nucleu oarecare K_1 și K_2 este de asemenea o funcție-nucleu*.

Răspuns:

a. Dată fiind o funcție-nucleu oarecare $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$, conform definiției există o altă funcție $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^n$ (numită în general *[funcție de] mapare a trăsăturilor*) astfel încât $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$, pentru orice x_i și x_j .¹⁴⁸ Datorită comutativității produsului scalar vom avea $K(x_i, x_j) \stackrel{\text{def.}}{=} \phi(x_i) \cdot \phi(x_j) = \phi(x_j) \cdot \phi(x_i) \stackrel{\text{def.}}{=} K(x_j, x_i)$ pentru orice x_i și x_j . Așadar, funcția-nucleu K este simetrică.

b. Scriind vectorul-coloană f și matricea-nucleu A pe componente,

$$f = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix} \text{ și } A = \begin{bmatrix} \phi(x_1) \cdot \phi(x_1) & \phi(x_1) \cdot \phi(x_2) & \dots & \phi(x_1) \cdot \phi(x_m) \\ \phi(x_2) \cdot \phi(x_1) & \phi(x_2) \cdot \phi(x_2) & \dots & \phi(x_2) \cdot \phi(x_m) \\ \vdots & \vdots & \ddots & \vdots \\ \phi(x_m) \cdot \phi(x_1) & \phi(x_m) \cdot \phi(x_2) & \dots & \phi(x_m) \cdot \phi(x_m) \end{bmatrix},$$

vom avea:

$$\begin{aligned} f^\top A f &= (f^\top A) f = \left[\left(\sum_{i=1}^m f_i \phi(x_i) \right) \cdot \phi(x_1), \dots, \left(\sum_{i=1}^m f_i \phi(x_i) \right) \cdot \phi(x_m) \right] \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix} \\ &= \left(\sum_{i=1}^m f_i \phi(x_i) \right) \cdot (f_1 \phi(x_1)) + \dots + \left(\sum_{i=1}^m f_i \phi(x_i) \right) \cdot (f_m \phi(x_m)) \\ &= \left(\sum_{i=1}^m f_i \phi(x_i) \right) \cdot \left(\sum_{j=1}^m f_j \phi(x_j) \right) = \left(\sum_{i=1}^m f_i \phi(x_i) \right)^2 \\ &\stackrel{\text{def.}}{=} \left\| \sum_{i=1}^m f_i \phi(x_i) \right\|^2 \geq 0. \end{aligned}$$

¹⁴⁷ Afirmația directă și reciproca ei constituie împreună *teorema lui Mercer* (în variantă discretă): $K : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}$ este funcție-nucleu dacă și numai dacă pentru $\forall m \in \mathbb{N}$ și $\forall x_1, \dots, x_m \in \mathbb{R}^d$, matricea A de tip $m \times m$ având elementele $A(i, j) \stackrel{\text{def.}}{=} K(x_i, x_j)$ este simetrică și pozitiv semidefinită (adică, pentru orice $f \in \mathbb{R}^m$ văzut ca vector-coloană, avem $f^\top A f \geq 0$).

Acest rezultat a fost publicat de către matematicianul englez James Mercer (1883-1932) în anul 1909 în articolul *Functions of positive and negative type and their connection with the theory of integral equations*, în revista *Philosophical Transactions of the Royal Society, London, series A (Mathematical, Physical and Engineering Sciences)*.

¹⁴⁸În general, $n > d$ sau chiar $n \gg d$, ultima notație însemnând că n este mult mai mare decât d .

Observație: Dacă vom nota cu Φ matricea care are coloanele $\phi(x_1), \dots, \phi(x_m)$ — cu un ușor abuz de notație, putem scrie $\Phi = [\phi(x_1), \dots, \phi(x_m)]$ —, matricea-nucleu A se va scrie simplu ca un produs, $A = \Phi^\top \Phi$, iar calculul de mai sus se poate exprima foarte succint astfel:

$$f^\top A f = f^\top (\Phi^\top \Phi) f = (f^\top \Phi^\top)(\Phi f) = (\Phi f)^\top (\Phi f) = \|\Phi f\|^2 \geq 0.$$

În acest calcul am ținut cont de asociativitatea înmulțirii matricelor, precum și de faptul că pentru orice două matrice înlănuite¹⁴⁹ operația de transpunere are proprietatea $(XY)^\top = Y^\top X^\top$.

c. Fie K_1 și K_2 două funcții-nucleu oarecare definite pe $\mathbb{R}^d \times \mathbb{R}^d$, iar A_1 și A_2 matricele-nucleu corespunzătoare lui K_1 și respectiv K_2 pentru un set de elemente x_1, \dots, x_m arbitrar alese din \mathbb{R}^d .

Conform punctelor a și b , rezultă că matricele-nucleu A_1 și A_2 sunt simetrice și pozitiv semidefinite. Conform *observației* din enunț, funcția $K_1 + K_2$ (căreia, vom vedea mai jos, îi corespunde matricea-nucleu $A_1 + A_2$) este funcție-nucleu dacă $A_1 + A_2$ este matrice simetrică și $f^\top (A_1 + A_2) f \geq 0$ pentru orice vector-colonă nenul $f \in \mathbb{R}^m$.

Faptul că $A_1 + A_2$ este matrice simetrică rezultă imediat din ipoteză și din punctul a (aplicat pe rând funcțiilor nucleu K_1 și K_2).

Întrucât înmulțirea matricelor este distributivă față de adunare, rezultă:

$$f^\top (A_1 + A_2) f = f^\top (A_1 f + A_2 f) = f^\top A_1 f + f^\top A_2 f \geq 0$$

fiindcă $f^\top A_1 f \geq 0$ și $f^\top A_2 f \geq 0$, inegalități care derivă imediat din ipoteză, conform rezultatului de la punctul b .¹⁵⁰

În concluzie, matricea $A_1 + A_2$ fiind simetrică și pozitiv semidefinită, rezultă că funcția $K_1 + K_2$ este funcție-nucleu.

Consecință: Pentru orice număr finit de funcții-nucleu K_1, K_2, \dots, K_l , suma lor este de asemenea o funcție-nucleu. (Demonstrația se face imediat prin inducție matematică.)

Observație: Proprietatea demonstrată la punctul c privește „calitatea“ de funcție-nucleu a sumei $K_1 + K_2$. Implicit, este dovedită existența unei „mapări“ ϕ care corespunde lui $K_1 + K_2$. Însă demonstrația aceasta nu este construcțivă, adică nu se dă efectiv expresia funcției ϕ . Totuși, este foarte ușor de verificat următorul fapt: dacă luăm ca definiție pentru $\phi(x)$ vectorul obținut prin concatenarea vectorilor $\phi_1(x)$ și $\phi_2(x)$ — unde ϕ_1 și ϕ_2 sunt „mapările“ corespunzătoare nucleelor K_1 și respectiv K_2 —, se verifică egalitatea:

$$\phi(x) \cdot \phi(x') = \underbrace{K_1(x, x')}_{\phi_1(x) \cdot \phi_1(x')} + \underbrace{K_2(x, x')}_{\phi_2(x) \cdot \phi_2(x')} \stackrel{\text{def.}}{=} (K_1 + K_2)(x, x') \text{ pentru orice } x \text{ și } x' \in \mathbb{R}^d.$$

¹⁴⁹ Adică, de dimensiune $n_1 \times n_2$ și respectiv $n_2 \times n_3$.

¹⁵⁰ Ambele egalități de mai sus au fost deduse pe baza proprietății de distributivitate a adunării matricelor față de operația de adunare.

69. (Funcții-nucleu: [alte] câteva proprietăți de „construcție“)

CMU, 2012 fall, T. Mitchell, Z. Bar-Joseph, final exam, pr. 7.a.1

MIT, 2009 fall, Tommi Jaakkola, lecture 3

Stanford, 2008 fall, Andrew Ng, HW2, pr. 1.ce

Fie

K_1 și K_2 două funcții-nucleu definite pe $\mathbb{R}^d \times \mathbb{R}^d$,

$\Phi_1 : \mathbb{R}^d \rightarrow \mathbb{R}^n$ și $\Phi_2 : \mathbb{R}^d \rightarrow \mathbb{R}^n$ funcțiile de mapare a trăsăturilor, care corespund funcțiilor-nucleu K_1 și respectiv K_2 .

a. Arătați cum se pot defini funcțiile de mapare ale următoarelor funcții-nucleu, în raport cu Φ_1 și Φ_2 :

- i. $K(x, z) = cK_1(x, z)$, unde c este o constantă oarecare, pozitivă.
- ii. $K(x, z) = f(x)K_1(x, z)f(z)$, unde f este o funcție cu valori în \mathbb{R} .
(Se observă ușor că punctul ii este o generalizare a punctului i.)
- iii. $K(x, z) = K_1(x, z)K_2(x, z)$.

b. Arătați că funcția $K(x, z) = p(K_1(x, z))$, unde p este un polinom cu coeficienți pozitivi este de asemenea funcție-nucleu.

Răspuns:

a. Se verifică imediat că, definind $\Phi(x) = \sqrt{c}\Phi_1(x)$ în cazul i, și $\Phi(x) = f(x)\Phi_1(x)$ în cazul ii, este satisfăcută relația din definiția funcției-nucleu: $K(x, z) = \Phi(x) \cdot \Phi(z)$ pentru orice x și z din \mathbb{R}^d .

Pentru cazul iii, soluția nu mai este chiar atât de simplă. Dată fiind o instanță x arbitrar aleasă din \mathbb{R}^d , vom nota componentele vectorului $\Phi_1(x)$ cu $\phi_{11}(x), \phi_{12}(x), \dots, \phi_{1n_1}(x)$ și, similar, componentele vectorului $\Phi_2(x)$ cu $\phi_{21}(x), \phi_{22}(x), \dots, \phi_{2n_2}(x)$. Vom defini $\Phi(x)$ ca fiind un vector cu $n_1 n_2$ componente, fiecare componentă fiind un produs (de numere reale) de tipul $\phi_{1i}(x)\phi_{2j}(x)$:

$$\begin{aligned} \Phi(x) &\stackrel{\text{def.}}{=} [\phi_{11}(x)\phi_{21}(x), \phi_{11}(x)\phi_{22}(x), \dots, \phi_{11}(x)\phi_{2n_2}(x), \\ &\quad \phi_{12}(x)\phi_{21}(x), \phi_{12}(x)\phi_{22}(x), \dots, \phi_{12}(x)\phi_{2n_2}(x), \\ &\quad \dots \\ &\quad \phi_{1n_1}(x)\phi_{21}(x), \phi_{1n_1}(x)\phi_{22}(x), \dots, \phi_{1n_1}(x)\phi_{2n_2}(x)] \end{aligned}$$

Considerând acum perechea de elemente x și z arbitrar alese din \mathbb{R}^d , va rezulta:

$$\begin{aligned} \Phi(x) \cdot \Phi(z) &= \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \phi_{1i}(x)\phi_{2j}(x)\phi_{1i}(z)\phi_{2j}(z) \\ &= \left(\sum_{i=1}^{n_1} \phi_{1i}(x)\phi_{1i}(z) \right) \left(\sum_{j=1}^{n_2} \phi_{2j}(x)\phi_{2j}(z) \right) \\ &= (\Phi_1(x) \cdot \Phi_1(z))(\Phi_2(x) \cdot \Phi_2(z)) \\ &= K_1(x, z)K_2(x, z) = (K_1 K_2)(x, z). \end{aligned}$$

La cea de-a doua egalitate de mai sus am ținut cont de distributivitatea înmulțirii numerelor reale față de adunare.

b. Faptul că funcția compusă $p(K_1(x, z))$, unde p este un polinom cu coeficienți pozitivi este funcție-nucleu decurge imediat din relațiile i și iii de mai sus și punctul c de la problema 68.

70. (Funcții-nucleu: [încă] o proprietate de „construcție“)

■ □ • · Stanford, 2009 fall, Andrew Ng, practice midterm exam, pr. 3.a

Considerăm K o funcție-nucleu definită pe $\mathbb{R}^d \times \mathbb{R}^d$. Arătați că funcția compusă K_e definită prin relația $K_e(x, z) = e^{K(x, z)}$ este de asemenea funcție-nucleu.

Sugestie: Puteți folosi dezvoltarea sub formă de *serie Taylor* a lui e^x :

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots + \frac{x^n}{n!} + \dots,$$

precum și faptul că pentru orice sir de numere nenegative $\{a_n\}_{n \in \mathbb{N}}$, dacă există $a \stackrel{\text{not.}}{=} \lim_{n \rightarrow \infty} a_n$, atunci $a \geq 0$.

Consecință: Pentru orice număr real $a > 1$, avem $\ln a > 0$, deci funcția $(\ln a)K(x, z)$ este și ea funcție-nucleu (vedeți problema 69.a.i). Apoi,

$$a^{K(x, z)} = (e^{\ln a})^{K(x, z)} = e^{(\ln a)K(x, z)}.$$

Prin urmare, va rezulta că și funcția $a^{K(x, z)}$ este funcție-nucleu, pentru orice $a > 1$.

Răspuns:

Observație: În rezolvarea pe care o dăm mai jos, vom demonstra proprietatea din enunț nu în forma dată (generală, adică $e^{K(x, z)}$), ci într-un caz particular care este simplu, dar semnificativ. În mod concret, vom arăta că funcția $e^{x \cdot z}$ de variabile x și z din \mathbb{R}^d este funcție-nucleu. (Evident, funcția $x \cdot z$ este funcție-nucleu.) După aceea, a arăta că $e^{K(x, z)}$ este funcție-nucleu revine la a urma exact firul demonstrației de mai jos — înlocuind $x \cdot z$ cu $K(x, z) = \phi(x) \cdot \phi(z)$ și respectiv $x_i \cdot x_j$ cu $K(x_i, x_j)$ —, fiindcă justificările în baza cărora se asigură validitatea raționamentului în cazul simplu pe care l-am ales mai sus rămân valabile și pentru cazul general.

Vom demonstra că $e^{x \cdot z}$ este funcție-nucleu nu în mod direct — adică, dovedind existența unei funcții de mapare ϕ astfel încât $e^{x \cdot z} = \phi(x) \cdot \phi(z)$ —, ci folosind teorema de „caracterizare“ a lui Mercer (a se vedea problema 68.ab și nota de subsol 147).

Conform acestei teoreme, pentru a dovedi că $e^{x \cdot z}$ este funcție-nucleu este suficient să arătăm că pentru orice $m \in \mathbb{N}$ și pentru orice elemente $x_1, \dots, x_m \in \mathbb{R}^d$, matricea G (Gram), care este de tip $m \times m$ și are elementul generic de forma $G(i, j) \stackrel{\text{not.}}{=} e^{x_i \cdot x_j}$ pentru $i, j = \overline{1, m}$, este simetrică și pozitiv semidefinită.

Fie deci un număr natural $m \in \mathbb{N}$, fixat, precum și numerele $x_1, \dots, x_m \in \mathbb{R}^d$, de asemenea fixate. Folosind dezvoltarea lui $e^{x_i \cdot x_j}$ ca serie Taylor, avem:

$$e^{x_i \cdot x_j} = \sum_{n=0}^{\infty} \frac{(x_i \cdot x_j)^n}{n!} \stackrel{\text{def.}}{=} \lim_{n \rightarrow \infty} \left(\sum_{l=0}^n \frac{(x_i \cdot x_j)^l}{l!} \right)$$

Cu aceasta, devine evident că putem scrie matricea G ca fiind limita unui sir de matrice, $\{G_n\}_{n \in \mathbb{N}}$, elementul generic al matricei G_n fiind

$$G_n(i, j) \stackrel{\text{not.}}{=} \sum_{l=0}^n \frac{(x_i \cdot x_j)^l}{l!} \text{ pentru } i, j = \overline{1, m}.$$

Matricea G_n astfel definită este simetrică, datorită proprietății de simetrie a produsului scalar. Evident, în urma trecerii la limită ($n \rightarrow \infty$) pe componente, se păstrează simetria matricei.¹⁵¹ Așadar, rezultă că matricea G este simetrică.

Matricea G_n este și pozitiv semidefinită. Justificarea provine din faptul că funcția $x \cdot z$ de variabile x și z din \mathbb{R}^d este funcție-nucleu și din aplicarea proprietăților de „construcție“ de noi nuclee: $K_1 + K_2$, $K_1 K_2$ și cK_1 cu $c > 0$ sunt funcții-nucleu dacă K_1 și K_2 sunt funcții-nucleu (a se vedea problemele 68.c și 69).¹⁵² Rămâne de arătat că prin trecere la limită ($G_n \rightarrow G$) se păstrează proprietatea de „pozitiv semidefinire“ a matricelor.

Fie $f \in \mathbb{R}^m$, văzut ca vector-colonă. Urmează:

$$f^\top G f = f^\top (\lim_{n \rightarrow \infty} G_n) f = \lim_{n \rightarrow \infty} (\underbrace{f^\top G_n f}_{\geq 0}) \geq 0.$$

Cea de-a doua egalitate de mai sus are loc datorită proprietății de liniaritate a limitei de siruri: limita combinației liniare a $m \times m$ siruri este combinația liniară a limitelor sirurilor respective. Inegalitatea $\lim_{n \rightarrow \infty} (f^\top G_n f) \geq 0$ are loc fiindcă $f^\top G_n f \geq 0$ pentru orice n (datorită faptului că matricea G_n este pozitiv semidefinită, după cum am arătat mai sus) și datorită proprietății de păstrare a pozitivității prin trecerea la limită (vedeți cea de-a două sugestie din enunț).

Așadar, matricea G este simetrică și pozitiv semidefinită. În concluzie, funcția $e^{x \cdot z}$ este funcție-nucleu.

Observație: Din demonstrația de mai sus rezultă imediat că printr-un raționament similar se poate dovedi următoarea proprietate de tip „construcție“: dacă avem un sir de funcții-nucleu $K_1, K_2, \dots, K_n, \dots$ definite pe $\mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ și există $\lim_{n \rightarrow \infty} K_n \stackrel{\text{not.}}{=} K$, atunci K este și ea funcție-nucleu.¹⁵³

71. (Calculul distanței euclidiene dintre imaginile a două instanțe
în spațiul de trăsături, cu ajutorul funcției-nucleu)

University of Utah, 2008 spring, Hal Daumé III, HW1C, pr. 3

Presupunem că pentru o funcție $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ există o funcție $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^n$ care satisfac proprietatea $K(x, z) = \phi(x) \cdot \phi(z)$ pentru orice $x, z \in \mathbb{R}^d$. (Operatorul \cdot reprezintă produsul scalar în \mathbb{R}^n .) Fie distanța euclidiană în spațiul n -dimensional,

$$\| \phi(x) - \phi(z) \| \stackrel{\text{def.}}{=} \sqrt{\sum_{i=1}^n (\phi(x)_i - \phi(z)_i)^2}.$$

Arătați cum se poate calcula această distanță folosind doar funcția K , adică fără a apela (în mod explicit) la valorile funcției ϕ .

¹⁵¹Altfel spus, $G_n(i, j) = G_n(j, i) \rightarrow G(i, j) = G(j, i)$, unde $G(i, j) = \lim_{n \rightarrow \infty} G_n(i, j)$ și $G(j, i) = \lim_{n \rightarrow \infty} G_n(j, i)$.

¹⁵²Se ține cont de relația de recurență între matricele-nucleu $G_n = G_{n-1} + \frac{1}{n!} [(x_i \cdot x_j)^n]_{i,j=1,m}$, cu $G_1 = I$, matricea identitate de ordin d . Este imediat că $(x \cdot z)^n$, unde $n \geq 2$, este funcție-nucleu, întrucât $x \cdot z$ este funcție-nucleu.

¹⁵³Cf. https://en.wikipedia.org/wiki/Positive-definite_kernel, accesat la data 30.03.2022.

Observație: Rezultatul acesta este util în cazul în care se aplică algoritmul de clasificare k -NN (sau algoritmi de clusterizare) după ce s-a făcut „maparea“ datelor de antrenament într-un nou „spațiu de trăsături“.¹⁵⁴

Răspuns:

$$\begin{aligned} \|\phi(x) - \phi(z)\| &= \sqrt{\sum_{i=1}^n (\phi(x)_i - \phi(z)_i)^2} = \sqrt{\sum_{i=1}^n (\phi(x)_i^2 + \phi(z)_i^2 - 2\phi(x)_i\phi(z)_i)} \\ &= \sqrt{\sum_{i=1}^n \phi(x)_i\phi(x)_i + \sum_{i=1}^n \phi(z)_i\phi(z)_i - 2\sum_{i=1}^n \phi(x)_i\phi(z)_i} \\ &= \sqrt{\phi(x) \cdot \phi(x) + \phi(z) \cdot \phi(z) - 2\phi(x) \cdot \phi(z)} = \sqrt{K(x, x) + K(z, z) - 2K(x, z)}. \end{aligned}$$

72. (Un exemplu de funcție-nucleu
care exprimă / măsoară similaritatea dintre două imagini oarecare)
□ • ○ CMU, 2014 fall, E. Xing+B. Poczos, HW2, pr. 3.1

Considerăm X o mulțime formată din imagini dreptunghiulare de dimensiuni arbitrară, în care fiecare pixel este reprezentat sub forma unui număr întreg din mulțimea $\{0, \dots, 255\}$. Fie $k_1 : X \times X \rightarrow \mathbb{R}$ o funcție de similaritate a imaginilor, definită astfel: $k_1(x, x') =$ numărul de zone pătratice de pixeli (engl., pixel patches), de dimensiune 16×16 , care apar atât în imaginea x cât și în imaginea x' .

- a. Demonstrați că funcția k_1 este funcție-nucleu.
- b. Demonstrați că funcția de similaritate definită mai jos nu este funcție-nucleu.

$$k_2(x, x') = \begin{cases} 1 & \text{dacă } k_1(x, x') \geq 1, \text{ adică, există cel puțin un "patch" (zonă pătratică) comun(ă) pentru } x \text{ și } x' \\ 0 & \text{în caz contrar.} \end{cases}$$

Sugestie: Arătați că există [un anumit set de] instanțe pentru care matricea Gram generată folosind funcția k_2 nu este pozitiv semidefinită. Conform teoremei a lui Mercer (vedeți problema 68) va rezulta că funcția k_2 nu este funcție-nucleu.

Răspuns:

- a. Vom demonstra că există o funcție de „mapare“ ϕ definită convenabil, astfel încât $k_1(x, x') = \phi(x) \cdot \phi(x')$ pentru oricare două imagini dreptunghiulare, x și x' .

Notăm cu d numărul tuturor zonelor pătratice constituite din pixeli (patch-uri), de dimensiune 16×16 . Întrucât fiecare pixel poate avea 256 de valori, urmează că

$$d = 256^{16 \times 16} = (2^8)^{256} = 2^{2048}.$$

¹⁵⁴Vedeți de exemplu problema 15 de la capitolul *Învățare bazată pe memorare și — mai ales!* — problema 51 de la capitolul *Clusterizare*.

Fie funcția ϕ definită pe mulțimea tuturor imaginilor dreptunghiulare posibile care ia valori în mulțimea $\{0, 1\}^d$, astfel: $\phi(x) \stackrel{\text{not.}}{=} (\phi(x)_1, \dots, \phi(x)_i, \dots, \phi(x)_d)$, unde $\phi(x)_i$ ia valoarea 1 dacă patch-ul i este prezent în imaginea x , și 0 în caz contrar. Se poate constata ușor că în baza acestei definiții rezultă $k_1(x, x') = \phi(x) \cdot \phi(x')$. Prin urmare, funcția k_1 este funcție-nucleu.

b. Fie A și B două patch-uri având toți pixelii 0 și, respectiv, toți pixelii 1. De asemenea, fie x_1, x_2 și x_3 trei imagini definite astfel: $x_1 = A$, $x_2 = B$, iar $x_3 = [AB]$, unde prin $[AB]$ am notat imaginea obținută prin concatenarea imaginilor A și B pe orizontală. Vom demonstra că matricea Gram care se obține pe acest set de imagini folosind funcția-nucleu k_2 nu este matrice pozitiv semidefinită.

Într-adevăr, respectiva matrice Gram este

$$G \stackrel{\text{not.}}{=} \begin{bmatrix} k_2(x_1, x_1) & k_2(x_1, x_2) & k_2(x_1, x_3) \\ k_2(x_2, x_1) & k_2(x_2, x_2) & k_2(x_2, x_3) \\ k_2(x_3, x_1) & k_2(x_3, x_2) & k_2(x_3, x_3) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

Conform definiției, G este matrice pozitiv semidefinită dacă pentru orice $f \in \mathbb{R}^3$ (văzut ca vector-colonă) urmează că $f^\top G f \geq 0$.

Considerând $f = (1, 1, -1)^\top$, rezultă că $f^\top G f = (0, 0, 1)(1, 1, -1)^\top = -1 < 0$. Prin urmare, matricea G nu este pozitiv semidefinită și, în consecință, conform teoremei lui Mercer, funcția k_2 nu este funcție-nucleu.

73.

(O metodă particulară de mapare a atributelor, care asigură separare liniară)

University of Utah, 2008 fall, Hal Daumé III, HW3, pr. 2

Considerăm N puncte într-un spațiu D -dimensional, nu neapărat separabile liniar. În acest exercițiu vom arăta că aceste puncte se pot mări într-un spațiu cu $D + N$ dimensiuni, în aşa fel încât să devină separabile liniar. Acest lucru se realizează transformând punctele $x_n \in \{x_1, x_2, \dots, x_D\}$ în puncte de forma $(x_1, x_2, \dots, x_D, 1_{\{n=1\}}, 1_{\{n=2\}}, 1_{\{n=3\}}, \dots, 1_{\{n=N-1\}}, 1_{\{n=N\}})$, unde de exemplu prin notația $1_{\{n=3\}}$ se înțelege valoarea 1 dacă $n = 3$ și 0 în cazul $n \neq 3$. Arătați că punctele obținute în urma acestei mapări sunt într-adevăr separabile liniar.

Răspuns:

Considerând x_1, \dots, x_N instanțele date și X_1, \dots, X_N imaginile lor din spațiul $(D+N)$ -dimensional, vom arăta că există $w \in \mathbb{R}^{D+N}$ astfel încât semnul expresiei $w \cdot X_i$ este pozitiv pentru instanțele etichetate pozitiv și negativ în rest.

Dacă luăm $w = (0, 0, \dots, 0, y_1, y_2, \dots, y_N) \in \mathbb{R}^{D+N}$, unde

$$y_i = \begin{cases} 1 & \text{dacă punctul } X_i \text{ este etichetat cu +} \\ -1 & \text{dacă punctul } X_i \text{ este etichetat cu -}, \end{cases}$$

rezultă că

$$w \cdot X_i = \sum_{i=1}^N y_i \cdot X_i^{(D+i)} = y_i, \forall i \in \{1, \dots, N\}$$

adică $w \cdot X_i = +1$ pentru toate instanțele x_i pozitive și $w \cdot X_i = -1$ pentru toate instanțele x_i negative.

Așadar, hiperplanul $w \cdot X = 0$ este un separator liniar pentru cele N puncte din spațiul $(D + N)$ -dimensional.

Observație: Deși are o anumită importanță teoretică,¹⁵⁵ maparea indicată în enunțul acestei probleme nu este eficientă din punct de vedere practic, în special atunci când N și / sau D sunt numere foarte mari. De asemenea, ea produce overfitting, datorită incapacității ei intrinseci de a generaliza.

74.

(RBF este funcție-nucleu: demonstrație)

■ □ • · · Stanford, 2009 fall, Andrew Ng, practice midterm exam, pr. 3.b

Arătați că funcția cu baza radială $e^{-\frac{1}{2\sigma^2}\|x-z\|^2}$, de variabile $x, z \in \mathbb{R}^d$, este într-adevăr funcție-nucleu.

Observație: În acest exercițiu nu se cere să se găsească efectiv funcția de mapare ϕ pentru nucleul RBF. Este suficient să se dovedească existența ei, folosind de exemplu teorema lui Mercer (a se vedea problema 68.ab și nota de sub-sol 147) sau proprietățile de „construcție“ de noi nuclee (problemele 68.c, 69, 70 și 153). Problema 75 va arăta că există o funcție de „mapare“ ϕ pentru nucleul RBF care asociază fiecărei instanțe x un „vector“ $\phi(x)$ a cărui dimensiune este infinită!¹⁵⁶

Răspuns:

Mai întâi vom pune expresia de definiție pentru funcția cu baza radială sub o formă mai convenabilă:

$$\begin{aligned} e^{-\frac{\|x-z\|^2}{2\sigma^2}} &= e^{-\frac{(x-z) \cdot (x-z)}{2\sigma^2}} = e^{-\frac{(x^2 - 2x \cdot z + z^2)}{2\sigma^2}} = e^{-\frac{x^2}{2\sigma^2}} e^{\frac{x \cdot z}{\sigma^2}} e^{-\frac{z^2}{2\sigma^2}} \\ &= f(x) e^{\frac{x \cdot z}{\sigma^2}} f(z), \end{aligned}$$

¹⁵⁵În afară de asigurarea separabilității în spațiul de trăsături (ceea ce am demonstrat în problema de față), această mapare — sau, chiar mai bine, o mapare similară cu aceasta — servește la determinarea numărului maxim de actualizări pe care le face perceptronul [Rosenblatt] cu margine (engl., margin perceptron) pe seturi de date neliniar-separabile, lucrând în spațiul de trăsături corespunzător. Pentru detalii, vedeti la capitolul Rețele neuronale artificiale problema 40.c (în conjuncție cu problema 18).

¹⁵⁶ Prof. Tommi Jaakkola de la MIT precizează (Machine Learning course, 2009 fall, lecture notes 3, pages 3-4) că

$$\exp(-\frac{\beta}{2} \|x - x'\|^2) = \int \phi(z; x)\phi(z; x')dz \quad (*)$$

unde $\phi(z; x) \stackrel{\text{def.}}{=} c(\beta, d) \mathcal{N}(z|x, \frac{1}{2\beta})$, cu $c(\beta, d)$ o constantă, și $\mathcal{N}(z|x, \frac{1}{2\beta})$ distribuția normală de medie x și varianță $\frac{1}{2\beta}$. $\phi(x)$ poate fi definit ca fiind „vectorul“ infinit $\{\phi(z; x)\}_{z \in \mathbb{R}^d}$. Așadar, indexarea trăsăturilor se face după $z \in \mathbb{R}^d$, nu după un indice întreg așa cum este cazul alteori. Semnificația relației (*) este următoarea: nucleul RBF măsoară probabilitatea ca o instanță z să fie generată de două gaussiene de medii x și respectiv x' și varianță comună, $\frac{1}{2\beta}$. Funcțiile-nucleu sunt adeseori definite dintr-o astfel de perspectivă.

unde prin $f(x)$ am notat expresia $e^{-\frac{x^2}{2\sigma^2}}$.

Acum putem aplica diverse proprietăți de „construcție“ a nucleelor, pornind de la un nivel simplu spre un nivel din ce în ce mai complex. Evident, funcția $x \cdot z$ este funcție-nucleu. La fel și funcția $\frac{x \cdot z}{\sigma^2}$ (vedeți problema 69 punctul a.i.).

A demonstra faptul că funcția $e^{\frac{x \cdot z}{2\sigma^2}}$ este funcție-nucleu revine la a reproduce linia demonstrației de la problema 70 (până la un factor constant). În sfârșit, datorită proprietății de la problema 69 punctul a.ii, rezultă că $f(x) e^{\frac{x \cdot z}{2\sigma^2}} f(z)$ — deci și $e^{-\frac{1}{2\sigma^2} \|x-z\|^2}$ — este funcție-nucleu.

75.

(Spații de „trăsături“ infinite)

■ □ • ○ CMU, 2011 fall, T. Mitchell, A. Singh, HW6, pr. 2.2

Pentru date reale de tipul $x \in \mathbb{R}$ putem să ne gândim la o transformare („mapare“) a trăsăturilor $\phi_n : \mathbb{R}^1 \rightarrow \mathbb{R}^{n+1}$ definită într-o manieră ceva mai complicată decât în mod obișnuit:

$$\phi_n(x) = \left\{ e^{-x^2/2}, e^{-x^2/2} x, e^{-x^2/2} \frac{x^2}{\sqrt{2}}, \dots, e^{-x^2/2} \frac{x^i}{\sqrt{i!}}, \dots, e^{-x^2/2} \frac{x^n}{\sqrt{n!}} \right\}.$$

Presupunem acum că $n \rightarrow \infty$ și definim o nouă „mapare“ a trăsăturilor, care produce un vector infinit:

$$\phi_\infty(x) = \left\{ e^{-x^2/2}, e^{-x^2/2} x, e^{-x^2/2} \frac{x^2}{\sqrt{2}}, \dots, e^{-x^2/2} \frac{x^i}{\sqrt{i!}}, \dots \right\}. \quad (93)$$

Se poate arăta — vedeți de exemplu problemele 9.c și 12.f de la capitolul *Mașini cu vectori-suport* — că în multe situații putem să exprimăm ecuația separatorului produs de un clasificator liniar folosind doar produse scalare de vectori (anumite instanțe de antrenament) într-un spațiu oarecare (sau în acesta-numitul spațiu de „trăsături“).¹⁵⁷ Ne punem problema dacă n-am putea să folosim cumva spațiul de trăsături obținut prin „maparea“ ϕ_∞ . Totuși, pentru aceasta ar trebui să putem calcula produsul scalar de [imagini de] instanțe în acest spațiu de „trăsături“ infinit. Definim produsul scalar dintre doi vectori infiniti $a = \{a_1, \dots, a_i, \dots\}$ și $b = \{b_1, \dots, b_i, \dots\}$ ca fiind suma infinită

$$a \cdot b = \sum_{i=1}^{\infty} a_i b_i \stackrel{\text{def.}}{=} \lim_{n \rightarrow \infty} \sum_{i=1}^n a_i b_i. \quad (94)$$

a. Putem oare să calculăm în mod explicit $\phi_\infty(a) \cdot \phi_\infty(b)$? Altfel spus, care este expresia explicită pentru funcția-nucleu $K(a, b) \stackrel{\text{not.}}{=} \phi_\infty(a) \cdot \phi_\infty(b)$?

Sugestie: Folosiți dezvoltarea în serie Taylor a lui e^x :

$$e^x = \lim_{n \rightarrow \infty} \sum_{i=0}^n \frac{x^i}{i!}. \quad (95)$$

¹⁵⁷Pentru caracterizări teoretice ale unor astfel de clasificatori (așa-numitele „teoreme de reprezentare“), vedeți problemele 88 și 177.

b. Ar trebui oare ca într-un asemenea spațiu de dimensiune infinită să ne punem problema *complexității* clasificatorului [care folosește această „mapare“]?

Răspuns:

a. Putem scrie imediat:

$$\begin{aligned} K(a, b) &\stackrel{\text{not.}}{=} \phi_\infty(a) \cdot \phi_\infty(b) \stackrel{(94)(93)}{=} \sum_{i=0}^{\infty} \frac{e^{-a^2/2} a^i}{\sqrt{i!}} \cdot \frac{e^{-b^2/2} b^i}{\sqrt{i!}} \\ &= \exp\left(-\frac{a^2 + b^2}{2}\right) \cdot \sum_{i=0}^{\infty} \frac{(ab)^i}{i!} \stackrel{(95)}{=} \exp\left(-\frac{a^2 + b^2}{2}\right) \cdot \exp(ab) = \exp\left(-\frac{(a-b)^2}{2}\right). \end{aligned}$$

Observație: Ultima expresie care a fost obținută aici corespunde unei funcții-nucleu cu baza radială, ale cărei argumente (a și b) sunt numere reale. Demonstrația pentru cazul *ceva mai* general al nucleului RBF $K(a, b) \stackrel{\text{def.}}{=} \exp(-\gamma(a-b)^2)$, unde $\gamma \in \mathbb{R}_+^*$, iar $a, b \in \mathbb{R}$, urmează îndeaproape linia acestei demonstrații.¹⁵⁸ Așadar, pentru funcțiile-nucleu de tip RBF cu argumente reale, „maparea“ ϕ ia valori într-un spațiu de dimensiune infinită. Pentru cazul nucleelor RBF cu argumente din \mathbb{R}^d , unde $d \geq 2$, este util să vedeti nota de subsol 156.

b. În general, ar trebui ca un model [de clasificare] bazat pe astfel de vectori de „trăsături“ să ne îngrijoreze (din perspectiva *overfitting*-ului), însă la clasificatorii care folosesc funcții-nucleu, calculele nu se fac în spațiul de „trăsături“, ci în spațiul de intrare (adică, \mathbb{R} în cazul de față, fiindcă $a, b \in \mathbb{R}$ și $\phi(a) \cdot \phi(b) = K(a, b)$.) În cazul mașinilor cu vectori-suport, complexitatea modelului este dată de numărul de vectori care intră *de facto* în definiția separatorului optimal (aşa-numiții „vectori-suport“), nu de dimensiunea spațiului de „trăsături“.¹⁵⁹

76.

(O proprietate simplă a nucleului de tip RBF)

CMU, 2012 fall, T. Mitchell, Z. Bar-Joseph, final exam, pr. 7.a.2

Aşa-numita funcție cu baza radială (RBF), $K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$ este una dintre cele mai folosite funcții-nucleu. Considerăm punctele x, z_1 și z_2 situate geometric astfel: z_1 foarte aproape de x , iar z_2 situat foarte departe de x .

Cum vor fi valorile lui $K(z_1, x)$ și $K(z_2, x)$? Alegeți una dintre următoarele variante:

- i. $K(z_1, x)$ va fi foarte aproape de 1, iar $K(z_2, x)$ va fi foarte aproape de 0.
- ii. $K(z_1, x)$ va fi foarte aproape de 0, iar $K(z_2, x)$ va fi foarte aproape de 1.
- iii. $K(z_1, x)$ va avea o valoare mult mai mare decât 1, iar $K(z_2, x)$ va avea o valoare mult mai mică decât 0.

¹⁵⁸Vedeți CMU, 2014 fall, Eric Xing, Barnabas Poczos, HW2, pr. 3.2.

¹⁵⁹Vedeți, mai concret, formulele (293) și (298) de la problemele 9 și respectiv 12 de la capitolul *Mașini cu vectori-suport*. Vectorii-suport sunt acele (în general, puține!) instanțe de antrenament x_i pentru care coeficienții α_i sunt nenuli.

iv. $K(z_1, x)$ va avea o valoare mult mai mică decât 0, iar $K(z_2, x)$ va avea o valoare mult mai mare decât 1.

Răspuns:

Funcția e^{-y} are valoarea 1 pentru $y = 0$. Fiind o funcție continuă, limita ei pentru $y \rightarrow 0$ este 1. Pentru $y \rightarrow +\infty$ limita acestei funcții este 0.

Prin urmare, pentru $z_1 \rightarrow x$ vom avea $\|z_1 - x\| \rightarrow 0$ și $K(z_1, x) \rightarrow 1$, iar pentru $\|z_2 - x\| \rightarrow +\infty$ vom avea $K(z_2, x) \rightarrow 0$. Așadar, doar afirmația de la punctul i. este adevărată; celelalte afirmații sunt false.

77.

(Funcții-nucleu: da sau nu?)

• CMU, 2017 fall, Nina Balcan, midterm, pr. 1.1

Care dintre următoarele funcții nu constituie funcții-nucleu valide?

- a. $c_1 K_1(x_1, x_2) + c_2 K_2(x_1, x_2)$, unde K_1 și K_2 sunt funcții-nucleu, iar c_1 și c_2 sunt constante reale;
- b. $K(x_1, x_2) \stackrel{\text{def.}}{=} x_1^\top A x_2$, unde $A \in \mathbb{R}^{d \times d}$ este o matrice simetrică și pozitiv semidefinită;¹⁶⁰
- c. $K(x_1, x_2) \stackrel{\text{def.}}{=} x_1^\top A x_2$, unde $A \in \mathbb{R}^{d \times d}$ este o matrice simetrică.

Răspuns:

a. Conform problemelor 69.a și 68.c, știm că atunci când c_1 și c_2 sunt pozitive, suma $c_1 K_1(x_1, x_2) + c_2 K_2(x_1, x_2)$ reprezintă o funcție-nucleu. Vom arăta că atunci când se renunță la condiția $c_1, c_2 \geq 0$ concluzia nu se mai menține (în general). Într-adevăr, dacă luăm $c_1 = 0$ și $c_2 = -1$, funcția $c_1 K_1(x_1, x_2) + c_2 K_2(x_1, x_2) = -K_2(x_1, x_2)$ va avea, pentru orice set de instanțe $\{x_i\}_{i=1}^m$, matricea Gram $-G_2$, unde prin G_2 am notat matricea Gram corespunzătoare pentru funcția-nucleu K_2 . Conform teoremei lui Mercer (vedeți problema 68.ab), G_2 este matrice pozitiv semidefinită. În cazul (particular) în care o astfel de matrice G_2 este chiar pozitiv definită rezultă că pentru orice vector-colonă nenul f din \mathbb{R}^m vom avea $f^\top G_2 f > 0$. Dacă presupunem prin *reducere la absurd* că funcția $-K_2$ este funcție-nucleu, conform aceleiasi teoreme ar rezulta că într-un astfel de context vom avea $f^\top (-G_2) f > 0 \Leftrightarrow -f^\top G_2 f > 0 \Leftrightarrow f^\top G_2 f < 0$ pentru orice vector-colonă nenul $f \in \mathbb{R}^m$, ceea ce evident intră în *contradicție* cu relația $f^\top G_2 f > 0$. Așadar, într-un astfel de caz funcția $-K_2$ nu este funcție-nucleu.¹⁶¹

b. În acest caz, răspunsul este pozitiv: funcția $x_1^\top A x_2$, unde A este matrice de numere reale, simetrică și pozitiv semidefinită, este funcție-nucleu. Jusificarea se face pe baza proprietății de *factorizare* care a fost demonstrată

¹⁶⁰Problema 160 indică două modalități de obținere de *noi* matrice pozitiv semidefinite pornind de la alte asemenea matrice.

¹⁶¹Observație: În cazul (complementar) în care orice astfel de matrice G_2 este pozitiv semidefinită (nu pozitiv definită, ca mai sus) rezultă — presupunând în continuare (prin reducere la absurd) că funcția K_2 este funcție-nucleu, ceea ce implică faptul că matricea $-G_2$ este pozitiv semidefinită — că $f^\top G_2 f = 0$ pentru orice $f \in \mathbb{R}^m$. În particular, atunci când luăm drept f vectorii de tip „one-hot“ din \mathbb{R}^m , rezultă imediat că vom avea $K(x_i, x_i) = 0$ pentru orice x_i . (Aceaștă înseamnă că funcția-nucleu K_2 este funcția constantă 0.) În consecință, $\Phi_2(x_i) = 0$ pentru orice x_i , ceea ce nu este deloc util din punctul de vedere al task-urilor principale din învățarea automată (clasificare, regresie ori clusterizare).

la problema 36.a (vedeți *Observația* (4) de la pagina 81): există o matrice B astfel încât $A = BB^\top$. Prin urmare,

$$x_1^\top Ax_2 = x_1^\top BB^\top x_2 = (B^\top x_1)^\top (B^\top x_2).$$

Definind funcția de „mapare“ $\phi(x) = B^\top x$, egalitatea precedentă implică imediat faptul că funcția $K(x_1, x_2) \stackrel{\text{def.}}{=} x_1^\top Ax_2 = \phi(x_1)^\top \phi(x_2)$ este funcție-nucleu.

c. În acest caz, chestiunea revine la a determina dacă renunțarea (în raport cu punctul b) la condiția de pozitiv semidefinire a matricei A (păstrând însă condiția de simetrie) are sau nu vreun efect asupra „nucleicității“ funcției $x_1^\top Ax_2$. Răspunsul este negativ.

Într-adevăr, dacă am presupune (prin *reducere la absurd*) că funcția $K(x_1, x_2) \stackrel{\text{def.}}{=} x_1^\top Ax_2$, unde A este o matrice de numere reale, simetrică dar fără a fi pozitiv semidefinită este funcție-nucleu, atunci în baza unui raționament absolut similar cu cel din rezolvarea problemei 68.b¹⁶² ar rezulta cu necesitate că A este matrice pozitiv semidefinită, ceea ce reprezintă o *contradicție*. Prin urmare, atunci când A nu este pozitiv semidefinită nu rezultă cu necesitate că funcția $K(x_1, x_2) \stackrel{\text{def.}}{=} x_1^\top Ax_2$ este funcție-nucleu.

0.1.7 Metode de optimizare în învățarea automată

78. (Definiții și proprietăți de bază ale funcțiilor convexe)
 • CMU, 2015 fall, A. Smola, B. Poczos, HW1, pr. 3.1

În anumiți algoritmi de optimizare, cum este de exemplu *metoda gradientului descendente*, convexitatea funcției “target” joacă un rol important în a stabili dacă algoritmul va converge (sau nu), cât de mare este rata / rapiditatea convergenței și.a.m.d. În această problemă, pornind de la *definiții*, vom pune în evidență câteva proprietăți de care ne vom putea servi atunci când va trebui să examinăm convexitatea unei funcții.¹⁶³

Definiție (mulțime convexă):

Spunem că o mulțime $C \subseteq \mathbb{R}^d$ este convexă dacă pentru orice pereche de puncte $x, y \in C$ și pentru orice $t \in [0, 1]$ urmează că și punctul $tx + (1-t)y \in C$.¹⁶⁴

Definiție (funcție convexă):

Spunem că o funcție $f : \mathbb{R}^d \rightarrow \mathbb{R}$ este convexă dacă domeniul ei de

¹⁶²Veti ține cont că ipoteza „ K este funcție-nucleu“ implică faptul că există o „mapare“ ϕ astfel încât $K(x_1, x_2) = \phi(x_1) \cdot \phi(x_2)$ pentru orice x_1 și x_2 din domeniul de definiție al lui K . Apoi, G , matricea Gram pentru instanțele x_1, \dots, x_m se va scrie generic $[x_i^\top Ax_j]_{i,j} = [\phi(x_i)^\top \phi(x_j)]_{i,j}$, după care inegalitatea $f^\top Gf \geq 0$ pentru orice $f \in \mathbb{R}^m$ va fi demonstrată exact ca la problema 68.b.

¹⁶³De exemplu, problema 163 examinează convexitatea unora dintre cele mai des folosite funcții obiectiv din domeniul învățării automate [profunde].

¹⁶⁴Această proprietate înseamnă că pentru orice pereche de puncte x și y din mulțimea convexă C , orice punct care este situat pe segmentul de dreaptă care unește perechea de puncte x, y aparține de asemenea mulțimii C . Este foarte util ca o mulțime să aibă această proprietate, deoarece putem ajunge la orice punct din mulțime pornind de la un punct dat [din aceeași mulțime], parcurgând o linie dreaptă, fără să atingem „marginea“ mulțimii respective. Adeseori, a atinge „marginea“ mulțimii înseamnă că poți ajunge să fi blocate într-un punct de minim local și să nu mai ai niciodată sansa să găsești soluția optimă globală.

definiție (notat de aici încolo cu $\text{dom}(f)$) este o mulțime convexă, și pentru orice $x, y \in \text{dom}(f)$ și orice $t \in [0, 1]$ urmează că:¹⁶⁵

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y). \quad (96)$$

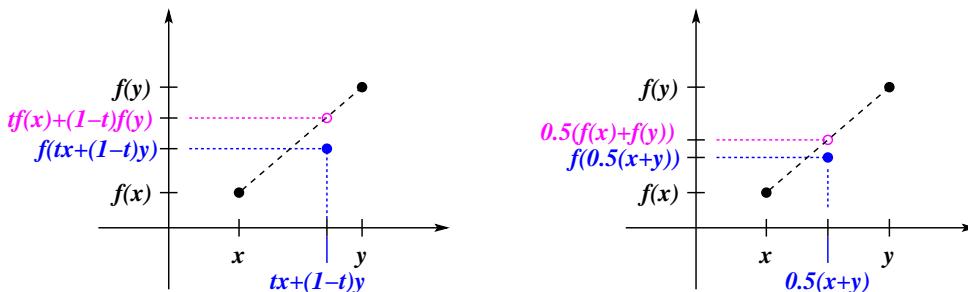
a. Uneori nu este ușor să demonstrăm convexitatea unei funcții în mod direct, adică bazându-ne doar pe definiția de mai sus. De aceea, în cele ce urmează vom formula câteva proprietăți ale funcțiilor convexe, fiecare dintre aceste proprietăți fiind echivalente cu definiția.¹⁶⁶ Pentru simplitate, vom presupune că $\text{dom}(f) = \mathbb{R}$.

1. Fie f o funcție continuă. Funcția f este convexă dacă și numai dacă

$$f\left(\frac{x+y}{2}\right) \leq \frac{f(x) + f(y)}{2}, \text{ pentru orice } x, y \in \mathbb{R}.$$

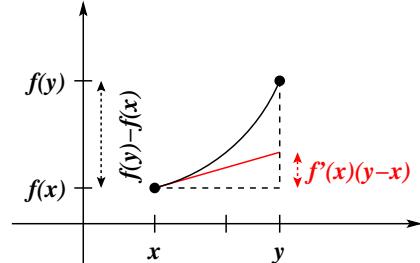
Aceasta este aşa-numita „proprietate a punctului de mijloc“ (engl., *mid-point property*).

Figura următoare oferă o reprezentare grafică a relației din definiția noțiunii de funcție convexă (relația (96)), precum și a relației din condiția 1.



2. Fie f o funcție despre care știm că este derivabilă, iar derivata sa o funcție continuă. f este convexă dacă și numai dacă inegalitatea $f(x) \geq f(y) + f'(y)(x - y)$ este satisfăcută, pentru orice $x, y \in \mathbb{R}$.

Observație: Interschimbând rolurile variabilelor x și y , ajungem la următoarea formulare echivalentă: $f(y) \geq f(x) + f'(x)(y - x)$, pentru orice $x, y \in \mathbb{R}$.¹⁶⁷



3. Fie f o funcție dublu derivabilă (adică, există atât f' cât și f''). Funcția f este convexă dacă și numai dacă $f''(x) \geq 0$, pentru orice $x \in \mathbb{R}$.

¹⁶⁵Inegalitatea (96) spune că segmentul de dreapta care unește două puncte situate pe graficul funcției f este întotdeauna situat deasupra [graficului] funcției f . O astfel de proprietate este foarte utilă atunci când urmărim să identificăm valoarea minimă a funcției, întrucât pentru orice două puncte, x și y , alese în mod arbitrar din $\text{dom}(f)$, există un punct z situat între x și y astfel încât $f(z)$ coincide cu minimul funcției f pe intervalul $[x, y]$. Așadar, în mod cert vom putea afla valoarea optimă a funcției f .

¹⁶⁶Astfel de proprietăți se numesc *caracterizări* ale definiției.

¹⁶⁷Semnificația geometrică a acestor relații este următoarea: dacă f este funcție convexă, tangenta la graficul lui f (în orice punct din domeniul de definiție) este situată sub graficul lui f .

Se poate spune că această „condiție“ (relație echivalentă cu definiția) este *de ordinul întâi*, spre deosebire de condiția 1, care este *de ordinul zero* (la fel ca și din condiția de definiție (96)), dar și de următoarea condiție (3), care este *de ordinul al doilea*.

Vă cerem ca pentru fiecare dintre punctele 1, 2 și 3 să faceți demonstrația echivalenței cu definiția noțiunii de funcție convexă, în ambele direcții.

b. Folosiți definiția sau proprietățile formulate mai sus pentru a demonstra următoarele afirmații:

4. Dacă f și g sunt funcții convexe având același domeniu de definiție, atunci funcția h definită prin relația $h(x) = \max(f(x), g(x))$ este de asemenea o funcție convexă.
5. Dacă f și g sunt funcții convexe având același domeniu de definiție, atunci funcția h definită prin relația $h(x) = f(x) + g(x)$ este de asemenea o funcție convexă.
6. Dacă f și g sunt funcții convexe, iar $\text{dom}(f) \supseteq \text{image}(g)$, atunci funcția compusă $f \circ g$ este oare în mod necesar o funcție convexă? Dacă nu, atunci ce tip de restricții ar trebui să mai adăugăm pentru ca funcția $f \circ g$ să fie convexă? (Justificați răspunsul în mod riguros.)

Răspuns:

a. Vom lua pe rând afirmațiile 1, 2 și 3 și vom demonstra că fiecare dintre ele este *echivalentă* cu *definiția* noțiunii de funcție convexă. (Așa cum s-a cerut în enunț, în fiecare din cele trei cazuri, vom demonstra echivalența în ambele sensuri.) Așadar, cele trei afirmații constituie *caracterizări* ale noțiunii de funcție convexă.

• Definiția implică afirmația 1:

Luând $t = 1/2$ în definiția noțiunii de funcție convexă, obținem imediat afirmația 1.

• Afirmația 1 implică definiția:

Este [mai] ușor să demonstrăm această implicație prin *reducere la absurd*. Presupunem că există o funcție f care satisfacă afirmația 1 dar nu este convexă. Prin urmare, există o pereche de puncte x_0, y_0 astfel încât

$$f(t_0x_0 + (1 - t_0)y_0) > t_0f(x_0) + (1 - t_0)f(y_0) \quad (97)$$

pentru un anumit $t_0 \in [0, 1]$. Vom defini funcția g cu ajutorul relației

$$g(t) = f(tx_0 + (1 - t)y_0) - tf(x_0) - (1 - t)f(y_0),$$

pentru $t \in [0, 1]$. Se constată imediat că $g(0) = 0$ și $g(1) = 0$. Evident, $g(t_0) > 0$. În continuare, vom folosi notațiile $M = \max_{t \in (0, 1)} g(t)$ și $t^* = \min\{t \in (0, 1) | g(t) = M\}$. Aceasta înseamnă că t^* identifică cea mai mică valoare a lui t pentru care se atinge maximul funcției g pe intervalul $(0, 1)$.¹⁶⁸ Mai întâi vom arăta că și funcția g satisfacă „proprietatea punctului de mijloc“, după care, folosind această proprietate a lui g vom arăta că se obține o contradicție.

Pentru orice a și b din $[0, 1]$,

$$g\left(\frac{a+b}{2}\right) \stackrel{\text{def.}}{=} f\left(\frac{a+b}{2}x_0 + \left(1 - \frac{a+b}{2}\right)y_0\right) - \frac{a+b}{2}f(x_0) - \left(1 - \frac{a+b}{2}\right)f(y_0).$$

¹⁶⁸Conform presupunerii noastre, rezultă că există cel puțin un t astfel încat $g(t) > 0$. (Vedeți relația (97).)

Întrucât

$$\frac{a+b}{2}x_0 + \left(1 - \frac{a+b}{2}\right)y_0 = \frac{1}{2}[ax_0 + (1-a)y_0] + \frac{1}{2}[bx_0 + (1-b)y_0],$$

iar f satisface „proprietatea punctului de mijloc“ (afirmația 1), rezultă

$$\begin{aligned} g\left(\frac{a+b}{2}\right) &\leq \frac{1}{2}f(ax_0 + (1-a)y_0) + \frac{1}{2}f(bx_0 + (1-b)y_0) \\ &\quad - \frac{1}{2}af(x_0) - \frac{1}{2}bf(x_0) - \frac{1}{2}(1-a)f(y_0) - \frac{1}{2}(1-b)f(y_0) \\ &\stackrel{\text{def.}}{=} \frac{1}{2}g(a) + \frac{1}{2}g(b). \end{aligned} \quad (98)$$

Așadar, am arătat că funcția g satisface (și ea) „proprietatea punctului de mijloc“. Fie acum un număr $\varepsilon > 0$ suficient de mic pentru ca $(t^* + \varepsilon)$ și $(t^* - \varepsilon)$ să aparțină [încă] intervalului $(0, 1)$.¹⁶⁹ În fine, folosind aceste proprietăți obținem:

$$\begin{aligned} g(t^*) &= g\left(\frac{(t^* + \varepsilon) + (t^* - \varepsilon)}{2}\right) \\ &\stackrel{(98)}{\leq} \frac{1}{2}g(t^* + \varepsilon) + \frac{1}{2}g(t^* - \varepsilon) < \frac{M+M}{2} = M. \end{aligned}$$

Ultima inegalitate de mai sus este justificată de faptul că $g(t^* - \varepsilon) < M$ și $g(t^* + \varepsilon) \leq M$. Inegalitatea obținută, $g(t^*) < M$, contrazice presupunerea pe care am făcut-o inițial, și anume că $g(t^*) = M$. Prin urmare, presupunerea făcută etse falsă. Așadar, concluzionăm că orice funcție f care satisface afirmația 1 este cu necesitate [o funcție] convexă.

- Definiția implică afirmația 2:

Definiția noțiunii de funcție convexă spune că pentru orice x și y din \mathbb{R} și pentru orice $t \in [0, 1]$, are loc inegalitatea

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y),$$

ceea ce (pentru $t \neq 0$) conduce la următoarea inegalitate:

$$\frac{f(y + t(x-y)) - f(y)}{t} \leq f(x) - f(y).$$

Prin împărțirea și înmulțirea termenului stâng cu $(x-y)$, obținem:

$$\frac{f(y + t(x-y)) - f(y)}{t(x-y)}(x-y) \leq f(x) - f(y).$$

Prin urmare, făcând $t \rightarrow 0$, va rezulta că

$$f'(y)(x-y) \leq f(x) - f(y),$$

ceea ce este echivalent cu

$$f(y) + f'(y)(x-y) \leq f(x).$$

¹⁶⁹Este imediat că există un astfel de ε .

Așadar, a rezultat că afirmația 2 este adevărată.

- Afirmația 2 implică definiția:

Trebuie să demonstrăm că pentru orice x și y din \mathbb{R} și pentru orice $t \in [0, 1]$ are loc inegalitatea $f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y)$.

Considerând un z arbitrar ales din \mathbb{R} și aplicând afirmația 2 mai întâi pentru x și z (în locul lui x și y) și apoi pentru y și z (în locul lui x și y), vom obține:

$$f(x) \geq f(z) + f'(z)(x - z) \quad (99)$$

$$f(y) \geq f(z) + f'(z)(y - z) \quad (100)$$

Înmulțind prima dintre aceste relații cu t și pe cea de-a doua cu $(1 - t)$,¹⁷⁰ va rezulta:

$$tf(x) + (1 - t)f(y) \geq f(z) + f'(z)(t(x - z) + (1 - t)(y - z)) = f(z) + f'(z)(tx + (1 - t)y - z).$$

În particular, pentru $z = tx + (1 - t)y$, rezultă:

$$tf(x) + (1 - t)f(y) \geq f(z) = f(tx + (1 - t)y).$$

Observație: Atunci când $f : \mathbb{R}^d \rightarrow \mathbb{R}$, afirmația 2 devine: dacă f este funcție derivabilă (adică există toate derivatele sale parțiale de ordinul întâi), atunci funcția f este convexă dacă și numai dacă pentru orice $x, y \in \mathbb{R}^d$ are loc inegalitatea $f(x) \geq f(y) + \nabla f(y)^\top(x - y)$, unde $\nabla f(y)$ este vectorul gradient (adică vectorul de derivate parțiale) al lui f calculat în punctul y .

- Definiția implică afirmația 3:

Am arătat mai sus că definiția implică afirmația 2, deci pentru orice x și y are loc inegalitatea

$$f(y) \geq f(x) + f'(x)(y - x).$$

Conform teoremei lui Taylor,¹⁷¹ într-o vecinătate suficient de mică a lui x îl putem înlocui pe $f(y)$ cu suma dintre aproximarea sa ca polinom Taylor de ordinul al doilea (notat cu $P(y)$) și un termen care reprezintă eroarea la aproximare:

$$f(y) = \underbrace{f(x) + f'(x)(y - x) + f''(x)(y - x)^2}_{P(y)} + h_2(y)(y - x)^2,$$

unde $h_2(y) \rightarrow 0$ pentru $y \rightarrow x$. Prin urmare, [atunci când y este în respectiva vecinătate a lui x] inegalitatea precedentă se poate scrie astfel:

$$f''(x)(y - x)^2 + h_2(y)(y - x)^2 \geq 0.$$

Împărțind ambii membri ai acestei inegalități cu cantitatea $(y - x)^2$, care este strict pozitivă pentru orice $y \neq x$, și făcând apoi $y \rightarrow x$, rezultă că $f''(x) \geq 0$.

- Afirmația 3 implică definiția:

Întrucât funcția $f(x)$ este dublu derivabilă, putem scrie din nou aproximarea lui $f(y)$ ca polinom Taylor de ordinul al doilea:

$$f(y) = f(x) + f'(x)(y - x) + f''(x)(y - x)^2 + h_2(y)(y - x)^2,$$

¹⁷⁰Atât t cât și $1 - t$ sunt ≥ 0 .

¹⁷¹Vedeți https://en.wikipedia.org/wiki/Taylor's_theorem.

unde, ca și mai sus, $h_2(y) \rightarrow 0$ pentru $y \rightarrow x$. Afirmăția 3 spune că $f''(x) \geq 0$ pentru orice x . Teorema lui Taylor afirmă că $|h_2(y)(y-x)^2|$ crește mult mai lent decât $f''(x)(y-x)^2$. Prin urmare, [putem considera că] suma ultimilor doi termeni din membrul drept al egalității precedente este mai mare sau egală cu zero. Așadar, vom avea

$$f(y) = f(x) + f'(x)(y-x) + f''(x)(y-x)^2 + h_2(y)(y-x)^2 \geq f(x) + f'(x)(y-x).$$

Deci este satisfăcută afirmația 2 și, în concluzie, f este funcție convexă.

Observație: Atunci când $f : \mathbb{R}^d \rightarrow \mathbb{R}$, afirmația 3 devine: dacă f este funcție dublu derivabilă (adică există toate derivatele sale parțiale de ordinul al doilea), atunci funcția f este convexă dacă și numai dacă pentru orice $x \in \mathbb{R}^d$ matricea sa hessiană¹⁷² $\nabla^2 f(x)$ (adică, matricea derivelor parțiale de ordinul al doilea ale lui f calculate în punctul x) este *pozitiv semidefinită*, ceea ce înseamnă că pentru orice $v \in \mathbb{R}^d$ urmează că $v^\top \nabla^2 f(x)v \geq 0$.¹⁷³

b.4 Vom face demonstrația pornind de la definiția noțiunii de funcție convexă. Pentru orice x și y , putem scrie

$$\begin{aligned} h(tx + (1-t)y) &\stackrel{\text{def.}}{=} \max\{f(tx + (1-t)y), g(tx + (1-t)y)\} \\ &\leq \max\{tf(x) + (1-t)f(y), tg(x) + (1-t)g(y)\} \\ &\leq \max\{tf(x), tg(x)\} + \max\{(1-t)f(y), (1-t)g(y)\} \\ &= t \max\{f(x), g(x)\} + (1-t) \max\{f(y), g(y)\} \\ &\stackrel{\text{def.}}{=} th(x) + (1-t)h(y). \end{aligned}$$

Așadar, funcția h este convexă.

b.5 și de data aceasta vom face demonstrația pornind de la definiție. Pentru orice x și y , putem scrie

$$\begin{aligned} h(tx + (1-t)y) &\stackrel{\text{def.}}{=} f(tx + (1-t)y) + g(tx + (1-t)y) \\ &\leq tf(x) + (1-t)f(y) + tg(x) + (1-t)g(y) \\ &= t(f(x) + g(x)) + (1-t)(f(y) + g(y)) \\ &\stackrel{\text{def.}}{=} th(x) + (1-t)h(y). \end{aligned}$$

Prin urmare, funcția h este convexă.

Observație: Proprietățile b.4 și b.5 sunt valabile și pentru cazul general $f, g : \mathbb{R}^d \rightarrow \mathbb{R}$, nu doar pentru cazul $d = 1$.

b.6 Afirmăția din enunț nu este adevărată întotdeauna. Să luăm de exemplu funcțiile $g(x) = x^2$ și $f(x) = -x$. Ambele funcții sunt convexe, dar $f(g(x)) = -x^2$ este o funcție concavă!

În cele ce urmează, pentru a identifica un set de condiții suficiente astfel încât funcția $f \circ g$ să fie convexă, vom porni de la afirmația a.3. Conform acestei afirmații, atunci când f și g sunt funcții dublu derivabile, compunerea lor

¹⁷²După numele lui Ludwig Otto Hesse (1811 – 1874), matematician german.

¹⁷³Se poate constata ușor că pentru cazul $d = 1$ relația $v^\top \nabla^2 f(x)v \geq 0$ este echivalentă cu condiția $f''(x) \geq 0$.

(notată $f \circ g$) este funcție convexă dacă $[f(g(x))]'' \geq 0$ pentru orice x . Derivata de ordinul al doilea al funcției $f \circ g$ se calculează astfel:

$$\begin{aligned}[f(g(x))]'' &= [f'(g(x)) g'(x)]' = (f'(g(x)))' g'(x) + f'(g(x)) g''(x) \\ &= f''(g(x))(g'(x))^2 + f'(g(x))g''(x).\end{aligned}$$

Întrucât funcțiile f și g sunt convexe și dublu derivabile, rezultă că $f''(x) \geq 0$ și $g''(x) \geq 0$ pentru orice x . Așadar, primul termen din expresia (101) este mai mare sau egal cu zero. Pentru a ne asigura că și al doilea termen din această expresie este mai mare sau egal cu zero, trebuie să impunem condiția ca $f'(x) \geq 0$, ceea ce înseamnă că f este o funcție nedecrescătoare.

În concluzie, putem formula următoarea *proprietate*:

Dacă f și g sunt funcții convexe și dublu derivabile, iar f este funcție nedecrescătoare, rezultă că $f \circ g$ este de asemenea funcție convexă.

79.

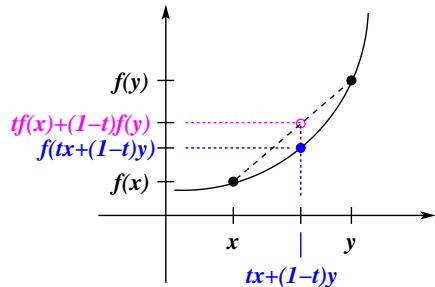
(Inegalitatea lui Jensen și câteva consecințe ale ei)

■ Liviu Ciortuz, 2020

Dacă $f : \mathbb{R} \rightarrow \mathbb{R}$ este o *funcție convexă*,¹⁷⁴ atunci, conform *definiției*,¹⁷⁵ pentru orice $t \in [0, 1]$ și orice $x_1, x_2 \in \mathbb{R}$ urmează

$$tf(x_1) + (1-t)f(x_2) \leq tf(x_1) + (1-t)f(x_2). \quad (101)$$

Dacă f este funcție strict convexă, atunci egalitatea are loc doar dacă $x_1 = x_2$.



a. Folosind definiția de mai sus, demonstrați *inegalitatea lui Jensen*:¹⁷⁵

Pentru orice $a_i \geq 0$, $i = 1, \dots, n$ cu $\sum_i a_i = 1$ și orice $x_i \in \mathbb{R}$, $i = 1, \dots, n$, dacă f este funcție convexă,¹⁷⁶ atunci

$$f\left(\sum_i a_i x_i\right) \leq \sum_i a_i f(x_i). \quad (102)$$

Mai general, pentru orice $a'_i \geq 0$, cu $i = 1, \dots, n$ și $\sum_i a'_i \neq 0$ avem

$$f\left(\frac{\sum_i a'_i x_i}{\sum_j a'_j}\right) \leq \frac{\sum_i a'_i f(x_i)}{\sum_j a'_j}. \quad (103)$$

Observații:

1. Dacă f este strict convexă, atunci în relațiile de mai sus egalitatea are loc doar dacă $x_1 = \dots = x_n$.

¹⁷⁴Vedeți problema 78.

¹⁷⁵Johan Jensen, inginer și matematician danez (1859–1925).

¹⁷⁶Dacă numerele a_i , cu $i = 1, \dots, n$, satisfac cele două proprietăți indicate, spunem că suma $\sum_i a_i x_i$ este o *combinație convexă* a numerelor x_1, \dots, x_n . Atunci când se folosește doar prima condiție ($a_i \geq 0$, $i = 1, \dots, n$), suma respectivă se numește *combinație canonică*. Când se folosește doar cea de-a doua condiție ($\sum_i a_i = 1$), suma se numește *combinație afină*. În fine, dacă se renunță la ambele condiții, suma se numește *combinație liniară*. (Cf. https://en.wikipedia.org/wiki/Linear_combination.)

2. Evident, rezultate similare cu cele de mai sus pot fi formulate și pentru funcții concave, înlocuind în relațiile (102) și (103) semnul \leq cu \geq .

b. Demonstrați *inegalitatea mediilor* folosind inegalitatea lui Jensen:¹⁷⁷

$$\frac{x_1 + x_2 + \dots + x_n}{n} \geq \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n} \quad \text{pentru orice } x_i \geq 0, i = 1, \dots, n.$$

c. În contextul teoriei probabilităților, inegalitatea lui Jensen este exprimată astfel: dacă X este o variabilă aleatoare și f este o funcție convexă, atunci $f(E[X]) \leq E[f(X)]$. (Similar, dacă f este funcție concavă, atunci $f(E[X]) \geq E[f(X)]$.)

Demonstrați această inegalitate în cazul în care X este variabilă aleatoare discretă cu un număr finit de valori (adică, $|Val(X)| < \infty$).

Răspuns:

a. Vom privi inegalitatea (102) ca pe o *ipoteză inductivă*, pe care o vom desemna cu $P(n)$, unde $n \in \mathbb{N}$, $n \geq 2$.¹⁷⁸ Vom demonstra că $P(n)$ este adevărată, folosind *principiul inducției matematice*:

$P(2)$: Dacă $x_1, x_2 \in \mathbb{R}$ și $a_1, a_2 \in [0, 1]$ astfel încât $a_1 + a_2 = 1$, iar f este funcție convexă, atunci inegalitatea $f(a_1x_1 + a_2x_2) \leq a_1f(x_1) + a_2f(x_2)$ se rescrie echivalent ca $f(a_1x_1 + (1 - a_1)x_2) \leq a_1f(x_1) + (1 - a_1)f(x_2)$. Această ultimă inegalitate coincide practic cu relația (101) din definiția funcției convexe, deci este adevărată.

Presupunem că proprietatea $P(n)$ este adevărată și vom demonstra $P(n+1)$, adică: dacă $x_1, \dots, x_n, x_{n+1} \in \mathbb{R}$ și $a_1, \dots, a_n, a_{n+1} \in [0, 1]$ astfel încât $\sum_{i=1}^{n+1} a_i = 1$, iar f o funcție convexă, atunci rezultă că

$$f\left(\sum_{i=1}^{n+1} a_i x_i\right) \leq \sum_{i=1}^{n+1} a_i f(x_i). \quad (104)$$

Vom rescrie acum membrul stâng al acestei inegalități într-o formă convenabilă:

$$\begin{aligned} f\left(\sum_{i=1}^{n+1} a_i x_i\right) &= f\left(\sum_{i=1}^n a_i x_i + a_{n+1} x_{n+1}\right) \\ &= f\left(\left(\sum_{i=1}^n a_i\right) \cdot \sum_{i=1}^n \frac{a_i}{\sum_{i=1}^n a_i} x_i + a_{n+1} x_{n+1}\right) \\ &= f\left((1 - a_{n+1}) \cdot \sum_{i=1}^n \frac{a_i}{1 - a_{n+1}} x_i + a_{n+1} x_{n+1}\right) \end{aligned} \quad (105)$$

Este imediat că dacă vom considera $A_i \stackrel{\text{not.}}{=} \frac{a_i}{1 - a_{n+1}}$, pentru $i = 1, \dots, n$, atunci rezultă că $A_i \geq 0$ pentru $i = 1, \dots, n$ și $\sum_{i=1}^n A_i = \frac{\sum_{i=1}^n a_i}{1 - a_{n+1}} = \frac{1 - a_{n+1}}{1 - a_{n+1}} = 1$.¹⁷⁹

¹⁷⁷Primul membru al *inegalității mediilor* este media aritmetică, iar cel de-al doilea este media geometrică.

¹⁷⁸Inegalitatea (102) se verifică și pentru $n = 1$, fiindcă $f(x_1) = f(x_1)$.

¹⁷⁹Evident, A_i este bine definit dacă $a_{n+1} \neq 1$. Dacă $a_{n+1} = 1$, atunci $a_1 = \dots = a_n = 0$, iar inegalitatea (104) devine în acest caz $f(x_{n+1}) \leq f(x_{n+1})$, care este adevărată în mod trivial.

Prin urmare,

$$\begin{aligned}
 f\left(\sum_{i=1}^{n+1} a_i x_i\right) &\stackrel{(105)}{=} f\left((1-a_{n+1}) \cdot \underbrace{\sum_{i=1}^n A_i x_i}_{x'_1} + a_{n+1} \underbrace{x_{n+1}}_{x'_2}\right) \\
 &\stackrel{f \text{ convexă}}{\leq} (1-a_{n+1}) \cdot f\left(\sum_{i=1}^n A_i x_i\right) + a_{n+1} \cdot f(x_{n+1}) \\
 &\stackrel{P(n)}{\leq} (1-a_{n+1}) \cdot \sum_{i=1}^n A_i f(x_i) + a_{n+1} \cdot f(x_{n+1}) \\
 &= (1-a_{n+1}) \cdot \left(\sum_{i=1}^n \frac{a_i}{1-a_{n+1}} f(x_i)\right) + a_{n+1} \cdot f(x_{n+1}) \\
 &= \sum_{i=1}^n a_i f(x_i) + a_{n+1} f(x_{n+1}) = \sum_{i=1}^{n+1} a_i f(x_i).
 \end{aligned}$$

Așadar, $f\left(\sum_{i=1}^{n+1} a_i x_i\right) \leq \sum_{i=1}^{n+1} a_i f(x_i)$, deci proprietatea $P(n+1)$ este adevărată.

Sumarizând, din faptul că $P(2)$ este adevărată, iar implicația $P(n) \Rightarrow P(n+1)$ este adevărată, conform principiului inducției complete rezultă că proprietatea $P(n)$ este adevărată pentru orice $n \in \mathbb{N}^* \setminus \{1\}$.

b. Fie numerele x_1, x_2, \dots, x_n , toate mai mari sau egale cu 0. Considerăm $a_1 = a_2 = \dots = a_n = \frac{1}{n}$ (ceea ce implică $\sum_{i=1}^n a_i = 1$). Conform inegalității lui Jensen, în care vom alege pe postul funcției f funcția \ln (logaritmul având ca bază numărul e) care este concavă, putem scrie:

$$\begin{aligned}
 \ln\left(\sum_{i=1}^n a_i x_i\right) &\geq \sum_{i=1}^n a_i \ln(x_i) \Leftrightarrow \ln\left(\frac{1}{n} \sum_{i=1}^n x_i\right) \geq \frac{1}{n} \sum_{i=1}^n \ln(x_i) \Leftrightarrow \\
 \ln\left(\frac{1}{n} \sum_{i=1}^n x_i\right) &\geq \frac{1}{n} \ln\left(\prod_{i=1}^n x_i\right) \Leftrightarrow \ln\left(\frac{1}{n} \sum_{i=1}^n x_i\right) \geq \ln\left(\prod_{i=1}^n x_i\right)^{\frac{1}{n}} \Leftrightarrow \\
 \ln\left(\frac{1}{n} \sum_{i=1}^n x_i\right) &\geq \ln(\sqrt[n]{x_1 x_2 \dots x_n}) \Leftrightarrow \frac{\sum_{i=1}^n x_i}{n} \geq \sqrt[n]{x_1 x_2 \dots x_n}.
 \end{aligned}$$

Ultima echivalentă are loc întrucât funcția \ln este strict crescătoare.

c. Fie $f : \mathbb{R} \rightarrow \mathbb{R}$ o funcție convexă și X o variabilă aleatoare discretă având următorul *tablou de repartīție*:

$$X : \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix}.$$

Conform inegalității lui Jensen (pe care o putem aplica luând în locul „probabilităților” a_i probabilitățile p_i , întrucât $p_i \geq 0$ și $\sum_{i=1}^n p_i = 1$), rezultă:

$$f\left(\sum_{i=1}^n p_i x_i\right) \leq \sum_{i=1}^n p_i f(x_i) \Leftrightarrow f(E_P[X]) \leq E_P[f(X)].$$

În mod similar, dacă f este funcție concavă, rezultă că $f(E_P[X]) \geq E_P[f(X)]$.

80. (Găsirea optimului unei funcții reale de gradul al doilea, folosind metoda analitică, metoda gradientului descendente și metoda lui Newton)

■ • *University of Utah, 2008 fall, Hal Daumé III, HW4, pr. 1*

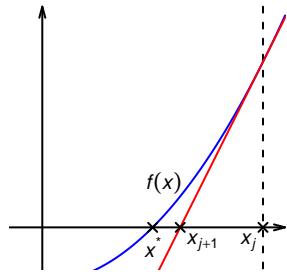
Să presupunem că dorim să găsim minimul funcției $f(x) = 3x^2 - 2x + 1$.

- Verificați că această funcție este convexă.
- Ca urmare a punctului precedent, rezultă că funcția f are un minim global. Găsiți acest minim, folosind cunoștințe de analiza matematică.
- La acest punct vom căuta optimul funcției f aplicând algoritmul gradientului descendente.¹⁸⁰ Efectuați trei pași ai acestui algoritm pornind de la punctul initial $x_0 = 1$ și folosind rata de învățare $\eta = 0.1$. Cât de aproape a ajuns algoritmul de soluția reală?
- În sfârșit, căutați din nou optimul funcției f aplicând de această dată metoda lui Newton,¹⁸¹ pornind tot de la punctul $x_0 = 1$. Ce observați? (Câte iterații sunt necesare pentru ca algoritmul să conveargă la soluția optimă?)

Comentariu [Metoda lui Newton]:

În forma sa cea mai simplă, metoda lui Newton — cunoscută în literatura de specialitate și sub numele de *metoda tangentei* — are ca obiectiv identificarea rădăcinii (sau, a rădăcinilor) unei funcții f care este convexă și derivabilă. (Rădăcinile unei funcții sunt acele valori x^* ale argumentului funcției pentru care se anulează funcția respectivă, adică $f(x^*) = 0$.) În acest scop, adică pentru aflarea rădăcinilor unei funcții, metoda lui Newton, care este o metodă iterativă, folosește următoarea relație de „actualizare“:

$$x_{j+1} = x_j - \frac{f(x_j)}{f'(x_j)},$$



cu x_0 ales în mod arbitrar. Facem *observația* că metoda lui Newton poate fi aplicată (modificată ușor) și în scopul găsirii optimului unei funcții. În acest caz, se caută valorile argumentului x pentru care se anulează prima derivată a funcției obiectiv f . Prin urmare, pentru a putea aplica metoda lui Newton se cere ca funcția f să fie dublu derivabilă (adică să existe atât prima cât și cea de-a doua derivată a lui f). Relația de actualizare acum devine:

$$x_{j+1} = x_j - \frac{f'(x_j)}{f''(x_j)}. \quad (106)$$

La problema de față, se folosește metoda lui Newton în varianta aceasta, adică pentru aflarea optimului (mai precis, a minimului) unei funcții.

Facem *observația* că formula (106) se folosește exact în aceeași formă și atunci când — în locul minimului — se caută maximul unei funcții, folosind *metoda lui Newton*. (Pentru a vă convinge, este suficient să înlocuiți f cu $-f$ în

¹⁸⁰Pentru o prezentare formală a acestui algoritm, vedeti enunțul problemei 164.

¹⁸¹Isaac Newton (1642-1727) a fost un matematician, fizician, astronom, alchimist și teolog englez.

relația (106).) În schimb, relația (107), care este folosită de metoda *gradientului descendant* (pentru căutarea unui punct de minim) trebuie înlocuită cu $\Delta x = +\eta \cdot f'(x)$ în cazul folosirii metodei *gradientului ascendent* (pentru căutarea unui punct de maxim).

Generalizarea metodei lui Newton la cazul multidimensional, numită și metoda Newton-Raphson,¹⁸² lucrează cu relația de actualizare

$$x_{j+1} = x_j - (H(x_j))^{-1} \nabla f(x_j).$$

Aici, $\nabla f(x)$ este, ca de obicei, vectorul gradient, format din derivatele parțiale ale lui $f(x)$, calculate în punctul x , iar $H(x)$ este *matricea hessiană* a lui f , formată din derivatele parțiale de ordinul al doilea, calculate în punctul x :

$$H_{ik}(x) = \frac{\partial^2 f(x)}{\partial x_i \partial x_k}.$$

În mod obișnuit, metoda lui Newton are o rată / viteză de convergență mai mare decât metoda gradientului (varianta “batch”) și necesită mult mai puține iterații pentru a ajunge foarte aproape de punctul de optim. Totuși, trebuie reținut că o [singură] iterație a algoritmului lui Newton poate fi mai costisitoare decât o iterație a metodei gradientului, fiindcă metoda lui Newton necesită atât calcularea cât și inversarea matricei hessiene. Cu toate acestea, pentru valori ale lui d (numărul de dimensiuni) nu prea mari, metoda lui Newton este în mod obișnuit mult mai rapidă decât metoda gradientului.

Răspuns:

a. Pentru a studia convexitatea funcției $f(x)$ se calculează derivata a două:

$$f''(x) = 6 > 0, \forall x \in \mathbb{R}.$$

Conform proprietății a.3 de la problema 78, rezultă că funcția f este convexă (de fapt, ea este *strict convexă*) pe întreg domeniul ei de definiție. Așadar, ea are un (singur) punct de minim.

¹⁸²Joseph Raphson (c.1668 - c.1715) a fost un matematician englez, contemporan cu Isaac Newton.

b. Pentru a calcula minimul funcției $f(x) = 3x^2 - 2x + 1$ utilizăm derivata de ordinul întâi:

$$f'(x) = 6x - 2.$$

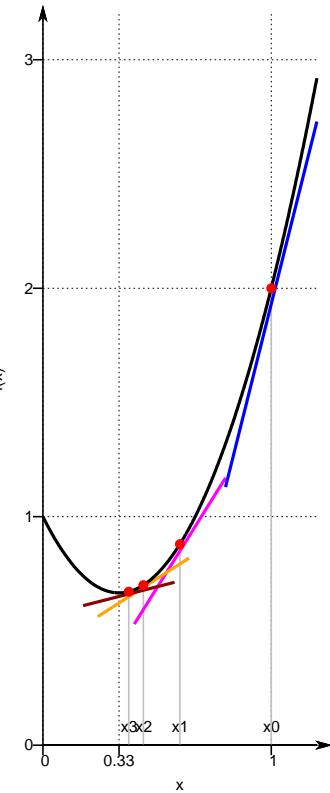
Punctul de minim este dat de soluția ecuației $f'(x) = 0$, și anume: $x = \frac{1}{3} \approx 0.33$.

c. Având punctul initial $x_0 = 1$ și $\eta = 0.1$, se poate aplica metoda gradientului descendente conform formulelor:

$$\begin{aligned} x_{n+1} &\leftarrow x_n + \Delta x \\ \Delta x &= -\eta \cdot f'(x) \end{aligned} \quad (107)$$

Primii trei pași ai algoritmului sunt următorii:

$$\begin{aligned} x_1 &= x_0 - \eta \cdot f'(x_0) = 1 - 0.1(6 - 2) \\ &= 1 - 0.4 = 0.6 \\ x_2 &= x_1 - \eta \cdot f'(x_1) = 0.6 - 0.1(6 \cdot 0.6 - 2) \\ &= 0.6 - 0.1 \cdot 1.6 = 0.6 - 0.16 = 0.44 \\ x_3 &= x_2 - \eta \cdot f'(x_2) = 0.44 - 0.1(6 \cdot 0.44 - 2) \\ &= 0.44 - 0.1 \cdot 0.64 = 0.44 - 0.064 = 0.376 \end{aligned}$$



Diferența dintre valoarea obținută după efectuarea a trei pași de gradient descendente și valoarea pre-calculată a punctului de minim este: $0.376 - 0.333 = 0.043$.

Observații: Se constată ușor că

i. apropierea de punctul de optim este [mai] rapidă atâtă timp cât valoarea primei derive (i.e., panta tangentei la graficul funcției) este mare în valoare absolută;

ii. dacă rata de învățare η a fost fixată la o valoare prea mare, atunci este posibil ca la un moment dat să depășim punctul de optim și apoi să „pendulăm“ în jurul lui. Acest punct slab al metodei gradientului poate fi contracararat reducând în mod dinamic mărimea lui η . Altăminteri, metoda gradientului are avantajul de a fi o tehnica de optimizare foarte simplă din punct de vedere conceptual și ușor de implementat.¹⁸³

d. La aplicarea metodei lui Newton pentru funcția dată în enunț, la prima iterare vom proceda astfel:

$$x_1 = x_0 - \frac{f'(x_0)}{f''(x_0)} = 1 - \frac{6 \cdot 1 - 2}{6} = 1 - \frac{2}{3} = \frac{1}{3}.$$

Se observă că $f'(x_1) = 0$, iar dacă am mai continua să facem și alte iterării, am obține $x_2 = x_3 = \dots = x_1$. Așadar, cu metoda lui Newton, soluția optimă se obține (pentru această funcție) într-o singură iterare.

¹⁸³Alte două puncte slabe ale metodei gradientului descendente sunt: imposibilitatea de a garanta găsirea optimului global, și numărul mare de iterări care trebuie executate pe unele seturi de date reale.

Observație: Se poate constata ușor că soluția optimă pentru funcția f se obține într-o singură iterație, indiferent de valoarea atribuită lui x_0 :

$$x_1 = x_0 - \frac{f'(x_0)}{f''(x_0)} = x_0 - \frac{6x_0 - 2}{6} = \frac{1}{3}.$$

Este interesant de reținut faptul că această proprietate, care este valabilă de fapt pentru orice funcție polinomială de gradul al doilea,¹⁸⁴ se întâlnește și în cazul rezolvării *regresiei liniare* cu ajutorul metodei lui Newton (vedeți problema 7.b de la capitolul *Metode de regresie*).

81.

(Noțiunile de *subderivată*, *subdiferențială* și *subgradient*: definiții, exemplificare)

*prelucrare de Liviu Ciortuz, după
□ • ○ CMU, 2009 fall, Carlos Guestrin, HW2, pr. 2.b
CMU, 2016 fall, N. Balcan, M. Gormley, HW4, pr. 3.a*

a. Fie $C \subseteq \mathbb{R}$ o mulțime convexă și $f : C \rightarrow \mathbb{R}$ o funcție convexă.¹⁸⁵ Numim *subderivată* lui f într-un punct $x_0 \in C$ un număr c astfel încât proprietatea $f(x) - f(x_0) \geq c(x - x_0)$ este satisfăcută pentru orice $x \in C$.¹⁸⁶ Cu alte cuvinte, un număr c este subderivată a funcției f în punctul x_0 dacă dreapta care trece prin punctul $(x_0, f(x_0))$ și are panta c este situată sub graficul funcției f .

Observație: Se poate demonstra că în cazul în care funcția f este derivabilă la stânga și la dreapta în punctul $x_0 \in C$, mulțimea subderivatelor funcției f în punctul x_0 este un interval închis, nevid, de forma $[c_1, c_2]$, unde:

$$c_1 = \lim_{x \nearrow x_0} \frac{f(x) - f(x_0)}{x - x_0} \text{ și } c_2 = \lim_{x \searrow x_0} \frac{f(x) - f(x_0)}{x - x_0}.$$

Mulțimea tuturor subderivateelor lui f în punctul x_0 (în cazul de mai sus, intervalul $[c_1, c_2]$) se numește *subdiferențială* lui f în x_0 și se notează cu $\partial f(x_0)$.¹⁸⁷

i. Fie funcția $f(x) = |x|$.¹⁸⁸ Care este mulțimea subderivateelor lui f în punctul 0?

ii. Folosind următoarea

Teoremă: Punctul x_0 este punct de minim global al funcției f dacă și numai dacă $0 \in \partial f(x_0)$.

¹⁸⁴Considerând $f(x) = ax^2 + bx + c$, rezultă $f'(x) = 2ax + b$ și $f''(x) = 2a$. Așadar, la aplicarea regulii [de actualizare] din metoda lui Newton, vom obține:

$$x_{j+1} = x_j - \frac{f'(x_j)}{f''(x_j)} = x_j - \frac{2ax_j + b}{2a} = x_j - x_j - \frac{b}{2a} = -\frac{b}{2a},$$

adică exact abscisa punctului de optim al funcției de gradul al doilea.

Proprietatea aceasta — adică faptul că metoda lui Newton converge într-o singură iterare — se generalizează la funcții de gradul al doilea cu un număr oarecare de variabile (nu doar una singură, cum a fost cazul mai sus); vedeți problema 169.

¹⁸⁵Pentru definițiile noțiunilor de *mulțime convexă* și *funcție convexă*, vedeți problema 78.

¹⁸⁶Vedeți proprietatea similară de la problema 78.a.2.

¹⁸⁷În cazul funcțiilor derivabile, subdiferențiala practic coincide cu derivata. Pentru alte câteva proprietăți ale subdiferențialei, vedeți <https://en.wikipedia.org/wiki/Subderivative>.

¹⁸⁸În legătură cu această funcție, la problema 162.b se cere să se arate că este funcție convexă.

arătați că $x_0 = 0$ este punctul [unic] de minim global al funcției $f(x) = |x|$.

b. Funcția de cost / pierdere *hinge* este definită astfel:¹⁸⁹

$$L(y, y') = \max(0, 1 - yy') \text{ pentru orice } y, y' \in \mathbb{R}.$$

Fie $x \in \mathbb{R}^d$ și $y \in \mathbb{R}$. Cu ajutorul lui L , definim o nouă funcție: $F(w) = L(y, w \cdot x)$, pentru orice $w \in \mathbb{R}^d$. Calculați subdiferențiala lui F în w (considerând de data aceasta că w este fixat în \mathbb{R}^d). *Indicație:* Trebuie să tratați 3 cazuri.

Notă: Noțiunile de subderivată și subdiferențială pot fi generalizate în mod natural la cazul funcțiilor de mai multe variabile. Dacă $f : U \rightarrow \mathbb{R}$ este o funcție convexă cu valori reale definită pe o mulțime convexă deschisă din spațiul euclidian \mathbb{R}^n , un vector v din acest spațiu este un *subgradient* al funcției f în punctul $x_0 \in U$ dacă pentru orice $x \in U$ are loc inegalitatea

$$f(x) - f(x_0) \geq v \cdot (x - x_0),$$

unde operatorul \cdot desemnează produsul scalar. Mulțimea tuturor subgradenților lui f în punctul x_0 se numește *subdiferențiala* lui f în x_0 și se notează cu $\partial f(x_0)$.

Observație: Pentru folosirea noțiunii / metodei subgradientului la conceperea [și implementarea] unor algoritmi de învățare automată, vedeti problema 168 de la prezentul capitol, problema 27 de la capitolul *Metode de regresie* și problemele 20, 21 și 51 de la capitolul *Mașini cu vectori-suport*.

Răspuns:

a.i. Calculăm cele două limite indicate în enunț:

$$\lim_{x \nearrow 0} \frac{f(x) - f(x_0)}{x - 0} = \lim_{x \nearrow 0} \frac{|x| - 0}{x} = \lim_{x \nearrow 0} \frac{-x}{x} = -1$$

$$\lim_{x \searrow 0} \frac{f(x) - f(x_0)}{x - 0} = \lim_{x \searrow 0} \frac{|x| - 0}{x} = \lim_{x \searrow 0} \frac{x}{x} = +1.$$

Așadar, subdiferențiala funcției $|x|$ în punctul 0 este intervalul $[-1, +1]$.

a.ii. Este imediat că

$$\partial f(x_0) = \begin{cases} -1 & \text{dacă } x_0 < 0 \\ 1 & \text{dacă } x_0 > 0. \end{cases}$$

Conform rezultatului de la punctul a.i., rezultă că $0 \in \partial f(0)$. Prin urmare, $x_0 = 0$ este punctul de minim global al funcției $f(x) = |x|$.

b. Fie $y, y' \in \mathbb{R}$, ca în enunț. Pentru a ușura calculele, vom nota $z = yy'$ și vom defini funcția $L_1(z) = L(y, y') = \max(0, 1 - z)$.¹⁹⁰ Este imediat că

$$\frac{\partial}{\partial z} L_1(z) = \begin{cases} -1 & \text{dacă } z < 1 \\ 0 & \text{dacă } z > 1. \end{cases}$$

Mulțimea subderivatelor lui L_1 în punctul $z = 1$ este intervalul $[-1, 0]$.

¹⁸⁹Pentru exemplificarea modului în care funcția de cost *hinge* este folosită în clasificarea automată, vedeti problema 20 de la capitolul *Mașini cu vectori-suport*.

¹⁹⁰Puteți vedea graficul acestei funcții la problema 88.

Acum, renotând $z \stackrel{\text{def.}}{=} yw \cdot x$, putem scrie:¹⁹¹

$$\frac{\partial}{\partial w} F(w) = \frac{\partial}{\partial w} L(y, w \cdot x) = \frac{\partial}{\partial w} L_1(yw \cdot x) = \frac{\partial}{\partial z} L_1(z) \frac{\partial}{\partial w} z,$$

în cazul în care $z \neq 1$. Prin urmare,

$$\frac{\partial}{\partial w} F(w) = \frac{\partial}{\partial w} \max(0, 1 - yw \cdot x) = \begin{cases} -yx & \text{dacă } yw \cdot x < 1 \\ 0 & \text{dacă } yw \cdot x \geq 1. \end{cases}$$

În cazul vectorilor w_0 pentru care $yw_0 \cdot x = 1$, subdiferențiala lui F în w_0 este formată din toți vectorii $v \in \mathbb{R}^d$ care au proprietatea

$$\begin{aligned} F(w) - F(w_0) \geq v \cdot (w - w_0) &\Leftrightarrow L_1(yw \cdot x) - L_1(yw_0 \cdot x) \geq v \cdot (w - w_0) \\ &\Leftrightarrow L_1(yw \cdot x) - \underbrace{L_1(1)}_0 \geq v \cdot (w - w_0) \Leftrightarrow L_1(yw \cdot x) \geq v \cdot (w - w_0) \\ &\Leftrightarrow \max(0, yw \cdot x) \geq v \cdot (w - w_0). \end{aligned}$$

Se observă că vectorul $v = 0$ din \mathbb{R}^d satisface inegalitatea de mai sus. Așadar, pentru aplicarea metodei subgradientului în conjuncție cu funcția de cost *hinge* definită prin $\max(0, yw \cdot x)$, este *suficient* (!) să considerăm

$$\frac{\partial}{\partial w} \max(0, 1 - yw \cdot x) = \begin{cases} -yx & \text{dacă } yw \cdot x < 1 \\ 0 & \text{dacă } yw \cdot x \geq 1. \end{cases}$$

În cazul $d = 1$, adică atunci când $w, x \in \mathbb{R}$, subdiferențiala acestei funcții *hinge* în punctul w_0 care satisface proprietatea $yw_0x = 1$ este intervalul $[-yx, 0]$.

82.

(Problema de optimizare [convexă] cu restricții — o variantă ușor simplificată: demonstrarea proprietății de dualitate „slabă“)

■ □ • ○ prelucrare de Liviu Ciortuz, după CMU, 2014 fall, E. Xing, B. Poczos, HW1, pr. 3.1

Introducere: În cazul general, o problemă de optimizare convexă cu restricții este definită sub forma primală astfel:

$$\begin{aligned} &\min_x f(x) \\ \text{a. î. } &g_i(x) \leq 0 \text{ pentru } i = 1, \dots, m \\ &h_j(x) = 0 \text{ pentru } j = 1, \dots, k, \end{aligned}$$

unde f (funcția obiectiv) și g_i sunt funcții convexe, iar h_j sunt funcții *afine* (adică funcții liniare augmentate cu termen liber), cu $m \geq 0$ și $k \geq 0$. Condițiile $g_i(x) \leq 0$ și $h_j(x) = 0$ se numesc *condiții de fezabilitate primală*.

Pentru problema de optimizare definită mai sus, *lagrangeanul generalizat* se definește astfel:¹⁹²

$$L(x, \alpha, \beta) \stackrel{\text{def.}}{=} f(x) + \sum_i \alpha_i g_i(x) + \sum_j \beta_j h_j(x),$$

¹⁹¹Este ușor de arătat că $\frac{\partial}{\partial w} z = \frac{\partial}{\partial w} yw \cdot x = yx$.

¹⁹²După numele matematicianului și astronomului Joseph-Louis Lagrange (în italiană, Giuseppe Lodovico Lagrangia), 1736-1813. El a lucrat mai întâi în Italia (la Torino, orașul său natal), apoi în Germania (la Berlin) și în final în Franța (la Paris).

unde $\alpha_i \geq 0$ pentru $i = 1, \dots, m$ și $\beta_j \in \mathbb{R}$ pentru $j = 1, \dots, k$. Condițiile $\alpha_i \geq 0$ se numesc *condiții de fezabilitate duală*. Unii autori numesc funcția *L lagrangeanul primal* și o (re)notează cu L_P .

Observație (1): Condiția de convexitate asupra funcțiilor f și g_i , precum și condiția ca funcțiile h_j să fie afine nu sunt necesare în demonstrațiile care urmează. Le-am scris aici pentru a asigura un cadru comun cu problema care urmează (83).

Fie următoarea problemă de optimizare cu restricții în formă primală, unde $f, h_1, h_2 : \mathbb{R}^d \rightarrow \mathbb{R}$:

$$\begin{aligned} & \min_x f(x) \\ \text{a. î. } & h_1(x) \leq 0, \\ & h_2(x) = 0. \end{aligned}$$

a. Scrieți *lagrangeanul generalizat* $L(x, \lambda, u)$, unde $\lambda \geq 0$ și $u \in \mathbb{R}$ sunt *variabilele duale* care corespund inegalității și respectiv egalității din cadrul problemei de optimizare.

b. Vom nota cu P așa-numita *regiunea fezabilă* a problemei de optimizare în formă primală, adică mulțimea formată din acele puncte $x \in \mathbb{R}^d$ pentru care $h_1(x) \leq 0$ și $h_2(x) = 0$.

Demonstrați că pentru orice $x \in P$ are loc egalitatea

$$f(x) = \max_{\lambda \geq 0, u} L(x, \lambda, u).$$

Observație (2): Se poate demonstra că pentru orice problemă de optimizare convexă (vedeți *Introducerea* de mai sus), adică în care funcțiile g_i sunt convexe, iar funcțiile h_j sunt afine, regiunea fezabilă este mulțimea convexă.

c. Datorită rezultatului de la punctul b, problema de optimizare dată în enunț (în formă primală) se poate scrie echivalent (în mod compact) astfel:

$$\min_{x \in P} \max_{\lambda \geq 0, u} L(x, \lambda, u).$$

Mai departe, renunțând la restricția $x \in P$ — adică, lucrând cu $x \in \mathbb{R}^d$, nerestricționat¹⁹³ — și inversând cei doi operatori, min și max, vom obține problema de optimizare convexă în formă duală:¹⁹⁴

$$\max_{\lambda \geq 0, u} \min_{x \in \mathbb{R}^d} L(x, \lambda, u).$$

Putem spune că funcția obiectiv a problemei duale este (în manieră „concepțională“) $g(\lambda, u) \stackrel{\text{not.}}{=} \min_{x \in \mathbb{R}^d} L(x, \lambda, u)$.

Arătați că

$$\min_{x \in P} f(x) \geq \max_{\lambda \geq 0, u} g(\lambda, u).$$

Atfel spus, valoarea optimă [a funcției obiectiv] din forma primală a problemei de optimizare este mai mare sau egală cu valoarea optimă [a funcției obiectiv]

¹⁹³De fapt, se poate constata — pentru justificare (chiar pentru cazul general al problemei de optimizare cu restricții), vedeți documentul *Support Vector Machines* de Andrew Ng, Stanford University, CS229 Lecture Notes, Part V, pag. 8-9 — că problema de optimizare $\min_{x \in P} \max_{\lambda \geq 0, u} L(x, \lambda, u)$ este echivalentă cu problema de optimizare $\min_{x \in \mathbb{R}^d} \max_{\lambda \geq 0, u} L(x, \lambda, u)$.

¹⁹⁴Observați că restricțiile asociate formei duale a problemei de optimizare (și anume, $\lambda \geq 0$) sunt mult mai simple decât cele din forma primală!

din problema duală. Veți putea constata ușor că acest rezultat se menționează și atunci când în problema de optimizare avem mai multe restricții de tip inegalitate sau egalitate. Proprietatea aceasta se numește *dualitate slabă* (engl., *weak duality*).

Observație (3): Se poate demonstra că pentru orice problemă de optimizare convexă (vedeți *Introducerea* de mai sus), adică în care funcțiile f și g_i sunt convexe, iar funcțiile h_j sunt affine, funcția $g(\lambda, u) \stackrel{\text{not.}}{=} \min_{x \in \mathbb{R}^d} L(x, \lambda, u)$ este concavă,¹⁹⁵ deci putem să-i calculăm maximul aplicând o metodă de optimizare convexă asupra funcției $-g(\lambda, u)$.

Răspuns:

a. Lagrangeanul generalizat se obține combinând (într-o singură funcție) *funcția obiectiv* și funcțiile cu ajutorul cărora se scriu *restricțiile* din problema de optimizare convexă:

$$L(x, \lambda, u) \stackrel{\text{def.}}{=} f(x) + \lambda h_1(x) + u h_2(x), \text{ unde } \lambda \in \mathbb{R}_+ \text{ și } u \in \mathbb{R}.$$

b. Dacă $x \in P$, regiunea fezabilă a problemei de optimizare în forma primală, atunci urmează că

$$f(x) \geq f(x) + \underbrace{\lambda}_{\geq 0} \underbrace{h_1(x)}_{\leq 0} + u \underbrace{h_2(x)}_{=0} \stackrel{\text{def.}}{=} L(x, \lambda, u) \text{ pentru orice } \lambda \geq 0 \text{ și orice } u \in \mathbb{R}.$$

Mai departe, se observă că inegalitatea $f(x) \geq L(x, \lambda, u)$, pe care tocmai am obținut-o, este satisfăcută cu egalitate dacă se ia $\lambda = 0$. Prin urmare, putem scrie $f(x) = \max_{\lambda \geq 0, u} L(x, \lambda, u)$.

c. Vom relua acum inegalitatea

$$f(x) \geq L(x, \lambda, u), \text{ pentru orice } x \in P, \lambda \geq 0, u \in \mathbb{R},$$

pe care am obținut-o la punctul b. Considerând acum că $\lambda \geq 0$ și $u \in \mathbb{R}$ sunt [aleși în mod arbitrar dar] fixați, și aplicând apoi operatorul $\min_{x \in P}$ la ambii membri ai acestei inegalități, vom obține

$$\min_{x \in P} f(x) \geq \min_{x \in P} L(x, \lambda, u) \geq \min_{x \in \mathbb{R}^d} L(x, \lambda, u) \stackrel{\text{not.}}{=} g(\lambda, u).$$

Ultima inegalitate se justifică prin faptul că $P \subseteq \mathbb{R}^d$. Așadar, summarizând, ceea ce am obținut până acum este inegalitatea

$$\min_{x \in P} f(x) \geq g(\lambda, u) \text{ pentru orice } \lambda \geq 0 \text{ și orice } u \in \mathbb{R}.$$

În particular, inegalitatea are loc pentru acei $\lambda \geq 0$ și $u \in \mathbb{R}$ pentru care se atinge maximul termenului drept al inegalității. Așadar,

$$\min_{x \in P} f(x) \geq \max_{\lambda \geq 0, u} g(\lambda, u).$$

Comentariu: Folosind notația din documentul *Support Vector Machines* de Andrew Ng, putem rescrie într-o manieră foarte sugestivă inegalitatea pe care tocmai am obținut-o (adică, proprietatea de *dualitate slabă*):

$$p^* \geq d^*,$$

unde prin p^* și d^* am notat *valorile optime* pentru problema primală, respectiv problema duală. Este imediat că proprietatea care a fost demonstrată în această problemă are loc pentru *orice* problemă de optimizare cu restricții.

¹⁹⁵Vedeți documentul *Convex Optimization Overview (cont'd)* de Chuong B. Do, 2009.

83. (Demonstrarea unei părți din **teorema Karush-Kuhn-Tucker**, folosind puncte-șă:
dacă sunt îndeplinite condițiile Karush-Kuhn-Tucker,
atunci dispunem de o soluție pentru problema primală)

prelucrare de Liviu Ciortuz, 2019, după CMU, 2015 fall, A. Smola, B. Poczos, HW3, pr. 4

Fie problema de optimizare convexă cu restricții

$$\begin{aligned} & \min_x f(x) \\ \text{a. i. } & g_i(x) \leq 0 \text{ pentru } i = 1, \dots, m \\ & h_j(x) = 0 \text{ pentru } j = 1, \dots, k, \end{aligned}$$

unde $x \in \mathbb{R}^d$, funcțiile f, g_1, \dots, g_m sunt convexe și derivabile în raport cu x , iar h_1, h_2, \dots, h_k sunt funcții affine (adică funcții liniare augmentate cu termeni liberi). În cele ce urmează, această problemă de optimizare convexă va fi desemnată cu (P).¹⁹⁶

Comentariu:

Pentru o problemă de optimizare convexă cu restricții, care este definită în forma primală (P), condițiile Karush-Kuhn-Tucker sunt următoarele:

- fezabilitate $\begin{cases} \text{primală} & g_i(x^*) \leq 0 \text{ pentru } i = 1, \dots, m \text{ și} \\ & h_j(x^*) = 0 \text{ pentru } j = 1, \dots, k \\ \text{duală} & \alpha_i^* \geq 0 \text{ pentru } i = 1, \dots, m \end{cases}$
- complementaritate [duală]: $\alpha_i^* g_i(x^*) = 0$ pentru $i = 1, \dots, m$
- staționaritate (sau, optimalitate): $\frac{\partial}{\partial x} L(x^*, \alpha^*, \beta^*) = 0,$

unde $L(x, \alpha, \beta)$ este lagrangeanul generalizat pentru problema de optimizare convexă cu restricții (P). (Vedeți *Introducerea la problema 82.*) Precizăm că orice punct în care se anulează $\frac{\partial}{\partial x} L$ se numește *punct de staționaritate*.

Teorema Karush-Kuhn-Tucker¹⁹⁷ afirmă următoarele:

- (1) Dacă este îndeplinită condiția de *dualitate tare* pentru problemele de optimizare (P) și duala sa (D) — adică valorile lor optime satisfac relația $p^* = d^*$ —, atunci soluțiile acestor două probleme satisfac condițiile Karush-Kuhn-Tucker.¹⁹⁸
- (2) Dacă x^* , α^* și β^* satisfac condițiile Karush-Kuhn-Tucker, atunci x^* este soluție a problemei (P), iar α^* , β^* constituie o soluție pentru problema (D).

Observație importantă: Sunt cunoscute anumite *condiții suficiente* pentru ca să aibă loc dualitatea tare ($p^* = d^*$) și, în consecință să fie satisfăcute condițiile Karush-Kuhn-Tucker. Aceste condiții suficiente sunt numite „condiție de regularitate“ (engl., *regularity constraint*, sau *constraint qualification*). Una dintre ele, pe care o vom folosi adeseori, se

¹⁹⁶Vă readucem aminte că această *formă* a problemei de optimizare convexă se numește [forma] *primală*.

¹⁹⁷Teorema Karush-Kuhn-Tucker (abreviat, KKT) are o istorie interesantă. Ea a fost descoperită în 1939 de către studentul american William Karush, care a inclus-o în teza sa de master, dar n-a fost cunoscută până când a fost redescoperită în 1950 de către matematicienii Harold Kuhn (american) și Albert Tucker (canadian). O variantă a acestei teoreme fusese descoperită și de către Fritz John (german) în 1948, însă *Duke Mathematical Journal* i-a refuzat publicarea. Vedeți Chuong B. Do, *Convex Optimization Overview (cont'd)*, 2009, pag. 7.

¹⁹⁸Pentru modul cum este definită / obținută forma duală (D) a problemei de optimizare convexă (P), vedeți problema 82.c.

enunță după cum urmează.

Condiția lui Slater pentru o problemă de optimizare convexă cu restricții (P): dacă există x astfel încât $g_i(x) < 0$ pentru orice i și $h_j(x) = 0$ pentru orice j , atunci $p^* = d^*$, adică are loc dualitatea tare.¹⁹⁹

În acest exercițiu vom demonstra prima parte a punctului (2) din Teorema Karush-Kuhn-Tucker: dacă x^* , α^* și β^* satisfac condițiile Karush-Kuhn-Tucker, atunci x^* este soluție a problemei (P).

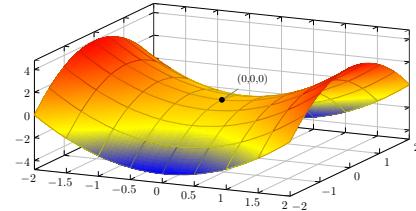
Pentru conveniență, în cele ce urmează, vom renota cu λ ansamblul variabilelor duale α și β .

A. Mai întâi vom demonstra că atunci când există $x^* \in \mathbb{R}^d$ și $\lambda^* \in \mathbb{R}_+^m \times \mathbb{R}^k$ care satisfac condițiile Karush-Kuhn-Tucker pentru problema (P), urmează că (x^*, λ^*) este un punct-șa (engl., saddle point) pentru lagrangeanul generalizat $L(x, \lambda)$.

Un punct (x^*, λ^*) este numit *punct-șa* al funcției $L(x, \lambda)$ dacă are loc dubla inegalitate

$$L(x^*, \lambda) \leq L(x^*, \lambda^*) \leq L(x, \lambda^*) \text{ pentru } \forall x \in \mathbb{R}^d \text{ și } \forall \lambda \in \mathbb{R}_+^m \times \mathbb{R}^k. \quad (108)$$

Exemplu: Fie funcția $g(x, y) = x^2 - y^2$. Se poate verifica imediat că originea sistemului de coordonate, adică punctul $(0, 0)$ este punct-șa pentru funcția g : $-y^2 \leq x^2 - y^2 \leq x^2$ pentru orice x și y din \mathbb{R} .



a. Introduceți multiplicatorii Lagrange $\lambda_1, \lambda_2, \dots, \lambda_{m+k}$ și scrieți funcția Lagrange generalizată care corespunde problemei (P). Scrieți apoi *condițiile Karush-Kuhn-Tucker* pentru problema de optimizare (P).

b. Folosiți condițiile de *fezabilitate primală* și *fezabilitate duală*, precum și condițiile de *complementaritate* (engl., complementary slackness) pentru a demonstra că — atunci când există $x^* \in \mathbb{R}^d$ și $\lambda^* \in \mathbb{R}_+^m \times \mathbb{R}^k$ care satisfac aceste condiții — prima jumătate a condiției din definiția noțiunii de punct-șa este satisfăcută.

c. Folosiți *condiția de fezabilitate duală* pentru a demonstra că $L(x, \lambda^*)$ — care este ultimul membru din dubla inegalitate (108) — este o funcție *convexă* în raport cu variabila x (considerând λ^* fixat) și, prin urmare, *condiția de staționaritate* (engl., the stationary condition) implică faptul că cea de-a doua jumătate a condiției din definiția noțiunii de punct-șa trebuie să fie satisfăcută. (*Sugestie:* Folosiți proprietățile funcțiilor convexe, pe care le-am demonstrat la problema 78.)

B. În cele ce urmează vom arăta că pentru orice $x^* \in \mathbb{R}^d$ pentru care există un $\lambda_i^* \in \mathbb{R}_+^m \times \mathbb{R}^k$ astfel încât (x^*, λ^*) este punct-șa al funcției Lagrange generalizate $L(x, \lambda)$, rezultă că x^* este soluție optimă pentru problema (P).

d. Arătați că atunci când prima jumătate a condiției (108) din definiția noțiunii de punct-șa este satisfăcută, rezultă cu necesitate că $h_j(x^*) = 0$ pentru $j =$

¹⁹⁹Cf. Morton Slater, *Lagrange multipliers revisited*, 1950.

$1, \dots, k$, apoi $g_i(x^*) \leq 0$ pentru $i = 1, \dots, m$ și $\sum_{i=1}^m \lambda_i^* g_i(x^*) = 0$, după care putem conchide că $f(x^*) = L(x^*, \lambda^*)$.

e. Completăți demonstrația [de la punctul precedent], arătând că $L(x^*, \lambda^*)$ — adică, membrul din mijloc din dubla inegalitate care apare în definiția noțiunii de punct-șa — este mărginit superior de p^* , valoarea optimă a funcției obiectiv pentru problema (P), și apoi că $f(x^*) = p^*$, deci x^* este soluție optimă a problemei (P).

Răspuns:

a. Folosind multiplicatorii Lagrange $\lambda_1, \lambda_2, \dots, \lambda_{m+k}$, scriem lagrangeanul generalizat pentru problema de optimizare (P) care a fost dată în enunț în modul următor:

$$L(x, \lambda) = f(x) + \sum_{i=1}^m \lambda_i g_i(x) + \sum_{j=1}^k \lambda_{m+j} h_j(x), \quad (109)$$

cu $\lambda_i \geq 0$ pentru $i \in [m]$, unde notația $[m]$ desemnează mulțimea $\{1, \dots, m\}$.

Condițiile Karush-Kuhn-Tucker pentru problema (P) se scriu astfel:

fezabilitate primală: $g_i(x^*) \leq 0, \forall i \in [m], h_j(x^*) = 0, \forall j \in [k]$;

fezabilitate duală: $\lambda_i^* \geq 0, \forall i \in [m]$;

complementaritate: $\lambda_i^* g_i(x^*) = 0, \forall i \in [m]$;

staționaritate / optimalitate: $\nabla f(x^*) + \sum_{i=1}^m \lambda_i^* \nabla g_i(x^*) + \sum_{j=1}^k \lambda_{m+j}^* \nabla h_j(x^*) = 0$,

unde operatorul / simbolul ∇ desemnează vectorul de derivate parțiale $\frac{\partial}{\partial x}$.

b. Pentru orice $\lambda \in \mathbb{R}^{m+k}$ astfel încât $\lambda_i \geq 0$ pentru $i \in [m]$, vom avea:

$$\begin{aligned} L(x^*, \lambda) &\stackrel{(109)}{=} f(x^*) + \sum_{i=1}^m \underbrace{\lambda_i}_{\geq 0} \underbrace{g_i(x^*)}_{\leq 0} + \sum_{j=1}^k \lambda_{m+j} \underbrace{h_j(x^*)}_{=0} \\ &\leq f(x^*) \\ &= f(x^*) + \sum_{i=1}^m \underbrace{\lambda_i^* g_i(x^*)}_{=0} + \sum_{j=1}^k \lambda_{m+j}^* \underbrace{h_j(x^*)}_{=0} \\ &\stackrel{(109)}{=} L(x^*, \lambda^*). \end{aligned}$$

c. Întrucât f și g_1, g_2, \dots, g_m sunt funcții convexe, $\lambda_i \geq 0$ pentru $i = 1, \dots, m$, iar funcțiile h_1, \dots, h_k sunt afine, deci convexe, ținând cont de proprietățile care au fost demonstreate la problema 78 — și anume, că produsul unei constante nenegative cu o funcție convexă este de asemenea funcție convexă, iar suma a două (sau mai multe) funcții convexe este tot o funcție convexă — rezultă că expresia

$$L(x, \lambda^*) \stackrel{(109)}{=} f(x) + \sum_{i=1}^m \lambda_i^* g_i(x) + \sum_{j=1}^k \lambda_{m+j}^* h_j(x)$$

reprezintă și ea o funcție convexă în raport cu variabila x .

Ca o consecință a faptului că $L(x, \lambda^*)$ este funcție convexă în raport cu variabila x , rezultă că valorile ei extreme sunt realizate în puncte pentru care vectorul gradient în raport cu x ia valoarea zero (mai exact, vectorul care are toate componente 0). Condiția de *stationaritate* Karush-Kuhn-Tucker afirmă că soluția (x^*, λ^*) satisfac această proprietate, așadar $L(x^*, \lambda^*) \leq L(x, \lambda^*)$ pentru orice x .²⁰⁰ Altfel spus, $L(x^*, \lambda^*)$ este o margine inferioară (engl., lower bound) pentru $L(x, \lambda^*)$.

$L(x, \lambda^*)$ este cel de-al treilea termen din dubla inegalitate (108), care constituie definiția punctului-șa. Înseamnă că, în condițiile date, și a doua jumătate a acestei dublei inegalități este satisfăcută.

Sumarizând această primă parte a exercițiului nostru, putem formula următoarea

Lemă (1): Dacă perechea (x^*, λ^*) satisfac condițiile Karush-Kuhn-Tucker pentru problema (P), atunci (x^*, λ^*) este punct-șa pentru lagrangeanul generalizat $L(x, \lambda)$ asociat lui (P).

d. Din prima jumătate a dublei inegalități (108), adică relația de definiție pentru noțiunea de punct-șa, rezultă imediat că

$$\sup_{\lambda_i \geq 0, i \in [m]} L(x^*, \lambda) \leq L(x^*, \lambda^*). \quad (110)$$

Dacă prin reducere la absurd ar exista un $j \in [k]$ astfel încât $h_j(x^*) \neq 0$, atunci în expresia

$$L(x^*, \lambda) \stackrel{(109)}{=} f(x^*) + \sum_{i=1}^m \lambda_i g_i(x^*) + \sum_{j=1}^k \lambda_{m+j} h_j(x^*)$$

am putea lua $\lambda_j = sign(h_j(x^*))(+\infty)$, iar marginea inferioară din relația (110) ar deveni $+\infty$, ceea ce ar contrazice relația de definiție pentru noțiunea de punct-șa (108). Așadar, $h_j(x^*) = 0$ pentru orice $j \in [k]$.

În ceea ce privește $g_i(x^*)$, dacă prin reducere la absurd ar exista vreun $i \in [m]$ astfel încât $g_i(x^*) > 0$, atunci marginea inferioară din relația (110) ar deveni $+\infty$ dacă se consideră $\lambda_i = +\infty$. Prin urmare, $g_i(x^*) \leq 0$ pentru orice $i \in [m]$.

Până aici am demonstrat că, în condițiile date, cele două condiții de fezabilitate primală din problema de optimizare convexă (P) sunt îndeplinite.

Acum, dacă $\lambda_i \geq 0$ pentru $i \in [m]$, rezultă $\sum_{i=1}^m \lambda_i g_i(x^*) \leq 0$. Mai mult, se observă că există o soluție trivială, și anume $\lambda_i = 0$ pentru $i \in [m]$, pentru ca această inegalitate să fie satisfăcută cu egalitate, adică $\sum_{i=1}^m \lambda_i g_i(x^*) = 0$. Așadar,

$$\sup_{\lambda_i \geq 0, i \in [m]} L(x^*, \lambda) = \sup_{\lambda_i \geq 0, i \in [m]} \left(f(x^*) + \underbrace{\sum_{i=1}^m \lambda_i g_i(x^*)}_{\leq 0} + \underbrace{\sum_{i=1}^k \lambda_{m+i} h_i(x^*)}_{=0} \right) = f(x^*).$$

Mai mult, relația (110) devine $\sup_{\lambda_i \geq 0, i \in [m]} L(x^*, \lambda) = L(x^*, \lambda^*)$ întrucât prin ipoteză avem $\lambda_i^* \geq 0$ pentru $i \in [m]$. Prin urmare, ajungem la concluzia care a fost formulată în enunț:

$$f(x^*) = \sup_{\lambda_i \geq 0, i \in [m]} L(x^*, \lambda) = L(x^*, \lambda^*).$$

²⁰⁰Vedeți proprietatea 2 de la problema 78.a.

e. La punctul precedent am demonstrat că x^* , prima componentă din punctul să (x^*, λ^*) , satisfac toate restricțiile din problema de optimizare dată și, de asemenea, că $L(x^*, \lambda^*) = f(x^*)$. Aici vom arăta că, în plus, $f(x^*) \leq \inf_{x \in X} f(x)$ (vedeți mai jos relația (111)), de unde va rezulta că x^* este soluție optimă pentru problema (P), deci $f(x^*) = p^* = L(x^*, \lambda^*)$.

Fie X spațiul / mulțimea tuturor soluțiilor fezabile pentru problema de optimizare dată. Are loc următoarea inegalitate multiplă:

$$L(x^*, \lambda^*) \leq \inf_x L(x, \lambda^*) \leq \inf_{x \in X} L(x, \lambda^*) \leq \inf_{x \in X} f(x). \quad (111)$$

Într-adevăr, prima dintre aceste inegalități este dată de cea de-a doua jumătate a relației (108). A doua inegalitate din (111) are loc pur și simplu fiindcă mulțimea X este o submulțime din \mathbb{R}^d . În fine, pentru a justifica ultima inegalitate din (111), vom demonstra următoarea proprietate (chiar dacă este ușor mai generală decât este nevoie aici):

[*Propoziție*]: Pentru orice soluție fezabilă \bar{x} pentru problema (P) și pentru orice $\lambda \in \mathbb{R}^{m+k}$ astfel încât $\lambda_i \geq 0$ pentru orice $i \in [m]$ rezultă că următoarea inegalitate este satisfăcută:

$$L(\bar{x}, \lambda) \leq f(\bar{x}).$$

Într-adevăr, dacă x o soluție fezabilă pentru problema (P), ea satisfac toate restricțiile, adică $g_i(\bar{x}) \leq 0$ pentru $i \in [m]$ și $h_{m+j}(\bar{x}) = 0$ pentru $j \in [k]$. Întrucât $\lambda_i \geq 0$ pentru orice $i \in [m]$, urmează că $\lambda_i g_i(\bar{x}) \leq 0$ pentru orice $i \in [m]$. Așadar, rezultă că

$$L(\bar{x}, \lambda) = f(\bar{x}) + \sum_{i=1}^m \underbrace{\lambda_i}_{\geq 0} \underbrace{g_i(\bar{x})}_{\leq 0} + \sum_{i=1}^k \lambda_{m+i} \underbrace{h_i(\bar{x})}_{=0} \leq f(\bar{x}). \quad (112)$$

Observație: Ultimul termen din relația (111) este exact p^* , valoarea optimă a funcției obiectiv pentru problema (P), problemă care se poate scrie echivalent sub forma $\inf_{x \in X} f(x)$. Așadar, $L(x^*, \lambda^*) \leq p^*$. Întrucât la punctul d am obținut că $L(x^*, \lambda^*) = f(x^*)$, rezultă că $L(x^*, \lambda^*) = p^*$.

Sumarizând cea de-a doua parte a exercițiului nostru, putem formula următoarea

Lemă (2): Dacă (x^*, λ^*) este punct-șa pentru lagrangeanul generalizat $L(x, \lambda)$ asociat problemei (P), atunci x^* este soluție optimă pentru (P).

Punând împreună cele două părți ale exercițiului nostru (altfel spus, combinând *Lema (1)* și *Lema (2)*), rezultă că putem formula următoarea

Propoziție: Dacă perechea (x^*, λ^*) satisfac condițiile Karush-Kuhn-Tucker pentru problema de optimizare convexă cu restricții (P), atunci x^* este soluție optimă pentru (P).²⁰¹

²⁰¹Teorema Karush-Kuhn-Tucker — veДЕti *Comentariul* de la pagina 186 — afirmă că în aceleași condiții (ca în această ultimă *Propoziție*) rezultă că λ^* este soluție optimă pentru duala problemei (P).

84. (Aplicarea metodei dualității Lagrange, folosind condițiile Karush-Kuhn-Tucker, la rezolvarea unei probleme de optimizare în care atât funcția obiectiv, cât și restricțiile sunt funcții pătratice; punerea în evidență a punctului-șa pentru lagrangeanul generalizat, și a legăturii sale cu soluțiile problemelor duală și primală)

*prelucrare de Liviu Ciortuz, 2019, după
• ○ University of Helsinki, 2014 spring, Jyrki Kivinen, HW5, pr. 1*

Considerăm funcțiile $f(x) = x^2$ și $g(x) = (x - 2)^2$. Fie problema de optimizare convexă

$$\begin{aligned} \min_x f(x) \\ \text{a. i. } g(x) \leq 1. \end{aligned} \tag{P}$$

a. Scrieți lagrangeanul generalizat $L_P(x, \alpha)$, unde $\alpha \geq 0$ este multiplicatorul Lagrange asociat restricției din problema (P). Folosind Maple (sau un alt soft matematic), faceți graficul acestei funcții. (Veți obține un grafic tridimensional.)

b. Rezolvați problema dată, folosind sistemul reprezentat de condițiile Karush-Kuhn-Tucker.

c. Dați expresia analitică a funcțiilor $x \mapsto \max_{\alpha \geq 0} L_P(x, \alpha)$ și $\alpha \mapsto \min_x L_P(x, \alpha)$. Indicați punctul (x^*, α^*) în care cele două funcții au aceeași valoare. Am folosit următoarele notații: $x^* = \arg \min_x \max_{\alpha \geq 0} L_P(x, \alpha)$ și $\alpha^* = \arg \max_{\alpha \geq 0} \min_x L_P(x, \alpha)$. (Vedeți problema 82.c.)

Sugestie: Ar fi util să reprezentați aceste două funcții pe un același sistem [de două axe] de coordonate, considerând atât variabila x cât și variabila α reprezentate pe axa orizontală.²⁰²

d. Scrieți forma duală a problemei (P), iar apoi rezolvați problema duală (D).

e. Înlocuiți restricția din problema de optimizare dată mai sus cu $g(x) \leq 8$ și apoi rezolvați noua problemă (notată cu (P')), urmând aceleași cerințe ca mai sus.²⁰³

Răspuns:

a. Expresia lagrangeanului generalizat al problemei de optimizare (P) se scrie astfel:

$$L_P(x, \alpha) = x^2 + \alpha((x - 2)^2 - 1).$$

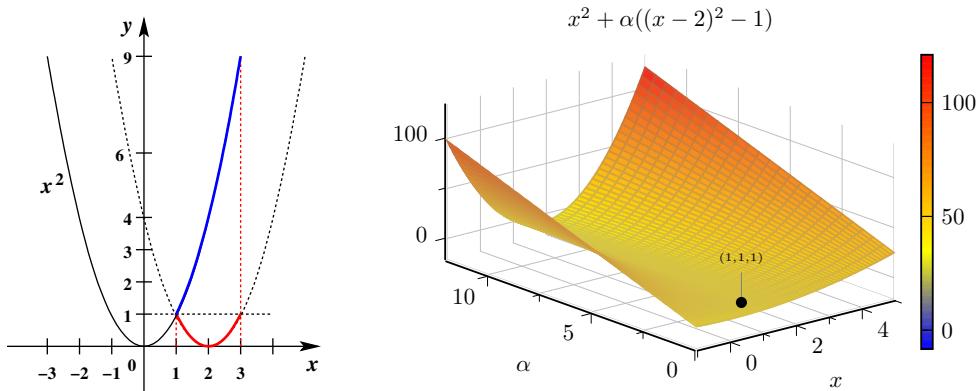
Am reprezentat funcția obiectiv, precum și funcția corespunzătoare restricției din cadrul acestei probleme de optimizare în graficul de mai jos, partea stângă, iar pentru lagrangeanul generalizat L_P am obținut graficul din partea dreaptă.²⁰⁴ Pe acest al doilea grafic am marcat (cu simbolul •) punctul-șa

²⁰² Atenție: Pentru prima funcție, $x \mapsto \max_{\alpha \geq 0} L_P(x, \alpha)$, nu veți reprezenta punctele în care ea ia valoarea $+\infty$.

²⁰³ Veți observa că spre deosebire de punctele $a - d$, unde restricția $g(x) \leq 1$ era „activă“ (fiindcă optimul global al funcției $f(x)$ nu este în interiorul mulțimii fezabile), aici restricția $g(x) \leq 8$ este superfluă.

²⁰⁴ Analizând graficul din partea stângă se poate constata că regiunea fezabilă pentru problema (P), adică mulțimea $\{x | g(x) \leq 1\}$ este intervalul $[1, 3]$. La efectuarea graficului lagrangeanului generalizat $L_P(x, \alpha)$ (și, de fapt, chiar mai înainte, la stabilirea domeniului de definiție pentru L_P), nu trebuie să se considere că x aparține doar intervalului $[1, 3]$. Corect este să considerăm $L_P : \mathbb{R} \times [0, +\infty) \rightarrow \mathbb{R}$. Vă readucem aminte că restricția $x \in [1, 3]$ a fost deja înglobată în expresia lagrangeanului generalizat $L_P(x, \alpha) \stackrel{\text{def.}}{=} f(x) + \alpha(g(x) - 1)$.

al acestui lagrangean, și anume punctul $(1, 1, 1)$, care corespunde lui $x^* = 1$, $\alpha^* = 1$, $L_P(x^*, \alpha^*) = 1$ (vedeți rezolvarea punctului c al problemei 83).



b. Condițiile Karush-Kuhn-Tucker corespunzătoare problemei (P) sunt următoarele:

$$\begin{aligned} (x-2)^2 - 1 &\leq 0 && \text{fezabilitate primală} \\ \alpha &\geq 0 && \text{fezabilitate duală} \\ \alpha((x-2)^2 - 1) &= 0 && \text{complementaritate} \\ \frac{\partial L_P}{\partial x} = 2x + \alpha(2x-4) &= 0 && \text{staționaritate (optimalitate)} \end{aligned}$$

Condiția de *staționaritate* se poate scrie echivalent astfel:

$$\alpha = \frac{x}{2-x}, \text{ dacă } x \neq 2 \quad (113)$$

sau, altfel:

$$x = \frac{2\alpha}{1+\alpha} = 2 - \frac{2}{1+\alpha} \text{ (remarcați că } \alpha \geq 0 \Rightarrow \alpha \neq -1\text{).} \quad (114)$$

Vom începe rezolvarea sistemului de condiții Karush-Kuhn-Tucker pornind de la condiția de *complementaritate*. Ca de obicei, vom avea de tratat două cazuri.

În primul caz, vom considera că $\alpha = 0$. (Observați că, în mod implicit, este satisfăcută condiția de *fezabilitate duală*.) Conform relației (114), rezultă $x = 0$, însă constatăm imediat că această valoare nu satisface condiția de *fezabilitate primală*.

În al doilea caz, vom considera că $\alpha \neq 0$ sau, mai precis (datorită condiției de *fezabilitate duală*) $\alpha > 0$. Din egalitatea $\alpha((x-2)^2 - 1) = 0$ rezultă că $(x-2)^2 - 1 = 0$. (Observați că, în mod implicit, este satisfăcută condiția de *fezabilitate primală*.) Ecuația $(x-2)^2 - 1 = 0$ are două soluții: $x_1 = 1$ și $x_2 = 3$. Conform relației (113), rezultă $\alpha_1 = \frac{1}{2-1} = 1 \geq 0$ și respectiv $\alpha_2 = \frac{3}{2-3} = -3 \leq 0$. Așadar, numai α_1 satisface condiția de *fezabilitate duală*.

Sumarizând, rezultă că singura soluție a sistemului de condiții Karush-Kuhn-Tucker este perechea $(x^* = 1, \alpha^* = 1)$. Conform teoremei Karush-Kuhn-Tucker

(mai exact, veДЕti proprietatea demonstrată la problema 83.B) rezultă că *soluția optimă* a problemei (P) este $x = 1$. Valoarea optimă a funcției obiectiv pentru această problemă este $p^* \stackrel{\text{not.}}{=} 1$.

c. Vom calcula mai întâi expresia analitică a funcției $x \mapsto \max_{\alpha \geq 0} L_P(x, \alpha)$.

Așadar, pentru un $x \in \mathbb{R}$ oarecare, dar fixat, vom avea:

$$\max_{\alpha \geq 0} L_P(x, \alpha) = \max_{\alpha \geq 0} (x^2 + \alpha((x-2)^2 - 1)) = x^2 + \max_{\alpha \geq 0} (\alpha((x-2)^2 - 1)). \quad (115)$$

Observați că

$$\begin{aligned} (x-2)^2 - 1 &= 0 && \text{pentru } x \in \{1, 3\} \\ (x-2)^2 - 1 &< 0 && \text{pentru } x \in (1, 3) \\ (x-2)^2 - 1 &> 0 && \text{pentru } x \in (-\infty, 1) \cup (3, +\infty), \end{aligned}$$

ceea ce implică

$$\max_{\alpha \geq 0} \alpha((x-2)^2 - 1) = \begin{cases} 0 & \text{pentru } x \in [1, 3] \\ +\infty & \text{în caz contrar.} \end{cases}$$

Tinând cont de acest rezultat intermediu, relația (115) devine:

$$\max_{\alpha \geq 0} L_P(x, \alpha) = \begin{cases} x^2 & \text{pentru } x \in [1, 3] \\ +\infty & \text{în rest.} \end{cases}$$

Este imediat că valoarea minimă a acestei funcții — vă reamintim că este vorba despre funcția $x \mapsto \max_{\alpha \geq 0} L_P(x, \alpha)$ — este $p^* = 1$, obținută pentru $x^* = 1$.

Observație importantă (1):

Mai general, se poate demonstra,²⁰⁵ elaborând un raționament similar cu cel de mai sus, că pentru orice problemă de optimizare convexă, pentru care funcția obiectiv este f , dacă notăm cu X mulțimea punctelor fezabile ale acestei probleme și cu α și β *multiplicatorii Lagrange* asociati restricțiilor (de tip inegalitate și respectiv egalitate) din problema de optimizare, rezultă că funcția $x \mapsto \max_{\alpha \geq 0, \beta} L_P(x, \alpha, \beta)$ este egală cu

$$\begin{cases} f(x) & \text{pentru orice } x \in X, \\ +\infty & \text{în rest.} \end{cases}$$

Mai mult, avem că

$$\min_{x \in X} f(x) = \min_x \max_{\alpha \geq 0, \beta} L_P(x, \alpha, \beta).$$

Observați că x apare nerestricționat în partea dreaptă a acestei ultime egalități, iar restricțiile asupra lui α sunt foarte simple. Așadar, noua problemă de optimizare (cea din partea dreaptă a ultimei egalități) se exprimă mai simplu decât problema inițială.

Acum vom calcula expresia analitică a funcției $\alpha \mapsto \min_x L_P(x, \alpha)$. Pentru un $\alpha \geq 0$ oarecare, dar fixat, vom avea:²⁰⁶

²⁰⁵Vedeți Chuong B. Do, *Convex Optimization Overview (cont'd)*, 2009, pag. 4-5.

²⁰⁶La prima egalitate de pe rândul al doilea (116) de la calculul expresiei lui $\min_x L_P(x, \alpha)$ am aplicat formula care ne dă optimul funcției de gradul al doilea $ax^2 + bx + c$, și anume $\frac{-\Delta}{4a}$, cu $\Delta = b^2 - 4ac$. De fapt, atunci când $b = 2b'$ (aşa cum se întâmplă în cazul de faţă), optimul se calculează mai simplu astfel: $-\frac{\Delta'}{a}$, unde $\Delta' = b'^2 - ac$. Vă readucem aminte că $\sqrt{\Delta}$ intervine în calculul rădăcinilor ecuației $ax^2 + bx + c = 0$:

$$\begin{aligned}
 \min_x L_P(x, \alpha) &= \min_x (x^2 + \alpha((x-2)^2 - 1)) = \min_x ((1+\alpha)x^2 - 4\alpha x + 3\alpha) \\
 &= -\frac{4\alpha^2 - 3\alpha(1+\alpha)}{1+\alpha} = \frac{\alpha(3-\alpha)}{1+\alpha} \\
 &= -\alpha + \frac{4\alpha}{1+\alpha} = -\alpha + 4 - \frac{4}{1+\alpha}.
 \end{aligned} \tag{116}$$

Vom nota această funcție cu $L_D(\alpha)$.²⁰⁷

Folosind mijloace clasice de analiză matematică din liceu, vom demonstra că valoarea maximă a acestei funcții este $d^* = 1$ și ea se obține pentru $\alpha^* = 1$.

$$\begin{aligned}
 L_D(\alpha) &= -\alpha + 4 - \frac{4}{1+\alpha} \\
 \Rightarrow L'_D(\alpha) &= -1 + \frac{4}{(1+\alpha)^2}
 \end{aligned} \tag{117}$$

$$\begin{aligned}
 &= \frac{4 - (1+\alpha)^2}{(1+\alpha)^2} = \frac{(2-(1+\alpha))(2+(1+\alpha))}{(1+\alpha)^2} = \frac{(1-\alpha)(3+\alpha)}{(1+\alpha)^2} \\
 \stackrel{(117)}{\Rightarrow} L''_D(\alpha) &= -\frac{8}{(1+\alpha)^3}.
 \end{aligned} \tag{118}$$

Rezultă că $L''_D(\alpha) < 0$ pentru orice $\alpha \geq 0$ (deci lagrangeanul dual este funcție concavă!), în vreme ce $L'_D(\alpha) = 0$ pentru $\alpha = 1$, $L'_D(\alpha) < 0$ pentru $\alpha = (1, +\infty)$, iar $L'_D(\alpha) > 0$ pentru $\alpha = [0, 1)$. Prin urmare, funcția $L_D(\alpha)$ este strict crescătoare pe intervalul $[0, 1]$ și strict descreșcătoare pe intervalul $[1, +\infty)$. Rezultă că, într-adevăr, valoarea maximă a funcției $L_D(\alpha)$ este 1, iar această valoare maximă se obține pentru $\alpha = 1$.

Întrucât $p^* = d^*$, conform primei părți a teoremei Karush-Kuhn-Tucker rezultă că sistemul de condiții Karush-Kuhn-Tucker este satisfăcut și putem conchide (vedeți rezultatul teoretic demonstrat la problema 83.A) că punctul $(x^* = 1, \alpha^* = 1)$ este punct-șa (ba mai mult, chiar unicul punct-șa) pentru lagrangeanul generalizat $L_P(x, \alpha)$.²⁰⁸

În imaginea alăturată am reprezentat într-un singur grafic cele două funcții calculate la acest punct.

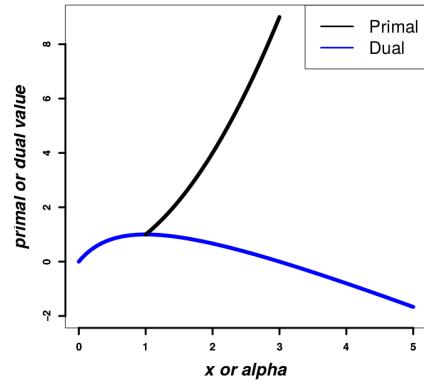
d. Funcția obiectiv pentru *duala unei probleme de optimizare convexă oarecare* se definește ca $L_D(\alpha) \stackrel{\text{not.}}{=} \min_x L_P(x, \alpha, \beta)$, unde $L_P(x, \alpha, \beta)$ este lagrangeanul

$x_{1,2} = -\frac{-b \pm \sqrt{\Delta}}{2a} = -\frac{-b' \pm \sqrt{\Delta'}}{a}$. Această ultimă variantă este întâlnită uneori în manuale sub numele de formula [cu Δ] pe jumătate, fiindcă $\sqrt{\Delta} = 2\sqrt{\Delta'}$.

Remarcați de asemenea faptul că expresia (114) reprezintă chiar abscisa punctului de optim pentru funcția $L_P(x, \alpha) = (1+\alpha)x^2 - 4\alpha x + 3\alpha$, atunci când parametrul α se consideră fixat. Valoarea pentru această abscisă se poate calcula folosind formula $x_{\min} = \frac{-b}{2a} = \frac{b'}{a}$. Mai târziu (vedeți punctul d), vom nota această abscisă cu $\arg \min_x L_P(x, \alpha)$.

²⁰⁷O astfel de funcție se numește îndeobște *lagrangeanul dual* corespunzător problemei (P). El ve fi funcția obiectiv a *formei duale* a problemei de optimizare (P). Vedeți punctul d.

²⁰⁸Faptul că punctul $(x^* = 1, \alpha^* = 1)$ satisface condiția de punct-șa (vedeți relația (108)) se poate verifica și



generalizat asociat problemei (P), cu α și β multiplicatorii Lagrange corespunzători restricțiilor de tip inegalitate și, respectiv, egalitate.

În cazul problemei noastre de optimizare (P), ținând cont de rezolvarea punctului c , rezultă că forma duală corespunzătoare lui (P) este următoarea:

$$\max_{\alpha \geq 0} L_D(\alpha) = \max_{\alpha \geq 0} \left(-\alpha + 4 - \frac{4}{1+\alpha} \right).$$

Calculele de la punctul c arată că soluția acestei probleme duale este $\alpha^* = 1$, iar valoarea optimă pentru funcția obiectiv a acestei probleme este $d^* = 1$.

Observație importantă (2):

Era de așteptat să obtinem $p^* = d^*$ (vedeți rezolvarea punctului c). Justificarea ține de faptul că pentru problema (P) este satisfăcută condiția lui Slater: $\exists x : g(x) < 0$, adică $\exists x : (x-2)^2 < 1$, deci $p^* = d^*$ (adică are loc *dualitatea tare*).²⁰⁹ Prima parte a teoremei Karush-Kuhn-Tucker afirmă că în aceste condiții soluțiile problemelor (P) și (D) (duala lui (P)) satisfac sistemul de condiții Karush-Kuhn-Tucker.

Pe noi însă la acest exercițiu nu ne-a interesat această chețiune / legătură, ci (în special la punctul b) cealaltă parte a teoremei Karush-Kuhn-Tucker: dacă sistemul de condiții Karush-Kuhn-Tucker este satisfăcut, atunci dispunem în mod natural de o soluție pentru problemele (P) și (D). Mai mult, la punctele c și d am ilustrat rezultatele teoretice care au fost demonstrate la problema 83.

Figura alăturată încearcă să ofere o imagine de sinteză (din punctul de vedere al reprezentărilor grafice) pentru raționamentele pe care le-am dezvoltat în cursul rezolvării problemei de optimizare convexă (P).

Cele două funcții care au fost calculate la punctul c sunt reprezentate aici astfel: prima (în negru) în planul de ecuație $\alpha = 0$, iar cea de-a doua (în bleu) în planul de ecuație $x = 0$.

Pe lângă acestea, am reprezentat funcția $\alpha \mapsto \arg \min_x L_P(x, \alpha)$ sub formă unei curbe punctate, în planul de ecuație $z = 0$, unde z notează cea de-a treia coordonată. Punctele de pe curba de culoare magenta (roz) au coordonatele $(\arg \min_x L_P(x, \alpha), \alpha, L_D(\alpha))$. Așadar, linia punctată reprezintă proiecția acestei curbe pe planul $z = 0$, în vreme ce curba în bleu este proiecția ei pe planul $x = 0$.

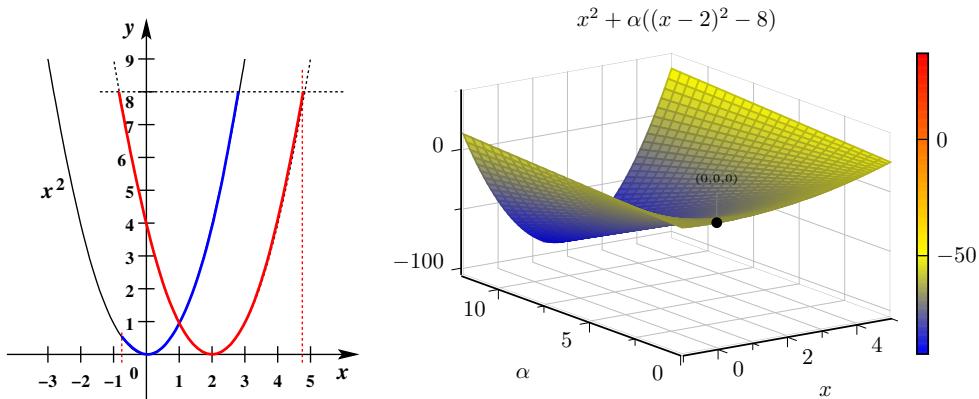
în mod direct:

$$\begin{aligned} L_P(x^*, \alpha) &\leq L_P(x^*, \alpha^*) \leq L_P(x, \alpha^*) \\ \stackrel{x^*=1, \alpha^*=1}{\Leftrightarrow} 1 + \alpha(1-1) &\leq 1 + 1(1-1) \leq x^2 + (x-2)^2 - 1 \\ &\Leftrightarrow 1 \leq 1 \leq 2(x-1)^2 + 1, \forall x, \forall \alpha \geq 0. \end{aligned}$$

²⁰⁹Vedeți *Observația importantă* de la pagina 186.

e. În cazul $g(x) \leq 8$, noul lagrangean generalizat este $L_{P'}(x, \alpha) = x^2 - \alpha((x - 2)^2 - 8)$.

Am reprezentat funcția obiectiv și funcția corespunzătoare restricției din cadrul acestei noi probleme de optimizare în graficul de mai jos, partea stângă, iar pentru lagrangeanul generalizat $L_{P'}$ am obținut graficul din partea dreaptă:



Condițiile de *fezabilitate primală* și respectiv *complementaritate* devin:

$$\begin{aligned} (x-2)^2 - 8 &\leq 0 \\ \alpha((x-2)^2 - 8) &= 0, \end{aligned}$$

iar celelalte condiții Karush-Kuhn-Tucker rămân neschimbate. Ca o consecință, se poate constata imediat că relațiile (113) și (114) sunt valabile și aici.

La rezolvarea noului sistem de restricții Karush-Kuhn-Tucker se constată, analizând condiția de *complementaritate*, că în cazul $\alpha > 0$ — care implică $(x-2)^2 - 8 = 0$ și mai departe $x_{1,2} = 2 \pm 2\sqrt{2}$ și $\alpha_1 = -(1 + \sqrt{2})/\sqrt{2} < 0$, $\alpha_2 = (1 - \sqrt{2})/\sqrt{2} < 0$ — condițiile de *fezabilitate duală* sunt încălcate, în vreme ce în cazul $\alpha = 0$ rezultă imediat că $x = 0$ și toate condițiile Karush-Kuhn-Tucker sunt satisfăcute.

Trecem acum la obținerea și analiza funcțiilor $x \mapsto \max_{\alpha \geq 0} L_{P'}(x, \alpha)$ și $\alpha \mapsto \min_x L_{P'}(x, \alpha)$.

Pentru $x \in \mathbb{R}$ oarecare, dar fixat, vom obține (fie în urma efectuării calculelor, fie ținând cont de *Observația importantă* (1) de la pag. 193):

$$\max_{\alpha \geq 0} L_{P'}(x, \alpha) = \begin{cases} x^2 & \text{pentru } x \in [2 - \sqrt{2}, 2 + \sqrt{2}] \\ +\infty & \text{în rest.} \end{cases}$$

Valoarea minimă a acestei funcții este 0, obținută pentru $x^* = 0$.

Pentru $\alpha \geq 0$ oarecare, dar fixat, vom avea:

$$\begin{aligned} \min_x L_{P'}(x, \alpha) &= \min_x (x^2 + \alpha((x-2)^2 - 8)) = \min_x ((1 + \alpha)x^2 - 4\alpha x - 4\alpha) \\ &= -\frac{(2\alpha)^2 + 4\alpha(1 + \alpha)}{1 + \alpha} = -\frac{4\alpha(2\alpha + 1)}{1 + \alpha} = -\left(8\alpha - 4 + \frac{4}{1 + \alpha}\right). \end{aligned}$$

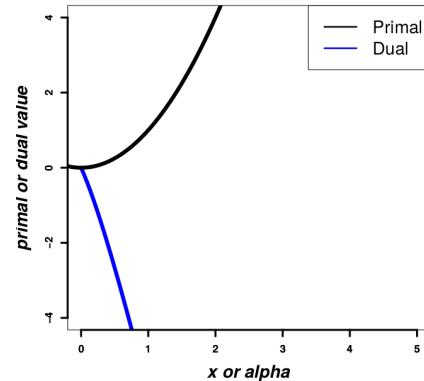
Așadar, lagrangeanul dual este $L_{D'}(\alpha) = -\frac{4\alpha(2\alpha + 1)}{1 + \alpha}$, iar problema duală se scrie ca $\max_{\alpha \geq 0} L_{D'}(\alpha)$. Folosind din nou cunoștințele de analiză matematică de liceu, se constată că

$$L'_{D'}(\alpha) = -8 + \frac{4}{(1+\alpha)^2} = \frac{4[1-\sqrt{2}(1+\alpha)][1+\sqrt{2}(1+\alpha)]}{(1+\alpha)^2} < 0, \forall \alpha \geq 0$$

și

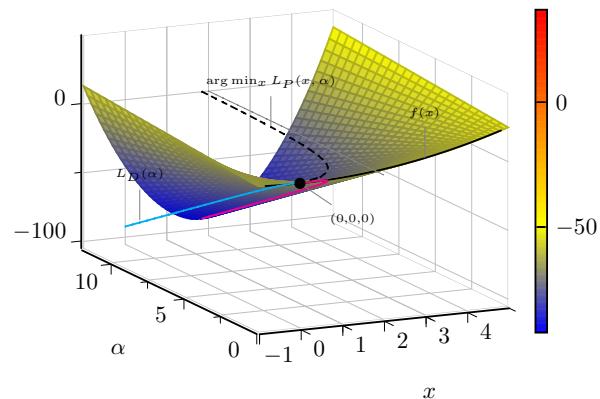
$$L''_{D'}(\alpha) = -\frac{8}{(1+\alpha)^3} < 0, \forall \alpha \geq 0,$$

deci funcția $L_{D'}(\alpha)$ este concavă, iar valoarea maximă a sa este $d^* = 0$, fiind obținută pentru $\alpha^* = 0$.²¹⁰



Prin urmare, punctul $(x^* = 0, \alpha^* = 0)$ este punct-șa (din nou, unicul punct-șa!) pentru lagrangeanul generalizat $L_{P'}(x, \alpha)$.²¹¹

În imaginea alăturată am reprezentat într-un singur grafic cele două funcții, $x \mapsto \max_{\alpha \geq 0} L_{P'}(x, \alpha)$ și $\alpha \mapsto \min_x L_{P'}(x, \alpha)$.



Ca și în cazul problemei (P), prezentăm alăturat o reprezentare grafică de *sinteză* pentru componentele-cheie folosite la rezolvarea problemei de optimizare convexă (P'). Semnificațiile culorilor diferențelor curbe din acest grafic sunt aceleași cu cele din figura corespunzătoare de la punctul d.

²¹⁰Se poate constata [din grafice] că în cazul restricției $g(x) \leq 8$ (deci pentru cea de-a doua problemă de optimizare dată), regiunea fezabilă include punctul $x = 0$ pentru care se atinge minimul funcției $f(x) = x^2$, deci în acest caz restricția nu [mai] afectează de fapt optimizarea funcției f . În cazul restricției $g(x) \leq 1$ (deci pentru prima problemă de optimizare dată), punctul $x = 0$ era situat în afara regiunii fezabile.

²¹¹Și aici, verificăm în mod direct faptul că punctul $(x^* = 1, \alpha^* = 1)$ satisfacă condiția de punct-șa:

$$\begin{aligned} L_{P'}(x^*, \alpha) &\leq L_{P'}(x^*, \alpha^*) \leq L_{P'}(x, \alpha^*) \\ \stackrel{x^*=0, \alpha^*=0}{\Leftrightarrow} \quad \alpha(4-8) &\leq 0 \leq x^2 \Leftrightarrow -4\alpha \leq 0 \leq x^2, \forall x, \forall \alpha \geq 0. \end{aligned}$$

85. (Metoda dualității Lagrange: aplicare în \mathbb{R}^2 , în două moduri:
mai întâi folosind condițiile Karush-Kuhn-Tucker,
iar apoi folosind corespondența cu soluțiile problemei duale)

*prelucrare de L. Ciortuz și S. Ciobanu, după
 University of Helsinki, 2014 spring, Jyrki Kivinen,
 Example of convex optimisation with Lagrange coefficients*

Fie următoarea problemă de optimizare cu restricții:

$$\min_{(x,y) \in \mathbb{R}^2} x^2 + y^2$$

a. Într-un sistem de coordonate bidimensional, indicați (prin hasurare) care este mulțimea punctelor din plan care satisfac restricția inclusă în cadrul problemei de optimizare din enunț.²¹²

- b. Indicați soluția acestei probleme de optimizare (însă fără a face calcule).
- c. Stabiliți dacă problema de optimizare dată este convexă.
- d. Este oare condiția lui Slater satisfăcută?²¹³ Are loc proprietatea de *dualitate tare*? Justificați.
- e. Scrieți problema duală (D) asociată problemei inițiale (P). Soluțiile acestor două probleme satisfac condițiile Karush-Kuhn-Tucker? (Atenție! Nu este nevoie să rezolvați cele două probleme ca să puteți răspunde la această întrebare.)
- f. Rezolvați problema (P) folosindu-vă doar de sistemul reprezentat de condițiile Karush-Kuhn-Tucker.²¹⁴
- g. Acum, rezolvați problema (P) în alt mod: aflați mai întâi soluția problemei duale (D) și apoi găsiți soluțiile problemei primale (P) folosindu-vă doar de *condiția de staționaritate* (sau, de *optimalitate*) din sistemul dat de condițiile Karush-Kuhn-Tucker, întrucât această condiție poate fi văzută ca o *relație de legătură* între soluțiile celor două probleme, (P) și (D).
- h. Rezolvați din nou punctele $a - g$, de data aceasta pentru problema de optimizare care se obține din problema inițială înlocuind restricția $x + y \leq 1$ cu $x + y \leq -1$. (Se poate observa că restricția inițială era de fapt irelevantă, pe când noua restricție este „activă“.)

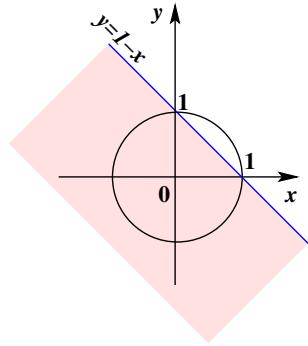
Răspuns:

²¹²Observație: Vă readucem aminte că această mulțime se mai numește mulțimea soluțiilor *fezabile* (engl., feasible) sau, mai simplu, *regiunea fezabilă* pentru problema de optimizare dată.

²¹³Vedeți *Observația importantă* de la pagina 186.

²¹⁴Teorema Karush-Kuhn-Tucker este formulată la problema 83, mai precis în *Comentariul* de la pagina 186.

- a. Figura alăturată prezintă regiunea fezabilă pentru problema de optimizare dată — problemă pe care de acum încolo o vom desemna cu (P) —, adică mulțimea punctelor din plan care satisfac restricția $x + y \leq 1$.



- b. Dacă nu ținem cont de restricția inclusă în problema (P), rezultă imediat că

$$\min_{x,y \in \mathbb{R}^2} (x^2 + y^2) = 0,$$

iar această valoare optimă se obține pentru $x^* = y^* = 0$. Întrucât această soluție satisfac restricția $x + y \leq 1$ (altfel spus, punctul (x^*, y^*) aparține regiunii fezabile pe care am determinat-o la punctul precedent), rezultă că $x^* = 0, y^* = 0$ este soluția optimă a problemei (P).

- c. Da, problema (P) este o problemă de optimizare convexă. Justificarea constă în faptul că pe de o parte funcția obiectiv $x^2 + y^2$ este convexă, iar pe de altă parte restricția $x + y \leq 1$ se poate scrie în mod echivalent $x + y - 1 \leq 0$, funcția $x + y - 1$ fiind liniară în ambele argumente, deci convexă.²¹⁵

- d. Da, condiția lui Slater este satisfăcută pentru problema de optimizare convexă (P), întrucât $\exists x, y$ astfel încât $x + y < 1$, de exemplu $x = 0, y = 0$. Condiția lui Slater implică proprietatea de *dualitate tare*, adică $p^* = d^*$, unde p^* este soluția optimă a problemei (P), iar d^* este soluția optimă a dualei sale, (D).

- e. Lagrangeanul generalizat al problemei de optimizare (P) este

$$L_P(x, y, \alpha) = x^2 + y^2 + \alpha(x + y - 1),$$

unde $\alpha \in \mathbb{R}_+$ este *variabila duală* (sau, multiplicatorul Lagrange) care corespunde restricției $x + y \leq 1$ din problema (P).²¹⁶ Calculând derivata parțială a lui L_P în raport cu variabila x și egalând-o apoi cu 0, vom avea:

$$\frac{\partial}{\partial x} L_P(x, y, \alpha) = 0 \Leftrightarrow 2x + \alpha = 0 \Leftrightarrow x = -\frac{\alpha}{2}.$$

Similar, calculând derivata parțială a lui L_P în raport cu variabila y și egalând-o apoi cu 0, va rezulta $y = -\frac{\alpha}{2}$.

²¹⁵ Funcția $x^2 + y^2$ este convexă întrucât ea are matricea hessiană

$$H \stackrel{\text{not.}}{=} \nabla^2(x^2 + y^2) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix},$$

iar această matrice este pozitiv semidefinită: este imediat că $v^\top H v \geq 0$ pentru orice $v \in \mathbb{R}^2$. (Vedeți *Observația* de la finalul rezolvării punctului a.3 de la problema 78.)

²¹⁶ Înțînd cont de condiția de staționaritate / optimalitate din sistemul Karush-Kuhn-Tucker pe de o parte, și de definiția formei duale a problemei de optimizare convexă cu restricții (vedeți problema 82.b) pe de altă parte, vom putea deduce lagrangeanul dual pe o altă cale (ceva mai directă!) decât am făcut-o la problema 84.c.

Înlocuind x și y cu $-\frac{\alpha}{2}$ în expresia lui $L_P(x, y, \alpha)$, vom obține *lagrangeanul dual*, care va fi funcția obiectiv a problemei de optimizare (D), forma duală a problemei (P):²¹⁷

$$L_D(\alpha) = 2 \cdot \frac{\alpha^2}{4} + \alpha(-\alpha - 1) = -\frac{\alpha^2}{2} - \alpha.$$

Așadar, duala problemei (P) este:

$$\max_{\alpha \geq 0} \left(-\frac{\alpha^2}{2} - \alpha \right).$$

Soluțiile problemelor (P) și (D) satisfac condițiile Karush-Kuhn-Tucker pentru că, aşa cum am arătat la punctul precedent, are loc dualitatea tare.²¹⁸

f. Sistemul reprezentat de condițiile Karush-Kuhn-Tucker pentru problema (P) este următorul:

$$\begin{cases} x + y \leq 1 & \text{fezabilitate primală} \\ \alpha \geq 0 & \text{fezabilitate duală} \\ \alpha(x + y - 1) = 0 & \text{complementaritate} \\ x = -\frac{\alpha}{2} \text{ și } y = -\frac{\alpha}{2} & \text{staționaritate (optimalitate)} \end{cases}$$

Remarcați faptul că forma aceasta a condițiilor de staționaritate / optimalitate a fost dedusă la punctul e , acolo unde am obținut soluțiile derivatelor parțiale ale lagrangeanului generalizat L_P în raport cu x și respectiv y .

Știm că, atunci când există, soluțiile x^*, y^* și α^* ale sistemului de condiții Karush-Kuhn-Tucker sunt totodată soluții optime ale problemelor (P) și respectiv (D).²¹⁹

Pornind de la *condiția de complementaritate*, putem rezolva sistemul de mai sus în felul următor:

Cazul I: $\alpha = 0$.

Conform *condiției de staționaritate*, rezultă $x = y = 0$. Condițiile de fezabilitate ($x + y \leq 1$ și $\alpha \geq 0$) sunt satisfăcute în acest caz, deci $\alpha^* = x^* = y^* = 0$ este soluție a sistemului constituit de condițiile Karush-Kuhn-Tucker. Urmează că $x^* = y^* = 0$ este soluție optimă a problemei (P), iar $\alpha^* = 0$ este soluție optimă a problemei (D).

Cazul II: $\alpha \neq 0$, deci $\alpha > 0$.

Din condiția de complementaritate, rezultă $x+y-1=0$, deci $x+y=1$. Urmează că este satisfăcută și condiția de fezabilitate primală, $x+y \leq 1$. Din relația $x+y-1=0$, ținând cont de condiția de staționaritate ($x=y=-\frac{\alpha}{2}$), rezultă că $-\alpha=1$, adică $\alpha=-1$, ceea ce contrazice condiția de fezabilitate duală ($\alpha \geq 0$). Prin urmare, în acest caz ($\alpha > 0$), sistemul Karush-Kuhn-Tucker nu are soluție.

²¹⁷Observați că spre deosebire de modul în care am procedat la problema precedentă, unde L_D a fost calculat apelând la definiția $\alpha \mapsto \min_{x,y} L_P(x, y, \alpha)$, aici obținem L_D din L_P pur și simplu particularizând L_P pentru acele valori ale lui x și respectiv y (exprimate în funcție de α) care au fost deduse din condiția de staționaritate / optimalitate.

²¹⁸Vedeți *Comentariul* de la pag. 186, teorema Karush-Kuhn-Tucker, prima parte.

²¹⁹Vedeți *Comentariul* de la pag. 186, teorema Karush-Kuhn-Tucker, a doua parte.

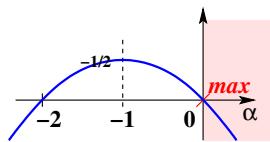
În concluzie, singura soluție a sistemului Karush-Kuhn-Tucker este $\alpha^* = x^* = y^* = 0$, iar singura soluție optimă a problemei (P) este $x^* = y^* = 0$.

g. Soluția optimă a problemei (D), a cărei formă a fost dedusă la punctul d,

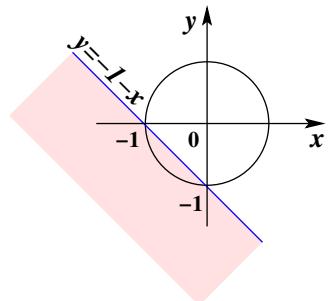
$$\max_{\alpha \geq 0} \left(-\frac{\alpha^2}{2} - \alpha \right) \Leftrightarrow \max_{\alpha \geq 0} \left(-\frac{\alpha}{2}(\alpha + 2) \right),$$

este $\alpha^* = 0$. Justificarea este imediată, dacă ținem cont de graficul funcției $-\frac{\alpha}{2}(\alpha + 2)$, din figura alăturată.

Din relațiile $x = y = -\frac{\alpha}{2}$, care au fost deduse la punctul e, rezultă că soluția optimă a problemei (P) este $x^* = y^* = 0$.



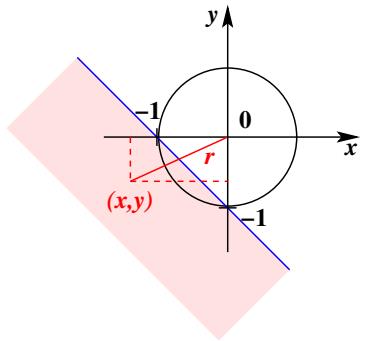
h. Figura alăturată prezintă regiunea fezabilă pentru noua problemă de optimizare convexă (pe care o vom nota cu (P')), adică mulțimea punctelor din plan care satisfac restricția $x + y \leq -1$.



Se poate observa că problema (P') este echivalentă cu o altă problemă de optimizare (care nu este însă convexă, fiindcă restricția de tip egalitate nu este afină):

$$\begin{aligned} \min_{r \in \mathbb{R}} \quad & r^2 \\ \text{a. i.} \quad & r^2 = x^2 + y^2 \\ & x + y \leq -1 \end{aligned}$$

și că valoarea minimă pentru r este corespunzătoare punctului $(x^* = -1/2, y^* = -1/2)$, deci $r^* = \frac{1}{\sqrt{2}}$.



Evident, noua restricție $(x + y \leq -1)$ nu schimbă tipul problemei (P') , ea rămânând una de optimizare convexă. De asemenea, condiția lui Slater este satisfăcută și aici: $\exists x, y$ astfel încât $x + y < -1$ (considerați, de exemplu, $x = y = -1$). În consecință, și în acest caz are loc *dualitatea tare*. Așadar, putem rezolva problema (P') rezolvând mai întâi duala sa (D') și folosind apoi relația de corespondență dintre soluțiile celor două probleme.

Deocamdată însă, vom rezolva problema (P') pornind de la sistemul de condiții Karush-Kuhn-Tucker:

$$\begin{cases} x + y \leq -1 & \text{fezabilitate primală} \\ \alpha \geq 0 & \text{fezabilitate duală} \\ \alpha(x + y + 1) = 0 & \text{complementaritate} \\ x = -\frac{\alpha}{2} \text{ și } y = -\frac{\alpha}{2} & \text{staționaritate (optimalitate)} \end{cases}$$

Remarcați faptul că forma aceasta a condițiilor de staționaritate / optimilitate a fost obținută calculând în prealabil soluțiile derivatelor parțiale ale lagrangeanului generalizat $L_{P'}$ în raport cu x și respectiv y .

Rezolvarea sistemului de condiții Karush-Kuhn-Tucker este următoarea:

Cazul I: $\alpha = 0$.

Conform condiției de staționaritate, rezultă $x = y = 0$, dar această soluție nu este fezabilă ($x + y \not\leq -1$).

Cazul II: $\alpha > 0$.

Din condiția de complementaritate rezultă $x + y = -1$, iar condiția de staționaritate va implica $x = y = -1/2$ și $\alpha = 1$. Aceste valori pentru x, y și α satisfac condițiile de fezabilitate.

Prin urmare, unica soluție a sistemului Karush-Kuhn-Tucker este $x^* = y^* = -1/2$ și $\alpha^* = 1$, de unde rezultă că unica soluție a problemei (P') este $x^* = y^* = -1/2$.

Alternativ, aşa cum de fapt am precizat mai sus, am fi putut găsi soluția problemei (P') rezolvând duala sa (D') și folosind apoi relația de corespondență dintre soluții. Este exact ceea ce vom face acum:

$$\begin{aligned} L_{P'}(x, y, \alpha) &= x^2 + y^2 + \alpha(x + y + 1) \\ \Rightarrow \frac{\partial}{\partial x} L_{P'}(x, y, \alpha) &= 0 \Leftrightarrow 2x + \alpha = 0 \Leftrightarrow x = -\frac{\alpha}{2} \\ \text{și, similar, } \frac{\partial}{\partial y} L_{P'}(x, y, \alpha) &= 0 \Leftrightarrow y = -\frac{\alpha}{2} \\ \Rightarrow L_{D'}(\alpha) &= \frac{\alpha^2}{2} + \alpha(-\alpha + 1) = -\frac{\alpha^2}{2} + \alpha = -\frac{\alpha}{2}(\alpha - 2). \end{aligned}$$

Prin urmare, problema duală (D') este

$$\max_{\alpha \geq 0} \left(-\frac{\alpha}{2}(\alpha - 2) \right).$$

Vă puteți convinge singuri (de exemplu, făcând graficul funcției $-\frac{\alpha}{2}(\alpha - 2)$, sau folosind formula generală de calcul pentru abscisa punctului de optim al funcției de gradul al doilea) că soluția problemei (D') este $\alpha^* = 1 \geq 0$. În concluzie, soluția problemei (P') este $x^* = y^* = -1/2$, exact ceea ce am găsit și anterior, însă pe altă cale (și anume, folosind doar condițiile Karush-Kuhn-Tucker).

86. (Metoda dualității Lagrange: aplicare pentru găsirea codificării optimale pentru [literele din] alfabetul unui limbaj)

• CMU, 2014 fall, E. Xing, B. Poczos, HW1, pr. 3.2

În acest exercițiu vom rezolva o problemă [relativ simplă] de optimizare convexă.

Alfabetul unui limbaj constă din litere, pe care le vom nota cu α_i , pentru $i \in [n]$, unde $[n] \stackrel{\text{not.}}{=} \{1, 2, \dots, n\}$. Presupunem că aceste litere apar în decursul folosirii [limbajului respectiv, în mod natural] cu probabilitățile p_i . Ne propunem să

găsim pentru aceste litere o *codificare optimală* sub formă de biți, în aşa fel încât să minimizăm *numărul mediu de biți* (engl., the expected number of bits) folosîți de litere, păstrând însă capacitatea de a decodifica secvența de biți transmiși pe un canal (engl., stream) de comunicare, fără a avea nevoie să folosim delimitatori (engl., markers) între litere.

De exemplu, dacă alfabetul considerat are $n = 32$ de litere și toate literele apar cu probabilități egale (deci $p_i = 1/32$), este normal să folosim pentru codificarea optimă 5 biți pentru fiecare literă.

Se poate arăta că o codificare optimă pentru alfabetul unui limbaj poate fi găsită rezolvând următoarea *problemă de optimizare cu restricții*:²²⁰

$$\begin{aligned} & \min_{x \in \mathbb{R}^n} \sum_{i=1}^n p_i x_i \\ \text{a. i. } & \sum_{i=1}^n 2^{-x_i} \leq 1 \\ & x_i \geq 0 \quad \forall i \in [n]. \end{aligned} \tag{119}$$

În această problemă, *funcția obiectiv* este numărul mediu de biți per literă, iar prima restricție este aşa-numita *inegalitate a lui Kraft*, care asigură faptul că literele vor putea fi *decodificate* în mod unic.²²¹

Observație: Deși numerele x_i ar trebui să fie întregi și pozitive, noi vom căuta soluția problemei de optimizare în multimea numerelor reale pozitive.

a. Demonstrați că *regiunea fezabilă* (engl., feasible region) corespunzătoare problemei de optimizare (119) — adică, multimea formată din toți acei $x \stackrel{\text{not.}}{=} (x_1, \dots, x_n) \in \mathbb{R}^n$ pentru care sunt satisfăcute cele două restricții din problema de optimizare dată — este convexă.

b. Arătați că prima restricție din problema (119) este satisfăcută cu egalitate în cazul soluției optime [a acestei probleme].

c. Datorită rezultatului de la punctul precedent, putem rezolva problema (119) înlocuind în prima restricție semnul \leq cu $=$. Rezolvați această nouă problemă de optimizare convexă.

Sugestie: Demonstrați în prealabil că funcția din membrul stâng din prima restricție din cadrul problemei de optimizare (119) este într-adevăr convexă, pentru a justifica aplicarea [ulterioră a] metodei dualității Lagrange.

Răspuns:

a. Notăm cu P regiunea fezabilă a problemei (119). Pentru a demonstra convexitatea mulțimii P , vom folosi *definiția* din enunțul problemei 78. Așadar, vom arăta că $\alpha x + (1 - \alpha)y \in P$ pentru orice $x, y \in P$ și orice $\alpha \in (0, 1)$. Notând $x \stackrel{\text{not.}}{=} (x_1, \dots, x_n)$ și $y \stackrel{\text{not.}}{=} (y_1, \dots, y_n)$, rezultă că $\alpha x + (1 - \alpha)y = (\alpha x_1 + (1 - \alpha)y_1, \dots, \alpha x_n + (1 - \alpha)y_n)$.

²²⁰ Atenție: Observați că nu am afirmat că problema (119) este o problemă de optimizare convexă! Rezolvarea punctelor a și b de mai jos nu depinde de această chestiune. Doar la punctul c vă vom cere să arătați (între altele) că într-adevăr aceasta este o problemă de optimizare convexă (cu restricții).

²²¹ Vedeți https://en.wikipedia.org/wiki/Kraft-McMillan_inequality.

$\alpha)y_1, \dots, \alpha x_n + (1 - \alpha)y_n$). Folosind *inegalitatea ponderată a mediilor*²²² precum și faptul că x și y aparțin regiunii fezabile (P), putem scrie:²²³

$$\begin{aligned} \sum_{i=1}^n 2^{-(\alpha x_i + (1 - \alpha)y_i)} &\leq \sum_{i=1}^n [\alpha 2^{-x_i} + (1 - \alpha) 2^{-y_i}] \\ &= \underbrace{\alpha \sum_{i=1}^n 2^{-x_i}}_{\leq 1} + \underbrace{(1 - \alpha) \sum_{i=1}^n 2^{-y_i}}_{\leq 1} \\ &\leq \alpha + (1 - \alpha) = 1. \end{aligned} \quad (120)$$

Așadar, $\alpha x + (1 - \alpha)y$ satisfac prima restricție (adică inegalitatea) din problema de optimizare (119). Apoi, este clar că $x_i, y_i \geq 0 \Rightarrow \alpha x_i + (1 - \alpha)y_i \geq 0$ pentru orice $\alpha \in [0, 1]$. Prin urmare, $\alpha x + (1 - \alpha)y \in P$. Cu aceasta, demonstrația pentru convexitatea mulțimii P este încheiată.

b. Presupunem (prin *reducere la absurd*) că o soluție optimă $x^* \stackrel{\text{not.}}{=} (x_1^*, \dots, x_n^*)$ a problemei (119) satisfac restricția de tip inegalitate în sens strict, adică $\sum_{i=1}^n 2^{-x_i^*} < 1$. Este imediat că pentru orice $j \in [n]$, avem $x_j^* > 0$, fiindcă altfel am avea $\sum_{i=1}^n 2^{-x_i^*} > 1$.²²⁴ Prin urmare, putem să micșorăm valoarea [măcar] pentru unul dintre acești x_j^* , fără a afecta condițiile de fezabilitate ale problemei, dar reușind în schimb să micșorăm valoarea funcției obiectiv. Așadar, x^* nu este soluție optimă a problemei (119), ceea ce intră în *contradicție* cu presupunerea pe care am făcut-o anterior.

În concluzie, orice soluție optimă a problemei (119) satisfac prima restricție cu egalitate. Ca o consecință directă a acestui fapt, a rezolva problema de optimizare (119) revine la a rezolva următoarea problemă de optimizare:²²⁵

²²²În forma cea mai simplă (cazul trivial), inegalitatea care leagă media aritmetică de media geometrică (engl., *inequality of arithmetic and geometric means*, abbr. *AM-GM inequality*), se formulează astfel: pentru orice două numere nenegative x și y , se poate arăta că

$$\frac{x+y}{2} \geq \sqrt{xy},$$

egalitatea având loc dacă și numai dacă $x = y$.

Varianta *ponderată* a acestei inegalități (engl., *weighted AM-GM inequality*) este următoarea: Fie numerele nenegative x_1, x_2, \dots, x_m , precum și ponderile nenegative w_1, w_2, \dots, w_m . Notăm $w = w_1 + w_2 + \dots + w_m$. Dacă $w > 0$, atunci este satisfăcută inegalitatea

$$\begin{aligned} \frac{w_1 x_1 + w_2 x_2 + \dots + w_m x_m}{w} &\geq \sqrt[w]{x_1^{w_1} x_2^{w_2} \dots x_m^{w_m}} \\ \Leftrightarrow \frac{w_1}{w} x_1 + \frac{w_2}{w} x_2 + \dots + \frac{w_m}{w} x_m &\geq x_1^{\frac{w_1}{w}} \dots x_m^{\frac{w_m}{w}}, \end{aligned}$$

egalitatea având loc dacă și numai dacă toți acei x_k pentru care $w_k > 0$ sunt egali. În acest context se folosește convenția $0^0 = 1$. Observați că atunci când toate ponderile w_k au valoarea 1, obținem inegalitatea mediilor care a fost prezentată mai sus.

(Mențiune: Conținutul acestei note de subsol a fost preluat de pe site-ul

https://en.wikipedia.org/wiki/Inequality_of_arithmetic_and_geometric_means#Weighted_AM-GM_inequality.)

²²³Pentru fiecare valoare a lui $i \in [n]$, fixată, dacă în inegalitatea ponderată a mediilor luăm $m = 2$, $x_1 = 2^{-x_i}$, $x_2 = 2^{-y_i}$, $w_1 = \alpha$ și, $w_2 = 1 - \alpha$, obținem

$$2^{-\alpha x_i} \cdot 2^{-(1-\alpha)y_i} = (2^{-x_i})^\alpha \cdot (2^{-y_i})^{(1-\alpha)} \leq \alpha 2^{-x_i} + (1 - \alpha) 2^{-y_i}.$$

Însușind aceste inegalități membru cu membru pentru $i = 1, \dots, n$, vom obține inegalitatea (120).

²²⁴LC: S-a considerat $n \geq 2$, fiindcă pentru $n = 1$ problema este trivială. (De fapt, chiar și pentru $n = 2$ problema este trivială: cele două litere ale alfabetului vor fi codificate prin biții 0 și 1.)

²²⁵Observație (LC): Deși restricția $\sum_{i=1}^n 2^{-x_i} = 1$ nu este de tip afin — vedeti definiția nețiunii de problemă

$$\begin{aligned} & \min_{x \in \mathbb{R}^n} \sum_{i=1}^n p_i x_i \\ \text{a. i. } & \sum_{i=1}^n 2^{-x_i} = 1 \\ & x_i \geq 0 \quad \forall i \in [n]. \end{aligned} \tag{121}$$

c. Vom arăta mai întâi că funcția $\left(\sum_{i=1}^n 2^{-x_i}\right) - 1$, care corespunde primei restricții din problema de optimizare (119), este convexă. Pentru aceasta, vom demonstra că matricea hessiană a acestei funcții este pozitiv semidefinită.

Vectorul gradient pentru această funcție este

$$\nabla \left(\sum_{i=1}^n 2^{-x_i} \right) = (\ln 2) [2^{-x_1}, \dots, 2^{-x_n}]^\top.$$

Prin urmare, matricea hessiană pentru aceeași funcție $\left(\sum_{i=1}^n 2^{-x_i}\right) - 1$ este

$$\nabla^2 \left(\sum_{i=1}^n 2^{-x_i} \right) = (\ln 2)^2 \begin{bmatrix} 2^{-x_1} & & 0 \\ & \ddots & \\ 0 & & 2^{-x_n} \end{bmatrix} \stackrel{\text{not.}}{=} (\ln 2)^2 \operatorname{diag} \left(2^{-x_i} \right)_{i=1,n}.$$

Pentru orice vector $v \in \mathbb{R}^n$, avem:

$$\begin{aligned} v^\top \nabla^2 \left(\sum_{i=1}^n 2^{-x_i} \right) v &= [v_1, \dots, v_n]^\top (\ln 2)^2 \operatorname{diag} \left(2^{-x_i} \right)_{i=1,n} [v_1, \dots, v_n] \\ &= (\ln 2)^2 [v_1 2^{-x_1}, \dots, v_n 2^{-x_n}] [v_1, \dots, v_n] = (\ln 2)^2 \sum_{i=1}^n v_i^2 2^{-x_i} \geq 0, \end{aligned}$$

ceea ce înseamnă că matricea hessiană $\nabla^2 \left(\sum_{i=1}^n 2^{-x_i} - 1 \right)$ este pozitiv semidefinită.

Prin urmare, problema de optimizare (119) este convexă și, pentru rezolvarea ei, putem aplica metoda dualității Lagrange.

Considerând variabilele duale $\lambda \in \mathbb{R}$ și $u_i \geq 0$ pentru $i \in [n]$, funcția lagrangeană generalizată pentru problema de optimizare convexă (121) este definită astfel:

$$L(x, \lambda, u) = \sum_{i=1}^n p_i x_i + \lambda \left(\sum_{i=1}^n 2^{-x_i} - 1 \right) - \sum_{i=1}^n u_i x_i. \tag{122}$$

Fie (x, λ, u) un tuplu care satisfac condițiile Karush-Kuhn-Tucker pentru problema (119). Una dintre aceste condiții, și anume *condiția de complementaritate* (engl., *complementary slackness*), afirmă că pentru orice $i \in [n]$, avem

de optimizare convexă cu restricții care a fost dată la ex. 83 —, pentru rezolvarea problemei (121) nu este greșit să folosim metoda multiplicatorilor lui Lagrange, întrucât schimbarea de variabilă $2^{-x_i} = y_i$ pentru $i = 1, \dots, n$ conduce la o reformulare a acestei probleme care satisfac cerințele din definiția neînunii de problemă de optimizare convexă cu restricții. (Puteți verifica singuri detaliile.)

$u_i x_i = 0$. Știm deja de la punctul b că $x_i > 0$, ceea ce conduce la concluzia $u_i = 0$.

Pentru orice $i \in [n]$, întrucât $u_i = 0$,

$$\begin{aligned} \frac{\partial}{\partial x_i} L(x, \lambda, u) &\stackrel{(122)}{=} p_i - \lambda 2^{-x_i} \ln 2 - u_i = p_i - \lambda 2^{-x_i} \ln 2, \\ \text{deci } \frac{\partial L}{\partial x_i} &= 0 \Leftrightarrow p_i = (\lambda \ln 2) 2^{-x_i}. \end{aligned} \quad (123)$$

Însumând egalitățile $p_i = (\lambda \ln 2) 2^{-x_i}$ membru cu membru, după toate valorile lui i , obținem

$$\sum_{i=1}^n p_i = \lambda \ln 2 \sum_{i=1}^n 2^{-x_i} \Rightarrow 1 = \lambda \ln 2,$$

întrucât p_i sunt probabilități, iar $\sum_{i=1}^n 2^{-x_i} = 1$ (vedeți punctul b). Așadar, $\lambda = 1/\ln 2$, iar din relația (123) obținem $p_i = 2^{-x_i}$, de unde prin logaritmare rezultă $x_i = -\log_2 p_i$. Este ușor de verificat că tuplul definit prin $x_i = -\log_2 p_i$, $\lambda = 1/\ln 2$, $u_i = 0$ satisface toate condițiile Karush-Kuhn-Tucker. Prin urmare, (conform proprietății care a fost demonstrată la problema 83) aceasta este soluție optimă pentru problema (121), deci și pentru problema (119).

Observație: Rezultă că funcția obiectiv a problemei (119) este $-\sum_{i=1}^n p_i \log_2 p_i$, adică exact *entropia* variabilei aleatoare a cărei distribuție probabilistă este reprezentată de p_i , $i \in [n]$.

87.

(O variantă a algoritmului Perceptron, pentru care relația de actualizare a ponderilor se obține rezolvând o problemă de optimizare convexă cu restricții)

■ □ • ○ University of Helsinki, 2014 spring, Jyrki Kivinen, HW5, pr. 2

În această problemă vom lucra cu o variantă a algoritmului Perceptron,²²⁶ în care vectorul „actualizat“ de ponderi w_{t+1} este definit ca fiind acel vector w care constituie soluția problemei de optimizare următoare:

$$\begin{aligned} \min_w \quad & \|w - w_t\|^2 \\ \text{a. i. } & y_t w \cdot x_t \geq 1. \end{aligned}$$

unde $w, w_t, x_t \in \mathbb{R}^d$ și $y_t \in \{-1, +1\}$.

- a. Determinați cum anume se poate scrie vectorul w_{t+1} în funcție de w_t , x_t și y_t . (Cu alte cuvinte, rezolvați problema de optimizare dată).
- b. Presupunând că inițial vectorul de ponderi w_1 are toate componente zero, arătați că w_{t+1} este o combinație liniară de instanțele x_1, \dots, x_t . Scrieți versiunea kernel-izată a acestui algoritm.

Răspuns:

²²⁶ Pentru o prezentare a algoritmului Perceptron [neciclic, nestochastic], vedeți problema 16 de la capitolul *Rețele neuronale artificiale*. Pentru varianta kernel-izată a acestui algoritm vedeți problema 19, tot de la capitolul *Rețele neuronale artificiale*. Pentru o variantă stochastică a algoritmului Perceptron, vedeți problema 56.A de la capitolul *Mașini cu vectori-suport*.

a. Fie funcțiiile $f : \mathbb{R}^d \rightarrow \mathbb{R}$ și $p : \mathbb{R}^d \rightarrow \mathbb{R}$, definite prin relațiile $f(w) = \|w - w_t\|^2 = (w - w_t) \cdot (w - w_t) = w \cdot w - 2w \cdot w_t + w_t \cdot w_t$ și respectiv $p(w) = 1 - y_t w \cdot x_t$. Vom demonstra mai întâi că f este funcție convexă.

Stim că o funcție care este definită pe o mulțime convexă și este dublu derivabilă (adică, admite toate derivatele parțiale de ordin secund) este convexă dacă și numai dacă matricea sa hessiană (notată cu $\nabla^2 f(w)$) este pozitiv semidefinită în interiorul mulțimii pe care este definită funcția respectivă.²²⁷ Fiindcă $f(w) = \sum_{i=1}^d (w_i - w_{t,i})^2$, este ușor de observat că

$$\frac{\partial f(w)}{\partial w_i} = 2(w_i - w_{t,i}) \quad \text{pentru orice } i \in \{1, 2, \dots, d\},$$

iar pentru orice $i, j \in \{1, 2, \dots, d\}$

$$\frac{\partial^2 f(w)}{\partial w_i \partial w_j} = 2 \text{ dacă } i = j \quad \text{și} \quad \frac{\partial^2 f(w)}{\partial w_i \partial w_j} = 0 \text{ dacă } i \neq j.$$

Așadar, matricea hessiană a lui f este $\nabla^2 f(w) = 2I_d$ pentru orice $w \in \mathbb{R}^d$. Este imediat că pentru orice vector-colonă $z \in \mathbb{R}^d$ avem $z^\top (2I_d)z = 2\|z\|^2 \geq 0$, ceea ce înseamnă că matricea $\nabla^2 f(w)$ este pozitiv semidefinită în orice punct $w \in \mathbb{R}^d$.

Problema de optimizare din enunț constă în a minimiza $f(w)$ ținând cont de restricția $p(w) \leq 0$. Vom rezolva această problemă rezolvând sistemul reprezentat de condițiile Karush-Kuhn-Tucker:²²⁸

$$\begin{cases} p(w) \leq 0 & (\text{fezabilitate primală}) \\ \lambda \geq 0 & (\text{fezabilitate duală}) \\ \lambda p(w) = 0 & (\text{complementaritate}) \\ \nabla f(w) + \lambda \nabla p(w) = 0 & (\text{staționaritate / optimalitate}) \end{cases}$$

sau, echivalent,

$$\begin{cases} 1 - y_t w \cdot x_t \leq 0 & (i) \\ \lambda \geq 0 & (ii) \\ \lambda(1 - y_t w \cdot x_t) = 0 & (iii) \\ 2w - 2w_t - \lambda y_t x_t = 0 & (iv) \end{cases}$$

Egalitatea (iv) implică

$$w = w_t + (1/2)\lambda y_t x_t. \quad (124)$$

Datorită egalității (iii), vom trata două cazuri.

În primul caz, $\lambda = 0$, deci inegalitatea (ii) este adevărată. De asemenea, din relația (124) rezultă $w = w_t$ și, datorită condiției (i), va trebui ca inegalitatea $1 - y_t w \cdot x_t \leq 0$ să fie satisfăcută.

În cazul al doilea, $\lambda > 0$, din egalitatea (iii) rezultă

$$\begin{aligned} 1 - y_t w \cdot x_t &\stackrel{(124)}{=} 1 - y_t \left(w_t + \frac{1}{2}\lambda y_t x_t \right) \cdot x_t \\ &= 1 - y_t w_t \cdot x_t - \frac{1}{2}\lambda x_t \cdot x_t = 0. \end{aligned} \quad (125)$$

²²⁷Vedeți *Observația* de la finalul rezolvării punctului a.3 de la problema 78.

²²⁸În cele ce urmează, simbolul ∇ desemnează vectorul gradient / derivata vectorială calculată în raport cu w .

Am ținut cont de faptul că $y_t^2 = 1$. Egalitatea $1 - y_t w \cdot x_t = 0$ implică faptul că relația (i) este satisfăcută cu egalitate. Scoțându-l pe λ din egalitatea (125), obținem

$$\lambda = 2 \cdot \frac{1 - y_t w_t \cdot x_t}{\|x_t\|^2}. \quad (126)$$

Observați că datorită relației (ii), din relația (126) rezultă că este necesar ca inegalitatea $1 - y_t w_t \cdot x_t > 0$ să fie satisfăcută.

Înlocuind în relația (124) expresia lui λ obținută la (126), vom obține:

$$\begin{aligned} w &= w_t + \frac{1}{2} \cdot 2 \cdot \frac{1 - y_t w_t \cdot x_t}{\|x_t\|^2} y_t x_t \\ &= w_t + y_t \frac{1 - y_t w_t \cdot x_t}{\|x_t\|^2} x_t. \end{aligned}$$

În sfârșit, punând toate acestea împreună, vom obține *regula de actualizare a ponderilor*:

$$w_{t+1} = w_t + \sigma_t y_t \frac{1 - y_t w_t \cdot x_t}{\|x_t\|^2} x_t,$$

unde

$$\sigma_t = \begin{cases} 0 & \text{dacă } y_t w_t \cdot x_t \geq 1 \\ 1 & \text{dacă } y_t w_t \cdot x_t < 1. \end{cases}$$

b. Ni s-a cerut să arătăm că dacă se ia $w_1 = 0$, atunci rezultă că w_{t+1} este o combinație liniară de vectorii x_1, x_2, \dots, x_t . Cu alte cuvinte, există $\alpha_1, \alpha_2, \dots, \alpha_t \in \mathbb{R}$ astfel încât $w_{t+1} = \alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_t x_t$. Pentru $t = 0$, suma este vidă, deci afirmația ($w_1 = 0 = \sum_{i=1}^0 w_i$) este adevărată. Fie acum $t \in \{1, 2, \dots\}$ și să presupunem că $w_t = \sum_{i=1}^{t-1} \alpha_i x_i$. Folosind regula de actualizare a ponderilor, vom putea scrie:

$$w_{t+1} = \sum_{i=1}^{t-1} \alpha_i x_i + \sigma_t y_t \frac{1 - y_t w_t \cdot x_t}{\|x_t\|^2} x_t = \sum_{i=1}^t \alpha_i x_i,$$

unde am notat $\alpha_t = \sigma_t y_t (1 - y_t w_t \cdot x_t) / \|x_t\|^2$. Folosind principiul inducției complete, rezultă că proprietatea de liniaritate este adevărată pentru orice $t \in \mathbb{N}$.

În algoritmul Perceptron stochastic kernel-izat, inlocuim vectorii / instanțele x_i cu $\phi(x_i)$, imaginile lor în spațiul de „trăsături“. Acest algoritm poate fi formalizat astfel:

For $t = 1, 2, \dots, T$ do

1. Get the instance x_t .
2. Let $p_t = w_t \cdot \phi(x_t) = \left(\sum_{i=1}^{t-1} \alpha_i \phi(x_i) \right) \cdot \phi(x_t) = \sum_{i=1}^{t-1} \alpha_i K(x_i, x_t)$.
Predict $\hat{y}_t = \text{sign}(p_t)$.
3. Get the correct answer y_t .
4. If $y_t p_t < 1$,
 - set $\alpha_t \leftarrow y_t (1 - y_t w_t \cdot \phi(x_t)) / \|\phi(x_t)\|^2 = y_t - \sum_{i=1}^{t-1} \alpha_i K(x_i, x_t) / K(x_t, x_t)$,
 - otherwise
 - set $\alpha_t \leftarrow 0$ (and x_t can be discarded).
5. Let $\|w_{t+1}\|^2 = \sum_{i=1}^t \alpha_i \phi(x_i) \cdot \sum_{j=1}^t \alpha_j \phi(x_j) = \sum_{i=1}^t \sum_{j=1}^t \alpha_i \alpha_j K(x_i, x_j)$.
If $\|w_{t+1}\| > 1$, set $\alpha_i \leftarrow \alpha_i / \|w_t\|$ for all $i \in \{1, 2, \dots, t\}$.

Remarcați faptul că operația de la punctul 5 obligă ponderile să rămână mici (în modul), ceea ce servește la contracararea overfitting-ului.

88.

(Teorema de reprezentare:
un rezultat foarte util pentru kernel-izarea
mai multor algoritmi de învățare automată
care produc separatori liniari)

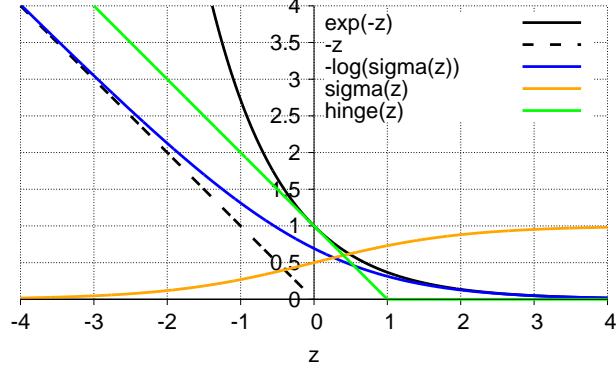
■ • Stanford, John Duchi, Supplemental lecture notes
CMU, 2015 spring, Alex Smola, midterm, pr. 5

Fie $L : \mathbb{R}^2 \rightarrow \mathbb{R}$. Vom spune că L este *funcție de cost* sau *funcție de pierdere* (engl., loss function) dacă ea este i. nenegativă și ii. convexă în raport cu primul argument, pentru orice valoare fixată a celui de-al doilea argument.

Spre exemplu,

- pentru regresia liniară și pentru perceptronul liniar se folosește – cu rol de funcție de cost – pătratul erorii, adică $L(r, y) = (y - r)^2$,
- pentru regresia logistică se folosește *funcția de cost logistică* $L(r, y) = \ln(1 + e^{-yr})$,
- pentru mașini cu vectori-suport (SVM) se folosește *funcția de cost hinge* $L(r, y) = \max\{0, 1 - yr\}$, iar
- pentru AdaBoost, se folosește *funcția de cost [negativ] exponentială* $L(r, y) = e^{-yr}$.

(În figura alăturată, am folosit notația $z = yr$.)



Fie de asemenea instanțele $x_i \in \mathbb{R}^d$ (cu $d \in \mathbb{N}^*$), precum și $y_i \in \mathbb{R}$ (desemnând fie clase, fie — ca în cazul regresiei liniare și al perceptronului liniar — valori reale oarecare pentru funcția de învățat), pentru $i = 1, \dots, n$.

a. Demonstrați că orice soluție w^* a următoarei *probleme de optimizare* (cu termen de regularizare de normă L_2)

$$\min_{w \in \mathbb{R}^d} \left(\frac{1}{n} \sum_{i=1}^n L(w \cdot x_i, y_i) + \frac{\lambda}{2} \|w\|^2 \right), \quad (127)$$

unde $\lambda \geq 0$, iar $\|\cdot\|$ desemnează norma euclidiană, poate fi scrisă sub forma următoare:

$$w^* = \sum_{i=1}^n \alpha_i x_i \text{ pentru anumite valori } \alpha_i \in \mathbb{R}. \quad (128)$$

Sugestie: Puteți rezolva problema în *cazul particular* când $\lambda > 0$, iar funcția de cost $L : \mathbb{R}^2 \rightarrow \mathbb{R}$ este derivabilă în raport cu primul argument, pentru orice valoare fixată a celui de-al doilea argument.

Observație: Acest rezultat se numește *teorema de reprezentare* și se datorează statisticienilor Grace Wahba și George Kimeldorf.²²⁹

²²⁹Vedeți articolul *A correspondence between Bayesian estimation on stochastic processes and smoothing by splines* din Annals of Mathematical Statistics, 41 (2): 495–502, 1970.

b. Se dă o funcție $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^m$, unde $m > d$ (de obicei m este chiar mult mai mare decât d). În continuare vom considera „imaginile“ instanțelor x_i prin funcția ϕ , adică $\phi(x_i)$, pentru $i = 1, \dots, n$. Dintre toate operațiile care pot fi (sau, sunt deja) definite între aceste „imagini“, singura care va fi permisă (pentru a fi folosită) aici este produsul scalar: $\phi(x_i) \cdot \phi(x_j) \stackrel{\text{not.}}{=} K(x_i, x_j)$. Funcția K astfel definită — și, de fapt, extinsă în mod natural la întreg spațiul $\mathbb{R}^d \times \mathbb{R}^d$ — se numește *funcție-nucleu*.²³⁰

Aplicați aşa-numitul “kernel-trick” problemei de optimizare

$$\tilde{w}^* = \arg \min_{\tilde{w} \in \mathbb{R}^m} \underbrace{\left(\frac{1}{n} \sum_{i=1}^n L(\tilde{w} \cdot \phi(x_i), y_i) + \frac{\lambda}{2} \|\tilde{w}\|^2 \right)}_{\text{not.}: J(\tilde{w})} \quad (129)$$

adică:

i. Înlocuiți în expresia $J(\tilde{w})$ din relația (129) pe \tilde{w} cu expresia lui

$$\tilde{w}^* = \sum_{j=1}^n \alpha_j \phi(x_j) \in \mathbb{R}^m, \quad (130)$$

existența coeficienților $\alpha_j \in \mathbb{R}$ care satisfac această egalitate fiind garantată de rezultatul de la punctul a. Veți nota această nouă expresie cu $J(\alpha)$, unde $\alpha \stackrel{\text{not.}}{=} (\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$.

ii. Transformați *forma* expresiei $J(\alpha)$ pe care ati obținut-o la punctul b.i într-o echivalentă, în care nu mai apare funcția $\phi()$ ci (doar) funcția-nucleu $K(,)$.

iii. Similar, scrieți ecuația *separatorului decizional* $\tilde{w}^* \cdot \phi(x) = 0$ — unde x este o instanță oarecare de test din \mathbb{R}^d —, înlocuind \tilde{w}^* conform relației (130) și, în final, eliminați $\phi()$, punând $K(x_i, x)$ în locul produselor scalare $\phi(x_i) \cdot \phi(x)$.

c. Fie problema de optimizare

$$\alpha^* = \arg \min_{\alpha \in \mathbb{R}^n} J(\alpha),$$

unde expresia $J(\alpha)$ a fost obținută la punctul b.ii. Presupunând că se doresc rezolvarea acestei probleme de optimizare folosind metoda gradientului descendente, deduceti *regula de actualizare* pentru vectorul de coeficienți α .

Sugestie: Pentru a vă facilita deducerea gradientului funcției $J(\alpha)$, vă sugerează ca în prealabil să rescrieți expresia funcției $J(\alpha)$, pe care ati obținut-o la punctul b.ii, sub formă matriceală. Concret, notând cu K_i vectorul-colonă $(K(x_i, x_1), \dots, K(x_i, x_n))^\top$, pentru $i = 1, \dots, n$, și apoi cu K matricea de dimensiune $n \times n$ obținută prin „alăturarea“ vectorilor K_i , adică $K = (K_1, \dots, K_n)$, veți putea demonstra egalitatea

$$J(\alpha) = \frac{1}{n} \sum_{i=1}^n L(K_i^\top \alpha, y_i) + \frac{\lambda}{2} \alpha^\top K \alpha.$$

Apoi, puteți, eventual, să folosiți următoarele formule de calcul cu derive de vectoriale:²³¹

²³⁰Pentru mai multe informații despre acest subiect, veți sectiunea *Funcții-nucleu* de la capitolul de *Fundamente*.

²³¹Cf. *Matrix Identities*, Sam Roweis, 1999, <http://www.cs.nyu.edu/~roweis/notes/matrixid.pdf>.

$$(5a) \frac{\partial}{\partial X} a^\top X = \frac{\partial}{\partial X} X^\top a = a$$

$$(5b) \frac{\partial}{\partial X} X^\top AX = (A + A^\top)X$$

Răspuns:

a. Conform Sugestiei din enunț, vom presupune că $\lambda > 0$ și, de asemenea, că funcția de cost $L : \mathbb{R}^2 \rightarrow \mathbb{R}$ este derivabilă în raport cu primul argument, pentru orice valoare fixată a celui de-al doilea argument. Pentru conveniență, vom nota costul (sau riscul empiric) mediu cu regularizare L_2 astfel:

$$J(w) = \frac{1}{n} \sum_{i=1}^n L(w \cdot x_i, y_i) + \frac{\lambda}{2} \|w\|^2.$$

De asemenea, în cele ce urmează vom folosi notația $L'(z, y) \stackrel{\text{not.}}{=} \frac{\partial}{\partial z} L(z, y)$. Gradientul funcției $J(w)$ în raport cu vectorul w este

$$\begin{aligned} \nabla J(w) &= \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial w} L(w \cdot x_i, y_i) + \frac{\lambda}{2} \frac{\partial}{\partial w} \|w\|^2 \\ &= \frac{1}{n} \sum_{i=1}^n L'(w \cdot x_i, y_i) \frac{\partial}{\partial w} (w \cdot x_i) + \frac{\lambda}{2} \frac{\partial}{\partial w} w^2 = \frac{1}{n} \sum_{i=1}^n L'(w \cdot x_i, y_i) x_i + \lambda w. \end{aligned}$$

La a doua egalitate de mai sus am folosit o generalizare a regulii de derivare pentru compunere de funcții, iar la ultima egalitate am folosit regulile de derivare vectorială (5a) și (5b), care au fost menționate la finalul enunțului acestei probleme.

Întrucât în punctul de minim al funcției J gradientul ei se anulează (adică, $\nabla J(w^*) = 0 \in \mathbb{R}^d$),²³² rezultă

$$w^* = -\frac{1}{n\lambda} \underbrace{\sum_{i=1}^n \frac{\partial}{\partial w} L(w^* \cdot x_i, y_i)}_{\in \mathbb{R}} x_i.$$

Considerând $\alpha_i \stackrel{\text{not.}}{=} -\frac{1}{n\lambda} \frac{\partial}{\partial w} L(w^* \cdot x_i, y_i)$, relația precedentă va putea fi rescrisă imediat sub forma egalității (128) din enunț.

b. Înlocuind în expresia funcției $J(\tilde{w})$, care a fost definită în relația (129), argumentul \tilde{w} cu expresia lui \tilde{w}^* din relația (130), vom obține funcția J exprimată de această dată în raport cu vectorul de coeficienți α :

$$\begin{aligned} J(\alpha) &= \frac{1}{n} \sum_{i=1}^n L\left(\left(\sum_{j=1}^n \alpha_j \phi(x_j)\right) \cdot \phi(x_i), y_i\right) + \frac{\lambda}{2} \left(\sum_{j=1}^n \alpha_j \phi(x_j)\right)^2 \\ &= \frac{1}{n} \sum_{i=1}^n L\left(\sum_{j=1}^n \alpha_j (\phi(x_j) \cdot \phi(x_i)), y_i\right) + \frac{\lambda}{2} \left(\sum_{i=1}^n \alpha_i \phi(x_i)\right) \cdot \left(\sum_{j=1}^n \alpha_j \phi(x_j)\right) \\ &= \frac{1}{n} \sum_{i=1}^n L\left(\sum_{j=1}^n \alpha_j K(x_i, x_j), y_i\right) + \frac{\lambda}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j K(x_i, x_j). \end{aligned} \tag{131}$$

²³²Proprietatea pe care am aplicat-o aici generalizează ceea ce știm deja din liceu despre funcțiile reale f de o singură variabilă, x , care sunt derivabile, și anume: $f'(x^*) = 0$, pentru orice punct de optim x^* .

Am ținut cont de faptul că funcția K este simetrică, întrucât produsul scalar al vectorilor este simetric: $K(x_i, x_j) \stackrel{\text{def.}}{=} \phi(x_i) \cdot \phi(x_j) = \phi(x_j) \cdot \phi(x_i) \stackrel{\text{def.}}{=} K(x_j, x_i)$.

Ecuatia *separatorului decizional* $\tilde{w}^* \cdot \phi(x) = 0$, unde x este o instanță oarecare de test din \mathbb{R}^d , se scrie astfel:

$$\begin{aligned} \tilde{w}^* \cdot \phi(x) = 0 &\Leftrightarrow \left(\sum_{i=1}^n \alpha_i \phi(x_i) \right) \cdot \phi(x) = 0 \Leftrightarrow \sum_{i=1}^n \alpha_i (\phi(x_i) \cdot \phi(x)) = 0 \Leftrightarrow \\ &\sum_{i=1}^n \alpha_i K(x_i, x) = 0. \end{aligned} \quad (132)$$

Observați că expresia *separatorului decizional* este liniară în *spațiul de trăsături* \mathbb{R}^m (și anume, $\tilde{w}^* \cdot \phi(x)$), însă în general ea este neliniară în spațiul inițial, \mathbb{R}^d . Pentru exemplificare — și pentru a înțelege mai bine afirmația aceasta —, vă sugerăm să-l înlocuiți în relația (132) pe $K(x_i, x)$ cu expresia nucleului gaussian, $\frac{\|x - x_i\|^2}{2\sigma^2}$, unde $\sigma \in \mathbb{R}_+$ este fixat.²³³

c. Folosind notațiile matriceale introduse în *Sugestia* din enunț, se verifică imediat că relația (131) se rescrie într-adevăr sub forma

$$J(\alpha) = \frac{1}{n} \sum_{i=1}^n L(K_i^\top \alpha, y_i) + \frac{\lambda}{2} \alpha^\top K \alpha.$$

Gradientul funcției $J(\alpha)$ se obține ușor folosind regulile de derivare vectorială indicate de asemenea în *Sugestia* din enunț:

$$\begin{aligned} \nabla_\alpha J(\alpha) &= \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \alpha} L(K_i^\top \alpha, y_i) + \frac{\partial}{\partial \alpha} \frac{\lambda}{2} \alpha^\top K \alpha \\ &\stackrel{(5a)}{=} \frac{1}{n} \sum_{i=1}^n L'(K_i^\top \alpha, y_i) \frac{\partial}{\partial \alpha} (K_i^\top \alpha) + \frac{\lambda}{2} (K + K^\top) \alpha \\ &\stackrel{(5b)}{=} \frac{1}{n} \sum_{i=1}^n L'(K_i^\top \alpha, y_i) K_i + \lambda K \alpha. \end{aligned} \quad (133)$$

Am ținut cont de faptul că matricea K este simetrică (deci $K = K^\top$), întrucât funcția-nucleu K este simetrică, după cum am justificat mai sus.

Acum putem scrie regula gradientului descendente, astfel:

$$\alpha \leftarrow \alpha - \eta \left[\frac{1}{n} \sum_{i=1}^n L'(K_i^\top \alpha, y_i) K_i + \lambda K \alpha \right],$$

unde $\eta > 0$ este *rata de învățare*.

Observație: Dacă ne propunem să obținem regula de actualizare corespunzătoare gradientului descendente *stochastic*, atunci este util (și chiar necesar!) de observat că produsul $K\alpha$ se poate scrie în mod echivalent astfel:

$$K\alpha = \sum_{i=1}^n \alpha_i K_i.$$

²³³Această funcție-nucleu mai este cunoscută și sub numele de “Radial Basis Function”, RBF. Vedeti problemele 74 și 75.

Tinând cont de această relație, precum și de relația (133), rezultă că putem scrie regula gradientului descendente stochastic astfel:²³⁴

$$\alpha \leftarrow \alpha - \eta \left[\frac{1}{n} L'(K_i^\top \alpha, y_i) K_i + \lambda \alpha_i K_i \right].$$

Comentariu:

La problemele 9 și 17 de la capitolul *Metode de regresie*, mai precis de la secțiunea Regresia liniară și respectiv de la secțiunea Regresia logistică am arătat cum anume se pot kernel-iza aceste metode, în vederea obținerii unor separatori neliniari. Pentru aceasta, în ambele cazuri a fost esențială *proprietatea* — care a fost argumentată acolo —, că ponderile w , care se combină liniar cu atributele de intrare, se pot scrie ca o combinație liniară de instanțele de antrenament x_i .

Pentru regresia logistică, echivalența dintre maximizarea funcției de log-verosimilitate condițională și minimizarea costului mediu folosind funcția de cost logistică este demonstrată la problema 13.c de la capitolul *Metode de regresie*. Pentru regresia liniară, echivalența dintre maximizarea funcției de log-verosimilitate condițională și minimizarea sumei pătratelor erorilor a fost demonstrată pentru cazul unidimensional la problema 1.a de la capitolul *Metode de regresie*, iar pentru cazul general la problema 3.d de la același capitol.

Sugestie: Vă recomandăm ca, după ce ati înțeles această problemă, să parcurgeți — pe lângă problemele 9 și 17 de la capitolul *Metode de regresie* — și problema 19 de la capitolul *Rețele neuronale artificiale*, problema 51 de la capitolul *Clusterizare* și problema 15 de la capitolul *Învățare bazată pe memorare*. Este foarte instructiv să analizați elementele pe care aceste probleme le au în comun cu problema de față. De asemenea, menționăm că această problemă poate constitui o foarte bună „poartă“ de intrare în capitolul de *Masini cu vectori-suport*.

²³⁴ *Observație:* În documentul *Supplemental lecture notes*, John Duchi recomandă următoarea variantă a acestei reguli:

$$\alpha \leftarrow \alpha - \eta [L'(K_i^\top \alpha, y_i) K_i + n \lambda \alpha_i K_i],$$

cu scopul de a evita ca estimatorul α calculat prin aplicarea metodei gradientului să fie *deplasat* (engl., biased). Se recomandă de asemenea ca valoarea lui λ să fie mică, iar η să se modifice în raport cu iterația. De exemplu, pentru iterația t se poate lua $\eta_t = 1/\sqrt{t}$ sau un multiplu al acestei valori.

0.2 Fundamente — Probleme propuse

0.2.1 Evenimente aleatoare și formula lui Bayes

89. (Calcul de probabilități elementare)

CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW1, pr. 1.3

Doi soldați A și B trag la țintă. Probabilitatea ca soldatul A să greșească țintă este de $1/5$. Probabilitatea ca soldatul B să greșească țintă este de $1/2$. Probabilitatea ca ambii soldați să greșească simultan țintă este de $1/10$.

- a. Care este probabilitatea ca cel puțin unul dintre soldați să greșească țintă?
- b. Care este probabilitatea ca exact unul dintre cei doi soldați să greșească țintă?

90. (Aplicarea formulei probabilității totale)

CMU, 2011 fall, T. Mitchell, A. Singh, HW1, pr. 1.e

O urnă conține w bile albe și b bile negre. Extragem o bilă dintre acestea, în mod aleatoriu. Apoi repunem bila în urnă, împreună cu alte d bile de aceeași culoare (ca a bilei extrase). După aceasta, extragem din urnă încă o bilă, în mod aleatoriu.

- a. Demonstrați că probabilitatea ca a doua bilă extrasă să fie albă nu depinde de d .
- b. Particularizați [raționamentul dumneavoastră] pentru cazul $w = 2$, $b = 3$ și $d = 7$.

91. (Probabilități condiționate, evenimente aleatoare independente: câteva proprietăți simple)

CMU, 2005 fall, T. Mitchell, A. Moore, HW1, pr. 3.1
 CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW1, pr. 1.1
 CMU, (?) spring, 10-701, HW1, pr. 1.2

Fie A și B două evenimente aleatoare.

- a. Arătați că $P(A | A, B) = 1$.
- b. Arătați că dacă $P(A) = 0$ atunci A și B sunt independente.
- c. Arătați că dacă $P(B) = 1$ atunci $P(A | B) = P(A)$.

Observație: În general, dacă $P(B) \neq 0$, din $P(A | B) = P(A)$ rezultă imediat $P(A \cap B) = P(A)P(B)$, ceea ce, conform definiției, înseamnă că A și B sunt independente. Din acest motiv, se poate spune că $P(A | B) = P(A)$ este o formă mai restrictivă (deși, mai aproape de intuiție!) sau mai „tare“ pentru condiția de independentă a două evenimente aleatoare.

92.

(Legătura dintre forma „slabă“ și forma „tare“ a definiției pentru evenimente aleatoare independente condițional)

• CMU, 2005 fall, T. Mitchell, A. Moore, HW1, pr. 3.2

Folosind doar definiția probabilității condiționate arătați că dacă $P(A | B, C) = P(A | C)$ sau $P(B | A, C) = P(B | C)$ atunci $P(A, B | C) = P(A | C) \cdot P(B | C)$.

93.

(Proprietăți ale funcției de probabilitate)

• ○ CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, midterm, pr. 1.d

Presupunem că evenimentele B_1, B_2, \dots, B_k formează o partitie a spațiului de esantionare (engl., sample space), Ω .²³⁵ Considerăm o funcție de probabilitate P definită pe 2^Ω și un eveniment A pentru care $P(A) > 0$.

Arătați că dacă $P(B_1 | A) < P(B_1)$, atunci $P(B_i | A) > P(B_i)$ pentru cel puțin o valoare a lui i din mulțimea $\{2, 3, \dots, k\}$.

Indicație: Una sau mai multe dintre următoarele proprietăți vă pot fi de folos:

a. $\sum_{i=1}^k P(B_i) = 1$

b. $\sum_{i=1}^k P(B_i \cap A) = P(A)$ o variantă a formulei probabilității totale

c. $P(B_i | A) \cdot P(A) = P(B_i \cap A)$ regula de înmulțire

d. $\sum_{i=1}^k P(B_i | A) = 1$

e. $P(B_i \cap A) + P(B_i \cap \bar{A}) = P(B_i)$.

Demonstrați în prealabil proprietățile de mai sus, în ordinea în care au fost date.

94.

(Formula lui Bayes)

• ○ CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW1, pr. 4

Într-o grupă de copii de la o creșă, 30% dintre ei au ochi căprui, 50% au ochi albaștri, iar restul de 20% au ochi de alte culori. Într-o zi, educatoarea organizează cu acești copii un joc. Pentru prima rundă a jocului, ea selecționează 65% dintre copiii cu ochi căprui, 82% dintre copiii cu ochi albaștri și 50% dintre copiii cu ochi de alte culori.

Care este probabilitatea ca un copil ales în mod aleatoriu din această grupă să aibă ochi albaștri, dacă știm că el n-a fost selecționat pentru prima rundă a jocului?

²³⁵Faptul că B_1, B_2, \dots, B_k (văzute ca mulțimi) formează o partitie a lui Ω revine la: $\bigcup_{i=1}^k B_i = \Omega$ și $B_i \cap B_j = \emptyset$ pentru orice $i \neq j$.

95.

(Formula lui Bayes)

• CMU, 2005 fall, T. Mitchell, A. Moore, HW1, pr. 3.3

Avem două urne. Prima urnă conține 11 bile albe și 4 bile roșii. Cea de-a doua urnă conține 8 bile albe și 5 bile roșii. Se alege în mod aleatoriu cu probabilitate uniformă una din cele două urne. Apoi se extrage o bilă din urna aleasă. Dacă bila extrasă este albă, care este probabilitatea ca ea să provină din prima urnă?

96.

(Probabilități elementare și probabilități condiționate:
Adevărat sau Fals?)

• * prelucrare de Liviu Ciortuz, după
CMU, 2014 fall, W. Cohen, Z. Bar-Joseph, HW1, pr. 4.a
CMU, 2014 fall, W. Cohen, Z. Bar-Joseph, midterm, pr. 1

Marcați cu *adevărat* sau *fals* fiecare dintre afirmațiile următoare:

a. $P(A \cup B \cup C) \geq P(A) + P(B) + P(C)$.

b. $P(A|B, C) = \frac{P(B|A, C) P(A|C)}{P(B|C)}$.

c. $P(A|C) = \sum_{i=1}^n P(A, B_i|C)$, unde $B_i \cap B_j = \emptyset$ pentru orice $i \neq j$ și $\bigcup_{i=1}^n B_i = \Omega$ (evenimentul sigur).²³⁶

d. $P(A|C) = \sum_{i=1}^n P(A|B_i, C) P(B_i|C)$, evenimentele B_i satisfac aceleasi proprietăți ca la punctul precedent.

Justificați în mod riguros fiecare răspuns. Pentru afirmațiile *false*, puteți da un contraexemplu. Pentru afirmațiile *adevărate*, exprimați în câteva cuvinte, dacă este posibil, *tipul* proprietății respective.

0.2.2 Variabile aleatoare

97.

(Variabile aleatoare: medii și varianțe;
exemplificări ale unor proprietăți)

• * CMU, 2016 fall, N. Balcan, M. Gormley, HW1, pr. 6.3.2

Fie X o variabilă aleatoare având media $E[X] = 1$ și varianța $Var(X) = 1$. Calculați:

i. $E[3X]$; ii. $Var(3X)$; iii. $Var(X + 3)$.

²³⁶În locul condiției $\bigcup_{i=1}^n B_i = \Omega$ se poate considera fie proprietatea $P(\bigcup_{i=1}^n B_i) = 1$ fie $A \subseteq \bigcup_i B_i^n$, deși ambele sunt mai laxe (i.e., mai puțin restrictive) decât condiția din enunț.

98.

(Variabila *indicator* pentru un eveniment aleator: calculul mediei)

• ○ CMU, 2015 spring, T. Mitchell, N. Balcan, HW2, pr. 1.c

Fie un eveniment aleatoriu oarecare A , iar X o variabilă aleatoare definită astfel:

$$X = \begin{cases} 1 & \text{dacă evenimentul } A \text{ se realizează} \\ 0 & \text{în caz contrar.} \end{cases}$$

Uneori, X este numită variabilă aleatoare *indicator* pentru evenimentul A . Arătați că $E[X] = P(A)$, unde $E[X]$ reprezintă *media* variabilei X .

99.

(Variabile aleatoare discrete: distribuții comune, distribuții marginale; medii, independentă)

○ CMU, 2009 fall, Carlos Guestrin, HW1, pr. 1.2

Fie două variabile aleatoare discrete cu distribuția (i.e., funcția masă de probabilitate) comună dată în tabelul de mai jos (partea dreaptă).

- a. Calculați valorile medii ale variabilelor X și Y în funcție de α, β, γ și δ .
 b. Găsiți condițiile necesare și suficiente astfel încât variabilele X și Y să fie independente.

X	Y	$P(X, Y)$
0	0	α
0	1	β
1	0	γ
1	1	δ

100.

(Covarianța nulă vs. independentă variabilelor aleatoare binare)

prelucrare de L. Ciortuz, după
• * CMU, 2009 fall, Geoff Gordon, HW1, pr. 3.2

Se știe că atunci când covarianța a două variabile aleatoare este nulă nu rezultă în mod neapărat că variabilele respective sunt independente (vedeți pr. 11.a). Însă, dacă X și Y sunt variabile aleatoare binare luând valori în multimea $\{0, 1\}$ și covarianța lor este nulă, rezultă că X și Y sunt independente (vedeți pr. 11.b).

Arătați că, în cazul în care doar variabila aleatoare X este binară, nu rezultă în mod neapărat că X și Y sunt independente, deși covarianța lor este nulă.

Indicație: Folosiți ca exemplu variabilele aleatoare ale căror distribuții sunt date în tabelul alăturat.

X	Y	$P(X, Y)$
0	-1	0.1
0	0	0.4
0	1	0.1
1	-1	0
1	0	0.4
1	1	0

101.

(Variabile aleatoare discrete: independentă, independentă condițională)

■ • ○ CMU, 2015 spring, T. Mitchell, N. Balcan, HW2, pr. 1.d

Fie X, Y și Z variabile aleatoare luând valori în mulțimea $\{0, 1\}$. Tabelul de mai jos conține probabilitățile corespunzătoare fiecărei asignări posibile (0 sau 1) pentru variabilele X, Y și Z :

	$Z = 0$		$Z = 1$	
	$X = 0$	$X = 1$	$X = 0$	$X = 1$
$Y = 0$	1/15	1/15	4/15	2/15
$Y = 1$	1/10	1/10	8/45	4/45

De exemplu, $P(X = 0, Y = 1, Z = 0) = 1/10$ și $P(X = 1, Y = 1, Z = 1) = 4/45$.

- Este oare variabila X *independentă* de variabila Y ? De ce da, sau de ce nu?
- Este X *independentă condițională* de Y în raport cu variabila Z ? De ce da, sau de ce nu?
- Calculați probabilitatea $P(X = 0 \mid X + Y > 0)$.

102.

(Variabile aleatoare: independentă condițională;
Formula lui Bayes)

o CMU, 2005 fall, T. Mitchell, A. Moore, midterm, pr. 1.1

- Fie variabilele aleatoare H , E_1 și E_2 . Presupunem că vrem să calculăm $P(H \mid E_1, E_2)$, dar nu avem informații referitoare la independentă condițională. Care din următoarele seturi de numere sunt suficiente pentru acest scop?
 - $P(E_1, E_2)$, $P(H)$, $P(E_1 \mid H)$, $P(E_2 \mid H)$
 - $P(E_1, E_2)$, $P(H)$, $P(E_1, E_2 \mid H)$
 - $P(H)$, $P(E_1 \mid H)$, $P(E_2 \mid H)$.
- Presupunem acum că $P(E_1 \mid E_2, H) = P(E_1 \mid H)$ pentru toate valorile posibile ale lui H , E_1 și E_2 . (Așadar, variabila E_1 este independentă condițional de E_2 , în raport cu variabila H .) În acest caz, care dintre seturile de numere de la punctul a sunt suficiente pentru a calcula $P(H \mid E_1, E_2)$?

103.

(Independentă condițională a variabilelor aleatoare:
o condiție suficientă)

prelucrare de Liviu Ciortuz, după

* CMU, 2009 fall, Geoff Gordon, HW2, pr. 2.2
CMU, 2008 fall, Eric Xing, final exam, pr. 9.3

Arătați că dacă distribuția de probabilitate comună a lui X și Y este independentă de W în raport cu Z , atunci variabilele X și W sunt independente în raport cu Z . Similar, variabila Y este independentă de W în raport cu Z .

În notație simplificată: $(X, Y) \perp W \mid Z$ implică $X \perp W \mid Z$ (și similar $Y \perp W \mid Z$).

104.

(Variabile aleatoare continue:
funcția densitate de probabilitate)

* CMU, 2008 spring, Eric Xing, HW1, pr. 1.1.a

Fie funcția $f(x) = ce^{-|x|}$, $-\infty < x < \infty$.

Ce valoare / valori poate avea constanta reală c în aşa fel ca f să poată reprezenta o funcție densitate de probabilitate?

105.

(Variabile aleatoare continue: calculul unei probabilități, folosind funcția densitate de probabilitate)

* CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, midterm, pr. 1.b

Presupunem că funcția densitate de probabilitate a unei variabile aleatoare continue X este definită astfel:

$$p(x) = \begin{cases} \frac{4}{3}(1-x^3) & \text{pentru } 0 \leq x \leq 1 \\ 0 & \text{în caz contrar.} \end{cases}$$

Cât este $P(X < 0)$?

106.

(Variabile aleatoare uniforme continue: p.d.f. condițională; independentă)

* CMU, 2003 fall, T. Mitchell, A. Moore, midterm, pr. 2

Fie X și Y variabile aleatoare continue având funcția densitate de probabilitate comună definită astfel:

$$p(x, y) = \begin{cases} 1 & \text{pentru } 0 < x < 1, |y| < x \\ 0 & \text{în caz contrar.} \end{cases}$$

a. Cât este $p(y | x = 0.5)$?

b. Este variabila X independentă de variabila Y ?

107.

(Variabile aleatoare discrete vs. variabile aleatoare continue; distincția dintre p.m.f. și p.d.f.)

□ • ○ CMU, 2012 spring, Ziv Bar-Joseph, HW1, pr. 1.1

Mickey nu este încă bine inițiat în teoria probabilităților și se confruntă cu chestiunea următoare, despre care el este înclinat să credă că este un paradox:

Fie X o variabilă aleatoare cu distribuția de probabilitate uniformă, definită pe intervalul $[0, \frac{1}{2}]$, și anume $f(x) = 2$ pentru $x \in [0, \frac{1}{2}]$. Întrucât suma probabilităților care corespund unei distribuții aleatoare trebuie să fie 1, Mickey este nedumerit de ce valoarea lui $f(x)$ este mai mare decât suma totală.

Explicați acest paradox. Altfel spus, arătați ce anume nu știe Mickey.

108.

(Variabile aleatoare uniforme continue, variabile aleatoare discrete; independentă, variabile aleatoare condiționate, medii)

○ CMU, 2009 spring, Ziv Bar-Joseph, final exam, pr. 1.1

Fie $X : \Omega \rightarrow [0, 1]$ o variabilă aleatoare continuă având distribuția uniformă (notație: $X \sim \text{Uniform}(0, 1)$). Fie a și b două numere reale, astfel încât $0 < a <$

$b < 1$. Definim (tot pe Ω) variabilele aleatoare discrete Y și Z în funcție de valorile lui X , astfel:

$$Y(\omega) = \begin{cases} 1 & \text{dacă } 0 \leq X(\omega) \leq a, \\ 0 & \text{altfel} \end{cases} \quad Z(\omega) = \begin{cases} 1 & \text{dacă } b \leq X(\omega) \leq 1, \\ 0 & \text{altfel.} \end{cases}$$

a. Sunt variabilele Y și Z independente? Justificați răspunsul.

Indicație: Pentru a găsi răspunsul corect vă sugerăm că ar fi util să completați tabelul de mai jos.

y	z	$P(Y = y)$	$P(Z = z)$	$P(Y = y)P(Z = z)$	$P(Y = y, Z = z)$
0	0				
0	1				
1	0				
1	1				

b. Pentru fiecare dintre valorile z ale variabilei Z , calculați media variabilei condiționate $Y | Z = z$. (Notație: $E_Y[Y | Z = z]$).

109. (Matricea de cross-covarianță pentru doi vectori de variabile aleatoare: o proprietate)

*prelucrare de Liviu Ciortuz, după
□ • ○ MIT, 2006 fall, Tommi Jaakkola, HW1, pr. 5.b*

Fie A și B două matrice de numere reale de dimensiune $p \times q$, iar x un vector aleatoriu de dimensiune $q \times 1$. Demonstrați egalitatea următoare:

$$\text{Cov}(Ax, Bx) = A \text{ Cov}(x) B^\top, \quad (134)$$

unde prin $\text{Cov}(u, v) \stackrel{\text{def.}}{=} E[(u - E[u])(v - E[v])^\top]$ am notat matricea de *cross-covarianță* pentru doi vectori aleatori oarecare u și v , iar prin $\text{Cov}(u) \stackrel{\text{def.}}{=} E[(u - E[u])(u - E[u])^\top]$ am notat matricea de covarianță pentru vectorul u .²³⁷

Observații:

0. Media unui vector (sau a unei matrice) de variabile aleatoare este vectorul (respectiv matricea) formată din mediile respective variabile aleatoare, presupunând că aceste medii există și sunt finite.
1. Proprietatea de *liniaritate a mediilor* (vedeți ex. 9.a) se extinde în mod natural atunci când vectorii (sau matricele) de variabile aleatoare se înmulțesc cu matrice de numere reale.
2. Este imediat că definiția pentru noțiunea de matrice de cross-covarianță generalizează noțiunea de matrice de covarianță: $\text{Cov}(u) = \text{Cov}(u, u)$ atunci când $u = v$.

²³⁷Definiția matricei de covarianță de aici coincide cu cea care a fost dată (în manieră descriptivă) la problema 20.

3. Proprietatea (134) reprezintă o generalizare a proprietății

$\text{Cov}(aX, bY) = ab \text{Cov}(X, Y)$, pentru $\forall a, b \in \mathbb{R}$ și $\forall X, Y$ variabile aleatoare,

proprietate care a fost introdusă la problema 9.c.

4. Știm de la ex. 9.c că pentru orice două variabile aleatoare X și Y sunt loc egalitatea $\text{Cov}(X, Y) = E[XY] - E[X] \cdot E[Y]$. Se poate demonstra că această relație se extinde în mod natural la matricea de cross-covarianță pentru vectori de variabile aleatoare: $\text{Cov}(u, v) = E[uv^\top] - E[u](E[v])^\top$.

110.

(Variabile aleatoare: Adevărat sau Fals?)

□ • ○ CMU, 2005 fall, T. Mitchell, A. Moore, midterm, pr. 1.2

Fie X și Y două variabile aleatoare. Care dintre următoarele afirmații sunt adevărate?

- Dacă X și Y sunt independente, atunci $E[2XY] = 2E[X]E[Y]$ și $\text{Var}[X+2Y] = \text{Var}[X] + \text{Var}[Y]$.
- Dacă X și Y sunt independente, iar $X > 1$, atunci $\text{Var}[X+2Y^2] = \text{Var}[X] + 4\text{Var}[Y^2]$ și $E[X^2 - X] \geq \text{Var}[X]$.
- Dacă X și Y nu sunt independente, atunci $\text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y]$.
- Dacă X și Y sunt independente, atunci $E[XY^2] = E[X]E[Y]^2$ și $\text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y]$.
- Dacă X și Y nu sunt independente și $f(X) = X^2$, atunci $E[f(X)|Y] = E[f(X)]E[Y]$ și $\text{Var}[X+2Y] = \text{Var}[X] + 4\text{Var}[Y]$.

0.2.3 Distribuții probabiliste uzuale

111.

(Distribuția binomială: exemple de aplicare, calcularea unor valori pentru funcția masă de probabilitate)

• ○ CMU, 2009 spring, Ziv Bar-Joseph, HW1, pr. 1.4

Un marinări încearcă să meargă pe o punte alunecoasă, însă datorită mișcărilor navei, el poate face exact un pas la fiecare interval de timp egal cu unitatea, și anume: fie un pas înainte (cu probabilitatea p), fie un pas înapoi (cu probabilitatea $1-p$). Marinări nu poate rămâne imobil.

Pozitia marinăriului la momentul de timp $i \in \{0, 1, \dots\}$ va fi exprimată cu ajutorul unei variabile aleatoare $X_i \in \{-\infty, \dots, -2, -1, 0, 1, 2, \dots, +\infty\}$, astfel:

$$\begin{aligned} X_0 &= 0 \text{ cu probabilitate } 1; \\ P(X_{t+1} = x_i + 1 | X_t = x_i) &= p, \text{ pentru } t \geq 0; \\ P(X_{t+1} = x_i - 1 | X_t = x_i) &= 1-p, \text{ pentru } t \geq 0. \end{aligned}$$

- a. Cât este $P(X_{16} = 8)$, adică probabilitatea ca marinarul să se afle în poziția +8 după exact 16 unități de timp?
- b. Generalizare: Cât este $P(X_n = r)$, adică probabilitatea ca marinarul să se afle în poziția r după exact n unități de timp (considerând $n \geq r$ și $n - r$ număr par)?
- c. Bazat pe formula de la punctul precedent, calculați $P(X_{32} = 16)$.
- d. Este probabilitatea de la punctul c aceeași cu $P(X_{32} = 16 | X_{16} = 8)$, adică probabilitatea ca marinarul să se afle la poziția +16 după exact 32 de unități de timp știind că la momentul de timp 16 se afla la poziția +8? Justificați riguros.

112.

(Distribuția categorială:
calcul de medii și probabilități)

■ • ○ CMU, 2009 fall, Geoff Gordon, HW1, pr. 4

Presupunem că avem n coșuri și m mingi. Aruncăm mingile în coșuri în mod independent și aleatoriu, aşa încât fiecare minge este la fel de probabil să cadă în oricare dintre coșuri. (Pentru simplitate, vom presupune că la orice aruncare mingea cade într-un coș oarecare.)

- a. Care este probabilitatea ca prima minge să cadă în primul coș?
- b. Care este, în medie, numărul de mingi care au căzut în primul coș?

Sugestii:

1. Puteți defini o variabilă aleatoare care să reprezinte (similar unei *funcții-indicator*; vedeti problema 98) faptul că mingea i a căzut în primul coș:

$$X_i = \begin{cases} 1 & \text{dacă mingea } i \text{ a căzut în primul coș,} \\ 0 & \text{altfel.} \end{cases}$$

- 2: Tineți cont de proprietatea de liniaritate a mediei (vedeti problema 9.a).
- c. Care este probabilitatea ca primul coș să rămână gol după aruncarea celor m mingi?
- d. Care este, în medie, numărul de coșuri care rămân goale după aruncarea celor m mingi?

113.

(O mixtură de distribuții Bernoulli; formula lui Bayes)

• ○ CMU, 2009 fall, Carlos Guestrin, HW1, pr. 1.3

Avem două monede dintre care una este perfectă, iar cealaltă nu. În cazul monedei imperfecte, probabilitatea de a obține stema este de $1/20$. Închizând ochii, alegem cu probabilitate de $1/2$ una din cele două monede, după care o aruncăm de două ori.

Calculați:

- a. probabilitatea să obținem stema la prima aruncare;
- b. probabilitatea să fi ales moneda perfectă știind că la ambele aruncări s-a obținut stema.

114.

(O mixtură de distribuții Bernoulli; formula lui Bayes)

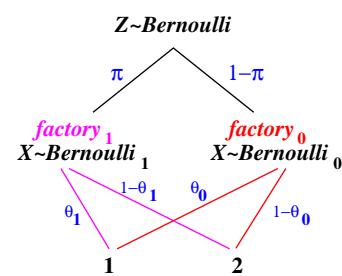
■ □ • ○ MIT, 2016 fall, R. Barzilay, S. Sra, Weekly exercises, week 2, pr. 3

Să zicem că tocmai ai cumpărat un zar cu două fețe — ca în figura alăturată —, la aruncarea căruia se poate produce fie rezultatul 1 fie rezultatul 2.



Intenționezi să folosești acest zar straniu la niște jocuri cu prietenii tăi în seara aceasta, dar mai întâi ai dori să știi care este probabilitatea ca zarul să producă rezultatul 1.

Cunoști faptul că zarul provine fie de la fabrica 0 fie de la fabrica 1, dar nu știi de la care anume dintre ele. Fabrica 0 livrează zaruri care produc rezultatul 1 cu probabilitatea θ_0 , iar fabrica 1 livrează zaruri care produc rezultatul 1 cu probabilitatea θ_1 . Inițial, tu ești de părere că zarul provine de la fabrica 1 cu probabilitatea π .



a. Fără să fi văzut încă vreo aruncare a acestui zar, ci doar ținând cont de cele spuse mai sus, ai putea să calculezi cât este $P(X = 1)$, adică probabilitatea ca acest zar să producă rezultatul 1?

b. Dacă aruncăm zarul și „observăm” rezultatul acestei aruncări, poți să inferezi ceva despre fabrica la care a fost el realizat? (Dedu expresia de calcul pentru probabilitățile $P(Z = 1|X = x)$ și $P(Z = 0|X = x)$; detaliază.)

c. În mod concret, să presupunem că

- $\theta_0 = 1$, adică orice zar livrat de fabrica 0 produce întotdeauna rezultatul 1;
- $\theta_1 = 0.5$, adică orice zar livrat de fabrica 1 este perfect (produce rezultatul 1 cu probabilitatea de 0.5);
- $\pi = 0.7$, adică ești de părere că zarul tău provine cu probabilitate de 0.7 de la fabrica 1.

Acum aruncăm zarul, iar rezultatul este 1. Cât este $P(Z = 1|X = x)$, adică probabilitatea a posteriori ca el să provină de la fabrica 1?

d. Aruncăm zarul din nou, iar rezultatul este iarăși 1! Cât este acum probabilitatea a posteriori ca el să provină de la fabrica 1, adică $P(Z = 1|X_1 = 1, X_2 = 1)$?

e. Răspundeți la aceeași întrebare ca la punctul d, considerând însă că la cea de-a doua aruncare s-a obținut rezultatul 2.

f. Demonstrează în cazul general — adică, fără a folosi valorile numerice date mai sus — că dacă se „observă” rezultatele a două aruncări (notate cu x_a și x_b), atunci

$$P(x_a, x_b) = P(x_b, x_a) \text{ și, de asemenea, } P(Z = z|x_a, x_b) = P(Z = z|x_b, x_a),$$

adică ordinea acestor „observații” nu influențează rezultatul cu privire la probabilitatea a posteriori ca el să provină de la fabrica z .

Indicție: Prima egalitate se demonstrează aplicând mai întâi regula de înmulțire și apoi independența condițională. La demonstrarea celei de-a două egalități se face uz de prima egalitate.

115.

(Calculul mediei și al varianței unei distribuții uniforme continue)

 • CMU, 2012 spring, Ziv Bar-Joseph, HW1, pr. 1.2

Fie X o variabilă aleatoare cu distribuția de probabilitate uniformă, definită pe intervalul $[0, \frac{1}{2}]$, și anume $f(x) = 2$ pentru $x \in [0, \frac{1}{2}]$. Calculați media și varianța acestei distribuții folosind definiția mediei și respectiv a varianței.

116.

(Exemple de distribuții probabiliste continue care nu au medie finită)

prelucrare de Liviu Ciortuz, 2019, după

 • CMU, 2017 fall, Nina Balcan, HW1, pr. 19

a. Distribuția Cauchy de parametru $\theta \in \mathbb{R}$ are p.d.f.-ul (adică, funcția de densitate de probabilitate; engl. probability density function):

$$f(x) = \frac{1}{\pi} \cdot \frac{1}{1 + (x - \theta)^2}.$$

Arătați că media oricărei variabile aleatoare care urmează o distribuție Cauchy este $+\infty$.

b. Procedați similar pentru distribuția continuă având următoarea densitate de probabilitate:

$$p(x) = \begin{cases} \frac{1}{x^2} & \text{dacă } x \geq 1 \\ 0 & \text{altfel.} \end{cases}$$

Indicație: Veți demonstra în prealabil că aceste două funcții sunt într-adevăr funcții de densitate de probabilitate.

117.

(Distribuția gaussiană bidimensională: eșantionare; o implementare simplă)

 • CMU, 2018 spring, N. Balcan, HW0

a. Generați 100 instanțe $x = (x_1, x_2)$ folosind o distribuție gaussiană bidimensională având ca medie vectorul $(0, 0)^\top$ și ca matrice de covarianță matricea identitate, adică $p(x) = -\frac{1}{\sqrt{(2\pi)^2}} \exp\left(-\frac{\|x\|^2}{2}\right)$, iar apoi faceți un “scatter plot” (x_1 vs. x_2).

b. Cum se schimbă scatter plot-ul atunci când înlocuim media cu vectorul $(-1, 1)^\top$?

c. Cum se schimbă scatter plot-ul atunci când în matricea de covarianță dublăm fiecare componentă?

d. Cum se schimbă scatter plot-ul atunci când înlocuim matricea de covarianță cu următoarea matrice?

$$\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

- e. Cum se schimbă scatter plot-ul atunci când înlocuim matricea de covarianță cu următoarea matrice?

$$\begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$$

118. (Mixturi de distribuții gaussiene multidimensionale: media (i.e., vectorul de medii) și matricea de covarianță)

prelucrare de Liviu Ciortuz, după
 • CMU, 2015 fall, Z. Bar-Joseph, E. Xing, HW4, pr. 3

Fie o mixtură de distribuții gaussiene multidimensionale formată din K componente:

$$p(x) = \sum_{k=1}^K \pi_k \mathcal{N}(x|\mu_k, \Sigma_k). \quad (135)$$

- a. Arătați că $E[x] = \sum_{k=1}^K \pi_k \mu_k$.

- b. Arătați că

$$Cov[x] \stackrel{\text{def.}}{=} E[(x - E[x])(x - E[x])^\top] = \sum_{k=1}^K \pi_k [\Sigma_k + \mu_k(\mu_k)^\top] - E[x](E[x])^\top.$$

Observații:

1. Pentru orice variabilă x care urmează o distribuție gaussiană multidimensională de medie μ și matrice de covarianță Σ , adică $x \sim \mathcal{N}(\mu, \Sigma)$ au loc egalitățile: $E[x] = \mu$ și $Cov(X) = \Sigma$.²³⁸

2. Știm de la ex. 9.b că pentru orice variabilă aleatoare X are loc egalitatea $Var[X] = E[X^2] - (E[X])^2$. Se poate demonstra că această relație se extinde în mod natural la [matricea de covarianță pentru] orice vector x de variabile aleatoare: $Cov(x) = E[xx^\top] - E[x](E[x])^\top$. Ca o consecință, se poate demonstra următorul rezultat: pentru orice distribuție gaussiană multidimensională de densitate de probabilitate $\mathcal{N}(\mu, \Sigma)$ urmează că $E[xx^\top] = \Sigma + \mu\mu^\top$.

- 119.

(Mixturi de distribuții [oarecare]: calculul mediilor și al varianțelor)

• CMU, 2010 fall, Aarti Singh, HW1, pr. 2.2.3-5

Modelele de mixturi (engl., mixture models) sunt adeseori folosite în învățarea automată și în statistică.

Presupunem că pe un anumit spațiu de eșantionare (engl., sample space) sunt definite câteva distribuții de probabilitate: $P_i(X) = P(X|C = i)$, $i = 1, \dots, k$. (Ca exemplu, gândiți-vă la cele două zaruri din problema 29; însă aici nu impunem restricții asupra distribuțiilor P_i , deci ele pot fi și distribuții continue.)

Considerăm că dispunem și de o distribuție probabilistă definită peste aceste „componente“ ale mixturii, identificată prin probabilitățile $P(C = i)$, unde C

²³⁸Pentru cazul unidimensional, demonstrația a fost făcută la ex. 32.bc

este o variabilă aleatoare discretă luând k valori. (La problema 29, unde $k = 2$, variabila C este reprezentată de aruncarea monedei.)

- Exprimăți $P(X)$ în funcție de $P_i(X)$ și $P(C = i)$ pentru $i = 1, \dots, k$.
- Exprimăți $E[X]$ în funcție de $E[X|C]$. Justificați în detaliu.
- Exprimăți $\text{Var}[X]$ în funcție de $\text{Var}[X|C]$ și $E[X|C]$. Justificați în detaliu.

120. (Funcția generatoare de *momente* pentru o variabilă aleatoare: definiție; două exemplificări; o proprietate)

• · Liviu Ciortuz, pornind de la Stanford, Machine Learning course, John Duchi, Supplemental Lecture Notes – Hoeffding's inequality

Fie Z o variabilă aleatoare. *Funcția generatoare de momente* (engl., moment generating function) a lui Z este definită astfel:²³⁹

$$M_Z(\lambda) \stackrel{\text{def.}}{=} E[\exp(\lambda Z)],$$

care poate fi infinită pentru anumite valori ale lui λ .

- Funcția generatoare de momente are o proprietate interesantă în raport cu suma de variabile aleatoare. Fie Z_1, \dots, Z_n variabile independente. Demonstrați că

$$M_{Z_1+\dots+Z_n}(\lambda) = \prod_{i=1}^n M_{Z_i}(\lambda).$$

Aceasta înseamnă că funcția generatoare de momente pentru suma unor variabile aleatoare independente este egală cu produsul funcțiilor generatoare de momente pentru respectivele variabile.

- Demonstrați că pentru $Z \sim \mathcal{N}(0, \sigma^2)$ avem

$$M_Z(\lambda) = \exp\left(\frac{\lambda^2 \sigma^2}{2}\right).$$

- Fie S o variabilă aleatoare *Rademacher*. Se mai spune că S este o variabilă aleatoare semn (engl., random sign variable). Prin definiție, $S = 1$ cu probabilitate $1/2$ și $S = -1$ cu probabilitate $1/2$. Demonstrați că

$$M_S(\lambda) \leq \exp\left(\frac{\lambda^2}{2}\right) \text{ pentru orice } \lambda \in \mathbb{R}.$$

Sugestie: Vă recomandăm să folosiți mai întâi dezvoltarea în serie Taylor pentru funcția exponențială, adică $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$, și apoi inegalitatea evidentă $(2k)! \geq 2^k$ pentru orice $k \in \mathbb{N}$.

²³⁹Pentru o prezentare succintă a noțiunii statistice de *moment*, vedeți intrarea *moments* de la pag. 282 din *The Cambridge Dictionary of Statistics*, B. S. Everitt, A. Skrondal, Cambridge University, 2010.

121.

(Distribuții probabiliste discrete și distribuții probabiliste continue)

 • ○ CMU, 2015 spring, T. Mitchell, N. Balcan, HW1, pr. 2.2

Faceți corespondența dintre numele de distribuții probabiliste din coloana din stânga cu funcțiile [masă, respectiv densitate] de probabilitate din coloana din dreapta.

- | | |
|-------------------|--|
| a. gaussiană | f. $p^{1-x}(1-p)^x$ cu $x \in \{0, 1\}$ |
| multidimensională | g. $\frac{1}{b-a}$ pentru $a \leq x \leq b$; 0 în caz contrar |
| b. exponentială | h. $C_n^x p^x (1-p)^{n-x}$ pentru $x \in \{0, \dots, n\}$ |
| c. uniformă | i. $\lambda e^{-\lambda x}$ pentru $x \geq 0$; 0 în caz contrar |
| d. Bernoulli | j. $\frac{1}{\sqrt{(2\pi)^d \Sigma }} \exp\left(-\frac{1}{2} - (x - \mu)^\top \Sigma^{-1} (x - \mu)\right)$ |
| e. binomială | |

122.

(Familia de distribuții exponențiale: exemplificare)

 • * Liviu Ciortuz, pornind de la Stanford, 2007 fall, Andrew Ng, ML course, Lecture Notes, Part III – Generalized Liniar Models

Spunem că o clasă de distribuții face parte din familia exponențială dacă ea poate fi scrisă sub forma următoare:

$$p(y|\eta) = b(y) \exp(\eta^\top T(y) - a(\eta))$$

În această expresie, η se numește *parametrul natural* (sau, *parametrul canonic*) al distribuției; $T(y)$ este *statistica suficientă* (adeseori vom avea $T(y) = y$); iar $a(\eta)$ este numit *funcția de log-partiție*. Cantitatea $e^{-a(\eta)}$ joacă rolul de constantă de normalizare, datorită căreia valorile distribuției $p(y|\eta)$ – indexate în raport cu y – se sumează / integrează la valoarea 1.

Atunci când se fixeză T , a și b , se definește o familie (sau o mulțime) de distribuții care este parametrizată prin intermediul lui η ; variind valorile lui η , obținem diferite distribuții din cadrul acestei famili.

Demonstrați că distribuția Bernoulli, distribuția gaussiană (normală) standard (mai general, distribuția gaussiană de medie μ și varianță 1),²⁴⁰ distribuția exponențială, precum și distribuția categorială sunt parte a familiei de distribuții exponențiale.²⁴¹

²⁴⁰Pentru distribuția gaussiană $\mathcal{N}(\mu, \sigma^2)$ ca membru în familia exponențială, vedeți problema 41.a de la capitolul *Metode de regresie*.

²⁴¹Similar, pentru distribuția multinomială, distribuția Dirichlet și distribuția gaussiană multivariată, vedeți problema 41 de la acest capitol, iar pentru distribuția geometrică și distribuția Poisson, vedeți problema 40.a de la capitolul *Metode de regresie*.

0.2.4 Estimarea parametrilor unor distribuții probabiliste

123.

(Distribuția Bernoulli: câteva chestiuni de bază; estimarea parametrului)

CMU, 2018 spring, Nina Balcan, HW0, pr. A.3.1-4

Considerăm setul de date $S = \{1, 1, 0, 1, 0\}$, obținut prin aruncarea unei monede X de cinci ori, unde 0 indică faptul că la aruncarea monedei a fost obținută față cu banul, iar 1 semnifică faptul că la aruncarea monedei a fost obținută față cu stema.

- a. Calculați $\bar{x} = \frac{1}{|S|} \sum_{x_i \in S} x_i$, *media empirică* (sau, media la eșantionare; engl., the sample mean) pentru acest set de date.
- b. Calculați $\sigma^2 = \frac{1}{|S|} \sum_{x_i \in S} (x_i - \bar{x})^2$, *varianța empirică* (sau, varianța la eșantionare; engl., the sample variance) pentru acest set de date.
- c. Calculați probabilitatea de a „observa“ / obține aceste date, presupunând că ele au fost generate prin aruncarea unei monede perfecte, adică având distribuția de probabilitate $P(X = 1) = 0.5$ și $P(X = 0) = 0.5$.
- d. Remarcați faptul că probabilitatea [generării] acestui set de date ar fi putut fi mai mare dacă valoarea lui $P(X = 1)$ n-ar fi fost 0.5, ci alta. Care este valoarea care maximizează probabilitatea [generării] setului S ? Justificați riguros.

124.

(Distribuția Bernoulli, estimarea parametrului în sens MLE: bias-ul și varianța estimatorului)

CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW2, pr. 1

Fie X o variabilă aleatoare binară, luând valoarea 0 cu probabilitatea p și valoarea 1 cu probabilitatea $1 - p$. Fie instanțele X_1, \dots, X_n produse în mod independent, conform distribuției variabilei X .

- a. Calculați o estimare în sens MLE pentru p ; o veți nota cu \hat{p} .
- b. Este oare \hat{p} , văzut ca variabilă aleatoare, un estimator nedeplasat (engl., unbiased estimator) al lui p ? Demonstrați.
- c. Calculați media erorii pătratice a lui \hat{p} (engl., the expected square error of \hat{p}) în raport cu p , adică $E[(\hat{p} - p)^2]$.

Observație: Veți vedea că, datorită rezultatului de la punctul precedent, această medie coincide cu varianța lui \hat{p} .

Cum evoluează valorile lui $Var[\hat{p}]$ pe măsură ce numărul de instanțe considerate (n) tinde la infinit?

- d. Demonstrați că atunci când știm că p se află în intervalul $[1/4; 3/4]$ și ne sunt date doar $n = 3$ „observări“ ale variabilei X , rezultă că \hat{p} este un estimator *inadmisibil* pentru parametrul p relativ la minimizarea mediei pătratelor erorilor (engl., expected square error) produse prin estimare.

Notă: Se zice că un estimator δ pentru parametrul θ este *inadmisibil* dacă există un alt estimator δ' astfel încât

$R(\theta, \delta') \leq R(\theta, \delta)$ pentru orice valoare a lui θ , și

$R(\theta, \delta') < R(\theta, \delta)$ pentru o valoare oarecare a lui θ ,

unde $R(\theta, \delta)$ este o *funcție de risc* (engl., risk function), care în problema de față va fi considerată media pătratelor erorilor estimatorului, $E[(\hat{p} - p)^2]$.

125.

(Distribuția Bernoulli:
estimarea parametrului (MLE și MAP),
într-un caz particular)

□ • ○ CMU, 2011 fall, T. Mitchell, A. Singh, midterm exam, pr. 2

În acest exercițiu veți estima probabilitatea ca la aruncarea unei monede să apară față *stema* (engl., head), făcând estimări în sensul verosimilității maxime (MLE) și, respectiv, în sensul probabilității maxime a posteriori (MAP).

Presupunem că dispunem de o monedă pentru care probabilitatea de apariție a feței cu stema este $p = 0.5$, adică este o monedă perfectă (engl., fair coin). Totuși, am dori să-i calculăm un estimator, $\hat{\theta}$. În mod obișnuit, un astfel de estimator — pentru o distribuție de tip Bernoulli — se calculează presupunând că el poate lua orice valoare din intervalul $[0, 1]$ (vedeți problema 43.c sau problema 124.a). Aici, în schimb, vom impune ca $\hat{\theta}$ să ia valori într-o mulțime finită, și anume $\{0.3, 0.6\}$.

- Presupunem că am aruncat moneda de 3 ori și am obținut de 2 ori *banul* și o dată *stema*. Calculați estimarea de verosimilitate maximă $\hat{\theta}_{MLE}$ a lui p în raport cu setul de valori posibile, $\{0.3, 0.6\}$.²⁴²
- Presupunem că folosim următoarea distribuție de probabilitate a priori pentru valorile parametrului p :

$$P(p = 0.3) = 0.3 \text{ și } P(p = 0.6) = 0.7.$$

Presupunem din nou că am aruncat moneda de 3 ori și am obținut de 2 ori *banul* și o dată *stema*. Calculați estimarea de probabilitate maximă a posteriori $\hat{\theta}_{MAP}$ a lui p în raport cu multimea de valori posibile, $\{0.3, 0.6\}$, folosind această distribuție a priori.

- Presupunem că aruncăm moneda de un număr de ori care tinde la infinit. În acest caz, care va fi estimarea de verosimilitate maximă $\hat{\theta}_{MLE}$ a lui p în raport cu setul de valori posibile, $\{0.3, 0.6\}$? Justificați răspunsul. (Vă readucem aminte că moneda este perfectă, deci $p = 0.5$.)
- Presupunem din nou că aruncăm moneda de un număr de ori care tinde la infinit. Calculați estimarea de probabilitate maximă a posteriori $\hat{\theta}_{MAP}$ a lui p în raport cu multimea de valori posibile, $\{0.3, 0.6\}$, folosind distribuția de probabilitate a priori care a fost definită la punctul b.

²⁴²LC: Se va considera că aceste două valori sunt echiprobabile.

126. (Estimare în sens MLE pentru distribuția Bernoulli și distribuția binomială: exemplificare)

• CMU, 2018 spring, Nina Balcan, HW2, pr. 2

Presupunem că „observăm“ valorile a n variabile aleatoare i.i.d. X_1, \dots, X_n , toate urmând o [aceeași] distribuție Bernoulli de parametru θ .²⁴³ Cu alte cuvinte, știm că pentru orice $i = 1, \dots, n$, au loc egalitățile

$$P(X_i = 1) = \theta \text{ și } P(X_i = 0) = 1 - \theta.$$

Scopul nostru este să estimăm [în sensul verosimilității maxime, MLE] valoarea parametrului θ , pornind de la aceste valori „observate“, X_1, \dots, X_n .

a. Scrieți formula funcției de log-verosimilitate, $\ell(\theta)$. Această funcție va depinde de variabilele aleatoare X_1, \dots, X_n și de parametrul θ . Aduceți expresia acestei funcții la cea mai simplă formă posibilă. (Așadar, nu vă limitați la a scrie doar definiția funcției de log-verosimilitate.)

Determinați o formulă analitică (engl., a closed form expression) pentru estimarea de verosimilitate maximă, $\hat{\theta}_{MLE}$.²⁴⁴

Folosiți formula pe care le-ați determinat-o mai sus pentru a calcula θ_{MLE} pentru următoarele 10 „observații“:

$$X = (0, 1, 1, 1, 1, 0, 1, 0, 1, 1).$$

b. Acum vom considera o altă distribuție, înrudită cu cea de mai sus. Presupunem că „observăm“ valorile a m variabile aleatoare i.i.d. Y_1, \dots, Y_m , toate urmând o aceeași distribuție binomială $B(n, \theta)$. O distribuție binomială modelizează numărul de apariții ale lui 1 dintr-o secvență de n variabile independente Bernoulli de parametru θ . Cu alte cuvinte,

$$P(Y_i = k) = C_n^k \theta^k (1 - \theta)^{n-k} = \frac{n!}{k!(n-k)!} \theta^k (1 - \theta)^{n-k}.$$

Scrieți expresia funcției de log-verosimilitate $\ell(\theta)$ pentru Y_1, \dots, Y_m , care au valorile k_1, \dots, k_m , respectiv. Această funcție depinde de m, n, k_1, \dots, k_m și θ .

Determinați o formulă analitică pentru estimarea de verosimilitate maximă a parametrului θ .

Apoi folosiți această formulă pentru a calcula θ_{MLE} în cazul a două variabile aleatoare binomiale Y_1 și Y_2 , ambele având parametrii $n = 5$ și θ . Variabilele Bernoulli pentru Y_1 și Y_2 au produs valorile $(0, 1, 1, 1, 1)$ și respectiv $(0, 1, 0, 1, 1)$, prin urmare $Y_1 = 4$ și $Y_2 = 3$.

c. Comparați estimările pe care le-ați obținut la punctele a și b . Dacă aceste două estimări diferă, care credeți că este motivul?

²⁴³Abrevierea i.i.d. înseamnă: independente și identic distribuite.

²⁴⁴Sugestie: Vă reamintim că x^* este un punct critic pentru funcția $f(x)$ dacă $f'(x^*) = 0$. Nu uitați să argumentați de ce a maximiza funcția $\ln f(x)$ este echivalent cu a maximiza funcția $f(x)$.

127. (Estimare în sens MLE pentru parametrul distribuției binomiale folosind metoda gradientului, metoda lui Newton și metoda analitică (prin calcul direct, cu ajutorul derivatei))
prelucrare de Liviu Ciortuz, după CMU, 2008 spring, T. Mitchell, W. Cohen, HW2, pr. 1.2

Funcția masă de probabilitate pentru distribuția binomială (engl., probability mass function, p.m.f.) este definită prin expresia $f(x) = C_n^x p^x (1-p)^{n-x}$.

În cele ce urmează veți considera $n = 100$ și $x = 8$. Găsiți [cât este] estimarea de verosimilitate maximă (engl., maximum likelihood estimation, MLE) pentru parametrul p al distribuției binomiale folosind mai întâi metoda gradientului și apoi metoda lui Newton. Faceți o implementare a acestor două metode în Python / R / Matlab sau în limbajul de programare pe care îl preferați. Folosiți condiția de oprire următoare: $|p_j - p_{j-1}| \leq 0.0001$.

Veți începe aplicarea ambelor metode dând parametrului p valoarea inițială $p_0 = 0.1$. Indicați rezultatul (p_j) care se obține la convergență. Ulterior veți verifica acest rezultat cu [cel obținut prin] metoda analitică.

128. (Distribuția categorială: estimare în sens MAP, folosind distribuția Dirichlet)
CMU, 2014 spring, B. Poczos, A. Singh, HW1, pr. 3

Presupunem că tocmai ai primit un zar cu 6 fețe de la prietenul tău care este statistician. Din păcate, el nu-și amintește exact ce valori au parametrii distribuției categoriale asociate acestui zar. În schimb, își amintește că a generat acești parametri (p_1, p_2, \dots, p_6) folosind următoarea distribuție Dirichlet:

$$P(p_1, p_2, \dots, p_6 | \theta_1, \dots, \theta_6) = \frac{\Gamma(\sum_{i=1}^6 \theta_i)}{\prod_{i=1}^6 \Gamma(\theta_i)} \cdot \prod_{i=1}^6 p_i^{\theta_i-1} \cdot \delta\left(\sum_{i=1}^6 p_i - 1\right),$$

și de asemenea că, în această formulă, el a ales $\theta_i = i$ pentru toate valorile lui $i = 1, \dots, 6$. În notația de mai sus, Γ desemnează funcția Gamma a lui Euler,²⁴⁵ iar δ este funcția delta, adică $\delta(a) = 1$ dacă $a = 0$ și $\delta(a) = 0$ în cazul contrar.

Pentru a estima probabilitățile p_1, p_2, \dots, p_6 , îți propui să arunci zarul de 1000 de ori și să „observi”/determini pentru fiecare față $i \in \{1, \dots, 6\}$ care este numărul (n_i) de a apariții ale feței respective din totalul de 1000 de aruncări ($\sum_{i=1}^6 n_i = 1000$).

- Arată că distribuția Dirichlet este [o] conjugată a priori pentru distribuția categorială.²⁴⁶
- Care este distribuția a posteriori pentru probabilitățile fețelor, adică $P(p_1, p_2, \dots, p_6 | n_1, n_2, \dots, n_6)$?

²⁴⁵Vedeți problema 31.b.

²⁴⁶Pentru definiția noțiunii de *distribuții conjugate* vedeți enunțul problemei 43.B.

129. (Distribuția geometrică: estimarea parametrului, în sens MLE și respectiv în sens MAP)

• prelucrare de Liviu Ciortuz, după
CMU, 2016 fall, N. Balcan, M. Gormley, HW2, pr. 2

Considerăm X_1, \dots, X_n variabile aleatoare independente, toate urmând *distribuția geometrică* (care este o distribuție discretă) de parametru θ . Aceasta înseamnă că pentru oricare variabilă X_i și pentru orice număr natural k avem $P(X_i = k) = (1 - \theta)^k \theta$.²⁴⁷

- Fie un set de date D , conținând „observațiile“ $D = \{X_1 = k_1, X_2 = k_2, \dots, X_n = k_n\}$. Scrieți expresia funcției de log-verosimilitate $\ell_D(\theta)$, ca funcție de D și θ . Este oare valoarea acestei funcții afectată de ordinea în care sunt „observate“ cele n variabile?
- Pornind de la funcția $\ell_D(\theta)$ dedusă la punctul precedent, calculați θ_{MLE} , estimarea de verosimilitate maximă (engl., Maximum Likelihood Estimation, MLE) pentru parametrul θ .
- Fie următoarea secvență de 15 „observații“:

$$X = (0, 21, 23, 8, 9, 2, 9, 0, 7, 8, 20, 9, 7, 4, 17).$$

Aplicând formula dedusă la punctul precedent, calculați valoarea lui θ_{MLE} , mai întâi pentru multimea formată din primele cinci „observații“, adică $(0, 21, 23, 8, 9)$, apoi pentru primele zece „observații“ și, în final, pentru toate cele cincisprezece „observații“.

- Pentru estimarea în sensul probabilității maxime a posteriori (engl., Maximum A posteriori Probability, MAP) a parametrului θ , vom folosi ca distribuție *a priori* distribuția (continuă) Beta.²⁴⁸

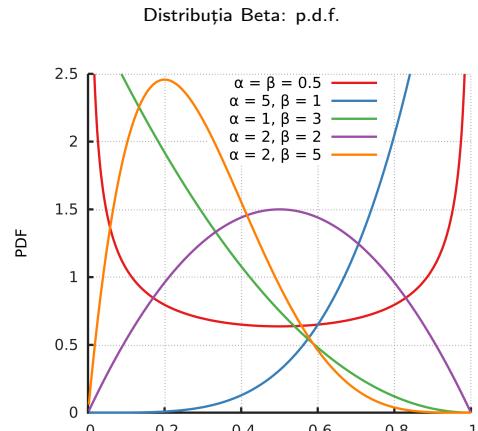
Funcția de densitate de probabilitate pentru distribuția Beta este

$$p(\theta) = \frac{\theta^{\alpha-1} (1-\theta)^{\beta-1}}{B(\alpha, \beta)},$$

unde $B(\alpha, \beta)$ este funcția Beta de argumente $\alpha, \beta \in \mathbb{R}_+$. Funcția Beta se definește astfel:

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)},$$

cu $\Gamma(x) = (x-1)!$ pentru orice $x \in \mathbb{N}^*$.



Deducreți formula de calcul pentru θ_{MAP} , estimarea de probabilitate maximă a posteriori pentru parametrul θ .

²⁴⁷Vedeți nota de subsol 38.

²⁴⁸Distribuția Beta este adeseori folosită ca „distribuție conjugată“ — în contextul estimării parametrilor în sens MAP — nu doar pentru distribuția geometrică, ci și pentru distribuția Bernoulli (vedeți ca exemplu pr. 43) și pentru distribuția binomială. Mai general, distribuția Dirichlet, care este o generalizare a distribuției Beta, este distribuție conjugată pentru distribuția categorială (vedeți pr. 128) și pentru distribuția multinomială. Pentru definiția noțiunii de *distribuții conjugate* vedeți enunțul problemei 43.B.

e. Similar cu cerința de la punctul c, calculați valorile celor trei estimări în sens MAP pentru parametrul θ , folosind de fiecare dată următoarele valori pentru parametrii distribuției Beta: $\alpha = 1$ și $\beta = 2$.

130.

(Estimarea parametrului unei distribuții continue particulare, în sensul verosimilității maxime (MLE))

CMU, 2014 fall, Z. Bar-Joseph, W. Cohen, midterm, pr. 2

Considerăm următoarea distribuție probabilistă, de parametru real nenuл α :

$$p(x|\alpha) = \begin{cases} (\alpha^2 + 1)x^{\alpha^2} & x \in [0, 1] \\ 0 & \text{în rest.} \end{cases}$$

Fiind dat un eșantion format din n puncte, notate X_1, \dots, X_n , generate în mod independent conform distribuției de mai sus, stabiliți care dintre expresiile de mai jos reprezintă valoarea lui α^2 care maximizează verosimilitatea datelor (engl., maximum likelihood estimator) în raport cu această distribuție.

$$\begin{array}{lll} i. & \frac{\sum_i X_i}{n} & ii. \quad \left(\frac{\sum_i X_i}{n} \right)^2 \\ iv. & \frac{-n - \sum_i \ln(X_i)}{\sum_i \ln(X_i)} & iii. \quad -\frac{\sum_i \ln(X_i)}{n} \\ & & v. \quad -\frac{\sum_i X_i - \sum_i \ln(X_i)}{n} \end{array}$$

Indicație: Folosiți funcția de log-verosimilitate.

131.

(Estimare în sens MLE pentru parametrul unei distribuții continue particulare)

CMU, 2001 fall, Andrew Moore, midterm, pr. 2

Fie o distribuție de probabilitate având următoarea funcție de densitate (p.d.f.):

$$p(x) = \begin{cases} p(x) = 0 & \text{dacă } x < 0; \\ p(x) = \frac{2}{w} - \frac{2x}{w^2} & \text{dacă } 0 \leq x \leq w; \\ p(x) = 0 & \text{dacă } x > w. \end{cases}$$

a. Reprezentați grafic funcția de densitate p dată mai sus.

b. Presupunând că variabila aleatoare X urmează distribuția de probabilitate definită de funcția p , care dintre expresiile următoare reprezintă media variabilei X ?

$$\begin{array}{lll} i. \int_{x=-\infty}^{\infty} \left(\frac{2}{w} - \frac{2x}{w^2} \right) dx & ii. \int_{x=-\infty}^{\infty} x \left(\frac{2}{w} - \frac{2x}{w^2} \right) dx & iii. \int_{x=-\infty}^{\infty} w \left(\frac{2}{w} - \frac{2x}{w^2} \right) dx \\ iv. \int_{x=0}^w \left(\frac{2}{w} - \frac{2x}{w^2} \right) dx & v. \int_{x=0}^w x \left(\frac{2}{w} - \frac{2x}{w^2} \right) dx & vi. \int_{x=0}^w w \left(\frac{2}{w} - \frac{2x}{w^2} \right) dx \\ vii. \int_{w=-\infty}^{\infty} \left(\frac{2}{w} - \frac{2x}{w^2} \right) dx & viii. \int_{w=-\infty}^{\infty} x \left(\frac{2}{w} - \frac{2x}{w^2} \right) dx & ix. \int_{w=-\infty}^{\infty} w \left(\frac{2}{w} - \frac{2x}{w^2} \right) dx \\ x. \int_{w=0}^x \left(\frac{2}{w} - \frac{2x}{w^2} \right) dx & xi. \int_{w=0}^x x \left(\frac{2}{w} - \frac{2x}{w^2} \right) dx & xii. \int_{w=0}^x w \left(\frac{2}{w} - \frac{2x}{w^2} \right) dx \end{array}$$

- c. Calculați $P(x = 1|w = 2)$.
- d. Cât este $p(x = 1|w = 2)$?
- e. Cât este $p(x = 0|w = 1)$?
- f. Presupunem că nu cunoaștem valoarea lui w , însă ni se furnizează o instanță / „observație“ a variabilei X : $x = 3$. Care este estimarea de verosimilitate maximă a lui w ?

132. (O distribuție uniformă continuă: estimarea parametrului, în sensul verosimilității maxime (MLE))

CMU, 2015 fall, Z. Bar-Joseph, E. Xing, midterm, pr. 1.2

Considerăm că datele X_1, \dots, X_n sunt generate în mod independent de către o distribuție uniformă p , definită pe discul cu raza θ și centrul în originea sistemului de coordonate din planul euclidian. Așadar, $X_i \in \mathbb{R}^2$ și

$$p(x|\theta) = \begin{cases} \frac{1}{\pi\theta^2} & \text{dacă } \|x\| \leq \theta \\ 0 & \text{în caz contrar,} \end{cases}$$

unde $\|x\| = \sqrt{x_1^2 + x_2^2}$.

Vă cerem să calculați estimarea de verosimilitate maximă pentru θ .

133. (Distribuția exponențială: estimarea parametrilor în sens MLE și respectiv în sens MAP, folosind ca distribuție a priori distribuția Gamma)

CMU, 2015 fall, A. Smola, B. Poczos, HW1, pr. 1.1

a. Distribuția exponențială de parameteru $\lambda > 0$ are funcția densitate de probabilitate $\text{Exp}(x) = \lambda e^{-\lambda x}$, definită pentru $x \geq 0$. Considerăm datele $\{x_i\}_{i=1}^n \sim \text{Exp}(\lambda)$, independente și identic distribuite (i.i.d.). Calculați estimarea de verosimilitate maximă (MLE) λ_{MLE} . Este oare acest estimator deplasat (engl., biased)? Vă readucem aminte că λ_{MLE} este nedeplasat (engl., unbiased) dacă $E[\lambda_{\text{MLE}}] = \lambda$.

b. **Remember:** Distribuția Gamma de parametri $r > 0$ și $\alpha > 0$ are funcția densitate de probabilitate (p.d.f.) următoare:

$$\text{Gamma}(x|r, \alpha) = \frac{\alpha^r}{\Gamma(r)} x^{r-1} e^{-\alpha x} \text{ pentru } x \geq 0,$$

unde Γ desemnează funcția gamma a lui Euler.²⁴⁹

Demonstrați că atunci când $X \stackrel{\text{not.}}{=} \{x_i\}_{i=1}^n$, iar $X \sim \text{Exp}(\lambda)$ și $\lambda \sim \text{Gamma}(r, \alpha)$, rezultă că $P(\lambda|X) \sim \text{Gamma}(r^*, \alpha^*)$ pentru anumite valori r^* și α^* . Cu alte cuvinte, arătați că distribuția Gamma este o distribuție a priori *conjugată* pentru distribuția $\text{Exp}(\lambda)$.²⁵⁰

²⁴⁹Vedeți problema 31.b.

²⁵⁰Vedeți definiția *distribuțiilor conjugate* dată la problema 43.B.

- c. Calculați estimarea de probabilitate maximă a posteriori (MAP) λ_{MAP} în funcție de r și α .
- d. Ce se întâmplă [cu λ_{MLE} și λ_{MAP}] atunci când n devine foarte mare [LC: adică, tinde la infinit]?

134.

(Distribuția gaussiană, cazul $\mu = 0$: estimarea varianței în sensul MLE) CMU, 2009 spring, Ziv Bar-Joseph, HW1, pr. 2.1

Fie X o variabilă aleatoare de distribuție normală (gaussiană) cu media 0 și varianța σ^2 , adică $X \sim \mathcal{N}(0, \sigma^2)$.

- a. Găsiți estimarea de verosimilitate maximă (engl., maximum likelihood estimate) pentru parametrul σ^2 , adică σ_{MLE}^2 .
- b. Este oare estimatorul obținut la punctul precedent *nedeplasat* (engl., unbiased)? Prin definiție, aceasta revine la a stabili dacă are loc (sau nu) egalitatea $E[\sigma_{\text{MLE}}^2] = \sigma^2$.

135.

(Distribuția gaussiană multidimensională: estimarea în sens MAP²⁵¹ a vectorului de medii μ și a matricei de precizie $\Lambda = \Sigma^{-1}$) CMU, 2010 fall, Aarti Singh, HW1, pr. 3.2.2-3

- a. Să zicem că aveți motive să credeți că următoarea distribuție²⁵²

$$gw(\mu, \Lambda) \stackrel{\text{not.}}{=} NW(\mu, \Lambda | \mu_0, s, V, \nu) \stackrel{\text{def.}}{=} \mathcal{N}(\mu | \mu_0, (s\Lambda)^{-1}) \cdot W(\Lambda | V, \nu)$$

este o distribuție a priori convenabilă pentru modelarea parametrilor μ și Λ ai distribuției gaussiene d -dimensionale date la problema 53, cu

$$W(\Lambda | V, \nu) \stackrel{\text{def.}}{=} \frac{|\Lambda|^{(\nu-d-1)/2}}{Z(V, \nu)} \exp\left(-\frac{\text{Tr}(V^{-1}\Lambda)}{2}\right),$$

unde notația $\text{Tr}(\cdot)$ desemnează *urma* (engl., trace) unei matrice diagonale, iar $1/Z(V, \nu)$ este *constanta de normalizare*. Presupunem de asemenea că știți valorile hiper-parametrilor $\mu_0 \in \mathbb{R}^d$, $s > 0$, $\nu > d + 1$, precum și matricea $V \in \mathbb{R}^{d \times d}$, care este pozitiv definită. Derivați estimările în sens MAP (engl., Maximum A posteriori Probability) pentru parametrii μ and Λ . (Le veți nota cu $\hat{\mu}_{\text{MAP}}$ și respectiv $\hat{\Lambda}_{\text{MAP}}$).

- b. Cum se comportă estimările $\hat{\mu}_{\text{MAP}}$ și $\hat{\Lambda}_{\text{MAP}}$ atunci când n tinde la zero sau la infinit? Cum sunt ele în raport cu distribuția a priori (gw) și cu estimările în sens MLE ($\hat{\mu}_{\text{MLE}}$ și $\hat{\Lambda}_{\text{MLE}}$)?

²⁵¹Pentru estimarea în sens MLE a parametrilor acestei distribuții, vedeți problema 53.a.

²⁵²Această distribuție este numită uneori distribuția [a priori] Gauss-Wishart (sau, distribuția normală Wishart).

136. (Distribuția Gamma: estimarea parametrilor în sens MLE folosind metoda gradientului și metoda lui Newton)

prelucrare de Liviu Ciortuz, după

*□ • ○ * CMU, 2015 fall, A. Smola, B. Poczos, HW1, pr. 1.2*

Este posibil ca pentru calcularea unor estimatori să nu existe soluții analitice (engl., closed forms).

Am ilustrat deja această chestiune la problema 52, în raport cu distribuția Gamma, și anume pentru parametrul β al acestei distribuții.²⁵³

Aici vom arăta că într-o astfel de situație putem folosi — în locul metodelor analitice (exacte) — metode de optimizare cum sunt metoda gradientului și metoda lui Newton.

a. Date fiind instanțele $\{x_i\}_{i=1}^n \sim \text{Gamma}(r, \beta)$, elaborați pașii necesari pentru a calcula estimatorii \hat{r}_{MLE} și $\hat{\beta}_{\text{MLE}}$ pentru parametrii distribuției Gamma.

Indicație: Pentru $\hat{\beta}_{\text{MLE}}$ veți putea folosi expresia analitică care a fost dedusă la problema 52. Apoi, pentru a calcula valoarea \hat{r}_{MLE} a parametrului r pentru care se maximizează funcția de log-verosimilitate $\ell(r, \hat{\beta}_{\text{MLE}})$, veți folosi metoda gradientului ascendent.²⁵⁴ Elaborați *regula de actualizare* corespunzătoare.

b. Funcția de log-verosimilitate la care ne-am referit la punctul precedent este dublu derivabilă. Așadar, pentru a optimiza această funcție (adică, pentru a-i calcula maximul) în locul metodei gradientului, putem să folosim metoda lui Newton.²⁵⁵

Elaborați *regula de actualizare* corespunzătoare metodei lui Newton pentru a calcula estimatorul \hat{r}_{MLE} pentru distribuția Gamma.

Observație: Regulile de actualizare obținute la rezolvarea acestui exercițiu vor putea face uz de funcțiile digamma ($\psi(r) \stackrel{\text{def.}}{=} \frac{\Gamma'(r)}{\Gamma(r)}$) și trigamma ($\psi''(r)$), unde $\Gamma(r)$ desemnează funcția a lui Euler.²⁵⁶

c. Pe site-ul acestei culegeri, în directorul corespunzător prezentei probleme,²⁵⁷ veți găsi fișierul estimators.mat, care conține un set de „observații“ / instanțe care au fost generate de către o distribuție Gamma, cu parametrii r și β fixați la anumite valori. Elaborați o implementare a metodei gradientului ascendent și a metodei lui Newton (conform punctului b) și rulați programul respectiv pentru a obține estimările de verosimilitate maximă (MLE) ale parametrilor acestei distribuții. Creați un grafic care să pună în evidență convergența acestor două metode. Care dintre ele a necesitat mai multe iterații? Indicați cele două estimări pe care le-ați obținut.

²⁵³Funcția densitate de probabilitate (p.d.f.) pentru distribuția Gamma $\left(r, \frac{1}{\beta}\right)$ a fost definită la problema 52.

²⁵⁴Metoda gradientului descendente a fost prezentată (prin intermediul unui exemplu) la problema 80.c.

²⁵⁵Pentru o introducere la metoda lui Newton, vedeți *Comentariul* din enunțul problemei 80.

²⁵⁶Pentru calculul acestor funcții există implementări disponibile. Pentru funcția Γ a lui Euler, vedeți problema 31.b.

²⁵⁷<https://profes.info.uaic.ro/~ciortuz/ML.ex-book/implementation-exercises/>

CMU.2015f.ASmola+BPoczos.HW1.pr.1.3.Gamma-distribution.grad-desc+Newton-method-for-MLE.data+sol/

137.

(MLE și parametrizarea alternativă)

 • CMU, 2009 fall, Carlos Guestrin, HW1, pr. 3.1

Estimarea în sensul verosimilității maxime (engl., maximum likelihood estimation, MLE) este o tehnică de folosită pentru estimarea parametrilor unui model. Totuși, trebuie să spunem că parametrizarea unui model nu este în mod neapărat unică. Să luăm, de exemplu, distribuția gaussiană unidimensională. Pentru această distribuție, parametrizarea cea mai de folosită este (μ, σ) , unde μ este media (engl., expected value) distribuției, iar σ este deviația standard (engl., standard deviation). În acest caz, funcția densitate de probabilitate (engl., probability density function, p.d.f.) a distribuției gaussiane este:

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) \quad (136)$$

O *alternativă* la această parametrizare este „parametrizarea naturală“ (numele acesta ține de la modul în care este parametrizată familia de distribuții exponentiale — vedeti problema 122 —, din care face parte și distribuția gaussiană). În acest caz, parametrii sunt (θ, η) , iar funcția densitate de probabilitate este:

$$f(x; \theta, \eta) = \exp\left(-\ln \sqrt{2\pi} + \frac{1}{2} \ln \eta - \eta \frac{x^2}{2} + \theta x - \frac{1}{2} \theta^2 / \eta\right) \quad (137)$$

În acest exercițiu vom vedea dacă un estimator în sensul verosimilității maxime pentru o *anumită parametrizare* ne poate spune ceva despre estimarea aceluiași model atunci când folosim o *altă parametrizare*.

a. Găsiți o funcție / transformare $S(\mu, \sigma) = (\theta, \eta)$ care să facă legătura între reprezentarea obișnuită pentru distribuția gaussiană (dată de formula (136)) și reprezentarea bazată pe parametrizarea naturală (dată de formula (137)). Este funcția aceasta o bijectie? (Aduceți-vă aminte că $\sigma > 0$.)

Indicație: Două polinoame $p(x) = \sum_{i=0}^n a_i x^i$ și $q(x) = \sum_{i=0}^n b_i x^i$ sunt identice pentru orice $x \in \mathbb{R}$ dacă și doar dacă $a_i = b_i$ pentru orice $i = 0, \dots, n$.

b. Fie $f(x; \alpha)$ o funcție densitate de probabilitate. Considerăm $S(\alpha) = \beta$ o reparametrizare a lui α , astfel încât S să fie o bijectie, și $g(x; \beta)$ funcția densitate de probabilitate pentru varianta reparametrizată. Demonstrați că dacă $\hat{\alpha}$ este estimarea de verosimilitate maximă (MLE) pentru $\{x_1, \dots, x_n\}$, un set de instanțe / „observații“ i.i.d., folosind parametrizarea α , atunci $\hat{\beta} = S(\hat{\alpha})$ este o estimare de verosimilitate maximă pentru funcția de densitate reparametrizată.

Indicația 1: Scrieți în mod explicit estimările în sens MLE folosind p.d.f.-urile g și respectiv f . Apoi arătați cum anume se poate trece de la una la cealaltă folosind bijectia S .

Indicația 2: Reparametrizarea înseamnă că $f(x; \alpha) = g(x; S(\alpha))$ pentru orice x și orice α .

c. Considerăm $\{x_1, \dots, x_n\}$ un set de n instanțe i.i.d. ale unei distribuții gaussiane unidimensionale. Cât este estimarea de verosimilitate maximă (MLE) pentru parametrii (θ, η) care au fost introdusi la punctul a? Justificați riguros răspunsul dumneavoastră, pe scurt.

Indicație: Aduceți-vă aminte cât este estimarea în sens MLE pentru (μ, σ) .

d. Considerăm o funcție densitate de probabilitate, $f(x; \alpha)$. Fie $T(\alpha) = \beta$ o reparametrizare a lui α , iar $g(x; \beta)$ o funcție densitate de probabilitate pentru forma reparametrizată a respectivei familii de distribuții. Notăm $\Theta_\beta = \{\alpha | T(\alpha) = \beta\}$ și definim *funcția de verosimilitate indușă*, $M(\beta)$, prin

$$M(\beta) = \sup_{\alpha \in \Theta(\beta)} L(\alpha),$$

unde $L(\alpha)$ desemnează funcția de verosimilitate corespunzătoare lui f (adică, $L(\alpha) = \prod_{i=0}^n f(x_i; \alpha)$ pentru o mulțime de instanțe $\{x_1, \dots, x_n\}$).

Arătați că dacă $\hat{\alpha}$ este estimarea de verosimilitate maximă (MLE) corespunzătoare lui $L(\alpha)$, atunci $\hat{\beta} = T(\hat{\alpha})$ maximizează $M(\beta)$. (*Indicație:* Când anume va conține mulțimea $\Theta(\beta)$ elementul $\hat{\alpha}$?)

Vom numi $\hat{\beta}$ *estimarea de verosimilitate maximă indușă*. Observați faptul că aceasta nu este identică cu *estimarea de verosimilitate maximă*, fiindcă $M(\beta)$ nu este în mod necesar totușa cu funcția de verosimilitate corespunzătoare noii parametrizări. Ce se întâmplă oare atunci când T este o bijecție? Arătați că în acest caz $M(\beta)$ coincide cu funcția de verosimilitate corespunzătoare reparametrizării.

0.2.5 Elemente de teoria informației

138. (Calculul entropiei unei variabile aleatoare discrete)

• CMU, (?) 15-781, midterm example questions, pr. 1.b

Cât este entropia următoarei distribuții de probabilitate (discretă): [0.0625, 0.0625, 0.125, 0.25, 0.5]?

139. (Câștigul de informație: câteva proprietăți și o exemplificare)

prelucrare de Liviu Ciortuz, după
 • CMU, 2011 fall, T. Mitchell, A. Singh, HW1, pr. 2
CMU, 2012 spring, Roni Rosenfeld, HW2, pr. 9

La problema 55 am definit câștigul de informație astfel:

$$IG(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X).$$

De asemenea, am arătat că entropia condițională medie a unei variabile aleatoare X în raport cu o altă variabilă aleatoare Y se poate calcula cu formula

$$H(X|Y) = - \sum_{x \in Val(X)} \sum_{y \in Val(Y)} P(X = x, Y = y) \log_2 P(X = x|Y = y).$$

Următoarea *demonstrație* ne arată că putem calcula câștigul de informație și în alt mod:

$$IG(X; Y) = H(X) - H(X|Y) \quad (138)$$

$$\begin{aligned} &= - \sum_{x \in Val(X)} P(X = x) \log_2 P(X = x) \\ &\quad - \left(- \sum_{x \in Val(X)} \sum_{y \in Val(Y)} P(X = x, Y = y) \log_2 P(X = x|Y = y) \right) \end{aligned} \quad (139)$$

$$\begin{aligned} &= - \sum_{x \in Val(X)} \sum_{y \in Val(Y)} P(X = x, Y = y) \log_2 P(X = x) \\ &\quad + \sum_{x \in Val(X)} \sum_{y \in Val(Y)} P(X = x, Y = y) \log_2 P(X = x|Y = y) \end{aligned} \quad (140)$$

$$= - \sum_{x \in Val(X)} \sum_{y \in Val(Y)} P(X = x, Y = y) (\log_2 P(X = x) - \log_2 P(X = x|Y = y)) \quad (141)$$

$$= - \sum_{x \in Val(X)} \sum_{y \in Val(Y)} P(X = x, Y = y) \log_2 \frac{P(X = x)}{P(X = x|Y = y)} \quad (142)$$

$$= - \sum_{x \in Val(X)} \sum_{y \in Val(Y)} P(X = x, Y = y) \log_2 \frac{P(X = x)P(Y = y)}{P(X = x, Y = y)}. \quad (143)$$

a. Justificați *de ce* anume au loc fiecare dintre egalitățile care intervin în demonstrația de mai sus.

b. Definiți independența a două variabile aleatoare X și Y .

Apoi, folosind rezultatul demonstrat mai sus, și anume

$$IG(X; Y) = \sum_{x \in Val(X)} \sum_{y \in Val(Y)} P(X = x, Y = y) \log_2 \frac{P(X = x, Y = y)}{P(X = x)P(Y = y)},$$

arătați că dacă X și Y sunt independente, atunci $IG(X; Y) = 0$.

c. Demonstrați că $IG(X; X) = H(X)$.

d. Fie X și Y două variabile aleatoare independente, care iau valorile 0 și 1 cu probabilități egale, adică $P(X = 0) = P(X = 1) = P(Y = 0) = P(Y = 1) = 1/2$. Considerăm încă o variabilă aleatoare Z , tot cu valori în multimea $\{0, 1\}$. Distribuția de probabilitate a variabilei condiționate $Z|X, Y$ este definită în tabelul următor:

X	Y	$P(Z = 0 X, Y)$	$P(Z = 1 X, Y)$
0	0	0.8	0.2
0	1	0.2	0.8
1	0	0.2	0.8
1	1	0.8	0.2

i. Folosind regula de înmulțire, precum și independența variabilelor X și Y , completați în tabelul următor distribuția de probabilitate comună $P(X, Y, Z)$.

X	Y	Z	$P(X, Y, Z)$
0	0	0	
0	0	1	
0	1	0	
0	1	1	
1	0	0	
1	0	1	
1	1	0	
1	1	1	

ii. Calculați apoi distribuțiile marginale $P(X, Z)$, $P(Y, Z)$ și $P(Z)$.

X	Z	$P(X, Z)$
0	0	
0	1	
1	0	
1	1	

Y	Z	$P(Y, Z)$
0	0	
0	1	
1	0	
1	1	

Z	$P(Z)$
0	
1	

iii. În final, arătați că $IG(X; Z) = IG(Y; Z) = 0$. (Sugestie: Folosiți rezultatul de la punctul b.)

140.

(Probabilități marginale, entropii, entropii condiționale medii, câștiguri de informație)

prelucrare de Liviu Ciortuz, după
• o CMU, 2011 spring, Roni Rosenfeld, HW2, pr. 1.d

Echipa de fotbal american The Steelers (din Pittsburgh) va juca în cupa Superbowl XLV contra echipei The Green Bay Packers. Pregătindu-și meciul, ei (fotbalistii echipei The Steelers) se gândesc să-și definească strategia de joc în funcție de doi factori majori:

- dacă jucătorul Ben Roethlisberger va fi (sau nu) accidentat la vremea meciului ($Injured = yes / no$), și
- cum anume va fi vremea ($Weather = foggy / rainy / clear sky$).

Iată distribuția comună a acestor două tipuri de evenimente:

	$Weather = foggy$	$rainy$	$clear sky$	$P(Injured)$
$Injured = no$	0.1	0.25	0.35	
$Injured = yes$	0.05	0.1	0.15	
$P(Weather)$				

a. Pornind de la distribuția comună dată, completați ultima linie și ultima coloană din tabelul de mai sus cu valorile corespunzătoare distribuțiilor marginale $P(Weather)$ și $P(Injured)$.

- b. Calculați entropiile $H(Weather)$ și $H(Injured)$.
- c. Calculați entropiile condiționale medii $H(Injured|Weather)$ și $H(Weather|Injured)$.

d. Calculați în entropia comună $H(\text{Injured}, \text{Weather})$ folosind definiția (vedeți problema 55). Verificați apoi că

$$\begin{aligned} H(\text{Injured}, \text{Weather}) &= H(\text{Injured}) + H(\text{Weather} | \text{Injured}) \\ &= H(\text{Weather}) + H(\text{Injured} | \text{Weather}). \end{aligned}$$

Indicație: Vedeți problema 55.c.

e. Calculați câștigurile de informație $IG(\text{Injured}, \text{Weather})$ și $IG(\text{Weather}, \text{Injured})$ folosind definiția (vedeți problema 55). Verificați apoi că

$$IG(\text{Injured}, \text{Weather}) = KL(P_{\text{Injured}}, \text{Weather} || (P_{\text{Injured}} \cdot P_{\text{Weather}})).$$

Indicație: Vedeți problema 139.a și/sau problema 63.b.

Sugestie: Veți putea folosi următoarele aproximări: $\log_2 3 = 1.585$, $\log_2 5 = 2.322$, $\log_2 7 = 2.807$, $\log_2 11 = 3.459$ și $\log_2 13 = 3.700$.

141.

(O margine superioară pentru valoarea entropiei unei variabile aleatoare discrete)

■ □ * CMU, 2003 fall, T. Mitchell, A. Moore, HW1, pr. 1.1

Comentariu: La problema 55.a am demonstrat că entropia oricărei variabile aleatoare discrete este nenegativă ($H(X) \geq 0$).²⁵⁸ La acest exercițiu veți demonstra — tot pentru cazul discret — că există și o margine superioară pentru $H(X)$.

Așadar, fie X o variabilă aleatoare discretă care ia n valori și urmează distribuția probabilistă P . Conform definiției, entropia lui X este

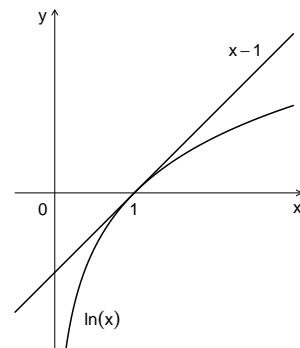
$$H(X) = - \sum_{i=1}^n P(X = x_i) \log_2 P(X = x_i).$$

Arătați că $H(X) \leq \log_2 n$.²⁵⁹

Sugestie (1): Puteți folosi inegalitatea $\ln x \leq x - 1$, care are loc pentru orice $x > 0$.

Sugestie (2): Puteți folosi inegalitatea lui Jensen (vedeți problema 79).

Sugestie (3): Puteți folosi metoda multiplicatorilor lui Lagrange.



²⁵⁸ Extensia acestei proprietăți la cazul variabilelor aleatoare continue este imediată.

²⁵⁹ *Observație:* Se poate demonstra că egalitatea $H(X) = \log_2 n$ are loc atunci și numai atunci când distribuția variabilei X este uniformă, adică $P(X = x_i) = \frac{1}{n}$ pentru $i = 1, \dots, n$, unde $n = |\text{Val}(X)|$.

142. (Entropia comună: forma particulară a relației de „înlănțuire“ în cazul variabilelor aleatoare independente)

■ □ • ○ * prelucrare de Liviu Ciortuz, după CMU, 2012 spring, Roni Rosenfeld, HW2, pr. 7.b

La problema 59 s-a demonstrat că dacă X și Y sunt variabile aleatoare independente, atunci are loc egalitatea $H(X, Y) = H(X) + H(Y)$ și reciproc. În demonstrație s-au folosit rezultate de la problema 144 (sau, echivalent / alternativ, problema 63).

Implicația directă din echivalența de mai sus, și anume X și Y independente $\Rightarrow H(X, Y) = H(X) + H(Y)$ (respectiv X și Y independente $\Rightarrow H(X) = H(X | Y)$ și $H(Y) = H(Y | X)$) se poate obține însă și în mod direct, pornind de la definiția independenței variabilelor aleatoare. Vă cerem să faceți astfel demonstrația acestei implicații. Veți trata mai întâi cazul variabilelor aleatoare discrete și apoi cazul variabilelor aleatoare continue.

Observație: Conform problemei 55.c, unde am demonstrat că $H(X, Y) = H(X) + H(Y | X) = H(Y) + H(X | Y)$ (indiferent dacă X și Y sunt sau nu independente), rezultă că egalitatea $H(X, Y) = H(X) + H(Y)$ este echivalentă cu egalitățile $H(X) = H(X | Y)$ și $H(Y) = H(Y | X)$ (și, de asemenea, cu egalitatea $IG(X; Y) = 0$).

- 143.

(Entropie comună și condițională: formula de „înlănțuire“ condițională)

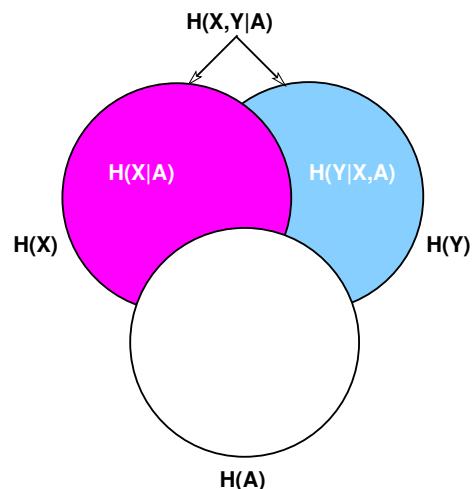
■ □ • ○ * prelucrare de Liviu Ciortuz, după CMU, 2012 spring, Roni Rosenfeld, HW2, pr. 4.b

- a. Demonstrați că proprietatea

$$\begin{aligned} H(X, Y|A) &= H(Y|A) + H(X|Y, A) \\ &= H(X|A) + H(Y|X, A) \end{aligned}$$

este adeverată pentru oricare 3 variabile aleatoare discrete X , Y și A .

Explicați în mod intuitiv, într-o singură frază, care este semnificația proprietății de mai sus.



- b. Arătați că varianta generalizată a regulii de înlănțuire pentru entropii (care a fost demonstrată deja la problema 55.c)

$$H(X_1, \dots, X_n) = H(X_1) + H(X_2|X_1) + \dots + H(X_n|X_1, \dots, X_{n-1})$$

poate fi dedusă [în manieră alternativă] pornind de la relația $H(X, Y) = H(X) + H(Y | X)$ — varianta de bază a regulii de înlănțuire pentru entropii, care a fost demonstrată tot la problema 55.c — și ținând cont de proprietatea care a fost demonstrată la punctul a.

144.

(Nenegativitatea câștigului de informație;
o condiție necesară și suficientă pentru anularea lui)**■ □ * Liviu Ciortuz**

Definiția *câștigului de informație* (sau: *a informației mutuale*) al unei variabile aleatoare X în raport cu o altă variabilă aleatoare Y este $IG(X, Y) = H(X) - H(X | Y) = H(Y) - H(Y | X)$.²⁶⁰ La problema 63 s-a demonstrat — pentru cazul în care X și Y sunt discrete — că $IG(X, Y) = KL(P_{X,Y} || P_X P_Y)$, unde KL desemnează *entropia relativă* (sau: *divergența Kullback-Leibler*), P_X și P_Y sunt distribuțiile variabilelor X și, respectiv, Y , iar $P_{X,Y}$ este distribuția comună a acestor variabile. Tot la problema 63 s-a arătat că divergența KL este întotdeauna nenegativă. În consecință, $IG(X, Y) \geq 0$ pentru orice X și Y .

a. Aici vă cerem să demonstrați inegalitatea $IG(X, Y) \geq 0$ în manieră directă, plecând de la prima definiție dată mai sus, fără a mai apela la divergența Kullback-Leibler.

b. Arătați tot într-o manieră directă că $IG(X, Y) = 0$ dacă și numai dacă X și Y sunt independente. (Într-o manieră indirectă, acest rezultat a fost demonstrat la problema 63.c.)

Sugestie: Puteți folosi următoarea formă [particulară] a inegalității lui Jensen:²⁶¹

$$\sum_{i=1}^n a_i \log x_i \leq \log \left(\sum_{i=1}^n a_i x_i \right)$$

unde baza logaritmului se consideră supraunitară, $a_i \geq 0$ pentru $i = 1, \dots, n$ și $\sum_{i=1}^n a_i = 1$.²⁶² Egalitatea are loc dacă și numai dacă $x_1 = x_2 = \dots = x_n$.

145.

(Cross-entropia — o aplicatie:
selecția modelelor probabiliste)**□ • ○ CMU, 2012 spring, Roni Rosenfeld, HW2, pr. 10**

Avem un zar măsluit (engl., unfair die). Probabilitățile [reale] de apariție pentru fiecare dintre fețele de la 1 la 6 sunt date de distribuția

$$P_{true} = (0.08, 0.55, 0.15, 0.12, 0.05, 0.05).$$

Fără să cunoască acest fapt, două persoane, identificate cu A și respectiv B , ne sugerează următoarele *modele* de probabilitate pentru zarul măsluit:

$$\begin{aligned} P_A &= (0.07, 0.14, 0.24, 0.24, 0.05, 0.26) \\ P_B &= (0.25, 0.13, 0.21, 0.03, 0.11, 0.27) \end{aligned}$$

- a. Elaborați câteva idei relativ la cum am putea măsura / determina care dintre aceste două modele este mai bun.
- b. Calculați cross-entropiile $CH(P_{true}, P_A)$, $CH(P_{true}, P_B)$ și $CH(P_{true}, P_{true})$. Presupunând că alegem ca măsură / mijloc de evaluare a modelelor cross-entropia, care dintre cele două modele (P_A și P_B) credeți că este mai bun?

²⁶⁰Vedeți problema 55.²⁶¹Vedeți problema 79.²⁶²Avantajul la această problemă, comparativ cu problema 63.a, este că aici se lucrează cu o singură distribuție (p), nu cu două distribuții (p și q). Totuși, demonstrația de aici va fi mai laborioasă.

146.

(Varianta condițională a divergenței KL;
 regula de înlănțuire pentru divergența KL;
 divergența condițională KL pentru un vector de variabile mutual independente în raport cu o altă variabilă)

*prelucrare de Liviu Ciortuz, după
 ■ □ • ○ Stanford, 2015 fall, Andrew Ng, HW3, pr. 5.b*

Vă readucem aminte — vedeți problema 63 — că entropia relativă sau divergența Kullback-Leibler (KL) dintre două distribuții discrete $P(X)$ și $Q(X)$ este definită astfel:²⁶³

$$KL(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}.$$

Pentru conveniență, vom presupune că $P(x) > 0$ și $Q(x) > 0$ pentru orice x . Uneori în cele ce urmează vom scrie $KL(P||Q)$ sub forma $KL(P(X)||Q(X))$.

Entropia relativă (sau, divergența KL) dintre două distribuții condiționale $P(X|Y)$ și $Q(X|Y)$ se definește în felul următor:²⁶⁴

$$KL(P(X|Y) || Q(X|Y)) = \sum_y P(y) \left(\sum_x P(x|y) \log \frac{P(x|y)}{Q(x|y)} \right).$$

Aceasta poate fi văzută ca fiind media [probabilistă a] divergenței KL dintre distribuțiile condiționale corespunzătoare pentru x (adică, dintre $P(X|Y = y)$ și $Q(X|Y = y)$), media fiind calculată în raport cu valorile y ale variabilei aleatoare Y .

a. Demonstrați că are loc următoarea proprietate (de tip „regulă de înlănțuire“) pentru divergența KL:

$$\begin{aligned} KL(P(X, Y) || Q(X, Y)) &= KL(P(X) || Q(X)) + KL(P(Y|X) || Q(Y|X)) \\ &= KL(P(Y) || Q(Y)) + KL(P(X|Y) || Q(X|Y)). \end{aligned} \quad (144)$$

b. Demonstrați că dacă $X = (X_1, \dots, X_n)$ și pentru orice $i \neq j$ avem că variabila X_i este independentă condițional de X_j în raport cu variabila Y , atunci

$$KL(P(X|Y) || Q(X|Y)) = \sum_{i=1}^n KL(P(X_i|Y) || Q(X_i|Y)). \quad (145)$$

147.

(Echivalența dintre maximizarea verosimilității unui set de date și minimizarea divergenței KL față de distribuția uniformă)

■ □ • ○ Stanford, 2015 fall, Andrew Ng, HW3, pr. 5.c

Presupunem că nici se dă o problemă de estimare a parametrilor unei distribuții probabiliste și că dispunem de setul de date „de antrenament“ $D = \{x^{(i)}; i = 1, \dots, m\}$. Considerăm distribuția empirică $\hat{P}(x) = \frac{1}{m}$, adică distribuția uniformă peste setul de date D . Așadar, a face eșantionare (engl., sampling) cu

²⁶³Atunci când P și Q sunt p.d.f.-uri (funcții de densitate de probabilitate) pentru variabile aleatoare continue, dacă în formula de mai sus înlocuim simbolul de sumare cu integrală, proprietatea enunțată în acest exercițiu rămâne valabilă. Totuși, din motive de simplitate, aici vom lucra doar cu funcții masă de probabilitate / distribuții discrete.

²⁶⁴Această noțiune se numește *divergență condițională KL*.

distribuția empirică înseamnă a alege în mod aleatoriu un exemplu din setul de date D .

Presupunem că avem o anumită familie de distribuții P_θ , indexată după parametrul θ . (Dacă doriți, puteți considera $P_\theta(x)$ ca notație alternativă pentru $P(x; \theta)$.)

Demonstrați că a găsi valoarea de verosimilitate maximă (engl., maximum likelihood estimate) pentru parametrul θ este echivalent cu a găsi acea distribuție P_θ pentru care entropia relativă (divergența KL) față de distribuția empirică \hat{P} este minimă. Așadar, vă cerem să demonstrați relația următoare:

$$\arg \min_{\theta} KL(\hat{P} || P_\theta) = \operatorname{argmax}_{\theta} \sum_{i=1}^m \ln P_\theta(x^{(i)}).$$

Comentariu: Ne vom referi acum la relația dintre proprietatea enunțată în acest exercițiu pe de o parte și estimarea parametrilor făcută de algoritmul de clasificare binară Bayes Naiv cu atrbute de tip Bernoulli de cealaltă parte (vedeți capitolul de *Clasificare bayesiană*). În modelul de clasificare Bayes Naiv se presupune că P_θ este de forma următoare: $P_\theta(x, y) = p(y) \prod_{i=1}^n p(x_i|y)$. Conform regulii de înlățuire pentru divergența KL (vedeți relațiile (144) și (145) de la ex. 146), avem:

$$KL(\hat{P} || P_\theta) = KL(\hat{P}(y) || p(y)) + \sum_{i=1}^n KL(\hat{P}(x_i|y) || p(x_i|y)).$$

Această relație arată că problema găsirii maximului verosimilității / minimumului divergenței KL (pentru a estima valorile parametrilor) se descompune în $2n + 1$ probleme de optimizare independente: una pentru distribuția a priori a clasei, $p(y)$, și câte una pentru fiecare distribuție condițională $p(x_i|y)$, corespunzătoare feacării trăsăturii x_i în raport cu fiecare din cele două valori posibile pentru eticheta y . Concret, găsirea estimării de verosimilitate maximă pentru fiecare dintre aceste probleme în mod individual rezultă în maximizarea verosimilității distribuției comune. (O observație similară se poate formula și în legătură cu estimarea parametrilor rețelelor bayesiene (engl., Bayesian networks).)

Observație: Pentru clasificatorul Bayes Naiv cu variabile booleene, legătura dintre maximizarea verosimilității datelor de antrenament și estimarea parametrilor în sens MLE este arătată la problema 38.ab de la capitolul *Clasificare bayesiană*.

148.

(Determinarea distribuțiilor unidimensionale care — în anumite condiții — au entropii maxime, folosind metoda dualității Lagrange)

*prelucrare de Andi Munteanu și Liviu Ciortuz, 2019, după □ • * CMU, 2013 spring, A. Smola, B. Poczos, HW2, pr. 1.bc*

a. Demonstrați că dintre toate distribuțiile continue care sunt definite pe un interval $[a, b] \subset \mathbb{R}$, cu a și b fixați ($a < b$), cea mai mare entropie o are distribuția continuă uniformă.²⁶⁵

²⁶⁵Vă reamintim că la problema 141 se demonstrează că, dată fiind o variabilă aleatoare $X : \Omega \rightarrow \mathbb{R}$ cu funcția de probabilitate P și $Val(X) \stackrel{\text{not.}}{=} \{x_1, \dots, x_n\}$, entropia variabilei X este maximă dacă $P(X = x_1) = \dots = P(X = x_n) = 1/n$. Acest rezultat este varianța discretă a proprietății care a fost enunțată la punctul a al problemei de față.

- b. Demonstrați că dintre toate distribuțiile continue care sunt definite pe intervalul $[0, +\infty)$, pentru care media este egală cu μ (un anumit număr real, pozitiv, fixat), cea mai mare entropie o are distribuția exponențială.
- c. Demonstrați că dintre toate distribuțiile continue care sunt definite pe intervalul \mathbb{R} , pentru care media este este egală cu μ (un anumit număr real, fixat) și varianța este egală cu σ^2 (un anumit număr real, pozitiv, fixat), cea mai mare entropie o are distribuția gaussiană.

Sugestii:

1. Folosiți metoda multiplicatorilor lui Lagrange.
2. Puteți folosi următoarele proprietăți ale derivatelor funcționale:²⁶⁶

- *liniaritatea:* $\frac{\delta(\lambda F + \mu G)(\rho)}{\delta\rho(x)} = \lambda \frac{\delta F(\rho)}{\delta\rho(x)} + \mu \frac{\delta G(\rho)}{\delta\rho(x)}$, unde λ și μ sunt constante reale;
- *regula produsului:* $\frac{\delta(FG)(\rho)}{\delta\rho(x)} = \frac{\delta F(\rho)}{\delta\rho(x)} G(\rho) + F(\rho) \frac{\delta G(\rho)}{\delta\rho(x)}$;
- *regula înlățuirii unei funcționale cu o funcție derivabilă:*

$$\frac{\delta F(g(\rho))}{\delta\rho(y)} = \frac{\delta F(g(\rho))}{\delta g(\rho(y))} \cdot \frac{dg(\rho)}{d\rho(y)};$$
- $\frac{\delta}{\delta\rho} \int_a^b F(\rho) dx = \frac{\delta F(\rho)}{\delta\rho}$.

149.

(Proprietăți ale entropiei: Adevărat sau Fals?)

- * CMU, 2011 spring, Roni Rosenfeld, HW2, pr. 2.a.1
CMU, 2008 fall, Eric Xing, final exam, pr. 1.4
CMU, 2012 spring, Roni Rosenfeld, HW2, pr. 7

Stabiliți dacă următoarele propoziții sunt adevărate sau false.

- a. Entropia nu este negativă.
- b. $H(X, Y) \geq H(X) + H(Y)$ pentru orice două variabile aleatoare X și Y .
- c. Dacă X și Y sunt variabile aleatoare independente, atunci $H(X, Y) = H(X) + H(Y)$.

²⁶⁶Pentru definiția noțiunii de derivată funcțională și pentru un exemplu de calculare a unei astfel de derive funcționale, vedeți problema 66.

În formulele care vor urma, F și G sunt *funcționale* (adică, funcții definite peste multimi de funcții care au anumite proprietăți, pe care însă nu le menționăm aici), ρ este o funcție de argument real x , iar g este o funcție care se aplică lui ρ sau, echivalent, se compune cu ρ .

0.2.6 Funcții-nucleu

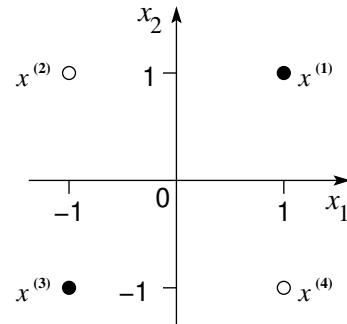
150.

(Conceptul \neg XOR: neseparabilitate vs. separabilitate liniară folosind o funcție de „mapare“ particulară)

prelucrare de Liviu Ciortuz, după CMU, 2020 fall, Pat Virtue, recitation 10, pr. 3

Conceptul \neg XOR corespunde unei probleme de clasificare neliniară; el poate fi reprezentat grafic ca în figura alăturată.

Vom considera funcția de „mapare“ (engl., feature transformation) ϕ definită prin relația $\phi((x_1, x_2)^\top) \stackrel{\text{def.}}{=} (x_1, x_1 x_2)^\top$, unde $x_1, x_2 \in \mathbb{R}$.



a. Calculați expresia funcției-nucleu $K(x, z)$ care corespunde „mapării“ ϕ .

b. Într-un nou sistem de coordonate, corespunzător spațiului de „trăsături“ determinat de „maparea“ ϕ , reprezentați grafic „imaginile“ instanțelor din setul / conceptul \neg XOR prin funcția ϕ .

c. Este oare setul de date corespunzător conceptului \neg XOR liniar separabil în spațiul de „trăsături“?

Dacă da, indicați care este *separatorul optimal* (adică, separatorul cu *magine geometrică maximă*) în spațiul de „trăsături“.

Apoi scrieți ecuația *graniței de decizie* (engl., decision boundary) din spațiul original care corespunde separatorului optimal pe care l-ați identificat în spațiul de „trăsături“.

Desenați această graniță de decizie pe figura de mai sus. Hașurați *zonele de decizie* corespunzătoare instanțelor pozitive.

151.

(Găsirea mapării care corespunde unei funcții-nucleu polinomiale particulare)

o CMU, 2010 fall, Aarti Singh, HW3, pr. 3.3.c

Se consideră funcția $K(u, v) = u \cdot v + 4(u \cdot v)^2$, unde u și v sunt vectori din \mathbb{R}^2 . Arătați că există o funcție ϕ astfel încât $K(u, v) = \phi(u) \cdot \phi(v)$.

Indicație: Veți determina efectiv $\phi(x)$, expresia funcției ϕ pentru argumentul $x = (x_1, x_2) \in \mathbb{R}^2$.

152.

(Funcții-nucleu: exemplificarea unor chestiuni de bază)

• CMU, 2016 fall, N. Balcan, M. Gormley, HW4, pr. 2.1

Presupunem că lucrăm cu instanțe de forma $x = (x_1, x_2, x_3)$ din \mathbb{R}^3 . Definim funcția de „mapare“ a trăsăturilor $\phi(x) = (x_1^2, x_2^2, x_3^2, \sqrt{2}x_1x_2, \sqrt{2}x_1x_3, \sqrt{2}x_2x_3)$.

- a. Deduceți expresia funcției-nucleu care corespunde acestei mapări, $K(x, z)$. Exprimăți răspunsul dumneavoastră mai întâi în funcție de $x_1, x_2, x_3, z_1, z_2, z_3$, iar apoi aduceți expresia lui $K(x, z)$ la forma cea mai simplă.
- b. Presupunem că vrem să calculăm valoarea funcției-nucleu $K(x, z)$ pentru două instanțe oarecare $x, z \in \mathbb{R}^3$. Câte adunări și câte înmulțiri sunt necesare dacă
- i. „mapăm“ vectorii de intrare (x și z) în spațiul de trăsături și apoi aplicăm produsul scalar?
 - ii. folosim expresia funcției-nucleu K care a fost obținută la punctul a?

153.

(Normalizarea funcțiilor-nucleu)

*prelucrare de Liviu Ciortuz, după*** University of Helsinki, 2005 spring, Jyrki Kivinen, HW9, pr. 1.b*

Fie $K : \mathbb{R}^{d \times d} \rightarrow \mathbb{R}$ o funcție-nucleu, iar ϕ „maparea“ corespunzătoare. Definim funcția \bar{K} printr-o procedură de „normalizare“:²⁶⁷

$$\bar{K}(x, z) = \frac{K(x, z)}{\sqrt{K(x, x)K(z, z)}}.$$

Arătați că funcția astfel obținută este de asemenea funcție-nucleu. Precizați care este expresia funcției de „mapare“ (notată cu $\bar{\phi}$) corespunzătoare lui \bar{K} . Ce puteți spune despre valoarea numerică a lui $\|\bar{\phi}(x)\|$, pentru x arbitrar din \mathbb{R}^d ? Ca o consecință, ce interpretare geometrică („vizualizare“) puteți să-i asociați lui $\bar{\phi}(x)$ în raport cu $\phi(x)$?

Credetă că are vreo importanță practică operația de „normalizare“ a instanțelor de antrenament în învățarea automată (în general, nu doar în contextul funcțiilor-nucleu)?

Sugestie: Pentru un exemplu de folosire a unei funcții-nucleu normalize pentru rezolvarea unei probleme de clasificare, vedeti exercițiul 41 de la capitolul *Mașini cu vectori-suport*.

154.

(Funcții-nucleu obținute prin aplicarea proprietăților de „construcție“: câteva exemple)

- CMU, 2015 spring, T. Mitchell, N. Balcan, HW6, pr. 4.2
- CMU, 2017 fall, Nina Balcan, HW3, pr. 1.2.a
- Stanford, 2020 summer, Andrew Ng, HW2, pr. 3.g

La problemele 68.c, 69 și 70 am arătat că putem folosi funcții-nucleu care au fost deja definite, ca să construim funcții-nucleu noi.

Argumentați de ce funcțiile propuse mai jos sunt sau nu sunt funcții-nucleu valide. Veți presupune că $x, z \in \mathbb{R}^d$, cu $d \in \mathbb{N}^*$.

- a. $K(x, z) = 5(x \cdot z)$.

²⁶⁷În cele ce urmează vom lucra doar în condițiile în care $\phi(x) \neq 0$ și $\phi(z) \neq 0$.

- b. $K(x, z) = (x \cdot z)^3 + (x \cdot z + 1)^2$.
- c. $K(x, z) = (x \cdot z)^2 + \exp(-\|x\|^2) \exp(-\|z\|^2)$.
- d. $K(x, z) = \|x\|^2 \|z\|^2 e^{(\|x\|^2 + \|z\|^2)}$.
- e. funcția $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ definită prin relația $K(x, z) \stackrel{\text{def.}}{=} K'(\psi(x), \psi(z))$, știind că $\psi : \mathbb{R}^d \rightarrow \mathbb{R}^p$, $p \in \mathbb{N}^*$, iar funcția $K' : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}$ este funcție-nucleu.

155.

(Funcții-nucleu: exemple de operații [cu funcții] care nu conduc la „construirea“ de noi funcții-nucleu)

 • * CMU, 2013 spring, A. Smola, B. Poczos, HW2, pr. 4.2

- a. Considerăm $K_1(x, x')$ și $K_2(x, x')$ două funcții-nucleu. Demonstrați că diferența lor, $K_1 - K_2$, nu este în mod neapărat funcție-nucleu.

Sugestie: Vă reamintim că pentru orice funcție-nucleu $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ și pentru orice set de instanțe x_1, \dots, x_m din \mathbb{R}^d , matricea-nucleu corespunzătoare (numită și matrice Gram), este pozitiv semidefinită.²⁶⁸

- b. Știm că $\exp(-\|x - y\|^2)$ este funcție-nucleu.²⁶⁹ Demonstrați că $\exp(\|x - y\|^2)$ nu este funcție-nucleu validă.

Sugestie: Construiți o matrice Gram care nu este pozitiv semidefinită.

156.

(Funcții-nucleu: o inegalitate / o margine superioară pentru valoarea absolută a unei funcții-nucleu oarecare)

 • ○ CMU, 2015 spring, Alex Smola, midterm, pr. 8.2

Demonstrați că pentru orice funcție-nucleu are loc inegalitatea

$$K^2(x, x') \leq K(x, x) K(x', x').$$

Sugestie: Folosiți inegalitatea Cauchy-Bunyakovsky-Schwarz: $|x \cdot y| \leq \|x\| \|y\|$, unde · desemnează produsul scalar al vectorilor, iar $\| \cdot \|$ reprezintă norma euclidiană.²⁷⁰

Consecință: Din inegalitatea de mai sus, folosind notația de la problema 153, rezultă $|\bar{K}(x, x')| \leq 1$, ceea ce înseamnă că valorile funcțiilor-nucleu normalizate, în modul, sunt întotdeauna mărginite superior de 1.

²⁶⁸Vedeți problema 68.b.²⁶⁹Vedeți problema 74.

²⁷⁰Un exemplu simplu: $(ax + by)^2 \leq (a^2 + b^2)(x^2 + y^2)$. În teoria probabilităților, inegalitatea Cauchy-Bunyakovsky-Schwarz este cunoscută sub forma următoare: $(E[XY])^2 \leq E[X^2] E[Y^2]$.

157.

(O funcție-nucleu particulară,
care asigură separabilitate liniară [în spațiul de „trăsături“]
pentru orice set de date de antrenament)

□ • CMU, 2015 spring, T. Mitchell, N. Balcan, HW6, pr. 4.1

Fie $X \stackrel{\text{not.}}{=} \{x_1, \dots, x_m\} \subset \mathbb{R}^d$. Considerăm funcția-nucleu $K : X \times X \rightarrow \mathbb{R}$ definită prin relația următoare:

$$K(x, x') = \begin{cases} 1, & \text{dacă } x = x' \\ 0, & \text{în caz contrar.} \end{cases}$$

a. Demonstrați că aceasta este o funcție-nucleu validă, folosind teorema lui Mercer. Apoi, găsiți o funcție de „mapare“ a trăsăturilor $\phi : X \rightarrow \mathbb{R}^m$ astfel încât $K(x, x') = \phi(x) \cdot \phi(x')$ pentru orice $x, x' \in X$.

b. Demonstrați că mulțimea $\phi(X) \stackrel{\text{not.}}{=} \{\phi(x) | x \in X\}$ — unde ϕ este maparea găsită la punctul a — este *liniar separabilă*, indiferent de modul cum s-ar face *etichetarea* instanțelor din mulțimea X (și, corespunzător, din mulțimea $\phi(X)$).

c. Întrucât cu ajutorul acestei funcții-nucleu obținem [în noul spațiul de trăsături] separabilitate liniară indiferent de cum a fost făcută etichetarea instanțelor date, suntem tentați să credem că funcția aceasta ne poate servi în mod perfect pentru a învăța orice concept-target. Este totuși posibil ca în practică să se ajungă la concluzia că ideea aceasta nu este atât de bună precum pare la prima vedere?

158.

(O proprietate simplă a nucleului de tip RBF)

CMU, 2011 spring, Tom Mitchell, HW6, pr. 1.1.b

Oricarei funcții-nucleu îi este asociată în mod implicit o anumită funcție („mapare“) ϕ care transformă instanțele de antrenament $x \in \mathbb{R}^d$ într-un spațiu Q de dimensiune [de obicei] mult mai mare decât d , și care satisfac proprietatea următoare: $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$. În continuare vom lua drept nucleu funcția cu baza radială (RBF)

$$K(x_i, x_j) = e^{-\frac{1}{2\sigma^2} \|x_i - x_j\|^2}$$

unde notația $\| \cdot \|$ corespunde normei euclidiene.

Demonstrați că $\|\phi(x_i) - \phi(x_j)\|^2 < 2$ unde ϕ este maparea asociată nucleului RBF definit mai sus.

Sugestie: Vă puteți inspira din rezultatul / rezolvarea problemei 71.

159.

(Crearea de trăsături noi, folosind funcții cu baza radială)

• MIT, 2018 spring, T. Jaakkola, R. Barzilay, midterm, pr. 4.a-f

În acest exercițiu vom pune în evidență cum anume, folosind funcții cu baza radială, putem crea noi [spații de] trăsături asociate instanțelor de antrenament sau de test. Pentru a defini astfel trăsături noi, avem nevoie de o mulțime de *instanțe* $E = (e_1, \dots, e_k)$,²⁷¹ unde fiecare e_i este un punct din \mathbb{R}^d , iar d este dimensiunea spațiului [de trăsături] original. Funcția pe care o vom folosi aici pentru crearea de trăsături este următoarea:

$$\phi(x) = \begin{bmatrix} \exp(-\beta\|e_1 - x\|^2) \\ \exp(-\beta\|e_2 - x\|^2) \\ \dots \\ \exp(-\beta\|e_k - x\|^2) \end{bmatrix}.$$

unde $\beta > 0$ este o constantă reală.

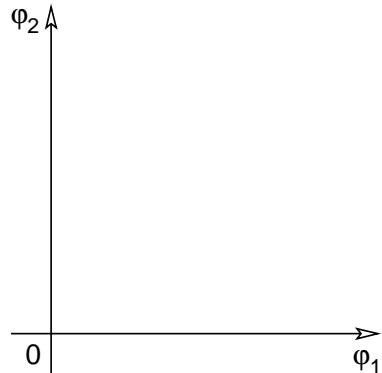
a. Cât este dimensiunea vectorului $\phi(x)$?

b. Considerăm următorul caz particular: $d = 1$, $k = 2$, $E = (1, 2)$, $\beta = 1$, iar setul de exemple de antrenament este format din instanțele etichetate $(0, +1), (1, -1), (2, +1)$.

Este acest set de exemple de antrenament liniar separabil în spațiul original? Da sau Nu? (Justificați răspunsul.) În cazul afirmativ, veți preciza ecuația unui separator liniar.

c. Pe sistemul de axe de coordonate alăturat, desenați punctele $\phi(0)$, $\phi(1)$ și $\phi(2)$.²⁷² Etichetați-le folosind *convenția* noastră obișnuită de notare: simbolul \bullet desemnează instanțe pozitive, iar simbolul \circ instanțe negative. Următorul tabel de valori vă poate fi de folos:

v	$\exp(-v)$
0	1.0
1	0.37
2	0.13
4	0.02
8	0.0003



Este setul de exemple de antrenament care a fost dat la punctul b liniar separabil în noul spațiu de trăsături? (În acest nou spațiu, instanțele etichetate sunt obținute prin „maparea“ / transformarea $(x_i, y_i) \rightarrow (\phi(x_i), y_i)$.) Răspundeți cu Da sau Nu și justificați. În cazul afirmativ, veți preciza ecuația unui separator liniar.

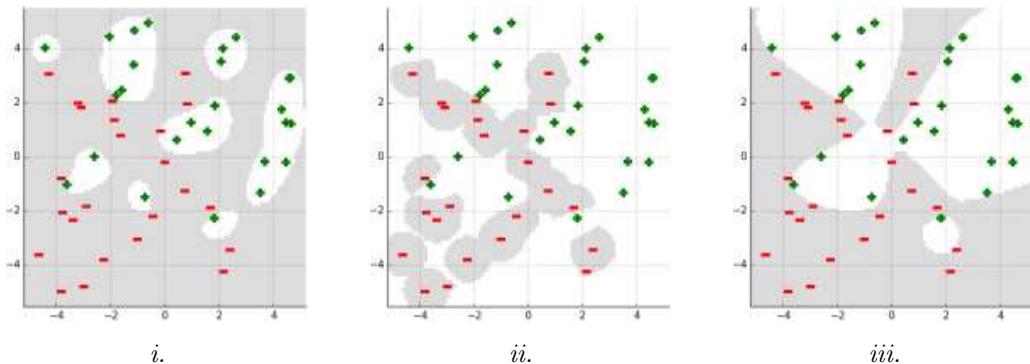
d. Pentru a alcătui mulțimea E , o modalitate des întâlnită constă în a folosi instanțe x_i din setul de date de antrenament.

Considerăm acum că se lucrează în spațiul euclidian bidimensional. Cele trei imagini de mai jos ne prezintă separatorii obținuți pentru cazul în care se

²⁷¹Ordinea acestor instanțe contează, de aceea am folosit notația cu paranteze în loc de accolade.

²⁷²În imagine am notat cu ϕ_1 și ϕ_2 cele două componente ale funcției ϕ , adică $\phi(x) = (\phi_1(x), \phi_2(x))$, unde $\phi_1(x) = \exp(-\beta\|e_1 - x\|^2)$ și $\phi_2(x) = \exp(-\beta\|e_2 - x\|^2)$.

rulează algoritmul perceptron [în varianta Rosenblatt] kernelizat,²⁷³ parametrul β luând valori în mulțimea $\{0.1, 1.0, 1000\}$. Zonele gri corespund regiunilor din spațiu care, după antrenare, au fost clasificate ca fiind negative. Pentru fiecare dintre aceste 3 imagini indicați ce valoare a lui β (dintre cele 3 valori menționate mai sus) corespunde respectivei imagini.



160. (Matrice pozitiv semidefinite: câteva proprietăți)

• CMU, 2013 spring, A. Smola, B. Poczos, HW2, pr. 3

a. [Produsul pe componente²⁷⁴ a două matrice pozitiv semidefinite]

Fie K_1 și $K_2 \in \mathbb{R}^{n \times n}$ două matrice pozitiv semidefinite. Demonstrați că *produsul lor pe componentă*, care este prin definiție matricea K cu proprietatea $K(i, j) = K_1(i, j)K_2(i, j)$ pentru orice $i, j \in \{1, \dots, n\}$, este de asemenea pozitiv semidefinită.

Sugestie: Pornind de la doi vectori n -dimensionali $u = (u_1, \dots, u_n)^\top \sim \mathcal{N}(0, K_1)$ și $v = (v_1, \dots, v_n)^\top \sim \mathcal{N}(0, K_2)$,²⁷⁵ puteți considera matricea de covarianță a vectorului $w = (u_1v_1, \dots, u_nv_n)^\top$. Vă readucem aminte că matricea de covarianță a oricărui vector de variabile aleatoare este pozitiv semidefinită.²⁷⁶

b. [Produsul a două matrice pozitiv semidefinite]

Fie A și $B \in \mathbb{R}^{n \times n}$ două matrice pozitiv semidefinite. Arătați că

- i. matricea AB nu este în mod neapărat pozitiv semidefinită.
- ii. matricea A^m este pozitiv semidefinită, pentru orice $m \in \mathbb{Z}_+$.

Sugestie: Puteți folosi un rezultat teoretic, cunoscut sub numele de *teorema de factorizare spectrală finit-dimensională*, care afirmă că orice matrice simetrică M [de numere reale] poate fi „diagonalizată” cu ajutorul unei matrice ortogonale. Mai exact, aceasta înseamnă că pentru orice matrice simetrică $M \in \mathbb{R}^{n \times n}$ există o matrice ortogonală $U \in \mathbb{R}^{n \times n}$ cu proprietatea că $D = U^\top MU \in \mathbb{R}^{n \times n}$ este matrice diagonală. (Faptul că matricea U este ortogonală înseamnă că $U^\top U = I$, unde I este matricea identitate.)

²⁷³Vedeți problema 19 de la capitolul *Rețele neuronale artificiale*.

²⁷⁴Engl., elementwise product.

²⁷⁵Așadar, vectorii u și v urmează fiecare căte o distribuție gaussiană multidimensională de medie 0 și matrice de covarianță K_1 , respectiv K_2 .

²⁷⁶Vedeți problema 20.

161.

(Funcții-nucleu: Adevărat sau Fals?)

 • ○ *Stanford, 2009 fall, Andrew Ng, practice midterm, pr. 6.c*

Fie x_1, x_2 și x_3 trei puncte oarecare din \mathbb{R}^p , cu $x_1 \neq x_2$, $x_1 \neq x_3$ și $x_2 \neq x_3$. Considerăm de asemenea punctele z_1, z_2 și z_3 din \mathbb{R}^q , arbitrar alese, dar fixate. În aceste condiții putem defini o funcție-nucleu $K : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}$ [LC: căreia îi va corespunde o anumită funcție de „mapare“ ϕ , definită, desigur, pe \mathbb{R}^p] astfel încât pentru orice $i, j \in \{1, 2, 3\}$ să aibă loc egalitatea $K(x_i, x_j) = z_i \cdot z_j$. Adevărat sau fals?

0.2.7 Metode de optimizare în învățarea automată

162. (Calcularea derivatei (respectiv a gradientului) unor funcții relativ simple.

Funcții convexe: câteva chestiuni simple, pornind de la definiție)

prelucrare de Liviu Ciortuz, după • ○ *CMU, 2019s, Nina Blacan, HW0, pr. Calculus**CMU, 2009 fall, Carlos Guestrin, HW2, pr. 2.a**CMU, 2013 fall, A. Smola, G. Gordon, midterm practice questions,*
*pr. 1.c*a. Dacă $y = x^3 + x - 5$, cât este derivata lui y în raport cu x ?Dacă $f(x_1, x_2) = x_1 (\sin x_2) e^{-x_1}$, cât este gradientul lui f (notație: $\nabla_f(x)$)?b. Fie funcție f cu valori reale, definită pe un interval $[a, b]$ (sau, mai general, pe o submulțime convexă a unui spațiu vectorial). Conform definiției,²⁷⁷ funcția f este convexă dacă pentru orice două puncte x_1 și x_2 din $[a, b]$ urmează că $f(tx_1 + (1 - t)x_2) \leq tf(x_1) + (1 - t)f(x_2)$, pentru orice $t \in [0, 1]$. Demonstrați că funcția $f(x) = |x|$, cu $x \in \mathbb{R}$, este convexă.c. Dacă o funcție f nu este dublu derivabilă, adică matricea sa hessiană nu este definită, atunci funcția f nu poate fi convexă. Adevărat sau fals?

163.

(Funcții de cost / pierdere în învățarea automată [profundă]: studiul convexității lor)

 • ○ *CMU, 2015 fall, A. Smola, B. Poczos, HW1, pr. 3.2*

În acest exercițiu vom lucra cu câteva dintre cele mai des folosite funcții de cost / pierdere (engl., loss functions) din învățarea automată [profundă]. Vom introduce mai întâi anumite *notații* pe care le vom folosi în continuare:

- $w \in \mathbb{R}^d$ reprezintă o pondere;
- $x_i \in \mathbb{R}^d$ este o instanță (vector de trăsături); vom considera că primul element din vectorul x_i are întotdeauna valoarea 1, deci în cele ce urmează nu vom folosi [niciun] termen liber (engl., bias term);
- $r_i = w \cdot x_i \in \mathbb{R}$;

²⁷⁷Vedeți problema 78.

- $y_i \in \mathbb{R}$ este eticheta instanței x_i .

Stabilități, pe bază de demonstrație, pentru fiecare dintre funcțiile de la punctele $a - d$ de mai jos dacă ele sunt (sau nu) convexe.²⁷⁸

Sugestie: Puteți face apel la proprietățile prezentate la problema 78.

- a. Funcția de cost *logistica* (engl., logistic loss function): $L(r_i) = \ln(1 + e^{-r_i})$.
- b. Funcția *ReLU* (engl., Rectified Linear function): $R(r_i) = \max(0, -r_i)$.
- c. Funcția de cost *hinge* (engl., hinge loss): $H([r_1, r_2, \dots, r_N]) = \sum_{i=1}^N \max(0, 1 - y_i r_i)$.
- d. Funcția de cost *softmax* (engl., softmax loss) pentru straturi cu conexiuni totale (engl., fully connected layers): considerăm matricea $W \in \mathbb{R}^{d \times k}$, pentru care vom nota coloana i cu w_i ; $W^\top x_i = [w_1 \cdot x_i, \dots, w_k \cdot x_i]^\top \stackrel{\text{not.}}{=} [r_i(1), r_i(2), \dots, r_i(k)]^\top$; $S_s(W) = \ln \left(\sum_{j=1}^k e^{r_i(j)} \right) - r_i(s)$.
- e. Bazându-vă pe punctele precedente, precizați motivele din cauza cărora se poate „rupe“ convexitatea funcțiilor obiectiv folosite din învățarea profundă. Justificați pe scurt de ce unele dintre funcțiile obiectiv folosite în învățarea profundă nu sunt convexe.

164.

(Metoda gradientului: exemple de aplicare,
și un rezultat teoretic: [relativ la] convergență)

□ • CMU, 2019 spring, Nina Blacan, HW3, pr. 1

Fie $f : \mathbb{R}^d \rightarrow \mathbb{R}$ o funcție derivabilă. Vă reamintim că algoritmul *gradientului descendente* — la care ne vom referi aici în forma abreviată, GD — pornește de la un anumit punct $x^{(0)} \stackrel{\text{not.}}{=} (x_1^{(0)}, \dots, x_d^{(0)}) \in \mathbb{R}^d$, iar după aceea algoritmul „actualizează“ în mod iterativ poziția $x^{(k)}$, folosind o *regulă de actualizare*, care se exprimă astfel pentru coordonata i (unde $i \in \{1, \dots, d\}$):

$$x_i^{(k+1)} \leftarrow x_i^{(k)} - \eta \frac{\partial}{\partial x_i} f(x^{(k)}).$$

Aici $\frac{\partial f}{\partial x_i} : \mathbb{R}^d \rightarrow \mathbb{R}$ este derivata parțială a funcției f în raport cu coordonata i , apoi $\frac{\partial}{\partial x_i} f(x^{(k)})$ reprezintă valoarea acestei derive parțiale în punctul $x^{(k)}$, iar $\eta > 0$ se numește *rata de învățare* (engl., learning rate).

A. Exemple de aplicare a metodei GD...

a. ...în \mathbb{R}

Fie funcția $f(x) = 4x^2 - 2x + 1$ și rata de învățare $\eta = 0.1$. Mai întâi, scrieți expresia derivatei $\frac{\partial}{\partial x} f(x)$. Apoi, pornind de la $x^{(0)} = 1$, pentru fiecare din pașii $k = 0, 1$ și 2 ai algoritmului GD, scrieți vectorul gradient $\frac{\partial}{\partial x} f(x^{(k)})$, noua poziție $x^{(k+1)}$, precum și noua valoare a funcției, $f(x^{(k+1)})$.

²⁷⁸Pentru punctele a și c , puteți vedea graficele din figura de la problema 88.

b. ...în \mathbb{R}^2

Fie $f(x_1, x_2) = x_1^2 + \sin(x_1 + x_2) + x_2^2$. Mai întâi, scrieți regula de actualizare pentru gradientul descendente, folosind o rată de învățare oarecare, $\eta > 0$. După aceea, pentru fiecare dintre următoarele *două cazuri* faceți grafice pentru funcția f , care va fi considerată ca fiind definită pe domeniul $[-4, +4] \times [-4, +4]$, folosind [soft pentru] *diagrame de izocontur*,²⁷⁹ precum și săgeți de la o poziție a algoritmului GD către următoarea poziție.²⁸⁰

- *Cazul i:* Determinați punctele (x_1, x_2) pentru primii 10 pași făcuți de GD, începând cu $(x_1^{(0)}, x_2^{(0)}) = (3, -3)$ și folosind $\eta = 0.4$,
- *Cazul ii:* Determinați punctele (x_1, x_2) pentru primii 10 pași făcuți de GD, începând cu $(x_1^{(0)}, x_2^{(0)}) = (3, -3)$ și folosind $\eta = 0.8$,

Ce observați?

B. Analiza algoritmului GD

Metoda gradientului descendente este în sine un *algoritm de căutare locală*: la fiecare pas, el folosește informația din punctul curent (și anume, vectorul gradient) pentru a determina mișcarea pe care trebuie să o facă, în direcția minimizării funcției (care este tocmai direcția descreșterii vectorului gradient). Algoritmul GD nu are cunoștință despre modul în care se comportă funcția pe întreg domeniul de definiție. Totuși, deși GD este un algoritm de căutare locală, dacă f are proprietăți convenabile — mai precis, dacă f este convexă și β -netedă (engl., β -smooth), noțiune care este definită mai jos — și are un punct de minim [de abscisă] x^* , atunci există un $\eta > 0$ pentru care $x^{(k)} \rightarrow x^*$ atunci când $k \rightarrow +\infty$. În cele ce urmează vă vom ghida cum să faceți demonstrația acestei proprietăți în cazul unidimensional.

c. [Lema descreșterii]

Presupunem că f este o funcție derivabilă pe tot domeniul de definiție [LC: cu derivata continuă peste tot²⁸¹] și β -netedă, unde $\beta > 0$. Prin *definiție*, faptul că f este β -netedă înseamnă că

$$f(y) \leq f(x) + f'(x)(y - x) + \frac{\beta}{2}(y - x)^2 \text{ pentru orice } x \text{ și } y. \quad (146)$$

Stim că $x^{(k+1)} = x^{(k)} - \eta f'(x^{(k)})$ la pasul curent al algoritmului GD. Substituind $x = x^{(k)}$ și $y = x^{(k+1)}$ în relația de mai sus, obținem

$$f(x^{(k+1)}) \leq f(x^{(k)}) + f'(x^{(k)}) \left(x^{(k+1)} - x^{(k)} \right) + \frac{\beta}{2} \left(x^{(k+1)} - x^{(k)} \right)^2.$$

Observație: Conform problemei 78.a, proprietatea 2 (vedeți *Observația* de acolo), faptul că f este convexă (în ipoteza în care f este derivabilă, cu derivata continuă peste tot) este echivalent cu:

²⁷⁹O *diagramă de izocontur* este o modalitate de reprezentare a unor suprafețe, care în mod normal ar trebui să fie reprezentate în spațiul 3D, în spațiul 2D. Ea constă în a desena în plan mai multe curbe, fiecare curbă fiind formată din acele puncte din domeniul de definiție al funcției în care aceasta (adică, funcția) ia o anumită valoare, fixată. Procedura aceasta este similară cu modul în care se face reprezentarea pe hărți a regiunilor muntoase, folosind curbele de nivel / altitudine.

²⁸⁰Exemple de *diagrame de izocontur* găsiți la capitolul de *Clusterizare*, la rezolvarea problemei 17.b, pagina 874.

²⁸¹Proprietatea aceasta este implicată dacă este satisfăcută o condiție mai „tare“: funcția f să fie dublu derivabilă, adică să existe $f''(x)$, pe tot domeniul de definiție. În acest caz, condiția de convexitate va implica $f''(x) \geq 0$ pentru orice x .

$$f(y) \geq f(x) + f'(x)(y - x) \text{ pentru orice } x \text{ și } y. \quad (147)$$

Remarcați faptul că relația (146) din definiția pentru proprietatea de β -netezire a lui f introduce o margine superioară pentru $f(y)$, pe lângă marginea inferioară din relația (147) dată de proprietatea de convexitate.

Vă cerem să rezolvați următoarele două *cerințe*:

i. Folosind regula de actualizare din algoritmul GD, arătați că din relația precedentă rezultă că

$$f(x^{(k+1)}) \leq f(x^{(k)}) - \eta \left(1 - \frac{\eta\beta}{2}\right) (f'(x^{(k)}))^2.$$

ii. Identificați intervalul în care trebuie să se situeze valorile ratei de învățare η astfel încât să rezulte $f(x^{(k+1)}) < f(x^{(k)})$.²⁸²

d. Presupunem că există un singur punct de minim pentru f , adică un punct x^* astfel încât $f(x^*) < f(x)$, $\forall x \neq x^*$. Arătați că pentru $\eta = \frac{1}{\beta}$ vom avea $f'(x^{(k)}) \rightarrow 0$ atunci când $k \rightarrow +\infty$. Explicați de ce această convergență implică faptul că $x^{(k)} \rightarrow x^*$ atunci când $k \rightarrow +\infty$.

Sugestie: Folosiți *lema descreșterii* (de la punctul c) pentru a arăta că pentru orice $K \geq 1$ are loc inegalitatea următoare: $\sum_{k=1}^K (f'(x^{(k)}))^2 \leq 2\beta(f(x^{(1)}) - f(x^*))$.

Observație: O demonstrație pentru convergența metodei gradientului descentant în cazul unor funcții cu proprietăți mai generale decât cele de mai sus poate fi găsită în cartea *Understanding Machine Learning: From Theory to Algorithms* de Shai Shalev-Schwartz și Shai Ben-David, Cambridge University Press, 2014.²⁸³

165. (Metoda gradientului descendat: aplicare / implementare)

• Caltech, 2012, Abu Mostafa, HW5, pr. 4

Considerăm că o anumită funcție de eroare are expresia $E(u, v) = (ue^v - 2ve^{-u})^2$.

a. Calculați derivatiile parțiale ale acestei funcții în raport cu u și respectiv v , adică $\frac{\partial E}{\partial u}(u, v)$ și $\frac{\partial E}{\partial v}(u, v)$.

b. Pentru a identifica minimul funcției de eroare E , veți aplica metoda gradientului descendat, începând cu punctul $(u_0, v_0) = (1, 1)$. Veți folosi *rata de învățare* $\eta = 0.1$.

Câte iterații se efectuează până când valoarea diferenței $E(u_{t+1}, v_{t+1}) - E(u_t, v_t)$ scade pentru prima dată sub pragul de 10^{-14} ? Aveți grijă ca în programul pe care îl veți implementa să folosiți [variabile în] dublă precizie, pentru a putea obține acuratețea cerută.

²⁸²Așadar, vrem să ne asigurăm că valoarea lui η este aleasă astfel încât, la executarea unei iterații a algoritmului GD, valoarea funcției obiectiv să descrească.

²⁸³<http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning>.

166.

(Alegerea unei funcții de eroare convenabilă pentru învățarea unor concepte geometrice și studierea aplicabilității metodei gradientului descendente în acest context)

- * T. Mitchell, "Machine Learning", 1997, pr. 4.12
CMU, 2011 spring, Roni Rosenfeld, HW4, pr. 3

Considerăm că avem de învățat o clasă de concepte definită de multimea dreptunghiurilor din planul bidimensional care au laturile paralele cu axele sistemului de coordinate.

Fiecare ipoteză este descrisă cu ajutorul a 4 coordinate: llx, lly, urx, ury . Semnificația acestor coordinate este următoarea: perechea (llx, lly) desemnează colțul din stânga-jos, iar (urx, ury) desemnează colțul din dreapta-sus al dreptunghiului de învățat. O instanță (x, y) este etichetată pozitiv de către ipoteza llx, lly, urx, ury dacă și numai dacă punctul (x, y) este situat în interiorul dreptunghiului respectiv.

Ne propunem să stabilim cum anume ar trebui să procedăm pentru a minimiza eroarea la clasificare automată, pentru a „învăța“ astfel de dreptunghiuri.

- a. Definiți o funcție de eroare (E) adecvată pentru această problemă de învățare.
- b. Se poate aplica un algoritm de tip gradient descendente pentru a identifica minimul acestei funcții de eroare? De ce da, sau de ce nu?
- c. Dacă răspunsul la punctul b este *da*, descrieți pe scurt procedura de tip gradient descendente. Dacă răspunsul la punctul b este *nu*, puteți sugera o aproximare a funcției de eroare de la punctul a astfel încât să puteți aplica metoda gradientului descendente pentru funcția de eroare nou-definită?

167.

(Metoda subgradientului stochastic descendente (engl., stochastic subgradient descent, SSGD): deducerea regulii de actualizare a parametrilor în cazul unei regresii liniare particulare)

- • ○ MIT, 2018 spring, Tommi Jaakkola, midterm review, pr. 6

Funcția de cost / pierdere (engl., loss) patratică care a fost definită la problema 88, poate fi generalizată astfel:

$$L(g, y) = \begin{cases} c_1(g - y)^2 & \text{dacă } g > y \\ c_2(g - y)^2 & \text{în cazul contrar,} \end{cases}$$

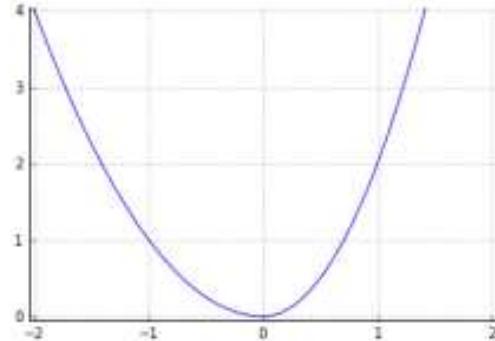
unde c_1 și c_2 sunt constante reale pozitive și, în contextul folosirii acestei funcții pentru regresie (adică, învățare de funcții cu valori reale), g este valoarea *prezisă* pentru funcția care este învățată [și anume, într-un anumit punct x], iar y este valoarea *adevărată* a funcției [în punctul respectiv].

O astfel de definiție este utilă, de exemplu, în situația următoare:

Niște organizatori de concerte vor să creeze *model* care să estimeze / prezică cât de mulți participanți vor avea la un eveniment oarecare, folosind ca trăsături precum sunt data concertului, genul [muzical al] formației invitate, etc.

Ei vor utiliza acest model pentru a decide ce formație să angajeze pentru următorul eveniment, dar ar vrea ca predicțiile pe care le fac să fie „protectoare“ (engl., conservative), în sensul următor: dacă modelul lor sub-estimează numărul de participanți, este mai bine aşa decât să-l supra-estimeze.

a. Cât trebuie să fie valorile lui c_1 și c_2 ca această funcție de cost să fie chiar loss-ul pătratic (engl., squared error)?



b. În figura alăturată am alcătuit un grafic al valorilor funcției de cost L în raport cu diferența $g - y$, dar am uitat care sunt valorile c_1 și c_2 care au fost folosite pentru a realiza acest grafic.

Conform acestui grafic, ce este mai rău:

- să sub-estimezi, ori să supra-estimezi,
dată fiind o aceeași diferență în valoare absolută, $|g - y|$? Justificați riguros.

c. Considerăm un model liniar în care $g(x) = w \cdot x + w_0$. Folosiți metoda subgradientului descedent pentru a deduce regulile de actualizare pentru w și w_0 , cu rata de învățare / pasul (engl., step size) η .²⁸⁴ Scrieți apoi pseudo-codul algoritmului subgradientului descedent stochastic care folosește aceste reguli de actualizare pentru a „învăța“ valorile optime pentru w și w_0 .

168. (O legătură între Perceptronul Rosenblatt și metoda subgradientului)

• ○ CMU, 2016 fall, N. Balcan, M. Gormley, HW4, pr. 3.b

Algoritmul Perceptron este unul dintre primii și cei mai simpli algoritmi de clasificare automată.²⁸⁵ El ia ca input un set de date de antrenament format din instanțele etichetate $\{(x_1, y_1), \dots, (x_n, y_n)\}$, cu $x_i \in \mathbb{R}^d$ și $y_i \in \{-1, +1\}$ pentru $i = 1, \dots, n$. Outputul algoritmului este un set de ponderi $w \in \mathbb{R}^d$. Pseudo-codul acestui algoritm este următorul:²⁸⁶

```

initialize  $w \leftarrow \bar{0}$  not  $(0, \dots, 0) \in \mathbb{R}^d$ 
for  $i = 1, \dots, n$ 
    if  $y_i(w \cdot x_i) \leq 0$  then
         $w \leftarrow w + y_i x_i$ 
    end if
end for

```

²⁸⁴ Sugestie: Dacă vă este greu să gândiți despre w și x ca fiind vectori, vă sugerăm să procedați la început considerând că $[w \text{ și } x]$ sunt valori reale / scalare.

²⁸⁵ Frank Rosenblatt. *The perceptron: A probabilistic model for information storage and organization in the brain*. Psychological Review, 65(6), 386–408, 1958.

²⁸⁶ În acest pseudo-cod nu se folosește termenul liber (w_0), deci separarea se face printr-un hiperplan care trece prin originea sistemului de coordonate. (Alternativ, se poate considera că toate instanțele x_i au pe poziția 1 constanta 1.) Pentru un exemplu de aplicare a algoritmului Perceptron, vedeți problema 16 de la capitolul *Rețele neuronale artificiale*.

Comentariu: Dacă setul de date de antrenament este liniar separabil, algoritmul Perceptron converge, adică outputul său, notat cu $\bar{w} \in \mathbb{R}^d$, satisfacă inegalitățile

$$y_i(\bar{w} \cdot x_i) \leq 0 \text{ pentru } i = 1, \dots, n.$$

Mai mult, se poate arăta că există o margine superioară pentru numărul de greșeli comise de Perceptron (deci, de ajustări ale ponderilor sale) până se ajunge la convergență. (Vedeți pr. 18 de la capitolul *Rețele neuronale artificiale*.)

Acest exercițiu vă va ajuta să identificați o legătură între algoritmul Perceptron și algoritmul subgradientului descendente.²⁸⁷

Considerând funcția $F(w) = \frac{1}{n} \sum_{i=1}^n F_i(w)$, *algoritmul subgradientului descendente (ciclic)* este următorul:

```

initialize the variable  $w \in \mathbb{R}^d$ 
do until convergence:
    for  $i = 1 \dots n$ 
        find a subgradient  $v_i(w)$  of  $F_i(w)$ 
         $w \leftarrow w - \eta v_i(w)$ , where  $\eta \in \mathbb{R}^+$  is the learning rate
    end for
end do

```

Arătați că Perceptronul este un caz particular de algoritm de tip subgradient descendente.

169.

(Metoda lui Newton: o proprietate interesantă
în cazul funcțiilor pătratice)

*prelucrare de Liviu Ciortuz, după
□ • ○ Stanford, 2009 fall, Andrew Ng, practice midterm, pr. 6.d*

Fie o funcție $f : \mathbb{R}^d \rightarrow \mathbb{R}$ definită prin expresia $f(x) = \frac{1}{2}x^\top Ax + bx + c$, unde A este o matrice simetrică și pozitiv definită.

a. Arătați că f este funcție convexă.

(*Sugestie:* Puteti folosi fie definiția fie proprietățile formulate la problema 78.)

b. Demonstrați că atunci când se folosește metoda lui Newton pentru a afla minimul funcției f , este suficient să se execute o singură iterație. Veți considera că metoda lui Newton face inițializarea cu vectorul 0 (din \mathbb{R}^d).

²⁸⁷Vă readucem aminte că, dată fiind o funcție $F : \mathbb{R}^d \rightarrow \mathbb{R}$ care poate fi nederivabilă, vectorul $v(x) \in \mathbb{R}^d$ este un *subgradient* al lui $F(x)$ dacă și numai dacă pentru orice x' are loc inegalitatea:

$$F(x') \geq F(x) + v(x) \cdot (x' - x).$$

170.

(Reparametrizarea liniară nu afectează metoda [de optimizare a] lui Newton, însă afectează metoda gradientului)

Stanfurd, 2014 fall, Andrew Ng, HW1, pr. 5

Presupunem că folosim un algoritm de optimizare iterativă (de exemplu, metoda lui Newton sau metoda gradientului descendente) pentru a calcula [punctul în care se atinge] minimul unei funcții derivabile în mod continuu $f(x)$, cu $x \in \mathbb{R}^n$. Presupunem că inițializăm algoritmul cu vectorul $x^{(0)} = \vec{0}$ din \mathbb{R}^n . La execuție, algoritmul va produce pentru un input oarecare $x \in \mathbb{R}^n$ (fixat) câte o valoare din \mathbb{R}^n la fiecare iterație: $x^{(1)}, x^{(2)}, \dots$.

Considerăm acum că ni se dă o matrice oarecare pătratică nesingulară (adică, inversabilă) $A \in \mathbb{R}^{n \times n}$ și, de asemenea, că definim o nouă funcție $g(z) = f(Az)$. Presupunem că folosim același algoritm de optimizare iterativă pentru a calcula optimul funcției g , cu inițializarea $z^{(0)} = \vec{0}$ din \mathbb{R}^n .

Dacă valorile $z^{(1)}, z^{(2)}, \dots$ produse folosind această metodă satisfac [în mod necesar] condițiile $z^{(i)} = A^{-1}x^{(i)}$ pentru orice i , vom spune că acest algoritm de optimizare este *invariant la reparametrizare liniară*.

a. Arătați că *metoda lui Newton* (aplicată pentru găsirea minimului unei funcții) este invariantă la reparametrizare liniară.

Sugestie: Remarcați faptul că deoarece $z^{(0)} = \vec{0} = A^{-1}x^{(0)}$, este suficient să demonstrezi următoarea *implicație*:

dacă la aplicarea metodei lui Newton asupra lui $f(x)$, din $x^{(i)}$ se obține $x^{(i+1)}$,
atunci când metoda lui Newton se va aplica asupra lui $g(z) = f(Az)$,
din $z^{(i)} = A^{-1}x^{(i)}$ se va obține $z^{(i+1)} = A^{-1}x^{(i+1)}$.

b. Este oare metoda *gradientului descendente* invariantă la reparametrizare liniară? Justificați în mod riguros răspunsul dumneavoastră.

171.

(Metoda multiplicatorilor lui Lagrange: aplicare într-un caz simplu)

prelucrare de Liviu Ciortuz, după CMU, 2021 fall, Aarti Singh, Recitation 7, pr. 4

Folosiți condițiile Karush-Kuhn-Tucker pentru a rezolva următoarea problemă de optimizare cu restricții:

$$\begin{aligned} & \min_{x,y} x^2 + y^2 \\ \text{a. i. } & x + y - 1 = 0. \end{aligned}$$

Cerințe preliminare rezolvării:

Veți face reprezentarea grafică corespunzătoare [rezolvării] acestei probleme. Folosind această reprezentare veți indica soluția problemei de optimizare date înainte de rezolvarea ei folosind metoda lui Lagrange. (Faceți justificarea în mod riguros.)

172.

(Metoda lui Lagrange: rezolvarea formei duale pentru o problemă de optimizare convexă simplă)

*prelucrare de Liviu Ciortuz, după**□ • ○ CMU, 2013 fall, A. Smola, G. Gordon,
midterm practice questions, pr. 2.c*

Fie următoarea problemă de optimizare cu restricții, în formă primală:

$$\min_x (x^2 + 1)$$

a. î. $(x - 2)(x - 4) \leq 0$.

Obțineți forma duală a acestei probleme de optimizare, apoi rezolvați această problemă duală prin 3 metode: *i.* direct; *ii.* rezolvând sistemul de condiții Karush-Kuhn-Tucker și folosind partea a doua din teorema Karush-Kuhn-Tucker; *iii.* rezolvând mai întâi problema primală și apoi folosind relația de legătură dintre soluțiile celor două probleme (primală și respectiv duală) obținută din condiția de staționaritate / optimalitate din sistemul de condiții Karush-Kuhn-Tucker.²⁸⁸

173.

(Metoda dualității Lagrange: aplicare în \mathbb{R} , rezolvând mai întâi problema duală și folosind apoi relația de corespondență cu soluția problemei primale)*□ • * Liviu Ciortuz, 2019, după
CMU, Aarti Singh, 2010 fall, SVM – Lecture notes*

Se dă următoarea problemă de optimizare convexă, scrisă în forma primală:

$$\min_x x^2$$

a. î. $x \geq b$,

unde b este o constantă reală.a. Rezolvați această problemă direct, în funcție de valorile lui b . (Faceți în prealabil reprezentarea grafică.)b. Este oare îndeplinită *condiția lui Slater* pentru această problemă de optimizare convexă? În cazul afirmativ, precizați care sunt implicațiile satisfacerii acestei condiții în contextul problemei date.c. Scrieți lagrangeanul generalizat $L_P(x, \alpha)$, unde α este multiplicatorul Lagrange corespunzător restricției din problema de optimizare convexă dată.Calculați x^* , rădăcina ecuației $\frac{\partial}{\partial x} L_P(x, \alpha) = 0$, în funcție de α .²⁸⁹Aflați expresia lagrangeanului dual $L_D(\alpha)$, înlocuind în expresia lagrangeanului generalizat, $L_P(x, \alpha)$, argumentul x cu expresia pe care tocmai ați obținut-o pentru x^* , soluția unică a ecuației $\frac{\partial}{\partial x} L_P(x, \alpha) = 0$.d. Scrieți forma duală a problemei de optimizare convexă din enunț.²⁹⁰ Rezolvați această problemă și apoi calculați soluția x^* a problemei date în enunț, folosind relația pe care ați obținut-o la punctul c.²⁸⁸Pentru varianta *iii* este necesar ca în prealabil să verificați faptul că este îndeplinită condiția lui Slater.²⁸⁹În general, rădăcinile ecuației $\frac{\partial}{\partial x} L_P(x, \alpha) = 0$ se mai numesc *punctele staționare* ale lui L_P .²⁹⁰Vedeți problema 82.b.

174. (Metoda multiplicatorilor lui Lagrange: un exemplu de formalizare a unei probleme de optimizare convexă pornind de la descrierea ei în limbaj natural, urmată de rezolvarea ei)

*prelucrare de Liviu Ciortuz, după
□ • CMU, 2021 fall, Aarti Singh, recitation 7, pr. 1²⁹¹*

Să zicem că o companie fabrică două produse (notează X și Y), folosind două mașini (notează A și B). Realizarea fiecărui produs X necesită 50 de minute (ca timp de procesare) pe mașina A și 30 de minute pe mașina B . Realizarea fiecărui produs Y necesită 24 de minute pe mașina A și 33 de minute pe mașina B .

Timpul total [de procesare] *disponibil* este estimat la 40 de ore în cazul mașinii A , și la 35 ore în cazul mașinii B .

Pentru săptămâna curentă, cererea [prognozată] este de cel puțin 45 de unități din produsul X și de cel puțin 5 unități din produsul Y .

Politica / *obiectivul* companiei este să maximizeze numărul total de unități produse (X și Y) la sfârșitul săptămânii.

a. Formulați ca o *problemă de optimizare convexă* problema care constă în a decide cât de multe unități din fiecare produs trebuie să producă această companie în săptămâna curentă pentru a atinge *obiectivul* formulat mai sus.

Atenție! Problema de optimizare convexă având — prin definiție — funcția obiectiv supusă operatorului \min , va trebui să aveți grijă cum definiți funcția obiectiv în cazul specific prezentat mai sus.

b. Stabiliți dacă problema de optimizare convexă pe care ați formulat-o la punctul a satisfacă condiția lui Slater. În cazul afirmativ, care sunt consecințele?

c. Scrieți sistemul de condiții Karush-Kuhn-Tucker pentru problema de optimizare convexă pe care ați formulat-o la punctul a .

d. Rezolvați sistemul de condiții Karush-Kuhn-Tucker de la punctul c , iar la final indicați soluția problemei de optimizare convexă de la punctul a .

Atenție! Veți ignora faptul că numărul de unități care trebuie realizate din fiecare dintre produsele X și Y trebuie să fie numere naturale; le veți considera numere reale pozitive. (La final veți putea aplica operatorul *parte întreagă inferioară* la soluțiile obținute.)

²⁹¹From <http://people.brunel.ac.uk/mastjjb/jeb/or/morelp.html>.

175.

(Metoda dualității Lagrange:
un exemplu de problemă de optimizare convexă
pentru care condițiile Karush-Kuhn-Tucker nu sunt satisfăcute)

· Sebastian Ciobanu, 2018²⁹²

Stim că atunci când o problemă de optimizare convexă cu restricții satisface condiția lui Slater,²⁹³ urmează că forma primală și forma duală a problemei de optimizare date au același optim (adică, $p^* = d^*$) și, conform teoremei Karush-Kuhn-Tucker,²⁹⁴ soluțiile acestor două forme ale problemei satisfac condițiile Karush-Kuhn-Tucker.

Obiectivul acestui exercițiu este să vă arate că în cazul nesatisfacerii condiției lui Slater, este posibil ca sistemul de condiții Karush-Kuhn-Tucker să nu fie satisfiabil (adică să nu aibă soluție), chiar dacă problema de optimizare convexă dată este satisfiabilă.

Fie următoarea problemă de optimizare cu restricții:

$$\begin{aligned} \min_x \quad & x \\ \text{a. î.} \quad & x^2 \leq 0 \end{aligned}$$

- a. Indicați soluția problemei de optimizare din enunț (fără a face calcule).
- b. Stabiliți dacă problema de optimizare este convexă.
- c. Stabiliți dacă este satisfăcută condiția lui Slater.
- d. Rezolvați sistemul reprezentat de condițiile Karush-Kuhn-Tucker.
- e. Orice soluție a problemei de optimizare care a fost dată în enunț satisface condițiile Karush-Kuhn-Tucker?

176.

(O [altă] variantă a algoritmului Perceptron,
pentru care relația de actualizare a ponderilor se obține rezolvând o problemă de optimizare convexă cu restricții)

* University of Helsinki, 2014 spring, Jyrki Kivinen, HW5, pr. 3

Asemănător cu problema 87, aici vom lucra asupra unui algoritm în care vectorul w_{t+1} este obținut prin rezolvarea problemei de optimizare următoare:

$$\begin{aligned} \min_w \quad & d_{\text{re}}(w, w_t) \\ \text{a. î.} \quad & y_t w \cdot x_t \geq 0 \\ & w_i \geq 0 \text{ pentru } i = 1, \dots, d \\ & \sum_{i=1}^d w_i = 1, \end{aligned}$$

unde $w \stackrel{\text{not.}}{=} (w_1, \dots, w_d)$, iar $d_{\text{re}}(u, v)$ desemnează entropia relativă²⁹⁵

²⁹²Surse: https://en.wikipedia.org/wiki/Karush%20%80%93Kuhn%20%80%93Tucker_conditions și https://math.stackexchange.com/questions/1346767/do-we-really-need-the-constraint-qualification?utm_medium=organic&utm_source=google_rich_qa&utm_campaign=google_rich_qa

²⁹³Vedeți Observația importantă de la pagina 186.

²⁹⁴Vedeți Comentariul de la problema 85, pag. 186.

²⁹⁵Pentru proprietăți elementare ale entropiei relative (care este cunoscută și sub numele de divergența Kullback-Leibler), vedeți problema 63.

$$d_{\text{re}}(u, v) = \sum_{i=1}^d u_i \ln \frac{u_i}{v_i}.$$

Demonstrați că atunci când există o soluție pentru această problemă, rezultă cu necesitate că există un număr real $\beta_t \geq 1$ astfel încât noul vector de ponderi satisface relația

$$w_{t+1,i} = \frac{1}{Z} w_{t,i} \beta_t^{y_t x_{t,i}}$$

unde

$$Z = \sum_{j=1}^d w_{t,j} \beta_t^{y_t x_{t,j}}.$$

Nu trebuie să găsiți ca atare valoarea lui β_t (fiindcă nu există o formulă analitică (engl., closed-form solution)); trebuie doar să arătați că soluția w_{t+1} are forma care a fost indicată.

Observație: Deși acest algoritm nu este [probabil] foarte practic, astfel de probleme de optimizare au condus la crearea algoritmului Winnow, precum și a altor algoritmi *multiplicativi* de învățare automată *online*.

177.

(Teorema de reprezentare generalizată)

prelucrare de Liviu Ciortuz după
□ • Stanford, John Duchi, Supplemental lecture notes

O generalizare a *teoremei de reprezentare* care a fost demonstrată la problema 88.a a fost publicată de Bernhard Schölkopf, Ralf Herbrich și Alex Smola în anul 2001.²⁹⁶ Generalizarea se referă — pe lângă faptul că nu vom mai cere ca funcția de cost $L(z, y)$ să fie derivabilă în raport cu argumentul z , aşa cum s-a considerat la rezolvarea problemei 88 — la faptul că termenul de regularizare $\frac{\lambda}{2} \|w\|^2$ din expresia (127) se înlocuiește cu termenul mai general $\rho(\|w\|)$, unde ρ este o funcție nedescrescătoare.²⁹⁷

Similar cu modul în care am procedat la problema 88, vom considera setul de date de antrenament $(x_i, y_i)_{i=1,\dots,n}$, cu $x_i \in \mathbb{R}^d$ și $y_i \in \mathbb{R}$. Acum vom desemna *costul* (sau, *riscul empiric*) *regularizat* prin funcția $J_\rho(w) \stackrel{\text{not.}}{=} \frac{1}{n} \sum_{i=1}^n L(w \cdot x_i, y_i) + \rho(\|w\|)$.

Concluzia teoremei de reprezentare generalizată este formulată astfel:

$$\begin{aligned} \exists \alpha \in \mathbb{R}^n \text{ a. i. pentru } w^{(\alpha)} \stackrel{\text{not.}}{=} \sum_{i=1}^n \alpha_i x_i \in \mathbb{R}^d \text{ și } \forall w \in \mathbb{R}^d \text{ are loc inegalitatea} \\ J_\rho(w^{(\alpha)}) \leq J_\rho(w). \end{aligned}$$

²⁹⁶Vedeți articolul *A Generalized Representer Theorem* din volumul Computational Learning Theory, Lecture Notes in Computer Science, 2111: 416–426, 2001.

²⁹⁷Evident, pentru cazul particular $\rho(\|w\|) = \frac{\lambda}{2} \|w\|^2$ cu $\lambda > 0$, se obține regularizarea de normă L_2 . Într-un alt caz particular remarcabil, ρ poate fi considerată chiar funcția constantă 0, deci *teorema de reprezentare generalizată* acoperă și cazul când nu se face deloc regularizare.

Observație: Importanța (foarte mare!) a acestor *teoreme de reprezentare* constă în faptul că datorită lor rezolvarea problemelor de învățare automată constând în minimizarea unei funcții de tip $J_\rho(w)$ se poate reduce la căutarea unor tupluri $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$,²⁹⁸ iar aceasta se poate realiza aplicând algoritmi clasici pentru minimizarea funcțiilor.²⁹⁹

Demonstrați teorema de reprezentare generalizată.

²⁹⁸ Altfel spus, căutarea soluțiilor se poate limita la *spațiul liniar generat* de instanțele de antrenament x_i .

²⁹⁹ Vedeți https://en.wikipedia.org/wiki/Representer_theorem, accesat la data de 30.05.2018.



© M. Romanică

1 Metode de regresie

Sumar

Noțiuni preliminare

- estimarea parametrilor unor distribuții probabiliste uzuale (în special distribuția Bernoulli, distribuția gaussiană, distribuția Laplace); vedeți secțiunea corespunzătoare de la cap. *Fundamente*;
- elemente de calcul vectorial (în particular, produsul scalar) și de calcul matriceal: ex. 36 de la cap. *Fundamente*; norma L_2 (euclidiană) și norma L_1 : ex. 3, ex. 25, ex. 11; calculul derivatelor parțiale [de ordinul întâi și al doilea]: ex. 7; reguli de derivare cu argumente vectoriale: ex. 1; calculul subderivatelor / subgradienților și subdiferențialelor: ex. 81 de la cap. *Fundamente*;
- metode de optimizare (în speță pentru aflarea maximului / minimului unei funcții reale, derivabile): metoda analitică; metoda gradientului și metoda lui Newton — exemplificare: ex. 80 de la cap. *Fundamente*; metoda subgradientului: ex. 168 de la capitolul de *Fundamente*; metoda descreșterii (resp., a creșterii) pe coordonate (engl., coordinate descent / ascent).

Regresia liniară

- prezentarea generală a metodei regresiei liniare:³⁰⁰
 - MLE și corespondența cu estimarea în sens LSE (least squared errors): ex. 3.A.abc; rezolvare folosind matricea de design și *optimizare analitică*: ex. 3.A.d; particularizare pentru cazul unidimensional: ex. 1.ab, ex. 19; exemplificare pentru cazul unidimensional (ex. 2, ex. 20) și pentru cazul bidimensional (ex. 22.a, ex. 31);
 - (P1) *scalarea atributelor* nu schimbă predicțiile obținute (pentru instanțele de test) cu ajutorul *formulelor analitice*: ex. 4, 39.a;
 - (P2) adăugarea de noi trăsături / atrbute nu mărește suma pătratelor erorilor: ex. 24;
 - o proprietate surprinzătoare a regresiei liniare: adăugarea câtorva „observații“ suplimentare poate conduce la modificarea radicală a valorilor optime ale parametrilor de regresie: CMU, 2014 fall, Z. Bar-Joseph, W. Cohen, HW2, pr. 4;
 - [rezolvarea problemei de] regresie liniară folosind *metoda lui Newton*: ex. 7;
 - MAP și corespondența cu *regularizarea* de normă L_2 (regresia *ridge*): ex. 3.C; particularizare pentru cazul unidimensional: ex. 1.c;
 - (P3) efectul de diminuare a ponderilor (engl., weight decay) în cazul regularizării de normă L_2 (respectiv L_1) a regresiei liniare, în comparație cu cazul neregularizat: ex 25.b (respectiv ex. 25.a);

³⁰⁰În mod implicit, în această secțiune se va considera că termenul-zgomot este modelat cu distribuția gaussiană (dacă nu se specifică altfel, în mod explicit).

- bias-ul și [co]varianța estimatorului regresiei liniare; bias-ul regresiei *ridge*: ex. 5;
- regresia polinomială [LC: mai general: folosirea așa-numitelor *funcții de bază*]: ex. 3.B;
exemplificare pentru cazul bidimensional: CMU, 2015 spring, T. Mitchell, N. Balcan, HW4, pr. 1;
- cazul regresiei liniare cu termen de regularizare L_2 (regresia *ridge*):
deducerea regulilor de actualizare pentru *metoda gradientului descendente*: varianta “batch” / “steepest descent”: ex. 6.a;
și varianta stochastică / secvențială / “online”: ex. 6.b; exemplu de aplicare: ex. 23;
- cazul regresiei liniare cu termen de regularizare de normă L_1 :
rezolvare combinând *metoda descreșterii pe coordonate* cu *metoda optimizării analitice*: ex. 11;
rezolvare combinând *metoda descreșterii pe coordonate* cu *metoda subgradientului* (aplicare la selecția de trăsături): ex. 27;
- regresia liniară în cazul zgromotului modelat cu distribuția *Laplace* (în locul zgromotului gaussian): ex. 8.B;
exemplificare pentru cazul bidimensional: ex. 22.c;
rezolvare în cazul unidimensional [chiar particularizat] cu ajutorul derivatei, acolo unde aceasta există: ex. 28;
- regresia liniară și *overfitting-ul*: ex. 12;
- regresie liniară folosită pentru *clasificare*: exemplificare: ex. 31;
- cazul *multivaluat* al regresiei liniare, reducerea la cazul uninomial: ex. 30;
- regresia liniară cu regularizare L_2 (regresia *ridge*), *kernel-izarea* ecuațiilor „normale“: ex. 9; aplicare: ex. 26;
(P4) folosind nucleu RBF, eroarea la antrenare devine 0 atunci când parametrul de regularizare λ tinde la 0: ex. 10;
- *regresia liniară ponderată*: ex. 8.A;
particularizare pentru cazul unidimensional [neparametric]: CMU, 2010 fall, Aarti Singh, midterm, pr. 4
particularizare / exemplificare pentru cazul bidimensional: ex. 22.b;
cazul multivaluat, cu regularizare L_2 : Stanford, 2015 fall, Andrew Ng, midterm, pr. 2;
○ proprietate a regresiei liniare local-ponderate [demonstrată în cazul unidimensional]: „netezirea“ liniară: ex. 29.

Regresia logistică

- prezentare generală,
- (•) calculul funcției de log-verosimilitate, estimarea parametrilor în sens MLE, folosind *metoda gradientului* (i.e., deducerea regulilor de actualizare a parametrilor): ex. 13, ex. 34, 39.b;
particularizare pentru cazul datelor din \mathbb{R}^2 : ex. 33 (inclusiv *regularizare* L_1 / estimarea parametrilor în sens MAP, folosind o distribuție a priori Laplace);
- (P0) *granița de decizie* pentru regresia logistică: ex. 33.d;

- (P1) funcția de log-verosimilitate în cazul regresiei logistice este concavă (deci are un maxim global), fiindcă matricea hessiană este pozitiv definită: ex. 14;
Observație: Demonstrația furnizează tot ce este necesar pentru obținerea [ulterioră] relației de actualizare a parametrilor la aplicarea *metodei lui Newton* în cazul regresiei logistice;
- (P2) analiza efectului duplicării atributelor: ex. 35;
- (P3) efectul de diminuare a ponderilor (engl., weight decay) în cazul regularizării de normă L_2 a regresiei logistice — adică la estimarea parametrilor în sens MAP, folosind ca distribuție a priori distribuția gaussiană multidimensională sferică —, în comparație cu cazul estimării parametrilor în sensul MLE: ex. 15;
- *Variante / extensii ale regresiei logistice:*
 - regresia logistică local-ponderată, cu regularizare L_2 : calcularea vectorului gradient și a matricei hessiene (necesare pentru aplicarea metodei lui Newton în acest caz): ex. 16;
 - regresia logistică kernel-izată: adaptarea metodei gradientului: ex. 17;
 - regresia logistică n -ară (așa-numita regresie *softmax*): calculul funcției de log-verosimilitate, cu regularizare L_2 , deducerea regulilor de actualizare a ponderilor, folosind metoda gradientului: ex. 18;
 - (P4) echivalența cu un anumit tip de mixtură de distribuții gaussiane multidimensionale: ex. 37;
- (P5) o [interesantă] proprietate comună pentru regresia liniară și regresia logistică: ex. 36;
- întrebări (cu răspuns A/F) cu privire la aplicarea *metodei lui Newton* comparativ cu *metoda gradientului* (în contextul rezolvării problemelor de regresie liniară și / sau regresie logistică): ex. 39.c;
- comparații între regresia logistică și alți clasificatori (Bayes Naiv, ID3): ex. 33.c, ex. 38.ab;

Modele liniare generalizate (GLM)

- condiții suficiente pentru *concavitatea* funcției de log-verosimilitate: ex. 42;
- particularizare pentru cazul distribuției geometrice: ex. 40;
- particularizare pentru cazul distribuției gaussiane unidimensionale: ex. 41.

1.1 Metode de regresie — Probleme rezolvate

1.1.1 Regresia liniară

1. (Estimare [în sens MLE și MAP] pentru parametrul unui model de regresie liniară unidimensională fără termen liber, în prezența unei componente-zgomot modelate de o distribuție gaussiană)
- • CMU, 2012 fall, T. Mitchell, Z. Bar-Joseph, midterm, pr. 3
 - CMU, 2011 spring, Tom Mitchell, midterm, pr. 4

Introducere: Obiectivul acestei probleme este acela de a face cunoștință cu metoda / problema de regresie liniară într-o dintre cele mai simple variante ale sale: cazul unidimensional. Aceasta înseamnă că variabila de ieșire (Y) depinde de o singură variabilă de intrare (X),³⁰¹ dar și de o componentă „zgomot“ (ε). Modelarea acestei componente va fi făcută în manieră probabilistă, și anume (aici) cu ajutorul unei distribuții gaussiene.

Fie variabilele aleatoare X și Y . Presupunem că valorile variabilei Y se generează în funcție de valorile variabilei X conform relației:

$$Y = aX + \varepsilon,$$

unde fiecare ε reprezintă valoarea unei variabile aleatoare (care este independentă de variabilele precedente), numită „zgomot“, și care urmează o distribuție gaussiană de medie 0 și deviație standard $\sigma > 0$. Acest fapt se notează îndeobște astfel: $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Acesta este un model probabilist (de tip „regresie liniară“, cu un singur atribut, X), în care a este singurul parametru (o pondere; engl., weight).

Probabilitatea condiționată a lui Y în raport cu X urmează distribuția $\mathcal{N}(aX, \sigma^2)$, deci poate fi descrisă ca

$$p(Y|X, a) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(Y - aX)^2\right).$$

Următoarele întrebări se referă la acest model de regresie liniară.

Estimare în sens MLE:

- a. Presupunem că avem un set de date de antrenament constând din n perechi (X_i, Y_i) , cu $i = 1, \dots, n$, și că varianța σ este cunoscută. Care dintre următoarele expresii reprezintă în mod corect problema de optimizare pentru estimarea lui a în sensul verosimilității maxime? Răspundeți cu *da* sau *nu* la fiecare dintre

³⁰¹Evident, dependența dintre Y și X va fi considerată de tip liniar.

ele. Nu neapărat una singură dintre aceste expresii trebuie să fie însotită de răspunsul *da*.

- i. $\arg \max_a \sum_i \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\right)$
- ii. $\arg \max_a \prod_i \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\right)$
- iii. $\arg \max_a \sum_i \exp\left(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\right)$
- iv. $\arg \max_a \prod_i \exp\left(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\right)$
- v. $\arg \max_a \sum_i (Y_i - aX_i)^2$
- vi. $\operatorname{argmin}_a \sum_i (Y_i - aX_i)^2$

b. Obțineți estimarea de verosimilitate maximă pentru parametrul a în funcție de exemplele de antrenament X_i și Y_i ($i = 1, \dots, n$). Vă recomandăm / cerem să porniți de la forma cea mai simplă a problemei de optimizare pe care ati identificat-o la punctul precedent.

Estimare în sens MAP:

Să presupunem că $a \sim \mathcal{N}(0, \lambda^2)$, deci

$$p(a|\lambda) = \frac{1}{\sqrt{2\pi}\lambda} \exp\left(-\frac{1}{2\lambda^2}a^2\right) \text{ cu } \lambda > 0.$$

Probabilitatea (de fapt, p.d.f.) a posteriori a parametrului a este

$$\begin{aligned} p(a|Y_1, \dots, Y_n, X_1, \dots, X_n, \lambda) &\stackrel{T.B.}{=} \frac{p(Y_1, \dots, Y_n|X_1, \dots, X_n, a) p(X_1, \dots, X_n, a|\lambda)}{p(Y_1, \dots, Y_n|X_1, \dots, X_n, \lambda)} \\ &\stackrel{F.P.T.}{=} \frac{p(Y_1, \dots, Y_n|X_1, \dots, X_n, a) p(a|\lambda)}{\int_{a'} p(Y_1, \dots, Y_n|X_1, \dots, X_n, a') p(a'|\lambda) da'}. \end{aligned}$$

Putem să ignorăm numitorul acestei fracții atunci când facem estimare în sens MAP, fiindcă el nu depinde de parametrul a .

c. Presupunem că $\sigma = 1$ și că parametrul a priori λ a fost fixat. Găsiți estimarea de probabilitate maximă a posteriori (engl., maximum a posteriori probability, MAP) pentru parametrul a :

$$\arg \max_a [\ln p(Y_1, \dots, Y_n|X_1, \dots, X_n, a) + \ln p(a|\lambda)].$$

Soluția pe care o veți da trebuie să fie în funcție de X_i , Y_i ($i = 1, \dots, n$) și λ .

Răspuns:

a. Funcția de *verosimilitate condițională* a datelor de ieșire $Y = (Y_1, \dots, Y_n)$ în raport cu datele de intrare $X = (X_1, \dots, X_n)$ și cu parametrul a este:

$$\begin{aligned} L(a) &\stackrel{\text{def.}}{=} p(Y_1, \dots, Y_n|X_1, \dots, X_n, a) \\ &\stackrel{i.i.d.}{=} \prod_{i=1}^n p(Y_i|X_i, a) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\right). \end{aligned}$$

Prin urmare,

$$a_{MLE} \stackrel{def.}{=} \arg \max_a L(a) = \arg \max_a \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp \left(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2 \right).$$

Așadar, soluția *ii* propusă în enunț este corectă.

Mai departe, se poate scrie imediat o formă mai simplă pentru expresia funcției de verosimilitate:

$$\begin{aligned} L(a) &= \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \prod_{i=1}^n \exp \left(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2 \right) \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp \left(-\sum_{i=1}^n \frac{1}{2\sigma^2}(Y_i - aX_i)^2 \right). \end{aligned}$$

Factorul $\frac{1}{(\sqrt{2\pi}\sigma)^n}$ este constant în raport cu parametrul a și este de asemenea pozitiv, pentru că deviația standard σ este prin definiție un număr pozitiv. În consecință,

$$a_{MLE} \stackrel{def.}{=} \arg \max_a L(a) = \arg \max_a \prod_{i=1}^n \exp \left(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2 \right),$$

ceea ce înseamnă că și soluția *iv* propusă în enunț este corectă.

Funcția de log-verosimilitate a datelor are expresia:

$$\begin{aligned} \ell(a) &\stackrel{def.}{=} \ln L(a) = \ln \left(\frac{1}{(\sqrt{2\pi}\sigma)^n} \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - aX_i)^2 \right) \right) \\ &= -n \ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - aX_i)^2. \end{aligned}$$

Datorită faptului că funcția \ln (ca și orice funcție \log_b cu $b > 1$) este strict crescătoare, rezultă că la compunere de funcții conservă / păstrează monotonia. În cazul nostru, această proprietate implică

$$\begin{aligned} a_{MLE} &\stackrel{def.}{=} \arg \max_a L(a) = \arg \max_a \ell(a) \\ &= \arg \max_a \left(-n \ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - aX_i)^2 \right). \end{aligned}$$

Termenul $-n \ln(\sqrt{2\pi}\sigma)$ fiind constant în raport cu a , rezultă că

$$\begin{aligned} a_{MLE} &= \arg \max_a \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - aX_i)^2 \right) = \arg \max_a \left(-\sum_{i=1}^n (Y_i - aX_i)^2 \right) \\ &= \arg \min_a \sum_{i=1}^n (Y_i - aX_i)^2. \end{aligned}$$

Penultima egalitate de mai sus are loc fiindcă factorul $\frac{1}{2\sigma^2}$ este pozitiv și constant în raport cu a . Ultima egalitate se explică prin faptul că la schimbarea

de semn (adică, prin renunțarea la semnul – din fața simbolului \sum), operatorul $\arg \max$ se transformă în $\arg \min$. Așadar, soluția v_i din enunț este, de asemenea, corectă. Celelalte soluții propuse în enunț (i , iii și v) sunt incorecte.

b. Vom face calculele pornind de la expresia v_i de la punctul precedent.

$$a_{MLE} = \arg \min_a \sum_{i=1}^n (Y_i - aX_i)^2 = \arg \min_a \left(a^2 \sum_{i=1}^n X_i^2 - 2a \sum_{i=1}^n X_i Y_i + \sum_{i=1}^n Y_i^2 \right).$$

Putem observa că expresia $a^2 \sum_{i=1}^n X_i^2 - 2a \sum_{i=1}^n X_i Y_i + \sum_{i=1}^n Y_i^2$ reprezintă o funcție de gradul 2 în raport cu parametrul / variabila a , care are — în ipoteza că există măcar un $X_i \neq 0$ — coeficientul dominant pozitiv, deci va avea un (singur) minim. Abscisa acestui punct de minim este chiar a_{MLE} . Prin urmare,³⁰²

$$a_{MLE} = -\frac{-2 \sum_{i=1}^n X_i Y_i}{2 \sum_{i=1}^n X_i^2} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}.$$

c. Tinând cont și de faptul că

$$\arg \max_a \prod_i \frac{1}{\sqrt{2\pi}\sigma} \exp \left(-\frac{1}{2\sigma^2} (Y_i - aX_i)^2 \right) = \arg \max_a \left(-\frac{1}{2} \sum_i (Y_i - aX_i)^2 \right),$$

problema de optimizare în sens MAP devine:

$$\begin{aligned} & \arg \max_a \left(-\frac{1}{2} \sum_{i=1}^n (Y_i - aX_i)^2 + \ln \frac{1}{\sqrt{2\pi}\lambda} - \frac{1}{2\lambda^2} a^2 \right) \\ &= \arg \max_a \left(-\frac{1}{2} \sum_{i=1}^n (Y_i - aX_i)^2 - \frac{1}{2\lambda^2} a^2 \right) = \arg \min_a \left(\sum_{i=1}^n (Y_i - aX_i)^2 + \frac{a^2}{\lambda^2} \right) \quad (148) \\ &= \arg \min_a \left(a^2 \left(\sum_{i=1}^n X_i^2 + \frac{1}{\lambda^2} \right) - 2a \sum_{i=1}^n X_i Y_i + \sum_{i=1}^n Y_i^2 \right). \end{aligned}$$

Ca și la punctul b, observăm că funcția de gradul al doilea căreia î se aplică operatorul $\arg \max$ are coeficientul dominant pozitiv, deci va avea un (singur) minim. Prin urmare, obținem pentru a_{MAP} următorul rezultat:

$$a_{MAP} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2 + \frac{1}{\lambda^2}}.$$

2.

(Regresie liniară [cu anumite restricții impuse]: exemplu de aplicare)

CMU, 2005 fall, T. Mitchell, A. Moore, midterm exam, pr. 3

Presupunem că vrem să „antrenăm“ (engl., fit) un model de regresie liniară pe următoarele date:

x	-1	0	2
y	1	-1	1

³⁰²Se știe că pentru funcția de gradul al doilea $ax^2 + bx + c$, abscisa punctului de optim este $-\frac{b}{2a}$.

- a. Antrenați modelul $Y_i = \beta_1 X_i + \varepsilon_i$ (reprezentând o regresie liniară unidimensională, dar fără termenul constant (β_0)); găsiți valoarea optimă pentru parametrul β_1 .³⁰³
- b. Antrenați modelul $Y_i = \beta_0 + \varepsilon_i$ (reprezentând o regresie liniară „degenerată“, fiindcă nu avem nicio variabilă de intrare); găsiți valoarea optimă pentru parametrul β_0 .

Răspuns:

a. Adaptând definiția regresiei liniare pentru acest caz particular, vom scrie: $Y_i|X_i \sim \mathcal{N}(\beta_1 X_i, \sigma^2)$ pentru $i = 1, 2, 3$. Apoi, pornind de la expresia funcției de verosimilitate a datelor — în scrierea căreia vom folosi notația $Y = (Y_1, Y_2, Y_3)$ și $X = (X_1, X_2, X_3)$ —, vom putea calcula $\hat{\beta}_1$, valoarea optimă a lui β_1 , astfel:

$$\begin{aligned}\hat{\beta}_1 &= \underset{\beta_1}{\operatorname{argmax}} P(Y|X; \beta_1) \stackrel{i.i.d.}{=} \underset{\beta_1}{\operatorname{argmax}} \prod_{i=1}^3 P(Y_i|X_i; \beta_1) \\ &= \underset{\beta_1}{\operatorname{argmax}} \prod_{i=1}^3 \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(Y_i - \beta_1 X_i)^2}{2\sigma^2}} = \underset{\beta_1}{\operatorname{argmax}} \prod_{i=1}^3 e^{-\frac{(Y_i - \beta_1 X_i)^2}{2\sigma^2}} \\ &= \underset{\beta_1}{\operatorname{argmax}} \ln \prod_{i=1}^3 e^{-\frac{(Y_i - \beta_1 X_i)^2}{2\sigma^2}} = \underset{\beta_1}{\operatorname{argmax}} \sum_{i=1}^3 -\frac{(Y_i - \beta_1 X_i)^2}{2\sigma^2} \\ &= \underset{\beta_1}{\operatorname{argmin}} \sum_{i=1}^3 (Y_i - \beta_1 X_i)^2 = \underset{\beta_1}{\operatorname{argmin}} \left(\beta_1^2 \sum_{i=1}^3 X_i^2 - 2\beta_1 \sum_{i=1}^3 X_i Y_i + \sum_{i=1}^3 Y_i^2 \right) \\ &= \frac{\sum_{i=1}^3 X_i Y_i}{\sum_{i=1}^3 X_i^2} = \frac{(-1) \cdot 1 + 0 \cdot (-1) + 2 \cdot 1}{(-1)^2 + 0^2 + 2^2} = \frac{1}{5}.\end{aligned}$$

b. În acest caz, $Y_i|X_i \sim \mathcal{N}(\beta_0, \sigma^2)$ pentru $i = 1, 2, 3$. Similar raționamentului de la punctul precedent, vom avea:

$$\begin{aligned}\hat{\beta}_0 &= \underset{\beta_0}{\operatorname{argmax}} P(Y|X; \beta_0) \stackrel{i.i.d.}{=} \underset{\beta_0}{\operatorname{argmax}} \prod_{i=1}^3 P(Y_i|X_i; \beta_0) \\ &= \underset{\beta_0}{\operatorname{argmax}} \prod_{i=1}^3 \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(Y_i - \beta_0)^2}{2\sigma^2}} = \underset{\beta_0}{\operatorname{argmax}} \prod_{i=1}^3 e^{-\frac{(Y_i - \beta_0)^2}{2\sigma^2}} \\ &= \underset{\beta_0}{\operatorname{argmax}} \ln \prod_{i=1}^3 e^{-\frac{(Y_i - \beta_0)^2}{2\sigma^2}} = \underset{\beta_0}{\operatorname{argmax}} \sum_{i=1}^3 -\frac{(Y_i - \beta_0)^2}{2\sigma^2} \\ &= \underset{\beta_0}{\operatorname{argmin}} \sum_{i=1}^3 (Y_i - \beta_0)^2 = \underset{\beta_0}{\operatorname{argmin}} \left(3\beta_0^2 - 2\beta_0 \sum_{i=1}^3 Y_i + \sum_{i=1}^3 Y_i^2 \right) \\ &= \frac{\sum_{i=1}^3 Y_i}{3} = \frac{1 + (-1) + 1}{3} = \frac{1}{3}.\end{aligned}$$

³⁰³Acest punct poate fi văzut ca o aplicație / exemplificare a modelului de regresie liniară de la problema 1.ab.

3. (Regresia liniară: prezentare generală, rezolvare folosind MLE;
 regresie neliniară [LC: polinomială];
 regularizare de normă L_2 (regresia ridge),
 corespondență cu estimarea MAP)
 prelucrare de Liviu Ciortuz, după
 ■ □ • ○ CMU, 2015 spring, Alex Smola, HW1, pr. 2

Introducere:

Obiectivul acestei probleme este acela de a ne furniza o privire mai generală³⁰⁴ asupra regresiei liniare, mai întâi din perspectiva estimării de verosimilitate maximă (engl., Maximum Likelihood Estimation, MLE) a parametrilor, iar apoi din perspectiva estimării de probabilitate maximă a posteriori (engl., Maximum-a-Posteriori Estimation probability, MAP). Vom vedea că în cel de-al doilea caz în expresia funcției obiectiv apar și niște termeni de *regularizare*.³⁰⁵

În fiecare dintre cele două cazuri vom pune în evidență *i.* cum anume problema de *estimare* a parametrilor pentru respectiva problemă de regresie liniară conduce la [și, chiar mai mult, este echivalentă cu] a rezolva o anumită problemă de *optimizare* scrisă sub formă matriceală și *ii.* cum se calculează *soluția analitică* a acelei probleme de optimizare. Vom prezenta succint și o variantă de regresie neliniară, și anume regresia polinomială.

A. Regresia liniară: prezentare generală, bazată pe estimarea în sens MLE și, echivalent, folosind criteriul LSE

Considerăm un *model liniar* care conține o componentă de tip „zgomot“ (engl., noise) care urmează o distribuție gaussiană:

$$Y_i = X_i \cdot w + b + \varepsilon_i \text{ cu } \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \text{ pentru } i = 1, \dots, n, \quad (149)$$

unde $Y_i \in \mathbb{R}$ este „răspunsul“ pentru intrarea $X_i \in \mathbb{R}^d$ (acesta din urmă fiind văzut ca vector-linie), $b \in \mathbb{R}$ este un număr real (care în terminologia de limbă engleză se numește *bias*, iar în limba română *termen liber*), $w \in \mathbb{R}^d$ este tot un vector-linie, format din *ponderi* (engl., weights) care „acționează“ asupra instanțelor X_i , iar ε_i sunt „zgomote“ i.i.d.³⁰⁶ de tip gaussian, cu varianță σ^2 .

Având instanțele $\{X_i | i = 1, \dots, n\}$, scopul nostru este să estimăm valorile parametrilor w și b , care specifică / determină modelul de regresie.

Vom arăta că a rezolva modelul liniar (149) prin metoda MLE revine în mod echivalent la a rezolva următoarea problemă de tip *Least [Sum of] Squared Errors* (rom., *suma celor mai mici pătrate*):

$$\arg \min_{\beta} \underbrace{(Y - X' \beta)^T (Y - X' \beta)}_{\|Y - X' \beta\|^2}, \quad (150)$$

unde notația $\| \cdot \|$ indică norma euclidiană (L_2),³⁰⁷ $Y = (Y_1, \dots, Y_n)^\top$, $X'_i = (1, X_i)^\top$, $X' = (X'_1, \dots, X'_n)^\top$, iar $\beta = (b, w)^\top$.³⁰⁸

³⁰⁴La precedentele două probleme (1 și 2) am lucrat cu o singură variabilă de intrare; aici vom lucra cu un număr arbitrar (d) de variabile de intrare (X_i , cu $i = 1, \dots, d$).

³⁰⁵La problema 1, termenul de regularizare apare sub forma $\frac{a^2}{\lambda^2}$ în relația (148).

³⁰⁶Adică independente și identic distribuite.

³⁰⁷Vă readucem aminte că dat fiind vectorul $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, norma sa de ordin k , cu $k \in \mathbb{N}$, se definește astfel: $\|x\|_k = (x_1^k + \dots + x_d^k)^{1/k} = \sqrt[k]{x_1^k + \dots + x_d^k}$. În general, atunci când nu se precizează indicele din scrierea lui $\| \cdot \|_k$, se consideră că este vorba despre valoarea $k = 2$, care corespunde normei euclidiene.

³⁰⁸Matricea X' se numește *matrice de design*. Linia i a acestei matrice este instanța X_i extinsă cu elementul 1 pe prima poziție, adică $(X'_i)^\top$.

- a. Pornind de la modelul (149), calculați funcția densitate de probabilitate (p.d.f.) condițională a „răspunsului“ Y_i văzut ca variabilă aleatoare, în raport cu intrarea X_i și cu parametrii w și b . (Notație: $\Pr(Y_i|X_i, w, b)$.) Remarcați faptul că putem să-l privim pe X_i pur și simplu ca pe o instanță (punct fixat) din \mathbb{R}^d , nu ca pe o variabilă aleatoare.
- b. Calculați în mod explicit expresia funcției de log-verosimilitate a datelor $\ell(\beta) \stackrel{\text{def}}{=} \Pr(Y|X, \beta)$, unde $X \stackrel{\text{not.}}{=} (X_1, \dots, X_n)$.
- c. Demonstrați că a găsi estimarea de tip MLE a parametrului β — adică, valoarea lui β pentru care funcția de log-verosimilitate condițională $\ell(\beta)$ își atinge maximul — este echivalent cu a rezolva problema de tip *Least Squared Errors* (150).
- d. Deducreți valoarea lui β care maximizează funcția de log-verosimilitate condițională. Puteți presupune că *matricea de design* X' , care a fost definită mai sus, are rangul maxim posibil (engl., full rank) în spațiul determinat de vectorii-coloană din care este formată această matrice.

Sugestie (1): Puteți, eventual, să folosiți următoarele formule de calcul cu derivate vectoriale:³⁰⁹

$$(5a) \quad \frac{\partial}{\partial X} a^\top X = \frac{\partial}{\partial X} X^\top a = a$$

$$(5b) \quad \frac{\partial}{\partial X} X^\top AX = (A + A^\top)X$$

B. Regresie neliniară [LC: polinomială]

- e. Considerăm vectorul $\phi(X_i) = (1, X_i, X_i^2, \dots, X_i^k)^\top$, cu $X_i \in \mathbb{R}$. În noul model de regresie pe care-l construim acum (și pe care-l putem numi / considera un *model de ordin k*), vom exprima vectorul Y astfel:

$$Y_i = \phi(X_i) \cdot \beta' + \varepsilon_i \text{ cu } \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \text{ pentru } i = 1, \dots, n, \quad (151)$$

unde $\beta' \in \mathbb{R}^{k+1}$.

Arătați că atunci când facem estimare în sens MLE și presupunem că matricea $\phi(X) \stackrel{\text{not.}}{=} (\phi(X_1), \phi(X_2), \dots, \phi(X_n))^\top$ are rangul maxim în spațiul determinat de coloanele acestei matrice, valoarea optimă a parametrului β' din relația (151) este

$$(\phi(X)^\top \phi(X))^{-1} \phi(X)^\top Y.$$

Sugestie (2): Nu este nevoie să elaborați toți pașii demonstrației, cum ați procedat în secțiunea A. Concentrați-vă pe schimbările care trebuie operate asupra expresiei funcției de log-verosimilitate, și deduceți noua formă a problemei de optimizare.

C. Regresia liniară cu „zgomot“ gaussian și regularizare de normă L_2 (*regresia ridge*); estimarea ponderilor în sensul MAP

*Comentariu:*³¹⁰ Multe probleme de regresie pot avea sute sau mii de atribute / variabile de intrare (numite și *variabile de predicție*). Dacă aceste atribute sunt corelate,

³⁰⁹Cf. *Matrix Identities*, Sam Roweis, 1999, <http://www.cs.nyu.edu/~roweis/notes/matrixid.pdf>.

³¹⁰Cf. CMU, 2008 fall, Eric Xing, HW1, pr. 4.2.

tehniciile standard de regresie (precum cele de la punctele A și B) vor conduce în faza de antrenament la modele foarte complexe, precum și la erori de generalizare mari în faza de testare / generalizare. O altă chestiune care poate să apară este faptul că unele probleme pot avea mai multe atribute decât numărul de instanțe din setul de date de antrenament. Atunci când se întâmplă aşa ceva, metodele regresie standard vor *eșua*. (Vom vedea mai jos de ce.) O metodă care ne permite să rezolvăm ambele chestiuni de mai sus este *regresia ridge*. Ideea pe care se bazează această metodă de regresie este una simplă: este bine să modificăm *funcția de cost / pierdere* (engl., loss function), adăugând un *termen de penalizare* aplicat ponderilor (engl., weights), pentru a le determina să ia valori mici. Acest termen de penalizare este cunoscut ca *regularizator*; el controlează *complexitatea modelului* prin faptul că permite ponderi mari (în valoare absolută) doar pentru *cele mai importante atribute* din model. Regresia *ridge* penalizează pătratul lungimii / mărimii vectorului de ponderi (β). Aceasta este numită uneori *penalizare L_2* , fiindcă este pătratul normei L_2 (adică, norma euclidiană) aplicate vectorului de ponderi (β).

f. În situația în care matricea $X'^\top X'$ din secțiunea A (respectiv matricea $\phi(X)^\top \phi(X)$ din secțiunea B) nu este inversabilă,³¹¹ se poate arăta că există $\lambda > 0$, astfel încât matricea $X'^\top X' + \lambda I$ — unde I este matricea identitate de dimensiune $d + 1$ — să fie inversabilă.³¹²

Arătați că

$$\hat{\beta}'' \stackrel{not.}{=} (X'^\top X' + \lambda I)^{-1} X'^\top Y$$

este soluția problemei de optimizare

$$\arg \min_{\beta''} (\|Y - X'\beta''\|^2 + \lambda \|\beta''\|^2). \quad (152)$$

g. Considerăm cazul în care β'' urmează o distribuție a priori $\beta'' \sim \mathcal{N}(0, \eta^2 I)$.

Scrieți distribuția a posteriori a parametrului β'' în raport cu Y_i și instanța X_i , precum și distribuția a posteriori a lui β'' în raport cu Y și întreg setul de date de antrenament, X . Veți considera că β'' și „zgomotele” $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, pentru $i = 1, \dots, n$ sunt independente.

Sugestie (3): Folosiți regula lui Bayes: $\Pr(\beta''|Y_i, X_i) = \frac{\Pr(Y_i|X_i, \beta'') \Pr(\beta''|X_i)}{\Pr(Y_i|X_i)} = \frac{\Pr(Y_i|X_i, \beta'') \Pr(\beta'')}{\Pr(Y_i|X_i)}$. Apoi urmați aceiași pași ca în secțiunea A.

Sugestie (4): Folosiți faptul că prin corelarea a două sau mai multe variabile gaussiene unidimensionale independente se obține o distribuție (multidimensională) tot de tip gaussian. (Vedeți problema 34 de la capitolul de *Fundamente*.)

h. $\hat{\beta}''$, estimarea în sens MAP a parametrului β'' , este definită ca fiind acea valoare a lui β'' pentru care se obține *modul* [adică, valoarea maximă a] probabilității a posteriori $\Pr(\beta''|Y, X)$. Arătați că identificarea acestei estimări MAP

³¹¹Aceasta este situația când, spre exemplu, numărul trăsăturilor, $d + 1$ din secțiunea A — respectiv $k + 1$ din secțiunea B —, este mai mare decât numărul de instanțe, n (adică, $d + 1 > n$). În acest caz, vom avea $\text{rang}(X'^\top X') = \text{rang}(X')$ (vedeți documentul *Matrix identities*, de Sam Roweis, formula (2f)). Așadar, $\text{rang}(X'^\top X')$ este mai mic sau egal cu $\min(n, d + 1) = n$, care este mai mic decât $d + 1$. Prin urmare, matricea $X'^\top X'$, care are dimensiunea $(d + 1) \times (d + 1)$, nu are rangul maxim și în consecință nu poate fi inversată.

³¹²Pentru a înțelege în mod intuitiv această chestiune, observați că dacă două coloane din matricea $X^\top X$ ar fi liniar dependente, atunci λI ar adăuga exact aceeași valoare (λ) la doar două dintre componente (diferite) ale acestor coloane. Așadar, aceste coloane ar deveni liniar independente în matricea $X^\top X + \lambda I$.

conduce la problema de optimizare (152) în cazul în care λ poate fi exprimat în funcție de σ și η , adică $\lambda = h(\sigma, \eta)$. Găsiți expresia funcției $h(\sigma, \eta)$.

i. Indicați o problemă care poate eventual să apară dacă am elimina termenul de regularizare din relația (152). Precizați cum anume poate termenul de regularizare să rezolve această potențială problemă.

Răspuns:

a. Observați că $Y_i|X_i; w, b \sim \mathcal{N}(X_i \cdot w + b, \sigma^2)$, aşadar putem scrie funcția densitate de probabilitate (p.d.f.) a lui $Y_i|X_i, w, b$ în forma următoare:

$$p(Y_i = y_i|X_i; w, b) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - X_i \cdot w - b)^2}{2\sigma^2}\right).$$

b. Notând $y = (y_1, \dots, y_n)^\top$, întrucât instanțele X_i sunt date, iar variabilele reprezentând „zgomotele“ ε_i sunt i.i.d., funcția de verosimilitate a lui Y în raport cu X și β se scrie astfel:

$$\begin{aligned} L(\beta) &\stackrel{\text{def}}{=} p(Y = y|X, \beta) = \prod_{i=1}^n p(y_i|X_i, w, b) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - X_i \cdot w - b)^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\sum_{i=1}^n (y_i - X_i \cdot w - b)^2}{2\sigma^2}\right). \end{aligned}$$

Apoi, aplicând logaritmul natural (\ln), rezultă că funcția de log-verosimilitate a lui $Y|\beta$ este următoarea:

$$\ell(\beta) \stackrel{\text{def}}{=} \ln L(\beta) = -n \ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - X_i \cdot w - b)^2. \quad (153)$$

c. Pentru a calcula maximul funcției de log-verosimilitate $\ell(\beta)$, ne vom concentra asupra celui de-al doilea termen din expresia (153), fiindcă primul termen este constant în raport cu β . Observăm că pentru a maximiza cel de-al doilea termen din expresia (153), trebuie să minimizăm $\sum_{i=1}^n (y_i - X_i \cdot w - b)^2$. Scriind această sumă în notație matriceală / vectorială, obținem:

$$\max_{\beta} \ell(\beta) = \min_{\beta} \sum_{i=1}^n (y_i - X_i \cdot w - b)^2 = \min_{\beta} (y - X'\beta)^\top (y - X'\beta), \quad (154)$$

unde, din nou, $Y = (Y_1, \dots, Y_n)^\top$, $X'_i = (1, X_i)^\top$, $X' = (X'_1, \dots, X'_n)^\top$, iar $\beta = (b, w)^\top$.

d. Definind funcția obiectiv

$$J(\beta) \stackrel{\text{def}}{=} (y - X'\beta)^\top (y - X'\beta), \quad (155)$$

și aplicând regulile (5a) și (5b) menționate în Sugestia (1) din enunț, rezultă:³¹³

³¹³Într-adevăr,

$$\begin{aligned} J(\beta) &= (y - X'\beta)^\top (y - X'\beta) = (y^\top - (X'\beta)^\top)(y - X'\beta) = (y^\top - \beta^\top X'^\top)(y - X'\beta) \\ &= y^\top y - y^\top X'\beta - \beta^\top X'^\top y + \beta^\top X'^\top X'\beta \end{aligned}$$

$$\stackrel{(5a),(5b)}{\Rightarrow} \nabla_{\beta} J(\beta) = 0 - (y^\top X')^\top - X'^\top y + 2X'^\top X'\beta = -2X'^\top y + 2X'^\top X'\beta = 2X'^\top (X'\beta - y).$$

$$\nabla_{\beta} J(\beta) = 2X'^{\top}(X'\beta - y).$$

Se poate demonstra relativ ușor că matricea hessiană corespunzătoare funcției $J(\beta)$ este pozitiv definită (vedeți problema 7.a), deci această funcție este convexă.

Valoarea $\hat{\beta}$ pentru care se obține maximul funcției de log-verosimilitate poate fi găsită rezolvând următoarea ecuație:

$$\begin{aligned} \nabla_{\beta} J(\hat{\beta}) = 0 &\Leftrightarrow X'^{\top}(X'\hat{\beta} - y) = 0 \Leftrightarrow X'^{\top}X'\hat{\beta} = X'^{\top}y \Leftrightarrow \\ \hat{\beta} &= (X'^{\top}X')^{-1}X'^{\top}y. \end{aligned} \quad (156)$$

Remarcați faptul că matricea $(X'^{\top}X')^{-1}$ există fiindcă s-a presupus că matricea X' are rangul maxim în spațiul determinat de coloanele sale.³¹⁴ (Acest fapt, coroborat cu observația de mai sus referitoare la convexitatea funcției $J(\beta)$, implică faptul că punctul de maxim al lui J este unic.)

e. Întrucât s-a modificat relația dintre Y_i și X_i , se vor modifica și distribuția condițională (adică, p.d.f.-ul variabilei $Y_i|X_i, \beta'$) și funcția de verosimilitate. Noile lor expresii pot fi obținute pur și simplu înlocuind X_i cu $\phi(X_i)$, după cum urmează:

$$\begin{aligned} p(Y_i = y_i | X_i; w, b) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{(y_i - \phi(X_i) \cdot \beta')^2}{2\sigma^2}\right) \\ \ell(\beta) &= -n \ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \phi(X_i) \cdot \beta')^2. \end{aligned}$$

La fel cum am procedat mai înainte, putem aplica și acum metoda estimării de verosimilitate maximă:

$$\max_{\beta'} \ell(\beta') = \min_{\beta'} \sum_{i=1}^n (y_i - \phi(X_i) \cdot \beta')^2 = \min_{\beta'} (y - \phi(X)\beta')^{\top} (y - \phi(X)\beta'),$$

iar estimatorul de verosimilitate maximă este

$$\hat{\beta}' = (\phi(X)^{\top} \phi(X))^{-1} \phi(X)^{\top} y. \quad (157)$$

f. Demonstrația urmează aproape același curs ca la punctul d. Mai întâi, vom defini funcția obiectiv J'' astfel:

$$\begin{aligned} J''(\hat{\beta}'') &= (y - X'\beta'')^{\top} (y - X'\beta'') + \lambda \underbrace{\|\beta''\|^2}_{\beta^{\top} \beta} \\ \stackrel{(5a),(5b)}{\implies} \nabla_{\beta''} J''(\beta'') &= 2X'^{\top}(X'\beta'' - y) + 2\lambda\beta''. \end{aligned}$$

³¹⁴LC: Afirmația „matricea X' are rangul maxim în spațiul determinat de coloanele sale“ trebuie interpretată astfel: numărul coloanelor din matricea X — matrice ale cărei linii sunt chiar instanțele X'_i — care sunt liniar independente este maxim, deci $d+1$ (dar și $n \geq d+1$, unde n este numărul de instanțe de antrenament, deci și numărul de linii din matricea X'). Folosind următoarea proprietate din documentul *Matrix Identities* de Sam Roweis

(2f) $\text{rang}(A^{\top} A) = \text{rang}(AA^{\top}) = \text{rang}(A)$ pentru orice matrice A , rezultă că rangul matricei $X'^{\top}X'$ este $d+1$. Prin urmare, matricea $X'^{\top}X'$ este inversabilă.

Similar cu observația făcută la punctul d, se poate arăta că și în cazul funcției J'' matricea hessiană este pozitiv definită, deci J'' este funcție convexă. Valoarea $\hat{\beta}''$ pentru care se obține maximul funcției de log-verosimilitate poate fi găsită rezolvând următoarea ecuație:

$$\begin{aligned} \nabla_{\beta''} J''(\hat{\beta}'') &= 0 \\ \Leftrightarrow X'^\top (X'\hat{\beta}'' - y) + \lambda\hat{\beta}'' &= 0 \\ \Leftrightarrow (X'^\top X' + \lambda I)\hat{\beta}'' &= X'^\top y \\ \Leftrightarrow \hat{\beta}'' &= (X'^\top X' + \lambda I)^{-1}X'^\top y. \end{aligned} \quad (158)$$

g. Stim că $Y_i|X_i, \beta'' \sim \mathcal{N}(X_i \cdot \beta'', \sigma^2)$, iar $\beta'' \sim \mathcal{N}(0, \eta^2 I)$. Folosind regula lui Bayes, rezultă că

$$f(\beta''|Y_i, X_i) \propto p(Y_i|X_i, \beta'') g(\beta'') \propto \exp\left(-\frac{1}{2\sigma^2}(Y - X_i \cdot \beta'')^2\right) \cdot \exp\left(-\frac{1}{2\eta^2}\beta''^\top \beta''\right),$$

unde semnul \propto înseamnă ‘proporțional cu’, $g(\cdot)$ este funcția de densitate de probabilitate (p.d.f.) pentru β'' , iar $f(\cdot)$ este funcția de densitate de probabilitate pentru $\beta''|Y_i, X_i$. Factorul de normalizare pentru p.d.f.-ul f este definit ca $Z_i = \int_{-\infty}^{\infty} f(Y_i|X_i, \beta'')g(\beta'')d\beta''$, aşadar putem scrie p.d.f.-ul $f(\beta''|Y_i, X_i)$ astfel:

$$f(\beta''|Y_i, X_i) = \frac{1}{Z_i} \exp\left(-\frac{1}{2\sigma^2}(Y - X_i \cdot \beta'')^2 - \frac{1}{2\eta^2}\beta''^\top \beta''\right).$$

În mod similar, ținând cont de Sugestia (4) din enunț, rezultă că p.d.f.-ul lui $\beta''|Y, X$ este

$$\begin{aligned} f(\beta''|Y) &= \frac{1}{Z} \exp\left(-\frac{1}{2\sigma^2}(Y - X'\beta'')^\top(Y - X'\beta'') - \frac{1}{2\eta^2}\beta''^\top \beta''\right) \\ &= \frac{1}{Z} \exp\left(-\frac{1}{2\sigma^2}\|Y - X'\beta''\|^2 - \frac{1}{2\eta^2}\|\beta''\|^2\right), \end{aligned}$$

unde Z este factorul de normalizare, definit ca $Z = \int_{-\infty}^{\infty} f(Y|X, \beta'')g(\beta'')d\beta''$.

h. De la punctul g este clar că

$$\max_{\beta''} f(\beta''|Y_i) = \min_{\beta''} \left(\frac{1}{2\sigma^2}\|Y - X\beta''\|^2 + \frac{1}{2\eta^2}\|\beta''\|^2 \right). \quad (159)$$

Dacă η^2 devine $\frac{\sigma^2}{\lambda}$, atunci problema de minimizare devine echivalentă cu problema (152),

$$\arg \min_{\beta''} (\|Y - X\beta''\|^2 + \lambda\|\beta''\|^2).$$

Altfel spus, prin stabilirea corespondenței termenilor, suntem conduși la a lăsa $\lambda = h(\sigma, \eta) = \frac{\sigma^2}{\eta^2}$, ceea ce forțează relația (159) să devină (152).

i. Termenul de regularizare penalizează componentele din β care au valori mari; în consecință, β va avea valori mici (în normă). Prin urmare, termenul de regularizare „încurajează“ modelul să evite *overfitting*-ul, împiedicându-l astfel să se aducă la instanțele care sunt *outlier*-e (altfel, acestea ar influența în mod drastic valoarea lui β).

4.

(O proprietate a regresiei liniare, varianta LSE:
la rezolvarea cu ajutorul *formulelor analitice*,
scalarea atributelor nu schimbă predicțiile obținute
pentru instanțele de test)

■ □ • ○ MIT, 2001 fall, Tommi Jaakkola, HW1, pr. 1.6

Considerăm un set de exemple de antrenament $D = \{(x_i, y_i) | i = 1, \dots, n\}$ cu $x_i \in \mathbb{R}^d$ și $y_i \in \mathbb{R}$, precum și o instanță de test $x_{test} \in \mathbb{R}^d$ și, în plus, factorii de scalare $\alpha_1, \dots, \alpha_d \in \mathbb{R}^*$.³¹⁵ Demonstrați următoarea egalitate:

$$\text{predicted } y_{test} = \text{predicted } \tilde{y}_{test},$$

unde

$\text{predicted } y_{test}$ este predicția pentru outputul sau *valoarea de răspuns*³¹⁶ pentru instanța x_{test} după ce am „antrenat“ / produs un model de regresie liniară de tip LSE pe setul D folosind *soluția analitică*,³¹⁷

$\text{predicted } \tilde{y}_{test}$ este predicția obținută în mod similar, antrenând un model de regresie LSE pe setul de date $D' = \{(\tilde{x}_i, y_i) | i = 1, \dots, n\}$, cu $\tilde{x}_{i,j} = \alpha_j x_{i,j}$ și facând apoi predicție pentru \tilde{x}_{test} , unde $\tilde{x}_{test,j} = \alpha_j x_{test,j}$, pentru $j \in \{1, \dots, d\}$.

Observație: O proprietate similară este valabilă și în cazul metodei lui Newton; vedeti problema 170 de la capitolul de *Fundamente*. Rezultatul respectiv este însă mai general; el nu este limitat la folosirea metodei lui Newton pentru rezolvarea problemelor de regresie [liniară].

Răspuns:

Putem să formalizăm relația dintre noua *matrice de design* (\tilde{X}) și cea veche (X) folosind o *matrice de transformare*, $A \in \mathbb{R}^{d \times d}$, pe cărei diagonala sunt plasati coeficienții de scalare α_j (și care arătă celelalte elemente 0):

$$\tilde{X} = XA.$$

Utilizând această relație în „ecuația normală“ care ne dă valoarea optimă a vectorului de ponderi β , vom obține:³¹⁸

$$\begin{aligned} \tilde{\beta} &= (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top y \\ &= ((XA)^\top XA)^{-1} (XA)^\top y \\ &= (A^\top X^\top X A)^{-1} A^\top X^\top y \\ &= A^{-1} (X^\top X)^{-1} \underbrace{(A^\top)^{-1} A^\top}_{I} X^\top y \\ &= A^{-1} (X^\top X)^{-1} X^\top y \\ &= A^{-1} \beta. \end{aligned}$$

³¹⁵Pentru atributele pe care nu le scalăm vom considera în mod implicit $\alpha_j = 1$.

³¹⁶Vedeți problema 5, unde pentru această notiune am folosit notația \hat{y} .

³¹⁷Adică, „ecuația normală“ reprezentată de relația (156) de la problema 3.d.

³¹⁸În continuare, X și y se referă la datele de antrenament.

Așadar, outputul care va fi prezis este următorul:

$$\begin{aligned} \text{predicted } \tilde{y}_{test} &= \tilde{X}_{test} \tilde{\beta} \\ &= (X_{test} A)(A^{-1} \beta) = X_{test} (AA^{-1}) \beta \\ &= X_{test} \beta \\ &= \text{predicted } y_{test}. \end{aligned}$$

Observație:

La același rezultat se poate ajunge făcând o altă demonstrație, pe care doar o schițăm:

Întrucât instanțele \tilde{x}_i sunt obținute prin scalarea atributelor instanțelor x_i , multimea de transformări liniare asupra [atributelor] unei instanțe oarecare $x \in \mathbb{R}^d$ este exact aceeași cu multimea de transformări liniare asupra [atributelor] instanței scalate \tilde{x} . Predictor-ul, în ambele cazuri, este [vectorul de ponderi β care determină] funcția liniară din această mulțime care minimizează eroarea la antrenare.

[LC:] Se poate demonstra ușor că valorile optime ale funcțiilor obiectiv în cele două cazuri sunt identice. Aceasta implică faptul că pătratele erorilor (de forma $\|y_i - x_i \beta\|^2$ și respectiv $\|y_i - \tilde{x}_i \tilde{\beta}'\|^2$) sunt identice pentru orice $i = 1, \dots, n$. Ca să arătăm că avem și $x_i \beta = \tilde{x}_i \tilde{\beta}'$ pentru $i = 1, \dots, n$ (ceea ce este adevărat dacă $\tilde{\beta}' = A^{-1} \beta$), observăm mai întâi că cele două funcții obiectiv sunt convexe, fiindcă matricele lor hessiene sunt pozitiv definite; vedeti problema 7.a), iar apoi că cele două matrice hessiene sunt chiar inversabile, deoarece conform enunțului putem scrie ecuațiile „normale“. Așadar, punctele de minim pentru cele două funcții obiectiv sunt unice. Rezultă cu necesitate că relația între cei doi predictori este cea de mai sus, $\tilde{\beta}' = A^{-1} \beta$.

În concluzie, cei doi predictori vor produce rezultate identice și atunci când sunt aplicăți pe instanțele de test X_{test} și respectiv \tilde{X}_{test} .

5.

(Regresia liniară, varianta LSE:
bias-ul și matricea de covarianță a estimatorului;
Regresia ridge: bias-ul)

■ □ • CMU, (?) spring, (10-701 course), HW1, pr. 3ab+4a

După cum știți, modelul regresiei liniare are forma

$$y = x\beta + \varepsilon,$$

unde $x = (x_1, \dots, x_d)^\top$, $\beta = (\beta_1, \dots, \beta_d)^\top$, $y \in \mathbb{R}$, iar ε este un „zgomot“, pe care aici îl vom considera de tip gaussian, având $E(\varepsilon) = 0$ și $Var(\varepsilon) = \sigma^2$. Dat fiind setul de date de antrenament $(x_1, y_1), \dots, (x_n, y_n)$, în care fiecare instanță x_i este un vector de forma $(x_{i1}, \dots, x_{id})^\top \in \mathbb{R}^d$, iar $y_i \in \mathbb{R}$, vrem să estimăm parametrii β . Notăm cu X matricea de dimensiune $n \times d$ în care linia i este vectorul x_i^\top ; în mod similar, vectorul-colonă Y , cu n -componente, este format din ieșirile / „răspunsurile“ corespunzătoare instanțelor din setul de antrenament.

Remember: Am arătat la problema 3.A că a minimiza suma pătratelor erorilor — vă reamintim că în notația matricială această sumă se scrie $\|Y - X\beta\|^2$ — revine la a calcula următoarea estimare a parametrului β :

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y. \quad (160)$$

Rezultă imediat că vectorul de *predictii* corespunzător lui Y este calculat astfel:

$$\hat{y} = X\hat{\beta} = X(X^\top X)^{-1}X^\top Y.$$

Remarcați faptul că vectorul Y este o variabilă aleatoare (n -ară), fiindcă pentru orice instanță x_i , valoarea outputului y_i este afectată de „zgomotul“ aleatoriu ε_i . Într-adevăr, vectorul $\hat{\beta}$ se calculează în funcție de vectorul Y deci și el ($\hat{\beta}$) este o cantitate aleatoare.

- a. Demonstrați că $\hat{\beta}$, estimatorul regresiei liniare în varianta LSE, este *nedeplasat* (engl., unbiased), adică $E(\hat{\beta}) = \beta$.
- b. Demostrați că matricea de covarianță a lui $\hat{\beta}$ este egală cu $\sigma^2(X^\top X)^{-1}$.
- c. **Remember:** La problema 3.C am arătat că putem obține estimatorul *ridge*, o variantă a estimatorului de tip LSE pentru regresia liniară, impunând un factor de penalizare asupra mărimii vectorului format din coeficienții de regresie ($\hat{\beta}$):

$$\|X\beta - Y\|^2 + \lambda\beta^\top\beta,$$

unde parametrul λ controlează contribuția / efectul termenului de regularizare. Minimizarea acestei funcții de cost ne conduce la următorul estimator al vectorului de coeficienți ai regresiei liniare regularizate:

$$\hat{\beta} = (X^\top X + \lambda I)^{-1}X^\top Y, \quad (161)$$

care în statistică este cunoscut sub numele de *estimator ridge*.³¹⁹ Predicția de tip *ridge* a lui Y este [dată de formula]

$$\hat{y} = X\hat{\beta} = X(X^\top X + \lambda I)^{-1}X^\top Y.$$

Demonstrați că estimatorul *ridge* este deplasat (engl., biased).

Răspuns:

- a. Pornind de la relația (160), putem scrie:

$$\hat{\beta} = (X^\top X)^{-1}X^\top Y = (X^\top X)^{-1}X^\top(X\beta + \varepsilon) = \beta + (X^\top X)^{-1}X^\top\varepsilon \quad (162)$$

Apoi, aplicând proprietatea de liniaritate a mediei, vom obține:

$$E[\hat{\beta}] = E[\beta + (X^\top X)^{-1}X^\top\varepsilon] = \beta + (X^\top X)^{-1}X^\top \underbrace{E[\varepsilon]}_0 = \beta.$$

Așadar, estimatorul $\hat{\beta}$ este nedeplasat.

- b. Vom demonstra mai întâi că pentru orice vector de variabile aleatoare $V = (V_1, \dots, V_d)$, matricea sa de covarianță, care este notată cu $Cov[V]$ și are ca element generic $Cov(V_i, V_j) \stackrel{\text{def.}}{=} E[(V_i - E[V_i])(V_j - E[V_j])]$, satisfac egalitatea $Cov[V] = E[VV^\top] - E[V](E[V]^\top)$:

$$\begin{aligned} Cov[V] &\stackrel{\text{def.}}{=} [Cov(V_i, V_j)]_{i,j \in \{1, \dots, d\}} \\ &\stackrel{\text{def.}}{=} [E[(V_i - E[V_i])(V_j - E[V_j])]]_{i,j \in \{1, \dots, d\}} \\ &= [E[V_i V_j] - E[V_i]E[V_j]]_{i,j \in \{1, \dots, d\}} \\ &= E[VV^\top] - E[V](E[V]^\top). \end{aligned}$$

³¹⁹ Remarcați faptul că atât estimatorul LSE cât și estimatorul *ridge* sunt *estimatori liniari*. Prin definiție, un estimator $\tilde{\beta}$ este liniar dacă el poate fi scris ca o combinație liniară de vectorul de ieșirile Y , adică $\tilde{\beta} = KY$, unde K este o matrice oarecare de dimensiune $d \times n$.

Observație: Această egalitate este o generalizare a egalității $\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$, unde X și Y sunt variabile aleatoare oarecare; veți problema 9.c de la capitolul de *Fundamente*.

Înlocuind V cu $\hat{\beta}$ în relația pe care am demonstrat-o mai sus, obținem

$$\text{Cov}[\hat{\beta}] = E[\hat{\beta}\hat{\beta}^\top] - E[\hat{\beta}](E[\hat{\beta}]^\top). \quad (163)$$

În continuare, făcând uz de relația (162), în membrul drept al egalității (163) îl vom înlocui pe $\hat{\beta}$ cu $\beta + (X^\top X)^{-1}X^\top \varepsilon$ și apoi vom folosi rezultatul $E[\hat{\beta}] = \beta$ demonstrat la punctul a. Așadar,

$$\begin{aligned} \text{Cov}[\hat{\beta}] &= E[(\beta + (X^\top X)^{-1}X^\top \varepsilon)(\beta + (X^\top X)^{-1}X^\top \varepsilon)^\top] - \beta\beta^\top \\ &= E[(\beta + (X^\top X)^{-1}X^\top \varepsilon)(\beta^\top + \varepsilon^\top X((X^\top X)^{-1})^\top)] - \beta\beta^\top \\ &= E[(\beta + (X^\top X)^{-1}X^\top \varepsilon)(\beta^\top + \underbrace{\varepsilon^\top X((X^\top X)^{-1})^\top}_{X^\top X})] - \beta\beta^\top \\ &\stackrel{\text{lin. med.}}{=} \underbrace{E[\beta\beta^\top]}_{\beta\beta^\top} + E[\beta\varepsilon^\top X(X^\top X)^{-1} + (X^\top X)^{-1}X^\top \varepsilon\beta^\top] + \\ &\quad E[(X^\top X)^{-1}X^\top \varepsilon \varepsilon^\top X(X^\top X)^{-1}] - \beta\beta^\top \\ &\stackrel{\text{lin. med.}}{=} \underbrace{\beta E[\varepsilon]^\top X(X^\top X)^{-1}}_0 + (X^\top X)^{-1}X^\top \underbrace{E[\varepsilon]\beta^\top}_0 + \\ &\quad E[(X^\top X)^{-1}X^\top \varepsilon \varepsilon^\top X(X^\top X)^{-1}] \\ &= E[(X^\top X)^{-1}X^\top \varepsilon \varepsilon^\top X(X^\top X)^{-1}] \\ &\stackrel{\text{lin. med.}}{=} (X^\top X)^{-1}X^\top \underbrace{E[\varepsilon \varepsilon^\top]}_{\sigma^2 I_{n \times n}} X(X^\top X)^{-1} \\ &= \sigma^2 (X^\top X)^{-1} (X^\top X) (X^\top X)^{-1} \\ &= \sigma^2 (X^\top X)^{-1}. \end{aligned}$$

c. Pentru a arăta că estimatorul $\hat{\beta}$ al regresiei *ridge* este deplasat, va trebui să demonstreăm că $E[\hat{\beta}] \neq \beta$ sau, alternativ, că $E[\hat{\beta}] - \beta$ este diferit de vectorul nul. Într-adevăr, folosind relația (161), putem scrie:

$$\begin{aligned} E[\hat{\beta}] - \beta &= E[(X^\top X + \lambda I)^{-1}X^\top y] - \beta \\ &= E[(X^\top X + \lambda I)^{-1}X^\top(X\beta + \varepsilon)] - \beta \\ &\stackrel{\text{lin. med.}}{=} (X^\top X + \lambda I)^{-1}X^\top X\beta + (X^\top X + \lambda I)^{-1}X^\top \underbrace{E[\varepsilon]}_0 - \beta \\ &= (X^\top X + \lambda I)^{-1}X^\top X\beta - \beta \\ &= (X^\top X + \lambda I)^{-1}(X^\top X + \lambda I - \lambda I)\beta - \beta \\ &\stackrel{\text{distrib.}}{=} (X^\top X + \lambda I)^{-1}(X^\top X + \lambda I)\beta - (X^\top X + \lambda I)^{-1}\lambda I\beta - \beta \\ &= \beta - \lambda(X^\top X + \lambda I)^{-1}\beta - \beta \\ &= -\lambda(X^\top X + \lambda I)^{-1}\beta \\ &\neq 0. \end{aligned}$$

6. (Regresia liniară cu regularizare L_2 (Regresia *ridge*): rezolvare cu metoda gradientului, varianta “batch” și varianta stochastică)

■ □ • ○ CMU, 2008 fall, Eric Xing, HW1, pr. 4.2.2

Fie $D = \{(x_i, y_i) | i = 1, \dots, n\}$ un set de n exemple de antrenament pentru o problemă de regresie liniară. Fie $y_i \in \mathbb{R}$ eticheta (sau, variabila de răspuns) pentru exemplul i , iar $y \in \mathbb{R}^n$ vectorul-coloană alcătuit din toate variabilele de răspuns. Fiecare exemplu de antrenament are d atribută (sau, variable de intrare / predicție). Fie $x_i \in \mathbb{R}^{d \times 1}$ vectorul-coloană format din toate variabilele de predicție pentru exemplul i , iar $X \in \mathbb{R}^{n \times d}$ matricea de design, formată din toate variabilele de predicție, linia i a matricei X fiind x_i^\top . Fie $\beta \in \mathbb{R}^{d \times 1}$ un vector-coloană care conține *ponderile* / *parameterii* pe care vrem să le / îi „învățăm“.

Comentariu: La problema 3.C am arătat (folosind calculul matriceal și derivatele vectoriale) că *soluția analitică* a problemei

$$\arg \min_{\beta} [\|y - X\beta\|^2 + \lambda\|\beta\|^2]$$

este

$$\hat{\beta}_{ridge} = (X^\top X + \lambda I)^{-1} X^\top y.$$

Deși această *soluție analitică* este ușor de implementat, ea poate deveni total nepractică atunci când numărul atributelor este foarte mare, fiindcă trebuie să calculăm atât produsul $X^\top X$ cât și inversa acestei matrice. O tehnică / soluție pe care o putem folosi pentru a depăși acest neajuns este metoda *gradientului descendente* (engl., gradient descent), care este o metodă de optimizare, iterativă.

a. La acest punct vom lucra cu varianta “batch” (sau “steepest descent”) a metodei gradientului descendente. La fiecare iterație se îmbunătățesc ponderile β , aplicând următoarea *regulă de actualizare*:

$$\beta^{(t)} = \beta^{(t-1)} - \eta \nabla_{\beta} \ell(\beta^{(t-1)}),$$

unde constanta η este așa-numita *rată de învățare* (engl., learning rate), iar $\nabla_{\beta} \ell(\beta)$ este transpusul vectorului gradient pentru funcția de cost / pierdere $\ell(\beta)$.

Deducreți regula de actualizare pentru metoda gradientului, varianta “batch”, știind că expresia funcției de log-verosimilitate $\ell(\beta)$ este:

$$\ell(\beta) = \frac{1}{2} \sum_{i=1}^n (y_i - x_i^\top \beta)^2 + \frac{\lambda}{2} \beta^\top \beta.$$

Sugestie: Puteți folosi următoarele formule de calcul pentru derivare vectorială (din documentul *Matrix Identities*, de Sam Roweis, 1999):³²⁰

$$(5a) \quad \frac{\partial}{\partial X} a^\top X = \frac{\partial}{\partial X} X^\top a = a \quad (5b) \quad \frac{\partial}{\partial X} X^\top A X = (A + A^\top) X.$$

b. Varianta secvențială / stochastică a metodei gradientului descendente actualizează parametrii β la procesarea fiecărui exemplu. Acest fapt este extrem de folositor în cazul seturilor de date de antrenament foarte mari, precum și atunci când exemplele sunt furnizate în manieră secvențială, printr-un

³²⁰<http://www.cs.nyu.edu/~roweis/notes/matrixid.pdf>.

“stream” de date. În acest caz, *regula de actualizare* a valorilor β este următoarea:

$$\beta^{(t)} = \beta^{(t-1)} - \eta \nabla_{\beta} \ell_i(\beta^{(t-1)}; x_i, y_i),$$

unde η este rata de învățare (o constantă), iar $\ell_i(\beta; x_i, y_i)$ este transpusul vectorului gradient pentru funcția de cost / pierdere calculată pentru exemplul particular i .

Calculați regula de actualizare pentru varianta stochastică a metodei gradientului descendente, știind că $\ell_i(\beta; x_i, y_i)$ este

$$\ell_i(\beta; x_i, y_i) = (y_i - x_i^T \beta)^2 + \frac{\lambda}{2} \beta^T \beta.$$

Răspuns:

a. Mai întâi vom aplica unele transformări elementare asupra funcției de log-verosimilitate ℓ (în aşa fel încât să putem aplica ulterior regulile de derivare vectorială):

$$\begin{aligned} \ell(\beta) &= \frac{1}{2} \sum_{i=1}^n (y_i - x_i^T \beta)^2 + \frac{\lambda}{2} \beta^T \beta = \frac{1}{2} \sum_{i=1}^n (y_i^2 - 2y_i x_i^T \beta + \underbrace{(x_i^T \beta)^2}_{(x_i^T \beta)^T (x_i^T \beta)}) + \frac{\lambda}{2} \beta^T \beta \\ &= \frac{1}{2} \sum_{i=1}^n (y_i^2 - 2y_i x_i^T \beta + \beta^T x_i x_i^T \beta) + \frac{\lambda}{2} \beta^T \beta. \end{aligned}$$

Acum vom calcula gradientul funcției ℓ , aplicând regulile de derivare vectoriale:

$$\nabla_{\beta} \frac{\partial \ell(\beta)}{\partial \beta} \stackrel{(5a)}{=} \frac{1}{2} \sum_{i=1}^n (-2y_i x_i + 2x_i x_i^T \beta) + \lambda \beta = \sum_{i=1}^n [-y_i x_i + x_i (x_i^T \beta)] + \lambda \beta. \quad (164)$$

Așadar, pentru regresia *ridge*, metoda gradientului folosește în varianta “batch” următoarea regulă de actualizare:

$$\beta^{(t)} = \beta^{(t-1)} - \eta \left[\sum_{i=1}^n [-y_i x_i + x_i (x_i^T \beta^{(t-1)})] + \lambda \beta^{(t-1)} \right].$$

b. Procedând similar cu rezolvarea de la punctul precedent, vom ajunge la următorul rezultat:

$$\beta^{(t)} = \beta^{(t-1)} - \eta \left[-y_i x_i + x_i (x_i^T \beta^{(t-1)}) + \lambda \beta^{(t-1)} \right].$$

7.

(Regresia liniară, varianta LSE:
rezolvare cu metoda lui Newton)

■ □ • ○ *Stanford, 2007 fall, Andrew Ng, HW1, pr. 1*

Introducere: La problema 3 am rezolvat problemele de regresie prezentate acolo folosind *metoda analitică* (și obținând așa-numitele *ecuații normale*, (156), (157), (158)). În schimb, aici (și la problema 6) vom folosi *metode de optimizare iterativă*: metoda lui Newton și respectiv metoda gradientului.

În această problemă vom demonstra că la folosirea *metodei lui Newton*³²¹ pentru rezolvarea problemei de regresie liniară în forma *sumei celor mai mici pătrate* (engl., least squared error, LSE), avem nevoie de o singură iterație pentru a ajunge la convergență (și, deci, pentru a obține soluția $\hat{\beta}$).

a. Calculați matricea hessiană³²² pentru funcția de cost / pierdere

$$J(\beta) = \frac{1}{2} \sum_{i=1}^n (\beta^\top x^{(i)} - y^{(i)})^2.$$

Observație: Funcția de cost $J(\beta)$ coincide (până la factorul 1/2) cu funcția obiectiv din problema de optimizare (150) de la problema 3.A.³²³

b. Arătați că indiferent de valoarea $\beta^{(0)}$ pe care o atribuim inițial vectorului β , la prima iterare a metodei lui Newton se obține vectorul $\beta^{(1)} = (X^\top X)^{-1} X^\top y = \hat{\beta}$, adică exact soluția problemei de regresie liniară de tipul (sau, în varianta) *suma celor mai mici pătrate*.

Răspuns:

a. Calculăm mai întâi derivatele parțiale ale funcției $J(\beta)$ în raport cu β_j , pentru $j = 1, \dots, d$, unde d este numărul de atrbute din instanțele $x^{(i)}$:

$$\frac{\partial J(\beta)}{\partial \beta_j} = \sum_{i=1}^n (\beta^\top x^{(i)} - y^{(i)}) x_j^{(i)}.$$

Rezultă că derivatele parțiale de ordin secund ale lui $J(\beta)$ sunt de forma următoare:

$$\frac{\partial^2 J(\beta)}{\partial \beta_j \partial \beta_k} = \sum_{i=1}^n \frac{\partial}{\partial \beta_k} (\beta^\top x^{(i)} - y^{(i)}) x_j^{(i)} = \sum_{i=1}^n x_j^{(i)} x_k^{(i)} = (X^\top X)_{jk}.$$

Așadar, matricea hessiană pentru funcția $J(\beta)$ este $X^\top X$.³²⁴

Observație: Se poate demonstra imediat că matricea $X^\top X$ este pozitiv definită: pentru orice $z \in \mathbb{R}^d$ avem $z^\top (X^\top X) z = (Xz)^\top (Xz) = (Xz)^2 \geq 0$. În consecință, J este funcție convexă, deci admite [măcar un punct de] minim. În anumite condiții — de exemplu, atunci când $X^\top X$ este matrice inversabilă, aşa cum s-a presupus la problema 3.A —, punctul respectiv de minim este unic.

b. Pornind de la o valoarea arbitrară $\beta^{(0)}$ pentru vectorul de ponderi β , metoda lui Newton calculează $\beta^{(1)}$ conform formulei

$$\beta^{(1)} = \beta^{(0)} - H^{-1} \nabla_\beta J(\beta^{(0)}).$$

³²¹Pentru o introducere la metoda lui Newton, vedeți *Comentariul* din enunțul problemei 80 de la capitolul de *Fundamente*.

³²²Denumirea de matrice hessiană este derivată de la numele matematicianului german Ludwig Otto Hesse (1811 - 1874), care l-a avut ca profesor pe un alt matematician german vestit, Carl Gustav Jacob Jacobi (1804 - 1851). Prin definiție, matricea hessiană a unei funcții — pentru care există derivatele parțiale de ordinul întâi și al doilea — este o matrice pătratică formată din derivatele parțiale de ordin secund ale funcției respective.

³²³În ce privește notațiile din aceste două probleme, avem corespondențele următoare: $x^{(i)} \rightarrow X'_i$, $y^{(i)} \rightarrow Y_i$ și $X \rightarrow X'$.

³²⁴Se poate ajunge la acest rezultat și aplicând pur și simplu regulile de derivare vectorială [care au fost învățate la cursul de algebră liniară].

Se poate arăta relativ ușor că $\nabla_{\beta} J(\beta) = X^T X \beta - X^T y$.³²⁵ Prin urmare,

$$\begin{aligned}\beta^{(1)} &= \beta^{(0)} - (X^T X)^{-1} (X^T X \beta^{(0)} - X^T y) \\ &= \beta^{(0)} - \beta^{(0)} + (X^T X)^{-1} X^T y \\ &= (X^T X)^{-1} X^T y.\end{aligned}$$

În consecință, indiferent de valoarea $\beta^{(0)}$ pe care o atribuim inițial vectorului β , metoda lui Newton obține soluția $\hat{\beta}$ (adică, optimul funcției obiectiv $J(\beta)$) după o singură iterare.

8.

([Alte] două variante ale regresiei liniare:
regresia liniară local-ponderată, cu „zgomot“ gaussian;
regresia liniară cu „zgomot“ modelat cu distribuția Laplace)

■ □ • ◦ CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh, HW2, pr. 3.1-2
CMU, 2007 fall, Carlos Guestrin, HW1, pr. 2.2

Remember: Pentru regresia liniară, se dă un set de date de antrenament de forma $D = (X, y) = \{(x_i, y_i) | i = 1, 2, \dots, n\}$, unde $x_i \in \mathbb{R}^{d \times 1}$, adică $x_i = (x_{i,1}, \dots, x_{i,d})^\top$, $y_i \in \mathbb{R}$, $X \in \mathbb{R}^{n \times d}$, linia i a matricei X fiind x_i^\top și, în sfârșit, $y = (y_1, \dots, y_n)^\top$. Folosind un model [parametrizat] de forma $y_i = x_i \beta + \varepsilon_i$, unde ε_i sunt „zgomote“ generate de o anumită distribuție probabilistă, metoda regresiei liniare caută să afle acea valoare a vectorului de parametri β care asigură cea mai bună „potrivire“ / adevarare (engl., *fit*) a modelului de mai sus pe datele de antrenament D . O modalitate (sau, un *criteriu*) de a măsura / evalua această „potrivire“ / adevarare este să găsim acel β care minimizează o funcție de cost / pierdere (engl., *loss function*) dată, $J(\beta)$. La problema 3.A am introdus modelul regresiei lineare în varianta *sumei pătratelor erorilor* (engl., least squared errors, LSE), iar la secțiunile B și C din aceeași problemă am prezentat două extensii / variante ale acestui model: regresia polinomială și regresia *ridge*.

În această problemă vom explora alte două extensii / variante ale modelului de regresie LSE: *regresia local-ponderată*, în cazul căreia asociem fiecărei instanțe x_i un coeficient w_i care reprezintă ponderea / importanța respectivei instanțe, precum și un model de regresie în care „zgomotele“ ε_i sunt modelate cu distribuția Laplace.

A. Regresie liniară *local-ponderată*, cu „zgomot“ gaussian³²⁶

Presupunem că „zgomotele“ $\varepsilon_1, \dots, \varepsilon_n$ sunt [tot de tip gaussian, tot] independente, dar acum varianțele nu mai sunt identice, deci $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$, pentru $i = 1, \dots, n$.

a. Stabiliti formula de calcul pentru estimarea de verosimilitate maximă (MLE) a vectorului de parametri β în această variantă de regresie.

³²⁵Cunoscând $\frac{\partial J(\beta)}{\partial \beta_j}$ de la punctul a , se poate scrie vectorul gradient $\nabla_{\beta} J(\beta) \stackrel{\text{def.}}{=} \left(\frac{\partial J(\beta)}{\partial \beta_1}, \dots, \frac{\partial J(\beta)}{\partial \beta_d} \right)$ astfel:

$$\nabla_{\beta} J(\beta) = \sum_{i=1}^n (\beta^\top x^{(i)} - y^{(i)}) x^{(i)} = \sum_{i=1}^n (\beta^\top x^{(i)}) x^{(i)} - \sum_{i=1}^n y^{(i)} x^{(i)} = X^\top X \beta - X^\top y.$$

Pentru a justifica ultima egalitate, este suficient să scrieți componentele generice ale celor două matrice, $X^\top X \beta$ și $X^\top y$.

³²⁶Acest model de regresie corespunde sumei ponderate a celor mai mici pătrate (engl., locally-weighted least squares).

Sugestie: Următoarele formule (din documentul *Matrix Identities*, de Sam Roweis, 1999)³²⁷ vă pot fi de folos:

$$(5a) \frac{\partial}{\partial X} a^\top X = \frac{\partial}{\partial X} X^\top a = a \quad (5b) \frac{\partial}{\partial X} X^\top AX = (A + A^\top)X.$$

b. Arătați că estimarea în sens MLE pe care ați calculat-o la punctul a coincide cu valoarea parametrului β pentru care se obține minimul unei funcții de cost / pierdere de forma următoare:³²⁸

$$J_W(\beta) = \sum_i w_i (y_i - x_i^\top \beta)^2.$$

Exprimați ponderea w_i în funcție de σ_i^2 , varianța „zgomotului” ε_i din exemplul de antrenament i .

c. Explicați de ce acest nou estimator — suma ponderată a celor mai mici pătrate — este [uneori] preferată în raport cu versiunea neponderată.

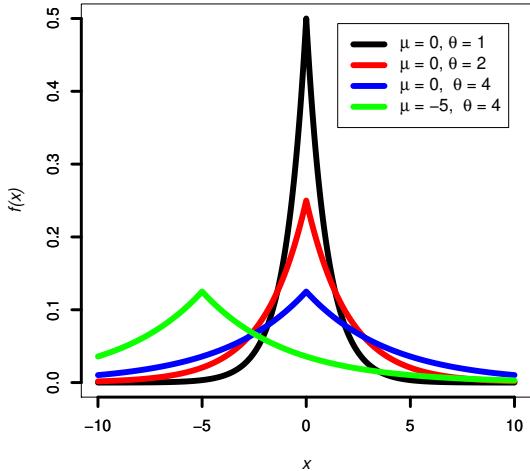
Sugestie: Analizați cazul / cazurile în care σ_i^2 are o valoare mare, comparativ cu cazul / cazurile în care valoarea sa este mică.

B. Regresie liniară [neponderată], cu „zgomot“ de tip *Laplace*

Distribuția Laplace: p.d.f.

Presupunem că „zgomotele” $\varepsilon_1, \dots, \varepsilon_n$ sunt distribuite în mod identic și independent, conform unei distribuții de tip Laplace de parametri $\mu = 0$ și $\theta > 0$. Vă readucem aminte că funcția de densitate de probabilitate (p.d.f.) pentru această distribuție este

$$\text{Laplace}(x; \mu, \theta) \stackrel{\text{def.}}{=} \frac{1}{2\theta} \exp\left(-\frac{|x - \mu|}{\theta}\right).$$



d. Găsiți funcția de cost / pierdere $J_{\text{Laplace}}(\beta)$ a cărei minimizare este echivalentă cu găsirea estimării în sens MLE pentru parametrul β în cazul acestui tip de modelare a „zgomotului”.

e. Care credeți că este avantajul acestui model, comparativ cu varianta standard (cea în care modelarea „zgomotului” se face folosind o distribuție gaussiană)?

Sugestie: Gândiți-vă la excepții / anomalii (engl., outliers).

³²⁷<http://www.cs.nyu.edu/~roweis/notes/matrixid.pdf>.

³²⁸Această funcție se numește *suma ponderată a celor mai mici pătrate*; engl., weighted least squares loss function.

Răspuns:

a. Știm că $y_i = x_i^\top \beta + \varepsilon_i$, iar $p(y_i|x_i; \beta) = \mathcal{N}(x_i^\top \beta, \sigma_i^2)$. Așadar, formula care ne dă estimarea în sens MLE pentru parametrul β este:

$$\begin{aligned}\beta_{MLE} &= \operatorname{argmax}_{\beta} \ln \prod_i p(y_i|x_i; \beta) = \operatorname{argmax}_{\beta} \sum_i \ln p(y_i|x_i; \beta) \\ &= \operatorname{argmax}_{\beta} \sum_i \ln \left(\frac{1}{\sqrt{2\pi}\sigma_i} \exp \left(-\frac{(y_i - x_i^\top \beta)^2}{2\sigma_i^2} \right) \right) \\ &= \operatorname{argmax}_{\beta} \sum_i \left(\ln \frac{1}{\sqrt{2\pi}\sigma_i} - \frac{(y_i - x_i^\top \beta)^2}{2\sigma_i^2} \right) = \operatorname{argmax}_{\beta} \sum_i -\frac{(y_i - x_i^\top \beta)^2}{2\sigma_i^2} \\ &= \operatorname{argmin}_{\beta} \sum_i \frac{(y_i - x_i^\top \beta)^2}{2\sigma_i^2} \\ &= \operatorname{argmin}_{\beta} \sum_i \frac{1}{\sigma_i^2} (y_i - x_i^\top \beta)^2.\end{aligned}\tag{165}$$

Vom scrie acum expresia (165) în notație matriceală. Considerând W matricea diagonală de dimensiune $n \times n$, în care $w_{ii} = \frac{1}{\sigma_i^2}$ pentru $i = 1, \dots, n$, vom obține:

$$\beta_{MLE} = \operatorname{argmin}_{\beta} (y - X\beta)^\top W (y - X\beta).\tag{166}$$

Ca și în cazul regresiei liniare neponderate în varianta LSE, se poate demonstra relativ ușor că matricea hessiană corespunzătoare funcției de cost $J_W(\beta) = (y - X\beta)^\top W (y - X\beta)$ este *pozitiv definită*, deci această funcție este *convexă* și prin urmare admite punct de minim.

Valoarea efectivă a estimatorului β_{MLE} este soluția ecuației care se obține egalând gradientul expresiei $(y - X\beta)^\top W (y - X\beta)$ cu vectorul 0:

$$\begin{aligned}0 &= \frac{\partial}{\partial \beta} ((y - X\beta)^\top W (y - X\beta)) \\ &\stackrel{distrib.}{=} \frac{\partial}{\partial \beta} \left(y^\top W y - \underbrace{y^\top W X \beta}_{\in \mathbb{R}} - \underbrace{\beta^\top X^\top W y}_{(y^\top W^\top X \beta)^\top} + \beta^\top X^\top W X \beta \right).\end{aligned}\tag{167}$$

Pentru orice număr real z are loc egalitatea $z^\top = z$, așadar $((\beta^\top X^\top) W y)^\top = y^\top W^\top X \beta = y^\top W X \beta$, întrucât $W^\top = W$. (Vă reamintim că matricea W este diagonală.) Înținând cont de aceasta, ecuația (167) va fi rescrisă astfel:

$$\begin{aligned}0 &= \frac{\partial}{\partial \beta} (y^\top W y - 2\beta^\top X^\top W y + \beta^\top X^\top W X \beta) \stackrel{(5a)(5b)}{\Leftrightarrow} 0 = -2X^\top W y + 2X^\top W X \beta \\ &\Leftrightarrow X^\top W y = X^\top W X \beta.\end{aligned}$$

În cazul în care matricea $X^\top W X$ este inversabilă, vom avea:

$$\beta_{MLE} = (X^\top W X)^{-1} X^\top W y.\tag{168}$$

Aceasta este ecuația „normală” corespunzătoare regresiei LSE local-ponderate.

b. La relația (165) am demonstrat că estimatorul MLE al parametrului β este valoarea pentru care se atinge minimul pentru suma ponderată a celor

mai mici pătrate, $\sum_i \frac{1}{\sigma_i^2} (y_i - x_i^\top \beta)^2$. Coroborând această expresie cu relația $J_W(\beta) \stackrel{\text{def.}}{=} \sum_i w_i (y_i - x_i^\top \beta)^2$ din enunț, obținem $w_i = \frac{1}{\sigma_i^2}$.

c. Atunci când varianța σ_i^2 este mare, „zgomotul“ ε_i poate fi oricât de mare, deci punctul (x_i, y_i) poate fi un outlier. Nu este de dorit ca în astfel de cazuri estimatorul β_{MLE} să fie deplasat (engl., biased) în aşa fel încât el să „explice“ aceste outlier-e (mai ales atunci când se foloseşte ca funcție de cost / pierdere suma pătratelor erorilor). Modelul regresiei liniare local-ponderate formulat în problema de față rezolvă această chestiune ponderând „importanța“ (sau, „contribuția“) fiecărui exemplu de antrenament la funcția obiectiv (J_W) cu ajutorul unui factor care este tocmai inversul varianței corespunzătoare exemplului respectiv. În consecință, instanțele care au o varianță mare nu vor afecta mult funcția de cost / pierdere; ele pot fi ignorate, sau [cel puțin] li se poate acorda o [mai] mică importanță atunci când se caută valoarea optimă a lui β .

d. Stim că $y_i = x_i^\top \beta + \varepsilon_i$, iar $p(y_i|x_i; \beta) = \text{Laplace}(x_i^\top \beta; 0, \theta)$. Așadar, formula care ne dă estimarea în sens MLE pentru parametrul β în acest caz este:

$$\begin{aligned} \beta_{MLE} &= \underset{\beta}{\operatorname{argmax}} \ln \prod_i p(y_i|x_i; \beta) = \underset{\beta}{\operatorname{argmax}} \sum_i \ln p(y_i|x_i; \beta) \\ &= \underset{\beta}{\operatorname{argmax}} \sum_i \ln \left(\frac{1}{2\theta} \exp \left(-\frac{|y_i - x_i^\top \beta|}{\theta} \right) \right) \\ &= \underset{\beta}{\operatorname{argmax}} \sum_i \left(\ln \frac{1}{2\theta} - \frac{|y_i - x_i^\top \beta|}{\theta} \right) = \underset{\beta}{\operatorname{argmax}} \sum_i -\frac{|y_i - x_i^\top \beta|}{\theta} \\ &\stackrel{\theta \geq 0}{=} \arg \min_{\beta} \sum_i |y_i - x_i^\top \beta|. \end{aligned}$$

În concluzie,

$$J_{\text{Laplace}}(\beta) = \sum_i |y_i - x_i^\top \beta|. \quad (169)$$

e. Dacă o instanță de antrenament este “outlier”, eroarea produsă atunci când se face predicția corespunzătoare acestei instanțe — folosind valoarea corectă pentru parametrul β — este mult mai mare în cazul în care se lucrează cu „zgomot“ de tip gaussian (pentru că atunci se folosește pătratul diferenței dintre valoarea “target” și valoarea prezisă de model) decât în cazul zgomotului de tip laplacian (fiindcă în acest caz se lucrează cu modulul diferenței respective). Prin urmare, outlier-ele afectează estimarea lui β mult mai mult în cazul „zgomotului“ gaussian decât în cazul celui laplacian.

Din punctul de vedere al modelării, ținând cont că $y_i = x_i^\top \beta + \varepsilon_i$, atunci când y_i este outlier, modelul poate „explica“ acest fapt dând lui ε_i o valoare mare (în modul), pentru a se putea „accepta“ diferența dintre valoarea “target” și valoarea prezisă pentru outlier-ul respectiv. Acest fapt este posibil în cazul modelului Laplace, fiindcă funcția densitate de probabilitate (p.d.f.) a distribuției Laplace are brațele (engl., tails) mai ridicate decât p.d.f.-ul distribuției gaussiene.

Observație: Făcând legătura cu secțiunea A a problemei de față, putem spune că pentru a obține un efect similar [cu cazul „zgomotului“ Laplace], acolo am presupus că fiecare instanță îi asociem o varianță proprie, σ_i^2 . Remarcați totuși că aceste varianțe trebuie estimate (folosind, de exemplu, un algoritm de tip EM — vedeti capitolul *Schema algoritmică EM*) fiindcă ele influențează problema de optimizare, în vreme ce în modelul Laplace nu trebuie să facem astfel de estimări. Pe de altă parte, este mai dificil să facem optimizare cu o funcție de cost / pierdere care folosește norma L_1 (cazul distribuției Laplace) decât atunci când se folosește norma L_2 (cazul distribuției gaussiene), fiindcă funcția modul nu este derivabilă pe tot domeniul ei de definiție.

9.

(Regresia liniară cu regularizare L_2 (Regresia *ridge*): kernel-izare)

□ • ○ CMU, 2014 spring, B. Poczos, A. Singh, HW2, pr. 4.B

În această problemă vom arăta cum anume poate fi obținută varianta kernelizată a regresiei *ridge*.³²⁹ După cum se procedează în general la regresia liniară, vom lucra cu un set de instanțe / date $\{x_i\}_{i=1}^n$ din \mathbb{R}^d și cu valorile de „răspuns“ asociate lor, $\{y_i\}_{i=1}^n$ din \mathbb{R} .

Pentru a obține rezultate mai bune pentru problema noastră de regresie, putem folosi o funcție de „mapare“ a atributelor (engl., attribute mapping function) ϕ , care asociază fiecărui vector d -dimensional x câte un nou vector, $\bar{x} = \phi(x)$ din \mathbb{R}^m , numit vector de *trsături* (engl., feature vector), unde m este mai mare decât d .³³⁰ Considerăm matricea de design $\Phi \in \mathbb{R}^{n \times m}$, în care linia i este vectorul de trăsături \bar{x}_i^\top corespunzător instanței i . Notăm cu y vectorul-coloană format din valorile de răspuns; în acest vector, componenta de pe poziția i este valoarea de răspuns y_i , care a fost asociată instanței i .

Așa cum am arătat la problema 3.C, pentru regresia *ridge* funcția obiectiv este

$$J(\beta) = \|y - \Phi\beta\|^2 + \lambda\|\beta\|^2, \quad (170)$$

unde $\lambda > 0$ este parametrul de *regularizare*. Vom nota cu $\hat{\beta}$ soluția / estimatorul acestei variante de regresie *ridge*. În cele ce urmează vom arăta cum anume se poate calcula această soluție.

a. Mai întâi, arătați că $\hat{\beta}$, care este un vector din \mathbb{R}^m , este în *spațiul* generat de liniile din matricea Φ . Aceasta înseamnă că $\hat{\beta}$ poate fi scris ca o combinație liniară de vectorii-linie care alcătuiesc matricea Φ . Altfel spus, există vectorul $\hat{\alpha} \in \mathbb{R}^{n \times 1}$ astfel încât $\hat{\beta}$ se poate scrie sub forma

$$\hat{\beta} = \Phi^\top \hat{\alpha}. \quad (171)$$

Arătați apoi că, în ipoteza că matricele $\Phi\Phi^\top$ și $\Phi\Phi^\top + \lambda I$ sunt inversabile, urmează că vectorul de coeficienți $\hat{\alpha}$ se poate exprima cu ajutorul matricei de design Φ astfel:³³¹

$$\hat{\alpha} = (\Phi\Phi^\top + \lambda I)^{-1}y. \quad (172)$$

³²⁹Pentru o introducere în subiectul funcțiilor-nucleu, vedeti secțiunea corespunzătoare din capitolul de *Fundamente*. Similar, pentru regresia *ridge*, vedeti problema 3.C.

³³⁰În general, m este mult mai mare decât d ; vom nota acest fapt prin $m \gg d$.

³³¹Pentru regresia liniară nekernelizată, am abținut deja o astfel de relație la problema 3.f.

Sugestie: Vectorul de ponderi β din expresia (170) — ca și orice alt vector din \mathbb{R}^m — poate fi descompus în două componente, și anume $\beta = \beta_{\parallel} + \beta_{\perp}$, unde componenta β_{\parallel} aparține spațiului generat de liniile din Φ , iar componenta β_{\perp} este ortogonală pe acest spațiu.³³² Această ultimă proprietate înseamnă că produsul scalar dintre orice vector din spațiul generat de liniile din Φ și β_{\perp} este 0.

b. În practică, este foarte dificil să construim funcții de „mapare“ ϕ . Dar chiar și atunci când putem să construim astfel de funcții, calcularea produselor $\phi(x)^T \phi(x')$ — care sunt necesare pentru aflarea vectorului $\hat{\alpha}$, conform relației (172) — poate fi ineficientă dacă se folosește maniera clasică de calcul.³³³ În schimb, putem calcula în mod *implicit* aceste produse, folosind o funcție-nucleu $k(x, x')$ care are proprietatea $k(x, x') = \phi(x)^T \phi(x')$. În principiu, a *kerneliza* înseamnă să folosim „apeluri“ la o funcție-nucleu pentru a înlocui produsele $\phi(x)^T \phi(x')$.

La acest punct, vă cerem să kernelizați regresia *ridge*, presupunând că vi se dă o funcție-nucleu $k(x, x')$, unde x și x' sunt vectorii originali, de dimensiune d , care corespund instanțelor de antrenament. Va trebui să lucrați atât asupra antrenării cât și asupra testării. Mai exact, va trebui ca pentru faza de antrenare să înlocuiți cu „apeluri“ la funcția-nucleu k toate produsele scalare care apar în calculul necesar obținerii vectorului $\hat{\alpha}$ (vedeți relația (172)). De asemenea, pentru faza de testare, dată fiind o instanță de test x , va trebui să înlocuiți toate produsele scalare care apar în calculul lui $\phi(x)^T \hat{\beta} = \phi(x)^T (\Phi^T \hat{\alpha}) = (\phi(x)^T \Phi^T) \hat{\alpha}$ cu „apeluri“ la funcția-nucleu k .³³⁴

Răspuns:

a. Tinând cont de *Sugestia* din enunț, vom putea scrie astfel funcția obiectiv a regresiei *ridge* kernel-izate:

$$\begin{aligned} J(\beta) &= \|y - \Phi\hat{\beta}\|^2 + \lambda\|\hat{\beta}\|^2 = \\ J(\beta_{\parallel} + \beta_{\perp}) &= \|y - \underbrace{\Phi\beta_{\perp}}_0 - \Phi\beta_{\parallel}\|^2 + \lambda\|\beta_{\perp}\|^2 + \lambda\|\beta_{\parallel}\|^2 \\ &= \|y - \Phi\beta_{\parallel}\|^2 + \lambda\|\beta_{\perp}\|^2 + \lambda\|\beta_{\parallel}\|^2. \end{aligned}$$

Minimizarea acestei funcții obiectiv va trebui să fie făcută în raport cu cele două componente, β_{\perp} și β_{\parallel} , care sunt ortogonale una în raport cu cealaltă. Din expresia lui J rezultă că, pentru minimizare, componenta β_{\perp} trebuie să fie vectorul $0 \in \mathbb{R}^d$. În consecință, $\hat{\beta} = \beta_{\parallel}$. (Și $J(\beta) = J(\hat{\beta}_{\parallel})$.) Rezultă, deci, că vectorul β aparține spațiului generat de liniile din matricea Φ , ceea ce înseamnă că există într-adevăr vectorul $\hat{\alpha} \in \mathbb{R}^{n \times 1}$ astfel încât $\hat{\beta}$ să se poată scrie sub forma $\hat{\beta} = \Phi^T \hat{\alpha}$.

Observație: Dacă analizați problema 88 de la capitolul de *Fundamente*, unde este prezentată *Teorema de reprezentare*, veți putea identifica o altă cale de

³³²Pentru explicații mai detaliate, puteți vedea documentul *Supplemental Lecture Notes* (pentru cursul de *Machine Learning*), de John Duchi, de la Universitatea Stanford, mai precis secțiunea *A more general representer theorem*, pag. 11-12.

³³³Observați că putem rescrie în mod echivalent produsul $\phi(x)^T \phi(x')$ cu ajutorul produsului scalar al vectorilor: $\phi(x) \cdot \phi(x')$.

³³⁴Veți observa că, din cauza acestei ultime expresii, la faza de antrenare este suficient să calculăm vectorul $\hat{\alpha}$, nu și vectorul $\hat{\beta} = \Phi^T \hat{\alpha}$.

demonstrare a faptului că $\hat{\beta}$ se poate scrie sub forma $\hat{\beta} = \Phi^\top \hat{\alpha}$, și anume, calculând și apoi egalând cu 0 derivata funcției $J(\beta)$, din relația (170).³³⁵

Acum, pentru a calcula efectiv vectorul de coeficienți $\hat{\alpha}$, vom folosi relația $\hat{\beta} = \Phi^\top \hat{\alpha}$ și-l vom înlocui pe $\hat{\beta}$ cu $\Phi^\top \hat{\alpha}$ în expresia lui $J(\hat{\beta})$:

$$J(\hat{\beta}) = J(\hat{\beta}_{||}) = \|y - \Phi\hat{\beta}_{||}\|^2 + \lambda\|\hat{\beta}_{||}\|^2 = \|y - \Phi\Phi^\top \hat{\alpha}\|^2 + \lambda\hat{\alpha}^\top \Phi\Phi^\top \hat{\alpha}. \quad (173)$$

Întrucât $\hat{\beta}$ minimizează funcția $J(\beta)$, putem calcula $\hat{\alpha}$ egalând cu vectorul $0 \in \mathbb{R}^n$ derivata în raport cu $\hat{\alpha}$ a expresiei rezultat din relația (173). Așadar, vom avea:³³⁶

$$-2\Phi\Phi^\top(y - \Phi\Phi^\top \hat{\alpha}) + 2\lambda\Phi\Phi^\top \hat{\alpha} = 0 \Leftrightarrow \Phi\Phi^\top(\Phi\Phi^\top + \lambda I)\hat{\alpha} = \Phi\Phi^\top y,$$

ceea ce implică, în ipoteza că matricele $\Phi\Phi^\top$ și $\Phi\Phi^\top + \lambda I$ sunt inversabile, faptul că

$$\hat{\alpha} = (\Phi\Phi^\top + \lambda I)^{-1}y.$$

b. La faza de antrenare, trebuie să calculăm vectorul $\hat{\alpha}$. Dacă notăm cu K matricea ale cărei elemente sunt $K_{ij} = k(x_i, x_j)$ rezultă că aceasta este chiar matricea $\Phi\Phi^\top$. Așadar, conform relației (172), rezultă că $\hat{\alpha} = (K + \lambda I)^{-1}y$. (Matricea $K = \Phi\Phi^\top$ se numește *matrice-nucleu* sau *matrice Gram*.)

La faza de testare va trebui să calculăm

$$\phi(x)^\top \hat{\beta} = \phi(x)^\top (\Phi^\top \hat{\alpha}) = (\phi(x)^\top \Phi^\top) \hat{\alpha} = \sum_{i=1}^n \hat{\alpha}_i k(x_i, x).$$

10.

(Regresia *ridge kernel*-izată, cu nucleu RBF; un rezultat interesant:

atunci când parametrul de regularizare λ tinde la 0, eroarea la antrenare devine 0)

prelucrare de Liviu Ciortuz, după
 MIT, 2006 fall, Tommi Jaakkola, HW2, pr. 2.a-c

La problema 9 am prezentat versiunea kernel-izată a regresiei *ridge*. Vă reamintim că regresie *ridge* înseamnă regresie liniară cu „zgomot“ gaussian și regularizare (L_2). Regula de predicție pentru acest tip de regresie kernel-izată este

$$y(x) = \sum_{i=1}^n \hat{\alpha}_i K(x_i, x), \quad (174)$$

unde $\hat{\alpha}_i$ este valoarea optimală a coeficientului α_i , pentru $i = 1, \dots, n$. Valorile $\hat{\alpha}_i$ au fost obținute astfel:

$$\hat{\alpha} = (\mathbf{K} + \lambda \mathbf{I})^{-1}y. \quad (175)$$

³³⁵Exact așa se procedează în problema de la Stanford, Andrew Ng, 2007 fall, HW2, pr. 1.b.

³³⁶Pentru derivarea efectivă, putem folosi mai întâi descompunerea

$$\begin{aligned} \|y - \Phi\Phi^\top \hat{\alpha}\|^2 &\stackrel{\text{def.}}{=} (y - \Phi\Phi^\top \hat{\alpha})^\top (y - \Phi\Phi^\top \hat{\alpha}) = (y^\top - \hat{\alpha}^\top \Phi\Phi^\top)(y - \Phi\Phi^\top \hat{\alpha}) \\ &= y^\top y - \hat{\alpha}^\top \Phi\Phi^\top y - y^\top \Phi\Phi^\top \hat{\alpha} + \hat{\alpha}^\top (\Phi\Phi^\top)(\Phi\Phi^\top) \hat{\alpha} \\ &= y^\top y - 2\hat{\alpha}^\top \Phi\Phi^\top y + \hat{\alpha}^\top (\Phi\Phi^\top)(\Phi\Phi^\top) \hat{\alpha} \end{aligned}$$

și apoi putem aplica regulile de derivare vectorială (5a) și (5b) care au fost folosite și la problema 6.

În această formulă, $\hat{\alpha} \stackrel{not.}{=} (\hat{\alpha}_1, \dots, \hat{\alpha}_n)^\top$, $y \stackrel{not.}{=} (y_1, \dots, y_n)^\top$, iar K este așa-numita *matrice-nucleu* (sau, *matricea Gram*) corespunzătoare instanțelor x_1, \dots, x_n .³³⁷

În acest exercițiu vrem să studiem evoluția erorii la antrenare pentru acest model de regresie liniară kernel-izată în condițiile următoare:

1. folosim ca nucleu așa-numita funcție cu baza radială (engl., *radial basis function*, RBF), care este definită astfel:

$$K(x, x') = \exp\left(-\frac{\gamma}{2}\|x - x'\|^2\right), \quad \gamma > 0 \text{ pentru orice } x, x' \in \mathbb{R}^d;$$

2. valorile parametrului de regularizare λ tind la 0.³³⁸

a. Arătați că, în ipoteza $x_i \neq x_j$ pentru orice $i \neq j$, matricea Gram corespunzătoare funcției-nucleu RBF este inversabilă (indiferent de valoarea parametrului γ).³³⁹

Sugestie: Puteți folosi *teorema lui Michelli* (1986): dacă o funcție ρ de variabilă t este monotonă pentru $t \in [0, \infty)$, atunci pentru orice set de instanțe distințe x_1, \dots, x_n , matricea pătratică definită prin relația $\rho_{ij} = \rho(\|x_i - x_j\|)$ pentru $i, j \in \{1, \dots, n\}$ este inversabilă.

b. Care este valoarea funcției $y(x)$ atunci când $\lambda \rightarrow 0$? Cu alte cuvinte, care este formula de calcul pentru predicțiile făcute de către modelul de regresie *ridge* kernel-izată atunci când se face această trecere la limită?

c. Demonstrați că eroarea la antrenare produsă la trecerea la limită (pentru $\lambda \rightarrow 0$) de către regresia *ridge* kernel-izată cu nucleu RBF — presupunând, ca mai înainte, că $x_i \neq x_j$ pentru orice $i \neq j$ — este exact zero.

Răspuns:

a. Funcția $f(t) = -\gamma t^2/2$ este monoton decrescătoare în raport cu t (pentru orice $\gamma > 0$ fixat). De asemenea, funcția $g(t) = e^t$ este monoton crescătoare în raport cu t . Prin compunere, funcția $(g \circ f)(t) \stackrel{not.}{=} h(t) = e^{-\gamma t^2/2}$ va fi monoton decrescătoare în raport cu t .

Funcția-nucleu RBF, care este definită prin relația

$$K(x, x') = e^{-\frac{\gamma}{2}\|x - x'\|^2} = (g \circ f)(\|x - x'\|), \quad \text{pentru orice } x, x' \in \mathbb{R}^d,$$

satisfacă condițiile teoremei lui Michelli. Prin urmare, ea va da naștere, pentru orice set de instanțe $x_1, \dots, x_n \in \mathbb{R}^d$ care sunt diferite între ele, la o matrice Gram care este inversabilă.

b. Calculăm mai întâi valoarea, la limită, a vectorului de parametri $\hat{\alpha}$:

$$\begin{aligned} \lim_{\lambda \rightarrow 0} \hat{\alpha} &= \lim_{\lambda \rightarrow 0} ((\mathbf{K} + \lambda \mathbf{I})^{-1} y) = \left(\lim_{\lambda \rightarrow 0} ((\mathbf{K} + \lambda \mathbf{I})^{-1}) \right) y \\ &= \left[\lim_{\lambda \rightarrow 0} (\mathbf{K} + \lambda \mathbf{I}) \right]^{-1} y = K^{-1} y. \end{aligned}$$

³³⁷ *Observație:* Spre deosebire de problema 9, aici am notat matricea Gram (care acolo era desemnată cu K) cu simbolul \mathbf{K} , iar funcția-nucleu (care acolo era $k(\cdot, \cdot)$) cu $K(\cdot, \cdot)$.

³³⁸ Facem aceasta, deși a seta $\lambda = 0$ — ceea ce înseamnă a lucra fără regularizare — nu este deloc convenabil, din cauza overfitting-ului.

³³⁹ *Justificare:* Trecerea la limită pentru $\lambda \rightarrow 0$ în relația (174) necesită calcularea valorii pe care o are vectorul de coeficienți $\hat{\alpha}$ la limită. Din cauza relației (175), întrucât $\lim_{\lambda \rightarrow 0} (\mathbf{K} + \lambda \mathbf{I}) = \mathbf{K}$, trebuie ca [în prealabil] să ne asigurăm că matricea Gram \mathbf{K} este inversabilă.

Această limită este [întotdeauna] bine-definită, fiindcă

- matricea \mathbf{K} este inversabilă (vedeți punctul a),
- matricea $\mathbf{K} + \lambda \mathbf{I}$, pentru valori mici $\lambda > 0$, este și ea inversabilă (LC: datorită continuității), iar
- pentru astfel de matrice, limita matricelor inverse este inversa matricei-limită.

Folosind acum relația (174), vom putea scrie:

$$\begin{aligned}\lim_{\lambda \rightarrow 0} y(x) &= \lim_{\lambda \rightarrow 0} \left(\sum_{i=1}^n \hat{\alpha}_i K(x_i, x) \right) = \sum_{i=1}^n \lim_{\lambda \rightarrow 0} \hat{\alpha}_i K(x_i, x) \\ &= \sum_{i=1}^n B_i K(x_i, x), \text{ unde } B \stackrel{\text{not.}}{=} [B_1, \dots, B_n]^T = \mathbf{K}^{-1}y.\end{aligned}$$

În consecință, prin trecerea la limita specificată în enunț, vom avea $y(x) = \sum_{i=1}^n B_i K(x_i, x)$.

c. Pentru a demonstra că eroarea la antrenare este zero, trebuie să arătăm că $y(x_t) = y_t$ pentru orice $t \in \{1, \dots, n\}$. Plecând de la rezultatul obținut la punctul b, putem scrie:

$$\begin{aligned}y(x_t) &= \sum_{i=1}^n B_i K(x_i, x_t) = \sum_{i=1}^n \left(\sum_{j=1}^n y_j \mathbf{K}^{-1}(i, j) \right) K(x_i, x_t) \\ &\stackrel{\text{asoc.}}{=} \sum_{j=1}^n y_j \sum_{i=1}^n \mathbf{K}^{-1}(i, j) K(x_i, x_t) \stackrel{\text{sim.}}{=} \sum_{j=1}^n y_j \sum_{i=1}^n \mathbf{K}^{-1}(j, i) K(x_i, x_t) \\ &= \sum_{j=1}^n y_j \sum_{i=1}^n \mathbf{K}^{-1}(j, i) \mathbf{K}(i, t) = \sum_{j=1}^n y_j \delta(j, t) \\ &= y_t,\end{aligned}$$

unde $\mathbf{K}^{-1}(i, j)$ elementul de pe poziția (i, j) a matricei \mathbf{K}^{-1} , iar

$$\delta(i, j) = \begin{cases} 0 & \text{pentru } i \neq j \\ 1 & \text{pentru } i = j. \end{cases}$$

În cele de mai sus, am ținut cont în primul rând de faptul că funcția-nucleu K (și, în consecință și matricea \mathbf{K}) este simetrică, ceea ce implică faptul că matricea \mathbf{K}^{-1} este simetrică, deci $\mathbf{K}^{-1}(i, j) = \mathbf{K}^{-1}(j, i)$ pentru orice indici i și j . În al doilea rând, din relația $\mathbf{K}^{-1}\mathbf{K} = \mathbf{I}$ rezultă că $\sum_i \mathbf{K}^{-1}(j, i) \mathbf{K}(i, t) = \delta(j, t)$ pentru orice indici j și t .

11. (Regresia liniară cu zgomot gaussian și cu regularizare L_1 : rezolvare cu metoda *descreșterii pe coordonate*)

• ○ *Stanford, 2007 fall, Andrew Ng, HW3, pr. 3.a*³⁴⁰

La problema 3.C am prezentat cazul regresiei *ridge*, adică regresia liniară cu „zgomot“ gaussian la care funcția obiectiv este augmentată cu termenul de *regularizare* $\lambda\|\beta\|^2$. De această dată, vom analiza cazul în care regularizarea se face folosind norma L_1 . Așadar, acum dorim să minimizăm funcția obiectiv

$$J(\beta) = \frac{1}{2} \sum_{i=1}^n (\beta^\top x^{(i)} - y^{(i)})^2 + \lambda\|\beta\|_1,$$

unde $\|\beta\|_1$ este definit ca fiind $\sum_{i=1}^d |\beta_i|$.

Regularizarea de normă L_1 este foarte utilă pentru că, aşa cum vom vedea, ea oferă avantajul creării de soluții „rare“ (engl., sparse), ceea ce înseamnă că multe dintre componentele vectorului rezultat β vor fi egale cu 0.

Comentariu: Acest tip de regresie este mai dificil de rezolvat decât regresia neregularizată ori cea cu regularizare de normă L_2 , pentru că termenul L_1 nu este derivabil. Totuși, în acest scop au fost dezvoltăți mai mulți algoritmi eficienți, care lucrează foarte bine în practică. Între aceștia, o abordare foarte directă este reprezentată de metoda *descreșterii pe coordonate* (engl., *coordinate descent*).³⁴¹

În această problemă veți concepe un algoritm de tipul descreșterii pe coordonate pentru regresia liniară de tip LSE cu regularizare L_1 . Mai specific, întrucât metoda descreșterii pe coordonate este iterativă — ca de altfel și metoda gradientului, care se aplică în cazul regresiei *ridge*, aşa cum ati văzut la problema 6 — veți deduce *regula de actualizare* pentru o pondere oarecare β_i , cu $i \in \{1, \dots, d\}$.

Considerând *matricea de design* X și vectorul y aşa cum au fost definite la problemele 3 și 6, precum și vectorul de parametri β , [ne întrebăm] cum am putea [oare] să „ajustăm“ valoarea lui β_i astfel încât să minimizăm funcția obiectiv?

Pentru a răspunde la această întrebare, vom rescrie funcția obiectiv de mai sus astfel:

$$\begin{aligned} J(\beta) &= \frac{1}{2}\|X\beta - y\|^2 + \lambda\|\beta\|_1 \\ &= \frac{1}{2}\|X\bar{\beta} + X_i\beta_i - y\|^2 + \lambda\|\bar{\beta}\|_1 + \lambda|\beta_i|, \end{aligned}$$

unde $X_i \in \mathbb{R}^n$ reprezintă coloana i a matricei X ,³⁴² iar vectorul $\bar{\beta}$ este identic cu vectorul β , doar că $\bar{\beta}_i = 0$. Tot ce am făcut aici a fost să rescriem expresia precedentă, în aşa fel încât termenul β_i să apară în mod explicit în funcția obiectiv.

³⁴⁰Punctele b și c ale acestei probleme sunt de tip implementare. Acolo sunteți solicitați să aplicați algoritmul dezvoltat în acest exercițiu pe un anumit set de date de antrenament.

³⁴¹Alți algoritmi de învățare automată care folosesc metoda descreșterii (sau, alternativ, a creșterii) pe coordinate sunt menționați în nota de subsol 671 de la capitolul *Masini cu vectori-suport*.

³⁴²Vă reamintim că în matricea de design $X \in \mathbb{R}^{n \times d}$, linia i este instanța $x^{(i)}$ (văzută ca vector-colonă transpusă).

Totuși, ultima expresie încă îl are în componență sa pe $|\beta_i|$, care este nederivable, deci este dificil de optimizat. Pentru a depăși acest impediment, *remarcăm* faptul că semnul parametrului β_i este fie pozitiv fie negativ (mai precis, fie negativ fie nenegativ). Dacă am ști semnul lui β_i , atunci $|\beta_i|$ ar deveni pur și simplu un termen liniar. Așadar, putem scrie funcția obiectiv astfel:

$$J(\beta) = \frac{1}{2} \|X\bar{\beta} + \beta_i X_i - y\|^2 + \lambda \|\bar{\beta}\|_1 + \lambda s_i \beta_i, \quad (176)$$

unde prin s_i am notat semnul lui β_i , deci $s_i \in \{-1, +1\}$.

Pentru a actualiza valoarea parametrului β_i , putem să calculăm valoarea sa optimă pentru fiecare dintre cele două valori posibile ale lui s_i în parte — asigurându-ne apoi că restricționăm aceste valori optime ale lui β_i astfel încât ele să satisfacă restricția de semn pe care am fixat-o pentru cazul respectiv —, iar apoi să vedem care dintre cele două soluții corespunde valorii celei mai bune pentru funcția obiectiv la iterată respectivă.

Vă cerem ca pentru fiecare dintre valorile posibile pentru semnul s_i , să calculați valorile optime corespunzătoare pentru parametrul β_i .

Sugestie: În acest scop, puteți fixa valoarea lui s_i în expresia (176), pe care apoi urmează să o derivați în raport cu β_i pentru a-i determina valoarea optimă. În final, impuneți restricții asupra valorii optime obținute pentru β_i , astfel încât ea să aparțină domeniului corespunzător. (De exemplu, pentru $s_i = +1$, trebuie să impuneți asupra lui β_i restricția $\beta_i \geq 0$.)

Răspuns:

În cazul $s_i = +1$,

$$\begin{aligned} J(\beta) &= \frac{1}{2} (X\bar{\beta} + \beta_i X_i - y)^\top (X\bar{\beta} + \beta_i X_i - y) + \lambda \|\bar{\beta}\|_1 + \lambda \beta_i \\ &= \frac{1}{2} (\beta_i^2 X_i^\top X_i + 2\beta_i X_i^\top (X\bar{\beta} - y) + \|X\beta - y\|^2) + \lambda \|\bar{\beta}\|_1 + \lambda \beta_i. \end{aligned}$$

Prin urmare,

$$\frac{\partial J(\beta)}{\partial \beta_i} = X_i^\top X_i \beta + X_i^\top (X\bar{\beta} - y) + \lambda,$$

ceea ce înseamnă că valoarea optimă pentru β_i în acest caz este dată de expresia

$$\beta_i = \max \left\{ \frac{-X_i^\top (X\bar{\beta} - y) - \lambda}{X_i^\top X_i}, 0 \right\}.$$

În mod similar, în cazul $s_i = -1$ valoarea optimă care se obține pentru β_i este:

$$\beta_i = \min \left\{ \frac{-X_i^\top (X\bar{\beta} - y) + \lambda}{X_i^\top X_i}, 0 \right\}.$$

12. (Regresia liniară și fenomenul de overfitting: Adevărat sau Fals?)

Stanford, 2015 fall, Andrew Ng, midterm, pr. 1.a

Presupunem că, folosind un set [fix] format din m exemple, antrenați un model de regresie liniară, $h_\beta(x) = \beta^\top x$, unde β și $x \in \mathbb{R}^{d+1}$. După antrenare, vă

dați seama că *varianța* modelului dumneavoastră este relativ mare, deci se produce *supra-specializare* (engl., *overfitting*). Pentru fiecare dintre metodele listate mai jos, asociați fie răspunsul *Adevărat* în caz că metoda respectivă poate contracara / reduce fenomenul de overfitting, fie răspunsul *Fals* în cazul contrar. Explicați de ce.

- Extindeți vectorul de trăsături prin adăugarea de noi atrbute / trăsături.
- Impuneți condiția ca parametrul (de fapt, vectorul de parametri) β să urmeze o distribuție a priori, și anume o distribuție de forma $\mathcal{N}(0, \tau^2 I)$, iar apoi deduceți valoarea lui β prin metoda estimării de probabilitate maximă a posteriori (engl., maximum a posteriori probability, MAP).

Răspuns:

- Fals. Adăugarea de noi trăsături va face ca modelul dumneavoastră să fie și mai *complex*. El va reuși să se adapteze la și mai multe excepții (engl., *outliers*) din setul de antrenament și, în consecință să producă și mai mult overfitting.
- Adevărat. Impunând / cerând ca parametrul β să urmeze o distribuție probabilistă a priori [precum cea indicată în enunț], se limitează în mod efectiv norma vectorului β , fiindcă vectorii mai mari [în normă] au o probabilitate mai mică.³⁴³ Astfel, acest procedeu face ca modelul nostru să fie mai puțin predispus la overfitting.

1.1.2 Regresia logistică

13.

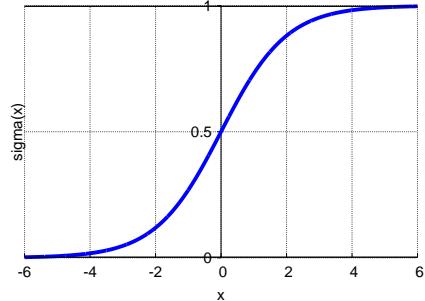
(Regresia logistică, chestiuni introductive:
estimare MLE; deducerea regulilor de actualizare a parametrilor,
folosind metoda gradientului)

*formulare de Liviu Ciortuz, după
■ □ ● ○ Stanford, 2016 spring, Chris Piech,
 Introduction to Probability for Computer Scientists course (CS109),
 Logistic Regression*

În învățarea automată, *algoritmii de clasificare* primesc ca date de intrare n instanțe de antrenament care sunt identice și independent distribuite, $(x^{(1)}, y^{(1)})$, $(x^{(2)}, y^{(2)})$, ..., $(x^{(n)}, y^{(n)})$, fiecare vector $x^{(i)}$ având d atrbute. În cazul de față vom presupune că $y^{(i)} \in \{0, 1\}$ pentru $i = 1, \dots, n$.

³⁴³Vedeți și problema 25.

Regresia logistică este un algoritm de clasificare, care lucrează urmărind să învețe o funcție care să aproximeze în mod convenabil distribuția $P(Y|X)$.³⁴⁴ Presupunerea ei de bază este că $P(Y|X)$ poate fi aproximată cu o funcție sigmoidală (numită și funcție logistică) aplicată unei combinații liniare de atributele de intrare.



Din punct de vedere matematic, dată fiind o (singură) instanță de antrenament (x, y) , regresia logistică va considera

$$P(Y = 1|X = x) = \sigma(z) \text{ și, echivalent, } P(Y = 0|X = x) = 1 - \sigma(z), \text{ unde} \quad (177)$$

$$\sigma(z) \stackrel{\text{def.}}{=} \frac{1}{1 + e^{-z}} = \frac{e^z}{1 + e^z}, \text{ cu}$$

$$z \stackrel{\text{not.}}{=} w_0 + \sum_{i=1}^d w_i x_i = w \cdot x \text{ și } w \stackrel{\text{not.}}{=} (w_0, w_1, \dots, w_d) \in \mathbb{R}^{d+1}, \text{ presupunând că}$$

vectorul x a fost extins cu componentă $x_0 = 1$.

Operatorul · desemnează produsul scalar al vectorilor.³⁴⁵ Pornind de la aceste formule pentru probabilitatea condiționată $P_{Y|X}$, putem crea un algoritm care selectează valori pentru vectorul w care maximizează această probabilitate pentru întreg setul de date de antrenament.

În acest exercițiu veți calcula mai întâi expresia funcției de log-verosimilitate [condițională] pentru datele $(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})$. Apoi vă vom arăta modul în care un algoritm iterativ bazat pe *metoda gradientului* poate să selecteze valorile optime pentru parametrii w . În final, veți descoperi cum anume se obțin *regulile de actualizare* a parametrilor din acest algoritm, folosind derivatele parțiale ale funcției de log-verosimilitate în raport cu componentele vectorului w .

a. Demonstrați că log-verosimilitatea condițională a întregului set de date de antrenament — ținând cont de *presupoziția de bază* a regresiei logisticice — este:³⁴⁶

$$\ell(w) = \sum_{i=1}^n \left(y^{(i)} \ln \sigma(w \cdot x^{(i)}) + (1 - y^{(i)}) \ln(1 - \sigma(w \cdot x^{(i)})) \right). \quad (178)$$

Observație: De fapt, funcția de *log-verosimilitate completă* este următoarea:

³⁴⁴La capitolul *Clasificare bayesiană* veți vedea că în acest tip de clasificare ideea de bază este că, dată fiind o instanță x , urmărим să alegem ca output acea etichetă y care maximizează probabilitatea condiționată $P(Y = y|X = x)$. Algoritmul Bayes Naiv este [tot] un algoritm de clasificare care modelează / aproximează probabilitatea $P(Y = y|X = x)$ folosind *presupoziția naivă* că toate trăsăturile / atributele sunt mutual independente în raport cu eticheta clasei. Regresia logistică nu folosește o astfel de presupozitie.

³⁴⁵În notație matricială, dacă ne referim la w și x ca vectori-colonă din același spațiu vectorial, am putea scrie $w^\top x$ în loc de $w \cdot x$. Operatorul \top desemnează operația de transpunere a matricelor / vectorilor.

³⁴⁶*Observație importantă:* Expresia acestei log-verosimilități condiționale, adică partea dreaptă a egalității (178) are forma unei cross-entropii, cu semnul schimbat. (De fapt, căte o cross-entropie (cu semnul schimbat) pentru fiecare instanță.) Pentru *definiția noțiunii de cross-entropie*, vedeți problema 64 de la capitolul de *Fundamente*. Valorile $y^{(i)}$ și $1 - y^{(i)}$ corespund unei distribuții de probabilitate, iar valorile $\sigma(w \cdot x^{(i)})$ și $1 - \sigma(w \cdot x^{(i)})$ corespund altrei distribuții de probabilitate (și anume, cea calculată de regresia logistică), care aproximează prima distribuție.

$$\begin{aligned}
\text{log-likelihood} &= \ln \prod_{i=1}^n P(x^{(i)}, y^{(i)}) = \ln \prod_{i=1}^n (P_{Y|X}(y^{(i)}|x^{(i)}) P_X(x^{(i)})) \\
&= \ln \left(\left(\prod_{i=1}^n P_{Y|X}(y^{(i)}|x^{(i)}) \right) \cdot \left(\prod_{i=1}^n P_X(x^{(i)}) \right) \right) \\
&= \ln \prod_{i=1}^n P_{Y|X}(y^{(i)}|x^{(i)}) + \ln \prod_{i=1}^n P_X(x^{(i)}) \stackrel{\text{not.}}{=} \ell(w) + \ell_x.
\end{aligned}$$

Observăm că ℓ_x nu depinde de parametrul w . Întrucât facem estimarea în sensul verosimilității maxime (MLE) a parametrului w , putem să ne limităm la a maximiza $\ell(w)$.

Comentariu: Dispunând de expresia (178) pentru funcția de log-verosimilitate condițională, putem să trecem la identificarea acelor valori ale lui w care maximizează această funcție. Spre deosebire de alte situații în care se pune(a) problema găsirii optimului funcției de log-verosimilitate, în acest caz nu există o *formulă analitică* (engl., closed form formula) care să ne permită să obținem în mod direct valoarea optimă a lui w . De aceea, vom identifica această valoare utilizând o *metodă de optimizare*. Mai precis, vom folosi un algoritm iterativ numit *gradientul ascendent*.³⁴⁷ Ideea acestui algoritm este că atunci când facem în mod continuu pași mici în direcția gradientului funcției de optimizat — gradientul fiind, conform definiției, vectorul de derivate parțiale al funcției respective —, vom ajunge la un moment dat într-un punct de maxim local al acelei funcții. În cazul regresiei logistice, se poate demonstra că rezultatul aplicării metodei / algoritmului gradientului ascendent va fi întotdeauna punctul de *maxim global*.³⁴⁸ Pașii mici pe care îi parcurem în mod iterativ, dat fiind setul de date de antrenament, sunt calculați / determinați folosind următoarea *regulă de actualizare*:

$$w_j^{new} = w_j^{old} + \eta \frac{\partial}{\partial w_j^{old}} \ell(w^{old}), \quad (179)$$

unde η este un factor constant care determină mărimea pasului pe care îl facem la fiecare iterație, cunoscut sub numele de *rată de învățare*.

b. Arătați că forma derivatelor parțiale ale funcției de log-verosimilitate condițională în raport cu fiecare componentă w_j a vectorului w este următoarea:

$$\frac{\partial}{\partial w_j} \ell(w) = \sum_{i=1}^n [y^{(i)} - \underbrace{\sigma(w \cdot x^{(i)})}_{P(Y=1|X=x;w)}] x_j^{(i)} \text{ pentru } j = 0, 1, \dots, d. \quad (180)$$

Sugestie: Puteți folosi următoarea *proprietate* pentru derivata funcției logistice σ în raport cu argumentul ei:

$$\frac{\partial}{\partial z} \sigma(z) = \sigma(z)[1 - \sigma(z)] \text{ pentru } \forall z \in \mathbb{R}.$$

Observație: Din relația (180) urmează că vectorul gradient $\nabla_w \ell(w)$ se poate scrie astfel:

$$\nabla_w \ell(w) = \sum_{i=1}^n [y^{(i)} - \sigma(w \cdot x^{(i)})] x^{(i)}. \quad (181)$$

³⁴⁷Pentru o ilustrare simplă a modului cum funcționează algoritmul gradientului *descendent*, vedeți problema 80.c de la capitolul de *Fundamente*.

³⁴⁸Vedeți pr. 14.

La problema 16 vom arăta că $\nabla_w \ell(w)$ se poate scrie sub o formă și mai compactă, folosind așa-numita *matrice de design* X .³⁴⁹

c. [Legătura cu *funcția de cost logistică*]³⁵⁰

Arătați că dacă în loc de $y^{(i)} \in \{0, 1\}$ vom considera $y'^{(i)} \in \{-1, 1\}$ pentru $i = 1, \dots, n$,³⁵¹ atunci se poate demonstra că următoarea *funcție de cost mediu* (sau de *risc empiric*; engl., empirical risk)

$$J(w) \stackrel{\text{def.}}{=} \frac{1}{n} \sum_{i=1}^n \ln(1 + \exp(-y'^{(i)} w \cdot x^{(i)})) \quad (182)$$

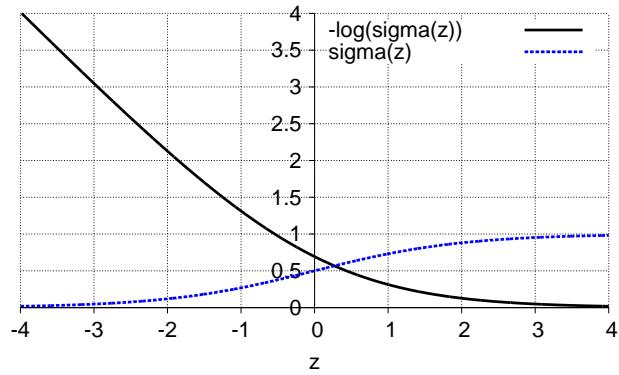
este chiar $-\frac{1}{n} \ell(w)$, unde $\ell(w)$ este funcția de log-verosimilitate condițională din relația (178). Ca o *consecință*, a maximiza funcția de log-verosimilitate $\ell(w)$ este echivalent cu a minimiza funcția de cost mediu $J(w)$.

Observație: Funcția $\phi(y'w \cdot x) \stackrel{\text{not.}}{=} \ln(1 + \exp(-y'w \cdot x))$ sau, mai simplu, $\phi(z) \stackrel{\text{not.}}{=} \ln(1 + e^{-z})$ este numită *funcția de cost* (sau, pierdere) *logistica*.³⁵²

Legătura dintre $\phi(z) = \ln(1 + e^{-z})$ pe de o parte și (atenție!) *funcția logistică* (sau *funcția sigmoidală*) $\sigma(z) \stackrel{\text{def.}}{=} \frac{1}{1 + e^{-z}}$ pe de altă parte este următoarea:

$$\phi(z) = \ln(1 + e^{-z}) = -\ln \sigma(z).$$

Din această cauză, funcția de cost logistică este numită uneori și *funcția log-sigmoidală*.



Răspuns:

a. Pentru început, iată o modalitate de a scrie în mod compact expresia (177) din enunț, care reprezintă probabilitatea condiționată a unei *singure* instanțe de antrenament:³⁵³

$$P(Y = y | X = x) = \sigma(w \cdot x)^y [1 - \sigma(w \cdot x)]^{1-y}, \text{ presupunând că } y \in \{0, 1\}.$$

Întrucât toate instanțele de antrenament sunt independente, verosimilitatea condițională a întregului set de date este următoarea:

$$\prod_{i=1}^n P(Y = y^{(i)} | X = x^{(i)}) = \prod_{i=1}^n \left(\sigma(w \cdot x^{(i)})^{y^{(i)}} [1 - \sigma(w \cdot x^{(i)})]^{1-y^{(i)}} \right). \quad (183)$$

³⁴⁹Acolo de fapt se lucrează cu o formă mai generală de regresie logistică, și anume *regresia logistică local-ponderată*.

³⁵⁰Cf. secțiunii 2 (Logistic Regression) din documentul *Supplemental lecture notes* (pentru cursul de *Machine Learning*), de John Duchi, de la Universitatea Stanford.

³⁵¹Definim $y'^{(i)} = -1$ dacă $y^{(i)} = 0$ și $y'^{(i)} = 1$ dacă $y^{(i)} = 1$.

³⁵²Atenție! În acest context, $z \stackrel{\text{not.}}{=} y'w \cdot x$. În enunț, la început (când $y \in \{0, 1\}$), am avut $z \stackrel{\text{not.}}{=} w \cdot x$.

³⁵³Acest procedeu este numit *artificiul exponentierii* (engl., the exponentiation trick).

Dacă aplicăm logaritmul natural acestei funcții, obținem imediat expresia indicată în enunț (178) pentru log-verosimilitatea datelor în cazul regresiei logistice.

b. Calculul derivatelor parțiale ale expresiei $\sigma(w \cdot x)$ în raport cu componentele vectorului w va fi făcut folosind formula de derivare pentru funcții compuse.

Concret, valoarea derivatei parțiale a lui $\sigma(w \cdot x)$ în raport cu componenta w_j a vectorului w , pentru o instanță (x, y) , se calculează astfel:

$$\begin{aligned}
& \frac{\partial}{\partial w_j} \ln[\sigma(w \cdot x)^y [1 - \sigma(w \cdot x)]^{1-y}] \\
&= \frac{\partial}{\partial w_j} y \ln \sigma(w \cdot x) + \frac{\partial}{\partial w_j} (1-y) \ln [1 - \sigma(w \cdot x)] \\
&= \left[\frac{y}{\sigma(w \cdot x)} - \frac{1-y}{1-\sigma(w \cdot x)} \right] \frac{\partial}{\partial w_j} \sigma(w \cdot x) \\
&= \left[\frac{y}{\sigma(w \cdot x)} - \frac{1-y}{1-\sigma(w \cdot x)} \right] \sigma(w \cdot x) [1 - \sigma(w \cdot x)] x_j \\
&= \frac{y - \sigma(w \cdot x)}{\sigma(w \cdot x) [1 - \sigma(w \cdot x)]} \sigma(w \cdot x) [1 - \sigma(w \cdot x)] x_j \\
&= [y - \sigma(w \cdot x)] x_j. \tag{184}
\end{aligned}$$

Întrucât derivata sumei este suma derivatelor, rezultă că derivata parțială a funcției de log-verosimilitate condițională $\ell(w)$ în raport cu w_j este suma acestui tip de termeni, câte unul pentru fiecare instanță de antrenament. Mai exact, după ce aplicăm funcția ln expresiei (183) și apoi calculăm derivele ei parțiale în raport cu w_j (pentru $j \in \{0, 1, \dots, d\}$), vom obține rezultatul (180), datorită relației (184), pe care am demonstrat-o mai sus.

c. Demonstrația se bazează în esență pe proprietatea $1 - \sigma(z) = \sigma(-z)$, care este imediată.

Cazul $y'^{(i)} = 1$: $P(Y = 1 | X = x^{(i)}) = \sigma(w \cdot x^{(i)}) = \sigma(y'^{(i)} w \cdot x^{(i)})$;

Cazul $y'^{(i)} = -1$: $P(Y' = -1 | X = x^{(i)}) = 1 - P(Y = 1 | X = x^{(i)}) = 1 - \sigma(w \cdot x^{(i)}) = \sigma(-w \cdot x^{(i)}) = \sigma(y'^{(i)} w \cdot x^{(i)})$.

Prin urmare,

$$\begin{aligned}
\ell(w) &= \ln \prod_{i=1}^n P(Y = y^{(i)} | X = x^{(i)}) = \sum_{i=1}^n \ln P(Y = y^{(i)} | X = x^{(i)}) \\
&= \sum_{i=1}^n \ln P(Y' = y'^{(i)} | X = x^{(i)}) = \sum_{i=1}^n \ln \sigma(y'^{(i)} w \cdot x^{(i)}) \\
&= \sum_{i=1}^n \ln \frac{1}{1 + \exp(-y'^{(i)} w \cdot x^{(i)})} = - \sum_{i=1}^n \ln(1 + \exp(-y'^{(i)} w \cdot x^{(i)})) = - \sum_{i=1}^n \phi(y'^{(i)} w \cdot x^{(i)}).
\end{aligned}$$

Așadar, $-\frac{1}{n} \ell(w) = J(w)$.

14.

(Regresia logistică, o proprietate importantă:
funcția de log-verosimilitate este concavă, deci are un maxim global;
calculul matricei hessiene pentru funcția de log-verosimilitate)

■ □ • ○ Stanford, 2008 fall, Andrew Ng, HW1, pr. 1.a

La problema 13.a am demonstrat că în cazul regresiei logistice funcția de log-verosimilitate condițională se scrie sub forma următoare:

$$\ell(w) = \sum_{i=1}^n \left(y^{(i)} \ln h(x^{(i)}) + (1 - y^{(i)}) \ln(1 - h(x^{(i)})) \right),$$

unde instanțele $x^{(i)}$ sunt vectori-coloană de forma $x = (x_1, \dots, x_d)^\top$, etichetele $y^{(i)} \in \{0, 1\}$, vectorul de ponderi $w \in \mathbb{R}^d$, iar $h(x) \stackrel{\text{def.}}{=} \frac{1}{1 + \exp(-w \cdot x)}$.

a. Calculați matricea hessiană (H) a acestei funcții, care este prin definiție matricea formată din derivatele parțiale de ordin secund ale funcției ℓ .

b. Arătați că pentru orice $z = (z_1, \dots, z_d)^\top$, are loc proprietatea

$$z^\top H z \leq 0.$$

Observație: Aceasta este una dintre modalitățile clasice de a arăta că matricea H este negativ semidefinită, ceea ce se notează sub forma $H \leq 0$. Aceasta implică faptul că ℓ este funcție concavă și, în consecință, ea nu are decât o singură valoare maximă (care este și maxim global).

Răspuns:

a. Vă reamintim o proprietate importantă a funcției logistice: $\sigma'(z) = \sigma(z)(1 - \sigma(z))$. Prin urmare, $\frac{\partial h(x)}{\partial w_k} = h(x)(1 - h(x)) x_k$, fiindcă $h(x) = \sigma(w \cdot x)$. Acest fapt este utilizat în cele ce urmează.

Calculăm mai întâi derivatele parțiale de ordinul întâi ale funcției de log-verosimilitate:

$$\begin{aligned} \frac{\partial \ell(w)}{\partial w_k} &= \frac{\partial}{\partial w_k} \left(\sum_{i=1}^n y^{(i)} \ln h(x^{(i)}) + (1 - y^{(i)}) \ln(1 - h(x^{(i)})) \right) \\ &= \sum_{i=1}^n \left[y^{(i)} \frac{1}{h(x^{(i)})} - (1 - y^{(i)}) \frac{1}{1 - h(x^{(i)})} \right] h(x^{(i)}) (1 - h(x^{(i)})) x_k^{(i)} \\ &= \sum_{i=1}^n \left[y^{(i)} - \cancel{y^{(i)} h(x^{(i)})} - h(x^{(i)}) + \cancel{y^{(i)} h(x^{(i)})} \right] x_k^{(i)} \\ &= \sum_{i=1}^n (y^{(i)} - h(x^{(i)})) x_k^{(i)}. \end{aligned} \tag{185}$$

În consecință, elementul generic al matricei hessiene (H) este

$$H_{lk} = \frac{\partial^2 \ell(w)}{\partial w_l \partial w_k} = \sum_{i=1}^n -\frac{\partial h(x^{(i)})}{\partial w_l} x_k^{(i)} = \sum_{i=1}^n -h(x^{(i)}) (1 - h(x^{(i)})) x_l^{(i)} x_k^{(i)}. \tag{186}$$

Tinând cont de faptul că putem scrie xx^\top ca o matrice X [dacă și numai dacă $X_{ij} = x_i x_j$], rezultă că matricea hessiană H este:³⁵⁴

$$H = - \sum_{i=1}^n \underbrace{h(x^{(i)})(1-h(x^{(i)}))}_{\in \mathbb{R}_+} x^{(i)}(x^{(i)})^\top. \quad (187)$$

b. Pentru a demonstra că matricea H este negativ semidefinită, vom arăta că pentru orice vector-coloană z are loc inegalitatea $z^\top H z \leq 0$.

$$\begin{aligned} z^\top H z &= -z^\top \left(\sum_{i=1}^n h(x^{(i)})(1-h(x^{(i)}))x^{(i)}(x^{(i)})^\top \right) z \\ &= -\sum_{i=1}^n h(x^{(i)})(1-h(x^{(i)}))z^\top x^{(i)} \underbrace{(x^{(i)})^\top z}_{(z^\top x^{(i)})^\top} \\ &= -\sum_{i=1}^n h(x^{(i)})(1-h(x^{(i)}))(z^\top x^{(i)})^2 \leq 0. \end{aligned}$$

Ultima inegalitate de mai sus se justifică astfel: $(z^\top x^{(i)})^2 \geq 0$ pentru orice z și orice $x^{(i)}$ din \mathbb{R}^d , iar din definiția funcției logistice σ rezultă $0 < h(x^{(i)}) < 1$, ceea ce implică $h(x^{(i)})(1-h(x^{(i)})) > 0$.

Observație importantă: Demonstrația de mai sus furnizează tot ce este necesar pentru obținerea [ulterioară a] relației de actualizare a parametrilor la aplicarea metodei lui Newton în cazul regresiei logistice.³⁵⁵

15. (Regresia logistică:
estimarea parametrilor în sens MAP — regularizare de normă L_2 ;
efectul de *diminuare a ponderilor* (engl., weight decay))
prelucrare de Liviu Ciortuz, după
 • Stanford, 2008 fall, Andrew Ng, HW3, pr. 4

Presupunem că folosim un model de *regresie logistică* $h_\theta(x) = \sigma(\theta \cdot x)$, unde σ este *funcția logistică* (sau *sigmoidală*), și că dispunem de un set de date de antrenament $D = \{(x^{(i)}, y^{(i)}); i = 1, \dots, n\}$ definit ca de obicei.

Remember: Estimarea de verosimilitate maximă a parametrilor θ este definită astfel:

$$\theta_{ML} \stackrel{\text{def.}}{=} \underset{\theta}{\operatorname{argmax}} P(D|\theta) \stackrel{i.i.d.}{=} \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^n P(y^{(i)}|x^{(i)}; \theta). \quad (188)$$

Dacă dorim să *regularizăm* regresia logistică, putem impune ca [vectorul de] parametri θ să urmeze o distribuție probabilistă a priori P . Estimarea de probabilitate maximă a

³⁵⁴La problema 16 vom arăta că H se poate scrie și mai compact, folosind așa-numita *matrice de design* X . (Vezi observă că coloanele de fapt se lucrează cu o formă mai generală de regresie logistică, și anume *regresia logistică local-ponderată*.)

³⁵⁵Pentru o descriere a metodei lui Newton, vezi *Comentariul* din enunțul problemei 80 de la capitolul de *Fundamente*. În prezentul capitol, la problema 7 se arată cum anume se aplică metoda lui Newton în cazul regresiei liniare, iar la problema 16 se calculează vectorul gradient și matricea hessiană pentru o variantă ușor generalizată (și anume, *local-ponderată* și *regularizată*) a regresiei logistice.

posteriori pentru θ este definită astfel:

$$\theta_{MAP} \stackrel{\text{def.}}{=} \underset{\theta}{\operatorname{argmax}} P(\theta|D) \stackrel{F.B.}{=} \underset{\theta}{\operatorname{argmax}} \frac{P(D|\theta)P(\theta)}{P(D)} = \underset{\theta}{\operatorname{argmax}} P(D|\theta)P(\theta) \quad (189)$$

$$\stackrel{i.i.d.}{=} \underset{\theta}{\operatorname{argmax}} P(\theta) \prod_{i=1}^m P(y^{(i)}|x^{(i)}; \theta). \quad (190)$$

Presupunem că alegem ca distribuție a priori pentru θ distribuția gaussiană multidimensională $\mathcal{N}(0, \tau^2 I)$, cu $\tau > 0$ și I matricea identitate de dimensiune $(d+1) \times (d+1)$.³⁵⁶

a. Arătați că atunci când parametrul θ urmează o distribuție probabilistă de genul celei indicate mai sus, efectul practic este următorul: maximizarea funcției de probabilitate a posteriori $P(\theta|D)$ este echivalentă cu maximizarea sumei dintre funcția de log-verosimilitate $\ln P(D|\theta)$ și un termen de forma $-\lambda\|\theta\|^2$, unde $\lambda = \frac{1}{2\tau^2}$. (Astfel se justifică denumirea de *regularizare de normă L₂* sau, mai simplu, *regularizare L₂*.)

b. Demonstrați următoarea inegalitate:

$$\|\theta_{MAP}\|^2 \leq \|\theta_{ML}\|^2.$$

Răspuns:

a. Tinând cont de faptul că funcția \ln este strict creșătoare pe întreg domeniul ei de definiție, putem rescrie egalitatea reprezentată de primul și ultimul termen din relația (188) sub forma următoare:

$$\theta_{ML} = \underset{\theta}{\operatorname{argmax}} \ln \prod_{i=1}^n P(y^{(i)}|x^{(i)}; \theta) = \underset{\theta}{\operatorname{argmax}} \underbrace{\sum_{i=1}^n \ln P(y^{(i)}|x^{(i)}; \theta)}_{\text{not.: } \ell(\theta)}.$$

Similar, rescriem relația (190) sub forma

$$\begin{aligned} \theta_{MAP} &= \underset{\theta}{\operatorname{argmax}} \ln P(\theta) \prod_{i=1}^m P(y^{(i)}|x^{(i)}; \theta) = \underset{\theta}{\operatorname{argmax}} (\ln P(\theta) + \sum_{i=1}^m \ln P(y^{(i)}|x^{(i)}; \theta)) \\ &= \underset{\theta}{\operatorname{argmax}} (\ln P(\theta) + \ell(\theta)) \\ &= \underset{\theta}{\operatorname{argmax}} \left(\ln \left(\frac{1}{(2\pi)^{(d+1)/2} \tau^{d+1}} \exp \left(-\frac{1}{2} \theta^\top \frac{1}{\tau^2} I \theta \right) \right) + \ell(\theta) \right) \\ &= \underset{\theta}{\operatorname{argmax}} \left(-\ln \left((2\pi)^{(d+1)/2} \tau^{d+1} \right) - \frac{1}{2\tau^2} \theta^2 + \ell(\theta) \right). \end{aligned}$$

Întrucât expresia $(2\pi)^{(d+1)/2} \tau^{d+1}$ nu depinde de θ , rezultă că

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}} \left(- \underbrace{\frac{1}{2\tau^2} \theta^2}_{\text{not.: } \lambda} + \ell(\theta) \right) = \underset{\theta}{\operatorname{argmax}} \left(\ell(\theta) - \lambda \|\theta\|^2 \right). \quad (191)$$

³⁵⁶Am scris $d+1$ în loc de d fiindcă se consideră că se lucrează și cu termenul liber θ_0 , deci $\theta = (\theta_0, \theta_1, \dots, \theta_d)$.

b. Presupunem prin reducere la absurd că $\|\theta_{MAP}\|^2 > \|\theta_{ML}\|^2$. Rezultă că

$$\begin{aligned} P(\theta_{MAP}) &= \frac{1}{(2\pi)^{\frac{d+1}{2}} |\tau^2 I|^{\frac{1}{2}}} \exp\left(-\frac{1}{2\tau^2} \|\theta_{MAP}\|^2\right) \\ &< \frac{1}{(2\pi)^{\frac{d+1}{2}} |\tau^2 I|^{\frac{1}{2}}} \exp\left(-\frac{1}{2\tau^2} \|\theta_{ML}\|^2\right) \\ &= P(\theta_{ML}). \end{aligned}$$

În consecință,

$$\begin{aligned} P(\theta_{MAP})P(D|\theta_{MAP}) &< P(\theta_{ML})P(D|\theta_{MAP}) \\ &\leq P(\theta_{ML})P(D|\theta_{ML}). \end{aligned}$$

Ultima inegalitate de mai sus are loc întrucât prin definiție θ_{ML} maximizează verosimilitatea datelor, $P(D|\theta)$. Însă rezultatul $P(\theta_{MAP})P(D|\theta_{MAP}) < P(\theta_{ML})P(D|\theta_{ML})$ constituie o contradicție, fiindcă θ_{MAP} prin definiție maximizează produsul $P(\theta)P(D|\theta)$. Prin urmare, presupunerea $\|\theta_{MAP}\|^2 > \|\theta_{ML}\|^2$ este falsă.

Observații:

1. Remarcați faptul că în vreme ce la regresia liniară termenul de regularizare era precedat de semnul + (vedeți relația (152) de la problema 3.C), în relația (191) termenul de regularizare este precedat de semnul -. Explicația rezidă în faptul că la regresia liniară — varianta LSE — se caută *minimul* sumei pătratelor erorilor, iar la regresia logistică se caută *maximul* funcției de verosimilitate condițională.
2. Inegalitatea $\|\theta_{MAP}\|^2 \leq \|\theta_{ML}\|^2$ indușă de regularizarea de normă L_2 este numită uneori și proprietatea de *diminuare a ponderilor* (engl., *weight decay*), fiindcă regularizarea aceasta încurajează ponderile / parametrii să ia valori care în general sunt mai mici (în valoare absolută) decât în cazul în care nu se folosește regularizare, ceea ce conduce la limitarea / prevenirea fenomenului de *overfitting*. Veți observa din demonstrație că această proprietate indușă de regularizarea L_2 este valabilă în general, nu doar pentru regresia logistică.³⁵⁷

16.

(Regresia logistică local-ponderată regularizată (L_2): exprimarea vectorului gradient și a matricei hessiene cu ajutorul matricei de design, X)

prelucrare de Liviu Ciortuz, după

Stanf ord, 2007 fall, Andrew Ng, HW1, pr. 2

Conform relațiilor (178) și (191) de la problemele 13 și respectiv 15, problema de regresie logistică cu regularizare L_2 constă în a maximiza o funcție de log-verosimilitate condițională extinsă cu un termen de regularizare, de forma

$$\sum_{i=1}^n [y^{(i)} \ln h_\theta(x^{(i)}) + (1 - y^{(i)}) \ln(1 - h_\theta(x^{(i)}))] - \frac{\lambda}{2} \|\theta\|^2, \quad (192)$$

³⁵⁷Mentionăm că în contextul regresiei liniare, proprietatea de diminuarea a ponderilor apare la problema 25.

unde $x^{(i)} \in \mathbb{R}^d$, $y^{(i)} \in \{0, 1\}$, $w_i \in \mathbb{R}_+$ pentru $i = 1, \dots, n$, $\theta \in \mathbb{R}^d$, iar h_θ desemnează o funcție de tip sigmoidal / logistic definită prin $h_\theta(x) = \frac{1}{1 + e^{-\theta \cdot x}}$ pentru orice $x \in \mathbb{R}^d$.³⁵⁸

Făcând o ușoară *generalizare*, vom putea spune că problema de regresie logistică *local-ponderată* cu regularizare L_2 constă în a maximiza o funcție de de forma

$$\ell(\theta) = \sum_{i=1}^n w_i [y^{(i)} \ln h_\theta(x^{(i)}) + (1 - y^{(i)}) \ln(1 - h_\theta(x^{(i)}))] - \frac{\lambda}{2} \|\theta\|^2,$$

unde ponderea w_i reprezintă „importanța“ acordată instanței de antrenament $(x^{(i)}, y^{(i)})$ în determinarea separatorului decizional produs de regresia logistică pe aceste date. Separatorul decizional determinat de acest model, și anume $\theta \cdot x = 0$, va fi tot liniar. Remarcați faptul că $-\frac{\lambda}{2} \|\theta\|^2 = -\frac{\lambda}{2} \theta^2$ din expresiile de mai sus este termenul de regularizare; el este necesar pentru a contracara fenomenul de *overfitting* (rom., supra-specializare).

a. Demonstrați că vectorul gradient pentru $\ell(\theta)$ este

$$\nabla_\theta \ell(\theta) = X^\top z - \lambda \theta, \quad (193)$$

unde X este *matricea de design* (care are dimensiunea $n \times d$, fiecare linie a acestei matrice fiind reprezentată de câte o instanță de antrenament), iar vectorul-coloană $z \in \mathbb{R}^n$ este de forma $(z_1, \dots, z_n)^\top$, cu

$$z_i = w_i(y^{(i)} - h_\theta(x^{(i)})), \text{ pentru } i = 1, \dots, n.$$

b. Demonstrați că matricea hessiană pentru $\ell(\theta)$ este

$$H = X^\top D X - \lambda I,$$

unde $D \in \mathbb{R}^{n \times n}$ este o matrice diagonală, în care elementul generic (notat cu D_{ii}) este

$$D_{ii} = -w_i h_\theta(x^{(i)})(1 - h_\theta(x^{(i)})).$$

Observație: Dacă se consideră $w_i = 1$ pentru $i = 1, \dots, m$ și se elimină termenul corespunzător regularizării — adică, se revine la modelul de regresie logistică introdus la problema 13 —, atunci expresia (193) va corespunde expresiei (180) de la problema 13 (și totodată expresiei (185) de la problema 14). De asemenea, elementul generic al matricei H de mai sus va corespunde expresiei (186) de la problema 14. În practică, folosirea matricei de design X este recomandată a fi folosită ori de câte ori este posibil, din motive de eficiență computațională.

Răspuns:³⁵⁹

a. Mai întâi vom calcula expresia termenului generic pentru vectorul gradient $\nabla_\theta \ell(\theta)$, generalizând ușor raționamentul de la problema 14, adică adăugând

³⁵⁸Vectorului w de la problemele 13 și 14 îi corespunde aici vectorul θ .

³⁵⁹Soluția care urmează a fost redactată într-o formă inițială de către Mădălina Racoviță, MSc student, UAIC, Iași, 2020.

ponderile w_i , precum și termenul de regularizare $-\frac{\lambda}{2}\theta^2$. Vă reamintim că $\frac{\partial}{\partial\theta_k}h_\theta(x) = h_\theta(x)(1 - h_\theta(x))x_k$.

$$\begin{aligned}
 \frac{\partial}{\partial\theta_k}\ell(\theta) &= \sum_{i=1}^m w_i \left[y^{(i)} \frac{\partial}{\partial\theta_k} \ln(h_\theta(x^{(i)})) + (1 - y^{(i)}) \frac{\partial}{\partial\theta_k} \ln(1 - h_\theta(x^{(i)})) \right] - \frac{\lambda}{2} \frac{\partial}{\partial\theta_k} \theta^2 \\
 &= \sum_{i=1}^m w_i \left[y^{(i)} \cdot \underbrace{\frac{1}{h_\theta(x^{(i)})}}_{1-h_\theta(x^{(i)})} \cdot h_\theta(x^{(i)}) (1 - h_\theta(x^{(i)})) \cdot x_k^{(i)} + \right. \\
 &\quad \left. + (1 - y^{(i)}) \cdot \underbrace{\frac{1}{1-h_\theta(x^{(i)})}}_{h_\theta(x^{(i)})} \cdot (-1) \cdot h_\theta(x^{(i)}) \underbrace{(1-h_\theta(x^{(i)}))}_{h_\theta(x^{(i)})} \cdot x_k^{(i)} \right] - \frac{\lambda}{2} \cdot 2\theta_k \\
 &= \sum_{i=1}^m w_i \left[y^{(i)} \cdot (1 - h_\theta(x^{(i)})) x_k^{(i)} - (1 - y^{(i)}) \cdot h_\theta(x^{(i)}) x_k^{(i)} \right] - \lambda\theta_k \\
 &= \sum_{i=1}^m w_i \left(y^{(i)} x_k^{(i)} - \underbrace{y^{(i)} h_\theta(x^{(i)}) x_k^{(i)}}_{z_i} - h_\theta(x^{(i)}) x_k^{(i)} + \underbrace{h_\theta(x^{(i)}) x_k^{(i)}}_{z_i} \right) - \lambda\theta_k \\
 &= \sum_{i=1}^m \underbrace{w_i (y^{(i)} - h_\theta(x^{(i)}))}_{z_i} x_k^{(i)} - \lambda\theta_k \\
 &= \sum_{i=1}^m z_i x_k^{(i)} - \lambda\theta_k, \text{ pentru } k = 1, \dots, d.
 \end{aligned}$$

În consecință, vectorul gradient, $\nabla_\theta\ell(\theta)$, se scrie astfel:

$$\nabla_\theta\ell(\theta) = \sum_{i=1}^m z_i x^{(i)} - \lambda\theta.$$

Acum vom arăta că $\nabla_\theta\ell(\theta) = X^\top z - \lambda\theta$, unde X este *matricea de design*. Conform rezultatului demonstrat mai sus, va fi suficient să arătăm că $X^\top z = \sum_{i=1}^m z_i x^{(i)}$. Într-adevăr,

$$\begin{aligned}
 X^\top z &= (x^{(1)\top}, \dots, x^{(m)\top}) \times (z_1, \dots, z_m)^\top = \begin{pmatrix} x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(m)} \\ \vdots & \vdots & & \vdots \\ x_d^{(1)} & x_d^{(2)} & \dots & x_d^{(m)} \end{pmatrix} \times \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_m \end{pmatrix} \\
 &= \left(\sum_{i=1}^m z_i x_1^{(i)}, \dots, \sum_{i=1}^m z_i x_d^{(i)} \right)^\top = \sum_{i=1}^m z_i x^{(i)}.
 \end{aligned}$$

b. Mai întâi vom calcula expresia termenului generic al matricei hessiene pentru funcția de log-verosimilitate condițională regularizată, $\ell(\theta)$.

$$\begin{aligned}
 H_{k,l} &= \frac{\partial^2}{\partial\theta_k\partial\theta_l}\ell(\theta) = \frac{\partial}{\partial\theta_l} \left[\sum_{i=1}^m w_i (y^{(i)} - h_\theta(x^{(i)})) x_k^{(i)} - \lambda\theta_k \right] \\
 &= \sum_{i=1}^m w_i \cdot (-1) \cdot \left(\frac{\partial}{\partial\theta_l} h_\theta(x^{(i)}) \right) \cdot x_k^{(i)} - \lambda \frac{\partial}{\partial\theta_l} \theta_k
 \end{aligned}$$

$$= - \sum_{i=1}^m w_i h_\theta(x^{(i)})(1 - h_\theta(x^{(i)})) x_l^{(i)} x_k^{(i)} - \lambda \frac{\partial}{\partial \theta_l} \theta_k,$$

pentru orice $k, l \in \{1, \dots, d\}$.

Așadar,

$$H_{k,l} = \begin{cases} - \sum_{i=1}^m w_i h_\theta(x^{(i)})(1 - h_\theta(x^{(i)})) (x_k^{(i)})^2 - \lambda & \text{pentru } k = l, \\ - \sum_{i=1}^m w_i h_\theta(x^{(i)})(1 - h_\theta(x^{(i)})) x_l^{(i)} x_k^{(i)} & \text{pentru } k \neq l. \end{cases}$$

Folosind notația $D_{ii} = -w_i h_\theta(x^{(i)})(1 - h_\theta(x^{(i)}))$ dată în enunț, putem scrie:

$$H_{k,l} = \begin{cases} \sum_{i=1}^m D_{ii} (x_k^{(i)})^2 - \lambda & \text{pentru } k = l, \\ \sum_{i=1}^m D_{ii} x_l^{(i)} x_k^{(i)} & \text{pentru } k \neq l. \end{cases}$$

Similar cu modul cum am procedat la punctul a, vom arăta că matricea H este egală cu $X^\top D X - \lambda I$, ceea ce revine la a demonstra că $X^\top D X = H + \lambda I$.

$$\begin{aligned} X^\top D X &= (x^{(1)}, \dots, x^{(m)}) \times \begin{bmatrix} D_{11} & \dots & 0 \\ & \ddots & \\ 0 & \dots & D_{mm} \end{bmatrix} \times (x^{(1)}, \dots, x^{(m)})^\top \\ &= (D_{11} x^{(1)}, \dots, D_{mm} x^{(m)}) \times (x^{(1)}, \dots, x^{(m)})^\top = \sum_{i=1}^m D_{ii} x^{(i)} x^{(i)^\top} \\ &= \begin{bmatrix} \sum_{i=1}^m D_{ii} \cdot (x_1^{(i)})^2 & \sum_{i=1}^m D_{ii} \cdot x_1^{(i)} x_2^{(i)} & \dots \\ \sum_{i=1}^m D_{ii} \cdot x_2^{(i)} x_1^{(i)} & \sum_{i=1}^m D_{ii} \cdot (x_2^{(i)})^2 & \dots \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^m D_{ii} \cdot x_d^{(i)} x_1^{(i)} & \dots & \sum_{i=1}^m D_{ii} \cdot (x_d^{(i)})^2 \end{bmatrix} \\ &= H + \lambda I. \end{aligned}$$

17.

(Regresia logistică: kernel-izare
în cazul folosirii metodei gradientului [ascendent])

*prelucrare de Liviu Ciortuz, după
■ □ • ○ CMU, 2005 fall, Tom Mitchell, HW3, pr. 2.ab*

Introducere: Ideea centrală din spatele metodei kernel-izării, aşa cum este folosită în clasificare,³⁶⁰ este transformarea / „maparea“ vectorilor care reprezintă instanțe de antrenament, provenind dintr-un spațiu \mathcal{X} care [în mod normal] are un număr mic de dimensiuni, într-un alt spațiu \mathcal{Z} (numit îndeobște *spațiu de trăsături*), care are [de obicei] un număr mare de dimensiuni. O astfel de transformare poate da naștere [în anumite condiții] unui clasificator mai flexibil, care moștenește însă eficiența computațională din spațiul inițial \mathcal{X} . Mai concret, *kernel-izarea* implică găsirea unei transformări (sau, a unei mapări, de la engl. mapping) $\phi : \mathcal{X} \rightarrow \mathcal{Z}$, astfel încât

³⁶⁰ Metoda kernel-izării a fost introdusă inițial pentru mașinile cu vectori-suport (engl., Support Vector Machines, SVM). Puteți consulta capitolul respectiv din prezenta culegere și (eventual, în prealabil) secțiunea *Funcții-nucleu* de la capitolul de *Fundamente*.

- i. spațiul \mathcal{Z} să aibă un număr de dimensiuni mai mare decât cel al spațiului \mathcal{X} ;
- ii. calculele pe care ar / va trebui să le facem cu vectori din spațiul \mathcal{Z} să se limiteze doar la folosirea produsului scalar;
- iii. să existe o funcție $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, astfel încât pentru orice x_i și x_j din \mathcal{X} produsul scalar dintre $\phi(x_i)$ și $\phi(x_j)$ să fie egal cu $K(x_i, x_j)$, însă valoarea $K(x_i, x_j)$ să poată fi calculată folosind în mod direct x_i și x_j (adică fără a face apel la $\phi(x_i)$ și $\phi(x_j)$). Vom numi K *funcție-nucleu* (engl., kernel function).³⁶¹

Veți vedea că un clasificator liniar din spațiul \mathcal{Z} va corespunde (în general) unui clasificator neliniar din spațiul \mathcal{X} .

Așa cum am arătat deja la problema 13, în varianta *standard* a regresiei logistice se consideră

$$P(Y = 1|X) = \sigma(w_0 + \sum_{i=1}^d w_i X_i)$$

$$P(Y = 0|X) = 1 - P(Y = 1|X),$$

unde σ este funcția logistică, definită prin relația $\sigma(a) \stackrel{\text{def.}}{=} 1/(1 + e^{-a})$ pentru orice $a \in \mathbb{R}$.

Fie o funcție ϕ care transformă o instanță oarecare X din spațiul \mathcal{X} (de dimensiune d) într-un element din spațiul \mathcal{Z} (a cărui dimensiune este m , cu $m > d$). Folosind „maparea“ ϕ , vom scrie:

$$P(Y = 1|\phi(X)) = \sigma(w_0 + \sum_{i=1}^m w_i \phi(X)_i). \quad (194)$$

unde $\phi(X)$ este vectorul de „trăsături“ (engl., feature vector) m -dimensional, care îi corespunde vectorului X din spațiul \mathcal{X} (d -dimensional). În cele ce urmează, vom nota cu $\phi(X)_i$ componenta i a vectorului $\phi(X)$.

a. Presupunem că vectorul de ponderi w este o combinație liniară de toți vectorii de trăsături $\phi(X^{(i)})$ cu $i = 1, \dots, n$, unde n este numărul instanțelor de antrenament.³⁶² Din punct de vedere formal, putem exprima acest fapt astfel:

$$(w_1, \dots, w_m)^\top = \sum_{j=1}^n \alpha_j \phi(X^{(j)}), \quad (195)$$

unde $X^{(j)}$ este instanță cu numărul de ordine j , iar α_j sunt constante reale ($j = 1, \dots, n$). Presupunem de asemenea că $w_0 = \alpha_0$.

Exprimăți probabilitatea condiționată $P(Y = 1|\phi(X))$ folosind procedeul kernel-izării (engl., kernel trick). (Astfel veți evita să faceți calculele în mod explicit în spațiul \mathcal{Z} .)

³⁶¹Spre exemplu, $K(x, x') \stackrel{\text{def.}}{=} \frac{1}{\sqrt{2\pi}\tau} \exp\left(\frac{\|x - x'\|^2}{2\tau^2}\right)$ este o astfel de funcție-nucleu. Ea se numește nucleu gaussian sau funcție cu baza radială (engl., Radial Basis Function, RBF).

³⁶²Dacă aplicăm metoda gradientului [ascendent] pornind cu valoarea $w = 0$, conform regulii de actualizare (179) și expresiei (181) din cadrul problemei 13, rezultă că într-adevăr vom obține vectorul de ponderi w sub forma unei combinații liniare de instanțe $X^{(j)}$ (respectiv $\phi(X^{(j)})$ în cazul kernel-izării).

Mai general, din *Teorema de reprezentare* care este demonstrată la problema 88 de la capitolul de *Fundamente*, rezultă că orice vector de ponderi w pentru care se atinge maximul funcției de log-verosimilitate condițională a datelor pentru regresia logistică — veți relația (178) — se poate scrie în modul indicat mai sus.

- b. Scrieți *regula de actualizare* specifică metodei gradientului ascendent pentru varianta kernel-izată a regresiei logistice.
- c. Scrieți *regula de decizie* pentru o instanță oarecare de test X din spațiul \mathcal{X} , presupunând că parametrii α_i (pentru $i = 1, \dots, m$) au fost deja învățați.

Răspuns:

- a. Pornind de la relația (194), putem exprima probabilitatea condiționată $P(Y = 1|\phi(X))$ astfel:

$$\begin{aligned} P(Y = 1|\phi(X)) &= \sigma(w_0 + \sum_{i=1}^m w_i \phi(X)_i) = \sigma(w_0 + w \cdot \phi(X)) \\ &\stackrel{(195)}{=} \sigma(w_0 + \left(\sum_{j=1}^n \alpha_j \phi(X^{(j)}) \right) \cdot \phi(X)) = \sigma(w_0 + \sum_{j=1}^n \alpha_j (\phi(X^{(j)})) \cdot \phi(X)) \\ &= \sigma(\alpha_0 + \sum_{j=1}^n \alpha_j K(X^{(j)}, X)). \end{aligned}$$

- b. În cele ce urmează, vom folosi deja-cunoscutul „artificiu al exponentierii“ necesar pentru a exprima în mod unitar probabilitățile $P(Y = 1|\phi(X))$ și $P(Y = 0|\phi(X))$:³⁶³

$$P(Y = y|\phi(X)) = \sigma(z)^y (1 - \sigma(z))^{1-y} \text{ pentru orice } y \in \{0, 1\},$$

unde

$$z \stackrel{\text{not.}}{=} \alpha_0 + \sum_{j=1}^n \alpha_j K(X^{(j)}, X). \quad (196)$$

De asemenea, vom ține cont de următoarele proprietăți ale funcției logistice:

$$\begin{aligned} \sigma(z) &= \frac{1}{1 + e^{-z}} = \frac{e^z}{1 + e^z} \\ \Rightarrow 1 - \sigma(z) &= \frac{e^{-z}}{1 + e^{-z}} = \frac{1}{1 + e^z} \\ \Rightarrow \ln(1 - \sigma(z)) &= -\ln(1 + e^z), \end{aligned}$$

pentru orice $z \in \mathbb{R}$.

Prin urmare, vom putea exprima log-verosimilitatea unei instanțe de antrenament oarecare $(\phi(X), Y = y)$ astfel:

$$\begin{aligned} \ln P(Y = y|\phi(X)) &= y \ln \sigma(z) + (1 - y) \ln(1 - \sigma(z)) \\ &= y \ln \frac{e^z}{1 + e^z} + (1 - y) \ln \frac{1}{1 + e^z} \\ &= yz - \cancel{y \ln(1 + e^z)} - \ln(1 + e^z) + \cancel{y \ln(1 + e^z)} \\ &= yz - \ln(1 + e^z). \end{aligned} \quad (197)$$

³⁶³Am folosit deja acest „artificiu“ la rezolvarea problemei 13.a.

Datele de antrenament fiind îndeobște considerate independente și identic distribuite (abrev., i.i.d.), vom putea scrie verosimilitatea condițională a setului de date $(\phi(X^{(1)}), Y^{(1)}), \dots, (\phi(X^{(n)}), Y^{(n)})$ astfel:

$$P(Y^{(1)}, \dots, Y^{(n)} | \phi(X^{(1)}), \dots, \phi(X^{(n)}); \alpha) \stackrel{i.i.d.}{=} \prod_{l=1}^n P(Y^{(l)} | \phi(X^{(l)}); \alpha),$$

unde $\alpha \stackrel{not.}{=} (\alpha_1, \dots, \alpha_n)$ este vectorul de coeficienți care au fost introdusi la relația (195).

Așadar, log-verosimilitatea condițională a setului de date $\{(\phi(X^{(1)}), Y^{(1)}), \dots, (\phi(X^{(n)}), Y^{(n)})\}$, exprimată ca funcție de α , este următoarea:

$$\begin{aligned} \ell(\alpha) &= \ln \prod_{l=1}^n P(Y^{(l)} | \phi(X^{(l)}); \alpha) = \sum_{l=1}^n \ln P(Y^{(l)} | \phi(X^{(l)}); \alpha) \\ (196) \stackrel{(197)}{=} & \sum_{l=1}^n \left[Y^{(l)} \left(\alpha_0 + \sum_{j=1}^n \alpha_j K(X^{(j)}, X^{(l)}) \right) - \ln \left(1 + \exp(\alpha_0 + \sum_{j=1}^n \alpha_j K(X^{(j)}, X^{(l)})) \right) \right]. \end{aligned}$$

De aici se obține ușor derivata parțială a funcției $\ell(\alpha)$ în raport cu componenta α_i , pentru fiecare $i = 1, \dots, n$:³⁶⁴

$$\begin{aligned} \frac{\partial \ell(\alpha)}{\partial \alpha_i} &= \sum_{l=1}^n \left(Y^{(l)} - \frac{\exp(\alpha_0 + \sum_{j=1}^n \alpha_j K(X^{(j)}, X^{(l)}))}{1 + \exp(\alpha_0 + \sum_{j=1}^n \alpha_j K(X^{(j)}, X^{(l)}))} \right) K(X^{(i)}, X^{(l)}) \\ &= \sum_{l=1}^n \left(Y^{(l)} - \sigma(\alpha_0 + \sum_{j=1}^n \alpha_j K(X^{(j)}, X^{(l)})) \right) K(X^{(i)}, X^{(l)}). \end{aligned} \quad (198)$$

și respectiv derivata parțială a funcției $\ell(\alpha)$ în raport cu componenta α_0 :

$$\begin{aligned} \frac{\partial \ell(\alpha)}{\partial \alpha_0} &= \sum_{l=1}^n \left(Y^{(l)} - \frac{\exp(\alpha_0 + \sum_{j=1}^n \alpha_j K(X^{(j)}, X^{(l)}))}{1 + \exp(\alpha_0 + \sum_{j=1}^n \alpha_j K(X^{(j)}, X^{(l)}))} \right) \\ &= \sum_{l=1}^n \left(Y^{(l)} - \sigma(\alpha_0 + \sum_{j=1}^n \alpha_j K(X^{(j)}, X^{(l)})) \right). \end{aligned}$$

Regula de actualizare specifică metodei gradientului ascendent aplicată la rezolvarea acestei probleme de regresie logistică kernel-izată este:

$$\alpha_i^{(t+1)} = \alpha_i^{(t)} + \eta \frac{\partial \ell(\alpha)}{\partial \alpha_i} \text{ pentru } i = 1, \dots, n,$$

sau, scriind vectorial: $\alpha^{(t+1)} = \alpha^{(t)} + \eta \nabla_{\alpha} \ell(\alpha)$.

c. Conform relației (194), modelul învățat de regresia logistică asociază instanței X eticheta $Y = 1$ dacă și numai dacă $\sigma(w_0 + \sum_{i=1}^m w_i \phi(X)_i) \geq 1/2$, ceea ce este echivalent cu $w_0 + \sum_{i=1}^m w_i \phi(X)_i \geq 0$, adică $w_0 + w \cdot \phi(X) \geq 0$. Tinând cont de relația (195) și de faptul că $w_0 = \alpha_0$, această ultimă inegalitate revine la $\alpha_0 + \sum_{j=1}^n \alpha_j \phi(X^{(j)}) \cdot \phi(X) \geq 0$, adică

$$\alpha_0 + \sum_{j=1}^n \alpha_j K(X^{(j)}, X) \geq 0.$$

³⁶⁴A se compara expresia (198) cu expresia (184) care a fost obținută la problema 13, pentru regresia logistică nekernel-izată.

18.

(Regresia logistică:
clasificare n -ară [regresie *softmax*] cu regularizare L_2 ,
folosind metoda gradientului ascendent)

■ □ • ○ CMU, 2012 fall, T. Mitchell, Z. Bar-Joseph, HW2, pr. 2
MIT, 2016 fall, R. Barzilay, S. Sra, Weekly Exercises, week 4, pr. 5.ad

În acest exercițiu vom arăta că putem extinde ușor modelul binar al regresiei logistice în aşa fel încât să realizăm clasificare n -ară. Să presupunem că avem K clase distincte și că probabilitatea a posteriori pentru clasa k este definită prin

$$\begin{aligned} P(Y = k|X = x) &= \frac{\exp(w_k \cdot x)}{1 + \sum_{l=1}^{K-1} \exp(w_l \cdot x)} \text{ pentru } k = 1, \dots, K-1 \\ P(Y = K|X = x) &= \frac{1}{1 + \sum_{l=1}^{K-1} \exp(w_l \cdot x)}, \end{aligned}$$

unde x și w_k cu $k = 1, \dots, K$ sunt vectori d -dimensionali. Remarcați că pentru a simplifica expresia acestor probabilități nu am folosit componentele w_{k0} .³⁶⁵ Remarcați de asemenea că pentru $K = 2$ se obține regresia logistică binară, aşa cum a fost prezentată la problema 13.

Observație: Putem simplifica expresia de mai sus pentru regresia logistică n -ară (engl., multiclass logistic regression) introducând un parametru fix, vectorul $w_K = 0$ (un vector d -dimensional format în întregime din zerouri).³⁶⁶

Scopul nostru este să estimăm parametrii / ponderile w_t folosind ca metodă de optimizare gradientul ascendent. Vom stabili de asemenea distribuțiile *a priori* pe care le urmează parametrii w_t , pentru a evita ca ei să [ajungă să] aibă valori foarte mari [în normă euclidiană] și să se producă *overfitting* (rom., supra-specializare).

a. Presupunem că se dă setul de date de antrenament $\{(x^1, y^1), \dots, (x^n, y^n)\}$, cu $x^i \in \mathbb{R}^d$ și $y^i \in \{1, \dots, K\}$ pentru $i = 1, \dots, n$. Scrieți expresia funcției de log-verosimilitate condițională, $\ell(w_1, \dots, w_K)$, folosind *regularizare* de normă L_2 pentru ponderi.³⁶⁷ Arătați pașii pe care i-ați urmat ca să deduceți această expresie.

b. Remarcați că, asemenea cazului clasificării binare (vedeți problema 13), nu există o expresie pentru calculul direct (engl., closed form) al valorii parametrului w_k care corespunde maximului funcției de log-verosimilitate condițională, $\ell(w_1, \dots, w_K)$. Totuși, putem găsi soluția folosind *metoda gradientului ascendent*, cu ajutorul derivatelor parțiale. Calculați expresia componentei

³⁶⁵ Alternativ, se poate considera că extindem „artificial“ vectorul x — și, la fel, fiecare vector x^l din setul de date de antrenament — cu componenta $x_0 = 1$ și, respectiv, $x_0^l = 1$.

³⁶⁶ Funcția $f : \{w_1 \cdot x, \dots, w_K \cdot x\} \rightarrow [0, 1]$ definită prin $f(w_k \cdot x) = \frac{\exp(w_k \cdot x)}{1 + \sum_{l=1}^{K-1} \exp(w_l \cdot x)}$ și numită *funcția softmax*, are proprietatea că menține ordinea, adică $w_k \cdot x < w_j \cdot x \Rightarrow f(w_k \cdot x) < f(w_j \cdot x)$. Întrucât, de asemenea, f este derivabilă în raport cu w_k — ceea ce justifică particula *soft* în denumirea softmax —, această funcție este folosită adeseori ca funcție de activare pentru stratul de ieșire al rețelelor neuronale cu output multiplu. În mecanica statistică funcția softmax este cunoscută sub numele *distribuția Boltzmann* (introdusă de fizicianul austriac Ludwig Boltzmann în anul 1868) sau *distribuția Gibbs*.

³⁶⁷ Termenii de *regularizare* și *overfitting* au fost introdusi pentru regresia liniară la problema 3.C. La secțiunea de *Regresie logistică*, am folosit pentru prima dată regularizarea (de normă) L_2 la problema 15.

de pe poziția k din vectorul gradient pentru $\ell(w_1, \dots, w_K)$, și anume derivata parțială a lui $\ell(w_1, \dots, w_K)$ în raport cu w_k .

c. Dând ponderilor valoarea inițială 0, scrieți *regula de actualizare* (engl., update rule) pentru w_k , folosind η ca rată a învățării (engl., step size). Va converge oare soluția la maximul global?

Răspuns:

a. Presupunem că dispunem de n instanțe de antrenament. Folosind funcția-indicator $1_{\{l=k\}}$, care este definită prin $1_{\{l=k\}} = 1$ dacă $Y^l = k$ și 0 în caz contrar, putem scrie funcția de verosimilitate condițională astfel:

$$L(w_1, \dots, w_K) = \prod_{l=1}^n \prod_{k=1}^K P(Y^l = k | X^l = x^l; w)^{1_{\{l=k\}}} = \prod_{l=1}^n \prod_{k=1}^K \left(\frac{\exp(w_k \cdot x^l)}{\sum_r \exp(w_r \cdot x^l)} \right)^{1_{\{l=k\}}}.$$

Aplicând funcția \ln , vom obține funcția de log-verosimilitate condițională:

$$\ell(w_1, \dots, w_K) = \sum_{l=1}^n \sum_{k=1}^K 1_{\{l=k\}} \left((w_k \cdot x^l - \ln \sum_r \exp(w_r \cdot x^l)) \right).$$

Adăugând termenul de regularizare corespunzător normei L_2 , rezultă:³⁶⁸

$$\ell(w_1, \dots, w_K) = \sum_{l=1}^n \sum_{k=1}^K 1_{\{l=k\}} \left(w_k \cdot x^l - \ln \sum_r (\exp(w_r \cdot x^l)) \right) - \frac{\lambda}{2} \sum_{k=1}^K \|w_k\|^2.$$

b. Calculăm derivata parțială — de fapt, vectorul de derive parțiale — pentru funcția de log-verosimilitate condițională, în raport cu parametrul w_i :³⁶⁹

$$\begin{aligned} \frac{\partial}{\partial w_i} \ell(w_1, \dots, w_K) &= \sum_{l=1}^n \left(1_{\{l=i\}} x^l - \frac{\exp(w_i \cdot x^l) x^l}{\sum_r \exp(w_r \cdot x^l)} \right) - \lambda w_i \\ &= \sum_{l=1}^n (1_{\{l=i\}} - P(Y^l = i | X^l = x^l)) x^l - \lambda w_i. \end{aligned}$$

c. Regula de actualizare pentru parametrul w_i , folosind gradientul ascendent, este următoarea:

$$w_i \leftarrow w_i + \eta \sum_{l=1}^n (1_{\{l=i\}} - P(Y^l = i | X^l = x^l)) x^l - \eta \lambda w_i.$$

Convergența la punctul de maxim global are loc întrucât funcția de log-verosimilitate condițională este concavă.³⁷⁰

³⁶⁸Vedeți problema 15.

³⁶⁹Vă reamintim că acest tip de derivată se numește *derivată vectorială*. Pentru expresia $w \cdot x = w^\top x$, derivata în raport cu vectorul w este vectorul x . (Vedeți documentul *Matrix Identities* de Sam Roweis.)

³⁷⁰Demonstrația acestei proprietăți extinde în mod natural demonstrația din cazul clasificării binare, care a fost prezentată la problema 14.

1.2 Metode de regresie — Probleme propuse

1.2.1 Regresia liniară

19.

(Regresia liniară unidimensională:
deducerea directă a soluției analitice MLE / LSE)

• CMU, 2011 fall, Eric Xing, HW1, pr. 2.1

Considerăm un caz simplu de regresie liniară, în care modelul [parametric] este de forma $y = ax + b + \varepsilon$, unde $x, y, a, b \in \mathbb{R}$, iar ε este un termen „zgomot“ care urmează o distribuție gaussiană de medie zero, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.³⁷¹ Avem n instanțe de antrenament, (x_i, y_i) , $i = 1, 2, \dots, n$, cu $x_i, y_i \in \mathbb{R}$.

a. Scrieți expresia funcției de cost reprezentând suma pătratelor erorilor în acest model. Nu folosiți forma matriceală. Funcția aceasta va fi exprimată folosind modelul parametric unidimensional (1-D).

b. Calculați valorile \hat{a} și \hat{b} care minimizează funcția de cost / pierdere și arătați că

$$\begin{aligned}\hat{a} &= \frac{n(\sum_i x_i y_i) - (\sum_i x_i)(\sum_i y_i)}{n(\sum_i x_i^2) - (\sum_i x_i)^2} \\ \hat{b} &= \frac{(\sum_i x_i^2)(\sum_i y_i) - (\sum_i x_i)(\sum_i x_i y_i)}{n(\sum_i x_i^2) - (\sum_i x_i)^2},\end{aligned}$$

bineînțeles, în condițiile în care expresia de la numitorul celor două fracții este nenulă. Ca și mai sus, nu folosiți notația matriceală. Minimizați în manieră directă funcția de cost / pierdere de la punctul a.

c. Deducreți formula pentru estimarea de verosimilitate maximă (MLE) pentru modelul 1-D. Arătați că estimarea aceasta (de tip MLE) coincide cu estimarea corespunzătoare minimizării criteriului sumei pătratelor erorilor, obținută la punctul a.

d. Arătați că soluția optimă calculată la punctul b este aceeași cu forma matriceală obținută în cazul modelului [mai] general de regresie liniară,

$$\hat{\beta} = (X^\top X)^{-1} X^\top y,$$

unde, în cazul nostru $\hat{\beta} \stackrel{\text{not.}}{=} (\hat{a}, \hat{b})$, $X \in \mathbb{R}^{n \times 2}$, cu linia i din X fiind $(1, x_i)$ pentru $i = 1, \dots, n$, iar $y = (y_1, \dots, y_n)^\top$.³⁷²

³⁷¹Așadar, acest model este ușor mai elaborat decât cel care a fost prezentat la problema 1, dar nu [atât de] general precum cel de la problema 3.A.

³⁷²Vedeți relația (156) de la problema 3.d.

20. (Compararea a două modele de regresie unidimensională de tip LSE)

• CMU, 2014 fall, Z. Bar-Joseph, W. Cohen, HW2, pr. 3

Fie X și Y două variabile aleatoare, iar $(x_1, y_1), \dots, (x_n, y_n)$ un set de n exemple de antrenament generate în mod independent.

Considerăm că modelele de regresie de mai jos sunt de tipul sumei celor mai mici pătrate (engl., least squared errors, LSE). Așadar, „zgomotul“ ε urmează o distribuție gaussiană $N(0, \sigma^2)$

- i. $Y = \beta X + \varepsilon$
- ii. $Y = \beta^2 X + \varepsilon$.

a. Calculați soluțiile acestor două modele de regresie.

b. Care dintre cele două modele de mai sus produce o eroare la antrenare mai mică?

Sugestie: Răspunsul poate să depindă de setul de date de antrenament. În consecință, explicați în ce situație un anumit model este mai bun decât celălalt.

21. (Regresia liniară: aplicație pentru „învățarea“ parametrului unei funcții nelineare, componenta „zgomot“ fiind modelată cu o distribuție gaussiană)

• CMU, 2009 spring, Tom Mitchell, midterm, pr. 6

Se dă un set de date constituit din n instanțe, $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Atât atributul de intrare x cât și atributul de ieșire y au ca valori numere reale. Valoarea atributului de ieșire este generată de o distribuție gaussiană având media $\sin(wx_i)$ și varianța 1:

$$p(y_i | x_i, w) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_i - \sin(wx_i))^2}{2}}$$

În modelul de mai sus avem un parametru necunoscut w și dorim să deducem / „învățăm“ din aceste date care este estimarea sa în sensul verosimilității maxime.

a. Scrieți expresia verosimilității [condiționale a] datelor ca funcție de parametrul w , date fiind valorile observate x_i și y_i , cu $i = 1, \dots, n$.

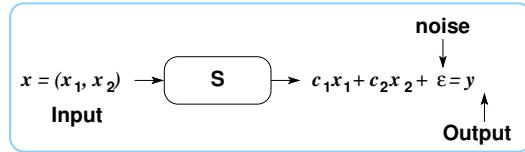
b. Precizați care dintre egalitățile de mai jos este adevărată atunci când se dorește / face estimarea parametrului w în sensul maximizării verosimilității datelor. În cazul în care niciuna dintre relațiile *i-v* nu este adevărată, marcați opțiunea *vi*.

- i. $\sum_i \cos^2(wx_i) = \sum_i y_i \sin(x_i)$
- ii. $\sum_i \cos^2(wx_i) = \sum_i y_i \sin(2wx_i)$
- iii. $\sum_i \sin(wx_i) \cos(wx_i) = \sum_i y_i \sin(wx_i/2)$
- iv. $\sum_i x_i \sin(wx_i) \cos(wx_i) = \sum_i x_i y_i \cos(wx_i)$
- v. $w \cos(x_i) = \sum_i x_i y_i \cos(x_i)$
- vi. Niciuna dintre variantele de mai sus.

22. (Estimarea parametrilor unui model de regresie liniară bidimensională, în care componenta-zgomot este modelată fie cu o distribuție gaussiană — inclusiv o variantă de regresie local-ponderată —, fie cu o distribuție Laplace)

■ □ • ○ CMU, 2009 fall, Carlos Guestrin, HW1, pr. 1.5.2
CMU, 2012 fall, E. Xing, A. Singh, HW1, pr. 2

Figura alăturată reprezintă un sistem S care ia ca intrări [valorile] x_1 și x_2 și produce la ieșire o combinație liniară a celor două intrări, $c_1x_1 + c_2x_2$, unde c_1 și c_2 sunt două numere reale necunoscute.



Dispozitivul folosit ca să măsoare outputul sistemului S introduce în plus o „eroare“ ε , care este [asimilată cu] o variabilă aleatoare ce urmează o anumită distribuție. Astfel, outputul observat (y) este dat de ecuația

$$y = c_1x_1 + c_2x_2 + \varepsilon.$$

Presupunem că avem $n > 2$ instanțe (x_{j1}, x_{j2}, y_j) , $j = 1, \dots, n$ sau, echivalent, (x_j, y_j) , $j = 1, \dots, n$, unde $x_j = [x_{j1}, x_{j2}]$. Prin urmare, a avea la dispoziție n măsurători revine la a avea n egalități de forma $y_j = c_1x_{j1} + c_2x_{j2} + \varepsilon_j$, cu $j = 1, \dots, n$. Scopul nostru este să estimăm parametrii c_1 și c_2 din aceste măsurători cu ajutorul criteriului verosimilității maxime, făcând [la fiecare dintre punctele de mai jos] diferite presupuneri în legătură cu „zgomotul“.

- Presupunem că ε_i pentru $i = 1, \dots, n$ sunt variabile aleatoare gaussiene i.i.d. (independente și identic distribuite) cu media 0 și varianță σ^2 . Calculați funcția de log-verosimilitate a datelor. Apoi folosiți această funcție pentru a demonstra că estimarea de verosimilitate maximă $c^* = [c_1^*, c_2^*]$ este soluția unei probleme de aproximare prin aşa-numita *metodă a celor mai mici pătrate* (engl., least squares approximation method). Nu este necesar să găsiți efectiv soluția problemei celor mai mici pătrate [pentru acest caz].
- [Varianta regresiei local-ponderate] Presupunem că ε_i pentru $i = 1, \dots, n$ sunt variabile aleatoare independente de tip gaussian având media 0 și varianța σ_i^2 . Calculați funcția de log-verosimilitate și găsiți perechea de valori $c^* = [c_1^*, c_2^*]$ care o maximizează, adică estimarea de verosimilitate maximă (MLE) a parametrilor c_1 și c_2 .
- Presupunem că pentru orice $i = 1, \dots, n$, variabila ε_i are funcția de densitate

$$f_{\varepsilon_i}(x) = f(x) = \frac{1}{2\theta} e^{-\frac{|x|}{\theta}}, \text{ cu } \theta > 0.$$

Cu alte cuvinte, zgomotul este i.i.d. și urmează o distribuție Laplace având parametrul de locație $\mu = 0$ și parametrul de scalare θ . Calculați funcția de log-verosimilitate a datelor pentru acest model al zgomotului și indicați motivul / motivele pentru care acest model conduce la *soluții mai robuste* [la „zgomote“ și outlier-e] decât în cazurile precedente.³⁷³

³⁷³Pentru reprezentarea grafică a mai multor distribuții Laplace, vedeți problema 8.B.

23. (Regresia liniară bidimensională: exemplu de aplicare a metodei gradientului, varianta stochastică)

• CMU, 2014 spring, Seyoung Kim, HW2, pr. 1.2.B

În această problemă vom presupune că ori de câte ori se înmulțește o instanță [reprezentată printr-un vector] x cu un vector de ponderi, x a fost deja augmentat [în aşa fel încât să înceapă] cu o componentă 1. De exemplu, x poate fi $(1, x_1)$ sau $(1, x_1, x_2)$, iar constanta 1 va fi înmulțită cu ponderea β_0 .

Considerăm că folosim metoda gradientului pentru a antrena un model de regresie liniară pentru a „învăța“ funcția $f : \mathbb{R}^2 \rightarrow \mathbb{R}$, ale cărei argumente (sau attribute) sunt notate cu x_1 și respectiv x_2 . Facem inițializarea vectorului de ponderi astfel: $\beta = (0.5, 0.5, 0.5)$. Primul exemplu considerat este $x = (2, 6)$ și, corespunzător, $y = 8.5$.

a. Folosind aceste valori inițiale pentru ponderi, cât este eroarea $y - f(x) = y - x \cdot \beta$ produsă de modelul de regresie?

b. Cât va fi β_1^{new} , noua valoare a ponderii β_1 (care corespunde atributului / trăsăturii x_1) la finalul primei iterării executate de algoritmul gradientului descendente, varianta stochastică,³⁷⁴ folosind exemplul indicat mai sus? Veți utiliza rata de învățare (engl., learning rate) $\eta = 0.05$.

În cele ce urmează vom presupune că am terminat de antrenat modelul de regresie liniară și că am obținut vectorul de ponderi $\hat{\beta} = (0.5, 1, 1)$. Acum vom considera instanța de test $x_q = (3, 4)$ și vom prezice $f(x_q)$. Presupunem că în prealabil am estimat în mod empiric varianța componentei-zgomot (ε) din modelul de regresie și am obținut $\sigma = 1$.

c. Cât este valoarea „așteptată“ (engl., expected value) pentru $f(x_q)$?

d. Cât este probabilitatea ca valoarea adevărată pentru $f(x_q)$ să fie în intervalul $[9, 10]$?

Sugestie: Puteti folosi tabela de valori ale funcției cumulative de distribuție (Φ) pentru distribuția gaussiană standard care apare la pr. 33 de la capitolul de *Fundamente*.

24. (Regresia liniară, varianta LSE: o proprietate interesantă: adăugarea de noi trăsături / attribute nu mărește suma pătratelor erorilor)

□ • ○ Stanford, 2014 fall, Andrew Ng, midterm, pr. 1

După cum am arătat la problema 3.A, la regresia liniară în varianta LSE (în care modelarea „zgomotului“ se face folosind distribuția gaussiană) se lucrează cu o funcție de cost de tipul sumei celor mai mici pătrate (engl., Least Squared Errors):³⁷⁵

$$J(\beta) = \sum_{i=1}^n (\beta^\top x^{(i)} - y^{(i)})^2 = (X\beta - y)^\top (X\beta - y).$$

³⁷⁴Pentru prezentarea metodei gradientului aplicată la rezolvarea regresiei liniare, vedeti pr. 6.

³⁷⁵Corespondența dintre notațiile de aici și cele din problema 3.A este următoarea: $x^{(i)} \rightarrow X'_i$ și $X \rightarrow X'$.

Obiectivul la aplicarea acestei metode de regresie este să găsim acea valoare a vectorului de parametri β pentru care se atinge minimul expresiei $J(\beta)$, pentru respectivul set de date de antrenament.

Presupunem că în acest set de date avem inițial d trăsături și, în consecință, inputul la antrenare este dat sub forma lui $X \in \mathbb{R}^{n \times (d+1)}$, așa-numita *matrice de design*.

Să zicem că acum avem acces la încă o trăsătură pentru fiecare exemplu de antrenament. Așadar, avem un vector adițional de trăsături $v \in \mathbb{R}^{n \times 1}$ pentru setul nostru de antrenament și dorim să-l folosim în problema noastră de regresie. Putem face acest lucru creând o nouă matrice de design: $\tilde{X} \stackrel{\text{not.}}{=} [X \ v] \in \mathbb{R}^{n \times (d+2)}$. Prin urmare, noul nostru vector de parametri este $\beta_{\text{new}} \stackrel{\text{not.}}{=} [\beta, p]^T$, unde $p \in \mathbb{R}$ este un parametru care corespunde noului vector de trăsături v .

Observație importantă: Pentru a asigura o anumită simplitate pentru demonstrația matematică, în cele ce urmează vom presupune că $X^T X = I \in \mathbb{R}^{(d+1) \times (d+1)}$ și $\tilde{X}^T \tilde{X} = I \in \mathbb{R}^{(d+2) \times (d+2)}$, $v^T v = 1$. Aceasta este așa-numita presupozitie de *ortonormalitate*. Altfel spus, coloanele matricii X (respectiv \tilde{X}) sunt ortonormale. Concluziile demonstrației din această problemă se mențin chiar și fără această presupozitie de ortonormalitate, însă cu ea demonstrația va fi mai ușoară.

a. Fie $\hat{\beta} \stackrel{\text{not.}}{=} \arg \min_{\beta} J(\beta)$ valoarea vectorului de parametri β care minimizează funcția obiectiv originală (cea care folosește matricea de design X). Folosind presupozitia de ortornormalitate, demonstrați că

$$J(\hat{\beta}) = (X X^T y - y)^T (X X^T y - y),$$

adică, arătați că aceasta este valoarea lui $\min_{\beta} J(\beta)$ (valoarea funcției obiectiv în punctul de minim).

b. Vom nota cu $\tilde{\beta}_{\text{new}}$ valoarea vectorului de parametri β pentru care se atinge minimul funcției

$$\tilde{J}(\beta_{\text{new}}) \stackrel{\text{not.}}{=} (\tilde{X} \beta_{\text{new}} - y)^T (\tilde{X} \beta_{\text{new}} - y).$$

Determinați $\tilde{J}(\tilde{\beta}_{\text{new}})$, valoarea minimă a noii funcții obiectiv, și scrieți această expresie sub forma

$$\tilde{J}(\beta_{\text{new}}) = J(\hat{\beta}) + f(X, v, y),$$

unde termenul $J(\hat{\beta})$ a fost obținut la punctul a , iar f este o anumită funcție de argumentele X , v și y .

c. Demonstrați că valoarea optimă a funcției obiectiv nu se mărește dacă adăugăm o nouă trăsătură la matricea de design. Altfel spus, arătați că

$$\tilde{J}(\hat{\beta}_{\text{new}}) \leq J(\hat{\beta}).$$

d. Arată oare rezultatul de mai sus că adăugând [din ce în ce] mai multe trăsături vom obține întotdeauna (sau, cu necesitate) un model de regresie care generalizează mai bine decât un model cu mai puține trăsături? Explicați de ce *da*, ori, dimpotrivă de ce *nu*.

25.

(Regresia liniară cu regularizare L_2 (*ridge*)
și respectiv regularizare L_1 :
efectul de *diminuare a ponderilor* (engl., weight decay))

*prelucrare de Liviu Ciortuz, după
□ • ○ CMU, 2011 fall, Eric Xing, HW1, pr. 2.2*

În regresia liniară, ni se dă un set de date de antrenament de forma $D = \{(x_i, y_i) | i = 1, \dots, n\}$, cu $x_i = (x_{i,1}, \dots, x_{i,d})^\top$ și $y_i \in \mathbb{R}$. Fie matricea de design $X \in \mathbb{R}^{n \times d}$, unde linia i este x_i^\top și, de asemenea, vectorul $y = (y_1, \dots, y_n)^\top$.

Considerând că lucrăm cu un model parametrizat (engl., parametric model) de forma $y_i = x_i^\top \beta + \varepsilon_i$, unde $\beta = (\beta_1, \dots, \beta_d)^\top$, iar ε_i este un termen „zgomot“ (engl., noise term) care urmează o [anumită] distribuție dată, regresia liniară caută să găsească un vector de parametri β care furnizează cel mai bun “fit” pentru modelul de regresie de mai sus. Un exemplu (de criteriu) pentru a măsura fitness-ul, este găsirea celui β care minimizează o anumită *funcție de cost / pierdere* (engl., loss function) $J(\beta)$, dată.

a. Pentru regresia liniară cu termen de regularizare de normă L_1 , minimizăm următoarea funcție de cost / pierdere:³⁷⁶

$$J_L(\beta) = \sum_i (x_i^\top \beta - y_i)^2 + \lambda \sum_{j=1}^d |\beta_j| = \underbrace{(X\beta - y)^\top (X\beta - y)}_{\|X\beta - y\|^2} + \lambda \|\beta\|_1.$$

Presupunem că $X^\top X = I$ (este așa-numita proprietate de *ortonormalitate* a matricei X). Calculați valoarea parametrului β care minimizează criteriul J_L . (Veți nota această valoare cu $\hat{\beta}_L$).

Indicație (1): Vă readucem aminte următoarea regulă de derivare:

$$\frac{\partial |\beta_a|}{\partial \beta_a} = \begin{cases} 1 & \text{dacă } \beta_a > 0 \\ -1 & \text{dacă } \beta_a < 0 \\ \text{nedefinit} & \text{dacă } \beta_a = 0. \end{cases}$$

Indicație (2): Pentru λ situat într-un anumit interval, nu există soluție analitică (ecuații „normale“) pentru această problemă de optimizare. Ce înseamnă acest fapt?

b. *Remember:* La problema 3.A am arătat că atunci când luăm ca funcție de cost suma pătratelor erorilor (engl., the square-error), adică

$$J(\beta) = \sum_i (x_i^\top \beta - y_i)^2 = (X^\top \beta - y)^\top (X^\top \beta - y),$$

obținem (atunci când matricea $X^\top X$ este inversabilă) următoarea soluție pentru problema de optimizare $\arg \min_{\beta} J(\beta)$:

$$\hat{\beta} = (X^\top X)^{-1} X^\top y.$$

³⁷⁶ Atenție! A nu se confundă acest tip de regresie regularizată cu varianta de regresie liniară în care zgomotul ε este modelat cu distribuția Laplace. Pentru regresia aceasta din urmă, criteriul de optimizat este — conform problemei 8.B — o sumă a erorilor în modul / norma L_1 (vezi relația (169)). În problema de față, doar termenul de regularizare este în norma L_1 (este vorba despre $\lambda \|\beta\|_1$), celălalt termen din criteriul J_L este o sumă de pătrate (deci este în norma L_2).

În cazul regresiei liniare cu termen de regularizare de normă L_2 (regresia *ridge*), minimizăm următoarea funcție de cost / pierdere:

$$J_R(\beta) = \sum_i (x_i^\top \beta - y_i)^2 + \lambda \sum_{j=1}^d \beta_j^2 = (X\beta - y)^\top (X\beta - y) + \lambda \|\beta\|^2.$$

La problema 3.C am arătat că valoarea parametrului β care minimizează criteriul J_R este următoarea (atunci când matricea $X^\top X + \lambda I$ este inversabilă):

$$\hat{\beta}_R = (X^\top X + \lambda I)^{-1} X^\top y.$$

Presupunem că $X^\top X = I$. Scrieți valorile parametrului β care minimizează criteriile J și J_R . Comparați și explicați cum anume cele două metode de regularizare, *ridge* și L_1 , afectează (eventual, diminuează) valorile parametrilor $\hat{\beta}_j$, pentru $j \in \{1, \dots, d\}$. (Am folosit notația $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_d)^\top$.)

26. (Regresia ridge kernel-izată: exemplu de aplicare)

□ • CMU, 2020 spring, Pat Virtue, recitation 9, pr. 1.2

Algoritmul regresiei ridge kernel-izate este următorul:³⁷⁷

Pasul 1 (Antrenare):

Calculează vectorul-coloană $\alpha = (K + \lambda I)^{-1}y$, unde $K_{ij} = k(x^{(i)}, x^{(j)})$, iar k este funcția-nucleu.

Pasul 1 (Predicție):

Dată fiind o instanță nouă x , calculează *predicția* $\hat{y} = \sum_{i=1}^N \alpha_i k(x, x^{(i)})$.

În cele ce urmează veți folosi funcția-nucleu

$$k(x, x^{(i)}) = \begin{cases} 1 & \text{dacă } \|x - x^{(i)}\|^2 \leq \frac{1}{2} \\ 0 & \text{altfel,} \end{cases}$$

unde simbolul $\|\cdot\|$ desemnează norma euclidiană.

Vi se dau următoarele instanțe etichetate (x, y) , cu x și y din \mathbb{R} : $(3, 4)$, $(3.7, 1)$, $(4.2, -2)$.

a. Calculați matricea K și vectorul α . Pentru simplitate, veți considera $\lambda = 0$.

b. Preziceți valoarea \hat{y} pentru punctul $x = 3.4$.

Remember:

Dacă A este matrice pătratică de ordin $n \times n$ *nesingulară* (adică, $\det(A) \neq 0$), atunci există A^{-1} , *inversa* matricei A , și se calculează astfel:

$$A^{-1} = \frac{1}{\det(A)} (A^*)^\top,$$

unde operatorul \top desemnează transpusa unei matrice oarecare, iar A^* este matricea de ordin $n \times n$ (numită matricea adjunctă) pentru care elementul generic $[A^*]_{i,j}$ este produsul dintre $(-1)^{i+j}$ și determinantul matricei obținute din matricea A în urma eliminării liniei i și a coloanei j .

³⁷⁷Vă readucem aminte că regresia eridge este regresia liniară cu termen de regularizare de normă L_2 . Pentru detalii privind kernel-izarea regresiei *ridge*, veți problema 9.

27. (Regresia liniară cu zgomot gaussian și cu regularizare L_1 : rezolvare folosind metoda descreșterii pe coordonate și noțiunile de *subderivată* / *subgradient* și *subdiferențială*)
 • CMU, 2009 fall, Carlos Guestrin, HW2, pr. 2.cde
 MIT, 2003 fall, Tommi Jaakkola, HW4, pr. 1.ab³⁷⁸

Se consideră un set de puncte împreună cu output-urile asociate: $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$, unde $x_i \in \mathbb{R}^d$, iar $y_i \in \mathbb{R}$ pentru $i = 1, \dots, n$. Vom folosi aceste date ca să antrenăm un predictor liniar $y = w \cdot F(x)$, unde $w = (w_1, \dots, w_d)$, iar $F(x) = (f_1(x), \dots, f_d(x))$, cu d finit. După cum știți, f_i este numită *trăsătură* (engl., feature) sau *funcția de bază* (engl., basis function) de indice i pentru problema noastră de învățare. Fie următoarea funcție-obiectiv:

$$J(w, \lambda) = \frac{1}{n} \sum_{j=1}^n \frac{1}{2} (y_j - w \cdot F(x_j))^2 + \lambda \|w\|_1, \quad (199)$$

unde $\|w\|_1 \stackrel{\text{def.}}{=} \sum_{i=1}^d |w_i|$ este norma L_1 pentru vectorul w , iar $\lambda > 0$. Vom folosi datele noastre ca să învățăm vectorul $w^* \stackrel{\text{not.}}{=} w^*(\lambda)$ minimizând funcția-obiectiv $J(w, \lambda)$, adică,

$$w^* = \arg \min_{w \in \mathbb{R}^d} J(w, \lambda). \quad (200)$$

Acest criteriu de optimizare conduce în mod tipic la realizarea unei bune selecții de trăsături atunci când se alege o valoare mare pentru parametrul λ . Cu alte cuvinte, dacă atribuim lui λ o valoare mare, atunci multe ponderi w_i vor deveni 0. În consecință, trăsăturile / funcțiile de bază corespunzătoare acestor ponderi w_i nu vor fi folosite de către predictorul nostru.

Comentariu: Pentru a rezolva această problemă de regresie liniară având termenul de regularizare de normă L_1 , putem folosi o metodă de optimizare simplă cunoscută sub numele de *metoda descreșterii pe coordonate* (engl., *coordinate descent*). Aceasta implică faptul că vom ajusta pe rând fiecare parametru w_i , minimizând funcția obiectiv în care toți ceilalți parametri sunt fixați:

$$w_i^* = \arg \min_{w_i} J(w, \lambda).$$

Făcând indicele i să ia valori succesive în mulțimea $\{1, \dots, d\}$ și ajustând pe rând parametrii w_i , cu repetarea de mai multe ori a acestui ciclu iterativ, vectorul de parametri va converge asymptotic la minimul global al funcției obiectiv convexe. În general, metoda descreșterii pe coordonate apelează o subrutină care realizează minimizarea pe coordonata aleasă la un moment dat — „minimizarea pe linie” (engl., *line minimization*). În problema noastră, această minimizare se poate realiza într-o formă analitică (engl., *closed form*).

a. Calculați $\partial_{w_i} J$, subdiferențiala funcției $J(w, \lambda)$ în raport cu parametrul w_i .³⁷⁹

Sugestie (1): Calculați derivata întâi a primului termen al funcției J în raport cu w_i și apoi subdiferențiala celui de-al doilea termen. Asigurați-vă că tratați toate cele trei cazuri: $w_i < 0$, $w_i = 0$ și $w_i > 0$. Scrieți derivata primului termen sub forma $a_i w_i - c_i$, unde c_i este dat (mai jos) de expresia (201).

³⁷⁸Punctele c și d ale acestei probleme sunt de tip implementare. Acolo sunteți solicitați să aplicați algoritmul dezvoltat în acest exercițiu pe un anumit set de date de antrenament.

³⁷⁹Pentru definiția noțiunii de *subderivată*, vedeți pr. 81 de la capitolul de *Fundamente*.

Sugestie (2): O subderivată a lui $J(w, \lambda)$ va fi egală cu suma dintre derivata primului termen (fiindcă acest termen este derivabil) și o subderivată celui de-al doilea termen.

Sugestie (3): Subderivata astfel calculată pentru $J(w, \lambda)$ va putea fi folosită ulterior în cadrul algoritmului subgradientului descendant (ciclic) care a fost introdus la pr. 168 de la capitolul de *Fundamente*, obținând astfel un algoritm de antrenare pentru regresia liniară cu regularizare de normă L_1 .

b. Calculați punctul de optim global w_i^* ținând cont de următoarea *proprietate*: w^* este un punct de optim global pentru funcția convexă f dacă și numai dacă $0 \in \partial f(w^*)$. Cu alte cuvinte, găsiți w_i^* astfel încât $0 \in \partial_{w_i^*} J$.

c. Să explorăm mai atent relația dintre parametrul de regularizare λ , ponderea w_i pentru trăsătura / funcția de bază de indice i și cantitatea c_i definită astfel:

$$c_i = \frac{1}{n} \sum_{j=1}^n f_i(x_j)(y_j - \sum_{m=1, m \neq i}^d w_m f_m(x_j)). \quad (201)$$

Ce semnificație are c_i ? Trasați graficul lui w_i^* în raport cu c_i . Unde apare λ în acest grafic?

28.

(Regresia liniară cu „zgomot“ modelat cu distribuția *Laplace*: rezolvare în cazul unidimensional [chiar particularizat] cu ajutorul derivatei, acolo unde există)

□ • CMU, 2009 fall, Carlos Guestrin, HW2, pr. 1

Presupunem că dorim să prezicem o valoare necunoscută Y , după ce ne-au fost puse la dispoziție [ca date de antrenament] o secvență de „observații“ x_1, \dots, x_n pentru Y , care sunt afectate de prezența unor „zgomote“ i.i.d., mai precis $y_i = x_i + \epsilon_i$, pentru $i = 1, \dots, n$.

Comentariu: Se poate arăta³⁸⁰ că atunci când presupunem că „zgomotele“ sunt independente și distribuite conform aceleiași distribuții gaussiene ($\epsilon_i \sim \mathcal{N}(0, \sigma^2)$), a găsi estimarea în sens MLE pentru Y echivalează cu a găsi acea valoare \hat{y} care minimizează suma pătratelor erorilor [adică, a diferențelor] în raport cu aceste valori date, x_i . Altfel spus,

$$\hat{y} = \arg \min_y \sum_{i=1}^n (y - x_i)^2,$$

iar această valoare se calculează cu o formulă analitică (engl., closed form solution) simplă:

$$\hat{y} = \frac{1}{n} \sum_{i=1}^n x_i.$$

În acest exercițiu veți demonstra că atunci când presupunem că „zgomotele“ ϵ_i sunt independente și distribuite conform distribuției $\text{Laplace}(0, b)$, al cărei p.d.f., vă readucem aminte, este

$$f_{\epsilon_i}(x) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right),$$

³⁸⁰Observați că acesta este un caz particular în raport cu regresia liniară de tip LSE prezentat la problema 3.A.

obținem o estimare mai robustă a soluției decât în cazul „zgomotului“ gaussian (vedeți problema 3.A).

La problema 8.B am demonstrat că a găsi estimarea în sensul MLE pentru Y , atunci când presupunem că „zgomotul“ este de tip Laplace, este echivalent cu a găsi valoarea \hat{y} care minimizează suma erorilor în valoare absolută (sau, în modul). Altfel spus,

$$L(y) = \sum_{i=1}^n |y - x_i| \text{ și } \hat{y} = \arg \min_y L(y).$$

O modalitate standard de a calcula minimul funcției de cost / pierdere este să calculăm derivata ei și să o egalăm cu 0. Problema este că funcția $L(y)$ nu este derivabilă pe tot domeniul său de definiție. Totuși, este ușor de observat că $L(y)$ este nederivabilă doar atunci când y ia aceeași valoare ca una dintre instanțele x_i .

- a. Presupunem că instanțele x_i sunt diferite între ele și că sunt sortate în ordine crescătoare ($x_i < x_j$ pentru orice $i < j$).

Calculați $L'(y)$, derivata lui $L(y)$ în raport cu y , pentru cazul când y este situat între două valori consecutive ale lui x (adică, $x_i < y < x_{i+1}$).

Sugestie: Analizați situația x -ilor care sunt mai mici decât y separat de cea a x -ilor care sunt mai mari decât y .

- b. Presupunând că n , numărul instanțelor de antrenament este par, care sunt valorile lui y pentru care $L'(y) = 0$?

- c. Dacă n este impar, [se poate arăta că] nu există nicio valoare a lui y astfel încât $L'(y) = 0$. Totuși, există o valoare y_0 astfel încât $L'(y) < 0$ pentru orice $y < y_0$ și $L'(y) > 0$ pentru orice $y > y_0$. Cât este acest y_0 ?

- d. Valoarea lui \hat{y} este determinată de răspunsurile pe care le-ați obținut la punctele b și c . Dați pe scurt o explicație pentru faptul că această soluție poate fi mai robustă la prezența outlier-elor din date în comparație cu soluția obținută în cazul regresiei de tip LSE (engl., least squared errors).

29.

(Regresia liniară local-ponderată, cazul unidimensional:
o proprietate: „netezirea“ liniară)

□ • ○ ★ CMU, 2014 spring, B. Poczos, A. Singh, HW2, pr. 4.A

Presupunem că lucrăm cu un model de regresie liniară unidimensională local-ponderată, în care „răspunsul“ prezis pentru o valoare oarecare a variabilei $x \in \mathbb{R}$ este $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$, unde parametrii $\hat{\beta}_0$ și $\hat{\beta}_1$ sunt definiți printr-o relație de forma următoare:

$$\hat{\beta}_0, \hat{\beta}_1 = \arg \min_{\beta_0, \beta_1} \left(\sum_{i=1}^n w_i(x) (y_i - \beta_0 - \beta_1 x_i)^2 \right),$$

cu ponderile $w_i(x) \in \mathbb{R}$.³⁸¹

a. Arătați că funcția obiectiv din definiția dată mai sus pentru ponderile $\hat{\beta}_0, \hat{\beta}_1$ poate fi rescrisă astfel:³⁸²

$$(y - X\beta)^\top W(x) (y - X\beta),$$

unde $y \stackrel{\text{not.}}{=} (y_1, \dots, y_n)^\top$, $\beta \stackrel{\text{not.}}{=} (\beta_0, \beta_1)^\top$, $W(x)$ este o matrice diagonală, care are diagonala $(w_1(x), \dots, w_n(x))$, iar

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}.$$

b. Arătați că funcția $\hat{f}(x)$, care a fost definită mai sus prin relația $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$, este o combinație liniară de valorile $\{y_i\}_{i=1}^n$.³⁸³ Mai precis, valoarea $\hat{f}(x)$ poate fi scrisă sub forma $\hat{f}(x) = \sum_{i=1}^n \ell_i(x) y_i = \ell(x)^\top y$, unde $\ell_i(x)$ desemnează o anumită cantitate / funcție definită în raport cu x , iar $\ell(x) \stackrel{\text{not.}}{=} (\ell_1(x), \ell_2(x), \dots, \ell_n(x))^\top$. Cu alte cuvinte, regresia liniară [unidimensională] local-ponderată manifestă proprietatea de *netezire liniară* (cf. CMU, 2007 fall, Carlos Guestrin, HW3, pr. 1).

30. (O extensie / generalizare a regresiei liniare, varianta LSE: cazul „răspunsului“ multivaluat)

□ • • Stanford, 2007 fall, Andrew Ng, HW1, pr. 3

La toate problemele de până acum, în ce privește regresia liniară, am considerat că variabila de „răspuns“ y are o valoare reală. Acum vom presupune că

³⁸¹Deși nu este neapărat necesar [pentru a demonstra rezultatele care fac obiectul acestui exercițiu], am putea considera că pentru fiecare $i \in \{1, \dots, n\}$, ponderea $w_i(x)$ are valoarea $\frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)}$, unde funcția K poate fi, de exemplu, așa-numitul *nucleu RBF*. Această funcție este definită pe \mathbb{R}^d prin relația $K\left(\frac{x}{h}\right) \stackrel{\text{def.}}{=} \exp\left(-\frac{\|x\|^2}{h^2}\right)$, unde $h > 0$ este [pentru problema de regresie] un hiper-parametru, numit *lățimea de bandă* (engl., *bandwidth*). De obicei, într-un astfel de caz, regresia liniară local-ponderată este numită regresie [locală] kernel-izată (engl., *kernelized [local] regression*).

Observație importantă: Vă rugăm să rețineți că termenul *kernel* (sau, *kernel-izat*) în acest context nu are nimic de a face cu metoda de *kernel-izare* (engl., *kernel trick*) care este folosită pentru mașinile cu vectori-suport (engl., Support Vector Machines, SVMs), regresia *ridge kernel-izată* (vedeți problema 9) și regresia logistică *kernel-izată* (problema 17).

³⁸²Remarcați faptul că expresia care urmează este, ca formă, identică cu relația (166) din cazul regresiei liniare (în varianta LSE) local-ponderate cu n atribute / variabile de intrare. Aici vă cerem însă să lucrați în cazul (particular) unidimensional.

³⁸³Pentru cazul regresiei liniare neponderate în varianta LSE (care a fost prezentată la problema 3.A), această proprietatea decurge imediat din relația (156). Pentru cazul regresiei liniare (în varianta LSE) local-ponderate în general (deci nu doar pentru cazul unidimensional despre care discutăm aici), proprietatea decurge imediat din relația (168). Totuși, ca și mai sus, în problema de față vă cerem să elaborați demonstrația pentru cazul (particular) al regresiei local-ponderate unidimensionale.

ni se cere să facem antrenarea pe un set de date având răspunsuri / outputuri multiple pentru toate exemplele de antrenament:

$$\{(x^{(i)}, y^{(i)}) | i = 1, \dots, n\}, x^{(i)} \in \mathbb{R}^d, y^{(i)} \in \mathbb{R}^p.$$

Așadar, pentru oricare exemplu de antrenament, $y^{(i)}$ este un vector cu p componente.

Vrem să folosim un *model liniar* pentru a prezice outputurile, asemănător celui definit prin estimarea în sensul MLE, adică prin minimizarea sumei pătratelor erorilor (engl., least squared errors, LSE),³⁸⁴ și anume specificând acum (în locul vectorului d dimensional) o matrice de parametri $\beta \in \mathbb{R}^{n \times p}$ în relația de definiție a modelului de regresie:

$$y = \beta^\top x.$$

a. Funcția de cost / pierdere în acest caz este

$$J(\beta) = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^p ((\beta^\top x^{(i)})_j - y_j^{(i)})^2.$$

Similar cu modul în care am procedat în cazul univaluat,³⁸⁵ scrieți $J(\beta)$ în notație vectorial-matriceală.

Sugestie: Începeți considerând așa-numita *matrice de design* (X), de dimensiune $n \times d$

$$X = \begin{bmatrix} \quad & (x^{(1)})^\top & \quad \\ \quad & (x^{(2)})^\top & \quad \\ \vdots & & \\ \quad & (x^{(n)})^\top & \quad \end{bmatrix}$$

și matricea de „răspunsuri“ (Y), de dimensiune $n \times p$

$$Y = \begin{bmatrix} \quad & (y^{(1)})^\top & \quad \\ \quad & (y^{(2)})^\top & \quad \\ \vdots & & \\ \quad & (y^{(n)})^\top & \quad \end{bmatrix},$$

iar apoi căutați să vedeți cum poate fi exprimat $J(\beta)$ cu ajutorul acestor matrice.

b. Găsiți *formula analitică* (engl., the closed form solution) pentru acea valoare a lui β care minimizează expresia $J(\beta)$.³⁸⁶ Altfel spus, aceasta înseamnă să obțineți așa-numitele ecuații „normale“ pentru cazul multivaluat al regresiei liniare în varianta LSE.

c. Să presupunem că în loc să considerăm vectorii $y^{(i)}$ ca atare, decidem să calculăm fiecare variabilă / componentă $y_j^{(i)}$ în mod separat, pentru $j = 1, \dots, p$. În această abordare, vom avea p modele liniare individuale, de forma

$$y_j = \beta_j^\top x^{(i)}, \text{ cu } \beta_j \in \mathbb{R}^d \text{ pentru } j = 1, \dots, p.$$

³⁸⁴Vedeți cum am procedat la problema 3.A.

³⁸⁵Adică, atunci când $y \in \mathbb{R}$; vedeți relațiile (154) și (155) de la problema 3.A.

³⁸⁶Această formulă va constitui o generalizare a relației (156) de la regresia liniară LSE univaluată (problema 3.A).

Care este relația dintre soluțiile (adică valorile optime pentru parametrii β_j) acestor p probleme independente de tip “least squared errors” (LSE) și soluția (adică valoarea optimă pentru matricea de parametri β) pentru varianta regresiei liniare LSE multivariate?

31.

(Regresie liniară folosită pentru clasificare; aplicare pe date din \mathbb{R}^2)

*prelucrare de Liviu Ciortuz, după
□ • ○ MIT, 2011 fall, Leslie Kälbling, HW1, pr. 2.2*

În acest exercițiu veți lucra cu un algoritm de regresie liniară de tipul sumei celor mai mici pătrate (engl., least squared errors, LSE), folosind ca etichete / ieșiri y doar valorile ± 1 . În acest fel, reducem problema de *clasificare* la una de *regresie*.³⁸⁷

Concepți un set de exemple de antrenament în planul euclidian (\mathbb{R}^2), care să fie separabile printr-o dreaptă care trece prin originea sistemului de coordinate, astfel încât, atunci când aceste exemple sunt folosite pentru antrenarea unui model de regresie liniară de tip LSE cu etichete binare, să rezulte [în urma antrenării] că unele dintre exemplele de antrenament vor fi clasificate în mod eronat.

1.2.2 Regresia logistică

32.

(Funcția sigmoidală / logistică: definiție și proprietăți de bază)

Liviu Ciortuz, 2024

Considerăm *funcția sigmoidală* (sau *logistică*)

$$\sigma : \mathbb{R} \rightarrow (0, 1), \text{ definită prin } \sigma(z) = \frac{1}{1 + e^{-z}} = \frac{e^z}{1 + e^z}, \text{ pentru orice } z \in \mathbb{R}.$$

a. Elaborați graficul acestei funcții. Dați toate justificările necesare. (*Indicație:* O parte dintre ele sunt enunțate la punctele b.i și b.ii de mai jos.)

b. Demonstrați următoarele trei proprietăți:

- i. $\sigma(-z) = 1 - \sigma(z)$.
- ii. $\sigma'(z) = \sigma(z)[1 - \sigma(z)]$.

iii. Funcția σ este inversabilă (veți justifica de ce este așa!), aşadar există funcția inversă $\sigma^{-1} : (0, 1) \rightarrow \mathbb{R}$, astfel încât

$$\sigma^{-1}(\sigma(z)) = \sigma(\sigma^{-1}(z)) = z \text{ pentru orice } z \in \mathbb{R}. \quad (202)$$

³⁸⁷Această metodă este descrisă în secțiunea 4.1.3 din cartea *Pattern Recognition and Machine Learning* de Ch. Bishop (Springer, 2006) și în secțiunea 4.2 din cartea *The Elements of Statistical Learning* de T. Hastie, R. Tibshirani and J. Friedman (Springer, 2009).

Arătați mai întâi că

$$\sigma^{-1}(z) = \ln \frac{z}{1-z} \text{ pentru orice } z \in (0, 1).$$

Aceasta este aşa-numita funcție *logit*.

Verificați apoi că dubla egalitate (202) este satisfăcută.

La final, faceți graficul funcției logit.

33. (Regresia logistică: particularizare în \mathbb{R}^2 ; deducerea regulilor de actualizare pentru metoda gradientului; regularizare L_1)

*prelucrare de Liviu Ciortuz, după
□ • * CMU, 2018 spring, Nina Balcan, HW3, pr. 1, 4*

A. În prima parte a acestei probleme, veți elabora algoritmul de regresie logistică bazat pe metoda gradientului, în cazul bidimensional. Considerăm setul de date de antrenament $\{(x^i, y^i), i = 1, \dots, n\}$ în care fiecare $x^i \in \mathbb{R}^2$ este un vector de trăsături, iar $y^i \in \{0, 1\}$ este o etichetă binară. Presupunând că folosim un model parametrizat de forma

$$p(y = 1|x; w) = \frac{1}{1 + \exp(-w_0 - w_1 x_1 - w_2 x_2)} = \frac{\exp(w_0 + w_1 x_1 + w_2 x_2)}{1 + \exp(w_0 + w_1 x_1 + w_2 x_2)},$$

obiectivul nostru este să găsim valorile w_i ale parametrilor w_i care maximizează verosimilitatea condițională (M(C)LE) a setului de date de antrenament.

a. Mai jos, vom face *noi* calculul pentru log-verosimilitatea condițională a datelor de antrenament. Relativ la acest calcul, *dumneavoastră* veți elabora câte o scurtă justificare pentru fiecare linie din demonstrație.

$$\ell(w) = \ln \prod_{j=1}^n p(y^j | x^j, w) \tag{203}$$

$$= \sum_{j=1}^n \ln p(y^j | x^j, w) \tag{204}$$

$$= \sum_{j=1}^n \ln (p(y^j = 1 | x^j, w)^{y^j} p(y^j = 0 | x^j, w)^{1-y^j}) \tag{205}$$

$$= \sum_{j=1}^n [y^j \ln p(y^j = 1 | x^j, w) + (1 - y^j) \ln p(y^j = 0 | x^j, w)] \tag{206}$$

$$= \sum_{j=1}^n \left[y^j \ln \frac{\exp(w_0 + w_1 x_1^j + w_2 x_2^j)}{1 + \exp(w_0 + w_1 x_1^j + w_2 x_2^j)} + (1 - y^j) \ln \frac{1}{1 + \exp(w_0 + w_1 x_1^j + w_2 x_2^j)} \right] \tag{207}$$

$$= \sum_{j=1}^n \left[y^j \ln (\exp(w_0 + w_1 x_1^j + w_2 x_2^j)) + \ln \frac{1}{1 + \exp(w_0 + w_1 x_1^j + w_2 x_2^j)} \right] \tag{208}$$

$$= \sum_{j=1}^n [y^j (w_0 + w_1 x_1^j + w_2 x_2^j) - \ln (1 + \exp(w_0 + w_1 x_1^j + w_2 x_2^j))]. \tag{209}$$

b. Acum veți calcula gradientul expresiei de mai sus în raport cu vectorul de ponderi w_0, w_1, w_2). Componentele vectorului gradient sunt derivatele parțiale $\frac{\partial \ell(w)}{\partial w_i}$, unde $\ell(w)$ este funcția de log-verosimilitate condițională de la punctul a. Arătați că

$$\frac{\partial \ell(w)}{\partial w_i} = \sum_{j=1}^n x_i^j (y^j - p(y^j = 1 | x^j; w)). \quad (210)$$

Observație: Am considerat că fiecare instanță/vector x^j (cu $j \in \{1, \dots, n\}$) a fost extins cu componenta $x_0^j = 1$, pentru ca expresia (210) să „acopere“ și cazul derivatei parțiale $\frac{\partial \ell(w)}{\partial w_0}$. În consecință, vectorul gradient va putea fi calculat astfel:

$$\nabla_w \ell(w) = \sum_{j=1}^n [y^j - p(y^j = 1 | x^j; w)] x^j.$$

c. Se poate demonstra că funcția de log-verosimilitate condițională ℓ este concavă,³⁸⁸ deci este ușor să determinăm maximul ei folosind metoda gradientului ascendent. *Algoritm*ul procedează după cum urmează:

Se fixează mai întâi rata de învățare $\eta > 0$ și se aleg niște valori inițiale w_0^0 , w_1^0 și w_2^0 pentru ponderile w_0 , w_1 și respectiv w_2 , după care se execută un număr de iterații [în care partea esențială este aplicarea *regulii de actualizare* a ponderilor w_0 , w_1 și w_2], până când diferența dintre valorile funcției ℓ la două iterații consecutive scade sub un anumit prag (fixat dinainte) $\varepsilon > 0$.

Scrieți regula (regulile) de actualizare a ponderilor w_0 , w_1 și w_2 , care sunt folosite de metoda gradientului ascendent pentru rezolvarea acestei probleme de regresie logistică.

d. Vă readucem aminte că regula de predicție / decizie pentru regresia logistică este următoarea:

dacă $p(y^j = 1 | x^j) > p(y^j = 0 | x^j)$, atunci alege predicția 1,
altfel alege predicția 0.

De ce tip este *granița de decizie* (engl., decision boundary) corespunzătoare modelului de regresie logistică? Scrieți ecuația corespunzătoare graniței de decizie în funcție de w_0 , w_1 , w_2 și x_1^j, x_2^j .

e. Explicați de ce regresia logistică este un clasificator de tip *discriminativ*, spre deosebire de clasificatorii de tip *generativ*, cum este de exemplu algoritmul Bayes Naiv (pentru acesta din urmă, vedeți capitolul *Clasificare bayesiană*).

B. În a doua parte a acestei probleme ne vom concentra asupra estimării parametrilor regresiei logistice în sensul probabilității maxime a posteriori (engl., maximum a posteriori probability, MAP).

f. Considerând o distribuție a priori de tip Laplace $p(w) = \prod_i \frac{1}{2b} e^{-\frac{|w_i|}{b}}$, calculați valoarea parametrului w care maximizează expresia funcției de probabilitate a posteriori. Așadar, veți calcula

³⁸⁸Pentru demonstrația acestei afirmații în cazul general (deci nu doar pentru cazul datelor din \mathbb{R}^2), vedeți problema 14.

$$\hat{w} = \operatorname{argmax}_w \ln[p(w) \prod_j p(y^j | x^j, w)].$$

Sugestie: Rezultatul pe care îl veți obține ar trebui să fie foarte asemănător cu expresia (209), cu singura diferență că noul rezultat va conține un termen suplimentar, care corespunde distribuției a priori.

g. Care este expresia pe care ar trebui să o aibe acum derivata parțială a estimării de probabilitate condiționată maximă a posteriori (M(C)AP)?

Sugestie: Ar trebui să puteți identifica / separa în expresia funcției de probabilitate condiționată maximă a posteriori (M(C)AP) termenul corespunzător distribuției a priori de termenul corespunzător funcției de verosimilitate condițională (M(C)LE). La final ar trebui să obțineți o expresie similară cu expresia (210), având însă un termen suplimentar.

34. (Regresia logistică, chestiuni introductive: exemplificare, folosind datele de la problema 2 de la capitolul *Arbore de decizie*)

■ * Liviu Ciortuz, 2024

Considerăm setul de date de mai jos, pe care l-am preluat de la problema 2 de la capitolul *Arbore de decizie*, cu următoarele schimbări referitoare la *notații* (pentru a face ușor legătura cu rezultatele de la problema 13, la care am descris fundamentele regresiei logistice):

Atributele de intrare au fost redenumite x_1, \dots, x_4 , atributul de ieșire a fost renoscut cu y , iar instanțele de antrenament au fost indexate cu $i \in \{1, \dots, 8\}$ și vor fi desemnate în continuare sub formă $x^{(1)}, \dots, x^{(8)}$. Valoarea atributului x_j pentru instanța de antrenament $x^{(i)}$ va fi desemnată cu $x_j^{(i)}$.

i	x_1	x_2	x_3	x_4	y
1	1	0	0	0	1
2	1	0	1	0	1
3	0	1	0	1	1
4	0	0	0	1	0
5	1	1	1	0	0
6	1	0	1	1	0
7	1	0	0	1	0
8	0	1	0	0	0

În acest exercițiu veți crea un model / clasificator discriminativ pentru acest set de date, aplicând metoda regresiei logistice care a fost descrisă la problema 13.

a. Scrieți funcția de log-verosimilitate condițională $\ell(w)$ corespunzătoare acestui set de date [în contextul regresiei logistice]. *Indicație:* Puteți să aplicați direct formula (178) de la problema 13. Nu uitați să extindeți vectorii $x^{(i)}$ care reprezintă instanțele de antrenament cu componenta $x_0^{(i)} = 1$ corespunzătoare ponderii w_0 !

b. La acest punct veți calcula vectorul gradient pentru funcția de log-verosimilitate condițională de la punctul a.

i. Aplicați direct formula (181) de la problema 13.

ii. Pentru o valoare oarecare a lui $j \in \{0, 1, \dots, 4\}$, calculați în mod clasic derivata parțială a funcției de log-verosimilitate condițională $\ell(w)$ în raport cu ponderea w_j (această derivată parțială se notează cu $\frac{\partial}{\partial w_j} \ell(w)$) și apoi verificați

că rezultatul pe care l-ați obținut coincide cu componenta de pe poziția j din vectorul gradient care a fost calculat mai sus (vedeți punctul i).

c. Considerând că pentru a afla maximul funcției de log-verosimilitate de la punctul a , se aplică metoda gradientului ascendent cu rata de învățare $\eta = 0.1$ (vedeți problema 80 de la capitolul de *Fundamente*), care este valoarea vectorului gradient după prima iterație, dacă la inițializare s-a fixat $w = 0$ (vectorul nul din \mathbb{R}^5)?

d. [Implementare] Folosind metoda gradientului ascendent, determinați valorile optime pentru ponderile w_0, w_1, \dots, w_4 din funcția de log-verosimilitate condițională de la punctul a .

Cum va clasifica modelul astfel obținut instanțele de test din tabelul alăturat?

	x_1	x_2	x_3	x_4	y
U	0	1	1	1	?
V	1	1	0	1	?
W	1	1	0	0	?

e. La acest punct veți calcula matricea hessiană corespunzătoare funcției de log-verosimilitate condițională $\ell(w)$ pe care ati obținut-o la punctul a .

i. Aplicați direct formula (187) de la problema 14:

$$H_w = - \sum_{i=1}^n \underbrace{h(x^{(i)})(1-h(x^{(i)}))}_{\in \mathbb{R}_+} x^{(i)}(x^{(i)})^\top.$$

(Și aici se consideră că vectorii $x^{(i)}$ au fost extinși cu componenta $x_0^{(i)} = 1$ corespunzătoare ponderii w_0 !) Nu este nevoie să aduceți matricea astfel obținută la o formă mai simplă.

ii. Pentru acea valoare $j \in \{0, \dots, 4\}$ care a fost aleasă la punctul b.i și pentru un $k \in \{0, \dots, 4\}$, calculați în mod clasic derivata parțială de ordin secund $\frac{\partial^2}{\partial w_k \partial w_j} \ell(w) \stackrel{\text{def.}}{=} \frac{\partial}{\partial w_k} \left(\frac{\partial}{\partial w_j} \ell(w) \right)$ și apoi verificați că rezultatul pe care l-ați obținut coincide cu componenta de pe linia k și coloana j din matricea hessiană H_w care a fost calculată mai sus (vedeți punctul i).

iii. Un rezultat matematic interesant afirmă că matricea hessiană este simetrică dacă derivatele pațiale de ordin secund sunt continue.³⁸⁹ Verificați (tot prin derivare directă) că pentru o pereche oarecare de indici $j \neq k$ are loc egalitatea $\frac{\partial^2}{\partial w_k \partial w_j} \ell(w) = \frac{\partial^2}{\partial w_j \partial w_k} \ell(w)$.

f. [Implementare] Folosind metoda Newton-Raphson (vedeți problema 80 de la capitolul de *Fundamente*), vectorul gradient $\nabla_w \ell(w)$ pe care l-ați obținut la punctul b și matricea hessiană H obținută la punctul e , determinați valorile optime pentru ponderile w_0, w_1, \dots, w_4 din funcția de cost logistică $J(w)$ de la punctul d .

Cum va clasifica modelul astfel obținut instanțele de test din tabelul de la punctul c?

Comparați numărul de iterații și respectiv timpii de execuție până la convergență pentru cele două metode, adică metoda gradientului și metoda Newton-Raphson, pe setul de date din enunț.

³⁸⁹https://en.wikipedia.org/wiki/Hessian_matrix.

g. Scrieți *funcția de cost logistică* corespunzătoare acestui set de date, notată cu $J(w)$.³⁹⁰ Pentru aceasta, vă cerem să aplicați în mod direct formula (182) de la problema 13. După ce veți aplica această formulă și veți compara rezultatul pe care l-ați obținut cu expresia funcției $\ell(w)$ pe care ați obținut-o la punctul a , veți constata că nu este imediat evidentă egalitatea $J(w) = -\frac{1}{n}\ell(w)$ care a fost demonstrată la problema 13. (Observați că în expresia lui $J(w)$ nu apare funcția σ .)

Ce transformări (adică, operații matematice) trebuie făcute asupra expresiei funcției $\ell(w)$ de la punctul a , pentru ca egalitatea $J(w) = -\frac{1}{n}\ell(w)$ să devină evidentă în mod direct?

Observație: Înând cont de rezultatul care a fost demonstrat la problema 13, minimul funcției de cost $J(w)$ poate fi obținut cu ajutorul metodei gradientului descendente³⁹¹ sau cu metoda Newton-Raphson.³⁹²

35.

(Regresia logistică:
analiza efectului duplicării atributelor de intrare)

■ □ • ○ CMU, 2011 spring, Tom Mitchell, midterm, pr 5.3

Vom considera o problemă de clasificare binară pe date caracterizate de variabilă / atributul $X_1 \in \{0, 1\}$ și eticheta $Y \in \{0, 1\}$.

Fie un set de date de antrenament format din n exemple: $D_1 = \{(x_1^1, y^1), \dots, (x_1^n, y^n)\}$. Presupunem că generăm un alt set de date de antrenament, D_2 , format tot din n exemple, $D_2 = \{(x_1^1, x_2^1, y^1), \dots, (x_1^n, x_2^n, y^n)\}$, unde în fiecare exemplu x_1 și y sunt aceleași ca în D_1 , iar x_2 este copia lui x_1 .

Învățăm un model de regresie logistică folosind setul de date D_1 ; acest model va avea doi parametri: w_0 și w_1 . De asemenea, învățăm un alt model de regresie logistică folosind setul de date D_2 ; acesta va avea trei parametri: w_0 , w_1 și w_2 .

Mai întâi, scrieți *funcția obiectiv* pe care o folosim pentru a estima (în sensul maximizării verosimilității condiționale) parametrii (w_0, w_1) și respectiv (w_0, w_1, w_2) , pe baza datelor de antrenament.

Apoi, analizând cele două funcții obiectiv, precizați care este *relația* dintre seturile de parametri (w_0, w_1) și (w_0, w_1, w_2) , care se estimează din datele D_1 și respectiv D_2 . Pornind de la această relație, argumentați de ce (sau cum anume) va fi afectată regresia logistică (sau *nu* va fi afectată) de duplicarea variabilei X_1 .

³⁹⁰Nu există nicio legătură între această funcție și criteriul J de la algoritmul K-means (vedeți capitolul de Clusterizare). Este pur și simplu o coincidență de notație.

³⁹¹Concret, puteți folosi același vector gradient care a fost calculat la punctul *b.i.*, dar regula de actualizare a ponderilor în acest caz va fi $w \leftarrow w - \frac{\eta}{n} \nabla_w \ell(w)$, unde $\eta > 0$ este rata de învățare, iar n este numărul instanțelor de antrenament.

³⁹²Forma regulii de actualizare pentru metoda lui Newton-Raphson este aceeași, indiferent dacă se calculează un punct de maxim pentru o funcție concavă ori un punct de minim pentru o funcție convexă.

36.

(Regresia liniară și regresia logistică:
definiții [,revizitate“],
o interesantă proprietate comună)

*prelucrare de Liviu Ciortuz, după
■ □ • ○ CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW2, pr. 4*

Dată fiind o instanță (sau, un input) $X \in \mathbb{R}^d$ împreună cu „răspunsul“ / outputul corespunzător $Y \in \mathbb{R}$, *regresia liniară* [cu „zgomot“ gaussian] construiește un model de forma

$$Y|X \sim \text{Normal}(\mu(X), \sigma^2),$$

unde media $\mu(X)$ este o funcție liniară de componente / atributelor inputului: $\mu(X) = \theta^\top X = \theta_0 + \theta_1 X_1 + \dots + \theta_d X_d$.

Dacă $Y \in \{0, 1\}$, *regresia logistică* — care, spre deosebire de regresia liniară servește pentru *clasificare* — modelează outputul Y astfel:

$$Y|X \sim \text{Bernoulli}(h_\theta(X)),$$

unde $h_\theta(X)$, parametrul acestei distribuții Bernoulli, este obținut din $\theta^\top X$ aplicând funcția *logistică* / *sigmoidală*:

$$h_\theta(X) = g(\theta^\top X),$$

unde prin g am notat *funcția logistică*

$$g(z) \stackrel{\text{def.}}{=} \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}}$$

sau, echivalent, folosind funcția *logit* (care este inversa funcției *logistice* / *sigmoidale*):³⁹³

$$\text{logit}(h_\theta(X)) \stackrel{\text{def.}}{=} \ln \frac{h_\theta(X)}{1 - h_\theta(X)} = \theta^\top X$$

Comentariu: Definițiile date mai sus pun în evidență un anumit „paralelism“ pentru cele două modele de regresie. În această problemă,

- mai întâi, veți putea vedea încă o *similaritate*, și anume între *vectorii gradient* corespunzători funcțiilor de log-verosimilitate condițională pentru cele două metode de regresie. Ne referim aici la expresia $\nabla_{\theta}\ell(\theta) = \sum_{i=1}^n (y_i - h_\theta(x_i))x_i$ (care corespunde relației (181) de la problema 13) pentru gradientul regresiei logistic și la expresia $\nabla_{\theta}\ell(\theta) = \sum_{i=1}^n (y_i - \theta^\top x_i)x_i$ pentru gradientul regresiei liniare, conform termenului principal din relația (164), care a fost obținută la problema 6;
- iar acum veți demonstra o *proprietate de tip probabilist*, care se scrie în mod identic(!) pentru cele două modele de regresie și care folosește media condițională a outputului Y în raport cu inputul X și cu $\hat{\theta}$, estimarea în sens MLE pentru parametrul θ .

Veți considera setul de date de antrenament $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$.

Demonstrați că pentru fiecare dintre cele două modele de regresie de mai sus, estimarea de verosimilitate maximă a parametrului θ (notație: $\hat{\theta}$) satisfac următoarea proprietate:

$$\sum_{i=1}^n y_i x_i = \sum_{i=1}^n E[Y|X = x_i, \theta = \hat{\theta}] x_i.$$

³⁹³Într-adevăr, $\sigma(z) = \frac{1}{1 + e^{-z}} = \frac{e^z}{1 + e^z} = y \in (0, 1) \Rightarrow e^z(1 - y) = y \Rightarrow z = \ln \frac{y}{1 - y}$.

Observații:

1. Remarcați faptul că lui y_i din partea stângă a acestei egalități îi corespunde în partea dreaptă o medie, și anume media condițională a outputului Y , dat fiind inputul $X = x_i$ și $\hat{\theta}$, estimarea în sens MLE pentru parametrul θ .
2. Datorită faptului că $y_i \in \{0, 1\}$, rezultă că membrul stâng al acestei egalități este de fapt suma acelor instanțe x_i pentru care $y_i = 1$. Membrul drept al egalității este o sumă ponderată a tuturor instanțelor x_i . Ponderile respective sunt mediile variabilelor aleatoare condiționate (de tip gaussian, respectiv de tip Bernoulli) scrise în mod unitar sub forma $Y|X = x_i, \theta = \hat{\theta}$.

37. (Regresia logistică multi-class (i.e., regresia softmax): echivalență cu un caz particular de mixtură de distribuții gaussiene multidimensionale)

prelucrare de Liviu Ciortuz, după

**■ □ • ○ MIT, 2001 fall, Tommi Jaakkola, HW2, pr. 3.7
MIT, 2016 fall, R. Barzilay, S. Sra, Weekly Exercises, week 4, pr. 5.bc**

Introducere: La problema 18 am introdus modelul de regresie logistică multi-class (i.e., regresia softmax). Distribuția gaussiană multidimensională a fost introdusă la problema 34 de la capitolul *Fundamente*. Definiția modelului de tip mixtură de distribuții gaussiene multidimensionale a fost prezentată la problema 118 tot de la capitolul *Fundamente*.

În acest exercițiu veți demonstra că pentru orice $K \in \mathbb{N}$, $K \geq 2$, modelul probabilist învățat de regresia softmax este echivalent cu un anumit tip de mixtură de distribuții gaussiene multi-variate.

Considerăm un model de regresie softmax (adică, regresie logistică multi-class) cu K clase, cu parametrii desemnați astfel: $w_i \in \mathbb{R}^d$ și $w_{i0} \in \mathbb{R}$ corespund clasei i , pentru $i = 1, \dots, K$, iar $(w_i)_j \stackrel{\text{not.}}{=} w_{ij}$, pentru $j = 1, \dots, d$. Folosind aceste notării, distribuția de probabilitate a posteriori softmax este definită astfel:³⁹⁴

$$Pr(y|x) = \frac{e^{w_y^\top x + w_{y0}}}{\sum_i e^{w_i^\top x + w_{i0}}}.$$

Arătați că acestui model softmax îi putem asocia o mixtură de distribuții gaussiene multi-variate de parametri $(\pi_i, \mu_i, \Sigma_i)_{i=1, \dots, K}$ astfel încât distribuția de probabilitate a posteriori softmax de mai sus să coincidă cu distribuția de probabilitate a posteriori a modelului de mixtură de distribuții gaussiene:

$$Pr(y|x; (\pi_i, \mu_i = w_i, \Sigma_i)_{i=1, \dots, K}) = \frac{\pi_y |\Sigma_y|^{-d/2} \exp\left(-\frac{1}{2}(x - \mu_y)^\top \Sigma_y^{-1}(x - \mu_y)\right)}{\sum_i \pi_i |\Sigma_i|^{-d/2} \exp\left(-\frac{1}{2}(x - \mu_i)^\top \Sigma_i^{-1}(x - \mu_i)\right)}.$$

³⁹⁴Nu este dificil de făcut legătura între definiția de aici și cea de la problema 18, chiar dacă ele diferă ușor. De asemenea, folosind definiția de aici se poate demonstra ușor că separatorii decizionali corespunzători regresiei softmax sunt de tip liniar.

38. (Comparații între regresia logistică și alți clasificatori... și încă o chestiune, legată de regresia liniară)

• CMU, 2009 spring, Tom Mitchell, midterm, pr. 1.1

Explicați pe scurt de ce

- a. pentru a rezolva o anumită problemă de clasificare automată, s-ar putea să fie mai bine să folosim [un algoritm de învățare de] arbori de decizie decât regresia logistică.
- b. pentru a rezolva o anumită problemă de clasificare automată, s-ar putea să fie mai bine să folosim regresia logistică decât algoritmul Bayes Naiv.
- c. atunci când aplicăm regresia liniară, selectăm acele valori ale parametrilor care minimizează suma pătratelor erorilor la antrenare.

39. (Regresia liniară și regresia logistică; comparație între metoda gradientului și metoda lui Newton: Adevărat sau Fals?)

• Stanford, 2014 fall, Andrew Ng, midterm, pr. 6.gba

a. Presupunem că tu și prietenul tău vreți să folosiți un model de *regresie liniară* pentru a prezice prețul locuințelor. Tu folosești în modelul tău trăsăturile $x_0 = 1$, $x_1 = \text{mărimea locuinței în metri pătrați și } x_2 = \text{înălțimea acoperișului măsurată în metri}$. Să zicem că prietenul tău face aceeași analiză, folosind exact același set de date de antrenament, doar că el reprezintă datele folosind trăsăturile $x'_0 = 1$, $x'_1 = x_1$ și $x'_2 = \text{înălțimea acoperișului măsurată în cm (așadar, } x'_2 = 100x_2\text{)}$.

i. Să zicem că atât tu cât și prietenul tău rezolvați problema de regresie liniară folosind metoda analitică (adică, așa-numitele *ecuații normale*, conform problemei 3). Presupunem că nu suntem într-un *cas de degenerare*, așadar aceste ecuații ne dau o soluție unică pentru parametrii regresiei liniare. Tu obții parametrii / soluția $\theta_0, \theta_1, \theta_2$, în vreme ce prietenul tău obține $\theta'_0, \theta'_1, \theta'_2$. Urmează că $\theta'_0 = \theta_0$, $\theta'_1 = \theta_1$, $\theta'_2 = \frac{1}{100}\theta_2$. Adevărat sau Fals?

ii. Presupunem că atât tu cât și prietenul tău rezolvați problema de regresie liniară inițializând parametrii cu 0 și comparând rezultatele pe care le-ați obținut după rularea unei singure iterării a algoritmului / metodei *gradientului [descendent]*, varianta *batch*. [LC: Amândoi folosiți aceeași rată a învățării η .] Tu obții parametrii / soluția $\theta_0, \theta_1, \theta_2$, iar prietenul tău obține $\theta'_0, \theta'_1, \theta'_2$. Urmează că $\theta'_0 = \theta_0$, $\theta'_1 = \theta_1$, $\theta'_2 = \frac{1}{100}\theta_2$. Adevărat sau Fals?

Observație: Punctele b și c care urmează în continuare nu au nicio legătură cu punctul a de mai sus.

b. Dacă utilizăm metoda *gradientului [ascendent]* în varianta *stochastică* pentru antrenarea *regresiei logistice* și folosim o rată fixă a învățării, atunci algoritmul va converge neapărat [în mod exact] la soluția optimă pentru parametrii θ , și anume $\hat{\theta} = \operatorname{argmax}_{\theta} \prod_{i=1}^n p(y^{(i)} | x^{(i)}; \theta)$, presupunând că s-a ales o rată a învățării rezonabilă. Adevărat sau Fals?

- c. Presupunem că dispuneți de o implementare pentru *metoda lui Newton* și metoda *gradientului [descendent]* pentru aflarea punctului de optim al unei funcții convexe. Să zicem că o iterație a metodei lui Newton durează dublu față de o iterație a metodei gradientului [descendent]. Prin urmare, rezultă că metoda gradientului [descendent] va converge mai repede la soluția optimă. Adevărat sau Fals?

1.2.3 Modele liniare generalizate

*Introducere:*³⁹⁵

Considerăm că avem o *problemă* de clasificare sau de regresie, în care am dori să prezicem valoarea unei variabile aleatoare y în funcție de [o altă variabilă aleatoare] x . Pentru a deriva un model liniar generalizat (engl., Generalized Linear Model, GLM) pentru această problemă, vom considera următoarele *trei presupozitii* în legătură cu distribuția condițională a lui y în raport cu x , precum și în legătură cu modelul nostru:

1. $y|x; \theta \sim \text{ExponentialFamily}(\eta)$. Așadar, dacă fiind x și θ , variabila y urmează o anumită distribuție din familia exponențială, caracterizată de parametrul η :³⁹⁶

$$p(y|\eta) = b(y) \exp(\eta^\top T(y) - a(\eta)).$$

2. *Scopul* nostru este ca, dat fiind x , să prezicem media valorii lui $T(y)$ în raport cu x . În cele mai multe dintre exemplele didactice, se consideră $T(y) = y$. Așadar, dorim ca predicția $h(x)$ generată (ca output) de către ipoteza învățată h să satisfacă relația $h(x) = E[y|x]$.³⁹⁷
3. Legătura dintre *parametrul natural* η și variabila de intrare x este de tip liniar: $\eta = \theta^\top x$. (Sau, dacă η este vector, atunci $\eta_i = \theta_i^\top x$.)

Cea de-a treia presupozitie poate părea mai puțin justificată decât celelalte două, și ar fi probabil mai indicat să o gândim ca fiind un element de *concepție* (engl., design choice) pentru modelele liniare generalizate (engl., Generalized Linear Models, GLM).

Aceste trei presupozitii / elemente de concepție ne permit să deducem o foarte elegantă clasă de algoritmi de învățare, numită GLM, care au multe proprietăți dezirabile, cum ar fi de pildă facilitatea învățării. Mai mult, modelele rezultate sunt adeseori foarte potrivite pentru a „caracteriza“ diverse tipuri de distribuții peste y . De exemplu, se poate arăta că atât regresia logistică cât și regresia liniară (în varianta celor mai mici pătrate) pot fi văzute ca fiind instanțe ale clasei de algoritmi de învățare automată GLM.

Vom mai adăga câteva *chestiuni de terminologie*, precizăm că funcția g care returnează media distribuției în raport cu parametrul natural η (adică, $g(\eta) =$

³⁹⁵Cf. Andrew Ng, Stanford University, Machine learning course, Lecture notes, Part III — Generalized Linear Models.

³⁹⁶Vedeți problema 122 de la capitolul de *Fundamente*.

³⁹⁷Veți putea observa că această presupunere este satisfăcută atât în cazul regresiei logistice cât și în cazul regresiei liniare, de către definiția respectivă aleasă pentru $h_\theta(x)$. De exemplu, la regresia logistică avem $h_\theta(x) \stackrel{\text{def.}}{=} \sigma(\theta^\top x) = p(y = 1|x; \theta) = 0 \cdot p(y = 0|x; \theta) + 1 \cdot p(y = 1|x; \theta) = E[y|x; \theta]$.

$E[T(y); \eta]$) este numită *funcția de răspuns canonică* (engl., canonical response function). Inversa sa, g^{-1} , este numită *funcția de legătură canonică* (engl., canonical link function). Se poate demonstra ușor că funcția de răspuns canonică pentru distribuția gaussiană [cu varianță 1] este funcția identitate, iar pentru distribuția Bernoulli este funcția logistică.

40. (GLM: particularizare pentru cazul distribuției geometrice și al distribuției Poisson)

• · Stanford, 2014 fall, Andrew Ng, midterm, pr. 3

Stanford, 2008 fall, Andrew Ng, HW1, pr. 3.abc

University of Chicago, 2004 spring, Michael Eichler, HW3, pr. 4.bd

Vă readucem aminte că forma standard a familiei de distribuții exponențiale este

$$p(y|\eta) = b(y) \exp(\eta^\top T(y) - a(\eta)).$$

De asemenea, vă readucem aminte că distribuția geometrică de parametru ϕ are funcția masă de probabilitate (p.m.f.) definită în felul următor:

$$p(y|\phi) = (1-\phi)^{y-1} \phi, \text{ pentru } y = 1, 2, \dots.$$

a. Arătați că distribuția geometrică face parte din familia de distribuții exponențiale și precizați valorile corespunzătoare pentru $b(y)$, η , $T(y)$ și $a(\eta)$.

b. Presupunem că vrem să facem regresie folosind un model liniar generalizat (engl., Generalized Liniar Model, GLM) cu o variabilă de răspuns geometrică. Precizați care este – în acest caz – *funcția canonică de răspuns*. (Puteți utiliza faptul că media distribuției geometrice este $\text{by } 1/\phi$.)

c. Vă readucem aminte că în modelul GLM, se consideră că $\eta = \theta^\top x$, cu θ și x din \mathbb{R}^d . De asemenea, dat fiind un set de date de antrenament $\{(x_i, y_i)\}_{i=1}^m$ cu $x_i \in \mathbb{R}^d$ și $y_i \in \mathbb{R}$, se știe că log-verosimilitatea unui exemplu (x_i, y_i) este $\ln p(x_i, y_i)$. Calculați derivata acestei log-verosimilități în raport cu θ_j , pentru a obține regula gradientului ascendent în varianta stochastică, în vederea antrenării unui model GLM cu răspunsuri geometrice y și funcție canonică de răspuns.

d. Calculați matricea hessiană H pentru $\ell(\theta) = \sum_{i=1}^m \ln p(y_i|x_i; \theta)$, funcția de log-verosimilitate a datelor de antrenament, și precizați cum va arăta aplicarea unui pas al metodei lui Newton pentru maximizarea acestei funcții de log-verosimilitate.

e. Vă cerem să satisfaceti din nou cerințele de la punctele $a-d$ de mai sus, însă de data aceasta lucrând cu distribuția Poisson de parametru λ . Vă readucem aminte că pentru această distribuție funcția masă de probabilitate are forma

$$p(y|\lambda) = \frac{e^{-\lambda} \lambda^y}{y!},$$

iar media este egală cu λ .

41.

(GLM: particularizare pentru cazul distribuției gaussiene unidimensionale)

 • · Stanford, 2015 fall, Andrew Ng, midterm, pr. 3

Vă readucem aminte că forma standard a familiei de distribuții exponențiale este

$$p(y|\eta) = b(y) \exp(\eta^\top T(y) - a(\eta)).$$

De asemenea, vă readucem aminte că funcția densitate de probabilitate pentru distribuția gaussiană [unidimensională] este definită în felul următor:

$$p(y|\mu, \sigma^2) = \frac{1}{\sigma^2} \exp\left(-\frac{1}{2\sigma^2}(y - \mu)^2\right).$$

a. Arătați că distribuția gaussiană face parte din familia de distribuții exponențiale. Specificați valorile corespunzătoare pentru $b(y)$, η , $T(y)$, și $a(\eta)$.

Indicație: η și $T(y)$ vor fi vectori bidimensionali; veți folosi notația $\eta = (\eta_1, \eta_2)^\top$. Veți exprima $a(\eta)$ în funcție de η_1 și η_2 .

b. Presupunem că dispunem de setul de date de antrenament $\{(x_i, y_i)\}_{i=1}^m$, cu $x_i \in \mathbb{R}^d$ și $y \in \mathbb{R}$. Pornind de la $p(y|\eta)$, forma standard a familiei exponențiale dată mai sus, calculați expresia [generală a] matricei hessiene pentru funcția de log-verosimilitate $\ell(\theta) = \sum_{i=1}^m \ln p(y_i|x_i; \theta)$, cu $\theta \in \mathbb{R}^d$. (Vă readucem aminte că în modelul GLM se lucrează cu $\eta = \theta^\top x$.) Veți scrie expresia acestei matrice hessiene în raport cu x , η_1 și η_2 .

c. Folosind rezultatul de la punctul b, arătați că matricea hessiană [pentru funcția de log-verosimilitate $\ell(\theta)$] este negativ semidefinită, adică $z^\top H z \leq 0$ pentru orice $z \in \mathbb{R}^d$.

42.

(GLM: calculul funcției de log-verosimilitate în cazul general, condiții suficiente pentru concavitate; calculul gradientului; calculul matricei hessiene)

prelucrare de Liviu Ciortuz, după

 • · Stanford, 2009 fall, Andrew Ng, practice midterm, pr. 1
Stanford, 2008 fall, Andrew Ng, HW1, pr. 3.d

Vă readucem aminte că în modelul liniar generalizat (engl., generalized linear models, GLM) se presupune că variabila de răspuns y (condiționată în raport cu x) urmează o distribuție care face parte din *familia de distribuții exponențiale*:

$$p(y|\eta) = b(y) \exp(\eta^\top T(y) - a(\eta)), \quad (211)$$

unde $\eta = \theta^\top x$. În acest exercițiu vom presupune că $\eta \in \mathbb{R}$.

a. Știți că fiind dat setul de date de antrenament $\{(x_i, y_i)\}_{i=1}^m$, funcția de log-verosimilitate [condițională] se definește astfel:

$$\ell(\theta) = \sum_{i=1}^m \ln p(y_i|x_i; \theta).$$

Formulați un set de condiții pe care trebuie să le satisfacă $b(y)$, $T(y)$ și $a(\eta)$ din relația (211) în aşa fel încât funcția $\ell(\theta)$ să fie concavă (asigurându-ne astfel că ea are o valoare maximă unică). Condițiile pe care le veți formula trebuie să fie rezonabile și, de asemenea, trebuie să fie cât se poate de laxe (engl., weak).³⁹⁸

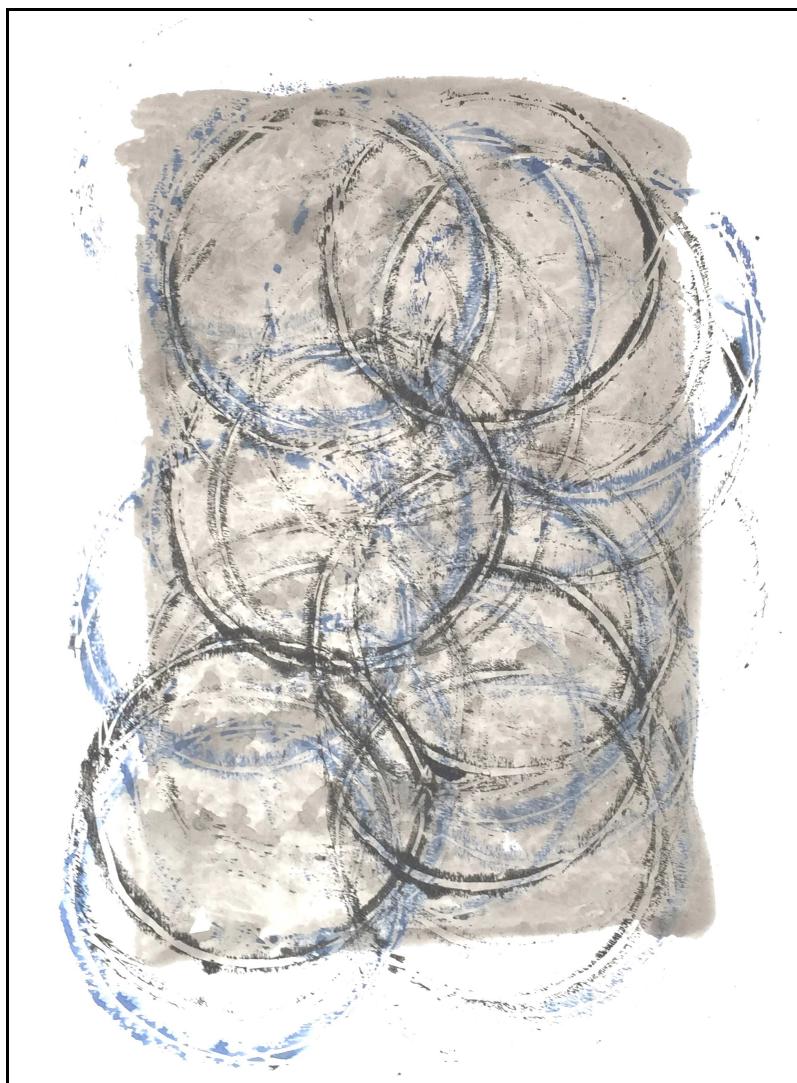
- b. Verificați că în cazul în care variabila de răspuns urmează distribuția gaussiană standard, avem $b(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$, $T(y) = y$ și $a(\eta) = \frac{\eta^2}{2}$. Verificați de asemenea că setul / sistemul de condiții pe care le-ați formulat la punctul a este satisfăcut în acest caz particular.
- c. Presupunem că folosim un model GLM având ca *variabilă de răspuns* orice membru al familiei de distribuții exponențiale, cu $T(y) = y$ și *funcția de răspuns canonic* pentru această familie.

Arătați că atunci când trebuie să aplicăm metoda gradientului ascendent în varianta stochastică pentru a determina maximul funcției de log-verosimilitate $\ln p(y|X, \theta)$, obținem următoarea *regulă de actualizare*:

$$\theta_i \leftarrow \theta_i - \alpha(h(x) - y)x_i,$$

unde α este *rata de învățare*.

³⁹⁸De pildă, răspunsul „orice $b(y)$, $T(y)$ și $a(\eta)$ astfel încât funcția $\ell(\theta)$ să fie concavă“ nu este rezonabil. Tot aşa, condiții foarte restrictive, care se aplică – de exemplu – doar pentru unele modele particulare de GLMs, nu sunt rezonabile.



© M. Romanică

2 Clasificare bayesiană

Sumar

Noțiuni preliminare

- probabilități și probabilități condiționate;
- formula lui Bayes: ex. 5.b; cap. *Fundamente*, ex. 6, ex. 7, ex. 94, ex. 95;
- independența [condițională a] evenimentelor aleatoare: cap. *Fundamente*, ex. 4, ex. 91, ex. 92;
- independența [condițională a] variabilelor aleatoare: ex. 9, ex. 10, ex. 12, ex. 33-43; vedeti și cap. *Fundamente*, ex. 15, ex. 30, ex. 99.b, ex. 108, ex. 106;
- distribuții probabiliste comune, marginale și condiționale: ex. 6, ex. 10, ex. 12, ex. 33; vedeti și cap. *Fundamente*, ex. 13, ex. 14;
- distribuția gaussiană: de la cap. *Fundamente*, ex. 32, ex. 33 (pentru cazul unidimensional), ex. 35 (pentru cazul bidimensional), ex. 20, ex. 34, ex. 36, ex. 37 (pentru cazul multidimensional);
- estimarea parametrilor pentru distribuții de tip Bernoulli (ex. 5.a), categorial și gaussian (ex. 15.a);³⁹⁹
- ipoteze MAP vs. ipoteze ML:
formulare [ca soluții la] probleme de optimizare:⁴⁰⁰ ex. 26;
exemplificare: ex. 1, ex. 2, ex. 3, ex. 24, ex. 25, ex. 42;
exemplificare în cazul arborilor de decizie: ex. 4;
- regresia logistică, chestiuni introductive:⁴⁰¹ de la cap. *Metode de regresie*, ex. 13.

Algoritmi de clasificare bayesiană

- Algoritmul Bayes Naiv și algoritmul Bayes Optimal:⁴⁰²
formulare ca probleme de optimizare / estimare în sens MAP: cartea ML, pag. 157;
pseudo-codul algoritmului Bayes Naiv (cf. cartea ML, pag. 177):

Training:

```
for each value  $v_j$  of the output attribute  
 $\hat{P}(v_j) \leftarrow \text{estimate } P(v_j)$ 
```

³⁹⁹ De la cap. *Fundamente*, pentru estimarea parametrului unei distribuții Bernoulli vedeti ex. 43 și ex. 124.a, pentru estimarea parametrilor unei distribuții categoriale vedeti ex. 44, iar pentru estimarea parametrilor unei distribuții gaussiene vedeti ex. 50, ex. 51, ex. 134 (pentru cazul unidimensional) și ex. 53 (pentru cazul multidimensional).

⁴⁰⁰ Vedeti cartea ML, pag. 156-157.

⁴⁰¹ Vedeti draftul capitolului suplimentar pentru cartea ML a lui T. Mitchell, *Generative and discriminative classifiers: Naive Bayes and logistic regression* (în special secțiunea 3).

⁴⁰² La secțiunea aceasta, precum și la următoarea secțiune, considerăm (implicit) că toate variabilele de intrare sunt de tip Bernoulli sau, mai general, de tip categorial. După aceea vom considera și variabile de intrare de tip continuu, în genere de tip gaussian. Variabila de ieșire se consideră întotdeauna de tip Bernoulli / categorial.

for each value a_i of each input attribute a
 $\hat{P}(a_i|v_j) \leftarrow \text{estimate } P(a_i|v_j)$

Classification of the test instance (a_1, a_2, \dots, a_n) :

$$v_{NB} = \operatorname{argmax}_{v_j \in V} \hat{P}(v_j) \prod_{a_i \in x} \hat{P}(a_i|v_j).$$

*Justificarea regulii de decizie pentru algoritmul Bayes Naiv:*⁴⁰³

$$\begin{aligned} v_{MAP} &= \operatorname{argmax}_{v_j \in V} P(v_j|a_1, a_2 \dots a_n) = \operatorname{argmax}_{v_j \in V} \frac{P(a_1, a_2 \dots a_n|v_j)P(v_j)}{P(a_1, a_2 \dots a_n)} \\ &= \operatorname{argmax}_{v_j \in V} P(a_1, a_2 \dots a_n|v_j)P(v_j) \stackrel{\text{indep. cdt.}}{=} \operatorname{argmax}_{v_j \in V} \prod_i P(a_i|v_j)P(v_j) \stackrel{\text{not.}}{=} v_{NB}. \end{aligned}$$

- exemple de aplicare: ex. 5, ex. 6, ex. 7, ex. 27, ex. 28, ex. 29;
- aplicarea / adaptarea algoritmului Bayes Naiv pentru clasificare de texte:⁴⁰⁴ ex. 8, ex. 30;
- folosirea regulii “add-one” [a lui Laplace] pentru „netezirea” parametrilor: ex. 8, ex. 31;
- calculul ratei medii a erorilor pentru algoritmii Bayes Naiv și Bayes Optimal: ex. 10, ex. 11, ex. 32, ex. 33, ex. 34, ex. 35.a-d, ex. 36, ex. 43;
- evidențierea grafică a neconcordanței predicțiilor făcute de clasificatorii Bayes Naiv și Bayes Optimal: ex. 12;
- exemple de clasificatori bayesieni care combină avantajele algoritmilor Bayes Naiv și Bayes Optimal: ex. 9, ex. 35.e.

Proprietăți ale algoritmilor Bayes Naiv și Bayes Optimal

- (P0) dacă proprietatea de independentă condițională a atributelor de intrare în raport cu variabila de ieșire se verifică, atunci rezultatele produse de către cei doi algoritmi (Bayes Naiv și Bayes Optimal) în faza de testare coincid; în caz contrar, rezultatele celor doi algoritmi pot să nu coincidă;
- (P1) numărul de parametri necesari de estimat din date: liniar pentru Bayes Naiv $(2d+1)$ și exponențial pentru Bayes Optimal $(2^{d+1}-1)$:⁴⁰⁵ ex. 7.e, ex. 27.b, ex. 29.ab, ex. 35.ac, ex. 37;
- (P2) complexitatea algoritmului Bayes Naiv:

complexitatea de spațiu: $\mathcal{O}(dK)$

complexitatea de timp:

la antrenare: $\mathcal{O}(dn)$

la testare: $\mathcal{O}(dK)$,

⁴⁰³ Justificarea regulii de decizie pentru algoritmul Bayes Comun / Optimal:

$$v_{MAP} = \operatorname{argmax}_{v_j \in V} P(v_j|a_1, a_2 \dots a_n) = \dots$$

$$= \operatorname{argmax}_{v_j \in V} P(a_1, a_2 \dots a_n|v_j)P(v_j) = \operatorname{argmax}_{v_j \in V} P(a_1, a_2 \dots a_n, v_j) \stackrel{\text{not.}}{=} v_{JB}.$$

⁴⁰⁴ Atenție: Noi am folosit aici versiunea de bază a algoritmului Bayes Naiv; varianta “bag of words” (vedeți cartea Machine Learning a lui Tom Mitchell, pag. 183) diferă ușor de aceasta.

⁴⁰⁵ Numărul de parametri indicați în paranteze se referă la cazul când atât atributele de intrare cât și atributul de ieșire sunt de tip Bernoulli.

unde n este numărul de exemple, d este numărul de atrbute de intrare, iar K este numărul de clase;

- (P3) algoritmul Bayes Optimal poate produce eroare [la clasificare] din cauza faptului că ia decizia în sensul unui vot majoritar. Algoritmul Bayes Naiv are și el această „sursă” de eroare; în plus el poate produce eroare și din cauza faptului că lucrează cu presupoziția de independentă condițională (care nu este satisfăcută în mod neapărat);
- (P4) Algoritmul Bayes Naiv nu garantează obținerea ratei medii a erorii 0 la antrenare pe date consistente: ex. 11.a;⁴⁰⁶
- (P5) este posibil ca rata medie a erorii la antrenare produsă de Bayes Naiv să fie 0 chiar dacă independentă condițională este încălcată: ex. 36;
- (P6) acuratețea [la clasificare] algoritmului Bayes Naiv scade atunci când unul sau mai multe atrbute de intrare sunt duplicate: ex. 10.d, ex. 33.def; în schimb, algoritmul Bayes Optimal nu este sensibil la duplicarea atrbutelor de intrare; atât Bayes Naiv cât și Bayes Optimal sunt sensibili la duplicarea exemplelor de antrenament;
- (P7) în cazul „învățării” unei funcții booleene (oarecare), rata medie a erorii produse la antrenare de către algoritmul Bayes Optimal (spre deosebire de Bayes Naiv!) este 0: ex. 35.d;⁴⁰⁷
- (P8) complexitatea de eșantionare: de ordin logaritmic pentru Bayes Naiv și de ordin exponențial pentru Bayes Optimal: ex. 13;
- (P9) corespondența dintre regula de decizie a algoritmului Bayes Naiv (când toate variabilele de intrare sunt de tip Bernoulli) și regula de decizie a *regresiei logistice* și, în consecință, liniaritatea granițelor de decizie: ex. 14, ex. 38.c.
- o perspectivă info-teoretică asupra algoritmului Bayes Naiv: ex. 147 (și ex. 146) de la capitolul de *Foundamente*.⁴⁰⁸
- *comparații* între algoritmul Bayes Naiv și alți algoritmi de clasificare automată: ex. 41, ex. 43.

Algoritmii Bayes Naiv și Bayes Optimal cu variabile de intrare de tip gaussian

- Aplicare: G[N]B: ex. 15, ex. 45 și ex. 53; GJB: ex. 50, ex. 51 și ex. 52; GNB vs. GJB: ex. 21.
- Numărul de parametri necesari de estimat din date: ex. 47.
- Proprietăți:
 - (P0') presupunem că variabila de ieșire este booleană, i.e. ia valorile 0 sau 1; dacă pentru orice atrbut de intrare, variabilele condiționale $X_i|Y = 0$ și $X_i|Y = 1$ au distribuții gaussiene de varianțe egale ($\sigma_{i0} = \sigma_{i1}$), atunci regula de decizie GNB (Gaussian Naive Bayes) este echivalentă (ca formă) cu cea a regresiei logistice, deci separarea realizată de către algoritmul GNB este de

⁴⁰⁶În contextul învățării conceptului XOR (care reprezintă un set de date consistent), algoritmul Bayes Naiv obține pentru rata medie a erorii la antrenare valoarea maximă posibilă în cazul clasificării binare: 50%!

⁴⁰⁷Rezultatul este valabil pentru cazul (mai general) al funcțiilor definite în sens matematic (corespondențe de tip “many-to-one”) cu argumente discrete și valori discrete.

⁴⁰⁸Vedeți legătura cu relația de echivalență max likelihood \Leftrightarrow min cross-entropy.

formă liniară: demonstrație: ex. 17; exemplificare în \mathbb{R} : ex. 45.a; exemplificare în \mathbb{R}^2 : ex. 46.c;

- (P1') similar, presupunem că variabila de ieșire este booleană; dacă variabilele de intrare (notație: $X = (X_1, \dots, X_d)$) au distribuțiile [comune] condiționale $X|Y = 0$ și $X|Y = 1$ de tip gaussian [multidimensional], cu matricele de covarianță egale ($\Sigma_0 = \Sigma_1$), atunci regula de decizie a algoritmului "full" / Joint Gaussian Bayes este și ea echivalentă (ca formă) cu cea a regresiei logistice, deci separarea realizată este tot de formă liniară: ex. 18, ex. 20.a.i – ii;
- (P2') când variabilele de intrare satisfac condiții mixte de tip (P0') sau (P7), atunci concluzia – separare liniară – se menține: ex. 49.b;
- (P3') dacă în condițiile de la propozițiile (P0')-(P2') presupozitia de independentă condițională este satisfăcută, iar numărul de instanțe de antrenament tinde la infinit, atunci rezultatul de clasificare obținut de către algoritmul Bayes Naiv gaussian este identic cu cel al regresiei logistice: ex. 22.a.

Atunci când presupozitia de independentă condițională nu este satisfăcută, iar numărul de instanțe de antrenament tinde la infinit, regresia logistică se comportă mai bine decât algoritmul Bayes Naiv [gaussian]: ex. 22.b;

- (P4') nu există o corespondență 1-la-1 între parametrii calculați de regresia logistică și între parametrii calculați de algoritmul Bayes Naiv [gaussian]: ex. 23.a;
- (P5') atunci când varianțele distribuțiilor gaussiene care corespund probabilităților condiționale $P(X_i|Y = k)$ depind și de eticheta k , separatorul decizional determinat de algoritmul Bayes Naiv gaussian nu mai are (în mod necesar) forma regresiei logistice: ex. 45.bc, ex. 48; similar, pentru algoritmul Bayes Optimal gaussian, atunci când $\Sigma_0 \neq \Sigma_1$, ecuația separatorului decizional este de ordin pătratic: ex. 19, ex. 20.a.iii – vi, ex. 47.e (separatorul decizional este un cerc), ex. 47.f (o hiperbolă), ex 52 (o reuniune de două drepte);
- (P6') parametrii algoritmilor Bayes Naiv gaussian și Bayes Optimal gaussian se pot estima în timp liniar în raport cu numărul de instanțe din setul de date de antrenament: ex. 23.b, ex. 46.bd.

2.1 Clasificare bayesiană — Probleme rezolvate

2.1.1 Ipoteze de probabilitate maximă a posteriori (MAP)

1. (Formula lui Bayes; medii ale unor variabile aleatoare discrete; [ipoteze MAP;] măsuri statistice folosite în clasificare)

■ • CMU, 2009 fall, Geoff Gordon, HW1, pr. 2

O anumită boală afectează una din 500 de persoane în medie. Identificarea persoanelor care au această boală se poate face cu ajutorul unei analize a săngelui, care costă 100 de dolari de persoană. Această analiză indică în cazul unui rezultat *pozitiv* faptul că se *poate* ca persoana respectivă să suferă de acea boală.

Testul / analiza are o *sensibilitate* (engl., *sensitivity* sau *recall*) perfectă — adică raportul dintre numărul instanțelor pozitive identificate ca atare de acel test și numărul total de instanțe pozitive este 1 —, ceea ce înseamnă că pentru orice persoană care are boala respectivă, rezultatul testului este pozitiv cu probabilitate de 100%. Pe de altă parte, testul are o *specificitate* — raportul dintre numărul instanțelor negative identificate ca atare de acel test și numărul total de instanțe negative — de 99%, adică o persoană care nu suferă de acea boală va avea cu probabilitate de 1% rezultatul testului pozitiv.

- a. Se testează o persoană selectată în mod aleatoriu, iar rezultatul este pozitiv. Care este probabilitatea ca persoana respectivă să suferă de acea boală?
- b. Există și un al doilea test, care costă 10.000 de dolari și are atât sensibilitatea cât și specificitatea de 100%. Dacă am cere ca toate persoanele detectate pozitiv la testul precedent să fie supuse acestui test mult mai scump, care ar fi costul mediu pentru testarea / analiza unui individ?
- c. O companie farmaceutică încearcă să reducă prețul celui de-al doilea test (care este perfect), adică are atât *sensibilitatea* cât și *specificitatea* de 100%. Cât ar trebui să fie prețul acesta pentru ca primul test să nu mai fie necesar? (Adică, la ce preț va rezulta că este mai ieftin să se utilizeze doar testul al doilea, decât să se facă ambele teste, ca la punctul b?)

Răspuns:

Definim următoarele variabile aleatoare:

B : ia valoarea 1 / adevărat pentru persoanele care suferă de această boală și 0 / fals în caz contrar

T_1 : rezultatul primului test, care poate fi + (în caz de boală) sau -

T_2 : rezultatul celui de-al doilea test, care poate fi tot + sau -.

Folosind aceste variabile aleatoare, datele problemei se pot scrie astfel:

$$\begin{aligned} P(B) &= \frac{1}{500} \\ P(T_1 = + | B) &= 1 \\ P(T_1 = + | \bar{B}) &= \frac{1}{100} \\ P(T_2 = + | B) &= 1 \\ P(T_2 = + | \bar{B}) &= 0 \end{aligned}$$

a. Probabilitatea ca o persoană oarecare să suferă de boala respectivă, știind că rezultatul primului test este pozitiv, este $P(B | T_1 = +)$ și se calculează cu ajutorul formulei lui Bayes:

$$\begin{aligned} P(B | T_1 = +) &= \frac{P(T_1 = + | B) \cdot P(B)}{P(T_1 = + | B) \cdot P(B) + P(T_1 = + | \bar{B}) \cdot P(\bar{B})} \\ &= \frac{1 \cdot \frac{1}{500}}{1 \cdot \frac{1}{500} + \frac{1}{100} \cdot \frac{499}{500}} = \frac{100}{599} \approx 0.1669 \end{aligned}$$

Observație: Remarcați faptul că $P(\bar{B} | T_1 = +) = 0.8331$. Este imediat că dintre cele două probabilități, $P(B | T_1 = +)$ și $P(\bar{B})$, este mai mare. În mod echivalent, pentru a stabili acest fapt era suficient să comparăm $P(B | T_1 = +)$ cu $1/2$, sau să stabilim care dintre produsele $P(T_1 = + | B) \cdot P(B)$ și $P(T_1 = + | \bar{B}) \cdot P(\bar{B})$ este mai mare. Aceste observații sunt utile pentru că ele fac legătura cu noțiunea de *ipoteză de probabilitate maximă a posteriori* (vedeți pr. 3).

b. Vom calcula costul mediu al testării unui individ folosind o nouă variabilă aleatoare, notată cu C , care reprezintă costul total de testare al unei persoane. Notând cu c_1 și c_2 costurile celor două teste, putem scrie:

$$C = \begin{cases} c_1 & \text{dacă persoana este testată doar cu primul test} \\ c_1 + c_2 & \text{dacă persoana este testată cu ambele teste} \end{cases}$$

O persoană este testată cu al doilea test doar dacă are rezultatul pozitiv la primul test, deci probabilitățile pentru variabila aleatoare C sunt:

$$P(C = c_1) = P(T_1 = -) \text{ și } P(C = c_1 + c_2) = P(T_1 = +)$$

Costul mediu cerut de problemă este media variabilei aleatoare C , deci se poate calcula astfel:

$$\begin{aligned} E[C] &= c_1 \cdot P(C = c_1) + (c_1 + c_2) \cdot P(C = c_1 + c_2) \\ &= c_1 \cdot P(T_1 = -) + (c_1 + c_2) \cdot P(T_1 = +) \end{aligned}$$

Stim că $P(T_1 = -) = 1 - P(T_1 = +)$, iar din formula probabilității totale avem:

$$\begin{aligned} P(T_1 = +) &= P(T_1 = + | B) \cdot P(B) + P(T_1 = + | \bar{B}) \cdot P(\bar{B}) \\ &= 1 \cdot \frac{1}{500} + \frac{1}{100} \cdot \frac{499}{500} = \frac{599}{50000} = 0.01198 \end{aligned}$$

Așadar, vom obține:

$$\begin{aligned}
 E[C] &= c_1 \cdot (1 - P(T_1 = +)) + (c_1 + c_2) \cdot P(T_1 = +) \\
 &= c_1 - c_1 \cdot P(T_1 = +) + c_1 \cdot P(T_1 = +) + c_2 \cdot P(T_1 = +) \\
 &= c_1 + c_2 \cdot P(T_1 = +) \\
 &= 100 + 10000 \cdot \frac{599}{50000} \\
 &= 219.8 \approx 220\$
 \end{aligned}$$

c. Notăm cu c_n noul preț pentru al doilea test (T'_2). Acest preț trebuie să fie mai mic sau egal cu costul mediu de aplicare al ambelor teste (T_1 și T'_2), deci:

$$\begin{aligned}
 c_n \leq E[C'] &= c_1 \cdot P(C = c_1) + (c_1 + c_n) \cdot P(C = c_1 + c_n) \\
 &= c_1 + c_n \cdot P(T_1 = +) = 100 + c_n \cdot \frac{599}{50000}
 \end{aligned}$$

Rezolvând ecuația $c_n = 100 + c_n \cdot 0.01198$, obținem $c_n \approx 101.2125$.

Așadar, dacă al doilea test ar costa cel mult 101.21 dolari, atunci primul test nu-ar fi necesar.

2.

(Formula lui Bayes;
[ipoteze MAP;] inferențe statistice)

■ • CMU, 2009 fall, Geoff Gordon, HW1, pr. 1
("Monty's haunted house" problem)

Fără să știi cum s-a întâmplat, ai nimerit într-o casă plină de fantome. Acum ești blocat în fața unui perete care are 3 uși (pentru conveniență, le vom nota cu numerele 1, 2, 3). Apare o fantomă care îți spune: „Scăparea ta este să ieși din casă printr-una din aceste uși. Însă doar una dintre ele dă în afară; celelalte două sunt păzite de câte un monstru care te va ucide imediat dacă încerci să ieși pe acolo. Trebuie să alegi o ușă!“

Decizi să alegi la întâmplare una din cele trei uși, să zicem ușa 1.

Observație: Probabilitatea *a priori* ca ușa aceasta să dea în afară este $1/3$. (La fel este și în cazul celorlalte două uși.) Prin adăgarea altor informații — vedeți continuarea problemei — probabilitatea *a posteriori* a aceluiasi eveniment se poate modifica, deci într-un caz fericit ea poate crește.

Într-adevăr, tocmai când ai pus mâna pe clanță ca să o deschizi, fantoma îți spune: „Așteaptă puțin! Îți voi mai da o informație.“ Zicând aceasta, fantoma intredeschide o altă ușă (să zicem ușa 2) și îți arată că în spatele ei se află un monstru groaznic. Apoi fantoma te întreabă: „Vrei acum să alegi ultima ușă (adică ușa 3) sau consideri că este mai bine să rămâi la alegerea pe care ai făcut-o inițial?“

Fantomă te mai ajută spunându-ți că în alegerea ei, ea a urmat o *strategie* bazată pe două *principii*:

P1. După ce tu ai făcut alegerea inițială, fantoma a ales una din celelalte două uși, mai precis o ușă în spatele căreia se află un monstru. (Este evident că întotdeauna există o astfel de ușă, indiferent de alegerea ta.)

P2. În cazul în care ambele uși între care are de ales fantoma au în spate câte un monstru, ea procedează după *una* din următoarele trei *variante* (pe care o alege *a priori* și îi aduce la cunoștință):

- Fantoma alege una din cele două uși cu probabilitate egală (1/2).
- Dacă, aşa cum a fost cazul mai sus, tu ai ales ușa 1, fantoma alege ușa 2 (cu probabilitate 1).
- Dacă, tot aşa, tu ai ales ușa 1, fantoma alege ușa 3 (cu probabilitate 1).
(*Notă:* Alte variante nu interesează pentru rezolvarea care se cere mai jos.)

Se cere ca, pentru fiecare din aceste 3 variante în parte, să determini probabilitățile ca ieșirea să se afle în spatele ușii 1, respectiv în spatele ușii 3, dacă fantoma a deschis ușa 2.

Indicație: Pentru rezolvare, vom folosi două *variabile aleatoare*:

- O (de la engl. *outside*), cu valori în multimea $\{1, 2, 3\}$, indicând unde este ieșirea cea bună;
- G (de la engl. *ghost*), pentru a desemna ușa aleasă de fantomă.

Pentru fiecare din variantele de strategie ale fantomei (a, b, c), folosind formula lui Bayes se calculează $P(O = 1 | G = 2)$ și $P(O = 3 | G = 2)$.

Răspuns:

Pentru variabila aleatoare O avem probabilități *a priori* egale pentru toate ieșirile:

$$P(O = 1) = P(O = 2) = P(O = 3) = \frac{1}{3}. \quad (212)$$

Folosind formula lui Bayes combinată cu formula probabilității totale, probabilitățile (*a posteriori*) cerute vor putea fi calculate astfel:

$$P(O = 1 | G = 2) = \frac{P(G = 2 | O = 1) \cdot P(O = 1)}{P(G = 2 | O = 1) \cdot P(O = 1) + P(G = 2 | O = 3) \cdot P(O = 3)}$$

$$P(O = 3 | G = 2) = \frac{P(G = 2 | O = 3) \cdot P(O = 3)}{P(G = 2 | O = 3) \cdot P(O = 3) + P(G = 2 | O = 1) \cdot P(O = 1)}.$$

Remarcăm că la fiecare din numitorii celor două fracții de mai sus ar fi trebuit să mai scriem $P(G = 2 | O = 2) \cdot P(O = 2)$, însă acesta este 0 fiindcă $P(G = 2 | O = 2) = 0$, conform principiului P1. Deci $P(O = 3 | G = 2) = 1 - P(O = 1 | G = 2)$.

Pentru a exprima succint probabilitățile condiționate necesare pentru calculul probabilităților $P(O = 1 | G = 2)$ și $P(O = 3 | G = 2)$ folosind formulele de mai sus, completăm următorul tabel:⁴⁰⁹

G	O	$P(G O)$		
		varianta a	varianta b	varianta c
2	1	1/2	1	0
3	1	1/2	0	1
2	2	0	0	0
3	2	1	1	1
2	3	1	1	1
3	3	0	0	0

⁴⁰⁹Se observă că pe fiecare dintre cele trei coloane din tabel, ultimele patru linii sunt identice, ceea ce este natural, conform principiilor P1 și P2.

În aceste condiții putem calcula ușor probabilitățile cerute în fiecare din cele trei variante:

Varianta a:

$$P(O = 1 | G = 2) = \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + 1 \cdot \frac{1}{3}} = \frac{1}{3} \text{ și } P(O = 3 | G = 2) = 1 - \frac{1}{3} = \frac{2}{3},$$

deci vei alege ușa a treia, fiindcă ea are probabilitatea mai mare de a te duce afară.

Varianta b:

$$P(O = 1 | G = 2) = \frac{1 \cdot \frac{1}{3}}{1 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3}} = \frac{1}{2} \text{ și } P(O = 3 | G = 2) = 1 - \frac{1}{2} = \frac{1}{2},$$

deci vei alege la întâmplare oricare din cele două uși rămase (ușa 1 și ușa 3), ele având aceeași probabilitate de salvare.

Varianta c:

$$P(O = 1 | G = 2) = 0 \text{ și } P(O = 3 | G = 2) = 1 - 0 = 1,$$

deci vei alege ușa a treia, care este cu siguranță ieșirea cea bună.

Observații:

1. Echivalent, pentru a determina maximul dintre $P(O = 1 | G = 2)$ și $P(O = 3 | G = 2)$ ar fi fost suficient, conform formulei lui Bayes, să comparăm $P(G = 2 | O = 1) \cdot P(O = 1)$ și $P(G = 2 | O = 3) \cdot P(O = 3)$. Mai mult, ținând cont de relația (212), aceasta revine la a compara $P(G = 2 | O = 1)$ și $P(G = 2 | O = 3)$. Răspunsul poate fi citit imediat din tabelul de mai sus (vedeți prima linie și penultima linie): în cazul variantei a, $O = 3$ este varianta pentru care se obține probabilitatea [a posteriori] maximă. Altfel spus, pentru varianta a, $O = 3$ este *ipoteza de probabilitate maximă a posteriori* (engl., maximum a posteriori probability (MAP) hypothesis). Absolut similar se poate proceda și pentru variantele b și c.⁴¹⁰

2. Alternativ, pentru variantele b și c putem răspunde la întrebare și făcând un raționament care nu folosește formula lui Bayes. Ieșirea cea bună se poate găsi în spatele uneia dintre cele trei uși (vedeți figura alăturată). Cum fantoma a deschis deja ușa cu numărul 2, una dintre aceste situații (și anume, a doua din figură) este eliminată, fiindcă în spatele ei este un monstru. În continuare, putem raționa în felul următor:

1	2	3
	M	M
M		M
M	M	

Varianta b: Întrucât fantoma alege ușa 2 cu probabilitate 1, vom putea afirma că ambele variante – 1 și 3 – au probabilități egale, și anume $\frac{1}{2}$. Într-adevăr,

⁴¹⁰Observație: În cazuri precum cel de mai sus ($P(O = 1) = P(O = 2) = P(O = 3) = 1/3$), ipoteza MAP coincide cu *ipoteza de verosimilitate maximă* (engl., maximum likelihood (ML) hypothesis).

- fie ușa 1, cea aleasă de mine, dă înspre afară, iar atunci fantoma trebuie, conform principiului P2, să aleagă ușa 2;
- fie ușa 3 dă înspre afară, iar atunci, din nou conform principiului P2, fantoma trebuie să aleagă ușa 2;
- conform principiului P1, nu există o a treia posibilitate;
- nu dispun de alte informații pentru a decide între cele două situații de mai sus.

Varianta c: Știind că fantoma nu a deschis ușa 3 (care ar fi opțiunea corespunzătoare principiului P2), ci a ales ușa 2, înseamnă că nu a putut face altfel, deci ușa 3 reprezintă ieșirea.

3. (Formula lui Bayes; inferențe statistice; exemplificarea noțiunii de ipoteză / ipoteze MAP (“Maximum A posteriori Probability”))

■ • ○ CMU, 2012 spring, Ziv Bar-Joseph, HW1, pr. 1.5

Mickey dă cu zarul de mai multe ori, sperând să obțină un 6. Secvența celor 10 rezultate obținute de el în urma acestor aruncări este următoarea: 1, 3, 4, 2, 3, 3, 2, 5, 1, 6. Mickey se întreabă dacă nu cumva zarul este măsluit (având tendința să producă de mai multe ori față 3 decât ar fi normal dacă zarul ar fi perfect).

Concepți o analiză simplă bazată pe teorema lui Bayes care să-i furnizeze lui Mickey informația care-l interesează: [în ce măsură putem spune că] zarul este măsluit [sau nu]?

Explicați raționamentul dumneavoastră.

Veți presupune că în general fiecare set de 100 de zaruri conține 5 zaruri măsluite (engl., unfair) în aşa fel încât este favorizată apariția feței 3, rezultând următoarea distribuție de probabilitate a celor șase fețe, (1, 2, 3, 4, 5, 6): $P = [0.1, 0.1, 0.5, 0.1, 0.1, 0.1]$.

Răspuns:

Acesta este un exercițiu [tipic] de punere în evidență a noțiunii de ipoteză de probabilitate maximă a posteriori (engl., Maximum A posteriori Probability, MAP).

Vă reamintim definiția:

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}} P(h|D),$$

unde D este setul de date cu care se lucrează, iar H este mulțimea de ipoteze considerate.

În cazul nostru, $D = \{1, 3, 4, 2, 3, 3, 2, 5, 1, 6\}$, iar $H = \{FD, LD\}$, unde am notat cu FD zarul corect / cinstit (engl., fair die) și cu LD zarul măsluit (engl., loaded die).

Folosind formula lui Bayes, definiția de mai sus, poate fi „rafinată“ astfel:

$$h_{MAP} \stackrel{\text{def.}}{=} \underset{h \in H}{\operatorname{argmax}} P(h|D) \stackrel{F.B.}{=} \underset{h \in H}{\operatorname{argmax}} \frac{P(D|h) \cdot P(h)}{P(D)} = \underset{h \in H}{\operatorname{argmax}} P(D|h) \cdot P(h),$$

ultima egalitate având loc datorită faptului că $P(D)$ este o cantitate pozitivă care nu depinde de h .⁴¹¹

Așadar, în cazul de față, a determina ipoteza de probabilitate maximă a posteriori (h_{MAP}) revine la a determina maximul dintre două produse: $P(D|FD) \cdot P(FD)$ și $P(D|LD) \cdot P(LD)$.

Facem observația că pentru a calcula $P(D|FD)$ și $P(D|LD)$, vom ține cont de faptul că aruncările zarului au fost independente unele de altele. Prin urmare, notând $D = \{x_1, x_2, \dots, x_{10}\}$, vom putea scrie:

$$\begin{aligned} P(D|FD) \cdot P(FD) &= P(x_1, x_2, \dots, x_{10}|FD) \cdot P(FD) \stackrel{i.i.d.}{=} \left(\prod_{i=1}^{10} P(x_i|FD) \right) \cdot P(FD) \\ &= \left(\frac{1}{6} \right)^{10} \cdot \frac{95}{100} = \frac{1}{2^{10} \cdot 3^{10}} \cdot \frac{19}{20}. \end{aligned}$$

Similar,

$$\begin{aligned} P(D|LD) \cdot P(LD) &= P(x_1, x_2, \dots, x_{10}|LD) \cdot P(LD) \stackrel{i.i.d.}{=} \left(\prod_{i=1}^{10} P(x_i|LD) \right) \cdot P(LD) \\ &= \left(\frac{1}{10} \cdot \frac{1}{2} \cdot \frac{1}{10} \cdot \frac{1}{10} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{10} \cdot \frac{1}{10} \cdot \frac{1}{10} \cdot \frac{1}{10} \right) \cdot \frac{5}{100} \\ &= \frac{1}{10^7 \cdot 2^3} \cdot \frac{1}{20} = \frac{1}{2^{10} \cdot 5^7} \cdot \frac{1}{20}. \end{aligned}$$

Așadar, a vedea care dintre produsele $P(D|FD) \cdot P(FD)$ și $P(D|LD) \cdot P(LD)$ este mai mare revine la a compara fracțiile $\frac{19}{3^{10}}$ și $\frac{1}{5^7}$. În loc să facem ridicările la putere (3^{10} și 5^7), este mai convenabil să logaritmăm, folosind ca bază un număr supra-unitar:

$$\ln \frac{19}{3^{10}} = \ln 19 - \ln 3^{10} = \ln 19 - 10 \ln 3 = 2.9444 - 10.9861 = -8.0417$$

și

$$\ln \frac{1}{5^7} = -\ln 5^7 = -7 \ln 5 = -11.2661.$$

Conchidem că ipoteza de probabilitate maximă a posteriori este FD , deci că zarul lui Mickey *nu* este măsluit.

Observație: Se obișnuieste ca, în loc să se lucreze cu cele două produse ($P(D|FD) \cdot P(FD)$ și $P(D|LD) \cdot P(LD)$) în mod separat, aşa cum am procedat noi mai sus, să se facă raportul lor,

$$\frac{P(D|LD) \cdot P(LD)}{P(D|FD) \cdot P(FD)} = \frac{P(LD|D)}{P(FD|D)}.$$

Acest raport se numește *raportul de şanse* (engl., odds ratio). (Dacă acest raport este supra-unitar, înseamnă că ipoteza LD este mai plauzibilă decât ipoteza FD .) Mai departe, aplicând logaritmul — fiindcă la calcule probabilitatea $P(D|\dots)$ se exprimă ca produs de n factori —, se obține ceea ce în limba

⁴¹¹Folosind formula probabilității totale, atunci când H este o mulțime discretă, putem exprima $P(D)$ ca $\sum_{h' \in H} P(D|h') \cdot P(h')$. Ar fi util să calculați această probabilitate *a priori* în cazul datelor din acest exercițiu.

engleză se numește *log-odds ratio*. (Evident, dacă *log-odds ratio* are valoare pozitivă, ipoteza *LD* este mai plauzibilă.) Pe datele noastre,

$$\ln \frac{P(LD|D)}{P(FD|D)} = -11.2661 - (-8.0417) = -3.2244 < 0.$$

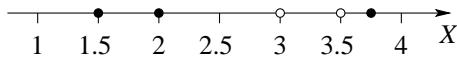
În concluzie, folosind un astfel de raport (de „potrivire“), vom putea spune nu doar care ipoteză este mai plauzibilă, ci și *în ce măsură* acea ipoteză (în cazul nostru, *FD*) este mai plauzibilă decât cealaltă ipoteză (*LD*).

4.

(Arbori ID3 cu decizii probabiliste,
ca ipoteze ML și respectiv MAP)

■ CMU, 2009 spring, T. Mitchell, midterm, pr. 2

Se consideră următorul set de date de antrenament din spațiul real unidimensional:



Este vorba de 5 date caracterizate de atributul real X , împărțite în două clase: clasa 0 constituuită din mulțimea $\{3, 3.75\}$, și clasa 1 – mulțimea $\{1.5, 2\}$.

Pe acest set de date se va aplica algoritmul ID3 pentru construirea unor arbori de decizie. Deoarece atributul are valori reale, testelete vor fi de forma $X > t$, unde t reprezintă o valoare-prag.

Se notează cu DT^* algoritmul care construiește arborele de decizie cu număr minim de noduri necesare pentru clasificarea perfectă a datelor de antrenament, și cu $DT1$ algoritmul care construiește un arbore de decizie cu un singur nod de test.

Indicație: Veți presupune că atunci când algoritmul de învățare de arbori de decizie găsește două praguri astfel încât testelete $x < t_1$ și $x < t_2$ produc același câștig de informație, el (adică, algoritmul) va alege pragul cel mai din stânga. (Așadar, dacă $t_1 < t_2$, atunci el va alege testul ($x < t_1$)).

- Care este eroarea la antrenare a lui $DT1$ pe datele specificate? Dar eroarea la cross-validation folosind metoda “Leave-One-Out” (CVLOO)?
- Care este eroarea la antrenare a lui DT^* pe datele specificate? Dar eroarea la cross-validation folosind metoda “Leave-One-Out”?

În continuare se consideră o nouă clasă de arbori de decizie, care au *etichete probabiliste*. Fiecare nod frunză specifică probabilitatea fiecărei etichete posibile, probabilitate scrisă sub forma raportului dintre datele cu acea etichetă din nodul respectiv și toate datele din acel nod.

De exemplu, un arbore de decizie neavând niciun nod de test, construit pe datele specificate mai sus, clasifică astfel: $P(Y = 1) = 3/5$ și $P(Y = 0) = 2/5$. Un arbore de decizie cu un singur nod de test (în raport cu valoarea / pragul 2.5) conține probabilitățile: $P(Y = 1) = 1$ dacă $X \leq 2.5$, și $P(Y = 1) = 1/3$ dacă $X > 2.5$.

c. Pentru setul de date de mai sus, determinați arborele de decizie de tip ML (engl., maximum likelihood), adică acel arbore cu decizii probabiliste care maximizează *verosimilitatea* datelor de antrenament:

$$T_{ML} = \operatorname{argmax}_T P_T(D), \text{ unde}$$

$$P_T(D) \stackrel{\text{def.}}{=} P(D|T) \stackrel{\text{indep.}}{=} \prod_{i=1}^5 P(Y = y_i | X = x_i, T),$$

cu y_i eticheta / clasa instanței $x_i \in \{1.5, 2, 3, 3.5, 3.75\}$.

d. Se consideră o distribuție a priori $P(T)$ care penalizează numărul de teste / split-uri din arborele de decizie T , și anume:

$$P(T) \propto \left(\frac{1}{4}\right)^{\text{splits}(T)^2}$$

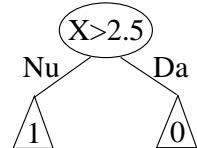
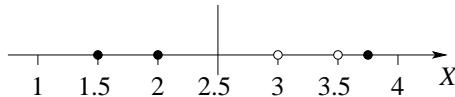
unde $\text{splits}(T)$ reprezintă numărul nodurilor de test din arborele T , iar simbolul \propto înseamnă „este proporțional cu”.

Pentru același set de date, folosind această distribuție a priori $P(T)$, găsiți arborele de decizie de tip MAP (engl., Maximum A posteriori Probability):

$$T_{MAP} = \operatorname{argmax}_T P_T(T|D)$$

Răspuns:

a. Dacă se aplică algoritmul DT1, întrucât câștigurile de informație calculate pentru testele $X > 2.5$ și $X > 3.625$ sunt 0.419 și respectiv 0.171, rezultatul este:



Eroarea la antrenare este $1/5$, deoarece punctul 3.75 este clasificat greșit.

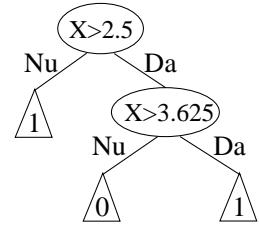
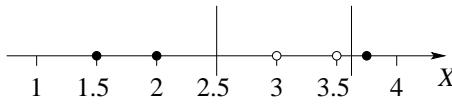
Eroarea la cross-validation cu metoda “Leave-One-Out” se determină astfel:

- $x_1 = 1.5$. Testul se face relativ la pragul 2.5 — ținând cont de *Indicația* din enunț —, deci punctul 1.5 este clasificat corect.
- $x_2 = 2$. Testul se face relativ la pragul 2.25 — ținând din nou cont de *Indicația* din enunț —, deci punctul $x_2 = 2$ este clasificat corect.
- $x_3 = 3$. Testul se face relativ la pragul 2.75, fiindcă se verifică imediat că IG-ul acestui prag este 0.311, deci mai mare decât IG-ul pragului 3.625, și anume 0.122. Prin urmare, punctul $x_3 = 3$ va fi clasificat fie corect (în cazul în care se consideră că decizia arborelui DT1 este 0 pentru $X > 2.75$), fie eronat (în cazul în care se consideră că decizia arborelui DT1 este 1 pentru $X > 2.75$). Însă în al doilea caz arborele DT1 s-ar reduce la un singur nod (care este nod frunză), ceea ce este contrar definiției sale, aşa că vom reține doar primul caz.

- $x_4 = 3.5$. Testul se face din nou relativ la pragul 2.5 — justificarea este similară cu cea de la punctul precedent —, iar punctul $x_4 = 3.5$ este clasificat fie corect (în cazul în care se consideră că decizia arborelui DT_1 este 0 pentru $X > 2.5$), fie eronat (în cazul în care se consideră că decizia arborelui DT_1 este 1 pentru $X > 2.5$). Însă (din nou!) în al doilea caz arborele DT_1 s-ar reduce la un singur nod (frunză), aşa că vom reține doar primul caz.
- $x_5 = 3.75$. Testul se face tot relativ la pragul 2.5, iar punctul $x_5 = 3.75$ este clasificat greșit.

Deci eroarea la cross-validation cu metoda “Leave-One-Out” este de $1/5$, punctul 3.75 fiind clasificat eronat.

- b. Dacă se aplică algoritmul DT^* , rezultatul este:



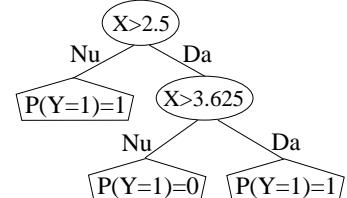
Eroarea la antrenare este bineînteleasă 0, deoarece datele de antrenament nu conțin inconistențe.

La cross-validation cu metoda “Leave-One-Out”, punctele care ar putea genera probleme sunt:

- $x_2 = 2$. Cele două teste se fac la pragurile 2.25 și la 3.625 (ordinea nu contează), deci punctul $x_2 = 2$ este corect clasificat.
- $x_3 = 3$. Primul test se face la pragul 2.75, deci punctul $x_3 = 3$ este corect clasificat.
- $x_4 = 3.5$. Al doilea test se face la pragul 3.375, deci punctul $x_4 = 3.5$ este clasificat greșit.
- $x_5 = 3.75$. Se face un singur test (la pragul 2.5), iar punctul $x_5 = 3.75$ este clasificat greșit.

Deci eroarea la cross-validation pentru arborele DT^* folosind metoda “Leave-One-Out” este $2/5$.

- c. Un arbore de decizie care maximizează probabilitățile datelor de antrenament va fi unul care clasifică în mod perfect aceste date. Deci un arbore de decizie ML poate fi obținut din arborele construit de DT^* , extins cu etichete probabiliste în frunze, conform figurii alăturate.



- d. Pentru a determina arborele de decizie MAP folosind distribuția a priori $P(T)$, va trebui să comparăm probabilitățile a posteriori ale celor 3 arbori de decizie posibili pe setul de date de antrenament. Vom scrie aceste probabilități a posteriori folosind formula lui Bayes:

$$P(T_j | D) = \frac{P(D | T_j) \cdot P(T_j)}{P(D)} = \frac{\prod_{i=1}^5 P(Y = y_i | T_j, X = x_i) \cdot P(T_j)}{P(D)}$$

Evident, la compararea propriu-zisă a probabilităților $P(T_j | D)$ cu $j = 0, 1, 2$, nu vom avea nevoie de numitorul $P(D)$. Așadar,

- pentru 0 noduri de test:

După cum s-a precizat și în enunț, acest arbore va avea $P(Y = 1) = 3/5$ și $P(Y = 0) = 2/5$. Deci

$$P(T_0 | D) \propto \left(\frac{3}{5}\right)^3 \cdot \left(\frac{2}{5}\right)^2 \cdot \left(\frac{1}{4}\right)^0 = \frac{3^3 \cdot 2^2}{5^5} = \frac{108}{3125} = 0.0336.$$

- pentru un nod de test:

Acest arbore va avea probabilitățile: $P(Y = 1) = 1$ dacă $X < 2.5$, și $P(Y = 1) = 1/3$ dacă $X \geq 2.5$. Prin urmare,

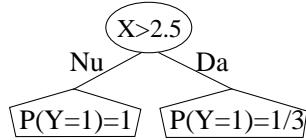
$$P(T_1 | D) \propto 1^2 \cdot \left(\frac{2}{3}\right)^2 \cdot \frac{1}{3} \cdot \left(\frac{1}{4}\right)^1 = \frac{1}{27} = 0.037$$

- pentru două noduri de test:

Este vorba de arborele construit la punctul c. Acesta clasifică perfect datele, dar

$$P(T_2) \propto \left(\frac{1}{4}\right)^4 \Rightarrow P(T_2 | D) \propto 1 \cdot \left(\frac{1}{4}\right)^4 = \frac{1}{256} = 0.0039$$

Probabilitatea a posteriori maximă o are arborele T_1 , deci acesta va fi arborele MAP. Reprezentarea grafică a acestui arbore este:



2.1.2 Algoritmii Bayes Naiv și Bayes Optimal

5.

(Algoritmul Bayes Naiv: aplicare; comparație cu estimarea MLE)

prelucrare de Liviu Ciortuz, după CMU, 2004 fall, T. Mitchell Z. Bar-Joseph, final exam, pr. 2

Fie setul de date alăturat, cu trei variabile booleene de intrare a, b, c și o variabilă booleană de ieșire K .

a. Estimați probabilitățile $P(K = 1 | a = 1, b = 1)$ și $P(K = 1 | a = 1, b = 1, c = 0)$ în sensul verosimilității maxime (engl., Maximum Likelihood Estimation, MLE).

b. Cum clasifică algoritmul Bayes Naiv instanța ($a = 1, b = 1$)?

Dar instanța ($a = 1, b = 1, c = 0$)?

a	b	c	K
1	0	1	1
1	1	1	1
0	1	1	0
1	1	0	0
1	0	1	0
0	0	0	1
0	0	0	1
0	0	1	0

Răspuns:

a. $P(K = 1 | a = 1, b = 1) = \frac{1}{2}$, fiindcă avem două instanțe în care ambele atrbute a, b sunt adevărate, dintre care una este clasificată $K = 1$, iar cealaltă $K = 0$.

$P(K = 1 | a = 1, b = 1, c = 0) = 0$, fiindcă există o singură instanță cu $a = 1, b = 1, c = 0$ în setul de antrenament și ea este clasificată $K = 0$.

b. Cazul ($a = 1, b = 1$):

$$\begin{aligned}\hat{k}_{MAP} &= \operatorname{argmax}_{k \in \{0,1\}} P(K = k | a = 1, b = 1) = \\ &= \operatorname{argmax}_{k \in \{0,1\}} \frac{P(a = 1, b = 1 | K = k) \cdot P(K = k)}{P(a = 1, b = 1)} = \\ &= \operatorname{argmax}_{k \in \{0,1\}} P(a = 1, b = 1 | K = k) \cdot P(K = k) = \\ &= \operatorname{argmax}_{k \in \{0,1\}} P(a = 1 | K = k) \cdot P(b = 1 | K = k) \cdot P(K = k)\end{aligned}$$

Avem:

$$\begin{aligned}p_0 &= P(a = 1 | K = 0) \cdot P(b = 1 | K = 0) \cdot P(K = 0) = \frac{2}{4} \cdot \frac{2}{4} \cdot \frac{4}{8} = \frac{1}{8} \\ p_1 &= P(a = 1 | K = 1) \cdot P(b = 1 | K = 1) \cdot P(K = 1) = \frac{2}{4} \cdot \frac{1}{4} \cdot \frac{4}{8} = \frac{1}{16}\end{aligned}$$

Prin urmare, $p_0 > p_1$. Așadar, clasificatorul Bayes Naiv va prezice $K = 0$ pentru instanța $(a = 1, b = 1)$ cu probabilitatea

$$\begin{aligned}P(K = 0 | a = 1, b = 1) &= \frac{P(a = 1, b = 1 | K = 0) \cdot P(K = 0)}{P(a = 1, b = 1 | K = 0) \cdot P(K = 0) + P(a = 1, b = 1 | K = 1) \cdot P(K = 1)} \\ &= \frac{p_0}{p_0 + p_1} = \frac{\frac{1}{8}}{\frac{1}{8} + \frac{1}{16}} = \frac{2}{3}\end{aligned}$$

Observație: Se constată că probabilitatea $P(K = 1 | a = 1, b = 1)$ estimată în sensul MLE (adică, $1/2$; vedeti punctul a) diferă de probabilitatea (*a posteriori*) cu care algoritmul Bayes Naiv clasifică instanța $(a = 1, b = 1)$ ca aparținând clasei $K = 1$ (adică, $1/3$). Explicația rezidă în faptul că presupozitia de independentă condițională asumată de către algoritmul Bayes Naiv nu este validă. Într-adevăr, $P(a = 0 | K = 1) = 2/4$ și $P(a = 0 | b = 1, K = 1) = 0$, deci $P(a = 0 | K = 1) \neq P(a = 0 | b = 1, K = 1)$.

Cazul ($a = 1, b = 1, c = 0$):

$$\begin{aligned}\hat{k}_{MAP} &= \operatorname{argmax}_{k \in \{0,1\}} P(K = k | a = 1, b = 1, c = 0) = \\ &= \operatorname{argmax}_{k \in \{0,1\}} \frac{P(a = 1, b = 1, c = 0 | K = k) \cdot P(K = k)}{P(a = 1, b = 1, c = 0)} = \\ &= \operatorname{argmax}_{k \in \{0,1\}} P(a = 1, b = 1, c = 0 | K = k) \cdot P(K = k) = \\ &= \operatorname{argmax}_{k \in \{0,1\}} P(a = 1 | K = k) \cdot P(b = 1 | K = k) \cdot P(c = 0 | K = k) \cdot P(K = k)\end{aligned}$$

Avem:

$$\begin{aligned}
 p_0 &= P(a = 1|K = 0) \cdot P(b = 1|K = 0) \cdot P(c = 0|K = 0) \cdot P(K = 0) = \\
 &= \frac{2}{4} \cdot \frac{2}{4} \cdot \frac{1}{4} \cdot \frac{4}{8} = \frac{1}{32} \\
 p_1 &= P(a = 1|K = 1) \cdot P(b = 1|K = 1) \cdot P(c = 0|K = 1) \cdot P(K = 1) = \\
 &= \frac{2}{4} \cdot \frac{1}{4} \cdot \frac{2}{4} \cdot \frac{4}{8} = \frac{1}{32}
 \end{aligned}$$

Întrucât $p_0 = p_1$, clasificatorul Bayes Naiv va prezice $K = 0$ sau $K = 1$ cu aceeași probabilitate ($1/2$).

Notă: Observația de mai sus este valabilă și pentru cazul ($a = 1, b = 1, c = 0$).

6. (Calculul parametrilor pentru clasificatorul Bayes Naiv pornind de la distribuția comună a variabilelor; comparație între algoritmii Bayes Naiv și Bayes Optimal)

*prelucrare de L. Ciortuz, după
■ • CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh, midterm, pr. 2.1*

Fie P o distribuție de *probabilitate comună* peste variabilele aleatoare booleene x_1, x_2 și y .

- a. Exprimăți $P(y = 0 | x_1, x_2)$ în funcție de $P(x_1, x_2, y = 0)$ și $P(x_1, x_2, y = 1)$.

x_1	x_2	y	$P(x_1, x_2, y)$
0	0	0	0.15
0	0	1	0.25
0	1	0	0.05
0	1	1	0.08
1	0	0	0.10
1	0	1	0.02
1	1	0	0.20
1	1	1	0.15

În cele ce urmează vom considera distribuția P , definită conform tabelului alăturat.

- b. Pornind de la această distribuție comună, calculați probabilitățile necesare pentru clasificare bayesiană naivă. Completăți tabelele de mai jos:

y	$P(y)$	$P(x_1 y)$	$x_1 = 0$	$x_1 = 1$	$P(x_2 y)$	$x_2 = 0$	$x_2 = 1$
$y = 0$		$y = 0$			$y = 0$		
$y = 1$		$y = 1$			$y = 1$		

- c. Cât este probabilitatea $P(y = 1 | x_1 = 1, x_2 = 0)$ calculată de către clasificatorul Bayes Naiv?

- d. Cât este probabilitatea $P(y = 1 | x_1 = 1, x_2 = 0)$ calculată de către clasificatorul Bayes Optimal? (Vă readucem aminte că acest algoritm este similar cu algoritmul Bayes Naiv, însă nu folosește presupozitia de independentă condițională a atributelor.)

- e. Răspunsurile la precedentele două întrebări ar trebui să fie diferite. Care este motivul? Justificați.

Răspuns:

a. Aplicând definiția probabilității condiționate și apoi proprietatea de aditivitate numărabilă din definiția funcției de probabilitate, obținem:

$$P(y = 0 | x_1, x_2) = \frac{P(x_1, x_2, y = 0)}{P(x_1, x_2)} = \frac{P(x_1, x_2, y = 0)}{P(x_1, x_2, y = 0) + P(x_1, x_2, y = 1)}$$

b. Probabilitățile cerute în enunț sunt $P(x_1 | y)$, $P(x_2 | y)$ și $P(y)$.

$P(y)$ este o *probabilitate marginală* a distribuției comune, deci se calculează astfel:

$$\begin{aligned} P(y = 0) &= P(x_1 = 0, x_2 = 0, y = 0) + P(x_1 = 0, x_2 = 1, y = 0) + \\ &\quad + P(x_1 = 1, x_2 = 0, y = 0) + P(x_1 = 1, x_2 = 1, y = 0) \\ &= 0.15 + 0.05 + 0.1 + 0.2 = 0.5 \\ P(y = 1) &= 1 - P(y = 0) = 0.5 \end{aligned}$$

$P(x_1 | y)$ se calculează folosind din nou definiția probabilității condiționate și formula probabilității totale:

$$P(x_1 = 0 | y = 0) = \frac{P(x_1 = 0, y = 0)}{P(y = 0)} = \frac{P(x_1 = 0, y = 0)}{P(x_1 = 0, y = 0) + P(x_1 = 1, y = 0)}$$

Probabilitățile implicate în formulă sunt probabilități marginale ale distribuției comune și se calculează astfel:

$$P(x_1 = 0, y = 0) = P(x_1 = 0, x_2 = 0, y = 0) + P(x_1 = 0, x_2 = 1, y = 0) = 0.15 + 0.05 = 0.2$$

$$P(x_1 = 1, y = 0) = P(x_1 = 1, x_2 = 0, y = 0) + P(x_1 = 1, x_2 = 1, y = 0) = 0.1 + 0.2 = 0.3$$

Prin urmare,

$$P(x_1 = 0 | y = 0) = \frac{0.2}{0.2 + 0.3} = 0.4,$$

iar

$$P(x_1 = 1 | y = 0) = 1 - P(x_1 = 0 | y = 0) = 0.6$$

Analog se calculează și celelalte probabilități corespunzătoare lui $P(x_1 | y)$:

$$\begin{aligned} P(x_1 = 0 | y = 1) &= \frac{P(x_1 = 0, y = 1)}{P(y = 1)} = \frac{0.25 + 0.08}{0.5} = 0.66 \\ P(x_1 = 1 | y = 1) &= 1 - P(x_1 = 0 | y = 1) = 0.34 \end{aligned}$$

$P(x_2 | y)$ se calculează în același mod.

Putem completa tabelele următoare cu valorile numerice ale probabilităților calculate la acest punct:

y	$P(y)$	$P(x_1 y)$	$x_1 = 0$	$x_1 = 1$	$P(x_2 y)$	$x_2 = 0$	$x_2 = 1$
$y = 0$	0.5	$y = 0$	0.40	0.60	$y = 0$	0.50	0.50
$y = 1$	0.5	$y = 1$	0.66	0.34	$y = 1$	0.54	0.46

c. Clasificatorul Bayes Naiv face presupunerea de *independență condițională* a variabilelor, deci:

$$\begin{aligned}
 P(y = 1 | x_1 = 1, x_2 = 0) &\stackrel{Bayes}{=} \frac{P(x_1 = 1, x_2 = 0 | y = 1) \cdot P(y = 1)}{P(x_1 = 1, x_2 = 0)} \\
 &= \frac{P(x_1 = 1, x_2 = 0 | y = 1) \cdot P(y = 1)}{P(x_1 = 1, x_2 = 0 | y = 1)P(y = 1) + P(x_1 = 1, x_2 = 0 | y = 0)P(y = 0)} \stackrel{\text{indep. cdt.}}{=} \\
 &= \frac{P(x_1 = 1 | y = 1) \cdot P(x_2 = 0 | y = 1) \cdot P(y = 1)}{P(x_1 = 1 | y = 1)P(x_2 = 0 | y = 1)P(y = 1) + P(x_1 = 1 | y = 0)P(x_2 = 0 | y = 0)P(y = 0)} \\
 &= \frac{0.34 \cdot 0.54 \cdot 0.5}{0.34 \cdot 0.54 \cdot 0.5 + 0.6 \cdot 0.5 \cdot 0.5} \approx 0.3796
 \end{aligned}$$

d. Clasificatorul Bayes Optimal nu face niciun fel de presupunere, deci folosind o formulă similară cu cea obținută la punctul a, vom avea:

$$\begin{aligned}
 P(y = 1 | x_1 = 1, x_2 = 0) &= \frac{P(x_1 = 1, x_2 = 0, y = 1)}{P(x_1 = 1, x_2 = 0, y = 1) + P(x_1 = 1, x_2 = 0, y = 0)} \\
 &= \frac{0.02}{0.02 + 0.1} = 0.1(6).
 \end{aligned}$$

e. Valorile calculate de clasificatorul Bayes Naiv și de clasificatorul Bayes Optimal pentru $P(y = 1 | x_1 = 1, x_2 = 0)$ sunt diferite deoarece presupunerea de independență condițională făcută de clasificatorul Bayes Naiv nu este adevărată. Într-adevăr, se observă că variabilele x_1 și x_2 nu sunt independente condițional în raport cu variabila y :

$$\begin{aligned}
 P(x_1 = 1, x_2 = 0 | y = 1) &= \frac{P(x_1 = 1, x_2 = 0, y = 1)}{P(y = 1)} = \frac{0.02}{0.5} = 0.04 \\
 P(x_1 = 1 | y = 1) \cdot P(x_2 = 0 | y = 1) &= 0.34 \cdot 0.54 = 0.1836 \neq 0.04
 \end{aligned}$$

7. (Algoritmul Bayes Naiv și algoritmul Bayes Optimal; comparație relativ la numărului de parametri)

■ • CMU, 2008 fall, Eric Xing, HW1, pr. 2

Fie următoarea problemă de clasificare: X_1 și X_2 sunt variabile aleatoare observabile, Y este eticheta clasei asignate fiecărei instanțe observate, conform tabelului alăturat.

În acest exercițiu veți compara rezultatele care se obțin în urma antrenării pe acest set de date de către doi algoritmi de clasificare: Bayes Naiv și Bayes Comun (engl., Joint Bayes), care este numit adeseori și *Bayes Optimal* (engl., Optimal Bayes).

X_1	X_2	Y	Nr. apariții
0	0	0	2
0	0	1	18
1	0	0	4
1	0	1	1
0	1	0	4
0	1	1	1
1	1	0	2
1	1	1	18

Comentariu: Pentru cel de-al doilea algoritm, denumirea *Bayes Comun* se datorează faptului că acest clasificator lucrează cu distribuția comună a variabilelor / atributelor de intrare, în vreme ce denumirea *Bayes Optimal* este justificată de faptul că nicio altă distribuție probabilistă asupra datelor de intrare nu poate conduce la o eroare medie la clasificare mai mică decât cea obținută de către acest clasificator. (În cele ce urmează, vom opta pentru cea de-a doua denumire.)

- a. Clasificați instanța $X_1 = 0, X_2 = 0$ folosind clasificatorul Bayes Naiv.
- b. Clasificați instanța $X_1 = 0, X_2 = 0$ folosind clasificatorul Bayes Optimal.
- c. Notăm cu P_{NB} și respectiv P_{JB} valoarea probabilității $P(Y = 1 | X_1 = 0, X_2 = 0)$ calculate pentru clasificatorul Bayes Naiv, respectiv pentru clasificatorul Bayes Optimal. De ce diferă cele două valori? Sugestie: Calculați $P(X_1, X_2 | Y)$.

X_1	X_2	Y	Nr. apariții
0	0	0	3
0	0	1	9
1	0	0	3
1	0	1	9
0	1	0	3
0	1	1	9
1	1	0	3
1	1	1	9

- d. Care ar fi situația pentru P_{NB} și P_{JB} de la întrebarea precedentă în situația în care datele observate ar proveni din tabelul alăturat?

- e. De câți parametri independenți (i.e., probabilități estimate) este nevoie în total pentru a construi clasificatorul Bayes Naiv? Dar în cazul clasificatorului Bayes Optimal?

Răspundeți la aceste întrebări și în cazul general, când se folosesc n variabile binare observate. Comentați rezultatul.

Răspuns:

- a. Clasificatorul Bayes Naiv face predicția pentru valoarea lui Y după formula de mai jos:

$$\hat{y}_{NB} = \operatorname{argmax}_{y \in \{0,1\}} P(X_1 = 0 | Y = y) \cdot P(X_2 = 0 | Y = y) \cdot P(Y = y)$$

Avem:

$$\begin{aligned} p_0 &\stackrel{\text{not.}}{=} P(X_1 = 0 | Y = 0) \cdot P(X_2 = 0 | Y = 0) \cdot P(Y = 0) \\ &\stackrel{\text{MLE}}{=} \frac{6}{12} \cdot \frac{6}{12} \cdot \frac{12}{50} = \frac{3}{50} = \frac{6}{100} \\ p_1 &\stackrel{\text{not.}}{=} P(X_1 = 0 | Y = 1) \cdot P(X_2 = 0 | Y = 1) \cdot P(Y = 1) \\ &\stackrel{\text{MLE}}{=} \frac{19}{38} \cdot \frac{19}{38} \cdot \frac{38}{50} = \frac{19}{100} \end{aligned}$$

Întrucât $p_0 < p_1$, clasificatorul Bayes Naiv va prezice $Y = 1$ pentru instanța $(X_1 = 0, X_2 = 0)$.

- b. Deoarece clasificatorul Bayes Optimal nu lucrează cu presupunerea de independentă condițională a atributelor de intrare în raport cu atributul de ieșire, el va face predicția folosind formula de mai jos:

$$\hat{y}_{JB} = \operatorname{argmax}_{y \in \{0,1\}} P(X_1 = 0, X_2 = 0 | Y = y) \cdot P(Y = y)$$

Probabilitățile din formulă sunt, ca și în cazul clasificatorului Bayes Naiv, cele estimate din distribuția datelor de antrenament.

Așadar, avem:

$$p'_0 \stackrel{not.}{=} P(X_1 = 0, X_2 = 0 \mid Y = 0) \cdot P(Y = 0) \stackrel{MLE}{=} \frac{2}{12} \cdot \frac{12}{50} = \frac{2}{50}$$

$$p'_1 \stackrel{not.}{=} P(X_1 = 0, X_2 = 0 \mid Y = 1) \cdot P(Y = 1) \stackrel{MLE}{=} \frac{18}{38} \cdot \frac{38}{50} = \frac{18}{50}$$

Fiindcă $p'_0 < p'_1$, clasificatorul Bayes Optimal va prezice tot $Y = 1$.

c. Vom calcula cele două valori pentru $P(Y = 1 \mid X_1 = 0, X_2 = 0)$:

$$\begin{aligned} P_{NB} &\stackrel{not.}{=} P(Y = 1 \mid X_1 = 0, X_2 = 0) \\ &\stackrel{F. Bayes}{=} \frac{P(X_1 = 0, X_2 = 0 \mid Y = 1) \cdot P(Y = 1)}{P(X_1 = 0, X_2 = 0 \mid Y = 1)P(Y = 1) + P(X_1 = 0, X_2 = 0 \mid Y = 0)P(Y = 0)} \\ &\stackrel{indep. cdt.}{=} \frac{P(X_1 = 0 \mid Y = 1) \cdot P(X_2 = 0 \mid Y = 1) \cdot P(Y = 1)}{P(X_1 = 0 \mid Y = 0) \cdot P(X_2 = 0 \mid Y = 0) \cdot P(Y = 0) + P(X_1 = 0 \mid Y = 1) \cdot P(X_2 = 0 \mid Y = 1) \cdot P(Y = 1)} \\ &= \frac{p_1}{p_0 + p_1} = \frac{\frac{19}{100}}{\frac{6}{100} + \frac{19}{100}} = \frac{19}{25} \end{aligned}$$

$$\begin{aligned} P_{JB} &\stackrel{not.}{=} P(Y = 1 \mid X_1 = 0, X_2 = 0) \\ &\stackrel{F. Bayes}{=} \frac{P(X_1 = 0, X_2 = 0 \mid Y = 1) \cdot P(Y = 1)}{P(X_1 = 0, X_2 = 0 \mid Y = 1)P(Y = 1) + P(X_1 = 0, X_2 = 0 \mid Y = 0)P(Y = 0)} \\ &= \frac{p'_1}{p'_0 + p'_1} = \frac{\frac{18}{50}}{\frac{2}{50} + \frac{18}{50}} = \frac{18}{20} \end{aligned}$$

Cele două valori diferă deoarece presupunerea de independență condițională a variabilelor X_1 și X_2 în raport cu Y făcută de clasificatorul Bayes Naiv este falsă. Acest lucru se poate vedea ușor din valorile estimate pentru $P(X_1 = 0, X_2 = 0 \mid Y = 0)$, $P(X_1 = 0 \mid Y = 0)$ și $P(X_2 = 0 \mid Y = 0)$:

$$\begin{aligned} P(X_1 = 0, X_2 = 0 \mid Y = 0) &\stackrel{MLE}{=} \frac{2}{12} \\ P(X_1 = 0 \mid Y = 0) \cdot P(X_2 = 0 \mid Y = 0) &\stackrel{MLE}{=} \frac{6}{12} \cdot \frac{6}{12} = \frac{1}{4} \end{aligned} \quad \Rightarrow$$

$$\Rightarrow P(X_1 = 0, X_2 = 0 \mid Y = 0) \neq P(X_1 = 0 \mid Y = 0) \cdot P(X_2 = 0 \mid Y = 0) \Rightarrow$$

$$\Rightarrow P(X_1, X_2 \mid Y) \neq P(X_1 \mid Y) \cdot P(X_2 \mid Y)$$

Așadar, variabilele X_1 și X_2 nu sunt independente condițional în raport cu variabila de ieșire Y .

d. Vom calcula cele două valori pentru $P(Y = 1 \mid X_1 = 0, X_2 = 0)$ în cazul noilor date:

$$\begin{aligned} P_{NB} &= P(Y = 1 \mid X_1 = 0, X_2 = 0) \\ &= \frac{P(X_1 = 0, X_2 = 0 \mid Y = 1) \cdot P(Y = 1)}{P(X_1 = 0, X_2 = 0 \mid Y = 1)P(Y = 1) + P(X_1 = 0, X_2 = 0 \mid Y = 0)P(Y = 0)} \\ &= \frac{P(X_1 = 0 \mid Y = 1) \cdot P(X_2 = 0 \mid Y = 1) \cdot P(Y = 1)}{P(X_1 = 0 \mid Y = 0) \cdot P(X_2 = 0 \mid Y = 0) \cdot P(Y = 0) + P(X_1 = 0 \mid Y = 1) \cdot P(X_2 = 0 \mid Y = 1) \cdot P(Y = 1)} \end{aligned}$$

$$\begin{aligned}
&= \frac{\frac{18}{36} \cdot \frac{18}{36} \cdot \frac{36}{48}}{\frac{6}{12} \cdot \frac{6}{12} \cdot \frac{12}{48} + \frac{18}{36} \cdot \frac{18}{36} \cdot \frac{36}{48}} = \frac{\frac{9}{48}}{\frac{3}{48} + \frac{9}{48}} = \frac{9}{12} = \frac{3}{4} \\
P_{JB} &= P(Y = 1 | X_1 = 0, X_2 = 0) \\
&= \frac{P(X_1 = 0, X_2 = 0 | Y = 1) \cdot P(Y = 1)}{P(X_1 = 0, X_2 = 0 | Y = 1)P(Y = 1) + P(X_1 = 0, X_2 = 0 | Y = 0)P(Y = 0)} \\
&= \frac{\frac{9}{36} \cdot \frac{36}{48}}{\frac{3}{12} \cdot \frac{9}{48} + \frac{9}{36} \cdot \frac{36}{48}} = \frac{\frac{9}{48}}{\frac{3}{48} + \frac{9}{48}} = \frac{9}{12} = \frac{3}{4}
\end{aligned}$$

Așadar, în acest caz avem $P_{NB} = P_{JB}$.

De fapt, se poate constata ușor că în cazul distribuției probabiliste date la acest punct al problemei se verifică independența condițională a variabilelor X_1 și X_2 în raport cu Y . Prin urmare, predicțiile făcute de cei doi clasificatori, Bayes Naiv și Bayes Optimal, vor coincide întotdeauna.

e. În contextul problemei noastre, clasificatorul Bayes Naiv are nevoie de estimările următoarelor probabilități:

$$\begin{aligned}
P(Y = 0) &\Rightarrow P(Y = 1) = 1 - P(Y = 0) \\
P(X_1 = 0 | Y = 0) &\Rightarrow P(X_1 = 1 | Y = 0) = 1 - P(X_1 = 0 | Y = 0) \\
P(X_1 = 0 | Y = 1) &\Rightarrow P(X_1 = 1 | Y = 1) = 1 - P(X_1 = 0 | Y = 1) \\
P(X_2 = 0 | Y = 0) &\Rightarrow P(X_2 = 1 | Y = 0) = 1 - P(X_2 = 0 | Y = 0) \\
P(X_2 = 0 | Y = 1) &\Rightarrow P(X_2 = 1 | Y = 1) = 1 - P(X_2 = 0 | Y = 1)
\end{aligned}$$

Avem nevoie, prin urmare, doar de 5 valori pentru a construi complet clasificatorul Bayes Naiv.

În cazul general, dacă avem n variabile de intrare, avem nevoie de estimări pentru probabilitățile $P(Y)$, $P(X_i | Y)$ și $P(X_i | \neg Y)$ pentru $i = \overline{1, n}$, deci $2n + 1$ valori.

Pentru clasificatorul Bayes Optimal avem nevoie de:

$$\begin{aligned}
P(Y = 0) &\quad P(Y = 1) = 1 - P(Y = 0) \\
P(X_1 = 0, X_2 = 0 | Y = 0) &\quad P(X_1 = 1, X_2 = 1 | Y = 0) \text{ se poate determina} \\
P(X_1 = 0, X_2 = 1 | Y = 0) &\quad \text{din celelalte 3 valori, aşa cum vom arăta mai} \\
P(X_1 = 1, X_2 = 0 | Y = 0) &\quad \text{jos.} \\
P(X_1 = 0, X_2 = 0 | Y = 1) &\quad \text{Similar, } P(X_1 = 1, X_2 = 1 | Y = 1) \text{ se poate} \\
P(X_1 = 0, X_2 = 1 | Y = 1) &\quad \text{determina din celelalte 3 valori.} \\
P(X_1 = 1, X_2 = 0 | Y = 1)
\end{aligned}$$

Notăm evenimentul $X_1 = 0$ cu A , $X_2 = 0$ cu B și $Y = 0$ cu C . Știm că:

$$\begin{aligned}
\Omega &= (A \wedge B) \vee (\neg A \wedge B) \vee (A \wedge \neg B) \vee (\neg A \wedge \neg B) \Rightarrow \\
\Omega \wedge C &= ((A \wedge B) \vee (\neg A \wedge B) \vee (A \wedge \neg B) \vee (\neg A \wedge \neg B)) \wedge C \Rightarrow \\
C &= ((A \wedge B) \wedge C) \vee ((\neg A \wedge B) \wedge C) \vee ((A \wedge \neg B) \wedge C) \vee ((\neg A \wedge \neg B) \wedge C)
\end{aligned}$$

De asemenea, deoarece toate evenimentele din partea dreaptă a egalității sunt disjuncte două câte două, putem scrie egalitatea de mai sus și cu probabilități:

$$\begin{aligned}
P(C) &= P((A \wedge B) \wedge C) + P((\neg A \wedge B) \wedge C) + P((A \wedge \neg B) \wedge C) + P((\neg A \wedge \neg B) \wedge C) \\
\Rightarrow 1 &= \frac{P((A \wedge B) \wedge C)}{P(C)} + \frac{P((\neg A \wedge B) \wedge C)}{P(C)} + \frac{P((A \wedge \neg B) \wedge C)}{P(C)} + \frac{P((\neg A \wedge \neg B) \wedge C)}{P(C)}
\end{aligned}$$

$$\Rightarrow 1 = P(A, B | C) + P(\neg A, B | C) + P(A, \neg B | C) + P(\neg A, \neg B | C)$$

$$\Rightarrow P(\neg A, \neg B | C) = 1 - (P(A, B | C) + P(\neg A, B | C) + P(A, \neg B | C))$$

Prin urmare, știind 3 valori o putem afla și pe a patra. La fel și pentru $\neg C$. Pentru a avea un clasificator Bayes Optimal complet avem deci nevoie de 7 valori diferite.

În cazul general, pentru n variabile de intrare, avem nevoie de probabilitățile $P(Y)$, $P(\tilde{X}_1, \dots, \tilde{X}_n | Y)$ și $P(\tilde{X}_1, \dots, \tilde{X}_n | \neg Y)$, unde

$$\tilde{X}_i \in \{X_i, \neg X_i\} \quad \forall i \in \overline{1, n} \quad \text{și} \quad (\tilde{X}_1, \dots, \tilde{X}_n) \neq (\neg X_1, \dots, \neg X_n).$$

Avem, deci, $2(2^n - 1) + 1 = 2^{n+1} - 1$ valori.

Se observă că algoritmul Bayes Naiv folosește un număr liniar de parametri (în raport cu n , numărul de atrbute de intrare), în vreme ce algoritmul Bayes Optimal folosește un număr exponential de parametri (în raport cu același n).

8.

(Algoritmul Bayes Naiv: aplicație la filtrarea emailurilor spam)

■ • CMU, 2009 spring, Ziv Bar-Joseph, midterm, pr. 2

Circa 2/3 dintre emailurile tale sunt spam, aşadar te-ai decis să descarci de pe internet un filtru spam open-source care utilizează un clasificator Bayes Naiv.

Presupunem că ai strâns următoarele emailuri spam și non-spam (engl., regular), și de asemenea că doar trei cuvinte sunt discriminative pentru această clasificare, deci fiecare email este reprezentat ca un vector de 3 componente binare, fiecare dintre ele indicând dacă respectivul cuvânt este conținut (sau nu) în email.

‘study’	‘free’	‘money’	Category	count
1	0	0	Regular	1
0	0	1	Regular	1
1	0	0	Regular	1
1	1	0	Regular	1
0	1	0	Spam	4
0	1	1	Spam	4

- a. Descrie că filtrul spam open-source folosește o probabilitate a priori $P(\text{spam}) = 0.1$. Explică în mod succint de ce crezi că această alegere este rezonabilă.
- b. Calculează următorii parametri ai modelului prin metoda estimării de verosimilitate maximă (MLE), folosind netezire (engl., smoothing) de tip “add-one” (regula lui Laplace).

$$P(\text{study|spam}) =$$

$$P(\text{study|regular}) =$$

$$P(\text{free|spam}) =$$

$$P(\text{free|regular}) =$$

$$P(\text{money|spam}) =$$

$$P(\text{money|regular}) =$$

- c. Folosind probabilitatea a priori și probabilitățile condiționate de mai sus, calculează probabilitatea ca mesajul $s = \text{"money for psychology study"}$ să fie spam, adică $P(\text{spam} | s)$.
- d. Care ar trebui să fie valoarea probabilității a priori $P(\text{spam})$ în cazul în care dorim ca mesajul de mai sus să aibă aceeași probabilitate de fi spam respectiv non-spam (i.e., el va fi clasificat ca spam cu probabilitatea 0.5)?

Răspuns:

- a. Diferența dintre $P_{MLE}(\text{Category}|\text{Spam}) = 2/3$ și probabilitatea a priori indicată în enunț (0.1) se explică prin faptul că se preferă trecerea prin filtru a unor emailuri spam, decât să fie marcate ca spam unele emailuri non-spam și astfel să nu ajungă în Inbox.
- b. Dacă nu am folosi regula de tip "add-one" a lui Laplace (pentru „netezirea“ probabilităților), parametrii modelului ar avea valorile (obținute prin metoda estimării de verosimilitate maximă – MLE) care apar mai jos în partea stângă. Folosind regula lui Laplace, parametrii primesc valorile indicate în partea dreaptă. Observați că aparițiile lui 2 de la numitorul fracțiilor corespund numărului de valori pentru fiecare dintre atributele / variabilele de intrare.

$$\begin{array}{ll} P(\text{study}|\text{spam}) = \frac{0}{8} = 0 & P(\text{study}|\text{spam}) \stackrel{\text{Laplace}}{=} \frac{0+1}{8+2} = \frac{1}{10} \\ P(\text{study}|\text{regular}) = \frac{3}{4} & P(\text{study}|\text{regular}) \stackrel{\text{Laplace}}{=} \frac{3+1}{4+2} = \frac{2}{3} \\ P(\text{free}|\text{spam}) = \frac{8}{8} = 1 & P(\text{free}|\text{spam}) \stackrel{\text{Laplace}}{=} \frac{8+1}{8+2} = \frac{9}{10} \\ P(\text{free}|\text{regular}) = \frac{1}{4} & P(\text{free}|\text{regular}) \stackrel{\text{Laplace}}{=} \frac{1+1}{4+2} = \frac{1}{3} \\ P(\text{money}|\text{spam}) = \frac{4}{8} = \frac{1}{2} & P(\text{money}|\text{spam}) \stackrel{\text{Laplace}}{=} \frac{4+1}{8+2} = \frac{1}{2} \\ P(\text{money}|\text{regular}) = \frac{1}{4} & P(\text{money}|\text{regular}) \stackrel{\text{Laplace}}{=} \frac{1+1}{4+2} = \frac{1}{3} \end{array}$$

- c. Avem mesajul $s = \text{"money for psychology study"}$, deci trebuie să calculăm $P(\text{spam} | s) = P(\text{spam} | \text{study}, \neg\text{free}, \text{money})$.

$$\begin{aligned} P(\text{spam} | s) &\stackrel{F. Bayes}{=} \\ &= \frac{P(\text{study}, \neg\text{free}, \text{money} | \text{spam}) \cdot P(\text{spam})}{P(\text{study}, \neg\text{free}, \text{money} | \text{spam})P(\text{spam}) + P(\text{study}, \neg\text{free}, \text{money} | \text{reg})P(\text{reg})} \end{aligned}$$

Calculăm probabilitățile folosind ipoteza de independentă condițională:

$$\begin{aligned} P(\text{study}, \neg\text{free}, \text{money} | \text{spam}) &= P(\text{study}|\text{spam}) \cdot P(\neg\text{free}|\text{spam}) \cdot P(\text{money}|\text{spam}) \\ &= \frac{1}{10} \cdot \frac{1}{10} \cdot \frac{1}{2} = \frac{1}{200} \end{aligned}$$

$$\begin{aligned} P(\text{study}, \neg\text{free}, \text{money} | \text{reg}) &= P(\text{study}|\text{reg}) \cdot P(\neg\text{free}|\text{reg}) \cdot P(\text{money}|\text{reg}) \\ &= \frac{2}{3} \cdot \frac{2}{3} \cdot \frac{1}{3} = \frac{4}{27} \end{aligned}$$

Înlocuind valorile în formulă, obținem: $P(\text{spam} | s) = \frac{\frac{1}{200} \cdot \frac{1}{10}}{\frac{1}{200} \cdot \frac{1}{10} + \frac{4}{27} \cdot \frac{9}{10}} \approx 0.0037$

Aceasta este o probabilitate mică. Se observă însă că dacă nu am fi folosit regula lui Laplace, probabilitatea ca emailul s să fie spam ar fi fost 0. Aceasta se datorează faptului că niciunul dintre emailurile spam din datele de antrenament nu conține cuvântul *study*, care apare însă în emailul s .

d. Dacă notăm cu p probabilitatea a priori cerută, $P(\text{spam})$, știind că $P(\text{spam}|s) = 0.5$, putem scrie:

$$\begin{aligned} 0.5 &= \frac{\frac{1}{200} \cdot p}{\frac{1}{200} \cdot p + \frac{4}{27} \cdot (1-p)} \Leftrightarrow \frac{1}{2} = \frac{\frac{p}{200}}{\frac{p}{200} + \frac{4}{27} - \frac{4p}{27}} \Leftrightarrow \frac{2p}{200} = \frac{p}{200} + \frac{4}{27} - \frac{4p}{27} \\ &\Leftrightarrow 54p = 27p + 800 - 800p \Leftrightarrow p = \frac{800}{827} \approx 0.9673 \end{aligned}$$

9.

(Clasificare bayesiană: un caz particular)

CMU, 2010 spring, E. Xing, A. Singh, T. Mitchell, midterm, pr. 2.2

Se consideră variabilele aleatoare X_1, X_2, X_3 și X_4 . Aceste variabile sunt independente condițional două câte două în raport cu variabila Y , cu excepția perechii X_3, X_4 . (Așadar, dacă am aplica algoritmul Bayes Naiv, acesta ar produce erori de clasificare.)

Cum am putea modifica *regula de decizie* a algoritmului Bayes Naiv pentru a ține cont de această particularitate a datelor?

Răspuns:

Întrucât are loc egalitatea

$$\begin{aligned} \underset{y}{\operatorname{argmax}} P(Y = y | X_1, X_2, X_3, X_4) &= \\ \underset{y}{\operatorname{argmax}} P(X_1 | Y = y) \cdot P(X_2 | Y = y) \cdot P(X_3, X_4 | Y = y) \cdot P(Y = y), \end{aligned}$$

urmează că regula de decizie a algoritmului Bayes Naiv în cazul dat este:

$$\underset{y}{\operatorname{argmax}} P_{MLE}(X_1 | Y = y) \cdot P_{MLE}(X_2 | Y = y) \cdot P_{MLE}(X_3, X_4 | Y = y) \cdot P_{MLE}(Y = y).$$

Pentru a justifica egalitatea de mai sus, se folosește regula de înlățuire condițională:

$$P(A_1, A_2, A_3 | B) = P(A_3 | B) \cdot P(A_2 | A_3, B) \cdot P(A_1 | A_2, A_3, B),$$

care se demonstrează ușor. Varianta necondițională a regulii de înlățuire este:

$$P(A_1, A_2, A_3) = P(A_3) \cdot P(A_2 | A_3) \cdot P(A_1 | A_2, A_3).$$

Apoi, regula de înlățuire condițională se particularizează pentru evenimentele $A_1 = (X_1 = x_1)$, $A_2 = (X_2 = x_2)$, $A_3 = (X_3 = x_3, X_4 = x_4)$ și $B = (Y = y)$. În fine, se va ține cont că $P(X_1 = x_1 | X_2 = x_2, X_3 = x_3, X_4 = x_4, Y = y) = P(X_1 = x_1 | Y = y)$ și $P(X_2 = x_2 | X_3 = x_3, X_4 = x_4, Y = y) = P(X_2 = x_2 | Y = y)$ datorită proprietății de independentă condițională din enunț.

Observație: Regula de decizie obținută mai sus constituie fundamentalul unui clasificator bayesian care poate fi situat din punct de vedere conceptual *între* (engl., in between) clasificatorii Bayes Naiv și Bayes Optimal, fiindcă el combină avantajele celorlalți doi clasificatori: produce o eroare la antrenare identică cu cea a lui Bayes Optimal și folosește o proprietate de independență condițională similară cu cea a lui Bayes Naiv (ceea ce conduce la un număr mai restrâns de parametri decât are Bayes Optimal). Pentru un exemplu similar, vedeti ex. 35.e.

10.

(Algoritmul Bayes Naiv:
calculul ratei medii a erorii la antrenare)*CMU, 2006 fall, T. Mitchell, E. Xing, midterm, pr. 6*

Considerăm o problemă de clasificare binară în care fiecare exemplu X are două attribute binare $X_1, X_2 \in \{0, 1\}$ și eticheta $Y \in \{0, 1\}$. Vom presupune că X_1 și X_2 sunt independente condițional în raport cu Y ,⁴¹² și că $P(Y = 0) = P(Y = 1) = 0,5$. De asemenea, probabilitățile condiționate sunt date în tabelele următoare:

$P(X_1 Y)$	$Y = 0$	$Y = 1$
$X_1 = 0$	0,7	0,2
$X_1 = 1$	0,3	0,8

$P(X_2 Y)$	$Y = 0$	$Y = 1$
$X_2 = 0$	0,9	0,5
$X_2 = 1$	0,1	0,5

a. Calculați predicția \hat{Y} făcută de clasificatorul Bayes Naiv pentru fiecare din cele patru combinații posibile de valori ale variabilelor X_1 și X_2 . Completăți următorul tabel:

X_1	X_2	$P(X_1, X_2, Y = 0)$	$P(X_1, X_2, Y = 1)$	$\hat{Y}(X_1, X_2)$
0	0	$0,7 \cdot 0,9 \cdot 0,5$	$0,2 \cdot 0,5 \cdot 0,5$	0
0	1			
1	0			
1	1			

b. Presupunând că se folosesc o infinitate de exemple, calculați *rata medie a erorii* (engl., the expected error rate) făcute de acest clasificator *la antrenare*, folosind formula:

$$P(Y \neq \hat{Y}(X_1, X_2)) = \sum_{X_1=0}^1 \sum_{X_2=0}^1 P(X_1, X_2, Y = 1) - \hat{Y}(X_1, X_2).$$

c. Care din următorii doi clasificatori are rata medie a erorii la antrenare mai mică:

- clasificatorul Bayes Naiv care prezice Y având ca input doar X_1 ;
- clasificatorul Bayes Naiv care prezice Y având ca input doar X_2 .

d. Presupunem că definim un nou atribut X_3 , care este o copie a lui X_2 . Care este rata medie a erorii la antrenare a clasificatorului Bayes Naiv care prezice Y folosind toate attributele X_1, X_2, X_3 ? (Se presupune că datele de antrenament sunt în număr infinit.)

⁴¹²Așadar, în această situație rezultatele algoritmilor Bayes Naiv și Bayes Optimal vor coincide.

- e. Explicați de ce rata erorii de la punctul d diferă față de cea de la punctul a .

Răspuns:

- a. Predicția \hat{Y} făcută de clasificatorul Bayes Naiv pentru fiecare din cele patru combinații posibile de valori ale variabilelor X_1 și X_2 este înregistrată în ultima coloană a tabelului de mai jos. Calculele necesare pentru justificare — folosind formula $P(X_1, X_2, Y) = P(X_1, X_2|Y) \cdot P(Y) = P(X_1|Y) \cdot P(X_2|Y) \cdot P(Y)$ — sunt conținute în coloanele a treia și a patra.

X_1	X_2	$P(X_1, X_2, Y = 0)$	$P(X_1, X_2, Y = 1)$	$\hat{Y}(X_1, X_2)$
0	0	$0,7 \cdot 0,9 \cdot 0,5 = 0,315$	$0,2 \cdot 0,5 \cdot 0,5 = 0,05$	0
0	1	$0,7 \cdot 0,1 \cdot 0,5 = 0,035$	$0,2 \cdot 0,5 \cdot 0,5 = 0,05$	1
1	0	$0,3 \cdot 0,9 \cdot 0,5 = 0,135$	$0,8 \cdot 0,5 \cdot 0,5 = 0,2$	1
1	1	$0,3 \cdot 0,1 \cdot 0,5 = 0,015$	$0,8 \cdot 0,5 \cdot 0,5 = 0,2$	1

Observație (1):

În acest tabel putem vedea valorile distribuției comune $P(X_1, X_2, Y)$. Spre deosebire de problema 6, unde distribuția comună era dată, iar distribuțiile marginale condiționale erau calculate pornind de la aceasta, aici se procedează invers, ținând cont [să] de presupozitia de independentă condițională.

- b. Rata medie a erorii este:

$$\begin{aligned} P(Y \neq \hat{Y}(X_1, X_2)) &= \sum_{X_1=0}^1 \sum_{X_2=0}^1 P(X_1, X_2, Y = 1 - \hat{Y}(X_1, X_2)) \\ &= P(X_1 = 0, X_2 = 0, Y = 1 - 0) + P(X_1 = 0, X_2 = 1, Y = 1 - 1) \\ &\quad + P(X_1 = 1, X_2 = 0, Y = 1 - 1) + P(X_1 = 1, X_2 = 1, Y = 1 - 1) \\ &= 0,05 + 0,035 + 0,135 + 0,015 = 0,235. \end{aligned}$$

Pentru justificarea penultimei egalități, vedetă tabelul de la punctul precedent.

- c. Făcând prezicerea doar cu X_1 ca atribut de intrare — folosind regula de înmulțire $P(X_1, Y) = P(X_1|Y) \cdot P(Y)$ —, obținem:

X_1	$P(X_1, Y = 0)$	$P(X_1, Y = 1)$	$\hat{Y}_1(X_1, X_2)$
0	$0,7 \cdot 0,5 = 0,35$	$0,2 \cdot 0,5 = 0,1$	0
1	$0,3 \cdot 0,5 = 0,15$	$0,8 \cdot 0,5 = 0,4$	1

Rata medie a erorii în acest caz va fi:

$$\begin{aligned} P(Y \neq \hat{Y}_1(X_1, X_2)) &= \sum_{X_1=0}^1 \sum_{X_2=0}^1 P(X_1, X_2, Y = 1 - \hat{Y}_1(X_1, X_2)) \\ &= P(X_1 = 0, X_2 = 0, Y = 1 - 0) + P(X_1 = 0, X_2 = 1, Y = 1 - 0) \\ &\quad + P(X_1 = 1, X_2 = 0, Y = 1 - 1) + P(X_1 = 1, X_2 = 1, Y = 1 - 1) \\ &= 0,05 + 0,05 + 0,135 + 0,015 = 0,1 + 0,15 = 0,25. \end{aligned}$$

Similar, dacă luăm în considerare doar variabila X_2 , avem:

X_2	$P(X_2, Y = 0)$	$P(X_2, Y = 1)$	$\hat{Y}_2(X_1, X_2)$
0	$0,9 \cdot 0,5 = 0,45$	$0,5 \cdot 0,5 = 0,25$	0
1	$0,1 \cdot 0,5 = 0,05$	$0,5 \cdot 0,5 = 0,25$	1

Acum, rata medie a erorii va fi:

$$\begin{aligned}
 P(Y \neq \hat{Y}_2(X_1, X_2)) &= \sum_{X_1=0}^1 \sum_{X_2=0}^1 P(X_1, X_2, Y = 1 - \hat{Y}_2(X_1, X_2)) \\
 &= P(X_1 = 0, X_2 = 0, Y = 1 - 0) + P(X_1 = 0, X_2 = 1, Y = 1 - 1) \\
 &\quad + P(X_1 = 1, X_2 = 0, Y = 1 - 0) + P(X_1 = 1, X_2 = 1, Y = 1 - 1) \\
 &= 0,05 + 0,035 + 0,2 + 0,015 = 0,25 + 0,05 = 0,3.
 \end{aligned}$$

Prin urmare, rata medie a erorii la antrenare este mai mică pentru clasificatorul Bayes Naiv care prezice Y având ca input doar X_1 (decât pornind doar de la X_2).⁴¹³

Observație (2):

Atât în cazul lui X_1 cât și în cazul lui X_2 , rata medie a erorii (pentru algoritmul Bayes Naiv) putea fi calculată folosind direct probabilitățile din cele două tabele de mai sus. Justificarea ține de faptul că distribuțiile calculate de Bayes Naiv în aceste două tabele sunt distribuții marginale în raport cu distribuția comună (reală!) din tabelul de la punctul a. La punctul d veți vedea că acolo trebuie procedat altfel, fiindcă în cazul respectiv distribuția calculată de către Bayes Naiv nu mai coincide cu distribuția reală a datelor!

d. Clasificatorul Bayes Naiv care prezice valoarea lui Y în funcție de toate cele trei variabilele $X_1, X_2, X_3 = X_2$ va lua deciziile conform tabelului următor:⁴¹⁴

X_1	X_2	X_3	$P(X_1, X_2, X_3, Y = 0)$	$P(X_1, X_2, X_3, Y = 1)$	$\hat{Y}_3(X_1, X_2)$
0	0	0	$0,7 \cdot 0,9 \cdot 0,9 \cdot 0,5 = 0,2835$	$0,2 \cdot 0,5 \cdot 0,5 \cdot 0,5 = 0,025$	0
0	1	1	$0,7 \cdot 0,1 \cdot 0,1 \cdot 0,5 = 0,0035$	$0,2 \cdot 0,5 \cdot 0,5 \cdot 0,5 = 0,025$	1
1	0	0	$0,3 \cdot 0,9 \cdot 0,9 \cdot 0,5 = 0,1215$	$0,8 \cdot 0,5 \cdot 0,5 \cdot 0,5 = 0,1$	0
1	1	1	$0,3 \cdot 0,1 \cdot 0,1 \cdot 0,5 = 0,0015$	$0,8 \cdot 0,5 \cdot 0,5 \cdot 0,5 = 0,1$	1

Observație (3):

Este util să observați că distribuția de probabilitate comună calculată aici de către algoritmul Bayes Naiv nu mai coincide cu distribuția „reală“ (vedeți tabelul de la punctul a). Rata erorii se calculează în raport cu distribuția „reală“ a datelor, nu cu cea calculată de către Bayes Naiv (deși pentru a identifica situațiile în care $\hat{Y} \neq Y$ se folosește ultimul tabel)!

Rata medie a erorii produse la antrenare va fi:

$$\begin{aligned}
 P(Y \neq \hat{Y}_3(X_1, X_2)) &= \sum_{X_1=0}^1 \sum_{X_2=0}^1 P(X_1, X_2, Y = 1 - \hat{Y}_3(X_1, X_2)) \\
 &= P(X_1 = 0, X_2 = 0, Y = 1 - 0) + P(X_1 = 0, X_2 = 1, Y = 1 - 1) \\
 &\quad + P(X_1 = 1, X_2 = 0, Y = 1 - 0) + P(X_1 = 1, X_2 = 1, Y = 1 - 1) \\
 &= 0,05 + 0,035 + 0,2 + 0,015 = 0,3.
 \end{aligned}$$

e. Diferența dintre cele două rate medii ale erorilor care au fost calculate la punctele a și d se datorează faptului că presupunerea de independență condițională a variabilelor nu este adevărată. Într-adevăr, X_2 nu este independent

⁴¹³Însă ambii clasificatori au rata medie a erorii mai mare decât clasificatorul Bayes Naiv care folosește atât X_1 cât și X_2 , ceea ce era de așteptat întrucât în condițiile date Bayes Naiv are același comportament ca și Bayes Optimal.

⁴¹⁴Este bine de observat că în această situație presupozitia de independență condițională este încălcată. Așadar, Bayes Naiv nu va mai furniza aceleași rezultate ca Bayes Optimal.

condițional față de X_3 deoarece cele două variabile au tot timpul valori identice.

Observații importante:

1. Este imediat că atunci când datele satisfac presupozitia de independență condițională, algoritmii Bayes Naiv și Bayes Optimal produc aceleași rezultate și au aceeași rată medie a erorilor.
2. Se poate arăta ușor că algoritmul Bayes Optimal nu produce în mod neapărat o rată medie a erorilor nulă, chiar dacă lucrează cu distribuția reală a datelor. De exemplu, în condițiile definite inițial de problema noastră, algoritmul Bayes Optimal produce rata medie 0.235, ca și algoritmul Bayes Naiv (vedeți punctul b). Erorile produse de către algoritmul Bayes Optimal se datoreză faptului că el aplică operatorul $\arg \max$, echivalent cu luarea unui vot majoritar (impus deci minorității).

11. (Cât de naiv / prost este algoritmul Bayes Naiv?)

■ • CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh, HW2, pr. 1.2

În mod evident, clasificatorul Bayes Naiv lucrează cu o presupozitie foarte restrictivă (engl., strong presupposition). Însă ne putem întreba dacă acest clasificator nu este totuși destul de folositor chiar și în cazul în care respectiva presupozitie nu este satisfăcută.

În consecință, în acest exercițiu ne propunem să folosim un exemplu simplu pentru a explora limitările algoritmului Bayes Naiv.

Fie X_1 și X_2 variabile aleatoare binare de tip Bernoulli de parametru $p = 0.5$, iar Y o funcție deterministă în raport cu valorile lui X_1 și X_2 , luând valori în multimea $\{1, 2\}$.

a. Definiți Y astfel încât (pe setul de date respectiv) algoritmul Bayes Naiv să aibă rata medie a erorii de 50% (adică maximă!).

Pe acest caz, observați cum se coreleză valorile lui X_1 și X_2 când valoarea lui Y este fixată. (Altfel spus, observați cât de (in)dependente sunt în acest context valorile lui X_1 și X_2 , dată fiind valoarea lui Y .)

X_1	X_2	Y
0	0	0
0	1	1
1	0	1
1	1	1

b. Există în total $2^4 = 16$ moduri în care poate fi definită funcția Y . Însă, datorită simetriei (relativ la valorile lui Y), problema se reduce la doar 4 cazuri,

X_1	X_2	Y									
0	0	1	0	0	1	0	0	1	0	0	1
0	1	1	0	1	1	0	1	2	0	1	2
1	0	1	1	0	1	1	0	1	1	0	2
1	1	1	1	1	2	1	1	2	1	1	1

dintre care un caz corespunde punctului a de mai sus. În fiecare din acele trei cazuri rămase după rezolvarea de la punctul a, arătați că rata erorii produsă de algoritmul Bayes Naiv este 0.

Răspuns:

a. Considerăm Y definit conform tabelului de mai jos.

Observație: Dacă se consideră valoarea lui Y fixată (fie 1, fie 2), atunci putem să stabilim o regulă astfel încât dacă îl cunoaștem pe X_1 să-l determinăm pe X_2 (și invers).⁴¹⁵ Altfel spus, X_1 este unic determinat de X_2 (și invers), dată fiind o valoare fixată a lui Y . Deci condiția de independentă condițională este încălcată. Mai mult, în acest caz avem maximul posibil de „dependență” între cele două variabile (în raport cu Y).

X_1	X_2	Y
0	0	1
0	1	2
1	0	2
1	1	1

Dorim să calculăm rata erorii înregistrate de algoritmul Bayes Naiv pe datele din tabelul de mai sus. Bayes Naiv estimează valoarea lui Y astfel:

$$\hat{y} = \operatorname{argmax}_{y \in \{1,2\}} P(X_1 | Y = y) \cdot P(X_2 | Y = y) \cdot P(Y = y)$$

Pentru $X_1 = 0, X_2 = 0$, algoritmul compară următoarele două valori:

$$\begin{aligned} p_1 &= P(X_1 = 0 | Y = 1) \cdot P(X_2 = 0 | Y = 1) \cdot P(Y = 1) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8} \\ p_2 &= P(X_1 = 0 | Y = 2) \cdot P(X_2 = 0 | Y = 2) \cdot P(Y = 2) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8} \end{aligned}$$

Cum $p_1 = p_2$, algoritmul va alege una dintre ele cu o probabilitate de 0.5. Deoarece valoarea lui Y din tabel este 1, înseamnă că algoritmul va alege greșit cu o probabilitate de 0.5.

Pentru celelalte 3 cazuri, $(X_1 = 0, X_2 = 1)$, $(X_1 = 1, X_2 = 0)$ și $(X_1 = 1, X_2 = 1)$, se observă ușor că se obțin de asemenea valori egale, iar algoritmul va alege pentru Y una dintre valorile 1 sau 2 cu o probabilitate de 0.5.

Deci pentru această definiție a lui Y rata erorii este de 50%.

b. Vom calcula rata erorii pentru fiecare dintre cele 3 moduri de definire a lui Y care nu a fost studiat.

<i>Cazul 1:</i> $\begin{array}{ c c c } \hline X_1 & X_2 & Y \\ \hline 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ \hline \end{array}$			<i>Este similar cu cazul:</i> $\begin{array}{ c } \hline Y \\ \hline 2 \\ 2 \\ 2 \\ 2 \\ \hline \end{array}$
--	--	--	--

- Pentru $X_1 = 0, X_2 = 0$, algoritmul compară:

$$\begin{aligned} p_1 &= P(X_1 = 0 | Y = 1) \cdot P(X_2 = 0 | Y = 1) \cdot P(Y = 1) = \frac{2}{4} \cdot \frac{2}{4} \cdot 1 = \frac{1}{4} \\ p_2 &= P(X_1 = 0 | Y = 2) \cdot P(X_2 = 0 | Y = 2) \cdot P(Y = 2) = 0 \cdot 0 \cdot 0 = 0 \end{aligned}$$

Cum $p_1 > p_2$ algoritmul alege pentru Y valoarea 1, ceea ce este corect.

- Pentru celelalte 3 cazuri, $(X_1 = 0, X_2 = 1)$, $(X_1 = 1, X_2 = 0)$ și $(X_1 = 1, X_2 = 1)$, se observă că se obțin aceleași valori pentru p_1 și p_2 ca mai sus, deci algoritmul alege (în mod corect) pentru Y valoarea 1.

⁴¹⁵Pentru $Y = 1$, regula este: X_2 are aceeași valoare ca și X_1 . Pentru $Y = 2$, regula este: X_1 și X_2 au valori complementare.

Așadar, am obținut că rata erorii este în acest caz 0.

Cazul 2:			Cazuri similare:		
X_1	X_2	Y	Y	Y	Y
0	0	1	1	2	2
0	1	1	1	2	2
1	0	1	2	1	2
1	1	2	1	1	2

- Pentru $X_1 = 0, X_2 = 0$:

$$\left. \begin{aligned} p_1 &= P(X_1 = 0 | Y = 1) \cdot P(X_2 = 0 | Y = 1) \cdot P(Y = 1) = \frac{2}{3} \cdot \frac{2}{3} \cdot \frac{3}{4} = \frac{1}{3} \\ p_2 &= P(X_1 = 0 | Y = 2) \cdot P(X_2 = 0 | Y = 2) \cdot P(Y = 2) = 0 \cdot 0 \cdot \frac{1}{4} = 0 \end{aligned} \right\} \Rightarrow \hat{y} = 1$$

- Pentru $X_1 = 0, X_2 = 1$:

$$\left. \begin{aligned} p_1 &= P(X_1 = 0 | Y = 1) \cdot P(X_2 = 1 | Y = 1) \cdot P(Y = 1) = \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{3}{4} = \frac{1}{6} \\ p_2 &= P(X_1 = 0 | Y = 2) \cdot P(X_2 = 1 | Y = 2) \cdot P(Y = 2) = 0 \cdot 1 \cdot \frac{1}{4} = 0 \end{aligned} \right\} \Rightarrow \hat{y} = 1$$

- Pentru $X_1 = 1, X_2 = 0$:

$$\left. \begin{aligned} p_1 &= P(X_1 = 1 | Y = 1) \cdot P(X_2 = 0 | Y = 1) \cdot P(Y = 1) = \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{3}{4} = \frac{1}{6} \\ p_2 &= P(X_1 = 1 | Y = 2) \cdot P(X_2 = 0 | Y = 2) \cdot P(Y = 2) = 1 \cdot 0 \cdot \frac{1}{4} = 0 \end{aligned} \right\} \Rightarrow \hat{y} = 1$$

- Pentru $X_1 = 1, X_2 = 1$:

$$\left. \begin{aligned} p_1 &= P(X_1 = 1 | Y = 1) \cdot P(X_2 = 1 | Y = 1) \cdot P(Y = 1) = \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{3}{4} = \frac{1}{12} \\ p_2 &= P(X_1 = 1 | Y = 2) \cdot P(X_2 = 1 | Y = 2) \cdot P(Y = 2) = 1 \cdot 1 \cdot \frac{1}{4} = \frac{1}{4} \end{aligned} \right\} \Rightarrow \hat{y} = 2$$

Deci rata erorii este 0 pentru această definiție a lui Y .

Cazul 3:			Cazuri similare:		
X_1	X_2	Y	Y	Y	Y
0	0	1	2	1	2
0	1	2	1	1	2
1	0	1	2	2	1
1	1	2	1	2	1

- Pentru $X_1 = 0, X_2 = 0$:

$$\left. \begin{aligned} p_1 &= \frac{1}{2} \cdot 1 \cdot \frac{1}{2} = \frac{1}{4} \\ p_2 &= \frac{1}{2} \cdot 0 \cdot \frac{1}{2} = 0 \end{aligned} \right\} \Rightarrow p_1 > p_2 \Rightarrow \hat{y} = 1 \text{ (corect)}$$

- Pentru $X_1 = 0, X_2 = 1$:

$$\left. \begin{aligned} p_1 &= \frac{1}{2} \cdot 0 \cdot \frac{1}{2} = 0 \\ p_2 &= \frac{1}{2} \cdot 1 \cdot \frac{1}{2} = \frac{1}{4} \end{aligned} \right\} \Rightarrow p_1 < p_2 \Rightarrow \hat{y} = 2 \text{ (corect)}$$

- Pentru $X_1 = 1, X_2 = 0$:

$$\left. \begin{aligned} p_1 &= \frac{1}{2} \cdot 1 \cdot \frac{1}{2} = \frac{1}{4} \\ p_2 &= \frac{1}{2} \cdot 0 \cdot \frac{1}{2} = 0 \end{aligned} \right\} \Rightarrow p_1 > p_2 \Rightarrow \hat{y} = 1 \text{ (corect)}$$

- Pentru $X_1 = 1, X_2 = 1$:

$$\left. \begin{array}{l} p_1 = \frac{1}{2} \cdot 0 \cdot \frac{1}{2} = 0 \\ p_2 = \frac{1}{2} \cdot 1 \cdot \frac{1}{2} = \frac{1}{4} \end{array} \right\} \Rightarrow p_1 < p_2 \Rightarrow \hat{y} = 2 \text{ (corect)}$$

Prin urmare, rata erorii este 0 și în acest caz.

Cazul 4: (cel de la punctul a)	X_1	X_2	Y	Este similar cu cazul:	Y
	0	0	1		2
	0	1	2		1
	1	0	2		1
	1	1	1		2

Concluzie și... o observație interesantă:

Se constată că doar pentru 2 moduri (și anume, cazul 4) de definire a lui Y rata erorii este de 50%; pentru celelalte 14 moduri (cazurile 1, 2, 3) rata erorii este 0. Este imediat că aceste două moduri de definire a funcției Y (XOR și respectiv \neg XOR) corespund exact acelor moduri / funcții — dintre toate cele 16 moduri posibile — care au reprezentări care sunt neseparabile liniar în spațiul \mathbb{R}^2 . Acest fapt este în concordanță cu rezultatul demonstrat la problema 14. Acolo se arată că regulii de decizie a algoritmului Bayes Naiv îi corespunde (în general) un separator liniar, similar cu cel din cazul regresiei logistice.

12.

(O reprezentare grafică a neconcordanței deciziilor luate de algoritmii Bayes Naiv și Bayes Optimal)

*CMU, 2009 fall, Geoff Gordon, HW4, pr. 1
CMU, 2009 fall, Carlos Guestrin, HW1, pr. 4.1.5*

Pentru un task de clasificare se consideră atributele X_1, X_2 și X_3 și eticheta Y . Toate acestea sunt variabile aleatoare binare. X_1 și X_2 sunt independente condițional în raport cu Y , iar X_3 este o copie a lui X_2 (așadar, întotdeauna $X_2 = X_3$).

Sunt date următoarele probabilități condiționate:

$$\begin{aligned} P(X_1 = T \mid Y = T) &= p, & P(X_1 = T \mid Y = F) &= 1 - p, \\ P(X_2 = F \mid Y = T) &= q, & P(X_2 = F \mid Y = F) &= 1 - q, \\ P(Y = T) &= 0.5. \end{aligned}$$

Se dă instanța de test $X_1 = T, X_2 = X_3 = F$. Vrem să clasificăm această instanță, adică să prezicem valoarea lui Y pentru ea.

- Arătați că dacă se folosește clasificatorul Bayes Naiv, atunci eticheta pentru instanța aceasta de test este T — ceea ce revine la $P(Y = T \mid X_1 = T, X_2 = F, X_3 = F) \geq 0.5$ — dacă $p \geq \frac{(1-q)^2}{q^2 + (1-q)^2}$.
- Ce devine inegalitatea de la punctul precedent dacă în schimbul clasificatorului Bayes Naiv se utilizează clasificatorul Bayes Optimal?
- Desenați cele două curbe de decizie obținute la punctele a și b . Pe axa Ox marcați valorile lui q , iar pe axa Oy marcați valorile lui p . Atât p cât și q

variază în intervalul $[0, 1]$. Indicați pe grafic zona în care clasificatorul Bayes Naiv produce un output (Y) diferit de cel al algoritmului Bayes Optimal.

Atenție! Acest exercițiu *nu* studiază în ce condiții cei doi algoritmi clasifică corect (ori dimpotrivă, eronat) instanța $X_1 = T, X_2 = X_3 = F$, ci doar când anume (în funcție de p și q) produc ei clasificări diferite pentru această instanță. Se va vedea (grafic!) că algoritmul Bayes Naiv nu se comportă deloc rău în comparație cu algoritmul Bayes Optimal.

Răspuns:

Se lucrează cu variabile aleatoare binare, iar pentru claritatea calculelor vom nota prin X faptul că valoarea variabilei aleatoare X este T , iar prin $\neg X$ faptul că $X = F$.

Folosind aceste notății, putem transcrie datele din enunț astfel:

$$\begin{aligned} P(X_1 | Y) &= p, & P(X_1 | \neg Y) &= 1 - p, \\ P(\neg X_2 | Y) &= q, & P(\neg X_2 | \neg Y) &= 1 - q, \\ P(Y) &= 0.5. \end{aligned}$$

a. Eticheta pentru instanța aceasta de test este T dacă $P(Y | X_1, \neg X_2, \neg X_3) \geq 0.5$. Vom calcula această probabilitate utilizând regula lui Bayes precum și ipoteza de independentă condițională făcută de algoritmul Bayes Naiv:

$$\begin{aligned} P(Y | X_1, \neg X_2, \neg X_3) &\stackrel{\text{form. Bayes}}{=} \\ &= \frac{P(X_1, \neg X_2, \neg X_3 | Y)P(Y)}{P(X_1, \neg X_2, \neg X_3 | Y)P(Y) + P(X_1, \neg X_2, \neg X_3 | \neg Y)P(\neg Y)} \stackrel{\text{indep. cdt.}}{=} \\ &= \frac{P(X_1 | Y)P(\neg X_2 | Y)P(\neg X_3 | Y)P(Y)}{P(X_1 | Y)P(\neg X_2 | Y)P(\neg X_3 | Y)P(Y) + P(X_1 | \neg Y)P(\neg X_2 | \neg Y)P(\neg X_3 | \neg Y)P(\neg Y)} \\ &= \frac{p \cdot q \cdot q \cdot 0.5}{p \cdot q \cdot q \cdot 0.5 + (1 - p) \cdot (1 - q) \cdot (1 - q) \cdot 0.5} = \frac{pq^2}{pq^2 + (1 - p)(1 - q)^2} \end{aligned}$$

Deci eticheta este T dacă $\frac{pq^2}{pq^2 + (1 - p)(1 - q)^2} \geq 0.5$, ceea ce înseamnă că

$$\begin{aligned} pq^2 \geq 0.5(pq^2 + (1 - p)(1 - q)^2) &\Leftrightarrow pq^2 - 0.5pq^2 \geq 0.5(1 - p)(1 - q)^2 \\ \Leftrightarrow pq^2 \geq (1 - q)^2 - p(1 - q)^2 &\Leftrightarrow p(q^2 + (1 - q)^2) \geq (1 - q)^2 \Leftrightarrow p \geq \frac{(1 - q)^2}{q^2 + (1 - q)^2} \end{aligned}$$

b. Dacă în locul clasificatorului Bayes Naiv se folosește clasificatorul Bayes Optimal, nu se mai folosește presupunerea de independentă condițională. În locul acesteia se folosesc informațiile furnizate în enunț, și anume că X_1 și X_2 sunt independente condițional în raport cu Y , iar X_3 este o copie a lui X_2 .

$$\begin{aligned} P(Y | X_1, \neg X_2, \neg X_3) &\stackrel{\text{form. Bayes}}{=} \\ &= \frac{P(X_1, \neg X_2, \neg X_3 | Y)P(Y)}{P(X_1, \neg X_2, \neg X_3 | Y)P(Y) + P(X_1, \neg X_2, \neg X_3 | \neg Y)P(\neg Y)} \\ &= \frac{P(X_1 | Y)P(\neg X_2, \neg X_3 | Y)P(Y)}{P(X_1 | Y)P(\neg X_2, \neg X_3 | Y)P(Y) + P(X_1 | \neg Y)P(\neg X_2, \neg X_3 | \neg Y)P(\neg Y)} \\ &= \frac{P(X_1 | Y)P(\neg X_2 | Y)P(Y)}{P(X_1 | Y)P(\neg X_2 | Y)P(Y) + P(X_1 | \neg Y)P(\neg X_2 | \neg Y)P(\neg Y)} \end{aligned}$$

$$= \frac{p \cdot q \cdot 0.5}{p \cdot q \cdot 0.5 + (1-p) \cdot (1-q) \cdot 0.5} = \frac{pq}{pq + (1-p)(1-q)}$$

S-a folosit egalitatea

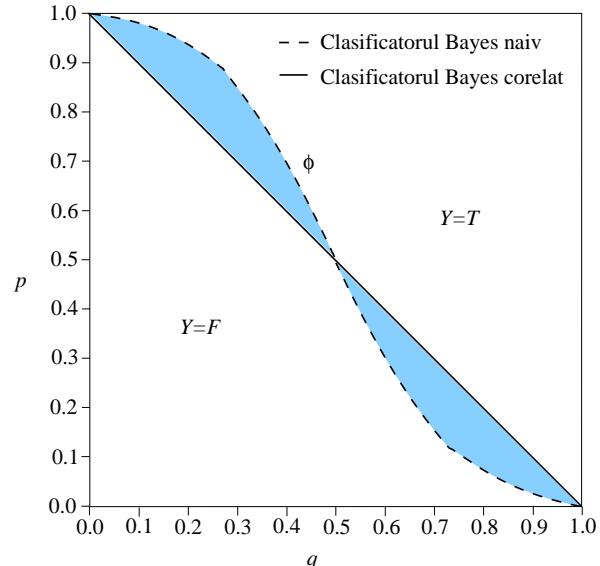
$$P(X_1, \neg X_2, \neg X_3 | Y) = P(X_1 | \neg X_2, \neg X_3, Y) \cdot P(\neg X_2, \neg X_3 | Y) = P(X_1 | Y) \cdot P(\neg X_2 | Y).$$

În acest caz eticheta este T dacă $\frac{pq}{pq + (1-p)(1-q)} \geq 0.5$, adică

$$\begin{aligned} pq \geq 0.5(pq + (1-p)(1-q)) &\Leftrightarrow pq - 0.5pq \geq 0.5(1-p-q+pq) \\ &\Leftrightarrow pq \geq 1-p-q+pq \Leftrightarrow p \geq 1-q. \end{aligned}$$

c. Dreapta de ecuație $p = 1 - q$ este ușor de reprezentat.

Notând $\phi(q) = \frac{(1-q)^2}{q^2 + (1-q)^2}$, se observă imediat că $\phi(q) + \phi(1-q) = 1 \Leftrightarrow \phi(1-q) = 1 - \phi(q)$. De aici, cu ajutorul unui raționament geometric simplu, se ajunge imediat la concluzia că graficul funcției ϕ este simetric față de punctul de coordonate $(1/2, 1/2)$. Așadar, va fi suficient să studiem graficul lui ϕ pe intervalul $[0, 1/2]$. Proprietățile funcției ϕ necesare elaborării graficului sunt ușor de studiat. Adițional, se poate arăta imediat că $\phi(q) \geq 1 - q$ pentru orice $q \in [0, 1/2]$ și $\phi(q) \leq 1 - q$ pentru orice $q \in [1/2, 1]$.



În figura de mai sus am reprezentat curbele $p = 1 - q$ și $p = \phi(q)$, precum și zonele de decizie pentru cei doi clasificatori obținuți la punctele a și b . Se observă ușor zona în care rezultatul produs de clasificatorul Bayes Naiv (Y) este diferit / „eronaț“ în raport cu algoritmul Bayes Optimal.

13.

(Cât de multe date de antrenament necesită algoritmul Bayes Naiv vs. algoritmul Bayes Optimal?
[LC: complexitatea la eșantionare])

■ • CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh, HW2, pr. 1.1

Unul dintre motivele pentru care folosim clasificatorul Bayes Naiv este faptul că el necesită mult mai puține date de antrenament (în vederea estimării parametrilor) decât clasificatorul Bayes Optimal.

Acest exercițiu te va ajuta să înțelegi cât de importantă este această diferență dintre cei doi algoritmi.

Presupunem că o *observație / instanță* este o valoare generată în mod aleatoriu de către vectorul de variabile aleatoare $\bar{X} = (X_1, \dots, X_{d-1}, X_d)$, unde fiecare

X_i este o variabilă aleatoare urmând distribuția probabilistă Bernoulli de parametru $p = 0.5$. Considerăm X_1, \dots, X_{d-1} variabilele de intrare, iar $X_d = Y$ variabila de ieșire.

Pentru a estima în sensul verosimilității maxime (MLE) parametrii clasificatorului Bayes Optimal, avem nevoie să observăm / întâlnim fiecare valoare a lui \bar{X} de un număr rezonabil de ori. Similar, pentru a antrena clasificatorul Bayes Naiv avem nevoie să întâlnim fiecare valoare a fiecărei variabile X_i ($i = \overline{1, d}$) de un număr rezonabil de ori.

Ne întrebăm cât de multe observații sunt necesare (a fi generate) pentru ca fiecare valoare a variabilei comune \bar{X} în cazul algoritmului Bayes Optimal, și respectiv fiecare valoare a variabilelor X_i ($i = \overline{1, d}$) în cazul Bayes Naiv să fie întâlnită cel puțin o dată. (În practică este nevoie de mult mai multe observații, dar în acest exercițiu ne limităm la câte o singură observație pentru fiecare valoare în parte.)

Indicație: La rezolvarea punctelor de mai jos vă sugerăm să folosiți următoarele două inegalități:

- pentru orice evenimente E_1, \dots, E_n , avem $P(E_1 \cup \dots \cup E_n) \leq \sum_{i=1}^n P(E_i)$.⁴¹⁶
- $(1 - \frac{1}{k})^k \leq \frac{1}{e}$ pentru orice $k \geq 1$, unde $e \approx 2.71828$ este baza logaritmului natural.

a. Începem cu algoritmul Bayes Naiv. Fie $i \in \{1, \dots, d\}$ fixat. Arătați că dacă s-au făcut N observații (având forma $\bar{x}_j = (x_1^j, \dots, x_{d-1}^j, x_d^j)$ cu $j = 1, \dots, N$), atunci probabilitatea să nu fi întâlnit ambele valori ale variabilei X_i este $\frac{1}{2^{N-1}}$. (Observați că această fracție reprezintă un număr foarte mic atunci când N este suficient de mare.)

b. Fie $\varepsilon > 0$ fixat. Folosind prima inegalitate din *indicația* de mai sus, arătați că dacă au fost făcute câte $N_{NB} = 1 + \log_2 \frac{d}{\varepsilon}$ observații de aceeași formă ca mai sus, atunci probabilitatea să nu fi întâlnit ambele valori pentru fiecare dintre variabilele X_i ($i = \overline{1, d}$) este mai mică sau egală cu ε .

c. Acum trecem la algoritmul Bayes Optimal. Fie \bar{x} o instanță (fixată) a variabilei comune \bar{X} . Folosind a doua inegalitate din *indicația* de mai sus, arătați că dacă s-au făcut N observații (fiecare observație implicând simultan toate variabilele X_i cu $i = \overline{1, d}$), atunci probabilitatea ca să nu se fi întâlnit niciodată \bar{x} este mai mică sau egală cu $e^{-\frac{N}{2^d}}$.

d. Arătați că dacă au fost făcute cel puțin $N_{JB} = 2^d \ln \frac{2^d}{\varepsilon}$ observații, atunci probabilitatea ca să nu se fi întâlnit toate instanțele variabilei comune \bar{X} este mai mică sau egală cu ε .

e. Dacă se fixează $\varepsilon = 0.1$, calculați valorile N_{NB} și N_{JB} pentru $d = 2$, $d = 5$ și $d = 10$. Ce concluzie puteți trage? (Altfel spus, cum interpretați rezultatele?)

Observații:

1. Știm că pentru clasificatorul Bayes Naiv, este necesar să estimăm din date

⁴¹⁶Aceasta se numește proprietatea de subaditivitate a probabilităților.

probabilitățile $P(Y = y)$ — adică, $P(X_d = x_d)$ — și $P(X_i = x_i|Y = y)$, pentru $i = 1, \dots, d - 1$. Putem considera că este suficient să întâlnim cu o probabilitate de cel puțin $1 - \varepsilon$ toate valorile y , precum și perechile de forma (x_i, y) , cu $i = 1, \dots, d - 1$. În mod implicit, problema noastră simplifică și mai mult cerințele, considerând că este suficient să găsim cu probabilitate de cel puțin $1 - \varepsilon$ toate valorile x_i , pentru $i = 1, \dots, d$ (subînțelegând că atunci vor apărea în date, cu probabilități semnificative, atât valorile y cât și fiecare dintre combinațiile (x_i, y) , cu $i = 1, \dots, d - 1$).

2. Pentru clasificatorul Bayes Optimal, în mod similar, este necesar să estimăm probabilitățile $P(Y = y)$ și $P(X_1 = x_1, \dots, X_{d-1} = x_{d-1}|Y = y)$, ceea ce, evident, va permite calcularea probabilităților de forma $P(x_1, \dots, x_{d-1}, y)$. Problema noastră consideră în mod implicit că este suficient să găsim cu probabilitate de cel puțin $1 - \varepsilon$ toate combinațiile (x_1, \dots, x_{d-1}, y) .

Răspuns:

a. Dacă s-au facut N observații și nu s-au întâlnit ambele valori ale variabilei X_i , înseamnă că ea are aceeași valoare în toate aceste observații (adică ea este fie 0 în toate observațiile, fie 1 în toate observațiile). Pentru fiecare dintre aceste două cazuri probabilitatea este $1/2^N$, deci:

$$P(\text{doar una dintre valorile variabilei } X_i \text{ a apărut în } N \text{ observații})$$

$$= \left(\frac{1}{2}\right)^N + \left(\frac{1}{2}\right)^N = \frac{2}{2^N} = \frac{1}{2^{N-1}}$$

b. Se cere să se calculeze probabilitatea să nu fi întâlnit ambele valori pentru fiecare dintre variabilele X_1, \dots, X_{d-1}, X_d . Pentru acesta vom folosi prima inegalitate din indicația dată în enunț:

$$P(\text{nu toate valorile variabilelor } X_i, i = \overline{1, d}, \text{ au apărut în } N_{NB} \text{ observații})$$

$$\begin{aligned} &\leq \sum_{i=1}^d P(\text{numai una dintre valorile variabilei } X_i \text{ a apărut în } N_{NB} \text{ observații}) \\ &= \sum_{i=1}^d \frac{1}{2^{N_{NB}-1}} = d \cdot \frac{1}{2^{N_{NB}-1}} = d \cdot \frac{1}{2^{1+\log_2 \frac{d}{\varepsilon}-1}} = d \cdot \frac{1}{2^{\log_2 \frac{d}{\varepsilon}}} = d \cdot \frac{1}{\frac{d}{\varepsilon}} = d \cdot \frac{\varepsilon}{d} = \varepsilon. \end{aligned}$$

c. Pentru algoritmul Bayes Optimal s-au făcut N observații. Trebuie să calculăm probabilitatea ca să nu se fi întâlnit niciodată instanța \bar{x} .

Cum există în total 2^d posibile observații, probabilitatea ca instanța \bar{x} să nu fie obținută la *una* dintre observații este $1 - \frac{1}{2^d}$. Observațiile fiind independente, probabilitatea ca \bar{x} să nu fie obținută *după* N observații este $\left(1 - \frac{1}{2^d}\right)^N$. Așadar,

$$\begin{aligned} P(\text{instanța } \bar{x} \text{ n-a fost întâlnită în } N \text{ observații}) &= \left(1 - \frac{1}{2^d}\right)^N \\ &= \left[\left(1 - \frac{1}{2^d}\right)^{2^d}\right]^{N/2^d} \leq \left(\frac{1}{e}\right)^{N/2^d} = e^{-N/2^d} \end{aligned}$$

d. Vom calcula probabilitatea ca să nu se fi întâlnit toate instanțele variabilei \bar{X} utilizând din nou prima inegalitate din *indicație*:

$$\begin{aligned} & P(\text{nu toate instanțele variabilei } \bar{X} \text{ au fost întâlnite în } N_{JB} \text{ observații}) \\ & \leq \sum_{\bar{x}} P(\text{instanța } \bar{x} \text{ n-a fost întâlnită în } N_{JB} \text{ observații}) \\ & \leq \sum_{\bar{x}} e^{-N_{JB}/2^d} = 2^d \cdot e^{-N_{JB}/2^d} = 2^d \cdot e^{-\ln \frac{2^d}{\varepsilon}} = 2^d \cdot \frac{1}{e^{\ln \frac{2^d}{\varepsilon}}} = \frac{2^d}{\varepsilon} = \varepsilon. \end{aligned}$$

e. Vom înlocui datele numerice în formulele $N_{NB} = 1 + \log_2 \frac{d}{\varepsilon}$, $N_{JB} = 2^d \ln \frac{2^d}{\varepsilon}$.

$$\begin{aligned} \varepsilon = 0.1, d = 2 \Rightarrow & \begin{cases} N_{NB} = 1 + \log_2 \frac{2}{0.1} = 1 + \log_2 20 \approx 5.32 \\ N_{JB} = 2^2 \cdot \ln \frac{2^2}{0.1} = 4 \cdot \ln 40 \approx 14.75 \end{cases} \\ \varepsilon = 0.1, d = 5 \Rightarrow & \begin{cases} N_{NB} = 1 + \log_2 \frac{5}{0.1} = 1 + \log_2 50 \approx 6.64 \\ N_{JB} = 2^5 \cdot \ln \frac{2^5}{0.1} = 32 \cdot \ln 320 \approx 184.58 \end{cases} \\ \varepsilon = 0.1, d = 10 \Rightarrow & \begin{cases} N_{NB} = 1 + \log_2 \frac{10}{0.1} = 1 + \log_2 100 \approx 7.64 \\ N_{JB} = 2^{10} \cdot \ln \frac{2^{10}}{0.1} = 1024 \cdot \ln 10240 \approx 9455.67 \end{cases} \end{aligned}$$

Acum se observă ușor [diferența dintre] numărul de date de antrenament necesare pentru cei doi algoritmi: de ordin logaritmice pentru Bayes Naiv, respectiv de ordin exponențial pentru Bayes Optimal.

14.

(Algoritmul Bayes Naiv: raportul cu regresia logistică și natura separatorului decizional; cazul când variabilele de intrare sunt de tip boolean)

■ □ • ○ CMU, 2005 fall, T. Mitchell, A. Moore, HW2, pr. 2
CMU, 2009 fall, Carlos Guestrin, HW1, pr. 4.1.2
CMU, 2009 fall, Geoff Gordon, HW4, pr. 1.2-3
CMU, 2012 fall, T. Mitchell, Z. Bar-Joseph, HW2, pr. 3.a

a. [Bayes Naiv și Regresia Logistică: relația dintre regulile de decizie]⁴¹⁷

Fie Y o variabilă aleatoare cu valori booleene, iar $X = (X_1, \dots, X_d)$ un vector de variabile aleatoare cu valori booleene. Demonstrați că distribuția condițională $P(Y|X)$ are forma funcției logistice de argument $z = -(w_0 + w_1 X_1 + \dots + w_d X_d)$,

⁴¹⁷ Pentru o introducere în chestiunea regresiei logistice, vedeti Tom Mitchell, *Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression*, draft pentru un capitol suplimentar pentru o nouă ediție a cărții *Machine Learning*, 2016. Puteți vedea de asemenea problema 13 de la capitolul *Metode de regresie* din prezenta culegere.

cu parametrii $w_0, w_1, \dots, w_d \in \mathbb{R}$,⁴¹⁸ adică

$$P(Y = 1|X) = \frac{1}{1 + \exp(w_0 + \sum_{i=1}^d w_i X_i)}$$

și, prin urmare

$$P(Y = 0|X) = \frac{\exp(w_0 + \sum_{i=1}^d w_i X_i)}{1 + \exp(w_0 + \sum_{i=1}^d w_i X_i)}.$$

Vă reamintim că *funcția logistică* (sau *sigmoidală*) este definită prin expresia $\sigma(z) = 1/(1 + e^{-z})$ pentru orice $z \in \mathbb{R}$.

Comentariu:

Regresia logistică (și, mai general, *clasificatorii „discriminativi“*) învață [în mod] direct parametrii distribuției $P(Y|X)$,⁴¹⁹ pe când algoritmul Bayes Naiv (și, mai general, *clasificatorii „generativi“*) învață [parametrii pentru] distribuțiile $P(X|Y)$ și $P(Y)$, cu ajutorul căror va calcula apoi $P(Y|X)$ și cea mai probabilă valoare pentru Y (atunci când X are o valoare fixată / dată). Vom spune că regresia logistică este corespondentul „discriminativ“ al clasificatorului „generativ“ Bayes Naiv.

Indicații:

1. Vom introduce o *notație* simplă, care ne va fi de folos în continuare. Întrucât variabilele X_i sunt booleene, odată fixată o valoare y_k pentru variabila Y , vom avea nevoie de un singur parametru pentru a defini distribuția condițională $P(X_i|Y = y_k)$, pentru fiecare $i = 1, \dots, d$. Așadar, vom desemna cu θ_{i1} probabilitatea $P(X_i = 1|Y = 1)$ și, prin urmare $P(X_i = 0|Y = 1) = 1 - \theta_{i1}$. În mod similar, vom desemna cu θ_{i0} probabilitatea $P(X_i = 1|Y = 0)$.

2. Remarcați că odată ce am introdus notațiile de mai sus, vom putea scrie $P(X_i|Y = 1)$ după cum urmează:

$$P(X_i|Y = 1) = \theta_{i1}^{X_i} (1 - \theta_{i1})^{(1-X_i)}, \quad (213)$$

bineînțeles, cu excepția cazurilor când $\theta_{i1} = 0$ și $X_i = 0$, respectiv $\theta_{i1} = 1$ și $X_i = 1$. Observați că atunci când X_i are valoarea 1, cel de-al doilea factor din partea dreaptă a egalității (213) este 1, pentru că exponentul lui este zero. Deci $P(X_i|Y = 1) = \theta_{i1}^{X_i} = \theta_{i1}$ pentru $X_i = 1$. În mod similar, atunci când $X_i = 0$ primul factor este egal cu 1, pentru că exponentul lui este zero. Deci $P(X_i|Y = 1) = (1 - \theta_{i1})^{1-X_i} = 1 - \theta_{i1}$ pentru $X_i = 0$.

b. [Relaxarea presupozitiei de independență condițională]⁴²⁰

Pentru a putea exprima interacțiunile dintre trăsături, modelul regresiei logistice poate fi extins cu niște termeni suplimentari. De exemplu, putem adăuga un termen care să exprime dependența dintre trăsăturile X_1 și X_2 :

$$P(Y = 1|X) = \frac{1}{1 + \exp(w_0 + w_{1,2}X_1 X_2 + \sum_{i=1}^d w_i X_i)}.$$

⁴¹⁸LC: Prin urmare, *separatorul decizional* (sau, *granița de decizie*) pentru algoritmul Bayes Naiv este — într-o astfel de situație — liniar (în funcție de argumentele X_1, \dots, X_d). Ecuția separatorului decizional va fi $w_0 + w_1 X_1 + \dots + w_d X_d = 0$.

⁴¹⁹LC: Parametrii distribuției $P(Y|X)$ sunt în acest caz $w_i \in \mathbb{R}$, cu $i = 0, 1, \dots, d$, iar învățarea lor se face prin maximizarea funcției de verosimilitate $\mathcal{L}(w) \stackrel{not.}{=} P(D|w)$, unde D este setul de date de antrenament. La rândul ei, maximizarea aceasta se realizează prin aplicarea unei metode de optimizare, de exemplu metoda gradientului ascendent sau metoda lui Newton. Vedeti de exemplu problemele 13 și 14 de la capitolul *Metode de regresie*.

⁴²⁰Vedeți de exemplu problema 9.

În mod similar, presupozitia de independentă condițională asumată de către algoritmul Bayes Naiv poate fi relaxată astfel încât trăsăturile X_1 și X_2 să nu mai trebuiască să satisfacă independentă condițională. Așadar, vom putea scrie:

$$P(Y|X) = \frac{P(Y) P(X_1, X_2|Y) \prod_{i=3}^d P(X_i|Y)}{P(X)}.$$

Demonstrați că în acest caz distribuția $P(Y|X)$ are aceeași formă ca și modelul de regresie logistică augmentat cu un termen suplimentar, care exprimă dependența dintre X_1 și X_2 (și, în acest fel, modelul extins al regresiei logistice rămâne corespondentul discriminativ al clasificatorului nostru generativ).

Indicații:

3. De data aceasta o altă notație simplă ne va ajuta. Vom avea nevoie de mai mulți parametri decât la punctul a pentru a defini distribuția comună $P(X_1, X_2|Y)$. Așa că vom nota $\beta_{ijk} = P(X_1 = i, X_2 = j|Y = k)$, pentru fiecare combinație posibilă de valori pentru indicii i, j și k .
4. Această nouă notație poate fi folosită acum pentru a exprima probabilitatea $P(X_1, X_2|Y = k)$ după cum urmează:

$$P(X_1, X_2|Y = k) = (\beta_{11k})^{X_1 X_2} (\beta_{10k})^{X_1(1-X_2)} (\beta_{01k})^{(1-X_1)X_2} (\beta_{00k})^{(1-X_1)(1-X_2)} \quad (214)$$

pentru $k \in \{0, 1\}$, cu excepția următoarelor cazuri: *i.* $\beta_{11k} = 0$ și $X_1 X_2 = 0$, *ii.* $\beta_{10k} = 0$ și $X_1(1 - X_2) = 0$, *iii.* $\beta_{01k} = 0$ și $(1 - X_1)X_2 = 0$ și *iv.* $\beta_{00k} = 0$ și $(1 - X_1)(1 - X_2) = 0$.

Răspuns:

- a. Mai întâi vom rescrie probabilitatea $P(Y = 1|X = x)$ ca o fracție, folosind formula lui Bayes, compusă cu formula probabilității totale, apoi vom împărți atât numărătorul cât și numitorul fracției astfel obținute cu expresia de la numărător:⁴²¹

$$\begin{aligned} P(Y = 1|X = x) &\stackrel{FB}{=} \frac{P(X = x|Y = 1) P(Y = 1)}{\sum_{y' \in \{0, 1\}} P(X = x|Y = y') P(Y = y')} \\ &= \frac{1}{1 + \frac{P(X = x|Y = 0) P(Y = 0)}{P(X = x|Y = 1) P(Y = 1)}}. \end{aligned}$$

După aceea, folosind formula $e^{\ln a} = a$ (valabilă pentru orice $a > 0$), vom forța punerea fracției de mai sus într-o formă apropiată de cea a funcției sigmoidale ($\sigma(x) = 1/(1 + e^{-x})$):⁴²²

$$\begin{aligned} P(Y = 1|X = x) &= \frac{1}{1 + \exp \left(\ln \frac{P(X = x|Y = 0) P(Y = 0)}{P(X = x|Y = 1) P(Y = 1)} \right)} \\ &= \frac{1}{1 + \exp \left(\ln \frac{P(X_1 = x_1, \dots, X_d = x_d|Y = 0) P(Y = 0)}{P(X_1 = x_1, \dots, X_d = x_d|Y = 1) P(Y = 1)} \right)}. \end{aligned}$$

⁴²¹Cu excepția cazului când $P(X = x|Y = 1) P(Y = 1) = 0$.

⁴²²Cu excepția cazului când $P(X = x|Y = 0) P(Y = 0) = 0$.

Mai departe, ținând cont de presupozitia de independență condițională și de proprietățile funcției logaritm, obținem:⁴²³

$$P(Y = 1|X = x) = \frac{1}{1 + \exp\left(\ln \frac{P(Y = 0)}{P(Y = 1)} + \sum_{i=1}^d \ln \frac{P(X_i = x_i|Y = 0)}{P(X_i = x_i|Y = 1)}\right)}.$$

Vom nota probabilitățile a priori $P(Y = 1)$ și $P(Y = 0)$ cu π și respectiv $1 - \pi$. Apoi, conform *indicatiei 2*, vom scrie $P(X_i|Y = 1)$ ca $\theta_{i1}^{X_i}(1 - \theta_{i1})^{(1-X_i)}$ și $P(X_i|Y = 0)$ ca $\theta_{i0}^{X_i}(1 - \theta_{i0})^{(1-X_i)}$. În consecință,

$$\begin{aligned} P(Y = 1|X = x) &= \frac{1}{1 + \exp\left(\ln \frac{1 - \pi}{\pi} + \sum_{i=1}^d \ln \frac{\theta_{i0}^{X_i}(1 - \theta_{i0})^{(1-X_i)}}{\theta_{i1}^{X_i}(1 - \theta_{i1})^{(1-X_i)}}\right)} \\ &= \frac{1}{1 + \exp\left(\ln \frac{1 - \pi}{\pi} + \sum_{i=1}^d \left(X_i \ln \frac{\theta_{i0}}{\theta_{i1}} + (1 - X_i) \ln \frac{1 - \theta_{i0}}{1 - \theta_{i1}}\right)\right)} \\ &= \frac{1}{1 + \exp\left(\ln \frac{1 - \pi}{\pi} + \sum_{i=1}^d \ln \frac{1 - \theta_{i0}}{1 - \theta_{i1}} + \sum_{i=1}^d X_i \left(\ln \frac{\theta_{i0}}{\theta_{i1}} - \ln \frac{1 - \theta_{i0}}{1 - \theta_{i1}}\right)\right)}. \end{aligned}$$

Pentru a pune această ultimă expresie sub forma dorită, adică

$$P(Y = 1|X = x) = 1/(1 + \exp(w_0 + \sum_{i=1}^d w_i X_i)),$$

vom alege valorile parametrilor w_i în mod natural:

$$w_0 = \ln \frac{1 - \pi}{\pi} + \sum_{i=1}^d \ln \frac{1 - \theta_{i0}}{1 - \theta_{i1}} \quad \text{și} \quad w_i = \ln \frac{\theta_{i0}}{\theta_{i1}} - \ln \frac{1 - \theta_{i0}}{1 - \theta_{i1}} \text{ pentru } i = 1, \dots, d.$$

b. Vom începe ca și la punctul precedent, prin a pune probabilitatea $P(Y = 1|X = x)$ sub o formă apropiată de cea a funcției sigmoidale:⁴²⁴

$$\begin{aligned} P(Y = 1|X) &\stackrel{FB}{=} \frac{P(X|Y = 1)P(Y = 1)}{P(X|Y = 1)P(Y = 1) + P(X|Y = 0)P(Y = 0)} \\ &= \frac{1}{1 + \frac{P(X|Y = 0)P(Y = 0)}{P(X|Y = 1)P(Y = 1)}} = \frac{1}{1 + \exp\left(\ln \frac{P(X|Y = 0)P(Y = 0)}{P(X|Y = 1)P(Y = 1)}\right)}. \end{aligned}$$

Până aici nu avem încă nicio diferență în raport cu calculul de la punctul a. Însă acum vom ține cont că toate variabilele X_i ci $i = 1, \dots, d$ sunt independente

⁴²³Cu excepția cazurilor când $P(X = x_i|Y = 0) = 0$ sau $P(X = x_i|Y = 1) = 0$ pentru $i = 1, \dots, d$.

⁴²⁴Cu excepția cazurilor când $P(X|Y = 1)P(Y = 1) = 0$ sau $P(X|Y = 0)P(Y = 0) = 0$.

condițional două câte două în raport cu Y , cu excepția perechii X_1, X_2 :⁴²⁵

$$\begin{aligned} P(Y = 1|X) &= \frac{1}{1 + \exp \left(\ln \frac{P(X_1, X_2|Y = 0)}{P(X_1, X_2|Y = 1)} \frac{\prod_{i=3}^d P(X_i|Y = 0)P(Y = 0)}{\prod_{i=3}^d P(X_i|Y = 1)P(Y = 1)} \right)} \\ &= \frac{1}{1 + \exp \left(\ln \frac{1 - \pi}{\pi} + \sum_{i=3}^d \ln \frac{P(X_i|Y = 0)}{P(X_i|Y = 1)} + \ln \frac{P(X_1, X_2|Y = 0)}{P(X_1, X_2|Y = 1)} \right)}. \end{aligned}$$

Ca și la punctul a , vom ține cont că

$$\ln \frac{P(X_i|Y = 0)}{P(X_i|Y = 1)} = \ln \frac{\theta_{i0}^{X_i}(1 - \theta_{i0})^{(1-X_i)}}{\theta_{i1}^{X_i}(1 - \theta_{i1})^{(1-X_i)}}$$

atunci când condițiile asociate cu relația (213) sunt satisfăcute. Mai departe, folosind *indicația 4*, vom putea înlocui $P(X_1, X_2|Y = 0)$ și $P(X_1, X_2|Y = 1)$ în funcție de β_{ijk} , obținând (atunci când condițiile asociate cu relația (214) sunt satisfăcute):

$$\ln \frac{P(X_1, X_2|Y = 0)}{P(X_1, X_2|Y = 1)} = \ln \frac{(\beta_{110})^{X_1 X_2} (\beta_{100})^{X_1(1-X_2)} (\beta_{010})^{(1-X_1)X_2} (\beta_{000})^{(1-X_1)(1-X_2)}}{(\beta_{111})^{X_1 X_2} (\beta_{101})^{X_1(1-X_2)} (\beta_{011})^{(1-X_1)X_2} (\beta_{001})^{(1-X_1)(1-X_2)}}.$$

Așadar, va rezulta:

$$P(Y = 1|X) = \frac{1}{1 + \exp \left(w_0 + \sum_{i=3}^d w_i X_i + w_1 X_1 + w_2 X_2 + w_{1,2} X_1 X_2 \right)},$$

unde

$$\begin{aligned} w_0 &= \ln \frac{1 - \pi}{\pi} + \sum_{i=3}^d \ln \frac{1 - \theta_{i1}}{1 - \theta_{i0}} + \ln \frac{\beta_{000}}{\beta_{001}} \\ w_1 &= \ln \frac{\beta_{100}}{\beta_{101}} + \ln \frac{\beta_{001}}{\beta_{000}} \\ w_2 &= \ln \frac{\beta_{010}}{\beta_{011}} + \ln \frac{\beta_{001}}{\beta_{000}} \\ w_{1,2} &= \ln \frac{\beta_{110}}{\beta_{111}} + \ln \frac{\beta_{101}}{\beta_{100}} + \ln \frac{\beta_{011}}{\beta_{010}} + \ln \frac{\beta_{000}}{\beta_{001}} \\ w_i &= \ln \frac{\theta_{i0}}{\theta_{i1}} + \ln \frac{1 - \theta_{i1}}{1 - \theta_{i0}} \text{ pentru } i = 3, \dots, d. \end{aligned}$$

2.1.3 Clasificare bayesiană [cu atrbute de intrare] de tip gaussian

15. (Algoritmul Bayes [Naiv] gaussian: aplicare pe date din \mathbb{R})

⁴²⁵Cu excepția cazurilor când $P(Y = 0) = 0$ sau $P(Y = 1) = 0$, respectiv $P(X_1 X_2|Y = 0) = 0$ sau $P(X_1 X_2|Y = 1) = 0$ și încă $P(X_i|Y = 0) = 0$ sau $P(X_i|Y = 1) = 0$ pentru $i = 3, \dots, d$.

*prelucrare de Liviu Ciortuz, după
■ • CMU, 2001 fall, Andrew Moore, midterm, pr. 3.a*

X	Y
0	A
2	A
3	B
4	B
5	B
6	B
7	B

Presupunem că dispunem de setul de date de antrenament din tabelul alăturat; singurul atribut de intrare (X) ia valori reale, iar atributul de ieșire (Y) este de tip Bernoulli, deci ia două valori, notate cu A și respectiv B .

a. Pornind de la acest set de date, va trebui mai întâi să învățați parametrii clasificatorului Bayes gaussian, prin metoda estimării de verosimilitate maximă (MLE).⁴²⁶ Centralizați rezultatele, completând tabelul următor:

$\mu_A =$	$\sigma_A^2 =$	$P(Y = A) =$
$\mu_B =$	$\sigma_B^2 =$	$P(Y = B) =$

b. Notăm $\alpha = p(X = 2|Y = A)$ și $\beta = p(X = 2|Y = B)$.

- Cât este $p(X = 2, Y = A)$ în funcție de α ?
- Cât este $p(X = 2, Y = B)$ în funcție de β ?
- Cât este $p(X = 2)$ în funcție de α și β ?
- Cât este $p(Y = A|X = 2)$ în funcție de α și β ?

c. Cum va clasifica algoritmul Bayes [Naiv] gaussian punctul $X = 2$? Puteți exprima răspunsul fie în funcție de α și β , fie — mai bine! — calculând în prealabil valorile lui α și β în funcție de parametrii calculați / estimăți la punctul precedent.

Răspuns:

a. Pentru a estima mediile μ_A și μ_B , vom folosi formula demonstrată la problema 50.a de la capitolul de *Fundamente*:

$$\mu_{MLE} = \frac{\sum_{i=1}^n x_i}{n},$$

unde n este numărul instanțelor de antrenament. Așadar, $\mu_A = \frac{\sum_{i=1}^2 X_i}{2} = \frac{0+2}{2} = 1$, iar $\mu_B = \frac{\sum_{i=3}^7 X_i}{5} = \frac{3+4+5+6+7}{5} = 5$.

Similar, pentru calculul varianțelor σ_A^2 și σ_B^2 , vom folosi formula care a fost demonstrată la problema 51.a de la capitolul de *Fundamente*:

$$\sigma_{MLE}^2 = \frac{\sum_{i=1}^n (x_i - \mu_{MLE})^2}{n}.$$

⁴²⁶Vedeți secțiunea corespunzătoare din capitolul de *Fundamente*, în speță (pentru acest caz) problemele 50.a și 51.a.

Așadar, $\sigma_A^2 = \frac{1}{2}[(0-1)^2 + (2-1)^2] = 1$, iar $\sigma_B^2 = \frac{1}{5}[(3-5)^2 + (4-5)^2 + 0^2 + (6-5)^2 + (7-5)^2] = \frac{1}{5} \cdot 2 \cdot [4+1] = 2$.

Pentru calculul probabilităților $P(Y = A)$ și $P(Y = B)$ se folosește formula clasică (numărul de cazuri favorabile împărțit la numărul de cazuri posibile; vedeti fie problema 43.c fie problema 124.a, ambele de la capitolul de *Fundamente*), fiindcă Y este variabilă de tip Bernoulli. Așadar, $P(Y = A) = 2/7$ și $P(Y = B) = 5/7$.

Centralizând aceste estimări, obținem:

$\mu_A = 1$	$\sigma_A^2 = 1$	$P(Y = A) = 2/7$
$\mu_B = 5$	$\sigma_B^2 = 2$	$P(Y = B) = 5/7$

b. Folosind regula de înmulțire a probabilităților, calculăm $p(X = 2, Y = A) = p(X = 2|Y = A) \cdot P(Y = A) = \frac{2\alpha}{7}$ și $p(X = 2, Y = B) = p(X = 2|Y = B) \cdot P(Y = B) = \frac{5\beta}{7}$.

Probabilitatea $p(X = 2)$ se poate obține aplicând formula probabilității totale: $p(X = 2) = p(X = 2|Y = A) \cdot P(Y = A) + p(X = 2|Y = B) \cdot P(Y = B) = \frac{1}{7}(2\alpha + 5\beta)$.

Probabilitatea condiționată $p(Y = A|X = 2)$ se calculează folosind definiția: $p(Y = A|X = 2) = \frac{p(Y = A, X = 2)}{p(X = 2)} = \frac{2\alpha}{2\alpha + 5\beta}$.

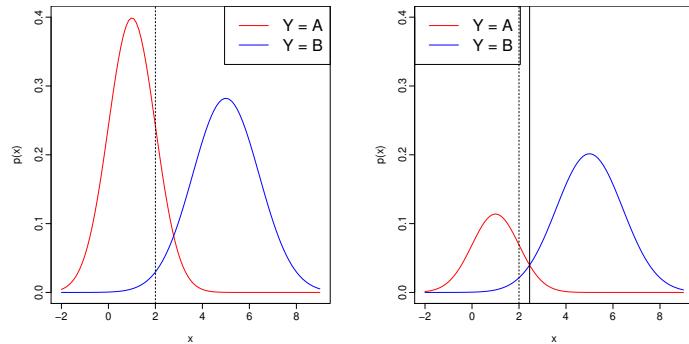
c. Algoritmul Bayes [Naiv] gaussian va asocia punctului $X = 2$ eticheta $Y = A$ dacă $p(Y = A|X = 2) \geq p(Y = B|X = 2) \Leftrightarrow \frac{2}{7}\alpha \geq \frac{5}{7}\beta$.

Folosind valorile estimate pentru parametrii μ_A , μ_B , σ_A și σ_B la punctul a , vom putea scrie: $\alpha = \frac{1}{\sqrt{2\pi}} e^{-\frac{(2-1)^2}{2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}} = 0.24197$ și $\beta = \frac{1}{\sqrt{2\pi} \cdot \sqrt{2}} e^{-\frac{(2-5)^2}{2 \cdot 2}} = \frac{1}{2\sqrt{\pi}} e^{-\frac{9}{4}} = 0.029732$. Deci,

$$\frac{2}{7}\alpha \geq \frac{5}{7}\beta \Leftrightarrow \frac{2}{7} \cdot 0.24197 \geq \frac{5}{7} \cdot 0.029732 \Leftrightarrow 0.06913 \geq 0.02123 \text{ (adevărat).}$$

Prin urmare, algoritmul Bayes [Naiv] gaussian va asocia punctului $X = 2$ eticheta $Y = A$, cu probabilitatea $\frac{0.06913}{0.06913 + 0.02123} = 0.76499$.

Observație: În cele două grafice alăturate am reprezentat $\mathcal{N}(\mu_A, \sigma_A^2)$ și $\mathcal{N}(\mu_B, \sigma_B^2)$, p.d.f.-urile gaussieneelor asociate claselor A și respectiv B (în partea stângă), precum și funcțiile obținute din aceste două p.d.f.-uri prin înmulțirea cu factorii de selecție $2/7$ și respectiv $5/7$ (în partea dreaptă).



Se poate constata relativ ușor că există două puncte de intersecție ($x_1 = -8.451$ și $x_2 = 2.451$) pentru graficele funcțiilor $\frac{2}{7}\mathcal{N}(\mu_A, \sigma_A^2)$ și $\frac{5}{7}\mathcal{N}(\mu_B, \sigma_B^2)$. Toate instanțele de test x situate între aceste puncte de intersecție ($x_1 < x < x_2$) vor apartine clasei A (acolo curba roșie este situată deasupra celei albastre). Instanțele situate fie la stânga lui x_1 fie la dreapta lui x_2 vor apartine clasei B (acolo curba albastră este situată deasupra celei roșii). *Separatorul decizional* este de tip pătratic, fiind constituit din punctele x_1 și x_2 .⁴²⁷

⁴²⁷LC: Mulțumesc studentului MSc Sergiu Dinu pentru această observație.

16.

(Clasificatorul Bayes [Naiv] gaussian,
cazul când se folosește un singur atribut de intrare:
zone de decizie și granițe de decizie;
analiza diferitelor *cazuri specifice*)

Sergiu Dinu, Liviu Ciortuz, 2020

În acest exercițiu ne propunem să identificăm *zonele de decizie și granițele de decizie* determinate de clasificatorul Bayes [Naiv] gaussian (abreviat *GB*), atunci când folosim un singur atribut de intrare X , iar atributul de ieșire, pe care îl vom nota cu Y , este binar, deci poate lua două valori, desemnate în continuare cu A și B . Vom desemna prin $p_A = p$ probabilitatea de selecție pentru clasa A . Prin urmare, probabilitatea de selecție pentru clasa B este $p_B = 1 - p$. Vom presupune că $p \in (0, 1)$ și $X|Y = A \sim \mathcal{N}(x|\mu_A, \sigma_A^2)$, iar $X|Y = B \sim \mathcal{N}(x|\mu_B, \sigma_B^2)$.

a. Arătați că *regula de decizie* a acestui clasificator Bayes [Naiv] gaussian

$$\hat{Y}_{GB}(X = x) = A \Leftrightarrow p \cdot \mathcal{N}(x|\mu_A, \sigma_A^2) \geq (1-p) \cdot \mathcal{N}(x|\mu_B, \sigma_B^2) \quad (215)$$

devine echivalentă cu

$$(\sigma_A^2 - \sigma_B^2)(x - x_1)(x - x_2) \geq 0,$$

unde

$$x_1 = \frac{\sigma_A^2 \mu_B - \sigma_B^2 \mu_A - \sqrt{\Delta'}}{\sigma_A^2 - \sigma_B^2} \quad \text{și} \quad x_2 = \frac{\sigma_A^2 \mu_B - \sigma_B^2 \mu_A + \sqrt{\Delta'}}{\sigma_A^2 - \sigma_B^2},$$

cu

$$\Delta' \stackrel{not.}{=} \sigma_A^2 \sigma_B^2 \left[(\mu_A - \mu_B)^2 + (\sigma_A^2 - \sigma_B^2) \ln \left(\frac{1-p}{p} \cdot \frac{\sigma_A}{\sigma_B} \right)^2 \right],$$

în *condițiile* în care $\sigma_A^2 \neq \sigma_B^2$ și $\Delta' \geq 0$.

b. Arătați că atunci când $\sigma_A^2 = \sigma_B^2 \stackrel{not.}{=} \sigma^2$ și $\mu_A \neq \mu_B$, regula de decizie (215) este echivalentă cu

$$x \geq x_0 \text{ dacă } \mu_A > \mu_B$$

și, respectiv

$$x \leq x_0 \text{ dacă } \mu_A < \mu_B,$$

unde

$$x_0 = \frac{\mu_A + \mu_B}{2} + \frac{\sigma^2}{\mu_A - \mu_B} \cdot \ln \frac{1-p}{p}.$$

c. Arătați că este posibil ca în anumite cazuri să avem $\Delta' < 0$, adică inegalitatea (215) să fie ori adevărată pentru orice $x \in \mathbb{R}$, ori falsă pentru orice $x \in \mathbb{R}$. Altfel spus, arătați că pot exista combinații de valori pentru parametrii σ_A și σ_B (ambii strict pozitivi), μ_A și μ_B din \mathbb{R} , precum și $p \in (0, 1)$, astfel încât

$$(\mu_A - \mu_B)^2 + (\sigma_A^2 - \sigma_B^2) \ln \left(\frac{1-p}{p} \cdot \frac{\sigma_A}{\sigma_B} \right)^2 < 0. \quad (216)$$

Răspuns:

a. Putem scrie următoarea succesiune de echivalențe:

$$\begin{aligned}
 p \cdot \mathcal{N}(x|\mu_A, \sigma_A^2) \geq (1-p) \cdot \mathcal{N}(\mu_B, \sigma_B^2) &\Leftrightarrow \\
 p \cdot \frac{1}{\sqrt{2\pi}\sigma_A} \exp\left(-\frac{(x-\mu_A)^2}{2\sigma_A^2}\right) \geq (1-p) \cdot \frac{1}{\sqrt{2\pi}\sigma_B} \exp\left(-\frac{(x-\mu_B)^2}{2\sigma_B^2}\right) &\Leftrightarrow \\
 \exp\left(\frac{1}{2} \left[\left(\frac{x-\mu_B}{\sigma_B}\right)^2 - \left(\frac{x-\mu_A}{\sigma_A}\right)^2 \right]\right) \geq \frac{1-p}{p} \cdot \frac{\sigma_A}{\sigma_B} &\Leftrightarrow \\
 \left(\frac{x-\mu_B}{\sigma_B}\right)^2 - \left(\frac{x-\mu_A}{\sigma_A}\right)^2 \geq \ln\left(\frac{1-p}{p} \cdot \frac{\sigma_A}{\sigma_B}\right)^2 &\Leftrightarrow \\
 (\sigma_A^2 - \sigma_B^2)x^2 + 2(\sigma_B^2\mu_A - \sigma_A^2\mu_B)x + \sigma_A^2\mu_B^2 - \sigma_B^2\mu_A^2 - \sigma_A^2\sigma_B^2 \ln\left(\frac{1-p}{p} \cdot \frac{\sigma_A}{\sigma_B}\right)^2 \geq 0. & \\
 \end{aligned} \tag{217}$$

Membrul stâng al acestei ultime inegalități este o funcție polinomială de gradul al doilea. Prin urmare, discriminantul ei „pe jumătate“ (adică, $\Delta' \stackrel{\text{not.}}{=} \Delta/2$) este:

$$\Delta' = (\sigma_B^2\mu_A - \sigma_A^2\mu_B)^2 - (\sigma_A^2 - \sigma_B^2) \left[\sigma_A^2\mu_B^2 - \sigma_B^2\mu_A^2 - \sigma_A^2\sigma_B^2 \ln\left(\frac{1-p}{p} \cdot \frac{\sigma_A}{\sigma_B}\right)^2 \right]$$

Însă,

$$\begin{aligned}
 &(\sigma_B^2\mu_A - \sigma_A^2\mu_B)^2 - (\sigma_A^2 - \sigma_B^2)(\sigma_A^2\mu_B^2 - \sigma_B^2\mu_A^2) \\
 &= \cancel{\sigma_B^4\mu_A^2} + \cancel{\sigma_A^4\mu_B^2} - 2\sigma_A^2\sigma_B^2\mu_A\mu_B - \cancel{\sigma_A^4\mu_B^2} + \sigma_A^2\sigma_B^2\mu_A^2 + \sigma_A^2\sigma_B^2\mu_B^2 - \cancel{\sigma_B^4\mu_A^2} \\
 &= \sigma_A^2\sigma_B^2(\mu_A - \mu_B)^2.
 \end{aligned}$$

Așadar,

$$\Delta' = \sigma_A^2\sigma_B^2 \left[(\mu_A - \mu_B)^2 + (\sigma_A^2 - \sigma_B^2) \ln\left(\frac{1-p}{p} \cdot \frac{\sigma_A}{\sigma_B}\right)^2 \right],$$

ceea ce ne conduce imediat la concluzia din enunț.

Consecință: În acest caz (a), separatorul decizional este format din punctele x_1 și x_2 . Dacă $\sigma_A^2 > \sigma_B^2$, atunci zona de decizie corespunzătoare clasei A este $(-\infty, x_1] \cup [x_2, +\infty)$, iar zona de decizie corespunzătoare clasei B este intervalul (x_1, x_2) . Similar, dacă $\sigma_A^2 < \sigma_B^2$, atunci zona de decizie corespunzătoare clasei A este (x_1, x_2) , iar zona de decizie corespunzătoare clasei B este intervalul $(-\infty, x_1] \cup [x_2, +\infty)$.

b. Dacă $\sigma_A^2 = \sigma_B^2 \stackrel{\text{not.}}{=} \sigma^2$, atunci este ușor de observat că inegalitatea (217) devine

$$2(\mu_A - \mu_B)x + \mu_B^2 - \mu_A^2 - \sigma^2 \ln\left(\frac{1-p}{p}\right)^2 \geq 0,$$

ceea ce fundamentează concluzia din enunț.

Consecință: În acest caz (b), separatorul decizional este punctul x_0 . Zonele de decizie se stabilesc imediat conform relațiilor din enunț.

c. Observăm că dacă vom considera cazul particular $\mu_A = \mu_B$, relația (216) va căpăta o formă mai simplă:

$$(\sigma_A^2 - \sigma_B^2) \ln\left(\frac{1-p}{p} \cdot \frac{\sigma_A}{\sigma_B}\right) < 0. \tag{218}$$

Această inegalitate este satisfăcută, de pildă atunci când σ_A și σ_B (care se consideră, prin convenție, strict pozitive) sunt în relația $\sigma_A > \sigma_B$, iar $\frac{1-p}{p} < \frac{\sigma_B}{\sigma_A}$. Se poate constata imediat că a două inegalitate este echivalentă cu $p > \frac{\sigma_A - \sigma_B}{\sigma_A + \sigma_B}$. Se observă că această ultimă fracție are valori în intervalul $(1/2, 1)$, fiindcă $\sigma_A > \sigma_B$.

Similar, dacă $\sigma_A < \sigma_B$, atunci pentru a satisface inegalitatea (218) se impune condiția $\frac{1-p}{p} > \frac{\sigma_B}{\sigma_A}$, care revine la $p < \frac{\sigma_A}{\sigma_A + \sigma_B}$. Această ultimă fracție ia valori în intervalul $(0, 1/2)$, întrucât $\sigma_A < \sigma_B$.

Consecință: În acest caz (c), nu există separator decizional. Una din cele două zone de decizie este vidă, iar cealaltă zonă coincide cu mulțimea tuturor numerelor reale, \mathbb{R} .

17.

(Algoritmul Bayes Naiv gaussian:
deducerea [formeи liniare a] regulii de decizie în cazul
matricelor de covarianță diagonale și identice,
i.e., $\sigma_{i0} = \sigma_{i1}$, pentru $i = 1, \dots, d$)

■ • ○ CMU, 2009 spring, Ziv Bar-Joseph, HW2, pr. 2

Considerăm un model de tip Bayes Naiv cu două clase ($Y \in \{0, 1\}$), definit peste spațiul real \mathbb{R}^d al atributelor de intrare X_1, \dots, X_d . Presupunem că în acest model distribuția [comună] condiționată $X|Y = 0$, unde $X = (X_1, \dots, X_d) \in \mathbb{R}^d$, poate fi definită ca un vector de distribuții gaussiene unidimensionale independente și-l vom desemna prin notația

$$\text{Gaussian}(\mu_0 = \langle \mu_{10}, \dots, \mu_{d0} \rangle, \sigma = \langle \sigma_1, \dots, \sigma_d \rangle)$$

și analog pentru $X|Y = 1$:

$$\text{Gaussian}(\mu_1 = \langle \mu_{11}, \dots, \mu_{d1} \rangle, \sigma = \langle \sigma_1, \dots, \sigma_d \rangle).$$

Observați că intrările X_1, \dots, X_d au — la condiționare în raport cu clasa — medii diferite dar varianțe (de fapt, matrice de covarianță, diagonale) identice pentru ambele clase.

În acest exercițiu vă vom arăta că, în modelul specificat mai sus, probabilitatea condiționată $P(Y = 1|X = x)$, unde $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, se poate scrie ca valoare a unei funcții sigmoidale / „logistice“, $f(x) = \frac{1}{1 + e^{-(w_0 + w \cdot x)}}$, cu parametrii $w_0 \in \mathbb{R}$ și $w = (w_1, \dots, w_d) \in \mathbb{R}^d$ aleși în mod convenabil.⁴²⁸

a. Folosiți regula lui Bayes (compusă cu formula probabilității totale) pentru a scrie $P(Y = 1|X = x)$ sub forma unei fracții. Împărțiți atât numărătorul cât și numitorul fracției astfel obținute cu expresia de la numărător.

b. La punctul a ar fi trebuit să ajungeți la un rezultat de forma

$$\frac{1}{1 + f(x, X, Y)},$$

⁴²⁸În consecință, [se poate arăta imediat că] regula de decizie a clasificatorului Bayes Naiv pentru acest model este de tip *liniar*.

unde f este o anumită funcție având argumentele x, X și Y . Folosind formula $e^{\ln a} = a$ (valabilă pentru orice $a > 0$), putem forța punerea acestei fracții într-o formă apropiată de funcția sigmoidală:

$$\frac{1}{1 + e^{\ln f(x, X, Y)}}.$$

Vă cerem să rescrieți sub o astfel de formă rezultatul de la punctul a .

c. Explicați presupozitia Bayes „naivă“ în cadrul modelului probabilist dat. Apoi folosiți-o pentru a converti exponentul care apare în fracția rezultată la punctul b la o sumă de forma

$$\ln g(Y) + \sum_{i=1}^d \ln h(x_i, X_i, Y).$$

Precizați expresiile funcțiilor g și h .

d. Acum folosiți specificul modelului dat — și anume, de tip gaussian, având pentru componentele condiționate (și anume, variabilele aleatoare condiționate $X_i|Y = 1$, respectiv $X_i|Y = 0$, pentru $i = 1, \dots, n$) medii diferite dar varianțe egale —, pentru a aduce expresia de la punctul c la o formă mai convenabilă.

e. Rescriind $P(Y = 1|X = x)$ conform expresiilor obținute la punctele b și d , rezultatul ar trebui să semene cu un alt model pe care l-am întâlnit la curs. Care anume? Exprimăți parametrii aceluia model în raport cu $P(Y = 1)$, μ_{i0} , μ_{i1} și σ_i , cu $i = 1, \dots, d$.

Răspuns:

a. Procedând conform cerințelor, vom scrie:

$$\begin{aligned} P(Y = 1|X = x) &= \frac{P(X = x|Y = 1)P(Y = 1)}{\sum_{y \in \{0,1\}} P(X = x|Y = y)P(Y = y)} \\ &= \frac{1}{1 + \frac{P(X = x|Y = 0)P(Y = 0)}{P(X = x|Y = 1)P(Y = 1)}}. \end{aligned}$$

Considerând $f(x, X, Y) = \frac{P(X = x|Y = 0)P(Y = 0)}{P(X = x|Y = 1)P(Y = 1)}$, se observă că ultima expresie obținută are într-adevăr formă $\frac{1}{1 + f(x, X, Y)}$.

b. Folosind formula $a = e^{\ln a}$, obținem:

$$P(Y = 1|X = x) = \frac{1}{1 + \exp\left(\ln \frac{P(X = x|Y = 0)P(Y = 0)}{P(X = x|Y = 1)P(Y = 1)}\right)}.$$

c. Conform presupozitiei specifice algoritmului Naive Bayes, vom considera că atrbutele X_i sunt independente condițional două câte două în raport cu variabila de ieșire. Așadar, vom avea egalitățile următoare:

$$\begin{aligned} P(X = x|Y = 1) &= \prod_{i=1}^d P(X_i = x_i|Y = 1) \\ P(X = x|Y = 0) &= \prod_{i=1}^d P(X_i = x_i|Y = 0) \end{aligned}$$

Prin urmare, vom scrie argumentul funcției $\exp(\cdot)$ din fracția de la punctul b astfel:

$$\ln \frac{P(X = x|Y = 0)P(Y = 0)}{P(X = x|Y = 1)P(Y = 1)} = \ln \frac{P(Y = 0)}{P(Y = 1)} + \sum_{i=1}^d \ln \frac{P(X_i = x_i|Y = 0)}{P(X_i = x_i|Y = 1)}.$$

În consecință, funcțiile g și h care au fost cerute în enunț vor fi:

$$g(Y) = \frac{P(Y = 0)}{P(Y = 1)} \quad h(x_i, X_i, Y) = \frac{P(X_i = x_i|Y = 0)}{P(X_i = x_i|Y = 1)} \text{ pentru } i = 1, \dots, d.$$

d. Tinând cont de specificul gaussian al modelului din enunț, vom putea scrie rezultatul de la punctul b astfel:

$$\begin{aligned} &\ln \frac{P(X = x|Y = 0)P(Y = 0)}{P(X = x|Y = 1)P(Y = 1)} \\ &= \ln \frac{P(Y = 0)}{P(Y = 1)} + \sum_{i=1}^d \ln \left(\frac{\frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x_i - \mu_{i0})^2}{2\sigma_i^2}\right)}{\frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x_i - \mu_{i1})^2}{2\sigma_i^2}\right)} \right) \\ &= \ln \frac{P(Y = 0)}{P(Y = 1)} + \sum_{i=1}^d \left(\frac{(x_i - \mu_{i1})^2}{2\sigma_i^2} - \frac{(x_i - \mu_{i0})^2}{2\sigma_i^2} \right) \\ &= \ln \frac{P(Y = 0)}{P(Y = 1)} + \sum_{i=1}^d \frac{2x_i(\mu_{i0} - \mu_{i1}) + (\mu_{i1}^2 - \mu_{i0}^2)}{2\sigma_i^2} \\ &= \ln \frac{P(Y = 0)}{P(Y = 1)} + \sum_{i=1}^d \left(\frac{x_i(\mu_{i0} - \mu_{i1})}{\sigma_i^2} + \frac{(\mu_{i1}^2 - \mu_{i0}^2)}{2\sigma_i^2} \right) \\ &= \ln \frac{P(Y = 0)}{P(Y = 1)} + \sum_{i=1}^d \frac{\mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2} + \sum_{i=1}^d \frac{\mu_{i0} - \mu_{i1}}{\sigma_i^2} x_i. \end{aligned}$$

e. Notând în expresia obținută la punctul d

$$w_0 = \ln \frac{P(Y = 0)}{P(Y = 1)} + \sum_{i=1}^d \frac{\mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2} \quad \text{și} \quad w_i = \frac{\mu_{i0} - \mu_{i1}}{\sigma_i^2} \text{ pentru } i = 1, \dots, d$$

și apoi revenind la expresia de la punctul b, vom putea scrie:

$$P(Y = 1|X = x) = \frac{1}{1 + \exp(w_0 + \sum_{i=1}^d w_i x_i)}.$$

Așadar, expresia probabilității a posteriori este exact cea a funcției sigmoidale, $\frac{1}{1 + e^{-x}}$. Rezultatul seamănă foarte bine cu modelul de *regresie logistică*.⁴²⁹

⁴²⁹Pentru o introducere la modelul regresiei logistice, vedeți problema 13 de la capitolul *Metode de regresie*.

18.

(Algoritmul Bayes Optimal gaussian:
raportul față de regresia logistică
în cazul $\Sigma_0 = \Sigma_1$)

■ □ • ○ CMU, 2011 spring, Tom Mitchell, HW2, pr. 2.2

În cele ce urmează, vom considera:

1. Y , o variabilă booleană care urmează o distribuție de tip Bernoulli, cu parametrul $\pi = P(Y = 1)$, ceea ce implică $P(Y = 0) = 1 - \pi$;
2. $X = (X_1, X_2, \dots, X_d)^\top$, un vector de variabile aleatoare care nu sunt independente condițional în raport cu variabila Y , probabilitatea [comună și] condiționată $P(X|Y = k)$ urmând o *distribuție gaussiană multidimensională*, $\mathcal{N}(\mu_k, \Sigma)$, unde $k \in \{0, 1\}$.

Rețineți faptul că μ_k , media acestei distribuții multidimensionale, este un vector-coloană (altfel spus, o matrice $d \times 1$) și există două astfel de medii, câte una pentru fiecare dintre cele două valori ale variabilei Y . De asemenea, Σ , matricea de covarianță are dimensiunea $d \times d$, iar ea nu depinde de valorile variabilei Y .⁴³⁰

În rezolvarea problemei, veți folosi funcția de densitate [de probabilitate] a distribuției gaussiene multidimensionale în notație matricială:⁴³¹

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1} (x - \mu)\right),$$

unde simbolul \top desemnează operația de transpunere a matricelor.

Răspundeți la următoarea întrebare:

Este oare distribuția $P(Y|X)$ corespunzătoare acestui clasificator de tip Bayes Optimal gaussian (nu naiv!) de aceeași formă cu cea a *regresiei logistice*?

Sugestie: Rafinați expresia distribuției probabiliste $P(Y|X)$.

Observație: La problema 17 am arătat că în cazul (particular!) în care intrările X_1, X_2, \dots, X_d sunt independente condițional în raport cu ieșirea Y — ceea ce, conform problemei 34 de la capitolul de *Fundamente*, este echivalent cu a spune că matricea Σ este diagonală —, răspunsul la întrebarea pusă în enunț este pozitiv.

Răspuns:

Vom demara calculele în maniera standard (adică, similar cu prima parte a rezolvării problemelor 14 și 17):

$$\begin{aligned} P(Y = 1|X) &= \frac{P(X|Y = 1) P(Y = 1)}{P(X|Y = 1) P(Y = 1) + P(X|Y = 0) P(Y = 0)} \\ &= \frac{1}{1 + \frac{P(Y = 0) P(X|Y = 0)}{P(Y = 1) P(X|Y = 1)}} = \frac{1}{1 + \exp\left(\ln \frac{P(Y = 0) P(X|Y = 0)}{P(Y = 1) P(X|Y = 1)}\right)} \\ &= \frac{1}{1 + \exp\left(\ln \frac{P(Y = 0)}{P(Y = 1)} + \ln \frac{P(X|Y = 0)}{P(X|Y = 1)}\right)}. \end{aligned}$$

⁴³⁰ Altfel spus, presupunând că Σ_k , pentru $k \in \{0, 1\}$, desemnează matricea de covarianță a distribuției gaussiene multidimensionale $\mathcal{N}(\mu_k, \Sigma_k)$, atunci considerăm că $\Sigma_0 = \Sigma_1$.

⁴³¹ Vedeți problema 37 de la capitolul de *Fundamente*.

Acum ne vom concentra atenția asupra termenului $\ln \frac{P(X|Y=0)}{P(X|Y=1)}$, ținând cont de faptul că $X|Y=0 \sim \mathcal{N}(\mu_0, \Sigma)$ și $X|Y=1 \sim \mathcal{N}(\mu_1, \Sigma)$:

$$\begin{aligned} & \ln \frac{P(X|Y=0)}{P(X|Y=1)} \\ &= \ln \frac{\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}}{\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}} + \ln \exp \left(\frac{1}{2} [(X - \mu_1)^\top \Sigma^{-1} (X - \mu_1) - (X - \mu_0)^\top \Sigma^{-1} (X - \mu_0)] \right) \\ &= \frac{1}{2} [(X - \mu_1)^\top \Sigma^{-1} (X - \mu_1) - (X - \mu_0)^\top \Sigma^{-1} (X - \mu_0)] \\ &= \frac{1}{2} [-X^\top \Sigma^{-1} \mu_1 - \mu_1^\top \Sigma^{-1} X + \mu_1^\top \Sigma^{-1} \mu_1 + X^\top \Sigma^{-1} \mu_0 + \mu_0^\top \Sigma^{-1} X - \mu_0^\top \Sigma^{-1} \mu_0] \\ &= \frac{1}{2} [\mu_1^\top \Sigma^{-1} \mu_1 - \mu_0^\top \Sigma^{-1} \mu_0 + X^\top \Sigma^{-1} (\mu_0 - \mu_1) + (\mu_0^\top - \mu_1^\top) \Sigma^{-1} X] \\ &= \frac{1}{2} \mu_1^\top \Sigma^{-1} \mu_1 - \frac{1}{2} \mu_0^\top \Sigma^{-1} \mu_0 + (\mu_0 - \mu_1)^\top \Sigma^{-1} X. \end{aligned}$$

Remarcați faptul că $((\mu_0^\top - \mu_1^\top) \Sigma^{-1} X)^\top = ((\mu_0 - \mu_1)^\top \Sigma^{-1} X)^\top = X^\top (\Sigma^{-1})^\top (\mu_0 - \mu_1) = X^\top (\Sigma^\top)^{-1} (\mu_0 - \mu_1) = X^\top \Sigma^{-1} (\mu_0 - \mu_1)$, întrucât matricea Σ^{-1} este simetrică (ca urmare a faptului că Σ însăși, ca matrice de covarianță, este simetrică; vedeti problema 20 de la capitolul de *Fundamente*).

Prin urmare,

$$\begin{aligned} P(Y=1|X) &= \frac{1}{1 + \exp \left(\ln \frac{1-\pi}{\pi} + \frac{1}{2} \mu_1^\top \Sigma^{-1} \mu_1 - \frac{1}{2} \mu_0^\top \Sigma^{-1} \mu_0 + (\mu_0 - \mu_1)^\top \Sigma^{-1} X \right)} \\ &= \frac{1}{1 + \exp(w_0 + w^\top X)}, \end{aligned}$$

cu $w_0 = \ln \frac{1-\pi}{\pi} + \frac{1}{2} \mu_1^\top \Sigma^{-1} \mu_1 - \frac{1}{2} \mu_0^\top \Sigma^{-1} \mu_0$ și $w = \Sigma^{-1} (\mu_0 - \mu_1)$. Evident, w_0 este un număr real (constant), iar w un vector-coloană (mai precis, o matrice de dimensiune $d \times 1$).

În concluzie, distribuția probabilistă $P(Y|X)$ are (și în acest caz!) aceeași formă cu cea din modelul regresiei logistice.

19.

(Clasificatorul Bayes Optimal Gaussian:
natura separatorului decizional în cazul în care $\Sigma_0 \neq \Sigma_1$)

■ • ○ *Stanford, 2014 fall, Andrew Ng, midterm, pr. 2.b*

Fie setul de date de antrenament $\{(x_1, y_1), \dots, (x_n, y_n)\}$, cu x_i vectori-colonă din \mathbb{R}^d și $y_i \in \{0, 1\}$ pentru orice i . Algoritmul Bayes Optimal gaussian estimatează următorii parametri: $\phi \in (0, 1)$, vectorii-colonă μ_0 și μ_1 din \mathbb{R}^d , precum și matricele de covarianță Σ_0 și Σ_1 de dimensiune $d \times d$, corespunzători următoarelor distribuții:

$$\begin{aligned} p(y) &= \phi^y (1 - \phi)^{1-y}, \text{ unde } \phi = p(y=1) \\ p(x|y=0) &= \frac{1}{(2\pi)^{d/2} |\Sigma_0|^{1/2}} \exp \left(-\frac{1}{2} (x - \mu_0)^\top \Sigma_0^{-1} (x - \mu_0) \right) \end{aligned}$$

$$p(x|y=1) = \frac{1}{(2\pi)^{d/2}|\Sigma_1|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^\top \Sigma_1^{-1}(x-\mu_1)\right)$$

Vă readucem aminte *regula de decizie* a acestui clasificator:

$$y = 1 \text{ dacă } p(y=1|x) \geq p(y=0|x) \text{ și } y = 0 \text{ în caz contrar.} \quad (219)$$

Arătați că atunci când $\Sigma_0 \neq \Sigma_1$, separatorul decizional determinat de algoritmul Bayes Optimal gaussian este de ordin pătratic, adică inegalitatea dintre probabilitățile condiționate a posteriori, $p(y=1|x) \geq p(y=0|x)$, este echivalentă cu o inegalitate numerică de forma

$$x^\top Ax + B^\top x + C \geq 0, \quad (220)$$

unde A este o anumită matrice de dimensiune $d \times d$, cu $A \neq 0$ (matricea nulă de dimensiune $d \times d$), $B \in \mathbb{R}^d$ (un anumit vector-coloană) și $C \in \mathbb{R}$ (o anumită constantă). Veți specifica în mod clar valorile pentru A , B și C .

Răspuns:

Elaborând relația dintre probabilitățile care determină decizia luată de către clasificatorul Bayes Optimal gaussian pentru o instanță oarecare de test x (vedeți relația (219)), putem obține următoarele echivalențe:

$$\begin{aligned} p(y=1|x) \geq p(y=0|x) &\Leftrightarrow \ln p(y=1|x) \geq \ln p(y=0|x) \\ &\Leftrightarrow \ln p(y=1|x) - \ln p(y=0|x) \geq 0 \Leftrightarrow \ln \frac{p(y=1|x)}{p(y=0|x)} \geq 0 \\ F. \underset{\text{Bayes}}{\Leftrightarrow} \ln \frac{p(x|y=1)p(y=1)}{p(x|y=0)p(y=0)} &\geq 0 \Leftrightarrow \ln \frac{p(y=1)}{p(y=0)} + \ln \frac{p(x|y=1)}{p(x|y=0)} \geq 0 \\ &\Leftrightarrow \ln \frac{\phi}{1-\phi} - \ln \frac{|\Sigma_1|^{1/2}}{|\Sigma_0|^{1/2}} - \frac{1}{2} \left((x-\mu_1)^\top \Sigma_1^{-1}(x-\mu_1) - (x-\mu_0)^\top \Sigma_0^{-1}(x-\mu_0) \right) \geq 0 \\ &\Leftrightarrow -\frac{1}{2} \left(x^\top (\Sigma_1^{-1} - \Sigma_0^{-1})x - 2(\mu_1^\top \Sigma_1^{-1} - \mu_0^\top \Sigma_0^{-1})x + \mu_1^\top \Sigma_1^{-1}\mu_1 - \mu_0^\top \Sigma_0^{-1}\mu_0 \right) \\ &\quad + \ln \frac{\phi}{1-\phi} - \ln \frac{|\Sigma_1|^{1/2}}{|\Sigma_0|^{1/2}} \geq 0 \\ &\Leftrightarrow x^\top \left(\frac{1}{2} (\Sigma_0^{-1} - \Sigma_1^{-1}) \right) x + \left(\mu_1^\top \Sigma_1^{-1} - \mu_0^\top \Sigma_0^{-1} \right) x \\ &\quad + \ln \frac{\phi}{1-\phi} + \ln \frac{|\Sigma_0|^{1/2}}{|\Sigma_1|^{1/2}} + \frac{1}{2} \left(\mu_0^\top \Sigma_0^{-1}\mu_0 - \mu_1^\top \Sigma_1^{-1}\mu_1 \right) \geq 0. \end{aligned} \quad (221)$$

Coroborând rezultatul pe care tocmai l-am obținut cu relația (219) din enunț, observăm că putem considera

$$\begin{aligned} A &= \frac{1}{2} (\Sigma_0^{-1} - \Sigma_1^{-1}) \\ B^\top &= \mu_1^\top \Sigma_1^{-1} - \mu_0^\top \Sigma_0^{-1} \Leftrightarrow \\ B &= (\mu_1^\top \Sigma_1^{-1} - \mu_0^\top \Sigma_0^{-1})^\top = (\Sigma_1^{-1})^\top \mu_1 - (\Sigma_0^{-1})^\top \mu_0 = \Sigma_1^{-1}\mu_1 - \Sigma_0^{-1}\mu_0 \\ C &= \ln \frac{\phi}{1-\phi} + \ln \frac{|\Sigma_0|^{1/2}}{|\Sigma_1|^{1/2}} + \frac{1}{2} \left(\mu_0^\top \Sigma_0^{-1}\mu_0 - \mu_1^\top \Sigma_1^{-1}\mu_1 \right). \end{aligned} \quad (222)$$

Se poate constata imediat că ipoteza $\Sigma_0 \neq \Sigma_1$ din enunț implică faptul că $\Sigma_0^{-1} \neq \Sigma_1^{-1}$, deci $A \neq 0$. Prin urmare, granița de separare (engl., decision boundary) determinată în acest caz de către algoritmul Bayes Optimal gaussian este de ordin pătratic.

20.

(Clasificatorul Bayes Optimal gaussian:
aplicare pe date din \mathbb{R}^2 ;
raportul cu regresia logistică)

*prelucrare de Liviu Ciortuz, după
• MIT, 2001 fall, Tommi Jaakkola, HW2, pr. 2.abe*

a. Precizați separatorii decizionali determinați de algoritmul Bayes Optimal gaussian pentru fiecare din cazurile de mai jos. Veți menționa în mod explicit, în fiecare caz în parte, tipul — liniar, pătratic (cerc, elipsă, parabolă, hiperbolă etc.), etc. — și ecuația acestor separatori decizionali.

(i.)

$$P_0 = 0.5, P_1 = 0.5$$

$$\mu_0 = (1, 1)^\top$$

$$\mu_1 = (-1, -1)^\top$$

$$\Sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

(ii.)

$$P_0 = 0.005, P_1 = 0.995$$

$$\mu_0 = (1, 1)^\top$$

$$\mu_1 = (-1, -1)^\top$$

$$\Sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

(iii.)

$$P_0 = 0.5, P_1 = 0.5$$

$$\mu_0 = (1, 1)^\top$$

$$\mu_1 = (-1, -1)^\top$$

$$\Sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\Sigma_1 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$$

(iv.)

$$P_0 = 0.5, P_1 = 0.5$$

$$\mu_0 = (0, 0)^\top$$

$$\mu_1 = (0, 0)^\top$$

$$\Sigma_0 = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}$$

$$\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$$

(v.)

$$P_0 = 0.5, P_1 = 0.5$$

$$\mu_0 = (1, 1)^\top$$

$$\mu_1 = (1, 1)^\top$$

$$\Sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\Sigma_1 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

(vi.)

$$P_0 = 0.5, P_1 = 0.5$$

$$\mu_0 = (-2, 1)^\top$$

$$\mu_1 = (1, 1)^\top$$

$$\Sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1/2 \end{pmatrix}$$

$$\Sigma_1 = \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix}$$

Sugestie: În cazul general, ecuația separatorului decizional determinat de clasificatorul Bayes Optimal gaussian este $x^\top Ax + B^\top x + C = 0$ (vedeți realția (220) de la problema 19), unde valorile pentru A , B și C au fost determinate prin relațiile (222).

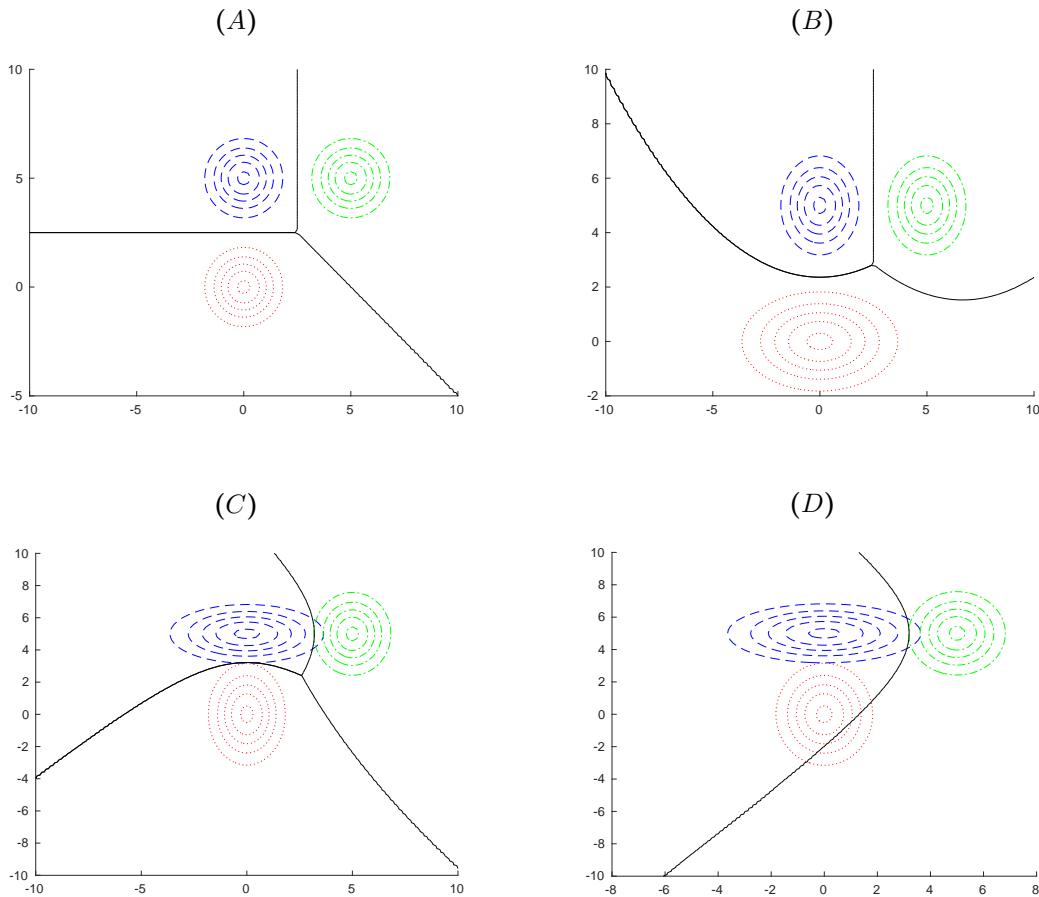
b. Arătați că este posibil ca întreg planul euclidian să fie alocat de către algoritmul Bayes Optimal gaussian⁴³² unei singure clase, situație în care, practic, nu avem separator decizional.

c. Care dintre *separatorii decizionali* de la punctul a poate corespunde și unui model de tip *regresie logistică*? Justificați răspunsul dumneavoastră.

d. Care dintre *zonele de decizie* din figurile de mai jos poate să corespundă unui model de tip *regresie logistică multinomială (softmax)*? Justificați răspunsul dumneavoastră.⁴³³

⁴³²LC: De fapt, este suficient să ne gândim la cazul $\Sigma_0 = \Sigma_1$. (Menționăm că acest caz particular de clasificare bayesiană gaussiană este numit în literatura de specialitate *analiză gaussiană discriminativă*.)

⁴³³Pentru o introducere în regresia logistică multinomială, vedeți problema 18 de la capitolul *Metode de regresie*.



Răspuns:

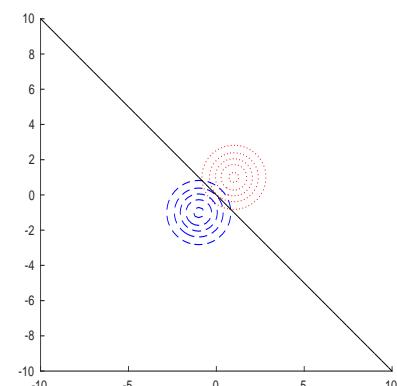
a. Folosind formulele (220), se calculează coeficienții A , B și C din ecuația $x^T A x + B^T x + C = 0$, pentru fiecare dintre cazurile date:⁴³⁴

i. $A = 0$ – matricea nulă de dimensiune 2×2 , $B^T = (2, 2)$, $C = 0$.

Ecuatia separatorului decizional este

$$2x_1 + 2x_2 = 0 \Leftrightarrow x_2 = -x_1.$$

Această ecuație reprezintă o dreaptă care trece prin originea sistemului de coordonate. Observați că această dreaptă este mediatoarea segmentului de dreaptă care unește punctele reprezentate de μ_1 și μ_2 (mediile celor două gaussiene).



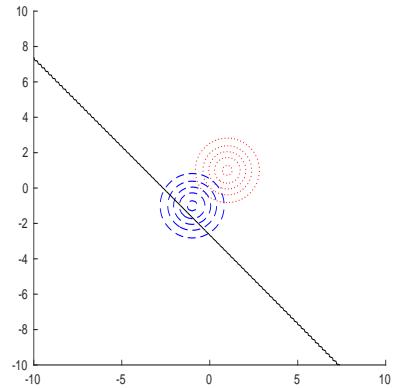
⁴³⁴ Am omis calculele efective, fiindcă nu sunt dificile. Invităm studentul să verifice că este capabil să obțină el însuși aceste rezultate.

ii. $A = 0$, $B^\top = (2, 2)$, $C = \ln 199$.

Ecuăția separatorului decizional este

$$2x_1 + 2x_2 + \ln 199 = 0 \Leftrightarrow x_2 = -x_1 - \underbrace{\ln \sqrt{199}}_{\approx 2.6466}.$$

Această ecuație reprezintă o dreaptă paralelă cu cea de la punctul precedent (i). Ea este situată acum mult mai aproape de punctul μ_1 — datorită faptului că în acest caz probabilitatea P_0 este mult mai mare decât probabilitatea P_1 .

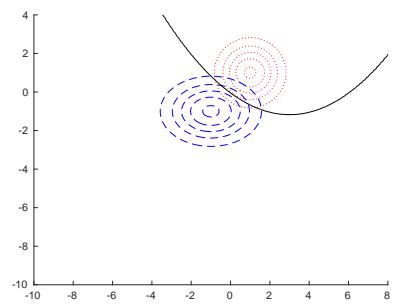


iii. $A = \begin{pmatrix} 1/4 & 0 \\ 0 & 0 \end{pmatrix}$, $B^\top = (-3/2, -2)$, $C = 1/4$.

În acest caz, spre deosebire de cazurile precedente, ecuația separatorului decizional este de ordin pătratic:

$$\frac{1}{4}x_1^2 - \frac{3}{2}x_1 - 2x_2 + \frac{1}{4} = 0 \Leftrightarrow x_2 = \frac{1}{8}x_1^2 - \frac{3}{4}x_1 + \frac{1}{8}.$$

Această ecuație reprezintă o parabolă, al cărei vârf are coordonatele $(3, -1)$.



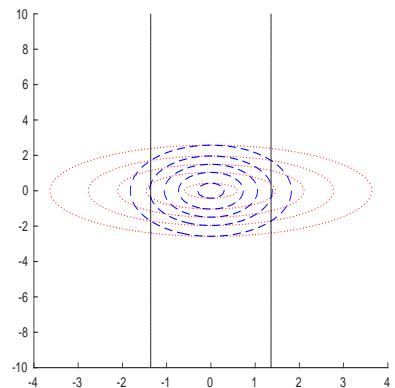
iv. $A = \begin{pmatrix} -3/8 & 0 \\ 0 & 0 \end{pmatrix}$, $B^\top = 0$ – vectorul nul din \mathbb{R}^2 , $C = \ln 2$.

Ecuăția separatorului decizional este

$$-\frac{3}{8}x_1^2 + \ln 2 = 0 \Leftrightarrow x_1^2 = \frac{8}{3} \ln 2.$$

Așadar, în acest caz, separatorul decizional este tot de ordin pătratic, dar este reprezentat de două drepte, având ecuațiile $x_1 = -\sqrt{\frac{8}{3} \ln 2} \approx -1.3595$ și respectiv $x_1 = \sqrt{\frac{8}{3} \ln 2} \approx 1.3595$.

Tinând cont de relația (221) care a fost demonstrată la problema 19, deducem că zona de decizie corespunzătoare clasei 1 este situată între aceste două drepte, iar zona de decizie corespunzătoare clasei 0 este în exteriorul lor.

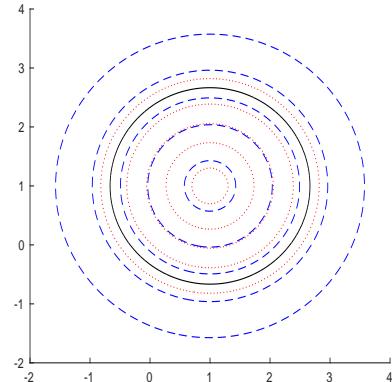


v. $A = \begin{pmatrix} 1/4 & 0 \\ 0 & 1/4 \end{pmatrix}$, $B^\top = (-1/2, -1/2)$, $C = -\ln 2 + 1/2$.

Ecuăția separatorului decizional este

$$\begin{aligned} \frac{1}{4}(x_1^2 + x_2^2) - \frac{1}{2}(x_1 + x_2) - \ln 2 + \frac{1}{2} = 0 &\Leftrightarrow \\ \frac{1}{4}(x_1^2 - 2x_1 + x_2^2 - 2x_2) = \ln 2 - \frac{1}{2} &\Leftrightarrow \\ (x_1 - 1)^2 + (x_2 - 1)^2 = 4\left(\ln 2 - \frac{1}{2}\right) + 2. & \end{aligned}$$

Această ecuație — tot de ordin pătratic — reprezintă un cerc cu centrul în punctul $(1, 1)$ și raza de $2\sqrt{\ln 2} \approx 1.3862$.

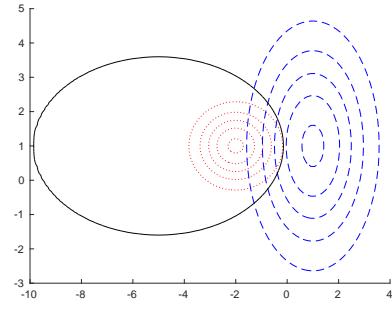


vi. $A = \begin{pmatrix} 1/4 & 0 \\ 0 & 7/8 \end{pmatrix}$, $B^\top = (5/2, -7/4)$, $C = -\ln(1/4) + 21/8$.

Ecuăția separatorului decizional este

$$\begin{aligned} \frac{1}{4}x_1^2 + \frac{7}{8}x_2^2 + \frac{5}{2}x_1 - \frac{7}{4}x_2 - 2\ln 2 + \frac{21}{8} = 0 &\Leftrightarrow \\ \frac{1}{4}(x_1^2 + 10x_1) + \frac{7}{8}(x_2^2 - 2x_2) = 2\ln 2 - \frac{21}{8} &\Leftrightarrow \\ \frac{(x_1 + 5)^2}{4} + \frac{(x_2 - 1)^2}{\frac{8}{7}} = 2\ln 2 - \frac{21}{8} + \frac{25}{4} + \frac{7}{8}. & \end{aligned}$$

Această ecuație — tot de ordin pătratic — reprezintă o elipsă care are centrul de simetrie în punctul $(-5, 1)$.



b. Pornind de la regula de decizie $y_{GNB} = \text{argmax}_{y \in \{0,1\}} P(Y = y|X = x)$, putem raționa astfel:

$$\begin{aligned} P(Y = 1|X = x) \geq P(Y = 0|X = x) &\stackrel{F. Bayes}{\Leftrightarrow} \\ P(X = x|Y = 1)P(Y = 1) \geq P(X = x|Y = 0)P(Y = 0). & \end{aligned}$$

Dacă alegem ca distribuțiile gaussiene corespunzătoare claselor $Y = 1$ și $Y = 0$ să fie identice — adică să aibă aceiași parametri μ și Σ —, în inegalitatea de mai sus factorii $P(X = x|Y = 1)$ și $P(X = x|Y = 0)$ vor fi egali. Prin urmare, atunci când $P(Y = 1) \neq P(Y = 0)$, va rezulta că întreg planul / spațiul va fi alocat unei singure clase, și anume acelei clase $y \in \{0, 1\}$ care are probabilitatea a priori ($P(Y = y)$) mai mare.

c. Se știe că regresia logistică determină separatori decizionali liniari.⁴³⁵ Prin urmare, doar separatorii din cazurile i și ii pot corespunde (și) unor modele de regresie logistică.

d. Justificarea este similară cu cea de la punctul c: doar în cazul (A) separatorii decizionali pot corespunde (și) unui model de regresie logistică multinomială / softmax.

⁴³⁵Vedeți problema 13 de la capitolul *Metode de regresie*, în special relațiile (177), din care rezultă:

$$P(Y = 1|X = x) \geq P(Y = 0|X = x) \Leftrightarrow \sigma(z) \geq 1/2 \Leftrightarrow z \geq 0,$$

unde $\sigma(z) \stackrel{\text{def.}}{=} \frac{1}{1+e^{-z}}$, $z \stackrel{\text{not.}}{=} w_0 + \sum_{i=1}^d w_i x_i$ și $x \stackrel{\text{not.}}{=} (x_1, \dots, x_d)$. Așadar, în cazul regresiei logistice separatorul decizional are ecuația de forma $w_0 + \sum_{i=1}^d w_i x_i = 0$, deci este de tip liniar.

21.

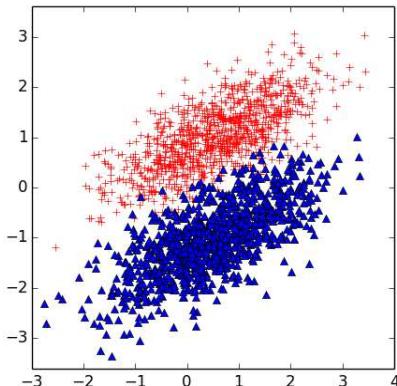
(Clasificatori de tip Bayes gaussian,
cazul datelor cu atrbute de intrare corelate:
exemplificare în \mathbb{R}^2 ; comparație)

■ □ • ○ CMU, 2014 fall, W. Cohen, Z. Bar-Joseph, HW2, pr. 5.c

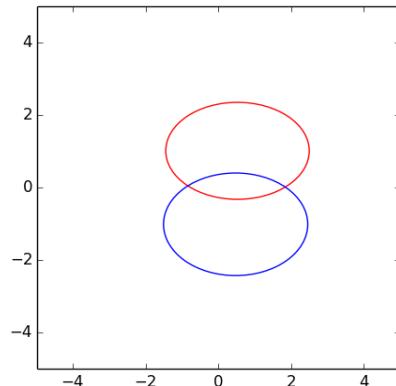
În cazul bidimensional, putem vizualiza modul în care se comportă algoritmul Bayes Naiv gaussian atunci când atrbutele de intrare (adică, trăsăturile; engl., features) sunt corelate (engl., correlated).⁴³⁶

Fie setul de date din figura (A) de mai jos, în care instanțele roșii sunt din clasa 0, iar instanțele albastre din clasa 1. Distribuțiiile comune condiționale $((X_1, X_2)|Y = 0)$ și $((X_1, X_2)|Y = 1)$ sunt de tip gaussian bidimensional. Elipsele din figurile (B), (C) și (D) reprezintă curbe de izocontur pentru [diverse] distribuții condiționale asociate celor două clase. Centrele elipselor corespund mediilor, iar curbele de izocontur sunt situate la o distanță de două deviații standard față de medii.⁴³⁷

- a. Care anume dintre perechile de elipse din figurile (B), (C) și (D) corespunde cel mai probabil distribuților condiționale care au generat datele din figura (A)?
- b. Care anume dintre aceste elipse corespunde [cel mai probabil] estimărilor de parametri făcute de către algoritmul Bayes Naiv gaussian?
- c. Dacă presupunem că probabilitățile a priori pentru cele două clase sunt egale, care dintre modelele (B), (C) și (D) va obține o acuratețe mai mare pe datele de antrenament? Care va fi natura separatorului decizional în acest caz?



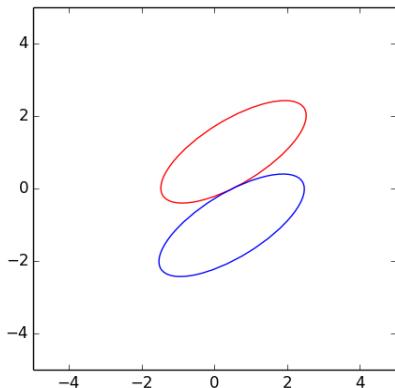
(A) Date



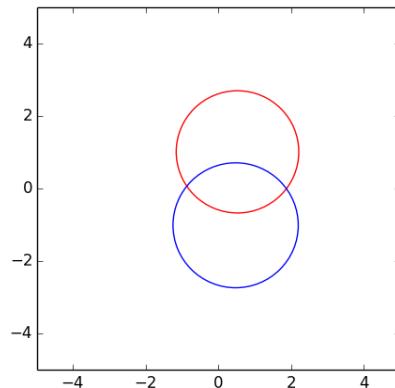
(B)

⁴³⁶Pentru definiția noțiunii de *corelare* pentru două variabile aleatoare, vedeți problema 19 de la capitolul de *Fundamente*.

⁴³⁷Mai exact, o astfel de curbă de izocontur este constituită din punctele x din planul euclidian pentru care $\Sigma^{-1/2}(x - \mu)^T \Sigma^{-1}(x - \mu) = 2 \Leftrightarrow (x - \mu)^T \Sigma^{-1}(x - \mu) = 4$, unde μ este media distribuției gaussiene considerate. Vedeți *Pattern Classification*, R. Duda, P. Hart, D. Stork, 2nd ed. (Wiley-Interscience, 2000), Appendix A, pag. 625.



(C)



(D)

Răspuns:

- a. Se observă în figura (A) că datele au fost generate de distribuții gaussiene bidimensionale având matrice de covarianță nediagonale și identice, însă având medii diferite. În plus, dacă notăm o instanță oarecare cu $x = (x_1, x_2)$ este evident că x_1 și x_2 nu sunt independente, ci există o anumită corelare între ele (mai precis, x_1 și x_2 sunt într-o relație de dependență de tip liniar). În consecință, curbele de izocontur pentru aceste distribuții sunt elipse identice ca mărime, având axe de simetrie neparalele cu axele sistemului de coordonate. Evident, doar desenul (C) corespunde acestei situații.
- b. Clasificatorul Bayes Naiv gaussian presupune independența condițională a celor două atribute (X_1 și X_2) în raport cu eticheta / variabila de ieșire (Y). Această independentă corespunde unor matrice de covarianță diagonale, respectiv unor curbe de izocontur reprezentate de elipse ale căror axe de simetrie sunt paralele cu axele sistemului de coordonate. Atât elipsele din desenul (B) cât și cele din desenul (D) satisfac aceste specificații, însă în cazul (D) elipsele sunt chiar cercuri, în vreme ce datele din figura (A) sunt dispuse în elipse având deviații standard diferite pe cele două axe de simetrie. Așadar, cazul (B) corespunde estimărilor făcute de clasificatorul Bayes Naiv gaussian pe aceste date.
- c. Evident, cazul (C) furnizează cea mai mică eroare la antrenare, deci cea mai mare acuratețe. În acest caz, se lucrează cu algoritmul Bayes Optimal gaussian. Întrucât matricele de covarianță sunt egale, separatorul decizional este de tip liniar. (Pentru justificare riguroasă, vedeti rezultatul teoretic demonstrat la problema 18.)

22. (Algoritmul Bayes Naiv [gaussian] vs. regresia logistică — comparații)

CMU, 2012 fall, T. Mitchell, Z. Bar-Joseph, HW2, pr. 1.f, 3.b-d

Fie un set de date caracterizate de atributele X_1, \dots, X_n (pe care le vom considera ca fiind fie toate Bernoulli, fie toate gaussiene) și de eticheta Y .⁴³⁸

⁴³⁸Așa este cazul problemei 14, în care variabilele condiționate $X_i|Y$ sunt de tip Bernoulli și sunt independente

Modelul Bayes Naiv [eventual de tip gaussian] va fi identificat în continuare cu abrevierea NB, iar modelul regresiei logistice cu LR.

- Presupunem că datele satisfac presupozitia de independentă condițională de tip Bayes Naiv. Atunci când numărul de exemple de antrenament tinde la infinit, care dintre cei doi clasificatori va produce rezultate mai bune, NB sau LR? Justificați.
- Presupunem acum că datele *nu* satisfac presupozitia de independentă condițională de tip Bayes Naiv. Ne punem aceeași întrebare ca mai sus: atunci când numărul de exemple de antrenament tinde la infinit, care dintre cei doi clasificatori va produce rezultate mai bune, NB sau LR? Justificați.
- Este oare posibil să calculăm distribuția $P(X)$ cu ajutorul parametrilor estimări de către algoritmul Bayes Naiv? Explicați în mod succint.
- Este oare posibil să calculăm distribuția $P(X)$ cu ajutorul parametrilor w calculați de către regresia logistică? Explicați în mod succint.

Răspuns:

- Regresia logistică este un clasificator probabilist de tip *discriminativ*, adică aproximează / modeleză $P(Y|X)$ cu ajutorul funcției logistice de argument $w \cdot X$ (unde w este vectorul de parametri, $w \in \mathbb{R}^d$, sau $w \in \mathbb{R}^{d+1}$ dacă extindem fiecare instanță X cu componenta $X_0 = 1$). În contrast cu acesta, NB este un clasificator probabilist de tip *generativ*, deci calculează distribuțiile $P(X|Y = y)$ (care în cazul GNB sunt de tip gaussian multidimensional) și de asemenea $P(Y)$, estimând parametrii acestor distribuții.

Atunci când numărul de instanțe tinde la infinit, pe de o parte aproximarea calculată de regresia logistică pentru distribuția $P(Y|X)$ tinde la distribuția reală $P(Y|X)$, iar pe de altă parte estimările făcute de clasificatorul Bayes Naiv pentru distribuțiile $P(Y)$ și $P(X|Y)$ vor tinde la distribuțiile reale corespunzătoare, dat fiind (în cazul lui $P(X|Y)$) că datele de antrenament satisfac presupozitia de independentă condițională. Corespondența dintre distribuțiile reale $P(Y|X)$ (pe de o parte) și $P(Y)$ și $P(X|Y)$ (pe de altă parte) este dată de formula lui Bayes. Prin urmare, în aceste condiții cei doi clasificatori vor produce rezultate echivalente.

- Regresia logistică va produce rezultate mai bune, fiindcă ea nu lucrează cu presupozitia de independentă condițională. (Vedeți pe de o parte problema 35 de la capitolul *Metode de regresie*, iar pe de altă parte problema 10 de la prezentul capitol.)
- Da, algoritmul Bayes Naiv este un clasificator de tip *generativ* (engl., generative classifier). Putem calcula $P(X)$ prin „marginalizarea“ distribuției condiționate $P(X|Y)$ în raport cu eticheta / clasa Y , și anume: $P(X) = \sum_y P(X|Y = y) \cdot P(Y = y)$.
- Nu, nu este posibil. Așa cum am precizat la punctul a, regresia logistică estimatează $P(Y|X)$ (nu $P(X|Y)$ și $P(Y)$), cum calculează clasificatorii de tip Bayes Naiv).

condițional două câte două, sau cel al problemei 17, în care toate variabilele condiționate, $X_i|Y$ pentru $i = 1, \dots, n$ urmează distribuții gaussiene — având varianțele $\sigma_{i0} = \sigma_{i1}$ — și sunt, de asemenea, independente condițional două câte două. (Similar este și cazul problemei 49, care constituie o combinație a precedentelor două tipuri.)

23.

(Clasificarea bayesiană gaussiană vs. regresia logistică:
Adevărat sau Fals?)

CMU, 2010 fall, Aarti Singh, midterm, pr. 1.2
CMU, (?) 15-781, midterm example questions, pr. 1.d

- a. Corespondența dintre regresia logistică și clasificatorul Bayes Naiv de tip gaussian înseamnă — în cazul în care matricele de covarianță corepunzătoare claselor sunt toate egale cu matricea-identitate⁴³⁹ — că există o corespondență 1-la-1 între parametrii celor doi clasificatori.
- b. Presupunând că lucrăm cu un număr fix de attribute, putem învăța un clasificator Bayes Optimal de tip gaussian în timp liniar în raport cu numărul de instanțe din setul de date de antrenament.

Răspuns:

- a. Fals. Se poate preciza de la început că deși cei doi clasificatori învăță separatori decizionali care au aceeași formă (și anume, o formă liniară, vedeți problema 17) nu rezultă în mod neapărat că pe un același set de date cei doi separatori decizionali învățați coincid. Faptul că matricele de covarianță sunt, toate, matrice identitate / diagonale înseamnă că presupozitia de independentă condițională este satisfăcută.⁴⁴⁰ Suntem, aşadar, în condiții similare cu cele de la problema 22.a, însă aici nu avem neapărat satisfăcută ipoteza că numărul de instanțe de antrenament tinde la infinit, caz în care rezultatele de clasificare furnizate de LR și NB ar fi echivalente. Așadar, nu există (în general) o corespondență de tip 1-la-1 între parametrii w_{LR} calculați de regresia logistică⁴⁴¹ și parametrii w_{NB} corespunzători clasificatorului Bayes Naiv gaussian.⁴⁴²

Observație: Rezultatul acesta este valabil și în cazul algoritmului Bayes Naiv cu variabile categoriale.

- b. Adevărat. În cazul clasificatorului Bayes Optimal de tip gaussian, regula de calcul pentru clasificarea unei instanțe noi $x \in \mathbb{R}^d$ este următoarea:

$$\begin{aligned} y_{GNB} &\stackrel{\text{def.}}{=} \underset{y \in \text{Val}(Y)}{\operatorname{argmax}} P(Y = y | X = x) = \underset{y \in \text{Val}(Y)}{\operatorname{argmax}} \frac{P(X = x | Y = y)P(Y = y)}{P(X = x)} \\ &= \underset{y \in \text{Val}(Y)}{\operatorname{argmax}} P(X = x | Y = y)P(Y = y) = \underset{y \in \text{Val}(Y)}{\operatorname{argmax}} \mathcal{N}(x; \mu_y, \Sigma_y)P(Y = y). \end{aligned}$$

Așadar, pentru fiecare $y \in \text{Val}(Y)$ vom avea de estimat câte o pereche de parametri, care caracterizează o distribuție gaussiană multidimensională: vectorul de medii μ_y și matricea de covarianță Σ_y . Conform problemei 53 de la capitolul de *Fundamente*, estimările celor doi parametri sunt media la eșantionare și respectiv matricea de covarianță la eșantionare, deci se calculează în timp liniar în raport cu numărul de instanțe de antrenament.

⁴³⁹LC: Chiar mai general, putem considera că aceste matrice sunt diagonale.

⁴⁴⁰Vedeți problema 34 de la capitolul de *Fundamente*.

⁴⁴¹Vedeți problema 13 de la capitolul *Metode de regresie*.

⁴⁴²În primul rând, din punctul de vedere al terminologiei, dacă ne referim la parametrii calculați de NB ca fiind parametrii distribuțiilor $P(X|Y = y)$ și $P(Y)$, este imediat că nu există o corespondență 1-la-1 între aceștia și parametrii w_{LR} calculați de LR, care servesc la estimarea / aproximarea distribuției $P(Y|X)$.

2.2 Clasificare bayesiană — Probleme propuse

2.2.1 Ipoteze de probabilitate maximă a posteriori (MAP)

24. (Formula lui Bayes; inferențe statistice; ilustrarea noțiunii de ipoteză MAP)

* CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW1, pr. 5

Imaginează-ți că te află în fața a trei cutii, care sunt etichetate cu literele A, B, C . Două dintre ele sunt goale, iar una conține un premiu. Tu nu știi în care dintre ele se află premiul; trebuie să ghicești. Procedezi în felul următor:

Mai întâi alegi la întâmplare o cutie X (să zicem că $X = A$). Totuși, chiar înainte de a deschide cutia X observi că altcineva, înaintea ta, a deschis cutia Y (să zicem că $Y = B$). Ai dreptul să privești în cutia Y , ca să vezi dacă ea conține sau nu premiul. Dacă ea nu conține premiul, vei avea dreptul să schimbi alegerea pe care ai făcut-o inițial.

În vederea luării unei decizii cât mai bune, îți se comunică *strategia* după care a fost aleasă cutia Y , și anume, folosind una dintre următoarele trei *variante*:

a. Dacă cutia pe care ai ales-o inițial conține premiul, atunci cutia Y este aleasă cu probabilitate de $1/2$ una dintre cele două cutii goale (diferite de cutia X). Dacă X este vidă, atunci Y se alege ca fiind cutia goală diferită de X .

b. Se alege aleatoriu cu probabilitate de $1/2$ una dintre cutiile diferite de cea pe care ai ales-o tu inițial, X . (În consecință, cutia Y poate sau nu să conțină premiul. Dacă Y conține premiul, ai pierdut jocul.)

c. Se alege în mod aleatoriu cu probabilitate de $1/2$ una dintre cutiile goale. (Deci este posibil ca $Y = X$. Așadar, în cazul $Y = X$ observi că a fost deschisă anterior chiar cutia X . În continuare vei putea alege una dintre celelalte două cutii.)

Considerând (pentru simplitate) că $X = A$, $Y = B$, iar cutia B este vidă, pentru fiecare dintre cele trei *variante* de mai sus decide ce cutie ar trebui să alegi în final pentru a-ți maximiza şansele de a obține premiul. Justifică-ți decizia, elaborând calculul probabilistic aferent.

25. (Probabilități condiționate; formula lui Bayes; formula probabilității totale; ilustrarea noțiunii de ipoteză MAP)

□ • ○ * CMU, 2014 fall, W. Cohen, Z. Bar-Joseph, HW1, pr. 3

Trei deținuți din SUA — desemnați aici în mod simplu prin literele a , b , și c — au fost condamnați la moarte și așteptau să fie execuțați. Guvernatorul statului în care se află închisoarea respectivă a decis să-l grățieze pe unul dintre acești trei deținuți (adică, să-i comute pedeapsa cu moartea) și a ales în mod uniform aleatoriu pe care anume dintre ei să îl grățieze. Guvernatorul l-a

informat pe gardianul încisorii despre decizia sa, precum și despre rezultatul alegerii sale aleatorii, însă i-a cerut să păstreze secretul.

După ce în încisoare au început să circule zvonuri despre grațiere, deținutul a a încercat să-l convingă pe gardian să-i spună care este soarta lui. Gardianul a refuzat. Atunci deținutul a i-a cerut gardianului să-i spună cine — mai precis, *unul* — dintre deținuții b sau c care va fi executat. Gardianul s-a gândit un pic și apoi i-a spus că deținutul b va fi executat.

Indicație importantă: La rezolvarea exercițiului va trebui să presupunem că, răspunzând la întrebarea deținutului a , gardianul a ales în mod aleatoriu — și uniform, dacă a fost cazul să aleagă între mai multe posibilități —, și că el a respectat totuși atât adevărul cât și cerința formulată de guvernator (înțeleasă într-un sens mai lax, și anume că *gardianul nu are voie să comunice niciunui deținut soarta sa, în mod direct*).

- a. Fie $X = a$, respectiv $X = b$ sau $X = c$ evenimentul care reprezintă faptul că deținutul a , respectiv b sau c a fost grațiat. Precizați care sunt valorile numerice pentru probabilitățile a priori $P(X = a)$, $P(X = b)$ și $P(X = c)$.

Notăm cu $Y = b$ evenimentul (comunicat de gardian) că deținutul b urmează să moară (adică, să fie executat). Calculați $P(X = a|Y = b)$. În urma obținerii de către deținutul a informației adiționale (de la gardian) că deținutul b va muri, a crescut oare probabilitatea ca el (deținutul a) să supraviețuiască?

Indicație: Comparați probabilitatea a posteriori $P(X = a|Y = b)$ cu probabilitatea a priori $P(X = a)$. În prealabil veți completa un tabel în care veți specifica valorile tuturor probabilităților condiționate $P(Y = y|X = x)$, cu x și $y \in \{a, b, c\}$.

	$P(Y = a X)$	$P(Y = b X)$	$P(Y = c X)$
$X = a$			
$X = b$			
$X = c$			

- b. Presupunem că deținutul a a comunicat toate cele de mai sus deținutului c . Calculați cât devine acum probabilitatea deținutului c de a supraviețui. (Așadar, calculați $P(X = c|Y = b)$.) Care dintre probabilitățile a posteriori $P(X = a|Y = b)$, $P(X = b|Y = b)$ și $P(X = c|Y = b)$ este cea mai mare?

26.

(Adevărat sau Fals?)

CMU, 2002 fall, Andrew Moore, final exam, pr. 11.a

Fie D un set de exemple (date de antrenament), iar H o mulțime de ipoteze pentru (un algoritm de) învățare automată pe datele D . Precizați care este *valoarea de adevăr* a următoarelor afirmații:

$\text{argmax}_{h \in H} P(D|h)$ este ipoteză de probabilitate maximă a posteriori, [iar] $\text{argmax}_{h \in H} P(h|D)$ este ipoteză de verosimilitate maximă.

2.2.2 Algoritmii Bayes Naiv și Bayes Optimal

27. (Algoritmii Bayes Naiv și Bayes Optimal; aplicare; numărul minimal de parametri de estimat)

• ○ *Liviu Ciortuz, 2017, pornind de la setul de date din Machine Learning, Tom Mitchell, 1997, ch. Decision Trees, page 59*

Considerăm următorul set de date de antrenament, în care variabila de ieșire este *EnjoyTennis*:

Day	Outlook	Temperature	Humidity	Wind	EnjoyTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

a. Determinați decizia luată de către algoritmul Bayes Naiv pentru instanța de test

$$X = \langle \text{Outlook} = \text{sunny}, \text{Temp} = \text{cool}, \text{Humidity} = \text{high}, \text{Wind} = \text{strong} \rangle,$$

precum și probabilitatea cu care este luată această decizie.

b. Care este numărul *minim* de parametri pe care trebuie să-l estimeze algoritmul Bayes Naiv pe aceste date [pentru a face apoi predicții pe un set oarecare de instanțe de test]? Dar în cazul clasificatorului Bayes Optimal?

c. Implementați algoritmul Bayes Naiv, iar apoi cu ajutorul acestei implementări calculați eroarea la antrenare și eroarea la CVLOO pe acest set de date.

28. (Algoritmul Bayes Naiv: aplicare)

* *CMU, 2004 fall, T. Mitchell Z. Bar-Joseph, midterm, pr. 6.a*

A	B	C	Y
0	0	1	0
0	1	0	0
1	1	0	0
0	0	1	1
1	1	1	1
1	0	0	1
1	1	0	1

Se dă setul de date din tabelul alăturat, în care A, B, C sunt atribută (de intrare) binare, iar Y este atribut de ieșire.

Care va fi răspunsul algoritmului de clasificare Bayes Naiv pentru intrarea $A = 0, B = 0, C = 1$?

29. (Algoritmul Bayes Naiv și algoritmul Bayes Optimal: aplicare)

*prelucrare de Liviu Ciortuz, după
• * CMU, 2002 fall, Andrew Moore, final exam, pr. 4.b-e*

Se dă setul de date alăturat, cu A și B variabile de intrare, iar C variabilă de ieșire.

a. Care este numărul minim de probabilități ce trebuie estimate pentru a putea construi după aceea (pe acest set de date) un clasificator de tip Bayes Naiv? Justificați.

b. Similar, pentru clasificatorul Bayes Optimal. Justificați.

c. Care este decizia clasificatorului Bayes Naiv pentru $A = 0, B = 1$? Precizați cu ce probabilitate este luată această decizie.

d. Care este decizia clasificatorului Bayes Optimal pentru $A = 0, B = 1$? Precizați cu ce probabilitate este luată această decizie.

e. Dacă rezultatele obținute la punctele c și d diferă (fie și numai în privința probabilităților cu care sunt luate deciziile), care este explicația? Justificați în mod riguros.

A	B	C	nr. apariții
0	0	1	3
0	1	0	1
0	1	1	4
1	0	0	5
1	1	0	2
1	1	1	1

30. (Aplicarea algoritmului Bayes Naiv la clasificarea de texte)

*• * Edinburgh, 2009 fall, C. Williams, V. Lavrenko, tutorial 2, pr. 2*

Firma Whizzco decide să implementeze un clasificator de texte. Pentru început, ei vor să clasifice documente aparținând fie clasei *sport* fie clasei *politică*. Ei decid să reprezinte fiecare document ca un vector de atrbute descriind prezența ori absența unor cuvinte-cheie:

goal, football, golf, defence, offence, wicket, office, strategy.

Datele de antrenament sunt reprezentate folosind o matrice în care fiecare linie este un vector de valori (0 sau 1) pentru cele 8 atrbute.

$xP=[1 \ 0 \ 1 \ 1 \ 0 \ 1 \ 1; % Politica$ $\quad \quad \quad 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 1 \ 1;$ $\quad \quad \quad 1 \ 0 \ 0 \ 1 \ 1 \ 0 \ 1 \ 0;$ $\quad \quad \quad 0 \ 1 \ 0 \ 0 \ 1 \ 1 \ 0 \ 1;$ $\quad \quad \quad 0 \ 0 \ 0 \ 1 \ 1 \ 0 \ 1 \ 1;$ $\quad \quad \quad 0 \ 0 \ 0 \ 1 \ 1 \ 0 \ 0 \ 1]$	$xS=[1 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0; % Sport$ $\quad \quad \quad 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0;$ $\quad \quad \quad 1 \ 1 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0;$ $\quad \quad \quad 1 \ 1 \ 0 \ 1 \ 0 \ 0 \ 0 \ 1;$ $\quad \quad \quad 1 \ 1 \ 0 \ 1 \ 1 \ 0 \ 0 \ 0;$ $\quad \quad \quad 0 \ 0 \ 0 \ 1 \ 0 \ 1 \ 0 \ 0;$ $\quad \quad \quad 1 \ 1 \ 1 \ 1 \ 1 \ 0 \ 1 \ 0]$
---	--

Folosind algoritmul Bayes Naiv, care este probabilitatea cu care documentul $x = (1, 0, 0, 1, 1, 1, 1, 0)$ va fi clasificat ca aparținând clasei *politică*?

31.

(Aplicarea algoritmului Bayes Naiv:
chestiunea valorilor lipsă (engl., missing values)
în datele de antrenament)

• CMU, 2013 fall, A. Smola, G. Gordon, midterm practice, pr. 9

Fie setul de date de antrenament (x, y) și datele de test z :

$$\begin{aligned} x_1 &= (0, 0, 0, 1, 0, 0, 1) & y_1 &= 1 \\ x_2 &= (0, 0, 1, 1, 0, 0, 0) & y_2 &= 1 \\ x_3 &= (1, 1, 0, 0, 0, 1, 0) & y_3 &= -1 \\ x_4 &= (1, 0, 0, 0, 1, 1, 0) & y_4 &= -1 \end{aligned}$$

$$\begin{aligned} z_1 &= (1, 0, 0, 0, 0, 1, 0) \\ z_2 &= (0, 1, 1, 0, 0, 1, 1) \end{aligned}$$

Ce *problemă* va întâmpina clasificatorul Bayes Naiv pe aceste date?

(*Indicație:* Pentru ca răspunsul dumneavoastră să fie cât mai bine justificat, veți estima toți parametrii necesari și veți aplica algoritmul pe cele două instanțe de test. Veți nota atributele cu $A1, A2, \dots$)

La curs am prezentat un „remediu“ standard pentru o astfel de *problemă*. Precizați cum se numește „tehnica“ respectivă și aplicați-o pe aceste date. După aceea, veți aplica algoritmul Bayes Naiv pentru a clasifica instanțele de test z_1 și z_2 .

32.

(Algoritmul Bayes Naiv:
calculul ratei medii a erorii – exemplificare)

• ○ CMU, 2011 spring, Tom Mitchell, midterm, pr. 5.1-2

Considerăm o problemă de clasificare binară în care se folosește o variabilă $X_1 \in \{0, 1\}$ și eticheta $Y \in \{0, 1\}$. Distribuția generativă „adevărată“ $P(X_1, Y) = P(Y) \cdot P(X_1|Y)$ este determinată conform tabelelor următoare:

Y	0	1
$P(Y)$	0.8	0.2

	$P(X_1 Y)$	$Y = 0$	$Y = 1$
$X_1 = 0$	0.7	0.3	
$X_1 = 1$	0.3	0.7	

a. Presupunem că am antrenat un clasificator Bayes Naiv folosind o infinitate de *date de antrenament* generate conform celor două tabele de mai sus. Scrieți în tabelul următor *predictiile* făcute de algoritmul Bayes Naiv pentru diferitele valori ale lui X_1 . Remarcați faptul că $\hat{Y}(X_1)$ din acest tabel este decizia [lui Bayes Naiv] cu privire la valoarea lui Y dat fiind X_1 . În coloanele care corespund probabilităților, veți scrie atât valorile concrete ale acestor probabilități, cât și modul cum au fost ele calculate (de exemplu, $0.8 \cdot 0.7 = 0.56$), iar în coloana care corespunde deciziei, veți scrie [pe fiecare linie] fie $\hat{Y} = 0$ fie $\hat{Y} = 1$.

	$P(X_1, Y = 0)$	$P(X_1, Y = 1)$	$\hat{Y}(X_1)$
$X_1 = 0$			
$X_1 = 1$			

- b. Cât este *rata medie a erorilor* (engl., expected error rate) produsă de acest clasificator Bayes Naiv pe *instanțele de test* care sunt generate conform primelor două tabele de mai sus? Cu alte cuvinte, calculați $P(\hat{Y}(X_1) \neq Y)$ unde perechile (X_1, Y) sunt generate conform celor două tabele.

Indicație: $P(\hat{Y}(X_1) \neq Y) = P(\hat{Y}(X_1) \neq Y, X_1 = 0) + P(\hat{Y}(X_1) \neq Y, X_1 = 1)$.

Pentru următoarele trei puncte ale acestui exercițiu vom considera două variabile, și anume $X_1 \in \{0, 1\}$ și $X_2 \in \{0, 1\}$, precum și eticheta $Y \in \{0, 1\}$. Y și X_1 sunt și de data aceasta generate conform primelor două tabele de mai sus, iar apoi X_2 este creat ca o *copie exactă* (adică, dupicat) după X_1 .

- c. Acum vom presupune că am antrenat un clasificator Bayes Naiv folosind o infinitate de *exemplu de antrenament* care au fost generate în conformitate cu primele două tabele de mai sus și cu regula de duplicare. Scrieți în tabelul următor *predictiile* făcute de către acest clasificator Bayes Naiv pentru diferitele valori ale perechii (X_1, X_2) . În privința probabilităților din tabel, puteți să scrieți doar cum anume sunt ele calculate (de exemplu, în loc de $0.8 \cdot 0.3 \cdot 0.3 = 0.072$ veți putea scrie doar $0.8 \cdot 0.3 \cdot 0.3$ pentru a economisi un pic de timp).

	$\hat{P}(X_1, X_2, Y = 0)$	$\hat{P}(X_1, X_2, Y = 1)$	$\hat{Y}(X_1, X_2)$
$X_1 = 0, X_2 = 0$			
$X_1 = 0, X_2 = 1$			
$X_1 = 1, X_2 = 0$			
$X_1 = 1, X_2 = 1$			

- d. Cât este *rata medie a erorilor* pentru acest clasificator Bayes Naiv pe instanțe de test care sunt generate în conformitate cu primele două tabele de mai sus și cu regula de duplicare?

- e. Comparativ cu cazul precedent (adică, fără X_2), cum s-a schimbat rata medie a erorilor (adică, a crescut ori a scăzut)? În tabelul de la punctul c, ce linie este responsabilă pentru această schimbare? Cum explicați ce s-a întâmplat?

33.

(Algoritmului Bayes Naiv:
calculul ratei medii a erorilor)

■ • ○ CMU, 2010 fall, Aarti Singh, HW1, pr. 4.2

Considerăm următoarea problemă de clasificare:

Fie variabila aleatoare $Y : Hike \in \{T, F\}$ care denotă faptul că Alice și Bob merg sau nu în drumeție în funcție de condițiile vremii: $X_1 : Sunny \in \{T, F\}$ și $X_2 : Windy \in \{T, F\}$.

Se presupune că au fost estimați următorii parametri:

$$\begin{aligned} P(Hike) &= 0.5 \\ P(Sunny | Hike) &= 0.8, \quad P(Sunny | \neg Hike) = 0.7 \\ P(Windy | Hike) &= 0.4, \quad P(Windy | \neg Hike) = 0.5 \end{aligned}$$

De asemenea, se consideră că este satisfăcută presupoziția de independentă condițională a algoritmului Bayes Naiv.

- a. Care este probabilitatea (comună) ca Alice și Bob să meargă în drumeție atunci când vremea este însorită și bate vântul, adică

$$P(\text{Sunny} = T, \text{Windy} = T, \text{Hike} = T) = ?$$

Care este decizia luată de algoritm Bayes Naiv în acest caz?

- b. Completați tabelul următor:

X_1	X_2	Y	$P(X_1, X_2, Y)$	$P_{NB}(Y X_1, X_2)$	<i>decizia algoritmului Bayes Naiv</i>
F	F	F			
F	F	T			
F	T	F			
F	T	T			
T	F	F			
T	F	T			
T	T	F			
T	T	T			

Observație: Calculele de la punctul a corespund ultimei linii din tabelul de mai sus.

- c. Care este rata medie a erorilor (engl., expected error rate) produse de algoritm Bayes Naiv? Vă reamintim că această (rată) medie este definită ca fiind suma probabilităților $P(X_1, X_2, Y)$ pentru acele (triplete de) valori ale variabilelor X_1, X_2, Y pentru care decizia luată de algoritm Bayes Naiv diferă de valoarea variabilei Y .

În cele ce urmează se presupune că se obțin mai multe informații despre vreme. Se introduce o nouă trăsătură $X_3: \text{Rainy} \in \{T, F\}$. Se presupune că în fiecare zi vremea poate fi fie *Rainy* fie *Sunny*, dar nu și *Rainy* și *Sunny*. Similar, se presupune că vremea nu poate fi într-o zi $\neg \text{Rainy}$ și $\neg \text{Sunny}$.

- d. În noile condiții, presupoziția de independentă condițională rămâne oare adevărată? Justificați.

- e. Calculați $P(\text{Sunny} = T, \text{Windy} = T, \text{Rainy} = F, \text{Hike} = T)$.

- f. Care este rata medie a erorilor produse de clasificatorul Bayes Naiv când se folosesc toate cele 3 atrbute de intrare?

- g. S-a îmbunătățit performanța algoritmului Bayes Naiv prin adăugarea atrbutoalui *Rainy*? Explicați de ce.

34.

(Algoritmul Bayes Naiv:
calculul ratei medii a erorii – exemplificare;
comparație cu regresia logistică)

• * CMU, 2009 fall, Carlos Guestrin, HW1, pr. 4.1.4

Considerăm o problemă de clasificare binară în care fiecare exemplu de antrenament are două atrbute binare $X_1, X_2 \in \{T, F\}$ și eticheta / clasa $Y \in \{T, F\}$.

Presupunem că $P(Y = T) = 0.5$, iar $P(X_1 = T|Y = T) = 0.8$, $P(X_1 = F|Y = F) = 0.7$, $P(X_2 = T|Y = T) = 0.5$ și $P(X_2 = F|Y = F) = 0.9$. (Se poate observa că atributul X_1 furnizează / constituie un indiciu întrucâtva mai puternic decât atributul X_2 în ce privește determinarea clasei unei instanțe oarecare.)

În cele ce urmează vom presupune că X_1 și X_2 sunt independente în raport cu Y .

a. Calculați probabilitățile $P(X_1 = F|Y = T)$, $P(X_1 = T|Y = F)$, $P(X_2 = F|Y = T)$ și $P(X_2 = T|Y = F)$. Asociați răspunsului dumneavoastră o justificare generală, sub forma unei formule din teoria probabilităților:

$$P(\neg A|B) = \dots, \text{ unde } A \text{ și } B \text{ sunt evenimente aleatoare oarecare.}$$

b. Scrieți regula de decizie a algoritmului Bayes Naiv pentru $X_1 = x_1$ și $X_2 = x_2$, justificând în mod succint obținerea ei.

c. Calculați rata medie a erorii produse de algoritmul Bayes Naiv, atunci când se folosesc ambele atrbute, X_1 și X_2 . (Veți da în prealabil definiția ratei medii a erorii.) Este oare această rată mai bună decât în cazul în care se folosește un singur atrbut (X_1 sau X_2)? De ce?

d. Să presupunem acum că se crează un nou atrbut, X_3 , care este o copie exactă a lui X_2 . Așadar, pentru fiecare exemplu de antrenament, atrbutele X_2 și X_3 au aceeași valoare, $X_2 = X_3$. Răspundeți la următoarele întrebări:

- Sunt X_2 și X_3 independente condițional în raport cu Y ?
- Cât este rata medie a erorii pentru Bayes Naiv acum? (Atenție! Distribuția „adevărată“ a datelor nu s-a modificat.)
- Explicați ce se întâmplă cu algoritmul Bayes Naiv. Oare *regresia logistică* are aceeași problemă? Explicați de ce.

35.

(Clasificare bayesiană: calculul ratei medii a erorilor pentru diverși clasificatori bayesieni)

prelucrare de L. Ciortuz, după
■ • ○ CMU, 2004 fall, T. Mitchell Z. Bar-Joseph, HW3, pr. 1.2

Fie funcția $Y = (A \wedge B) \vee \neg(B \vee C)$, unde A, B și C sunt variabile aleatoare binare independente, fiecare dintre ele având posibilitatea să ia valoarea 0 cu probabilitate de 50%.

a. Câți parametri trebuie să estimeze clasificatorul Bayes Naiv pentru a învăța funcția Y ? Enumerați acești parametri. Atenție: $P(\neg x)$ nu va fi socotit ca parametru dacă $P(x)$ a fost deja estimat ca parametru.

b. Care este rata medie a erorii la antrenare pentru clasificatorul Bayes Naiv la învățarea conceptului Y , presupunând că avem o infinitate de date de antrenament?

Indicație: Scrieți mai întâi tabela de adevăr a funcției Y , apoi estimați valorile parametrilor (în sensul verosimilității maxime, MLE). Pentru conveniență, centralizați toate calculele făcute de către algoritmul Bayes Naiv într-un tabel.

Convenție: În cazul în care, pentru o setare oarecare a variabilelor A , B și C , cele două probabilități calculate de către algoritmul Bayes Naiv în vederea determinării valorii y_{NB} sunt egale, convenim că algoritmul va lua decizia $y_{NB} = 1$.

- c. Câți parametri trebuie să estimeze clasificatorul Bayes Optimal pentru a „învăța” funcția Y ? Justificați în detaliu.
- d. Care este rata medie a erorii la antrenare pentru clasificatorul Bayes Optimal la învățarea conceptului Y , presupunând același lucru ca mai sus? *Atenție:* Nu este nevoie să calculați efectiv această rată; este suficient să indicați valoarea ei și să o justificați printr-un *raționament calitativ*.⁴⁴³
- e. Considerăm un alt clasificator de tip Bayes, care presupune că A este independent de C , condiționat de B și Y — în contrast cu clasificatorul Bayes Naiv, care presupune că variabilele A , B și C sunt independente două câte două în raport cu Y .

Arătați că acest clasificator Bayes va avea nevoie să estimeze mai puțini parametri decât clasificatorul Bayes Optimal la învățarea conceptului Y , și totuși va obține aceeași rată medie a erorii la antrenare (considerând că este valabilă aceeași presupoziție în legătură cu datele de antrenament).

Indicații:

- Folosind tabelă de adevăr a funcției Y , dovediți că într-adevăr proprietatea de independentă condițională formulată mai sus este satisfăcută. Și anume: pentru fiecare pereche de valori b și y ale variabilelor aleatoare B și respectiv Y , arătați că

$$P(A = a, C = c | B = b, Y = y) = P(A = a | B = b, Y = y) \cdot P(C = c | B = b, Y = y),$$

pentru $\forall a \in Val(A)$ și $\forall c \in Val(C)$.

- Scrieți regula de decizie pentru acest nou clasificator de tip Bayes și arătați că ea este echivalentă cu regula de decizie a clasificatorului Bayes Optimal.
- Calculați numărul minim de parametri de estimat de către noul nostru clasificator Bayes și enumerați-i.

36.

(Algoritmul Bayes Naiv:
independentă [condițională a] atributelor de intrare;
calculul ratei medii a erorii)

• CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW3, pr. 1.1

Presupunem că A și B sunt variabile aleatoare binare independente, fiecare dintre ele având posibilitatea de a lua valoarea 0 cu o probabilitate de 50%.

Definiți o funcție booleană $y = f(A, B)$ în aşa fel încât variabila A să nu fie independentă de variabila B în raport cu y (văzut și el ca variabilă aleatoare), însă clasificatorul Bayes Naiv să producă o rată medie a erorii de 0% (presupunând că datele de antrenament sunt în număr infinit).

⁴⁴³LC: Totuși, este recomandabil să procedați aşa după ce în prealabil ați văzut ce valoare produce algoritmul Bayes Optimal pentru [măcar] una dintre combinațiile de valori ale variabilelor, de exemplu, $A = B = C = 0$.

Demonstrați că acest clasificator are într-adevăr rata erorii de 0%.

Indicație: Puteți să identificați o astfel de funcție făcând căutare exhaustivă între cele 16 (de fapt, suficient, 14) funcții de la problema 11.

37. (Algoritmii Bayes Naiv și Bayes Optimal: complexitatea modelului)

□ • CMU, 2020 fall, E. Xing, Z. Bar-Joseph, HW1, pr. 4

Fie $X = (x_1, x_2, \dots, x_n)$ o instanță oarecare, cu x_1, x_2, \dots, x_n desemnând valorile trăsăturilor / atributelor, iar $y \in \{0, 1\}$ eticheta asociată respectivei instanțe.

Vă readucem aminte că în orice variantă / model de *clasificare de tip generativ*, probabilitatea a posteriori a etichetei (și anume, $P(y|X)$) este exprimată cu ajutorul probabilității condiționate a trăsăturilor în raport cu eticheta, $P(X|y)$:

$$P(y|X) \propto P(X|y)P(y), \quad (223)$$

unde semnul \propto înseamnă „proporțional cu“.

a. Scrieți expresia probabilității condiționate $P(X|y)$ din partea dreaptă a relației (223) în condițiile în care se folosește presupoziția [algoritmului Bayes Naiv] că trăsăturile sunt independente condițional în raport cu eticheta.

b. Să presupunem că pentru fiecare trăsătură / atribut de intrare, valorile x_i sunt luate în mulțimea $\{1, 2, \dots, K\}$. De asemenea, presupunem că distribuția probabilistă asociată etichetei y este de tip Bernoulli, iar distribuția probabilistă condiționată a fiecarei trăsături x_i în raport cu eticheta y (adică, $P(X|y)$) este de tip categorial. Vă cerem să elaborați [în mod riguros] răspunsuri la următoarele întrebări.

i. Cât este numărul total de parametri ai acestui model dacă se folosește presupoziția de independentă condiționată a algoritmului Bayes Naiv?

ii. Cât este numărul total de parametri când nu se folosește presupoziția de independentă condiționată a algoritmului Bayes Naiv?

iii. Presupunând acum că eticheta y ia valori în mulțimea $\{0, 1, \dots, M-1\}$, cum se vor schimba răspunsurile date la întrebările precedente (cu, și respectiv fără presupoziția de independentă condiționată a algoritmului Bayes Naiv)?

38. (Legătura dintre calculul parametrilor algoritmului Bayes Naiv și estimarea [lor] în sensul verosimilității maxime (MLE); caracterul liniar al separatorului decizional determinat de Bayes Naiv)

■ □ • ○ Stanford, 2007 fall, Andrew Ng, HW1, pr. 4

În această problemă vom face estimarea parametrilor algoritmului Bayes Naiv în sensul verosimilității maxime (MLE) folosind, bineînteles, presupoziția de independentă condițională.

Considerăm că în modelul nostru atrbutele de intrare x_j , cu $j = 1, \dots, n$, sunt variabile aleatoare discrete, luând valori binare: $x_j \in \{0, 1\}$. Vom numi $x = (x_1, x_2, \dots, x_n)^\top$ vector de intrare (engl., input vector). Pentru fiecare exemplu de antrenament, ieșirea asociată lui (engl., output target) este o

valoare binară: $y \in \{0, 1\}$. Ca urmare, *modelul nostru* va fi descris cu ajutorul *parametrilor* $\theta_{j0} = P(x_j = 1|y = 0)$, $\theta_{j1} = P(x_j = 1|y = 1)$ și $\theta_y = P(y = 1)$.

Vom modela *distribuția probabilistă comună* (engl., joint distribution) a perechii (x, y) în felul următor:

$$\begin{aligned} P(y) &= (\theta_y)^y (1 - \theta_y)^{1-y} \\ P(x|y=0) &= \prod_{j=1}^n P(x_j|y=0) = \prod_{j=1}^n ((\theta_{j0})^{x_j} (1 - \theta_{j0})^{1-x_j}) \\ P(x|y=1) &= \prod_{j=1}^n P(x_j|y=1) = \prod_{j=1}^n ((\theta_{j1})^{x_j} (1 - \theta_{j1})^{1-x_j}). \end{aligned}$$

a. Notând cu θ întregul set de parametri $\{\theta_y, \theta_{j0}, \theta_{j1}, j = 1, \dots, n\}$, determinați expresia funcției de verosimilitate comună (engl., joint likelihood function), $\ell(\theta) = \ln \prod_{i=1}^M P(x^{(i)}, y^{(i)}; \theta)$, în raport cu parametrii modelului care au fost prezentati mai sus.

b. Arătați că valorile parametrilor pentru care funcția de verosimilitate își atinge maximul sunt următoarele:

$$\begin{aligned} \theta_{j0} &= \frac{\sum_{i=1}^m 1_{\{x_j^{(i)}=1 \wedge y^{(i)}=0\}}}{\sum_{i=1}^m 1_{\{y^{(i)}=0\}}} \\ \theta_{j1} &= \frac{\sum_{i=1}^m 1_{\{x_j^{(i)}=1 \wedge y^{(i)}=1\}}}{\sum_{i=1}^m 1_{\{y^{(i)}=1\}}} \\ \theta_y &= \frac{\sum_{i=1}^m 1_{\{y^{(i)}=1\}}}{m}. \end{aligned}$$

Comentariu: Pentru o perspectivă mai generală asupra maximizării verosimilității datelor de antrenament în cazul folosirii clasificatorului Bayes Naiv cu variabile Bernoulli, vedeți *Comentariul* de la problema 147 de la capitolul de *Fundamente*.

c. Știm că modul în care algoritmul Bayes Naiv face predicție pentru o instanță [nouă] oarecare x constă în determina cea mai probabilă clasă pentru x .

Demonstrați că *ipoteza* produsă de către Bayes Naiv este un clasificator liniar — adică, presupunând că $P(y = 0|x)$ și $P(y = 1|x)$ sunt probabilitățile a posteriori ale claselor care sunt calculate de Bayes Naiv pentru x , atunci există un anumit vector-colonă $w \in \mathbb{R}^{n+1}$ astfel încât

$$P(y = 1|x) \geq P(y = 0|x) \text{ dacă și numai dacă } w^\top \begin{bmatrix} 1 \\ x \end{bmatrix} \geq 0 \ (\forall x).$$

(Veți considera că w_0 este termenul liber (engl., intercept term).)

39.

(Modele probabiliste [generative] corespunzătoare algoritmilor Bayes Naiv și Bayes Optimal; estimarea parametrilor)

*prelucrare de Liviu Ciortuz, după
□ • ○ CMU, 2009 spring, Tom Mitchell, HW2, pr. 2*

Precizare: Pe tot parcursul acestui exercițiu, atunci când vom folosi termenul de estimare, se va considera că este vorba despre estimare în sensul verosimilității maxime (engl., maximum likelihood estimation, MLE), nu despre estimare în sensul probabilității maxime a posteriori (engl., maximum a posteriori probability, MAP).

Presupunem că dispunem de un set de date de antrenament S , format din exemple pozitive (etichetate cu $y = 1$) și exemple negative (etichetate cu $y = 0$), fiecare instanță având $n = 10$ atribute cu valori binare, generate conform următorului model, M_{indep} , care folosește independența condițională a acestor atribute, în raport cu eticheta / clasa:

$$M_{indep} : \forall i : 1 \leq i \leq n, \forall x, y \in \{0, 1\} \quad P(X_i = x | Y = y) = p_{i,x,y}^{indep}.$$

Cu alte cuvinte, fiecare exemplu $((x_1, x_2, \dots, x_n), y)$ este generat alegând mai întâi o valoare y pentru clasa Y și, după aceea, câte o valoare x_i pentru fiecare atribut X_i , conform probabilității $p_{i,x_i,y}^{indep}$. Fiecare atribut x_i este astfel determinat în mod independent față de celealte atribute. Vom presupune de asemenea că probabilitatea de a alege clasa $Y = 1$ este de 0.5, adică $P(Y = 1) = P(Y = 0) = 0.5$.

Partea A.

- Care este numărul parameterilor „liberi“ (engl., free) din acest model, $p_{i,x,y}^{indep}$?
- Vom presupune acum că ni se dă o „instanță“ particulară a acestui model, în care parametrii au următoarele valori: $\forall i : p_{i,1,1}^{indep} = 0.8$ și $p_{i,1,0}^{indep} = 0.6$. Așadar, probabilitatea ca un atribut oarecare să aibă valoarea 1 este 0.8 pentru exemple pozitive, și respectiv 0.6 pentru exemple negative. Presupunem de asemenea că ni se dă un singur exemplu de test din clasa pozitivă, $(\bar{x}_{test}, y_{test}) = ((1, 1, 0, 0, 1, 1, 0, 1, 1, 1), 1)$. Care este probabilitatea ca instanța $\bar{x}_{test} = (1, 1, 0, 0, 1, 1, 0, 1, 1, 1)$ să fie generată de către (sau: în) clasa pozitivă? Cu alte cuvinte, cât este $P(\bar{x}_{test}|y = 1, M_{indep})$?
- Cât este $P(y = 1|\bar{x}_{test}, M_{indep})$? Așadar, cât este probabilitatea a posteriori ca instanța \bar{x}_{test} să aparțină clasei 1, conform modelului definit la punctul b?⁴⁴⁴
- Folosind acum setul de date de antrenament S (vedeți precizările de mai sus), care este regula de calcul pentru estimarea de verosimilitate maximă, $\hat{p}_{i,x,1}^{indep}$, pentru parametrul $p_{i,x,1}^{indep}$ al modelului? Care este regula de calcul pentru estimarea de verosimilitate maximă, $\hat{p}_{i,x,0}^{indep}$, pentru parametrul $p_{i,x,0}^{indep}$ al modelului? Exprimăți răspunsul în funcție de datele de antrenament, nu în funcție de parametrii „instanței“ modelului care a fost definită la punctul b de mai sus.

⁴⁴⁴Odată calculată probabilitatea $P(y = 1|\bar{x}_{test}, M_{indep})$, vom putea spune cum va clasifica algoritmul Bayes Naiv instanța \bar{x}_{test} (în modelul M_{indep}).

Partea B.

Considerăm următorul set de date de antrenament:

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	Y
0	1	0	1	1	1	0	1	1	0	0
1	1	0	1	1	1	0	1	1	0	0
1	1	1	0	1	1	1	0	1	0	0
1	1	1	1	0	1	0	0	1	0	1
0	1	0	0	0	1	1	1	0	0	1
1	1	0	0	0	1	1	0	0	0	1

e. Bazat pe acest set de date de antrenament, calculați estimările de verosimilitate maximă $\hat{p}_{i,x,y}^{indep}$ pentru parametrii modelului M_{indep} .

f. Distribuția de probabilitate Dirichlet⁴⁴⁵ este adeseori folosită ca distribuție a priori pentru a evita ca estimările parametrilor să aibă valoarea zero. Dacă presupunem că folosim probabilități a priori Dirichlet pentru fiecare dintre parametrii modelului M_{indep} și considerăm că parametrii distribuției Dirichlet sunt $\alpha_0 = \alpha_1 = 2$,⁴⁴⁶ cât sunt estimările în sens MAP pentru parametrii $p_{i,x,y}^{indep}$?

Partea C.

Vom considera acum un nou model, M_{dep} , pentru care nu se face nicio presupunere privitoare la posibilele dependențe dintre atrbute:

$$M_{dep} : \forall \bar{x} : \bar{x} \in \{0, 1\}^n, \forall y \in \{0, 1\} \quad P(\bar{X} = \bar{x} | Y = y) = p_{\bar{x},y}^{dep}.$$

Cu alte cuvinte, fiecare exemplu $((x_1, x_2, \dots, x_n), y)$ este generat alegând mai întâi o valoare y pentru clasa Y și, după aceea, alegând un întreg vector $\bar{x} = (x_1, x_2, \dots, x_n)$. Probabilitatea de a alege acest vector este dată de parametrul $p_{\bar{x},y}^{dep}$. Ca și mai sus, vom presupune că probabilitatea de a alege clasa $Y = 1$ este de 0.5. Așadar, $P(Y = 1) = P(Y = 0) = 0.5$.

g. Cât este numărul parametrilor „liberi“, $p_{\bar{x},y}^{dep}$, din acest model? Cum este acest număr, comparat cu numărul parametrilor „liberi“ din modelul M_{indep} ?

h. Prin \bar{x}_{test} ne vom referi aici la instanța de test de la punctul b. În raport cu noul model, M_{dep} , pentru a calcula probabilitatea a posteriori $P(Y|\bar{x}_{test})$ — așa cum se face atunci când se aplică algoritmul Bayes Optimal — avem nevoie ca mai întâi să estimăm parametrii $p_{\bar{x}_{test},1}^{dep}$ și $p_{\bar{x}_{test},0}^{dep}$.

Vom presupune că sunt date câte 500 de exemple de antrenament în fiecare dintre cele două clase ($y = 0$ și respectiv $y = 1$), care sunt generate conform modelului M_{indep} . Când vi se va cere să calculați estimările $\hat{p}_{\bar{x}_{test},1}^{dep}$ și $\hat{p}_{\bar{x}_{test},0}^{dep}$ în modelul M_{dep} — veți mai jos — veți folosi aceste date de antrenament.

(i.) Care este regula de calcul pentru estimarea de verosimilitate maximă $\hat{p}_{\bar{x}_{test},1}^{dep}$ pentru parametrul $p_{\bar{x}_{test},1}^{dep}$? Ca și la punctul d, veți exprima această estimare în funcție de datele de antrenament.

⁴⁴⁵Vedeți problema 41.b de la capitolul de *Fundamente*.

⁴⁴⁶Vedeți problema 128 de la capitolul de *Fundamente*. De remarcat că în acest context, folosirea distribuției de probabilitate a priori Dirichlet de parametri $\alpha_0 = \alpha_1 = 2$ este echivalentă cu folosirea distribuției de probabilitate a priori Beta, cu exact aceiași parametri; vedeți problema 43.B tot de la capitolul de *Fundamente*.

(ii.) Considerând că datele de antrenament au fost generate conform modelului M_{indep} folosind parametrii dați la punctul b , care este probabilitatea ca această estimare în sens MLE să fie zero? Altfel spus, determinați $P(\hat{p}_{\bar{x}_{test},1}^{dep} = 0)$, probabilitatea aceasta calculându-se în raport cu diferitele rezultate ale „experimentului“ care se referă la generarea datelor de antrenament conform modelului M_{indep} . (Cu alte cuvinte, se cere să se calculeze probabilitatea ca instanța \bar{x}_{test} să nu fie prezentă între cele 500 de exemple pozitive care există în setul de date de antrenament.)

(iii.) Cât este $P(\hat{p}_{\bar{x}_{test},0}^{dep} = 0)$, presupunând din nou că datele au fost generate folosind modelul M_{indep} de la punctul b ? (Cu alte cuvinte, se cere să se calculeze probabilitatea ca instanța \bar{x}_{test} să nu fie prezentă între cele 500 de exemple negative din setul de date de antrenament.)

- i. Formulați în una sau două fraze cum anume se leagă problema aceasta de discuția pe care am făcut-o la curs în legătură cu algoritmul Bayes Naiv și independentă condițională.

40.

(Algoritmul Bayes Naiv:
comparație cu alți clasificatori)

prelucrare de L. Ciortuz, după

** CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, midterm, pr. 6.b*

Considerăm un clasificator Bayes Naiv care lucrează pe un set de date descrise de atributele de intrare A și B și atributul de ieșire Y . (Exemplu: $Y = A \text{ XOR } B$). A și B sunt variabile aleatoare independente între ele.

- a. Este posibil ca, în situația aceasta, vreun alt clasificator — de exemplu, ID3, regresia logistică (eventual kernel-izată) sau SVM — să lucreze mai bine decât clasificatorul Bayes Naiv?
b. Care este motivul?

41.

(Comparație între clasificatorul Bayes Naiv
și algoritmul ID3)

CMU, 2010 fall, Ziv Bar-Joseph, midterm, pr. 5.b

Care dintre afirmațiile de mai jos sunt adevărate atât pentru clasificatorul Bayes Naiv cât și pentru algoritmul ID3 pentru învățarea de arbori de decizie? (Veți putea alege nu neapărat una singură dintre aceste afirmații.)

1. În cazul ambilor clasificatori se presupune că orice pereche de atrbute X_i și X_j cu $i \neq j$ — văzute ca variabile aleatoare — sunt independente.
2. În cazul ambilor clasificatori se presupune că orice pereche de atrbute X_i și X_j cu $i \neq j$ sunt dependente.
3. În cazul ambilor clasificatori se presupune că orice pereche de atrbute sunt independente în raport cu eticheta (adică variabila care reprezintă clasa).
4. În cazul ambilor clasificatori se presupune că orice pereche de atrbute sunt dependente în raport cu eticheta.

42. (Algoritmul Bayes Naiv – clasificator de tip MAP;
 o condiție [suficientă] pentru echivalența cu clasificarea de tip ML)
 CMU, 2012 fall, T. Mitchell, Z. Bar-Joseph, HW3, pr. 4.2

Algoritmul Bayes Naiv asociază unui exemplu x clasa c dacă aceasta maximizează probabilitatea $P(c|x)$.
 Când este această condiție [din formularea algoritmului Bayes Naiv] echivalentă cu selecționarea acelei clase c care maximizează probabilitatea $P(x|c)$?

43. (Algoritmii Bayes Naiv și Bayes Optimal: Adevărat sau Fals?)
 CMU, 2005 spring, C. Guestrin, T. Mitchell, midterm, pr. 2.b.5
 CMU, 2011 spring, Tom Mitchell, midterm, pr. 1.1.ab

- a. Clasificatorul Bayes Optimal poate să obțină rata de eroare 0 [la antrenare] pentru orice set de date. Justificați.
- b. Dacă antrenăm un clasificator Bayes Naiv folosind un număr infinit de date de antrenament care satisfac toate presupozitiile luate în calcul de acest tip de modelare (de exemplu, independentă condițională), atunci acest clasificator va produce eroare zero pe setul de exemple de antrenament considerat.
- c. Dacă antrenăm un clasificator Bayes Naiv folosind un număr infinit de date de antrenament care satisfac toate presupozitiile luate în calcul de acest tip de modelare (de exemplu, independentă condițională), atunci acest clasificator va produce eroare „adevărată“ zero pentru exemple de test generate conform aceleiași distribuții.

2.2.3 Clasificare bayesiană [cu atrbute de intrare] de tip gaussian

44. (Distribuția gaussiană unidimensională: exemplificare pentru estimarea mediei în sens MLE și calcularea unei probabilități a posteriori)
 * U. Edinburgh, 2009 fall, C. Williams, V. Lavrenko, HW2, pr. 1

Considerăm un set de date [de antrenament] care constă din *exemplu* pentru două clase. Exemplele pentru clasa 1 sunt 0.5, 0.1, 0.2, 0.4, 0.3, 0.2, 0.2, 0.1, 0.35, 0.25, iar exemplele pentru clasa 2 sunt 0.9, 0.8, 0.75, 1.0.

- a. Vă cerem să antrenați (engl., fit) câte o distribuție gaussiană unidimensională pentru fiecare din cele două clase, folosind metoda estimării în sensul verosimilității maxime (engl., Maximum Likelihood estimation, MLE). Veți presupune că varianța distribuției corespunzătoare clasei 1 este 0.0149, iar varianța distribuției pentru clasa 2 este 0.0092.
- b. Estimați de asemenea în sens MLE probabilitățile *a priori* de selecție pentru [generarea de exemple din] cele două clase.
- c. Care este probabilitatea *a posteriori* ca punctul $x = 0.6$ să aparțină clasei 1?

45.

(Algoritmul Bayes [Naiv] gaussian:
exemplificare pe date din \mathbb{R} , granițe de decizie) CMU, 2009 spring, Tom Mitchell, midterm, pr. 5

În acest exercițiu vom considera mai mulți clasificatori de tip Bayes Naiv gaussian (GNB) pentru un set de date având un singur atribut x și două clase, 0 și 1.⁴⁴⁷ Ca de obicei pentru clasificatori bayesieni, vom clasifica o instanță x ca aparținând clasei 1 dacă

$$P(y = 1|x) \geq P(y = 0|x) \Leftrightarrow \ln \frac{P(y = 1|x)}{P(y = 0|x)} \geq 0.$$

Vom folosi notația $\mathcal{N}(\mu, \sigma^2)$ pentru a desemna o distribuție normală / gaussiană de medie μ și varianță σ^2 . Vi se va cere să scrieți expresia analitică a separatorului decizional (sau, a graniței de decizie) pentru fiecare model de tip GNB prezentat mai jos.

a. Fie următorul clasificator Bayes [Naiv] gaussian:

$$\begin{aligned} x|y = 0 &\sim \mathcal{N}(0, 1) \\ x|y = 1 &\sim \mathcal{N}(2, 1) \\ P(y = 1) &= 0.5 \end{aligned}$$

Este oare granița de decizie a acestui clasificator GNB liniară? Altfel spus, puteți scrie o expresie de forma $w_0 + w_1x \geq 0$ care reprezintă granița de decizie a acestui model GNB? În cazul afirmativ, calculați valorile w_0 și w_1 .⁴⁴⁸ În cazul negativ, justificați.

b. Acum vom considera un alt clasificator de tip GNB. Noii parametri pentru cele două distribuții gaussiene sunt:

$$\begin{aligned} x|y = 0 &\sim \mathcal{N}(0, 1/4) \\ x|y = 1 &\sim \mathcal{N}(0, 1) \\ P(y = 1) &= 0.5 \end{aligned}$$

Este granița de decizie a acestui model GNB liniară? Dacă da, marcați care dintre opțiunile de mai jos constituie granița de decizie corectă. Dacă alegeți opțiunea (v), dați o scurtă explicație.

- (i) Alege clasa 1 dacă $x \geq 1/2$.
- (ii) Alege clasa 1 dacă $x \leq -1/2$.
- (iii) Alege clasa 1 dacă $x \leq 1$.
- (iv) Alege clasa 1 dacă $x \geq -1$.
- (v) Granița de decizie nu este liniară.

c. Acum vom considera o graniță de decizie pătratică. La o graniță de decizie pătratică adăugăm o nouă trăsătură, x^2 , instanței de antrenament $\langle x, y \rangle$.

⁴⁴⁷Fiind dat un singur atribut de intrare, putem renunța la termenul „Naiv“ din expresia care desemnează tipul clasificatorului.

⁴⁴⁸Adică, găsiți w_0 și $w_1 \in \mathbb{R}$ astfel încât $P(y = 1|x) \geq P(y = 0|x) \Leftrightarrow w_0 + w_1x \geq 0$ pentru $\forall x \in \mathbb{R}$.

Astfel, instanța $\langle x, y \rangle$ va fi transformată în $\langle x^2, x, y \rangle$. O graniță de decizie liniară pentru acest set de date modificat — aşadar, determinată de $w_0, w_1, w_2 \in \mathbb{R}$ cu proprietatea $w_0 + w_1x + w_2x^2 \geq 0 \Leftrightarrow P(y = 1|x) \geq P(y = 0|x)$ pentru $\forall x \in \mathbb{R}$ — produce o graniță de decizie pătratică pentru setul de date original.

Este oare posibil să găsim o graniță de decizie pătratică care corespunde exact graniței de decizie a modelului GNB de la punctul b? Dacă da, marcați care dintre opțiunile de mai jos constituie granița de decizie corectă. Dacă alegeti opțiunea (iv), dați o scurtă explicație.

(i) Alege clasa 1 dacă $x \leq -0.68$ or $x \geq 0.68$.

(ii) Alege clasa 1 dacă $-0.95 \leq x \leq 0.95$.

(iii) Alege clasa 1 dacă $-0.68 \leq x \leq 0.68$.

(iv) Granița de decizie nu este pătratică.

46.

(Algoritmul Bayes Naiv gaussian: chestiuni legate de estimarea parametrilor distribuțiilor condiționale; aplicare; algoritmul Bayes Optimal gaussian: numărul parametrilor necesari de estimat)

■ □ • ○ *prelucrare de Liviu Ciortuz, după CMU, 2014 fall, W. Cohen, Z. Bar-Joseph, HW2, pr. 5.ab CMU, 2007 spring, midterm, pr. 4.A*

Fie un set de date de antrenament pentru care $Y \in \{0, 1\}$ reprezintă clasele / etichetele, iar $X \in \mathbb{R}^d$ reprezintă vectorii de trăsături (d -dimensionali).

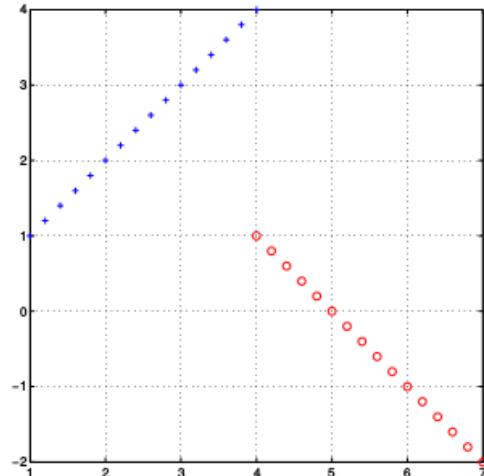
La punctele $a-c$ ale acestui exercițiu vom lucra cu modelul Bayes Naiv gaussian, în care distribuția probabilistă condițională pentru fiecare trăsătură este o distribuție gaussiană unidimensională / univariată.

a. Considerând n instanțe de antrenament generate în mod independent, $\{(X^{(1)}, Y^{(1)}), \dots, (X^{(n)}, Y^{(n)})\}$, calculați estimările de verosimilitate maximă (MLE) pentru parametrii distribuțiilor probabiliste următoare de variabilele condiționale $X^{(j)}|Y$, pentru $j = 1, \dots, d$.

b. Câtări parametri trebuie estimati pentru modelul Bayes Naiv gaussian? Veți presupune că parametrul pentru distribuția a priori a lui Y este deja cunoscut.

c. Considerăm setul de date din figura alăturată, în care sunt două clase, desemnate cu + și respectiv o. Număr de instanțe din cele două clase este același. Fiecare instanță are două trăsături, care corespund coordonatelor X_1 și X_2 , ambele având valori numere reale.

Trasați granița de decizie pe care o va învăța clasificatorul Bayes Naiv gaussian pe acest set de date.



d. În modelul Bayes Optimal [complet] Gaussian, presupunem că distribuția probabilistă condițională $Pr(X|Y)$ este o gaussiană multidimensională / multivariată, $X|Y \sim \mathcal{N}(\mu_Y, \Sigma_Y)$, unde μ_Y este vectorul medie, iar $\Sigma_Y \in \mathbb{R}^{d \times d}$ este matricea de covarianță.

Câți parametri trebuie estimati pentru modelul Bayes complet gaussian? Veți presupune, ca și la punctul b, că parametrul pentru distribuția a priori a lui Y este deja cunoscut.

Observație: Prin numărul de *parametri* se va înțelege aici *nu* numărul de vectori de medii — din \mathbb{R}^d — și numărul de matrice de covarianță — din $\mathbb{R}^{d \times d}$ — de estimat, ci numărul total de componente (adică, numere reale) ale tuturor acestor vectori de medii și matrice de covarianță, care sunt de estimat.

47.

(Algoritmul Bayes Naiv gaussian:
legătura dintre estimarea parametrilor și
maximizarea log-verosimilității datelor)

□ • CMU, 2017 fall, Nina Balcan, Practice questions, pr. 1.2

După cum știți, în clasificarea binară obiectivul este să putem prezice un „target” binar $Y \in \{0, 1\}$ pentru o instanță oarecare de test $X = (X_1, \dots, X_d)$. Aici vom presupune că $X \in \mathbb{R}^d$. Pentru a prezice Y , este suficient să estimăm funcția $P(Y|X)$. Algoritmul Bayes Naiv calculează $P(Y|X)$ astfel:

$$P(Y|X) = \frac{P(Y, X)}{P(X)} = \frac{P(Y) \prod_{j=1}^d P(X_j|Y)}{P(X)}.$$

Aici vom presupune că toate distribuțiile condiționate în raport cu clasa sunt de tip gaussian.⁴⁴⁹ Aceasta înseamnă:

$$P(X_j|Y = k) = \mathcal{N}(X_j|\mu_{j,k}, \sigma_{j,k}^2),$$

unde $\mu_{j,k}$ și $\sigma_{j,k}^2$ sunt media și respectiv varianța. Să presupunem că avem n instanțe de antrenament $\{X^{(i)}, Y^{(i)}\}_{i=1}^n$. Pentru a estima parametrii modelului

⁴⁴⁹LC: De fapt, concluzia pe care o vom formula mai jos se menține și atunci când în locul distribuțiilor gaussiene lucrăm cu distribuții Bernoulli (sau, mai general, cu distribuții categoriale).

pornind de la aceste date, vom învăța (engl., fit) câte o distribuție gaussiană univariată (sau, unidimensională) pentru fiecare trăsătură / atribut de intrare (X_j) și pentru fiecare clasă ($Y = k$).

Arătați că acest estimator de tip Bayes Naiv gaussian de fapt rezolvă următoarea *problemă* de estimare în sensul verosimilității maxime (engl., maximum likelihood estimation, MLE):

$$\max \frac{1}{n} \sum_{i=1}^n \ln P(X^{(i)}, Y^{(i)}).$$

48. (Algoritmul Bayes Naiv gaussian, cazul $\sigma_{i0} \neq \sigma_{i1}$: raportul cu regresia logistică)

□ • · CMU, 2011 spring, Tom Mitchell, HW2, pr. 2.1

În această problemă vom modifica clasificatorul Bayes Naiv gaussian care a fost prezentat la problema 17 și-l vom face ceva mai general, eliminând presupunerea că σ_i , deviația standard pentru distribuția urmată de $P(X_i|Y = k)$ nu depinde de eticheta k . Prin urmare, pentru fiecare X_i și fiecare k , unde $i \in \{1, 2, \dots, n\}$ și $k \in \{0, 1\}$, distribuția urmată de $P(X_i|Y = k)$ este de tipul $\mathcal{N}(\mu_{ik}, \sigma_{ik}^2)$.

Este oare și noua formă a probabilității condiționate $P(Y|X)$, care este implicată de acest clasificator Bayes Naiv gaussian [mai general], la fel cu forma regresiei logistice? Pentru a justifica răspunsul dumneavoastră, calculați noua formă a probabilității condiționate $P(Y|X)$.

49. (Algoritmul Bayes Naiv: deducerea regulii de decizie pentru cazul când toate atributele sunt gausiene, în afară de unul singur, care este de tip boolean)

□ • · CMU, 2010 fall, Aarti Singh, HW1, pr. 4.1

Considerăm funcția de învățare $X \rightarrow Y$, unde $Y \in \{T, F\}$ reprezintă eticheta clasei, iar X este n -uplul (X_1, X_2, \dots, X_n) , cu X_1 variabilă booleană și X_2, \dots, X_n variabile continue. Presupunem că în cazul fiecărei variabile continue X_i , distribuția $P(X_i|Y = y)$ este de tip gaussian de medie $\mu_{i,y}$ și varianță $\sigma_{i,y}^2$. Vrem ca în acest cadru să antrenăm un clasificator de tip Bayes Naiv, deci care să lucreze cu presupozitia de independență condițională a variabilelor X_i (cu $i = 1, \dots, n$) în raport cu variabila Y .

a. Enumerați toți parametrii care trebuie estimati pentru a putea clasifica o instanță nouă. Care este numărul total al acestor parametri?

b. Elaborați formula de calcul pentru $P(Y|X)$ în funcție de acești parametri și de atributele / trăsăturile X_i . Tratați cazul particular când $\sigma_{i,T}^2 = \sigma_{i,F}^2$.

50. (Algoritmul Bayes Optimal gaussian: câteva chestiuni de bază)

• CMU, 2017 fall, Nina Balcan, midterm, pr. 4.1-4,6-7

Fie două variabile aleatoare, $X \in \mathbb{R}^d$ și $Y \in \{0, 1\}$. Vom considera că atât $p_0(X = x) \stackrel{\text{not.}}{=} p(x|Y = 0)$ cât și $p_1(X = x) \stackrel{\text{not.}}{=} p(x|Y = 1)$ sunt distribuții gaussiene multidimensionale:

$$p_y(x) = \frac{1}{(2\pi)^d |\Sigma_y|^{1/2}} \exp \left(-\frac{1}{2} (x - \mu_y)^\top \Sigma_y^{-1} (x - \mu_y) \right), \text{ pentru } y \in \{0, 1\}.$$

Așadar, putem scrie $X|(Y = y) \sim \mathcal{N}(\mu_y, \Sigma_y)$. Mai departe, vom nota *probabilitățile de selecție* ale celor două clase cu $\pi_y \stackrel{\text{not.}}{=} P(Y = y)$ pentru $y \in \{0, 1\}$. Vom lucra cu un set de n „realizări“ i.i.d. ale cuplului de variabile (X, Y) , și anume $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$.

a. Clasificatorul Bayes Optimal gaussian va prezice eticheta / clasa unei noi instanțe x conform următoarei *regule de decizie*: $\hat{y} = I(\pi_1 p_1(x) > \pi_0 p_0(x))$, unde $I(\cdot)$ este funcție-indicator.⁴⁵⁰ Demonstrați că această regulă se poate scrie în mod echivalent astfel:

$$\begin{aligned} \hat{y} = & \\ I((x - \mu_1)^\top \Sigma_1^{-1} (x - \mu_1) - 2 \ln \pi_1 + \ln(|\Sigma_1|) < (x - \mu_0)^\top \Sigma_0^{-1} (x - \mu_0) - 2 \ln \pi_0 + \ln(|\Sigma_0|)). & \end{aligned} \tag{224}$$

b. (Adevărat sau fals?)

Este oare *granița de decizie* determinată de algoritmul Bayes Optimal gaussian liniară? Dacă da, explicați de ce este așa. Dacă nu, precizați în ce condiții / cazuri această graniță de decizie este liniară.

c. Dat fiind setul de date de antrenament S , precizați care este estimarea de verosimilitate maximă (MLE) pentru vectorii de medii μ_y , precum și pentru probabilitățile de selecție π_y , pentru $y \in \{0, 1\}$.

Sugestie: Pentru estimarea mediilor μ_y , vă sugerăm să faceți referire la problema 53.a de la capitolul *Fundamente*, iar pentru estimarea probabilităților de selecție π_y , vedeti problema 124.a (sau problema 43.c) de la capitolul *Fundamente*.

d. (Adevărat sau fals?)

Dat fiind setul de date de antrenament S , un *estimator nedeplasat* (engl., unbiased) al matricelor de covarianță este

$$\hat{\Sigma}_y = \frac{1}{n_y - 1} \sum_{i: y_i = y} (x_i - \hat{\mu}_y)(x_i - \hat{\mu}_y)^\top, \text{ pentru } y \in \{0, 1\},$$

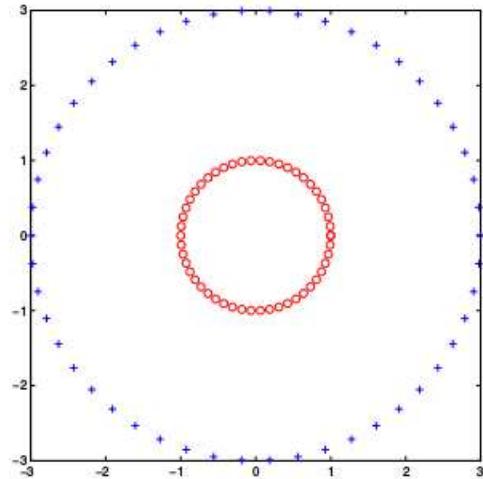
unde $n_1 = \sum_{i=1}^n y_i$ și $n_0 = \sum_{i=1}^n (1 - y_i)$.

Sugestie: Vedeți problemele 124.a și 51 de la capitolul *Fundamente*.

⁴⁵⁰Aceasta înseamnă că $\hat{y} = 1$ dacă x satisfacă condiția care a fost scrisă ca argument al funcției $I(\cdot)$, și $\hat{y} = 0$ în cazul contrar.

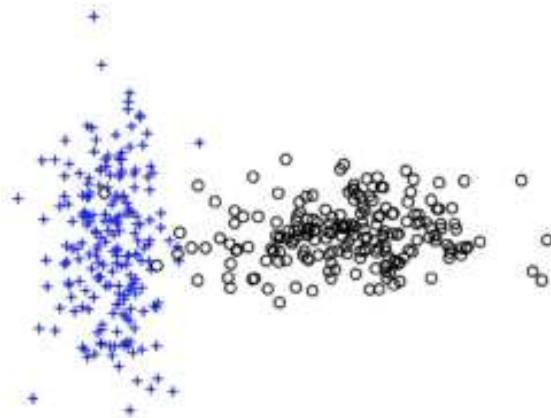
e. Presupunem că $\pi_0 = \pi_1 = 0.5$. Desenați granița de decizie determinată de algoritmul Bayes Optimal gaussian pentru setul de instanțe [ale celor două clase] din figura alăturată. Justificați cât se poate de riguros.

Sugestie: Faceți referire la relația (224).



f. Similar cu punctul precedent, veți presupune că $\pi_0 = \pi_1 = 0.5$ și veți desena granița de decizie determinată de algoritmul Bayes Optimal gaussian pentru setul de instanțe din figura alăturată.

Care dintre cele două trăsături (corespunzătoare celor două coordonate din planul euclidian) are o putere discriminativă mai mare?



51.

(Algoritmul Bayes Optimal gaussian: aplicare pe setul de date XOR)

• ○ * CMU, (?) 15-781, midterm example questions, pr. 3.2.b

În acest exercițiu ne vom referi la învățarea conceptului XOR. Așadar, vom considera că avem două atrbute de intrare, x și y , care iau valorile -1 și $+1$, iar outputul va fi pozitiv dacă și numai dacă $x \neq y$.

Ce se întâmplă dacă încercăm să aplicăm clasificatorul Bayes Optimal gaussian pe setul de date corespunzător conceptului XOR? (Vom presupune că acest clasificator este capabil să estimateze orice matrice de covarianță.)

52.

(Clasificatorul Bayes Optimal gaussian: aplicare în \mathbb{R}^2 ; granițe de decizie, în cazul $\Sigma_0 \neq \Sigma_1$)

• * CMU, 2010 fall, Aarti Singh, midterm, pr. 2.4

Fie următoarea problemă de clasificare în \mathbb{R}^2 :

Mai întâi presupunem că $P(y=0) = P(y=1) = 1/2$. De asemenea, presupunem

că funcțiile densitate de probabilitate (p.d.f.) condiționale în raport cu clasa / eticheta sunt de tip gaussian, cu media $\mu_0 \in \mathbb{R}^2$ și matricea de covarianță Σ_0 pentru clasa 0, respectiv media $\mu_1 \in \mathbb{R}^2$ și matricea de covarianță Σ_1 pentru clasa 1. Mai mult, presupunem că $\mu_0 = \mu_1 \stackrel{\text{not.}}{=} (\mu^1, \mu^2) \in \mathbb{R}^2$.

Pentru cazul

$$\Sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix} \quad \Sigma_1 = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$$

trasați într-un sistem de coordonate cartezian *curbe de izocontur* corespunzătoare celor două p.d.f.-uri condiționale, pe care le veți nota cu $p(x|y=0)$ și respectiv $p(x|y=1)$.

Apoi trasați *granițele de decizie* (engl., decision boundaries) determinate de clasificatorul Bayes Optimal gaussian. Determinați în mod riguros — adică, plecând de la *regula de decizie* a acestui clasificator — ecuația satisfăcută de aceste granițe de decizie.

În final, indicați regiunile în care clasificatorul va prezice clasa 0, respectiv clasa 1.

Indicație: Vă reamintim că pentru o variabilă aleatoare $X : \Omega \rightarrow \mathbb{R}^n$, care este reprezentată pe componente ca vector-colonă $X = (X_1, \dots, X_n)^\top$ și care urmează o distribuție gaussiană având media $\mu \in \mathbb{R}^n$ și matricea de covarianță Σ , funcția densitate de probabilitate are forma analitică următoare:

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right),$$

unde \top reprezintă operația de transpunere.⁴⁵¹

53.

(Clasificarea bayesiană gaussiană vs. regresia logistică și regresia liniară: Da sau nu?)

- ○ CMU, (?) 15-781, midterm example questions, pr. 3.3
CMU, 2009 spring, Tom Mitchell, midterm, pr. 1.b

a. Ne întrebăm dacă există vreun clasificator de tip Bayes gaussian pentru [date cu] un singur atribut de intrare x astfel încât, atunci când va fi folosit, să facă următoarele predicții:

- clasa 1 dacă $x < -1$;
- clasa 2 dacă $-1 < x < 1$;
- clasa 1 dacă $x > 1$.

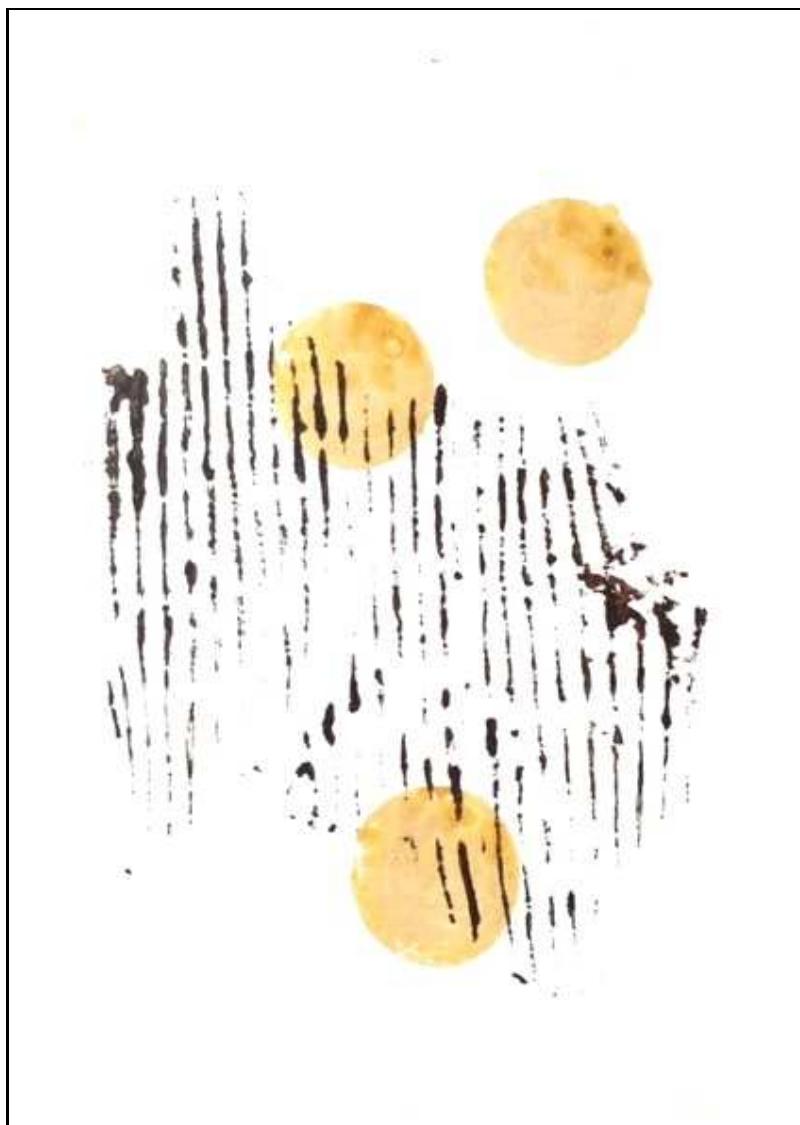
În cazul afirmativ, precizați cum anume se poate construi un astfel de clasificator.

b. Presupunem că antrenăm mai mulți clasificatori pentru a învăța funcția $f : X \rightarrow Y$, unde $X = \langle X_1, X_2, X_3 \rangle$ este vectorul de trăsături. Pentru fiecare dintre următorii clasificatori,

⁴⁵¹Observați că notația $(x - \mu)^\top \Sigma^{-1}(x - \mu)$ este matriceală. În urma efectuării operațiilor de înmulțire rezultă o matrice de tip 1×1 . În expresia $\exp(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu))$, matricea menționată mai sus este de fapt substituită cu un număr real, și anume cu unicul ei element.

- i. algoritmul Bayes Naiv [gaussian]
- ii. regresia logistică
- iii. regresia liniară

indicați dacă respectivul clasificator oferă suficiente informații ca să putem calcula probabilitățile $P(X_1, X_2, X_3, Y)$. Atunci când răspundeți cu *da*, precizați detaliile. Invers, atunci când răspundeți cu *nu*, spuneți [și] ce anume lipsește pentru a calcula $P(X_1, X_2, X_3, Y)$.



© M. Romanică

3 Învățare bazată pe memorare

Sumar

Noțiuni preliminare

- măsuri de distanță, măsuri de similaritate: ex. 2;
- normă într-un spațiu vectorial; [măsura de] distanță indusă de către o normă: ex. 7;
- k -NN vecinătate a unui punct din \mathbb{R}^d .

Algoritmul k -NN

- pseudo-cod (cf. cartea ML, pag. 232):

Training:

Store all training examples.

Classification:

Given a query/test instance x_q ,
first locate the k nearest training examples x_1, \dots, x_k ,
then take a vote among its k nearest neighbors

$$\operatorname{argmax}_{v \in V} \sum_{i=1}^k 1_{\{f(x_i)=v\}}$$

unde $1_{\{\cdot\}}$ este bine-cunoscuta *funcție-indicator*, iar V este multimea de valori pentru variabila de ieșire.

- bias-ul inductiv: „Cine se asemănă se adună“ (sau: „Spune-mi cu cine te împrietenești, ca să-ți spun cine ești“): ex. 16.a;
- exemple (simple) de aplicare: ex. 1, ex. 2;
- complexitate de spațiu: $\mathcal{O}(dn)$
complexitate de timp:

la antrenare: $\mathcal{O}(dn)$

la testare: $\mathcal{O}(dn \log n)$

[LC: $\mathcal{O}(dn k \log k)$ pt. $k > 1$ (worst case) și $\mathcal{O}(dn)$ pt. $k = 1$],

unde d este numărul de atrbute, iar n este numărul de exemple;

- arbori kd (engl., kd -trees): *Statistical Pattern Recognition*, Andrew R. Webb, 3rd ed., 2011, Wiley, pag. 163-173;
- k -NN ca algoritm ML “lazy” (vs. “eager”):
suprafețe de decizie și granițe de decizie:
diagrame Voronoi pentru 1-NN: ex. 4, ex. 11.a, ex. 20, ex. 21, ex. 22.a, ex. 23.ac;
- analiza erorilor:

- 1-NN pe date consistente: eroarea la antrenare este 0: ex. 2, ex. 12.a;

- CVLOO: ex. 3, ex. 12.b, ex. 16.bc, ex. 18, ex. 24.a, ex. 22.b, ex. 23.b;
- sensibilitatea / robustețea la „zgomote“: ex. 5, ex. 16;
- eroarea asimptotică: ex. 10, ex. 27.
- efectul trăsăturilor redundante sau irelevante;
- alegerea valorii convenabile pentru k : ex. 25;
variația numărului de erori (la antrenare și respectiv testare) în funcție de valorile lui k : ex. 26.ab;
cross-validation ca metodă neparametrică pentru alegerea lui k : ex. 26.c.

Proprietăți ale algoritmului k -NN

- (P0) output-ul algoritmului k -NN pentru o instanță oarecare de test x_q depinde de valoarea lui k : ex. 1, ex. 17;
- (P1) pe seturi de date de antrenament *consistent*, eroarea la antrenare produsă de algoritm 1-NN este 0, indiferent de *măsura de distanță* folosită: ex. 2, ex. 12.a, ex. 16.b;
- (P2) output-ul algoritmului k -NN, precum și suprafețele de decizie și separatoare decizionale depind de *măsura de distanță* folosită: ex. 7;
- (P3) „blestemul marilor dimensiuni“ (engl., the curse of dimensionality): în anumite condiții, numărul de instanțe de antrenament necesare pentru a avea un *cel mai apropiat vecin* situat la distanță *rezonabilă* față de instanța de test x_q crește exponențial în funcție de numărul de atrbute folosite: ex. 9;
- (P4) în anumite condiții, rata medie a *erorii asimptotice* a algoritmului 1-NN este mărginită superior de dublul ratei medii a erorii algoritmului Bayes Optimal: ex. 10, ex. 27;
- (P5): Pe orice set de exemple *consistent* din \mathbb{R} (deci cu un singur atribut de intrare, care este numeric și continuu), atât algoritm 1-NN cât și algoritm ID3 produc aceleși rezultate la testare / generalizare: ex. 14.a.

Comparații cu alți algoritmi de clasificare automată

- ID3: ex. 11.b, ex. 12.c, ex. 13, ex. 14, ex. 23.d, ex. 24.b;
- SVM: ex. 12.d, ex. 24.b;
- regresia logistică: ex. 24.b;
- 1-NN cu mapare cu RBF: ex. 15.

Variante ale algoritmului k -NN

- k -NN folosind alte măsuri de distanță (decât distanța euclidiană): ex. 7;
- k -NN cu *ponderarea distanțelor* (engl., distance-weighted k -NN): carte ML, pag. 236-238 (formulele 8.2, 8.3, 8.4);⁴⁵²
- algoritmul lui Shepard: ex. 8.

⁴⁵²Sectiunea 8.3 din carte ML (pag. 236-238) se referă la regresia [liniară] local-ponderată ca o formă mai generală de aproximare a [valorilor] funcțiilor, în raport cu cele calculate de către algoritm k -NN atunci când se folosește ponderarea distanțelor.

Alte metode de tip IBL

- rețele RBF: cartea ML, pag. 238-240;
- raționare bazată pe cazuri (engl., case-based reasoning): cartea ML, pag. 240-244.

3.1 Învățare bazată pe memorare — Probleme rezolvate

1.

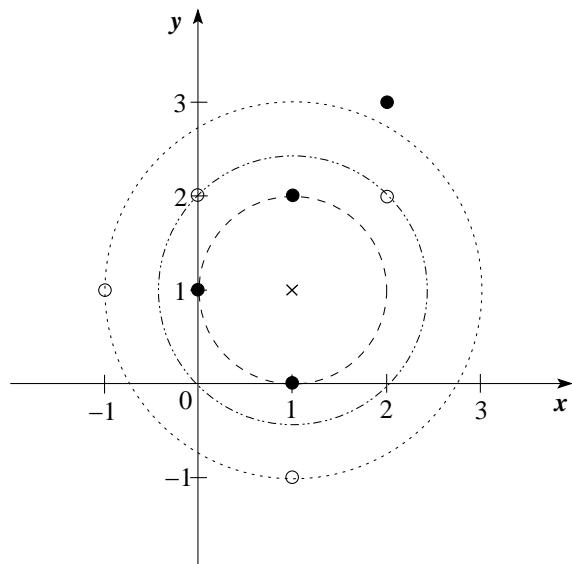
(Algoritmul k -NN: aplicare în \mathbb{R}^2 pentru diverse valori ale lui k)

■ • CMU, 2006 fall, E. Xing, T. Mitchell, final exam, pr. 2

Fie setul de instanțe de antrenament din tabelul alăturat:

- Vizualizați datele într-un sistem de axe din \mathbb{R}^2 .
- Presupunând că se folosește distanța euclidiană, care va fi predicția făcută pentru punctul $(1,1)$ de către
 - clasificatorul 3-NN?
 - clasificatorul 5-NN?
 - clasificatorul 7-NN?

x	y	
-1	1	-
0	1	+
0	2	-
1	-1	-
1	0	+
1	2	+
2	2	-
2	3	+

Răspuns:

În figura alăturată am reprezentat:

- datele de antrenament, folosind cerculețe albe pentru cele clasificate cu $-$ și cerculețe negre pentru cele clasificate cu $+$;
- punctul care trebuie clasificat, marcat cu \times ;
- vecinătățile luate în considerare de către cei trei clasificatori precizați în enunț; aceste vecinătăți sunt reprezentate prin cele trei cercuri concentrice având centrul în punctul $(1,1)$.

Analizând pe rând etichetele instanțelor din aceste trei vecinătăți, ajungem la concluzia că rezultatele obținute de cei trei clasificatori vor fi următoarele:

- 3-NN: $+$
- 5-NN: $+$
- 7-NN: $-$

Observație: Acest exercițiu pune în evidență faptul că la clasificarea cu algoritmul k -NN rezultatul poate să difere în funcție de diversele valori ale lui k , adică în funcție de cum se lărgește (sau se restrângă) vecinătatea punctului de test considerat.

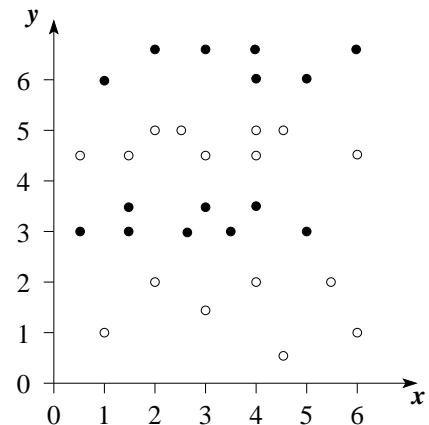
2.

(Algoritmul 1-NN: eroarea la antrenare)

CMU, 2002 fall, Andrew Moore, final exam, pr. 6.e

Figura alăturată prezintă un set de date având două atrbute de intrare x și y , și un atrbut de ieșire, ale cărui valori sunt reprezentate prin culoarea punctului (alb sau negru).

Putem alege o *metrică* (adică, *măsură* sau *funcție de distanță*) astfel încât, folosind algoritmul de învătare 1-NN (engl., [one] nearest neighbour), să obținem eroare 0 la antrenare pe setul acesta de date?



Răspuns:

Fie $d : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}$ o metrică oarecare. Conform definiției matematice, d îndeplinește următoarele *condiții*:

$$\begin{array}{ll} d(x, y) \geq 0, \forall x, y \in \mathbb{R}^2 & (\text{nenegativitatea}) \\ d(x, y) = 0 \Leftrightarrow x = y & (\text{identitatea indiscernabilor}) \\ d(x, y) = d(y, x) & (\text{simetria}) \\ d(x, y) \leq d(x, z) + d(z, y), \forall x, y, z \in \mathbb{R}^2 & (\text{inegalitatea triunghiului}) \end{array}$$

În particular, d poate fi metrica euclidiană, definită astfel:

$$d((x_1, y_1), (x_2, y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2},$$

pentru orice $(x_1, y_1), (x_2, y_2) \in \mathbb{R}^2$.

În general, când algoritmul 1-NN primește o instanță de test (x, y) , el caută în setul de antrenament punctul (x', y') cu proprietatea că distanța $d((x, y), (x', y'))$ este minimă.

În cazul particular în care punctul (x, y) însuși aparține datelor de antrenament, urmează că $(x', y') = (x, y)$, datorită primelor două proprietăți ale lui d enunțate mai sus.

Cum setul de date din enunț este *consistent* — adică nu există nicio instanță care să apară de două sau mai multe ori, însă cu etichete diferite —, vom avea de luat în considerare o singură valoare pentru calculul atrbutului de ieșire pentru punctul de test (x, y) . Evident, algoritmul 1-NN o va folosi pe aceasta ca rezultat al clasificării. Concluzia acestui raționament este că eroarea la antrenare produsă de algoritmul 1-NN pe acest set de date este 0.

Facem *observația* că nu doar pe acest set de date ci pe orice set de date de antrenament fără zgromote / inconsistențe în ce privește etichetarea, algoritmul 1-NN va avea eroarea la antrenare 0, indiferent de metrică folosită (bineînțeles, dacă există o metrică în spațiul instanțelor respective).

3.

(Algoritmul k -NN: calculul erorii la CVLOO pentru diferite valori ale lui k)*CMU, 2003 fall, T. Mitchell, A. Moore, final exam, pr. 5.ab*

Fie setul de date de antrenament din tabelul alăturat.

Vom folosi algoritmul k -NN cu distanța euclidiană (neponderată) pentru a prezice valorile atributului de ieșire Y (de tip boolean) plecând de la atributul de intrare X , care ia valori reale.Care este eroarea produsă de algoritmul k -NN la cross-validation cu metoda "Leave-One-Out" în cazurile următoare:

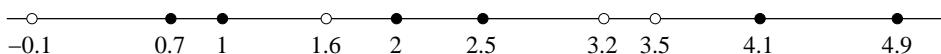
- a. $k = 1$.
- b. $k = 3$.

Exprimăți răspunsul sub forma numărului de instanțe clasificate eronat.

X	Y
-0.1	-
0.7	+
1.0	+
1.6	-
2.0	+
2.5	+
3.2	-
3.5	-
4.1	+
4.9	+

Răspuns:

Figura următoare vizualizează pe o axă datele de antrenament, precum și clasificările acestora (○ pentru instanțe negative și ● pentru instanțe pozitive):



Comportamentul celor doi clasificatori la cross-validation prin metoda "Leave-One-Out" este cel explicitat mai jos:

- Clasificatorul 1-NN:

Data	Eticheta	Vecinătate	Clasificare la CVLOO	Eroare?
-0.1	-	{0.7}	+	da
0.7	+	{1.0}	+	nu
1.0	+	{0.7}	+	nu
1.6	-	{2.0}	+	da
2.0	+	{1.6}	-	da
2.5	+	{2.0}	+	nu
3.2	-	{3.5}	-	nu
3.5	-	{3.2}	-	nu
4.1	+	{3.5}	-	da
4.9	+	{4.1}	+	nu

- Clasificatorul 3-NN:

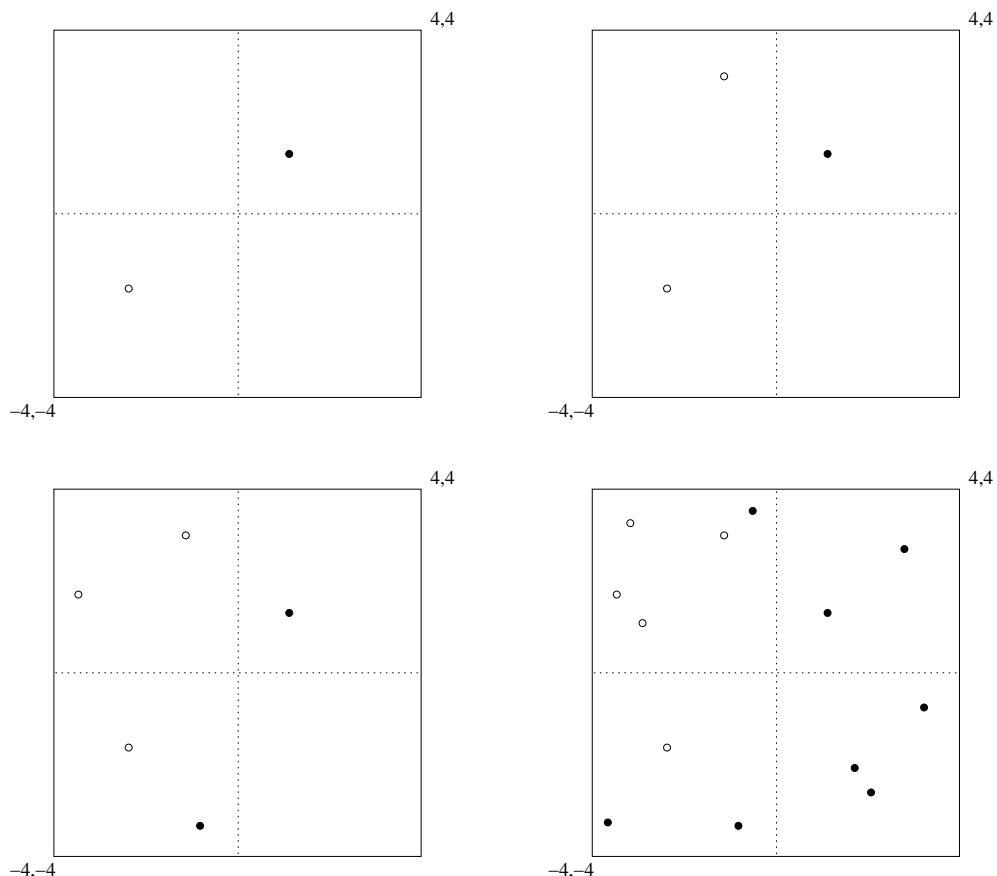
Data	Eticheta	Vecinătate	Clasificare la CVLOO	Eroare?
-0.1	-	{0.7; 1.0; 1.6}	+	da
0.7	+	{-0.1; 1.0; 1.6}	-	da
1.0	+	{0.7; 1.6; 2.0}	+	nu
1.6	-	{1.0; 2.0; 0.7/2.5}	+	da
2.0	+	{1.0; 1.6; 2.5}	+	nu
2.5	+	{1.6; 2.0; 3.2}	-	da
3.2	-	{2.5; 3.5; 4.1}	+	da
3.5	-	{2.5; 3.2; 4.1}	+	da
4.1	+	{3.2; 3.5; 4.9}	-	da
4.9	+	{3.2; 3.5; 4.1}	-	da

În concluzie, la cross-validation cu metoda “Leave-One-Out” avem 4 erori la clasificarea cu algoritmul 1-NN și 8 erori la clasificarea cu algoritmul 3-NN. Putem concluziona că rezultatul algoritmului 3-NN este foarte afectat de către puternica „mixare“ (adică, de frecvențele schimbări de clasă, de la o instanță oarecare la vecinii ei) din setul de antrenament.

4. (Algoritmul 1-NN: granițe / suprafete de decizie; diagrame Voronoi ca modalitate de învățare rapidă / “eager”)

■ • CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh, HW1, pr. 3.1-2

În fiecare din figurile următoare se dau câteva puncte în spațiul euclidian bidimensional. Fiecare dintre aceste puncte este etichetat fie pozitiv (cerc plin) fie negativ (cerc simplu).



a. Presupunând că folosim ca metrică distanța euclidiană, desenați suprafetele de decizie corespunzătoare clasificatorului 1-NN, în fiecare din aceste patru cazuri.

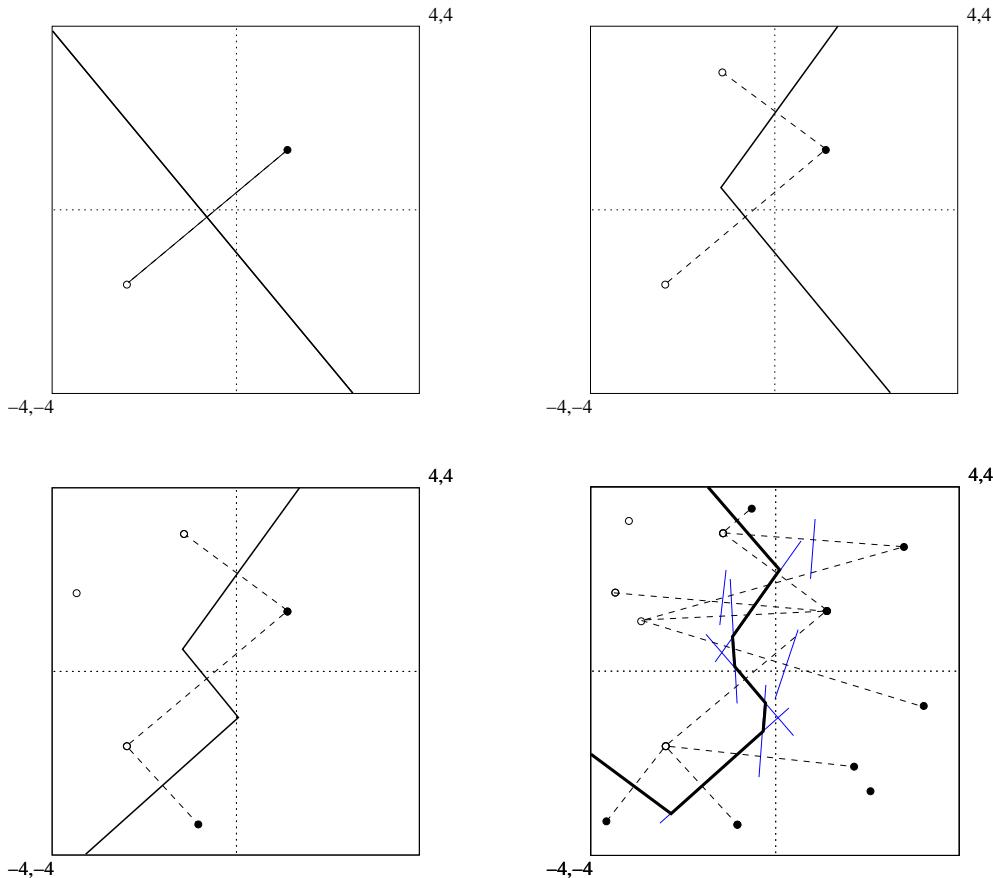
b. La curs am afirmat despre k -NN că este un clasificator lent (engl., lazy), care doar memorează toate instanțele de antrenament până ajunge la faza de test.

Totuși, la punctul precedent am văzut că putem să trasăm suprafețele de decizie pentru clasificatorul 1-NN, înainte de a intra în faza de test / generalizare. Atunci, în această fază, în loc să calculăm diverse distanțe și apoi să determinăm care sunt cei mai apropiati vecini față de punctul de test dat (notat x_q), pur și simplu î se va asigna lui x_q clasa / eticheta corespunzătoare suprafeței de decizie în care se placează.

Dacă am decide să memorăm toate aceste suprafețe de decizie (ca linii poligonale) în loc să memorăm toate datele de antrenament, am obține *întotdeauna* o îmbunătățire în ceea ce privește consumul de memorie necesar pentru acest clasificator?

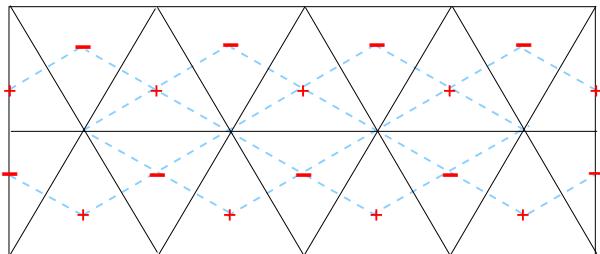
Răspuns:

- a. Suprafețele de decizie corespunzătoare clasificatorului 1-NN pentru cele patru cazuri sunt:



- b. Răspunsul este negativ, adică: nu întotdeauna memorarea separatorului decizional este mai convenabilă decât memorarea datelor de antrenament. Vom justifica dând un exemplu, care reprezintă o situație-limită, și anume, cazul care apare atunci când se creează tot atâtea suprafețe de decizie căte instanțe sunt în setul de antrenament.

În figura alăturată, pentru cele n instanțe de antrenament avem nevoie să memorăm $3\frac{n}{2}$ puncte care determină triunghiurile – suprafețele de decizie pentru clasa +. (Clasa – va fi determinată prin excludere / complementaritate.)



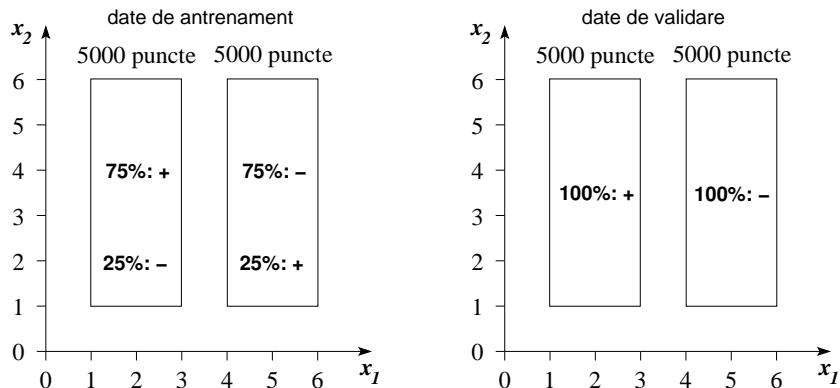
Se poate demonstra ușor următoarea proprietate: Chiar dacă am memoria o singură dată cele $3\frac{n}{2}$ puncte — fiecare punct având câte 2 coordonate — care determină contururile suprafețelor de decizie și apoi am folosi indecsi pentru a indica ce puncte determină fiecare suprafață de decizie, consumul de memorie ar fi mai mare decât dacă am memoria doar instanțele de antrenament.

5.

(Algoritmul k -NN:
rata medie a erorii la antrenare, la CVLOO și la testare
în prezența unor „zgomote“ în datele de antrenament)

• CMU, 2002 fall, Andrew Moore, final exam, pr. 7

Se constituie un set de date de antrenament și un set de date de validare, alegând în mod uniform aleatoriu puncte situate în anumite regiuni dreptunghihulare, conform imaginii următoare:



Se observă că datele de antrenament sunt „bruiate“ (engl., noisy): în fiecare regiune, 25% din date provin din clasa adversă. Datele de validare nu sunt bruiate.

Pentru fiecare dintre întrebările următoare, încercuiți fracția care este *cea mai apropiată de media / rata erorii* în cazul respectiv. Justificați în mod riguros alegerea făcută.

a. Care este rata medie a erorii la antrenare pentru clasificatorul 1-NN ?

0 1/8 1/4 3/8 1/3 1/2 5/8 2/3 3/4 7/8 1

b. Care este rata medie a erorii la cross-validation cu metoda “Leave-One-Out” pentru clasificatorul 1-NN pe setul de antrenare?

0 1/8 1/4 3/8 1/3 1/2 5/8 2/3 3/4 7/8 1

- c. Care este rata medie a erorii la testare pentru clasificatorul 1-NN pe setul de validare? (Antrenarea se face pe setul de antrenare.)

0 1/8 1/4 3/8 1/3 1/2 5/8 2/3 3/4 7/8 1

- d. Care este rata medie a erorii la antrenare pentru clasificatorul 21-NN ?

0 1/8 1/4 3/8 1/3 1/2 5/8 2/3 3/4 7/8 1

- e. Care este rata medie a erorii la cross-validation cu metoda “Leave-One-Out” pentru clasificatorul 21-NN pe setul de antrenare?

0 1/8 1/4 3/8 1/3 1/2 5/8 2/3 3/4 7/8 1

- f. Care este rata medie a erorii la testare pentru clasificatorul 21-NN pe setul de validare? (Antrenarea se face pe setul de antrenare.)

0 1/8 1/4 3/8 1/3 1/2 5/8 2/3 3/4 7/8 1

Răspuns:

- a. În acest caz, rata medie a erorii este 0, întrucât orice punct din datele de antrenament este cel mai apropiat vecin în raport cu el însuși, iar probabilitatea ca două puncte care au fost generate în mod aleatoriu (în oricare din cele două dreptunghiuri) să aibă exact aceleași coordonate și aceeași etichete diferite este foarte mică (este practic 0).

b. 3/8.

Să analizăm cazul primului dreptunghi. Constituie eroare cazul când un exemplu pozitiv (care apare cu probabilitatea „așteptată“ de 3/4) este clasificat negativ (iar aceasta se întâmplă cu probabilitatea 1/4) sau, invers, când un exemplu negativ (probabilitate: 1/4) este clasificat pozitiv (probabilitate: 3/4). Așadar, probabilitatea de a clasifica greșit un exemplu din primul dreptunghi o aflăm astfel: $3/4 \cdot 1/4 + 1/4 \cdot 3/4 = 3/8$. Pentru dreptunghiul din dreapta raționamentul este similar, iar selecția exemplelor din cele două dreptunghiuri se face cu aceeași probabilitate (1/2), deci rezultatul final este 3/8.

c. 1/4.

Probabilitatea de eroare este 1/4 pentru fiecare dreptunghi, fiindcă (în general) pentru un punct (x_1, x_2) selectat în mod aleatoriu din datele de validare din dreptunghiul respectiv aceasta (adică 1/4) este probabilitatea ca vecinul cel mai apropiat de (x_1, x_2) în setul de date de antrenament să aibă semnul opus (față de semnul lui (x_1, x_2)). Așadar, media erorii la testare cu clasificatorul 1-NN pe setul de validare este $1/2 \cdot 1/4 + 1/2 \cdot 1/4 = 1/4$.

d. 1/4.

Folosind algoritmul 21-NN, în dreptunghiul din stânga 3/4 din datele de antrenament sunt (în general) clasificate corect, iar restul de 1/4 (și anume, cele având semnul -) sunt clasificate eronat. Într-adevăr, la testarea unui punct oarecare (x_1, x_2) din acel dreptunghi, în 21-NN vecinătatea lui (x_1, x_2) , unul (cel mai apropiat vecin) este însuși (x_1, x_2) , iar dintre ceilalți 20 cei mai apropiati vecini 3/4 sunt (în general) de semn +, iar 1/4 de semn -. Analog se raționează pentru dreptunghiul din dreapta. Deci rata medie a erorii la antrenare este:

$$\frac{1}{2} \left(\frac{3}{4} \cdot 0 + \frac{1}{4} \cdot 1 \right) + \frac{1}{2} \left(\frac{3}{4} \cdot 0 + \frac{1}{4} \cdot 1 \right) = \frac{1}{4}.$$

e. 1/4.

În cazul primului dreptunghi, apare eroare atunci când un exemplu pozitiv (probabilitate de selectare: 3/4) este clasificat negativ (probabilitate: 0, fiind vorba de 21-NN) sau când un exemplu negativ (probabilitate de selectare: 1/4) este clasificat pozitiv (probabilitate: 1).⁴⁵³ Așadar, probabilitatea de a clasifica greșit un exemplu din primul dreptunghi este $3/4 \cdot 0 + 1/4 \cdot 1 = 1/4$. Cazul dreptunghiului din dreapta este similar, prin urmare rezultatul final este 1/4.

f. 0.

Datorită distribuției uniforme a datelor de antrenament, fiecare instanță (x_1, x_2) din setul de validare va avea majoritatea vecinilor — dintre cei mai apropiati, selectați de către algoritmul 21-NN din setul de date de antrenament — de același semn cu semnul lui (x_1, x_2) . (Și anume, în medie de 3 ori mai mulți vecini decât cei de semn contrar.) Prin urmare, fiecare punct din setul de validare va fi clasificat corect.

Observație (1): Comparând rezultatele obținute la punctele a și c pe de o parte cu cele de la punctele d și f (sau chiar b și e) pe de altă parte, se observă că are loc o relație exact de același gen cu cea din definiția fenomenului de *overfitting* (sau, *supra-specializare*):⁴⁵⁴ erorile produse de algoritmii 1-NN și 21-NN la antrenare sunt în relația $0 < 1/4$ dar în relație inversă ($1/4 > 0$) la testare (respectiv $3/8 > 1/4$ la cross-validation). Această manifestare a fenomenului de overfitting va fi pusă în evidență foarte clar în graficul de la finalul *Observației (2)*; vedeti curba roșie (ascendentă) și curba neagră (descendentă), în special în zona corespunzătoare valorilor $k \in \{1, \dots, 22\}$.

Observație (2): Remarcați „sublinierea“ din expresia „fracția cea mai apropiată de media / rata erorii“ din enunț. Generarea aleatorie a datelor poate conduce la rezultate ușor diferite față de cele pe care le-am obținut mai sus. În urma realizării unei implementări,⁴⁵⁵ au fost obținute următoarele rezultate de tip *eroare medie*, calculată în urma repetării de 100 de ori a generării datelor și aplicării algoritmilor k -NN, conform cerințelor din enunț:

- a. 0, b. 0.374022, c. 0.250472, d. 0.249342, e. 0.253088, f. 0.006436.

Constatăm că se verifică „previziunile“ noastre din rezolvarea de mai sus.

⁴⁵³În 21-NN vecinătatea punctului considerat (pentru testare), spre deosebire de cazul erorii la antrenare, la CVLOO nu se mai include punctul respectiv.

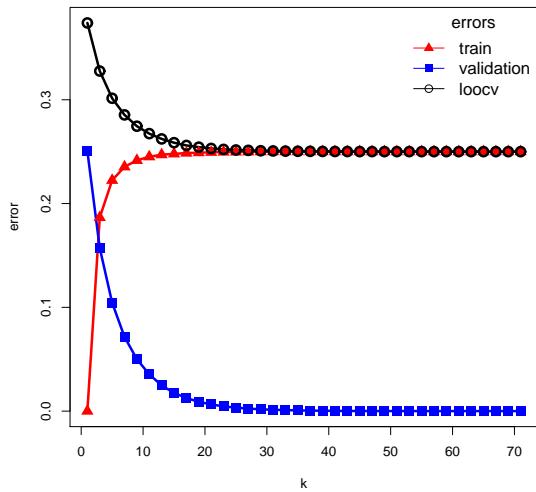
⁴⁵⁴Vă readucem aminte definiția lui Tom Mitchell pentru fenomenul de *overfitting* (vedeți cartea *Machine Learning*, pag. 67): două ipoteze h și h' obținute (eventual cu un același algoritm de clasificare automată) pe un set de date sunt în raport de overfitting dacă

$$\text{error}_{\text{train}}(h) < \text{error}_{\text{train}}(h'), \quad \text{dar} \quad \text{error}_{\mathcal{D}}(h) > \text{error}_{\mathcal{D}}(h'),$$

unde \mathcal{D} este distribuția reală a datelor.

⁴⁵⁵Implementarea a fost făcută de către studentul Sebastian Ciobanu de la Facultatea de Informatică a Universității „Al. I. Cuza“ din Iași în semestrul I al anului universitar 2016-2017.

Evoluția celor trei tipuri de eroare (la antrenare, la validare și cross-validation cu metoda “leave-one-out”), pentru $k = 1, 3, \dots, 69, 71$ este prezentată în figura alăturată. Se observă convergența la aceeași valoare (aprox., 0.25) pentru eroarea la antrenare și eroarea CV-LOO (pe setul de date de antrenare) începând din jurul valorii $k = 31$, precum și convergența la valori foarte apropiate de 0 pentru eroarea la testare pe setul de date de validare, începând din jurul valorii $k = 35$.



6.

(O versiune ipotetică pentru algoritmul k -NN:
selectarea celor mai apropiati vecini
nu pe baza calculării funcției de distanță,
ci folosind un oracol / “black box”)

CMU, 2009 spring, Ziv Bar-Joseph, final exam, pr. 2.2-3

Se încearcă clasificarea unor puncte în spațiul euclidian bidimensional. Sunt date n instanțe P_1, P_2, \dots, P_n , precum și etichetările corespunzătoare c_1, c_2, \dots, c_n , unde c_1, c_2, \dots, c_n reprezintă valori dintr-o mulțime C . În schema de clasificare k -NN, fiecare element nou Q este clasificat cu eticheta majoritară obținută în cadrul vecinătății formate din cei mai apropiati k vecini.

Să presupunem că măsura de distanță nu este dată în mod explicit. În locul acesteia, aveți la dispoziție un “black box”. Dacă se introduc instanțele $P_{i_1}, P_{i_2}, \dots, P_{i_l}$ (unde l este un număr natural oarecare) și un punct Q , black box-ul returnează cel mai apropiat vecin al lui Q , adică un element $P_{i_0} \in \{P_{i_1}, P_{i_2}, \dots, P_{i_l}\}$, precum și clasificarea corespunzătoare (c_{i_0}).

- Este posibil să se construiască un algoritm de tip k -NN bazat doar pe acest black box? Dacă da, explicați cum anume, iar dacă nu, explicați de ce nu este posibil.
- Dacă, în schimb, acel black box returnează cei mai apropiati j vecini (și etichetele corespunzătoare), iar $j \neq k$, este posibil să se construiască un algoritm de tip k -NN bazat doar pe acest black box? Dacă da, explicați cum anume, iar dacă nu, explicați de ce nu este posibil.

Răspuns:

- Da. Se folosește black box-ul dându-i ca intrare mai întâi mulțimea de exemple P_1, P_2, \dots, P_n ; se obține cel mai apropiat vecin al lui Q și eticheta sa. Apoi se scoate instanța / punctul returnat din mulțimea de exemple. Se repetă acest procedeu de k ori, iar la final se va alege pentru Q eticheta corespunzătoare majorității din mulțimea de k vecini care au fost identificați de către black box.

b. Dacă $j < k$, atunci se folosește black box-ul de $[k/j]$ ori, obținându-se $j * [k/j]$ cei mai apropiati vecini și clasificările acestora. Notăm cu V_1 mulțimea formată din acești vecini. În cazul în care $k \neq j * [k/j]$, pentru a obține restul de vecini necesari pentru k -NN, adică încă $k - j * [k/j]$ vecini, vom folosi black box-ul încă o dată. Vom nota cu V_2 mulțimea acestor noi j vecini. Dintre aceștia va trebui să alegem doar $k - j * [k/j]$ instanțe. Vom proceda astfel: considerăm o nouă mulțime de instanțe alcătuită din elementele lui V_2 și $j - (k - j * [k/j])$ dintre toți ceilalți vecini obținuți anterior (V_1). Aplicăm black box-ul pe acest nou set de date și vom obține cei mai apropiati j vecini pentru punctul Q . Printre aceștia se vor afla cei $k - j * [k/j]$ vecini căutați.

Dacă $j > k$, iar black box-ul nu acceptă intrări duplicate, atunci el nu poate fi folosit pentru a determina cei mai apropiati k vecini ai instanței de test.

Dar dacă $j > k$ și black box-ul acceptă intrări duplicate, este posibil să rezolvăm problema, cel puțin în situația în care cei mai apropiati j vecini ai lui Q se află toți la distanțe diferite față de acesta.⁴⁵⁶ De exemplu, pentru $k = 1$ și $j = 3$ vom proceda astfel:

Pasul 1: Aplicăm black box-ul inputului P_1, \dots, P_n . Vom obține outputul $P_{i_1}, P_{i_2}, P_{i_3}$.

Pasul 2: Corespunzător fiecărui punct $P_{i_1}, P_{i_2}, P_{i_3}$, vom aplica pe rând black box-ul inputului

- $P_{i_1}, P_{i_1}, P_{i_2}, P_{i_3}$,
- $P_{i_1}, P_{i_2}, P_{i_2}, P_{i_3}$ și respectiv
- $P_{i_1}, P_{i_2}, P_{i_3}, P_{i_3}$.

Se poate constata că într-unul singur din aceste cazuri black box-ul returnează outputul $P_{i_1}, P_{i_2}, P_{i_3}$.⁴⁵⁷ De pildă, dacă outputurile în aceste trei cazuri sunt

- $P_{i_1}, P_{i_1}, P_{i_2}$,
- $P_{i_1}, P_{i_2}, P_{i_2}$,
- $P_{i_1}, P_{i_2}, P_{i_3}$,

rezultă că P_{i_3} este cel mai distanță între cei trei vecini ai lui Q .

Pasul 3: Considerând că lucrurile stau ca la finalul pasului precedent, pentru fiecare dintre punctele P_{i_1}, P_{i_2} vom aplica black box-ului următorul input:

- $P_{i_1}, P_{i_1}, P_{i_1}, P_{i_2}$ și respectiv
- $P_{i_1}, P_{i_2}, P_{i_2}, P_{i_2}$.

Dacă vom obține outputul $P_{i_1}, P_{i_1}, P_{i_1}$, va rezulta că P_{i_1} este mai apropiat de Q decât punctul P_{i_2} . Invers, dacă obținem outputul $P_{i_2}, P_{i_2}, P_{i_2}$, va rezulta că P_{i_2} este mai apropiat de Q decât punctul P_{i_1} .

⁴⁵⁶Rămâne de analizat varianta contrară.

⁴⁵⁷Am presupus că black box-ul returnează exact j instanțe, chiar dacă între P_1, \dots, P_n există mai multe instanțe egal depărtate față de punctul Q , în spățiu mai multe instanțe situate la maximumul distanțelor dintre fiecare din cele j pe de o parte și punctul Q pe de altă parte.

7.

(Algoritmul 1-NN: suprafețele de decizie
[și separatorii decizionali] depind de măsurile de distanță folosite)

CMU, 2008 fall, Eric Xing, HW1, pr. 3.1.2

Se dau două puncte din spațiul euclidian bidimensional: punctul $(-1, 0)$ clasificat negativ și punctul $(1, 0)$ clasificat pozitiv.

Clasificatorul 1-NN care folosește distanța euclidiană și are ca set de date de antrenament cele două puncte de mai sus are următoarea *formă analitică* (ușor de dedus):

- dat fiind un punct arbitrar (x, y) ,
în cazul în care $x > 0$, eticheta asignată punctului respectiv este +, iar în cazul $x < 0$ eticheta asignată este -;
 - dreapta $x = 0$ este *granița de decizie* (engl., decision boundary) corespunzătoare acestui clasificator.
- a. Care va fi forma analitică a clasificatorului 1-NN dacă în locul distanței euclidiene (indusă de norma L_2) se folosește distanța Manhattan (indusă de norma L_1)? Vă reamintim că distanța Manhattan dintre două puncte (x_1, y_1) și (x_2, y_2) este $|x_1 - x_2| + |y_1 - y_2|$.
 - b. Dar dacă se folosește distanța Chebyshev, care este indușă de norma L_∞ și este definită în \mathbb{R}^2 prin

$$d((x_1, y_1), (x_2, y_2)) = \max\{|x_1 - x_2|, |y_1 - y_2|\}$$

Răspuns:

a. Putem exprima distanța Manhattan dintre un punct oarecare (x, y) din plan și de fiecare dintre cele două puncte date în enunț — punctul $(1, 0)$ clasificat pozitiv și punctul $(-1, 0)$ clasificat negativ — astfel:

$$\begin{aligned} d_+ &\stackrel{\text{not.}}{=} d((x, y), (1, 0)) = |x - 1| + |y| \\ d_- &\stackrel{\text{not.}}{=} d((x, y), (-1, 0)) = |x + 1| + |y| \end{aligned}$$

În consecință, a stabili care dintre cele două distanțe (d_+ și d_-) este mai mică revine la a compara (doar) expresiile $|x - 1|$ și $|x + 1|$.

Avem următoarele cazuri:

- dacă $x > 1$, atunci

$$\left. \begin{aligned} d_+ &= x - 1 + |y| \\ d_- &= x + 1 + |y| \end{aligned} \right\} \Rightarrow d_+ < d_- \Rightarrow \text{punctul } (x, y) \text{ va fi clasificat +}$$

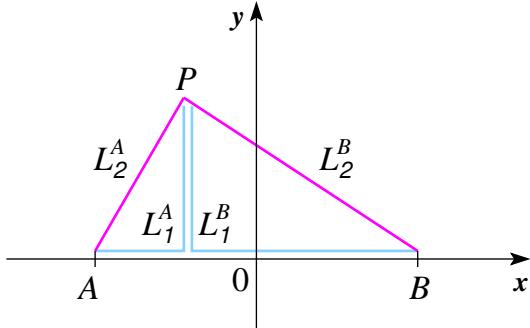
- dacă $x < -1$:

$$\left. \begin{aligned} d_+ &= -x + 1 + |y| \\ d_- &= -x - 1 + |y| \end{aligned} \right\} \Rightarrow d_+ > d_- \Rightarrow \text{punctul } (x, y) \text{ va fi clasificat -}$$

- dacă $-1 \leq x \leq 1$:

$$\left. \begin{aligned} d_+ &= -x + 1 + |y| \\ d_- &= x + 1 + |y| \end{aligned} \right\} \Rightarrow \begin{cases} d_+ < d_- \text{ pentru } x > 0 \text{ deci } (x, y) \text{ va fi clasificat +} \\ d_+ > d_- \text{ pentru } x < 0 \text{ deci } (x, y) \text{ va fi clasificat -} \\ d_+ = d_- \text{ dacă și numai dacă } x = 0. \end{cases}$$

Rezultă că granița de decizie a acestui clasificator este dreapta $x = 0$. Chiar mai mult: forma analitică a clasificatorului 1-NN care folosește distanța Manhattan pe setul de date din enunț este aceeași cu a clasificatorului 1-NN care folosește distanța euclidiană.



Observație: La concluzia de mai sus se putea ajunge mult mai ușor ținând cont de următoarea proprietate (demonstrabilă imediat pe cale geometrică): pentru orice două puncte A și B din \mathbb{R}^2 care au aceeași ordonată au loc următoarele două relații de echivalentă:

$$L_2(P, A) < L_2(P, B) \Leftrightarrow L_1(P, A) < L_1(P, B) \text{ pentru } \forall P \in \mathbb{R}^2 \text{ și}$$

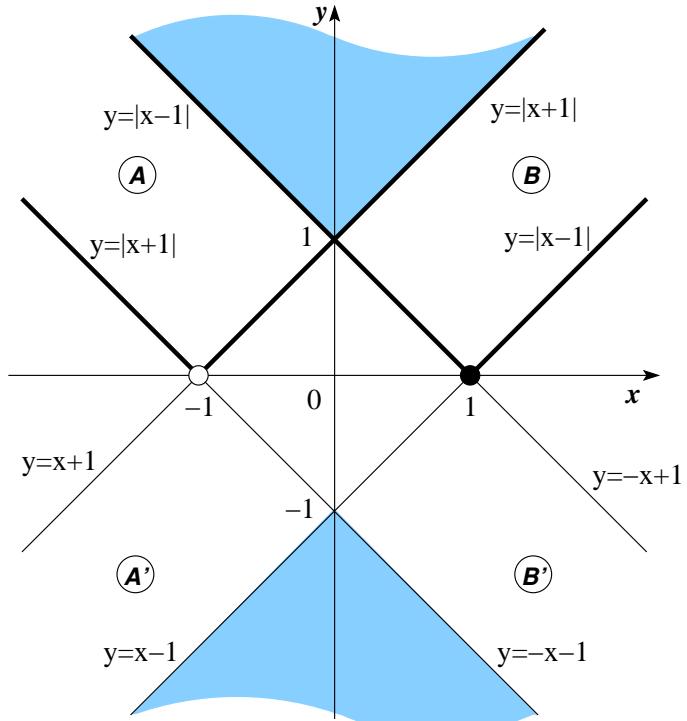
$$L_2(P, A) = L_2(P, B) \Leftrightarrow L_1(P, A) = L_1(P, B) \text{ pentru } \forall P \in \mathbb{R}^2.$$

b. Dacă se folosește distanța Chebyshev, atunci conform definiției acestei metriki vom avea:

$$\begin{aligned} d_+ &= d((x, y), (1, 0)) \\ &= \max\{|x - 1|, |y|\} \end{aligned}$$

$$\begin{aligned} d_- &= d((x, y), (-1, 0)) \\ &= \max\{|x + 1|, |y|\}. \end{aligned}$$

Mai întâi vom trasa graficele funcțiilor $y = |x - 1|$ și $y = |x + 1|$, și vom obține rezultatul din figura alăturată.



Cazul i: Se observă că $|y| \geq |x - 1|$ și $|y| \geq |x + 1|$ pentru toate punctele (x, y) situate în zonele unghiulare hașurate. În consecință, d_+ va fi egal cu d_- pentru orice punct din aceste zone. Așadar, zonele hașurate vor apartine graniței / suprafeței de decizie a clasificatorului nostru.

Cazul ii: Considerăm acum zonele identificate prin literele A, A', B și B' în figura de mai sus.⁴⁵⁸ Se observă ușor că pentru zonele A și A' avem $d_+ =$

⁴⁵⁸Analitic, zona A este definită de punctele (x, y) care satisfac inecuațiile $|x + 1| < |y| < |x - 1|$ și $y > 0$; zona A' : $|x + 1| < |y| < |x - 1|$ și $y < 0$; zona B : $|x - 1| < |y| < |x + 1|$ și $y > 0$; zona B' : $|x - 1| < |y| < |x + 1|$ și $y < 0$.

$\max\{|x - 1|, |y|\} = |x - 1|$ și $d_- = \max\{|x + 1|, |y|\} = |y|$, iar pentru zonele B și B' avem $d_+ = \max\{|x - 1|, |y|\} = |y|$ și $d_- = \max\{|x + 1|, |y|\} = |x + 1|$. În consecință, $d_- = |y| < |x - 1| = d_+$ pentru zonele A și A' , iar $d_+ = |y| < |x + 1| = d_-$ pentru zonele B și B' . Așadar, zonele A și A' vor fi clasificate negativ, iar zonele B și B' vor fi clasificate pozitiv.

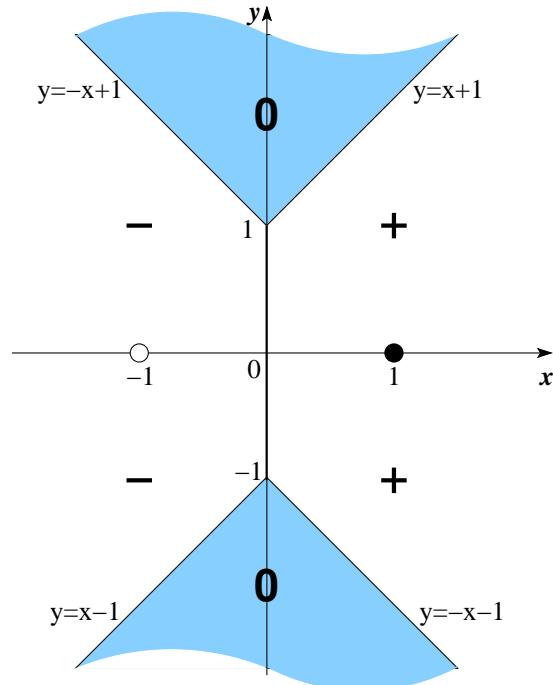
Cazul iii: Pentru toate celelalte zone rămase în discuție — adică pentru orice punct (x, y) situat în afara celor două zone hașurate și în afara zonelor A, A', B, B' —, vom avea $d_+ = \max\{|x - 1|, |y|\} = |x - 1|$ și $d_- = \max\{|x + 1|, |y|\} = |x + 1|$. Din grafic se observă că $|x + 1| > |x - 1|$ pentru $x > 0$ și $|x - 1| > |x + 1|$ pentru $x < 0$, iar $|x + 1| = |x - 1|$ pentru $x = 0$. Așadar, sumarizând, în zone avem: $d_+ = \min\{d_+, d_-\}$ pentru $x > 0$ și $d_- = \min\{d_+, d_-\}$ pentru $x < 0$, iar $d_+ = d_-$ pentru $x = 0$.

Concluzionând, forma analitică a clasificatorului 1-NN care folosește distanța L_∞ este următoarea:

- (x, y) va fi etichetat cu $+$ dacă $x > 0$ și $-x - 1 < y < x + 1$;
- (x, y) va fi etichetat cu $-$ dacă $x < 0$ și $x - 1 < y < -x + 1$;
- în rest este vorba de suprafață de separare, adică locul geometric al punctelor (x, y) pentru care distanța față de cele două puncte din enunț este egală.

Reprezentarea grafică a suprafețelor de decizie este dată în figura alăturată.

Este de remarcat faptul că pentru acest clasificator granița / suprafața de decizie nu este formată doar din drepte, ci este reuniunea unei drepte (axa Oy) cu două intersecții de (câte două) semiplane.



8.

(Algoritmul / metoda lui Shepard – aplicare)
CMU, 2002 fall, Andrew Moore, final exam, pr. 6.a-d

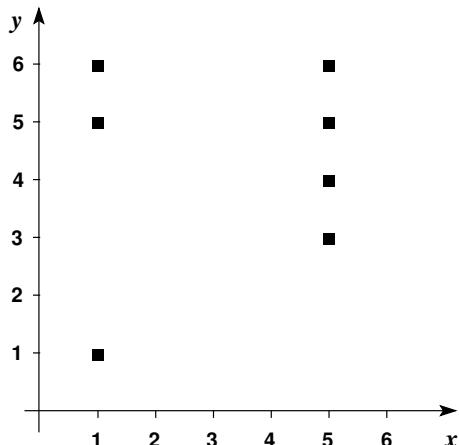
Figura de mai jos prezintă un set de date de antrenament cu un atribut de intrare $x \in \mathbb{R}$ și un atribut de ieșire $y \in \mathbb{R}$.

Vom estima din aceste date câteva valori ale unei funcții continue $f : \mathbb{R} \rightarrow \mathbb{R}$, folosind *metoda lui Shepard*. Aceasta este o variantă (de tip *regresie*) a algoritmului k -NN în care se iau în considerare toate punctele de antrenament, dar se aplică ponderi în funcție de distanță:

$$\hat{f}(x) \leftarrow \frac{\sum_i w(x, x_i) f(x_i)}{\sum_i w(x, x_i)}$$

Se va considera

$$w(x, x_i) = \begin{cases} 1, & \text{dacă } |x - x_i| \leq 3 \\ 0, & \text{în rest.} \end{cases}$$



Care va fi valoarea prezisă pentru funcția f pentru

- a. $x = 1?$
- c. $x = 5?$
- b. $x = 3?$
- d. $x = 6?$

Răspuns:

Din modul cum au fost definite ponderile w ajungem la concluzia că valoarea lui f pentru un punct oarecare x va fi calculată ca medie aritmetică a valorilor / componentelor y din acele date de antrenament pentru care abscisa (x') este situată la distanță de cel mult 3 unități de punctul x care ne interesează.

a. Avem $|1 - 1| = 0 \leq 3$ și $|1 - 5| = 4 > 3$, prin urmare vor fi luate în considerare doar valorile învățate pentru $x = 1$.

$$\hat{f}(1) = \frac{1 + 5 + 6}{3} = \frac{12}{3} = 4.$$

b. Avem $|3 - 1| = 2 \leq 3$ și $|3 - 5| = 2 \leq 3$, prin urmare vor fi luate în considerare și valorile învățate pentru $x = 1$ și cele pentru $x = 5$.

$$\hat{f}(3) = \frac{1 + 5 + 6 + 3 + 4 + 5 + 6}{7} = \frac{30}{7}.$$

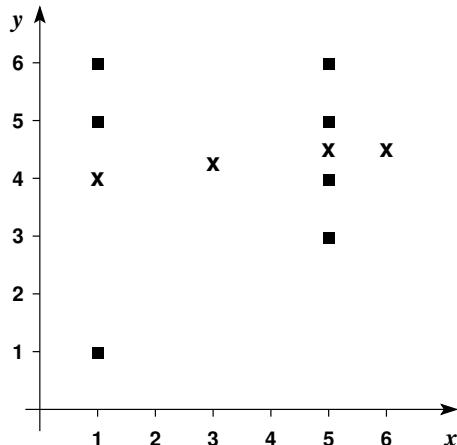
c. Avem $|5 - 1| = 4 > 3$ și $|5 - 5| = 0 \leq 3$, prin urmare vor fi luate în considerare doar valorile învățate pentru $x = 5$.

$$\hat{f}(5) = \frac{3 + 4 + 5 + 6}{4} = \frac{18}{4} = \frac{9}{2}.$$

d. Avem $|6 - 1| = 5 > 3$ și $|6 - 5| = 1 \leq 3$, prin urmare vor fi luate în considerare doar valorile învățate pentru $x = 5$, ca și în cazul precedent.

$$\hat{f}(6) = \hat{f}(5) = 4.5$$

Dacă vom plasa rezultatele de mai sus pe grafic și vom reprezenta punctele $(x, f(x))$ sub formă unor cruciulițe, vom obține figura alăturată.



9.

(Asupra folosirii algoritmului k -NN în spații (\mathbb{R}^p) de dimensiune (p) mare: un avertisment: „blestemul marilor dimensiuni“)

■ □ • CMU, 2010 fall, Aarti Singh, HW2, pr. 2.2

Considerăm punctele x_1, x_2, \dots, x_n distribuite în mod independent și uniform într-o sferă (notată cu B) care are raza egală cu unitatea⁴⁵⁹ și centrul în O , originea spațiului \mathbb{R}^p . Așadar, $B = \{x : \|x\|^2 \leq 1\} \subset \mathbb{R}^p$, unde $\|x\| = \sqrt{x \cdot x}$, iar operatorul · desemnează produsul scalar din \mathbb{R}^p .

În această problemă veți studia „mărimea“ vecinătății de tip 1-NN pentru originea O și cum anume variază ea în raport cu dimensiunea p . În acest fel, veți putea vedea care sunt dezavantajele folosirii algoritmului k -NN într-un spațiu de dimensiune mare.

Din punct de vedere formal, „mărimea“ menționată mai sus va fi identificată cu d^* , distanța de la O la cel mai apropiat vecin din mulțimea $\{x_1, x_2, \dots, x_n\}$:

$$d^* \stackrel{\text{not.}}{=} \min_{1 \leq i \leq n} \|x_i\|.$$

Observație: Din moment ce eșantionul $\{x_1, x_2, \dots, x_n\}$ este generat în mod aleatoriu, distanța d^* poate fi văzută ca fiind [produsă de către] o variabilă aleatoare.

a. În cazul particular $p = 1$, calculați expresia *funcției de distribuție cумулativă*⁴⁶⁰ a lui d^* (văzută ca variabilă aleatoare), și anume $P(d^* \leq t)$ pentru $t \in [0, 1]$.

b. Determinați expresia *funcției de distribuție cумултивă* (c.d.f.) a lui d^* în cazul general, adică pentru $p \in \{1, 2, 3, \dots\}$.

Sugestie: Puteți folosi următoarea formulă pentru volumul unei sfere de rază r din \mathbb{R}^p :

$$V_p(r) = \frac{(r\sqrt{\pi})^p}{\Gamma\left(\frac{p}{2} + 1\right)},$$

unde Γ reprezintă funcția Gamma a lui Euler, care are proprietățile:

⁴⁵⁹Termenul folosit în limba engleză pentru o astfel de sferă este *unit ball*.

⁴⁶⁰Engl., cumulative distribution function, c.d.f.

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}, \quad \Gamma(1) = 1, \quad \text{iar} \quad \Gamma(x+1) = x\Gamma(x) \quad \text{pentru } x > 0.$$
⁴⁶¹

c. Care este *mediana* variabilei aleatoare d^* (adică, valoarea lui t pentru care $P(d^* \leq t) = 1/2$)? Va trebui ca răspunsul să fie formulat în funcție de n și p (dimensiunea eșantionului și, respectiv, dimensiunea spațiului din care se face extragerea instanțelor, \mathbb{R}^p).

Pentru $n = 100$, alcătuiți un grafic cu valorile [funcției] mediane pentru $p = 1, 2, 3, \dots, 100$. Valorile lui p vor fi plasate pe axa Ox , iar valorile mediane pe axa Oy . Ce observați?

d. Folosind funcția de distribuție cumulativă (c.d.f.) de la punctul b , determinați cât de mare ar trebui să fie n (mărimea eșantionului) astfel încât

$$P(d^* \leq 0.5) \geq 0.9,$$

adică, cu probabilitate de cel puțin 9/10, distanța d^* de la originea O la cel mai apropiat vecin să fie mai mică decât 1/2 (adică, jumătate din distanța de la O la marginea sferei). Va trebui să formulați răspunsul ca expresie a unei funcții în raport cu variabila p .

Reprezentați grafic valorile acestei funcții pentru $p = 1, 2, \dots, 20$, plasând valorile lui p pe axa Ox și valorile funcției pe axa Oy . Ce observați?

Sugestie: Pentru $\ln(1-x)$, puteți face apel la dezvoltarea sa sub formă de *serie Taylor*:

$$\ln(1-x) = -\sum_{i=1}^{\infty} \frac{x^i}{i} \quad \text{pentru } -1 \leq x < 1.$$

e. În urma rezolvării punctelor de mai sus, ce puteți spune despre dezavantajele algoritmului k -NN în raport cu [diferitele valori posibile pentru] n și p ?

Răspuns:

a. Pentru $p = 1$, sfera de rază 1 este intervalul $[-1, 1]$, iar funcția de distribuție cumulativă va avea expresia:

$$F_{n,1}(t) \stackrel{\text{not.}}{=} P(d^* \leq t) = 1 - P(d^* > t) = 1 - P(\|x_i\| > t, i = 1, 2, \dots, n)$$

Tinem cont de presupozitia de independentă la generarea punctelor x_i , rezultă:

$$F_{n,1}(t) = 1 - \prod_{i=1}^n P(\|x_i\| > t) = 1 - (1-t)^n.$$

b. În cazul general, adică pentru p un număr natural oarecare nenul, fixat, vom exprima mai întâi $P(d^* \leq t)$ exact ca mai înainte:

$$\begin{aligned} F_{n,p}(t) \stackrel{\text{not.}}{=} P(d^* \leq t) &= 1 - P(d^* > t) = 1 - P(\|x_i\| > t, i = 1, 2, \dots, n) \\ &\stackrel{\text{indep. cdt.}}{=} 1 - \prod_{i=1}^n P(\|x_i\| > t). \end{aligned}$$

⁴⁶¹Se verifică ușor că pentru $p = 3$ se obține volumul sferei: $V_3(r) = \frac{(r\sqrt{\pi})^3}{\frac{3}{4}\sqrt{\pi}} = \frac{4\pi r^3}{3}$. Pentru demonstrarea proprietăților funcției Γ indicate mai sus, vedeti problema 31.b de la capitolul de *Fundamente*.

Apoi, ținând cont de presupozitia de uniformitate la generarea punctelor x_i și, folosind notația $V_p(t)$ pentru volumul sferei de rază t , obținem:

$$F_{n,p}(t) = 1 - \left(\frac{V_p(1) - V_p(t)}{V_p(1)} \right)^n = 1 - \left(1 - \frac{V_p(t)}{V_p(1)} \right)^n.$$

În sfârșit, folosind formula sugerată în enunț pentru V_p , rezultă imediat că $F_{n,p}(t) = 1 - (1 - t^p)^n$.

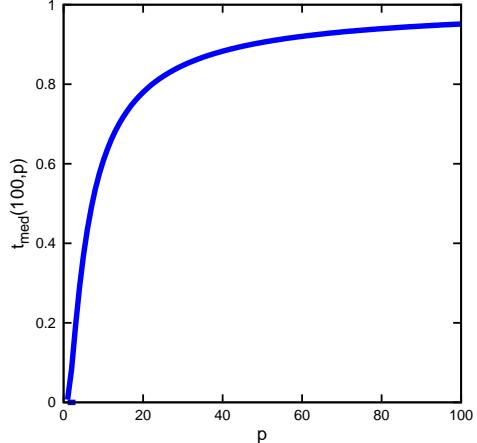
c. Pentru a afla valoarea mediană corespunzătoare variabilei aleatoare d^* , vom rezolva ecuația $P(d^* \leq t) = 1/2$ în funcție de t :

$$\begin{aligned} P(d^* \leq t) = \frac{1}{2} &\Leftrightarrow F_{n,p}(t) = \frac{1}{2} \stackrel{b}{\Leftrightarrow} 1 - (1 - t^p)^n = \frac{1}{2} \\ &\Leftrightarrow (1 - t^p)^n = \frac{1}{2} \Leftrightarrow 1 - t^p = \frac{1}{2^{1/n}} \\ &\Leftrightarrow t^p = 1 - \frac{1}{2^{1/n}} \end{aligned}$$

Prin urmare,

$$t_{med}(n, p) = \left(1 - \frac{1}{2^{1/n}} \right)^{1/p}.$$

Graficul funcției $t_{med}(100, p)$ pentru $p = 1, 2, \dots, 100$ este cel din figura alăturată. Se observă că sfera minimală care conține (cu probabilitate de $1/2$) cel mai apropiat vecin (un x_i , cu $i \in \{1, 2, \dots, n\}$) al originii O se lărgește foarte repede pe măsură ce p crește. Pentru valori ale lui p mai mari decât 10, majoritatea dintre cele 100 de instanțe de antrenament sunt mai aproape de conturul sferei de rază 1 decât de originea O (întrucât $P(D^* > t_{med}(n, p)) = 1/2$).



d. Putem scrie urmatorul sir de echivalențe:

$$\begin{aligned} P(d^* \leq 0.5) \geq 0.9 &\Leftrightarrow F_{n,p}(0.5) \geq 0.9 \Leftrightarrow \\ &\stackrel{b}{\Leftrightarrow} 1 - \left(1 - \frac{1}{2^p} \right)^n \geq \frac{9}{10} \Leftrightarrow \left(1 - \frac{1}{2^p} \right)^n \leq \frac{1}{10} \\ &\Leftrightarrow n \cdot \ln \left(1 - \frac{1}{2^p} \right) \leq -\ln 10 \\ &\Leftrightarrow n \geq \frac{\ln 10}{-\ln \left(1 - \frac{1}{2^p} \right)} \end{aligned}$$

Se poate vedea imediat că membrul din partea dreaptă a inegalității de mai sus tinde la $+\infty$ pentru $p \rightarrow \infty$. Este necesar să vedem cât de repede are

loc această tindere la infinit. Pentru aceasta, vom folosi descompunerea lui $-\ln(1 - 1/2^p)$ sub forma unei serii Taylor (luând $x = 1/2^p$):⁴⁶²

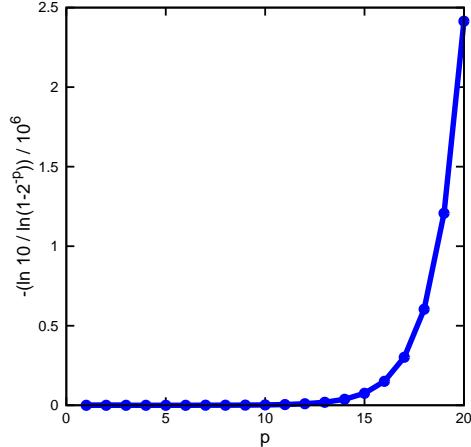
$$\begin{aligned} P(d^* \leq 0.5) \geq 0.9 &\Rightarrow n \geq (\ln 10) 2^p \frac{1}{1 + \frac{1}{2} \cdot \frac{1}{2^p} + \frac{1}{3} \cdot \frac{1}{2^{2p}} + \dots + \frac{1}{n} \frac{1}{2^{(n-1)p}} + \dots} \\ &\Rightarrow n > \frac{4}{3} 2^{p-1} \ln 10. \end{aligned}$$

Pentru obținerea ultimei inegalități de mai sus am ținut cont de faptul că inegalitatea $\frac{1}{n \cdot 2^{(n-1)p}} \leq \frac{1}{2^n} \Leftrightarrow 2^n \leq n \cdot 2^{(n-1)p}$ are loc pentru orice p și n cu $p \geq 1$ și $n \geq 2$,⁴⁶³ deci

$$\begin{aligned} &1 + \frac{1}{2} \cdot \frac{1}{2^p} + \frac{1}{3} \cdot \frac{1}{2^{2p}} + \dots + \frac{1}{n} \cdot \frac{1}{2^{(n-1)p}} + \dots \\ &\leq 1 + \frac{1}{2^2} + \frac{1}{2^3} + \dots + \frac{1}{2^n} + \dots \\ &< \left[1 + \frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \dots + \frac{1}{2^n} + \dots \right] - \frac{1}{2} \\ &\rightarrow \frac{1}{1 - \frac{1}{2}} - \frac{1}{2} = 2 - \frac{1}{2} = \frac{3}{2}. \end{aligned}$$

Observați că limita pe care tocmai am calculat-o ($3/2$) este o limită *superioară*, de aceea inegalitatea se păstrează până la final.

Așadar, rezultă că n crește în mod exponențial în raport cu p .⁴⁶⁴ Graficul pentru marginea inferioară dedusă mai sus ($-\ln 10 / \ln(1 - 2^{-p})$) este cel din figura alăturată.



Se observă că, într-adevăr, creșterea acestei margini inferioare (deci și a lui n) este exponențială.

e. Conform intuiției, clasificatorul k -NN se comportă bine atunci când instanța de test x_q este situată într-o vecinătate densă de instanțe de antrenament. Totuși, analiza teoretică de mai sus ne arată — în ipoteza că datele de antrenament sunt distribuite uniform — că pentru a ne asigura că punctul x_q

⁴⁶² $-\ln(1 - x) = x + \frac{1}{2}x^2 + \frac{1}{3}x^3 + \dots + \frac{1}{n}x^n + \dots$ pentru orice $x \in (-1, +1)$.

⁴⁶³ Demonstrația se poate face prin inducție după valorile lui p .

⁴⁶⁴ Mai detaliat: n , numărul de instanțe de antrenament necesare pentru a ne asigura că d^* (distanța până la cel mai apropiat vecin al originii O) este cu o probabilitate mare (și anume, $9/10$) mai mică decât 0.5 crește în mod exponențial în raport cu p .

are o vecinătate densă, numărul tuturor instanțelor de antrenament trebuie să crească exponențial în raport cu p , ceea ce nu este fezabil pentru valori mari ale lui p . (Parametrul p este dimensiunea spațiului în care se lucrează, adică numărul de trăsături ale instanțelor de antrenament și, respectiv, de test).

În consecință, pentru aplicațiile practice în care se folosesc date cu multe atrbute, este recomandat ca execuția algoritmului k -NN să fie precedată de efectuarea unei „selecții de trăsături“ (engl., feature selection).

10. (Algoritmul 1-NN [comparativ cu clasificatorul Bayes Optimal]: o margine superioară pentru eroarea medie asimptotică [la antrenare])

■ □ • CMU, 2005 spring, C. Guestrin, T. Mitchell, HW3, pr. 1

Un rezultat interesant obținut de Cover și Hart (1967) arată că, atunci când numărul datelor de antrenament tinde la infinit, iar datele de antrenament umplu spațiul în mod dens, rata medie a erorii produsă de către clasificatorul 1-NN este mărginită superior de dublul ratei medii a erorii pentru clasificatorul Bayes Comun (engl., Joint Bayes), care este numit adeseori și *Bayes Optimal* (engl., Optimal Bayes).

În acest exercițiu vi se va arăta, pas cu pas, cum se demonstrează rezultatul lui Cover și Hart în cazul particular al clasificării binare. Așadar, fie x_1, x_2, \dots instanțele de antrenament, iar y_1, y_2, \dots etichetele corespunzătoare, cu $y_i \in \{0, 1\}$. Putem considera instanțele x_i ca fiind puncte într-un spațiu euclidian d -dimensional.

Notăm $p_y(x) = P(X = x | Y = y)$ probabilitatea condiționată care reprezintă distribuția instanțelor din clasa y . Vom presupune că aceste probabilități condiționate sunt continue în raport cu variabila x și că $p_y(x) \in (0, 1)$ pentru orice x și orice y . Notăm cu θ probabilitatea ca un exemplu de antrenament selectat în mod aleatoriu să fie din clasa 1, așadar $\theta \stackrel{\text{not.}}{=} P(Y = 1)$. Din nou, presupunem că $\theta \in (0, 1)$.

a. Calculați probabilitatea ca o instanță oarecare x să aparțină clasei 1: $q(x) \stackrel{\text{not.}}{=} P(Y = 1 | X = x)$. Exprimăți $q(x)$ în funcție de $p_0(x)$, $p_1(x)$ și θ .

b. Clasificatorul Bayes Optimal asignează unui punct dat x cea mai probabilă clasă, $\text{argmax}_y P(Y = y | X = x)$. (Aceasta implică faptul că algoritmul Bayes Optimal maximizează probabilitatea clasificării corecte a tuturor datelor.) Considerând o instanță oarecare x , calculați probabilitatea ca x să fie clasificat greșit folosind clasificatorul Bayes Optimal, în funcție de probabilitatea $q(x) \stackrel{\text{not.}}{=} P(Y = 1 | X = x)$ care tocmai a fost calculată la punctul precedent. Veți desemna această nouă probabilitate cu $Error_{Bayes}(x)$.

c. Acum considerăm clasificatorul 1-NN. Aceasta îi asignează unei instanțe oarecare de test x eticheta celei mai apropiate instanțe de antrenament x' . Dată fiind o instanță de antrenament x (aleasă în mod arbitrar, dar fixată), calculați eroarea „așteptată“ (engl., expected error) produsă de către clasificatorul 1-NN, adică probabilitatea ca instanța x să fie clasificată greșit. Notați această probabilitate cu $Error_{1-NN}(x)$ și exprimați-o sub forma unei funcții definită în raport cu probabilitățile $q(x)$ și $q(x')$.

d. În *cazul asimptotic*, numărul de exemple de antrenament al fiecărei clase tinde la infinit, iar datele de antrenament umplu spațiul în mod dens. Atunci $q(x') \rightarrow q(x)$, unde, ca și mai sus, x' este cel mai apropiat vecin al lui x .⁴⁶⁵ Făcând această substituție în rezultatul obținut la punctul anterior, deduceți expresia *erorii asimptotice* pentru clasificatorul 1-NN în punctul x , adică $\lim_{x' \rightarrow x} Error_{1-NN}(x)$, în funcție de probabilitatea $q(x)$.

e. Arătați că eroarea asimptotică obținută la punctul d este mai mică decât dublul erorii clasificatorului Bayes Optimal obținută la punctul b , adică:

$$\lim_{x' \rightarrow x} Error_{1-NN}(x) \leq 2Error_{Bayes}(x).$$

În final, din această inegalitate deduceți relația corespunzătoare între ratele medii ale erorilor:⁴⁶⁶

$$E[\lim_{n \rightarrow \infty} Error_{1-NN}] \leq 2E[Error_{Bayes}].$$

Răspuns:

a. Conform enunțului, $p_1(x) \stackrel{not.}{=} P(X = x|Y = 1)$, $p_0(x) \stackrel{not.}{=} P(X = x|Y = 0)$, $q(x) \stackrel{not.}{=} P(Y = 1|X = x)$ și $\theta \stackrel{not.}{=} P(Y = 1)$. Putem calcula probabilitatea $q(x)$ în funcție de $p_1(x)$, $p_0(x)$ și θ folosind formula lui Bayes:

$$\begin{aligned} q(x) &\stackrel{Bayes}{=} \frac{P(X = x|Y = 1)P(Y = 1)}{P(X = x)} \\ &= \frac{P(X = x|Y = 1)P(Y = 1)}{P(X = x|Y = 1)P(Y = 1) + P(X = x|Y = 0)P(Y = 0)} \\ &= \frac{p_1(x)\theta}{p_1(x)\theta + p_0(x)(1 - \theta)} \end{aligned}$$

b. Este imediat faptul următor: probabilitatea ca algoritmul Bayes Optimal să greșească este $P(Y = 0|X = x)$ în cazul în care $P(Y = 1|X = x) \geq P(Y = 0|X = x)$, respectiv $P(Y = 1|X = x)$ atunci când $P(Y = 0|X = x) \geq P(Y = 1|X = x)$. Altfel spus,

$$\begin{aligned} Error_{Bayes}(x) &= \min\{P(Y = 0|X = x), P(Y = 1|X = x)\} \\ &= \min\{1 - q(x), q(x)\} = \begin{cases} q(x) \text{ în cazul } q(x) \in [0, 1/2] \\ 1 - q(x) \text{ în cazul } q(x) \in (1/2, 1]. \end{cases} \end{aligned}$$

c. Algoritmul 1-NN greșește atunci când instanța de antrenament x are eticheta 1, iar x' , cel mai apropiat vecin al lui x , are eticheta 0, sau invers, adică atunci când x are eticheta 0 iar x' are eticheta 1. În consecință, folosind algoritmul 1-NN, eroarea „asteptată“ la clasificarea lui x este:

$$\begin{aligned} Error_{1-NN}(x) &= P(Y = 1|X = x)P(Y = 0|X = x') + \\ &\quad P(Y = 0|X = x)P(Y = 1|X = x') \\ &= q(x)(1 - q(x')) + (1 - q(x))q(x'). \end{aligned}$$

⁴⁶⁵Adică, $P(Y = 1|X = x') \rightarrow P(Y = 1|X = x)$. Aceasta se justifică tînând cont de continuitatea lui $p_y(x) \stackrel{not.}{=} P(X = x|Y = y)$ care a fost asumată în enunț și, de asemenea, de rezultatul obținut la punctul a.

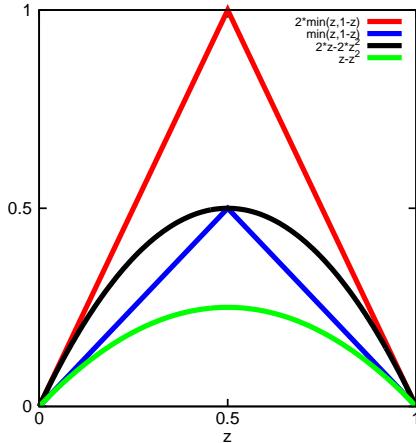
⁴⁶⁶Cititorul atent va remarcă faptul că în expresia de mai jos ($E[\lim_{n \rightarrow \infty} Error_{1-NN}]$) s-a înlocuit $\lim_{x \rightarrow x'}$ (folosită anterior) cu $\lim_{n \rightarrow \infty}$, pentru că se face trecerea la medii. Când $n \rightarrow \infty$, conform presupozиțiilor din enunț, rezultă $x \rightarrow x'$ pentru orice x .

d. Este imediat că $\lim_{x' \rightarrow x} Error_{1-NN}(x) = 2q(x)(1 - q(x))$.

e. Se arată imediat că $z - z^2 \leq z$ pentru $\forall z$, deci și pentru $z \in [0, 1/2]$, iar $z - z^2 \leq 1 - z$ pentru $\forall z$, deci și pentru $z \in [1/2, 1]$. Așadar, pentru orice x , vom avea:

$$q(x)(1 - q(x)) \leq \begin{cases} q(x) & \text{dacă } q(x) \in [0, 1/2] \\ 1 - q(x) & \text{dacă } q(x) \in (1/2, 1]. \end{cases}$$

Corelând cu rezultatul de la punctul b, obținem: $2q(x)(1 - q(x)) \leq 2Error_{Bayes}(x)$ pentru orice x .



Combinând acest rezultat cu egalitatea de la punctul d, rezultă că inegalitatea

$$\lim_{n \rightarrow \infty} Error_{1-NN}(x) = \lim_{x' \rightarrow x} Error_{1-NN}(x) \leq 2Error_{Bayes}(x)$$

este adevărată pentru orice x . Înmulțind ambii membri ai acestei inegalități cu $P(x)$ și însumând după toate valorile lui x — de fapt, integrând în raport cu x —, obținem:

$$E[\lim_{n \rightarrow \infty} Error_{1-NN}] \leq 2E[Error_{Bayes}].$$

Așadar, am demonstrat că media (sau: rata medie a) erorii asimptotice a algoritmului 1-NN este cel mult dublul mediei (sau: ratei medii a) erorii algoritmului Bayes Optimal.

Observația 1: Această margine a erorii asimptotice nu se păstrează și în cazul neasimptotic, unde numărul de exemple de antrenament este finit.

Observația 2: La fel, se poate arăta că $2z - 2z^2 \geq z$ pentru $\forall z \in [0, 1/2]$ și $2z - 2z^2 \geq 1 - z$ pentru $\forall z \in [1/2, 1]$. Luând din nou $z = q(x)$ și ținând cont de rezultatul de la punctul b, obținem că $2q(x)(1 - q(x)) \geq Error_{Bayes}(x)$, pentru orice x . Combinând această inegalitate cu egalitatea de la punctul d, rezultă o nouă inegalitate:

$$\lim_{n \rightarrow \infty} Error_{1-NN}(x) = \lim_{x' \rightarrow x} Error_{1-NN}(x) \geq Error_{Bayes}(x) \text{ pentru orice } x.$$

În final, trecând la medii, obținem următoarea inegalitate (care era de altfel de așteptat):

$$E[\lim_{n \rightarrow \infty} Error_{1-NN}] \geq E[Error_{Bayes}].$$

Observația 3 (preluată din *An Elementary Introduction to Statistical Learning Theory*, de Sanjeev Kulkarni și Gilbert Harman, 2011, pag. 69): În mod intuitiv, dacă mărim valoarea lui k , ar trebui ca eroarea medie a algoritmului k -NN să se reducă. Într-adevăr, în anumite condiții (dar nu în orice condiții!) se poate arăta că are loc următoarea inegalitate dublă:

$$E[Error_{Bayes}] \leq E[\lim_{n \rightarrow \infty} Error_{k-NN}] \leq \left(1 + \frac{1}{k}\right) E[Error_{Bayes}].$$

Este de remarcat faptul că există distribuții probabilistice ale datelor pentru care clasificatorul 1-NN se comportă mai bine decât k -NN pentru orice $k \neq 1$.

Observația 4 (preluată din aceeași lucrare, *An Elementary Introduction to Statistical Learning Theory*, citată mai sus): Dacă lucrăm cu k_n -NN, adică îl fixăm pe k în funcție de n (numărul instanțelor de antrenament), se poate demonstra că în cazul în care $\frac{k_n}{n} \rightarrow 0$ pentru $n \rightarrow \infty$ (de exemplu, $k_n = \sqrt{n}$), se obține:

$$E[\lim_{n \rightarrow \infty} \text{Error}_{k_n\text{-NN}}] = E[\text{Error}_{\text{Bayes}}].$$

Aceasta înseamnă că, la limită, algoritmul k_n -NN se comportă la fel de bine ca algoritmul Bayes Optimal!

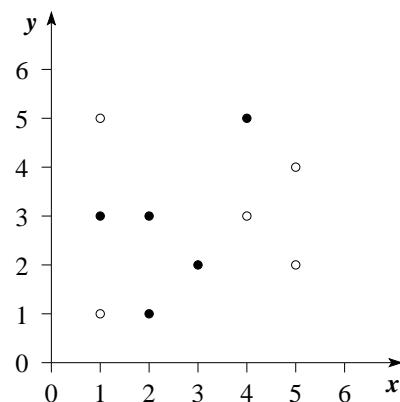
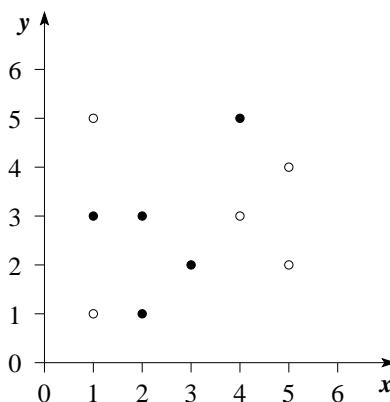
11.

(Comparație între algoritmii 1-NN și ID3:
zone de decizie și separatori decizionali)

*prelucrare de Liviu Ciortuz, după
■ • CMU, 2007 fall, Carlos Guestrin, HW2, pr 1.4*

Pe setul de date de mai jos desenați *granițele de decizie* și apoi hașurați *suprafețele de decizie* produse de

- algoritmul 1-NN (veți obține deci diagrama Voronoi);
- algoritmul ID3 extins cu capacitatea de a procesa atrbute cu valori continue.



Răspuns:

a. *Algoritmul 1-NN*:

Pentru a defini suprafețele de decizie în acest caz se procedează în felul următor:

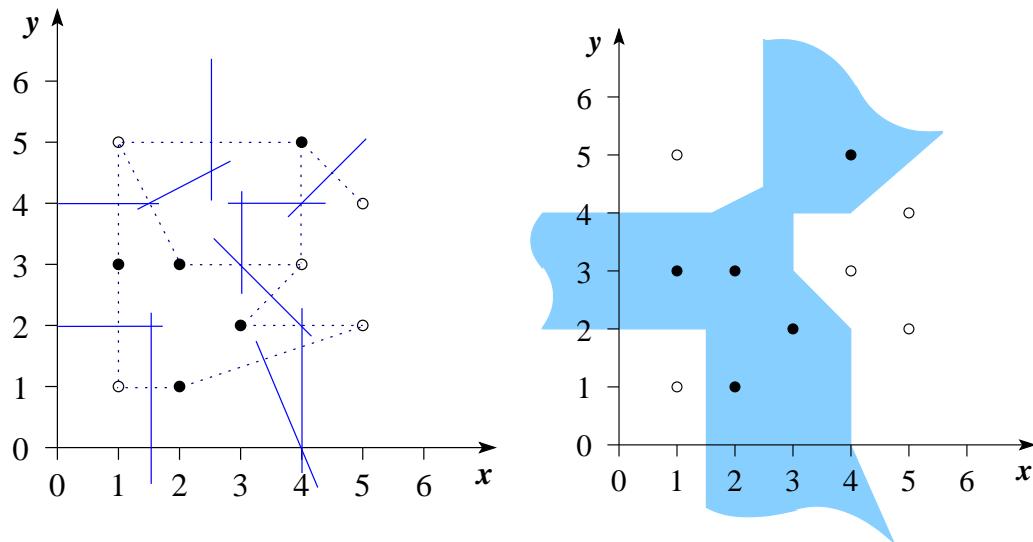
- se trasează mediatoarele segmentelor de dreapta determinate de perechi de puncte din setul de antrenament care sunt etichetate în mod diferit;
- se stabilesc intersecțiile acestor mediatoare; acest lucru este reprezentat în figura de mai jos, partea stângă;

- apoi se marchează pe aceste mediatoare acele segmente (determinate de intersecții) care determină zonele de decizie corespunzătoare clasificatorului 1-NN.

Observații:

1. De fapt, întrucât nu este necesar să se lucreze cu toate perechile de instanțe cu etichete diferite sunt relevante pentru clasificarea unei instanțe / zone, în timpul „execuției“ punctelor de mai sus este foarte util să se țină cont de următoarea regulă / euristică de *ghidare*: alegera perechilor de instanțe și apoi a segmentelor de pe mediatoare se va face urmărind delimitarea zonelor corespunzătoare instanțelor negative (○) de zonele corespunzătoare instanțelor pozitive (●).
2. Suplimentar, în jurul fiecărui punct de antrenament A se poate identifica câte o zonă [convexă] care va constitui mulțimea punctelor mai apropiate de A decât de oricare alt punct din setul de date de antrenament. (Toate punctele din această zonă convexă vor avea aceeași clasificare / etichetă ca și punctul A .) Aceasta este ușor de văzut pentru instanțele negative (○) din cazul de față.

În figura următoare, în partea dreaptă am hașurat zona corespunzătoare instanțelor pozitive (●).



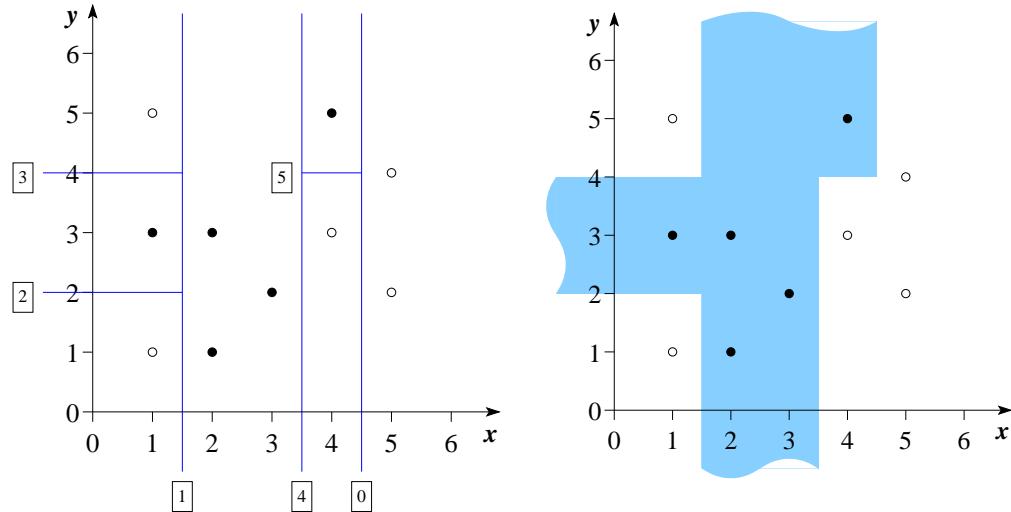
b. *Algoritmul ID3:*

Pentru construirea suprafețelor de decizie ale algoritmului ID3,

- mai întâi se determină valorile-prag pentru teste (adică, punctele de splitare de pe fiecare axă) și apoi se alege testul corespunzător nodului rădăcină din arborele ID3; se trasează o dreaptă prin punctul respectiv, paralelă cu cealaltă axă. În cazul nostru, testul din nodul rădăcină va fi $x < 4.5$.
- pentru fiecare test / split ulterior se trasează o semidreaptă (sau un segment de dreaptă) mărginit(ă) la un capăt de dreapta corespunzătoare nodului-părinte.

În figura de mai jos, în partea stângă am vizualizat toate testele / split-urile realizate de algoritm ID3 pentru a învăța complet arborele de decizie, iar

în partea dreaptă, ca și mai sus, am păstrat doar frontierele dintre zonele cu clasificări diferite.



Se observă că suprafețele de decizie determinate de cei doi algoritmi nu sunt identice, dar sunt totuși asemănătoare într-o anumită măsură (pentru că ambele sunt consistente cu datele de antrenament).

Observații:

3. Este de reținut faptul că suprafețele de decizie produse de algoritm ID3 nu sunt neapărat unic determinate, fiindcă sunt situații în care două teste diferite pot conduce la același câștig de informație. De exemplu, dacă în exercițiul nostru am avea de partionat la un moment dat mulțimea formată din instanțele de antrenament $(1, 1), (1, 3), (2, 1), (2, 3), (3, 2)$, atunci testele $x > 1.5$ și $y > 1.5$ ar produce același câștig de informație, iar suprafețele de decizie rezultate ar fi determinate (în mod diferit!) de ce anume alegem ca prim test.
4. De asemenea, trebuie să scoatem în evidență faptul că pragurile / spliturile care sunt calculate de algoritm ID3 pentru un același atribut continuu pot difera de la un nod de test la altul. De exemplu, la nodul rădăcină (nodul 0), atunci când se calculează câștigul de informație maxim, pentru atributul y se iau în calcul pragurile 1.5, 2.5, 3.5, și 4.5, în vreme ce la nodurile 2 și / sau 3 (vedeți figura de mai sus, partea stângă) se analizează pragurile 2 și 4, întrucât seturile / partitiile de instanțe asignate acestor noduri sunt diferite!

12. (Comparație între algoritmii 1-NN, ID3 cu atrbute continue și SVM: eroarea la antrenare, eroarea la CVLOO)

prelucrare de Liviu Ciortuz, după

• CMU, 2003 fall, T. Mitchell, A. Moore, midterm exam, pr. 5

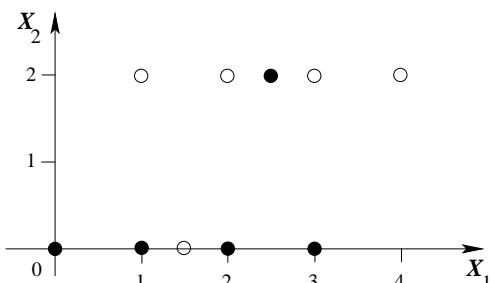
Folosind metoda 1-NN cu distanță euclidiană, învățăm un clasificator cu două valori pentru atributul de ieșire, $Y = 0$ și $Y = 1$, pornind de la datele de antrenament din tabelul de mai jos (X_1 și X_2 sunt atrbute de intrare).

- a. Care este eroarea la antrenare (exprimată ca număr de exemple clasificate eronat)?
- b. Care este eroarea la cross-validation folosind metoda “Leave-One-Out”?
- c. Răspundeți la întrebările de mai sus, considerând acum arbori de decizie cu atributuri numerice continue în locul metodei 1-NN.
- d. În sfârșit, răspundeți la întrebările *a* și *b*, considerând mașini cu vectori-suport în locul metodei 1-NN. (Se vor considera doar SVM-uri în cazul liniar cu margine “soft”, cu un parametru C suficient de mare pentru a minimiza numărul de instanțe de antrenament clasificate eronat.)

X_1	X_2	Y
0	0	1
1	0	1
2	0	1
2.5	2	1
3	0	1
1	2	0
1.5	0	0
2	2	0
3	2	0
4	2	0

Răspuns:

Reprezentarea datelor în planul euclidian este cea din figura alăturată.



- a. După cum am explicat și la exercițiul 2, întrucât datele de antrenament nu conțin inconistențe, eroarea la antrenare produsă de algoritmul 1-NN va fi 0.
- b. Comportamentul algoritmului la cross-validation cu metoda “Leave-One-Out” este cel descris în tabelul următor:

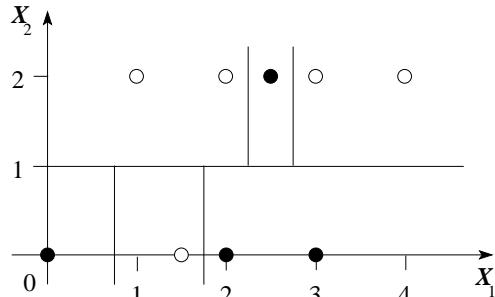
Data	Eticheta	Vecinătate	Clasificare la CVLOO	Eroare?
(0; 0)	1	(1; 0)	1	nu
(1; 0)	1	(1.5; 0)	0	da
(2; 0)	1	(1.5; 0)	0	da
(2, 5; 2)	1	(2; 2)/(3; 2)	0	da
(3; 0)	1	(2; 0)	1	nu
(1; 2)	0	(2; 2)	0	nu
(1, 5; 0)	0	(1; 0)/(2; 0)	1	da
(2; 2)	0	(2.5; 2)	1	da
(3; 2)	0	(2.5; 2)	1	da
(4; 2)	0	(3; 2)	0	nu

Deci în total avem 6 erori (din totalul de 10 instanțe), ceea ce indică faptul că algoritmul 1-NN este foarte puțin adecvat pentru acest gen de date.

- c. Eroarea la antrenare cu algoritmul ID3 pe acest set de date este 0, fiindcă instanțele sunt etichetate în mod consistent. Eroarea la CVLOO produsă de algoritmul ID3 pe același set de date se poate calcula destul de ușor, ținând cont că este suficient să determinăm granițele de decizie — deci și zonele de decizie — corespunzătoare celor două clase.⁴⁶⁷

⁴⁶⁷Datorită mixării puternice a datelor în raport cu separatorii de forma $X_1 = t$, la fiecare din cele 10 cazuri se va obține în nodul rădăcină același test: $X_2 < 1$.

De exemplu, atunci când se va elimina instanța $(1,0)$, se vor obține granițele de decizie ca în figura alăturată. Este imediat că *modelul* rezultat în acest caz va clasifica instanța $(1,0)$ în mod eronat.



Se constată că eroarea de tip CVLOO produsă de ID3 cu atrbute continue este de 6 (din totalul de 10) instanțe. Ba chiar, — este încă o coincidență! — ID3 produce la CVLOO exact aceleși erori (mai precis, el clasifică în mod eronat exact aceleși instanțe) ca și algoritmul 1-NN.

d. Se observă ușor că separatorul liniar care minimizează eroarea la antrenare este cel reprezentat în figura alăturată. Datele clasificate eronat de către acest separator sunt cele două puncte încercuite din figură; ele sunt de asemenea singurele puncte care generează eroare și la testare cu metoda CVLOO.

Mai concret, în cazul CVLOO folosind SVM, deoarece separatorul optimal este „susținut” de mai mulți vectori-suport pe fiecare parte, lipsa unuia singur dintre ei nu influențează cu nimic construirea separatorului. În fiecare caz în parte se va învăța ca separator dreapta paralelă cu Ox_1 care trece prin punctul $(0, 1)$.

Avem deci în ambele situații, adică atât la antrenare cât și la cross-validation cu metoda “Leave-One-Out”, (doar) două puncte clasificate eronat de către SVM.

Comparând cei trei clasificatori, 1-NN, ID3 cu atrbute continue și SVM, rezultă în mod clar că SVM este cel mai convenabil pe acest set de date (deși la antrenare SVM produce două erori, iar 1-NN și ID3 nicio eroare), fiindcă nu produce *overfitting* (aici!).

13. (Comparații între algoritmii 1-NN și ID3: Da sau nu?)

- ○ CMU, 2009 spring, Ziv Bar-Joseph, final exam, pr. 2.1
- CMU, 2015 fall, Z. Bar-Joseph, E. Xing, HW1, pr. 3.a

a. Este posibil să se construiască un arbore de decizie (având în fiecare nod intern teste de forma $x > a$, $x \leq b$, $y > c$, sau $y \leq d$, unde a , b , c și d sunt numere reale oarecare) care să producă la clasificare aceleși rezultate ca și algoritmul 1-NN folosind distanța euclidiană? Justificați răspunsul.

(Învățare rapidă / “eager” vs. învățare lentă / “lazy”; k -NN vs. ID3)

CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh, HW1, pr. 3.3

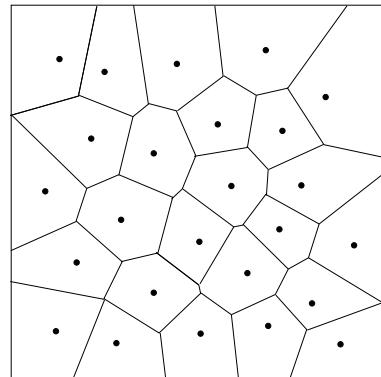
b. Algoritmul ID3 este o metodă de învățare de tip “batch”, care solicită ca toate datele de antrenament să-i fie puse la dispoziție pentru a putea elabora arborele de decizie. Așadar, în situația în care date de antrenament suplimentare ne sunt puse ulterior la dispoziție, acestea trebuie tratate cu atenție fiindcă ele pot modifica arborele de decizie rezultat în urma învățării.⁴⁶⁸ Algoritmul k -NN suferă și el de această problemă? Justificați.

Răspuns:

a. Nu.

Suprafețele de decizie pentru algoritmul 1-NN corespund diagramei Voronoi și nu sunt neapărat paralele cu axele de coordonate, după cum se observă în figura alăturată.

Suprafețele de decizie pentru un arbore de decizie cu atribută cu valori continue sunt întotdeauna paralele cu axele de coordonate, deoarece deciziile din fiecare nod sunt de forma $x > a$, $x \leq b$, $y > c$, sau $y \leq d$, $\forall a, b, c, d \in \mathbb{R}$.



Așadar, răspunsul este negativ pentru cazul general, deși există situații în care cei doi algoritmi produc exact aceeași clasificare.

b. Nu.

Răspunsul decurge din modul de lucru a algoritmului k -NN, care este un algoritm de învățare de tip “lazy”: k -NN estimează valoarea locală a unei funcții-target \hat{f} pentru una sau mai multe instanțe de test x_q . Valoarea $\hat{f}(x_q)$ nu depinde decât de cei mai apropiati vecini ai lui x_q ; celelalte instanțe de antrenament nu sunt necesare pentru calculul lui $\hat{f}(x_q)$. Evident, dacă pe măsură ce se acumulează noi date de antrenament se modifică și vecinătatea lui x_q , atunci se prea poate să se modifice și $\hat{f}(x_q)$. Însă în sine, algoritmul procedează exact la fel ca mai înainte. Se poate modifica doar outputul lui. Spre deosebire de algoritmul k -NN, la învățarea arborilor de decizie adăugarea de noi instanțe modifică în general și modelul / arborele rezultat, nu doar decizia pentru o instanță de test particulară.

14.

(Comparații între algoritmii 1-NN și ID3:
[o clasă de] seturi de date de antrenament pe care
cei doi clasificatori obțin rezultate identice)

prelucrare de Liviu Ciortuz, după

- ○ *CMU, 2009 spring, Ziv Bar-Joseph, final exam, pr. 2.1*
- CMU, 2015 fall, Z. Bar-Joseph, E. Xing, HW1, pr. 3.a*

Considerăm instanțele de antrenament $x_1, x_2, \dots, x_n \in \mathbb{R}^d$ având respectiv etichetele y_1, y_2, \dots, y_n .

Întrebare: Este oare posibil ca aplicând algoritmul ID3 cu atribută numerice continue să obținem aceleași rezultate la testare / generalizare (deci și aceleași

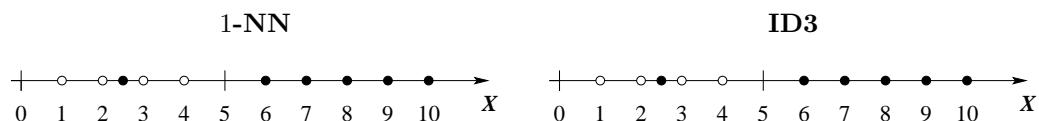
⁴⁶⁸Vedeți problema 16 de la capitolul *Arbore de decizie*.

zone de decizie) ca și cele produse de clasificatorul 1-NN folosind distanța euclidiană?

Veți aborda separat cazurile $d = 1$ și $d > 1$.

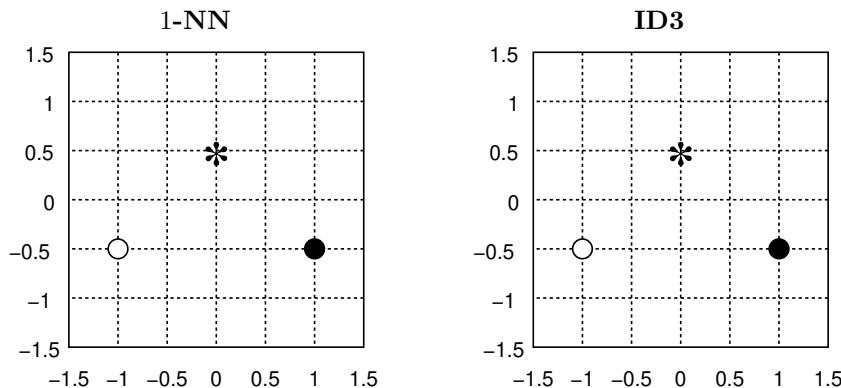
a. În cazul $d = 1$, pentru a înțelege mai bine cerința problemei, pentru setul de date de mai jos trasați diagrama Voronoi pe figura din stânga. Veți identifica în mod clar separatorii decizionali.

Pe figura din dreapta, desenați granițele de decizie și zonele de decizie determinate de algoritm ID3. (Vă readucem aminte convenția noastră: semnul \circ desemnează un exemplu negativ, iar semnul \bullet desemnează un exemplu pozitiv.)



Ce concluzie puteți trage referitor la calculul erorilor de tip CVLOO pentru cei doi clasificatori (1-NN și ID3 cu atrbute numerice continue) pe seturi de date din \mathbb{R} ?

b. În cazul $d > 1$, procedați similar pentru setul de date reprezentat mai jos, unde simbolii \circ , \bullet și $*$ denotă trei clase diferite. Adică, trasați diagrama Voronoi pe figura din stânga, iar pe figura din dreapta desenați granițele de decizie și zonele de decizie determinate de algoritmul ID3 (justificați riguros!).

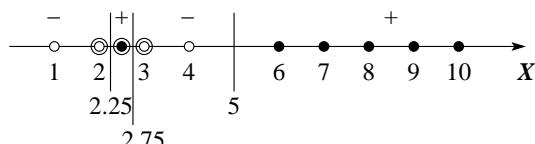


În cazul algoritmului ID3, rezultatul este unic determinat? Dacă da, explicați de ce. Dacă nu, arătați căte variante se pot obține în total.

c. Ce răspuns puteți formula acum referitor la *Întrebarea* de mai sus, din enunt? Încercați să *generalizați*, pornind de la exemplele de la punctele a și b.

Răspuns:

a. Atât pentru 1-NN cât și pentru ID3 zonele de decizie ai separatorii decizionali sunt conform figurii următoare:



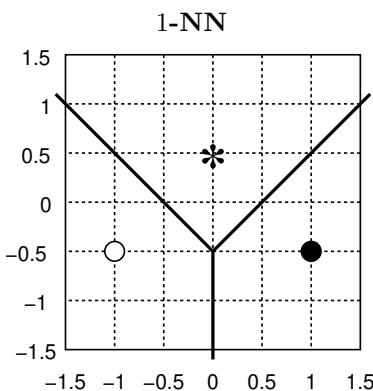
În această figură am încercuit exemplele care conduc la eroare la cross-validarea de tip “Leave-One-Out”, atât pentru algoritmul 1-NN cât și pentru algoritmul ID3.

Se constată că întotdeauna, pe orice set de exemple *consistente* din \mathbb{R} (deci cu un singur atribut de intrare, care este numeric și continuu), atât algoritmul 1-NN cât și algoritmul ID3 produc aceleași rezultate la testare / generalizare.

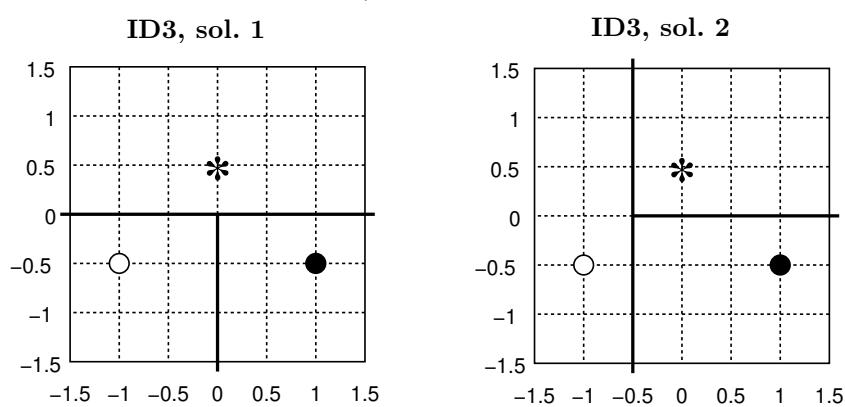
Explicația ține de folosirea distanței euclidiene la algoritmul 1-NN și respectiv modul cum se stabilesc pragurile de splitare de către algoritmul ID3 (și anume, la jumătatea distanței dintre două instanțe consecutive pe axa reală, care au etichete diferite).

În consecință, pe seturi de date *consistente* din \mathbb{R} , cei doi algoritmi produc exact aceleași zone de decizie și aceiași separatori decizionali, precum și aceleași erori la CVLOO.

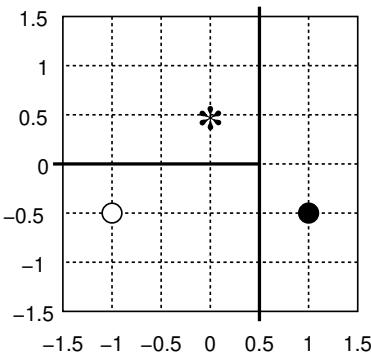
b. Diagrama Voronoi produsă de algoritmul 1-NN este cea din figura următoare:



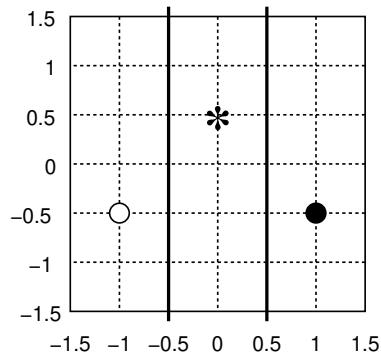
În cazul algoritmului ID3, split-urile sunt $x = -0.5$, $x = +0.5$, și $y = 0$, iar compașii de decizie care se formează au toți același tip de structură. (De exemplu: unul dintre ei are partitia $[1, -1+, 1^*]$ în rădăcină, iar descendenții lui au partitii $[1-, 0+, 0^*]$ și respectiv $[0-, 1+, 1^*]$. Ceilalți compași de decizie sunt ușor de „vizualizat“.) Asta face să se obțină mai multe soluții (a se vedea mai jos). Toate aceste soluții comportă zone de decizie diferite de cele determinate de algoritmul 1-NN (vedeți diagrama Voronoi de mai sus).



ID3, sol. 3



ID3, sol. 4



c. Spre deosebire de cazul $d = 1$, atunci când $d > 1$ se constată că în general algoritmul 1-NN și algoritmul ID3 pot produce la testare / generalizare rezultate diferite (deci și zone de decizie diferite). Există însă și cazuri în care rezultatele celor doi algoritmi sunt identice.

15.

(1-NN cu mapare cu RBF: Adevărat sau Fals?)

■ • ○ CMU, 2003 fall, T. Mitchell, A. Moore, final exam, pr. 7.f

Algoritmul 1-NN folosind distanța euclidiană neponderată este capabil să obțină rezultate mai bune dacă în prealabil intrările sale sunt mapate într-un „spațiu de trăsături“ folosind o funcție-nucleu cu baza radială (RBF).⁴⁶⁹ Adevarat sau Fals?

Răspuns:**Fals.**

Fie $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^n$ funcția de mapare în spațiul de trăsături, astfel încât să avem $K(x, y) \stackrel{\text{not.}}{=} e^{-\frac{\|x-y\|^2}{2\sigma^2}} = \phi(x) \cdot \phi(y)$, $\forall x, y \in \mathbb{R}^d$. (\mathbb{R}^d reprezintă spațiul inițial, \mathbb{R}^n spațiul de trăsături în care se face maparea, iar $e^{-\frac{\|x-y\|^2}{2\sigma^2}}$ este funcția-nucleu cu baza radială.) Avem:

$$\begin{aligned} \|\phi(x) - \phi(y)\|^2 &= (\phi(x) - \phi(y)) \cdot (\phi(x) - \phi(y)) \\ &= \phi(x) \cdot \phi(x) + \phi(y) \cdot \phi(y) - 2\phi(x) \cdot \phi(y) = e^{-\frac{\|x-x\|^2}{2\sigma^2}} + e^{-\frac{\|y-y\|^2}{2\sigma^2}} - 2e^{-\frac{\|x-y\|^2}{2\sigma^2}} \\ &= e^0 + e^0 - 2e^{-\frac{\|x-y\|^2}{2\sigma^2}} = 2 - 2e^{-\frac{\|x-y\|^2}{2\sigma^2}} = 2 - 2K(x, y). \end{aligned}$$

Prin urmare, pentru orice $x, x_i, x_j \in \mathbb{R}^d$ vom avea:

$$\begin{aligned} \|\phi(x) - \phi(x_i)\| \leq \|\phi(x) - \phi(x_j)\| &\Leftrightarrow \|\phi(x) - \phi(x_i)\|^2 \leq \|\phi(x) - \phi(x_j)\|^2 \Leftrightarrow \\ 2 - 2K(x, x_i) \leq 2 - 2K(x, x_j) &\Leftrightarrow K(x, x_i) \geq K(x, x_j) \Leftrightarrow e^{-\frac{\|x-x_i\|^2}{2\sigma^2}} \geq e^{-\frac{\|x-x_j\|^2}{2\sigma^2}} \Leftrightarrow \\ -\frac{\|x-x_i\|^2}{2\sigma^2} \geq -\frac{\|x-x_j\|^2}{2\sigma^2} &\Leftrightarrow \|x-x_i\|^2 \leq \|x-x_j\|^2 \Leftrightarrow \|x-x_i\| \leq \|x-x_j\|. \end{aligned}$$

⁴⁶⁹Vedeți problema 74 de la capitolul de *Fundamente*.

Cu alte cuvinte, dacă o instanță de test x are drept cel mai apropiat vecin punctul x_i în spațiul inițial, acest lucru rămâne valabil și în spațiul de trăsături. Așadar, decizia algoritmului 1-NN este identică în ambele spații. Aceeași concluzie este valabilă și pentru k -NN.

Observație: Este însă posibil ca folosind ponderarea sau alte măsuri de distanță (decât cea euclidiană) kernel-izarea să funcționeze cu succes pentru k -NN.

3.2 Învățare bazată pe memorare — Probleme propuse

16.

(Algoritmul k -NN: acuratețe; comparație cu un simplu clasificator aleator)

prelucrare de Liviu Ciortuz, după CMU, 2014 fall, W. Cohen, Z. Bar-Joseph, HW1, pr. 5.bc

- a. Enunțați [succint] *regula de decizie* a algoritmului k -NN pentru o instanță de test x_q .

Care este *bias-ul inductiv* al algoritmului k -NN?

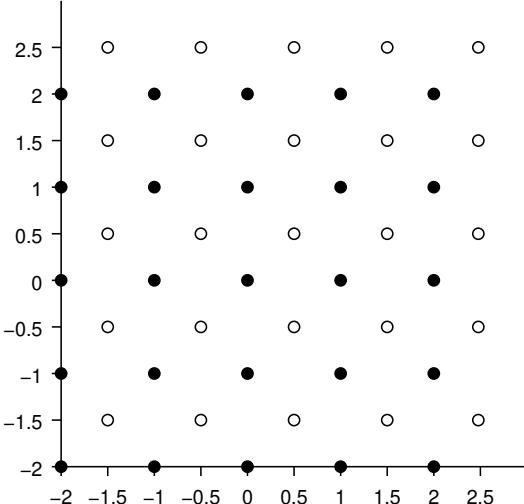
În continuare veți lucra pe datele din figura de mai jos. Veți aplica algoritmul k -NN, folosind distanța euclidiană.

k -NN funcționează bine atunci când instanțele dintr-o aceeași clasă sunt placeate într-una sau mai multe zone din spațiu relativ bine delimitate, fără întrepătrunderi puternice.

Obiectivul nostru acum este să analizăm ce se întâmplă atunci când datele sunt puternic mixate. Rezultatele pe care le veți obține la calculul erorilor vor fi exprimate sub formă de numere [fracționare] din intervalul $[0, 1]$.

Observație importantă:

În cazul în care există două sau mai multe instanțe situate exact pe „marginea“ [adică, pe conturul circular al] k -NN-vecinătății asociate instanței de clasificat, se va considera că toate aceste instanțe aparțin respectivei vecinătăți, iar fiecare dintre ele dispune de un vot întreg.



- b. Pentru $k = 1$, calculați eroarea la antrenare și eroarea la cross-validation cu metoda “leave-one-out” (CVLOO).

Ce puteți spune comparând cele două rezultate? (Care este legătura între *bias-ul inductiv* al lui k -NN și puterea de generalizare a lui 1-NN pe astfel de date?)

- c. Pentru $k = 2$, calculați eroarea la cross-validation cu metoda “leave-one-out” (CVLOO).

- d. Considerăm $k = 50$. (Remarcați faptul că în total în setul nostru de date sunt 50 de instanțe.) De această dată, vom impune ca algoritmul k -NN să ia decizia în mod *probabilist*. Aceasta înseamnă că dacă în vecinătatea k -NN a unei instanțe de test există n vecini pozitivi și m vecini negativi, atunci algoritmul k -NN va returna (pentru instanța respectivă) decizia + cu

probabilitatea $n/(n+m)$ și decizia – cu probabilitatea $m/(n+m)$. În consecință, pentru întreg setul de date vom putea calcula o *eroare medie*.

Calculați *eroarea medie* la antrenare pentru algoritmul 50-NN pe datele de mai sus. Cunoașteți o metodă de clasificare foarte simplă care obține pe aceste date rezultate la fel de bune / proaste precum 50-NN?

17.

(Algoritmul k -NN: acuratețe; comparații pentru diferite valori ale lui k)

• ○ CMU, 2014 fall, W. Cohen, Z. Bar-Joseph, HW1, pr. 5.a

Considerăm două clase, notează cu C_1 și C_2 , în spațiul euclidian bidimensional. Datele din clasa C_1 sunt distribuite în mod uniform într-un cerc de rază r . Datele din clasa C_2 sunt distribuite în mod uniform într-un alt cerc de rază r . (*Observație:* Numărul de date din cele două clase nu este neapărat același.) Centrele celor două cercuri sunt situate la o distanță strict mai mare decât $4r$.

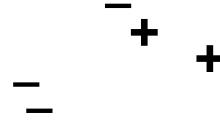
Arătați că este posibil ca [la antrenare] acuratețea algoritmului 1-NN aplicat pe aceste date să fie *strict* mai mare decât acuratețea algoritmului k -NN, pentru un anumit număr întreg $k \geq 3$, impar, ales în mod convenabil.

18.

(Algoritmul 1-NN: calculul erorii la CVLOO)

* CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh, midterm exam, pr. 1.7

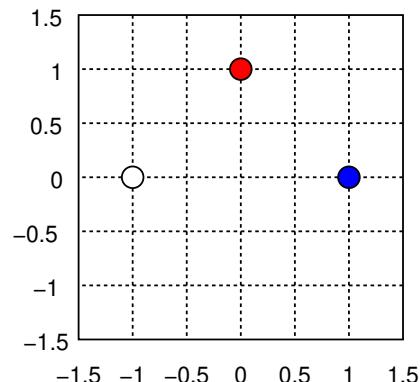
Care este eroarea clasificatorului 1-NN la cross-validation de tip “Leave-One-Out” pe setul de date alăturat?



19.

(Algoritmul 1-NN: granițe / suprafețe de decizie)

• CMU, 2013 fall, W. Cohen, E. Xing, final exam, pr. 3.6



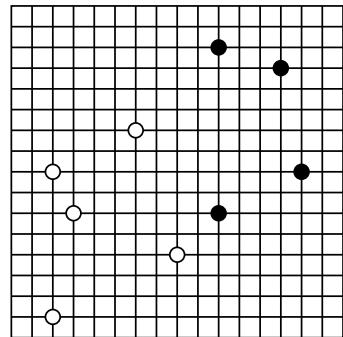
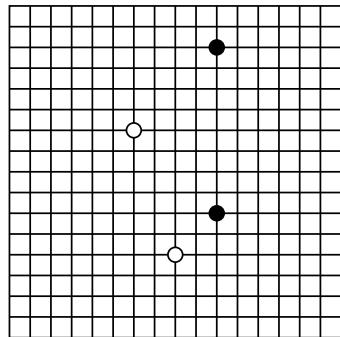
Trasați granițele de decizie (engl., *decision boundaries*) produse de clasificatorul 1-NN la aplicarea pe datele din figura alăturată. Diversele culori ale punctelor reprezintă clase diferite.

20.

(Algoritmul 1-NN: granițe / suprafețe de decizie)

* CMU, 2008 fall, Eric Xing, HW1, pr. 3

În fiecare din figurile următoare se dau câteva puncte / instanțe în spațiul bidimensional, care sunt etichetate cu • (instanțe pozitive) sau ○ (instanțe negative). Indicați în fiecare caz granițele / suprafețele de decizie pentru algoritmul 1-NN presupunând că se folosește distanța euclidiană.

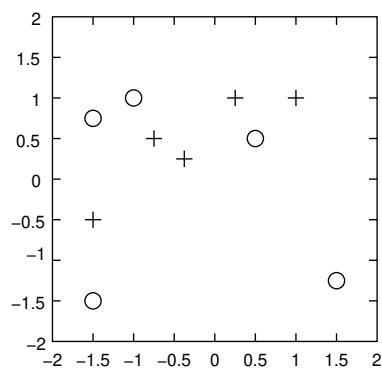
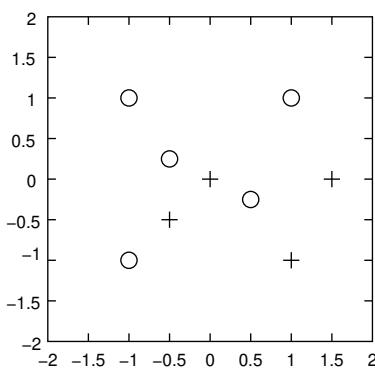
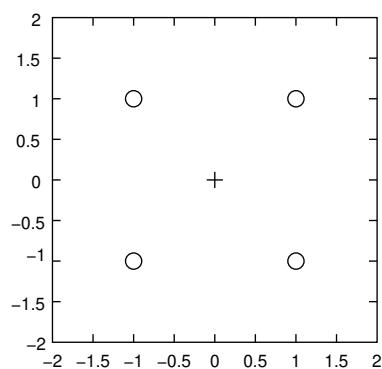
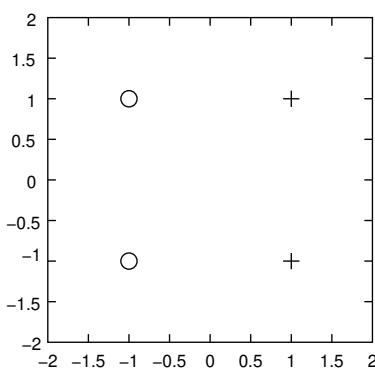


21.

(Algoritmul 1-NN: diagrame Voronoi)

* CMU, 2010 fall, Ziv Bar-Joseph, HW1, pr. 3.1

Desenați suprafețele de decizie pentru clasificatorul 1-NN pentru fiecare dintre seturile de date din figurile de mai jos. Folosiți distanța euclidiană. Hașurați fin zonele corespunzătoare clasei +.



22.

(Algoritmul k -NN: diagrama Voronoi, eroarea la CVLOO; comparație pentru diferite valori ale lui k)

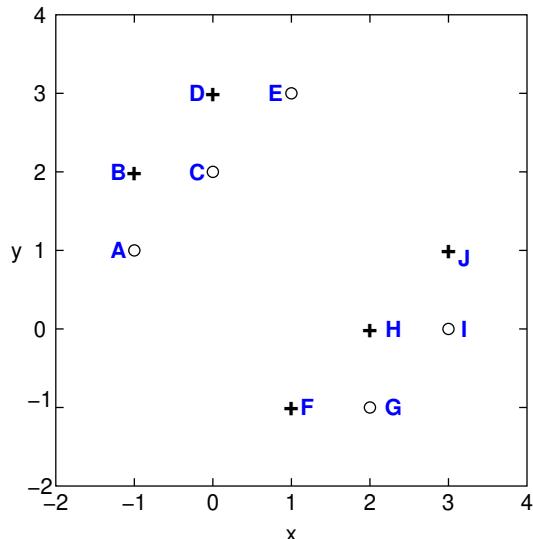
• prelucrare de L. Ciortuz, după
CMU, 2012 spring, Ziv Bar-Joseph, midterm exam, pr. 2

La acest exercițiu veți aplica algoritmul k -NN folosind distanța euclidiană pe setul de date din figura de mai jos. Fiecare punct aparține la una din două clase, desemnate cu $+$ și respectiv \circ .

- a. Pentru $k = 1$, trasați diagrama Voronoi și hașurați zona / zonele de decizie corespunzătoare etichetei $+$.

- b. Care este eroarea la cross-validation cu metoda "Leave-One-Out" (CVLOO) dacă se folosește algoritmul 1-NN?

- c. Care dintre următoarele valori ale lui k va conduce la o valoare minimă a erorii de tip CVLOO: 3, 5, 7 sau 9? Comentați succint rezultatul.



Indicații:

1. k -NN-vecinătățile vor fi construite în manieră *inclusivă*.⁴⁷⁰ Vă cerem(!) să puneti în evidență toate cazurile de acest tip. Vedeți, spre exemplu, liniile 2 și 3 din tabelul de mai jos, coloana 1-NN vecinătăților.
2. În caz de *paritate la voturi* (dar doar în acest caz!), se va considera că se aplică (în mod intuitiv) ponderarea distanțelor în sensul prezentat la curs.
3. Dacă veți ști să exploatați *simetriile*, veți avea mult mai puțin de elaborat la nivel de detaliu!
4. Pentru conveniență, puteți completa tabelul de mai jos. Din cauza spațiului restrâns, vă sugerăm ca la rubrica *Vecinătate*, în dreptul fiecărui punct (A, B, \dots, J), pentru fiecare din valorile 3, 5 și 9 ale lui k să menționați doar punctele care constituie *extensia* de la vecinătatea 1-NN la vecinătatea 3-NN și aşa mai departe.

⁴⁷⁰ Adică, dacă notăm cu

- x_1, x_2, \dots, x_n instanțele de antrenament,
- x o instanță oarecare căreia î se aplică la un moment dat procedura de cross-validation LOO cu algoritmul k -NN (unde k este fixat),
- $d(x, x_{i_1}) \leq d(x, x_{i_2}) \leq \dots \leq d(x, x_{i_n})$ secvența ordonată a distanțelor de la x la fiecare din instanțele de antrenament,

și există $l > 0$ astfel încât $x_{i_k} = x_{i_{k+1}} = \dots = x_{i_{k+l}} < x_{i_{k+l+1}}$ sau $k + l = n$, atunci în k -NN vecinătatea lui x vor fi incluse toate instanțele $x_{i_{k+1}}, \dots, x_{i_{k+l}}$.

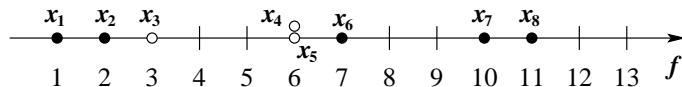
Data	Et.	Vecinătate					Clasif. la CVLOO					Eroare? (da/nu)				
		1-	3-	5-	7-	9-NN	1-	3-	5-	7-	9-NN	1-	3-	5-	7-	9-NN
A	o	B														
B	+	A C														
C	o	B D														
...																

23. (Diagrame Voronoi pentru dataset-uri din \mathbb{R} , folosind 1-NN și 5-NN; calcularea erorii la CVLOO pentru algoritmul 1-NN; comparație cu algoritmul ID3 cu atrbute numerice continue)

prelucrare de L. Ciortuz, după
• o * MIT, ML 6036 course, Review Material Problems, ex. 37

Comentariu: În acest exercițiu, spre deosebire de toate celelalte cazuri când am construit diagrame Voronoi (întotdeauna pentru algoritmul 1-NN și seturi de date din \mathbb{R}^2), vom construi astfel de diagrame pentru seturi de date din \mathbb{R} , mai întâi pentru algoritmul 1-NN și apoi pentru algoritmul 5-NN. (Absolut similar se poate proceda și pentru alte valori ale lui $k \neq 1$, atunci când datele de antrenament sunt din \mathbb{R} .)

În desenul de mai jos este reprezentat un set de antrenament care conține 8 instanțe, fiecare dintre ele având doar o trăsătură (engl., feature), notată cu f .



Remarcați faptul că sunt două instanțe pentru care valoarea trăsăturii f este aceeași, și anume 6. Aceste două instanțe sunt reprezentate prin două simboluri \circ , situate unul deasupra celuilalt, dar de fapt ele ar fi trebuit să fie reprezentate ca două simboluri \circ suprapuse (unul peste celălalt), întrucât aceste instanțe au exact aceeași valoare pentru trăsătura f .

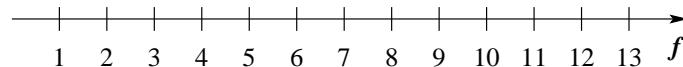
Vă readucem aminte *convenția* noastră de notare: simbolul \bullet desemnează instanțe pozitive (+), iar simbolul \circ instanțe negative (-).

- a. La acest punct al problemei veți folosi algoritmul 1-NN.

Convenții / reguli:

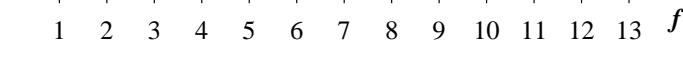
1. În cazul în care există două sau mai multe instanțe situate exact pe „marginea“ [adică, pe conturul circular al] k -NN-vecinătății asociate instanței de clasificat, se va considera că toate aceste instanțe aparțin respectivei vecinătăți, iar fiecare dintre ele dispune de un vot întreg.
2. În caz de paritate de voturi, veți alege eticheta instanței de antrenament care este situată la distanță minimă către stânga față de instanța de test respectivă. (Distanța minimă este 0 atunci când instanța de test coincide cu o instanță de antrenament!)
3. Aceste reguli vor fi aplicate și la punctele următoare.

Vă cerem ca pe linia de mai jos să hașurați — și, bineînțeles, să delimitați prin separatori decizionali — zonele în care algoritmul 1-NN va prezice semnul +, dat fiind setul de date de antrenament din figura de mai sus.



În prealabil, vă cerem să stabiliți cu ce etichetă vor fi clasificate instanțele de test $f = 2.5$ și $f = 6.5$.⁴⁷¹ Justificați.

- b. Dacă faceți cross-validation folosind metoda “leave-one-out” pe acest set de date, în conjuncție cu algoritmul 1-NN, cât va fi eroarea produsă? Veți arăta în mod clar — alcătuind un *tabel* care să conțină 1-NN-vecinătățile respective — cum anume ați ajuns la rezultatul pe care l-ați indicat.
- c. Similar cu punctul a, însă aici veți folosi algoritmul 5-NN.⁴⁷²



Indicație: Justificați în mod riguros rezultatul, referindu-vă la diverse intervale de valori (sau valori particulare) ale lui f :

Cazul 1: $f \in (-\infty, 4)$: ...

Cazul 2: ...

...

În prealabil, vă cerem să stabiliți cu ce etichetă vor fi clasificate instanțele de test $f = 4$, $f = 6$ și $f = 7$.⁴⁷³ Justificați.

Observație: Se poate demonstra faptul că în cazul seturilor de date de antrenament $x_1, x_2, \dots, x_n \subset \mathbb{R}$ (adică, atunci când se lucrează pe axa reală) k -NN-vecinătățile nu sunt discontinue, adică: dat fiind un punct de test oarecare $x_q \in \mathbb{R}$, dacă instanțele de antrenament x_i și x_j , cu $x_i \leq x_j$, se află în k -NN-vecinătatea lui x_q , atunci orice instanță de antrenament x_l care satisfac proprietatea $x_i \leq x_l \leq x_j$ aparține și ea respectivei k -NN-vecinătăți a lui x_q .⁴⁷⁴

- d. Construiți — în mod riguros — arborele de decizie corespunzător setului de date din enunț, considerând f ca fiind atribut numeric continuu. Comparați rezultatul obținut de data aceasta cu rezultatele care au fost produse de clasificatorii 1-NN și 5-NN.

⁴⁷¹ Acestea sunt [toate] cazurile când 1-NN-vecinătatea unui punct de test de pe axa reală conține instanțe având etichete diferite.

⁴⁷² Fără să aplicați 5-NN pentru instanțele de test $f = 2.5$ și $f = 6.5$.

⁴⁷³ Acestea sunt [toate] cazurile când 5-NN-vecinătatea unui punct de test de pe axa reală conține mai mult de 5 instanțe de antrenament.

⁴⁷⁴ O proprietate similară are loc în cazul clusterizării de instanțe pe axa reală. Vedeti problema 47.c de la capitolul *Clusterizare*.

24.

(Algoritmul k -NN:CVLOO: comparație pentru diferite valori ale lui k ;
eroarea la antrenare: comparație cu alți clasificatori) CMU, 2010 fall, Aarti Singh, midterm exam, pr. 2

a. Care dintre clasificatorii de mai jos realizează o eroare de tip CVLOO (Leave-One-Out Cross-validation) mai mare pe setul de date alăturat?

 1-NN 3-NN+ + - -
- - - -

b. Considerăm setul de date din figura alăturată. Care dintre clasificatorii de mai jos obține / obțin eroare nulă la antrenare pe acest set de date?

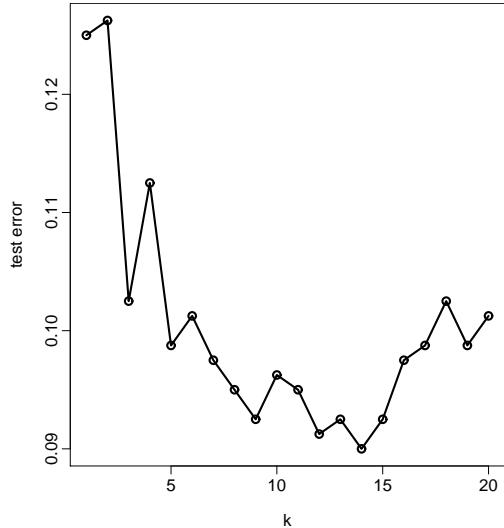
○ +
+ ○ arborii de decizie ID3 de adâncime 2 regresia logistică clasificatorul 3-NN SVM (cu nucleu pătratic)

25.

(Algoritmul k -NN: alegerea valorii convenabile pentru k)prelucrare de Liviu Ciortuz, după
CMU, (?) spring, ML course 10-701, HW1, pr. 5

Pe un anumit set de date format din date de antrenament, date de validare și date de test, după ce a fost antrenat algoritmul k -NN pentru diferite valori ale lui k , rezultatele obținute la validare au fost reprezentate în graficul care urmează.

Care este — în conformitate cu aceste rezultate — valoarea optimă care trebuie aleasă pentru k , în vederea folosirii ulterioare pe datele de test? Justificați alegerea făcută.



26.

(Algoritmul k -NN: întrebări de ordin calitativ) CMU, 2012 spring, Ziv Bar-Joseph, HW1, pr. 3

Știm că la aplicarea algoritmului k -NN, clasificarea unei instanțe date se face pe baza votului majoritar obținut în „vecinătatea“ instanței respective. Pre-supunem că se dau două clase de instanțe, fiecare clasă având $n/2$ puncte, întrepătrunse într-o anumită măsură, într-un spațiu bidimensional.

a. Descrieți ce se întâmplă cu eroarea la antrenare (folosind toate datele disponibile) când numărul k al vecinilor considerați variază de la n la 1.

- b. Schițați grafic cum anume ar evoluă *eroarea la generalizare* (de exemplu, reținând o parte din date pentru testare) atunci când k variază. Explicați modul în care ați raționat.
- c. Propuneți o metodă de determinare a unei valori adecvate pentru k .
- d. La folosirea algoritmului k -NN, odată ce s-a stabilit valoarea lui k , toți cei mai apropiati k vecini ai punctului de clasificat au ponderi egale (adică, aceeași importanță) la stabilirea etichetei respectivului punct. Sugerați o modificare a algoritmului k -NN care elimină această limitare.
- e. Dați două motive pentru care este de preferat să nu folosim algoritmul k -NN atunci când dimensiunea spațiului datelor de intrare este mare.

27.

(Compararea clasificatorilor 1-NN și Bayes Optimal:
o margine superioară mai bună
pentru *rata medie a erorii asymptotice* a lui 1-NN)

* Liviu Ciortuz, 2014, bazat pe un rezultat din
■ “An Elementary Introduction to Statistical Learning Theory”,
S. Kulkarni, G. Harman, 2011, pag. 68-69

La problema 10 am demonstrat că *rata medie a erorii* clasificatorului 1-NN este mărginită asymptotic⁴⁷⁵ de dublul ratei medii a erorii clasificatorului Bayes Optimal.

Arătați că — în aceleasi condiții ca la problema 10 — se poate obține o margine chiar mai bună:

$$E\left[\lim_{n \rightarrow \infty} Error_{1-NN}\right] \leq 2E[Error_{Bayes}](1 - E[Error_{Bayes}]).$$

28.

(Adevărat sau Fals?)

○ CMU, 2010 fall, Ziv Bar-Joseph, midterm, pr. 1.bc

Care dintre următoarele afirmații sunt adevărate pentru clasificatorii k -NN?
(Justificați pe scurt răspunsul, în dreptul fiecărui punct.)

- a. Acuratețea la antrenare crește pe măsură ce crește valoarea lui k .
- b. Granița de decizie este mai netedă (engl., smoother) pe măsură ce valoarea lui k scade.
- c. k -NN nu necesită o procedură explicită de antrenare.
- d. Granița de decizie este liniară.
- e. Este posibil ca un clasificator binar 1-NN să clasifice întotdeauna orice instanță de test ca fiind pozitivă, chiar dacă în setul de date de antrenament există instanțe negative.

⁴⁷⁵Adică, atunci când $n \rightarrow \infty$, unde n este numărul de instanțe de antrenament.

29. (Întrebări calitative despre design-ul unor experimente din Învățarea Automată: OK ori ...problematic?)
 • o CMU, 2009, Geoff Gordon, midterm exam, pr. 3

Fiecare din punctele de mai jos prezintă pe scurt design-ul unui experiment practic de învățare automată. Analizați fiecare din aceste cazuri, indicând apoi dacă respectivul experiment este *ok* ori *problematic* (încercuiți varianta pe care o alegeti). Dacă este *problematic*, identificați TOATE defectele [de concepție ale] design-ului respectiv.

- a. O echipă de proiectare raportează o eroare mică la antrenare și susține că metoda folosită este bună.

Ok

Problematic

- b. O echipă de proiectare susține că este un mare succes faptul că a obținut 98% acuratețe la antrenare pentru un task de clasificare binară care are următorul specific: unul din cele două cazuri se întâlnește foarte rar comparativ cu celălalt caz. (O astfel de problemă o constituie, de exemplu, identificarea tranzacțiilor bancare frauduloase.) Datele lor au constat din 50 de exemple pozitive și 4950 de exemple negative.

Ok

Problematic

- c. O echipă de proiectare și-a împărțit datele de care dispune în date de antrenament și date de test. Folosind datele de antrenament, ei au construit un *model* de clasificare caracterizat de anumiți *parametri*. Apoi, făcând *cross-validation*, au ales cea mai bună setare a parametrilor. La final, au raportat eroarea obținută pe *datele de test*.

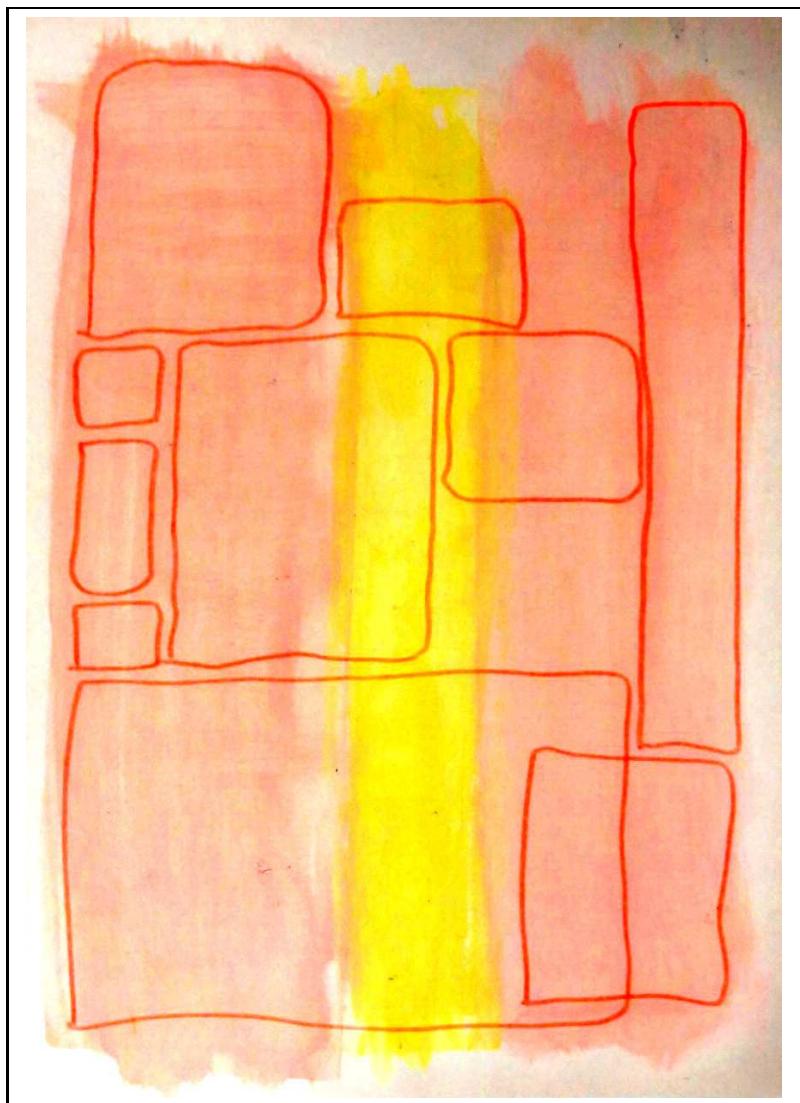
Ok

Problematic

- d. O echipă de proiectare a efectuat o procedură de *selecție a atributelor* (engl., features) pe toate datele și apoi a redus setul mare de atrbute la un set mai mic. După aceea, membrii echipei au împărțit datele în date de test și date de antrenament. Au construit *modelul* de clasificare pe datele de antrenament folosind mai multe setări ale parametrilor modelului, și au raportat cea mai bună eroare la *testare* pe care au obținut-o.

Ok

Problematic



© M. Romanică

4 Arbori de decizie

Sumar

Noțiuni preliminare

- partiție a unei multimi: ex. 93 de la cap. *Fundamente*;
- proprietăți elementare ale funcției logaritm; formule uzuale pentru calcule cu logaritmi;
- Elemente de *teoria informației* (vedeți secțiunea corespunzătoare din cap. *Fundamente*, în special ex. 55):
 - entropie, definiție: T. Mitchell, *Machine Learning*, 1997 (desemnată în continuare simplu prin *cartea ML*), pag. 57; ex. 2.a, ex. 41.a, ex. 36.a;
 - entropie condițională specifică: ex. 14.a;
 - entropie condițională medie: ex. 2.cd, ex. 36.c;
 - câștig de informație (definiție: *cartea ML*, pag. 58): ex. 2.cd, ex. 5.a, ex. 34, ex. 41.b, ex. 36.e;
- *arbori de decizie*, văzuți ca structură de date: ex. 1, ex. 32 și, respectiv, ca program în logica propozițiilor: ex. 2.e, ex. 40.bc, ex. 44.c;
 - (P0) expresivitatea arborilor de decizie cu privire la *funcții booleene*: ex. 33;
- *spațiu de versiuni* pentru un concept (de învățat): ex. 1, ex. 3, ex. 32, ex. 39.

Algoritmul ID3

- pseudo-cod, versiune simplificată:

```
create the root node;
assign all training examples to the root;
```

Main loop:

1. $A \leftarrow$ the “best” decision attribute for the current *node*;
2. for each value of the attribute *A*, create a new descendant of the *node*;
3. sort the training examples to the leaf nodes;
4. if the training examples are perfectly classified, then STOP;
else iterate over the new leaf nodes;

pseudo-cod, versiune completă: *cartea ML*, pag. 56;

- *bias-ul inductiv*: *ibidem*, pag. 63-64;
- exemple simple de aplicare: ex. 2, ex. 3, ex. 4.a, ex. 5, ex. 37, ex. 38, ex. 39.a, ex. 40, ex. 41, ex. 42;
- ID3 ca algoritm *per se*:
 - este un *algoritm de căutare*;
 - *spațiul de căutare* — mulțimea tuturor arborilor de decizie care se pot construi cu atributele de intrare în *nodurile de test* și cu valorile atributului de ieșire în *nodurile de decizie* — este de dimensiune exponențială în raport cu numărul de atrbute: ex. 1, ex. 3, ex. 32, ex. 39;
 - ID3 are ca *obiectiv* căutarea unui arbore / *model* care i. să explice cât mai

bine datele (în particular, atunci când datele sunt *consistente*, modelul trebuie să fie *consistent* cu acestea), *ii.* să fie cât mai *compact*, din motive de *eficiență* la generalizare / testare și *iii.* în final să aibă o [cât mai] bună putere de *generalizare*;⁴⁷⁶

- de tip *divide-et-impera* (\Rightarrow “*Iterative Dichotomizer*”), recursiv;
- *1-step look-ahead*;
- ID3 ar putea fi văzut și ca algoritm de *optimizare*;⁴⁷⁷
- *greedy* \Rightarrow nu garantează obținerea soluției optime d.p.v. al numărului de niveluri / noduri:
ex. 4, ex. 21.a, ex. 40 (vs. ex. 39.b, ex. 3.b), ex. 50;
- complexitate de timp, cf. *Weka book*:⁴⁷⁸
la antrenare, în anumite condiții: $\mathcal{O}(d m \log m)$; la testare $\mathcal{O}(d)$, unde d este numărul de atrbute, iar m este numărul de exemple;
- ID3 ca algoritm de învățare automată:
 - *bias-ul inductiv* al algoritmului ID3:
[dorim ca modelul să aibă structură ierarhică, să fie compatibil / consistent cu datele de antrenament dacă acestea sunt consistente (adică, necontradictorii), iar] arborele produs de ID3 trebuie să aibă un număr cât mai mic de niveluri / noduri;
 - algoritm de învățare de tip “*eager*”;
 - *analiza erorilor*:
la antrenare: ex. 7.a, ex. 10.a, ex. 46;⁴⁷⁹ [acuratețe la antrenare: ex. 6;]
la validare: ex. 44.d;
la *n-fold cross-validation*
la *cross-validation leave-one-out* (CVLOO): ex. 10.b, ex. 49.bc;⁴⁸⁰
 - *robustețe la „zgomote“ și overfitting*: ex. 10, ex. 21.bc, ex. 49, ex. 57.c;⁴⁸¹
 - *zone de decizie și granițe de separare / decizie* pentru arbori de decizie cu variabile continue: ex. 10, ex. 48, ex. 49, ex. 50.

Extensii / variante ale algoritmului ID3

- *atribute cu valori continue*: ex. 10-12, ex. 14.c, ex. 47-51; cap. *Învățare bazată pe memorare*, ex. 11.b;
- *atribute discrete cu multe valori*: ex. 13, ex. 52;
- *atribute cu valori nespecificate / absente* pentru unele instanțe;
- *atribute cu diferite costuri asociate*: ex. 14;
- *reducerea caracterului “eager”* al învățării: ex. 16;

⁴⁷⁶LC: Alternativ, putem spune că algoritmul ID3 produce o *structură* de tip *ierarhie* (arbore) între diferențe *partiționări* ale setului de instanțe de antrenament, această ierarhie fiind generată pe baza *corespondenței* dintre atributul de *iesire* și atrbutele de *intrare*, care sunt adăugate la model câte unul pe rând.

⁴⁷⁷LC: Am putea să-l interpretăm pe ID3 ca fiind un algoritm care caută între diferențele *distribuții probabiliste* discrete care pot fi definite pe setul de date de antrenament una care să satisfacă cerința de *ierarhizare*, și pentru care entropia să fie *minimală* (vedeți proprietatea de structuralitate de la ex. 62 de la cap. *Fundamente*). Cerința ca arborele ID3 să fie *minimal* (ca număr de niveluri / noduri) este însă mai importantă, mai practică și mai ușor de înțeles.

⁴⁷⁸“Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations”, Ian Witten, Eibe Frank (3rd ed.), Morgan Kaufmann Publishers, 2011.

⁴⁷⁹De asemenea, ex. 4.ab de la capitolul *Clasificare bayesiană*.

⁴⁸⁰De asemenea, ex. 4.ab de la capitolul *Clasificare bayesiană*.

⁴⁸¹De asemenea, ex. 4.ab de la capitolul *Clasificare bayesiană*.

- reducerea caracterului “greedy” al învățării:
IG cu “2-step look-ahead”: ex. 17, ex. 18;
- folosirea altor *măsuri de „impuritate“* în locul câștigului de informație:
Gini Impurity, Misclassification Impurity: ex. 15;
- reducerea *overfitting-ului*:
reduced-error pruning (folosind un set de date de validare): cartea ML, pag. 69-71; A. Cornuéjols, L. Miclet, *Apprentissage artificiel*, 2010, pag. 418-421;
rule post-pruning: cartea ML, pag. 71-72; ex. 54
top-down vs. bottom-up pruning, folosind un criteriu bazat pe câștigul de informație: ex. 19, ex. 53;
pruning folosind testul statistic χ^2 : ex. 20, ex. 55.

Proprietăți ale arborilor ID3

- (P1) arborele produs de algoritm ID3 este *consistent* (adică, în concordanță) cu datele de antrenament, dacă acestea sunt *consistente* (adică, necontradictori). Altfel spus, *eroarea la antrenare* produsă de algoritm ID3 pe orice set de *date consistente* este 0: ex. 2-4, ex. 37, ex. 39;
- (P2) arborele produs de algoritm ID3 *nu* este în mod neapărat *unic*: ex. 3, ex. 39;
- (P3) arborele ID3 *nu* este neapărat *optimal* (ca nr. de noduri / niveluri): ex. 4, ex. 21.a, ex. 40, ex. 50;
- (P4) influența *atributelor identice* și, respectiv, a *instanțelor multiple* asupra arborelui ID3: ex. 8;
- (P5) o *margine superioară* pentru *eroarea la antrenare* a algoritmului ID3, în funcție de numărul de valori ale variabilei de ieșire): ex. 7.b;
- (P6) o aproximare simplă a numărului de *instanțe greșit clasificate* din totalul de M instanțe care au fost asignate la un nod frunză al unui arbore ID3, cu ajutorul entropiei (H) nodului respectiv: ex. 45;
- (P7) *granițele de separare / decizie* pentru arborii ID3 cu *attribute de intrare continue* sunt întotdeauna paralele cu axele de coordonate: ex. 10, ex. 48, ex. 49, ex. 50, precum și cap. *Învățare bazată pe memorare*, ex. 11.b;
- (P8) *Zonele de decizie* produse de algoritm ID3 nu sunt în mod neapărat unice, fiindcă arboarele de decizie creat de ID3 nu este determinat în mod unic (vedeți proprietatea (P2)).

Observație: Următoarele trei proprietăți se referă la arbori de decizie în general, nu doar la arbori ID3.

- (P9) *adâncimea maximă* a unui arbore de decizie, când attributele de intrare sunt *categoriale*: numărul de attribute: ex. 56.c;
- (P10) o *margine superioară* pentru *adâncimea* unui arbore de decizie când attributele de intrare sunt continue, iar datele de antrenament sunt (ne)separabile liniar: ex. 11;
- (P11) o *margine superioară* pentru numărul de *noduri-frunză* dintr-un arbore de decizie, în funcție de numărul de exemple și de numărul de attribute de intrare, atunci când acestea (attributele de intrare) sunt binare: ex. 9.

- Alte metode de învățare automată bazate pe arbori: arbori de regresie (CART).

Învățare automată de tip ansamblist folosind arbori de decizie: Algoritmul AdaBoost

- Noțiuni preliminare:
distribuție de probabilitate discretă, factor de normalizare pentru o distribuție de probabilitate, ipoteze „slabe“ (engl., weak hypothesis), compas de decizie (engl., decision stump), prag de separare (engl., threshold split) pentru un compas de decizie, prag exterior de separare (engl., outside threshold split), eroare ponderată la antrenare (engl., weighted training error), vot majoritar ponderat (engl., weighted majority vote), overfitting, ansambluri de clasificatori (vedeți ex. 70), funcții de cost / pierdere (engl., loss function) (vedeți ex. 23.b și ex. 29);
- pseudo-codul algoritmului AdaBoost: ex. 22,⁴⁸²
proprietăți de bază: ex. 22;
alte proprietăți, precum și două margini superioare pentru eroarea la antrenare: ex. 23.a-d;
convergența algoritmului, precum și o condiție suficientă pentru învățabilitate empirică γ -slabă: ex. 23.e, ex. 67;
- exemple de aplicare: ex. 24, 59, 60, 61, 62, 63, 64 și 65.
- AdaBoost ca algoritm *per se*:
algoritm *iterativ*,
algoritm *de căutare* (spațiul de căutare este multimea combinațiilor liniare care se pot construi peste clasa de ipoteze „slabe“ considerate),
algoritm *greedy* (dacă la fiecare iterație se alege cea mai bună ipoteză „slabă“),
algoritm de *optimizare secvențială* (minimizează o margine superioară pentru eroarea la antrenare);
varianta de pseudo-cod de la pr. 26 (și mai ales varianta de AdaBoost generalizat de la pr. 29) evidențiază faptul că AdaBoost face *optimizare pe coordonate* (engl., coordinate descent), în raport cu componentele h_t și respectiv α_t ;
- *învățabilitate empirică γ -slabă*:
definiție: ex. 23.e
exemplificarea unor cazuri când nu există *garanție* pentru învățabilitate γ -slabă: ex. 25, 66;
- AdaBoost ca algoritm de *optimizare secvențială* în raport cu funcția de cost / „pierdere“ negativ-exponențială: ex. 26;
- marginea de votare: ex. 27, 28 și 68;
proprietăți ale marginilor de votare:
 $Margin_k(i) \in [-1, +1]$
 $Margin_k(x_i) = y_i \bar{f}_k(x_i)$ unde $\bar{f}_k(x_i) \stackrel{\text{not.}}{=} \sum_{t=1}^k \bar{\alpha}_t h_t(x_i)$
 x_i este corect clasificat la iterarea $k \Leftrightarrow Margin_k(i) \geq 0$
 $Margin_k(x_i) > Margin_k(x_j) \Leftrightarrow D_{k+1}(i) < D_{k+1}(j)$;
- *selectarea trăsăturilor* folosind AdaBoost; aplicare la clasificarea de documente: ex. 69;

⁴⁸²Vedeți pseudo-codul algoritmului AdaBoost.M1 din cartea „Probabilistic machine learning: An introduction“ de Kevin Murphy, pag. 610, MIT Press, 2022. (Cf. <https://probml.github.io/pml-book/book1.html>, accesat la 24.02.2022.)

- AdaBoost și overfitting: ex. 72;
- o margine superioară pentru eroarea la generalizare produsă de algoritmul AdaBoost: ex. 30.
- AdaBoost poate folosi “confidence-rated classifiers” ca ipoteze “slabe” (în locul compașilor de decizie): ex. 73;
- o variantă generalizată a algoritmului AdaBoost [în raport cu funcția de cost]:⁴⁸³ ex. 29 și ex. 75;
- AdaBoost multi-class: ex. 76;
- recapitulare (întrebări cu răspuns *adevărat / fals*): ex. 31 și 77.

• **Proprietăți ale algoritmului AdaBoost:**

(P0) AdaBoost poate produce rezultate diferite atunci când are posibilitatea să aleagă între două sau mai multe [cele mai bune] ipoteze „slabe”: ex. 24, 59;

(P1) $\varepsilon_i > \varepsilon_j \Leftrightarrow \alpha_i < \alpha_j$ (consecință imediată din relația $\alpha_t = \ln \sqrt{\frac{1 - \varepsilon_t}{\varepsilon_t}}$): ex. 22.v;

(P2) Din relația de definiție pentru distribuția D_{t+1} rezultă $Z_t = e^{-\alpha_t} \cdot (1 - \varepsilon_t) + e^{\alpha_t} \cdot \varepsilon_t = 2\sqrt{\varepsilon_t(1 - \varepsilon_t)}$ (ex. 22.i și 22.ii) și $\varepsilon_t \in (0, 1/2) \Rightarrow Z_t \in (0, 1)$ (ex. 22.iii).

(P3) O consecință din relația (227) și (P2): $D_{t+1}(i) = \begin{cases} \frac{1}{2\varepsilon_t} D_t(i), & i \in M \\ \frac{1}{2(1 - \varepsilon_t)} D_t(i), & i \in C \end{cases}$ (ex. 22.iv);

(P4) $\text{err}_{D_{t+1}}(h_t) = \frac{1}{2}$ (ex. 22.vi);

ca o consecință, rezultă că ipoteza h_t nu poate fi reselectată și la iterată $t+1$; ea poate fi reselectată la o iterată ulterioară;

(P5) $D_{t+1}(i) = \frac{1}{m \prod_{t'=1}^t Z_{t'}} e^{-y_i f_t(x_i)}$, unde $f_t(x_i) \stackrel{\text{def.}}{=} \sum_{t'=1}^t \alpha_{t'} h_{t'}(x_i)$ (ex. 23.a).

Produsul $y_i f_t(x_i)$ se numește *margine algebraică*;

(P6) $\text{err}_S(H_t) \leq \prod_{t'=1}^t Z_{t'}$, adică eroarea la antrenare comisă de ipoteza combinată produsă de AdaBoost este majorată de produsul factorilor de normalizare: ex. 23.b;

(P7) AdaBoost nu optimizează în mod direct $\text{err}_S(H_t)$, ci marginea sa superioară, $\prod_{t'=1}^t Z_{t'}$; optimizarea se face în mod secvențial (greedy): la iterată t se minimizează valoarea lui Z_t ca funcție de α_t , ceea ce conduce la $\alpha_t = \ln \sqrt{\frac{1 - \varepsilon_t}{\varepsilon_t}}$ (ex. 23.c);

(P8) $\text{err}_S(H_t)$ nu neapărat descrește de la o iterată la alta; în schimb, descresc marginile sale superioare: $\prod_{t'=1}^t Z_{t'}$ și $\exp(-\sum_{t'=1}^t \gamma_{t'}^2)$ (ex. 23.be);

(P9) O proprietate de monotonie a *costurilor minime* negativ-exponențiale determinate de către AdaBoost la iterării succesive: $J_t^* \leq J_{t-1}^*$, $\forall t \geq 1$, unde $J_t \stackrel{\text{not.}}{=} \frac{1}{m} \sum_{i=1}^m \exp(-y_i \sum_{t'=1}^t \alpha_{t'} h_{t'}(x_i))$, iar $J_t^* \stackrel{\text{not.}}{=} \min_{h \in \mathcal{H}, \alpha' \in \mathbb{R}_+} J_t(h, \alpha')$, $\forall t \in \mathbb{N}^*$:

⁴⁸³Pentru AdaBoost văzut ca instanță a unui algoritm general de învățare ansamblistă bazat pe minimizarea secvențială a unei funcții de cost / pierdere, vedeți ex. 26.d și ex. 74.

ex 26.e;

(P10) O condiție suficientă pentru învățabilitate γ -slabă, bazată pe marginea de votare: la fiecare iterație a algoritmului AdaBoost, media marginilor de votare ale instanțelor de antrenament în raport cu distribuția D_t să fie de cel puțin 2γ : ex. 28, ex. 68.d;

(P11) Orice mulțime formată din m instanțe din \mathbb{R} care sunt etichetate în mod consistent poate fi corect clasificată de către o combinație liniară formată din cel mult m compași de decizie: ex. 70.ac;

(P11') Orice mulțime de instanțe distincte [și etichetate] din \mathbb{R} este γ -slab învățabilă cu ajutorul compașilor de decizie: ex. 71.d.

- Alte metode de învățare ansamblistă bazate pe arbori de decizie: Bagging, Random Forests.

4.1 Arbori de decizie — Probleme rezolvate

4.1.1 Algoritmul ID3

1. (Arbori de decizie; optimalitate, relativ la numărul de noduri)

Reprezentați arborele / arborii de decizie care are / au numărul minim de noduri posibile și corespunde / corespund funcției booleene $(\neg A \vee B) \wedge \neg(C \wedge A)$ definită peste atributele booleene A, B și C .

Răspuns:

Vom determina arborele de decizie optimal (ca număr de noduri) parcurgând în mod *exhaustiv spațiul de versiuni*, adică multimea tuturor arborilor de decizie (construiți cu variabilele A, B și C) care sunt *consistenți* cu funcția dată. Așadar, vom examina ce se întâmplă când în nodul rădăcină se pun pe rând atributele A, B și respectiv C .

Notăm cu X funcția $(\neg A \vee B) \wedge \neg(C \wedge A)$, ale cărei valori sunt date în tabelul alăturat.

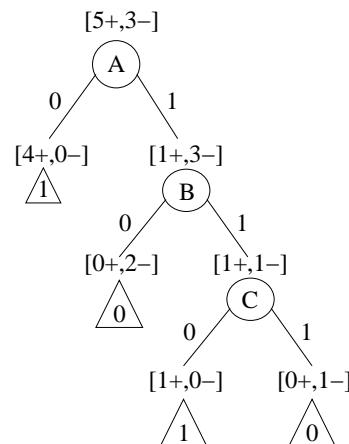
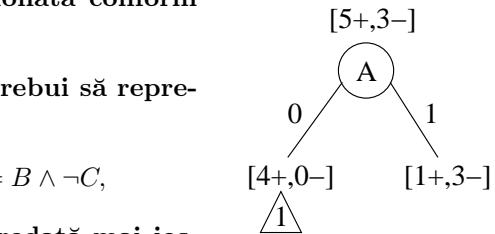
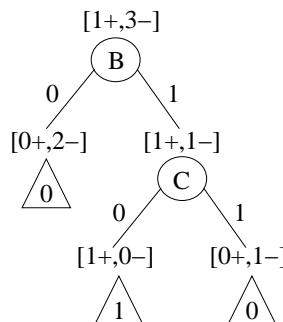
A	B	C	X
0	0	0	1
0	0	1	1
0	1	0	1
0	1	1	1
1	0	0	0
1	0	1	0
1	1	0	1
1	1	1	0

- Cazul 1:* Dacă în nodul rădăcină se plasează atributul A , multimea de exemple va fi re-partiționată conform reprezentării alăturate.

Subarborele drept al acestui arbore va trebui să reprezinte arborele de decizie pentru funcția

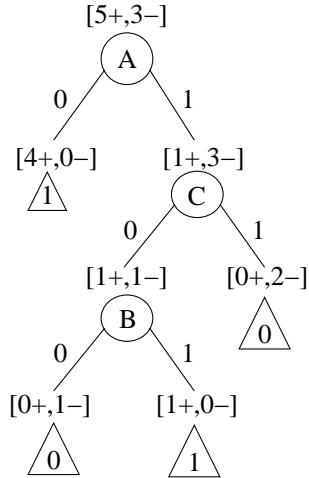
$$X_1 = X[A/1] = (\neg 1 \vee B) \wedge \neg(C \wedge 1) = B \wedge \neg C,$$

pentru care o reprezentare optimă este redată mai jos, în partea stângă:

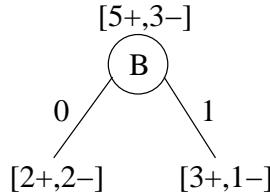


Prin urmare, un arbore optim (ca număr de noduri) care are variabila A în nodul rădăcină este cel reprezentat mai sus, în partea dreaptă.

Observație: Evident, există încă un arbore optim care are variabila A în nodul rădăcină (el corespunde unei alte reprezentări optimale a conjuncției $B \wedge \neg C$ față de cea de mai sus). Vedeți desenul alăturat.



- *Cazul 2:* Dacă în nodul rădăcină se alege atributul B , se obține partiția



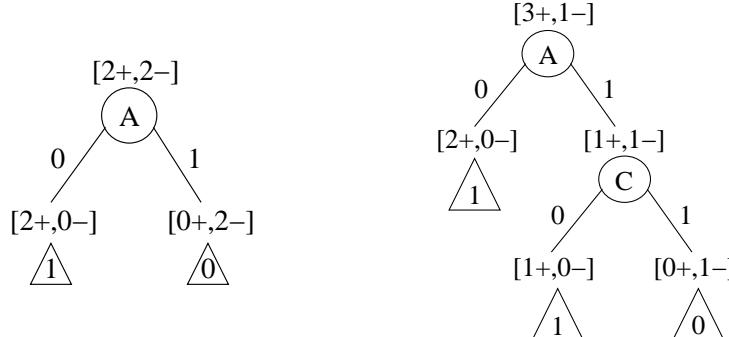
Subarborele stâng și subarborele drept trebuie să reprezinte arborele de decizie pentru funcțiile

$$X_2 = X[B/0] = (\neg A \vee 0) \wedge \neg(C \wedge A) = \neg A \wedge (\neg C \vee \neg A) = (\neg A \wedge \neg C) \vee \neg A = \neg A,$$

și respectiv

$$X_3 = X[B/1] = (\neg A \vee 1) \wedge \neg(C \wedge A) = 1 \wedge (\neg C \vee \neg A) = \neg C \vee \neg A,$$

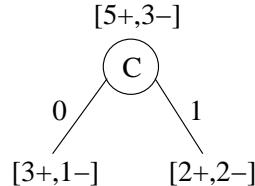
care au ca reprezentări optime arborii de mai jos:



Pentru arborele din dreapta există un arbore de decizie echivalent, obținut prin interschimbarea lui A cu C .

Prin urmare, orice arbore optim având variabila B în rădăcină are 3 niveluri și 4 noduri, aşadar cu un nod (de test) mai mult decât cel determinat în primul caz.

- **Cazul 3:** În sfârșit, când în nodul rădăcină se alege atributul C , obținem partitia



Subarborele stâng și subarborele drept trebuie să reprezinte arborele de decizie pentru funcțiile

$$X_4 = X[C/0] = (\neg A \vee B) \wedge \neg(0 \wedge A) = (\neg A \vee B) \wedge \neg 0 = \neg A \vee B$$

și respectiv

$$X_5 = X[C/1] = (\neg A \vee B) \wedge \neg(1 \wedge A) = (\neg A \vee B) \wedge \neg 1 = \neg A.$$

Urmând un raționament similar cu cel de la cazul anterior, putem spune că orice arbore optim cu atributul C în rădăcină are 3 niveluri și 4 noduri, cu un nod (de test) mai mult decât cel determinat în primul caz.

Așadar, putem concluziona că arborii de decizie optimi corespunzători funcției date sunt cei determinați în primul caz.

2.

(Algoritmul ID3: aplicare)

■ • CMU, 2002 spring, A. Moore, midterm example questions, pr. 2

Ai naufragiat pe o insulă pustie, unde nu găsești niciun alt fel de hrană decât ciuperci. Despre unele dintre aceste ciuperci se știe că sunt otrăvitoare, despre altele se știe că sunt comestibile, iar despre restul nu se știe ce fel sunt. Ai rămas singur pe insulă — foștii tăi camarazi, fiind epuizați de foame, au folosit metoda ‘trial and error’... — și ai la dispoziție următoarele date:

Exemplu	Ușoară	Mirositoare	ArePete	Netedă	Comestibilă
A	1	0	0	0	1
B	1	0	1	0	1
C	0	1	0	1	1
D	0	0	0	1	0
E	1	1	1	0	0
F	1	0	1	1	0
G	1	0	0	1	0
H	0	1	0	0	0
U	0	1	1	1	?
V	1	1	0	1	?
W	1	1	0	0	?

Atunci când nu vei mai avea la dispoziție pentru a supraviețui decât ciuperci U , V , sau W , ai putea estima care dintre ele sunt comestibile, folosind arbori de decizie.

În primele trei întrebări care urmează, ne vom referi la ciupercile $A - H$:

- Care este entropia atributului *Comestibilă*?
- Doar privind datele — adică fără a face explicit calculul câștigului de informație (engl., information gain) pentru cele patru atribute — poți determina ce atribut vei alege ca rădăcină a arborelui de decizie?
- Calculează câștigul de informație pentru atributul pe care l-ai ales la întrebarea precedentă.
- Elaborează întregul arbore de decizie ID3 bazat pe datele din tabel și apoi clasifică ciupercile U, V, W.
- Exprimă cu ajutorul calculului propozițional (logica predicatelor de ordinul 0) clasificarea produsă de arborele de decizie obținut. (*Comestibilă* $\leftrightarrow \dots$)
- Există vreun risc dacă vei consuma ciuperci care au fost clasificate de arborele de decizie ca fiind comestibile? De ce da? sau, de ce nu?

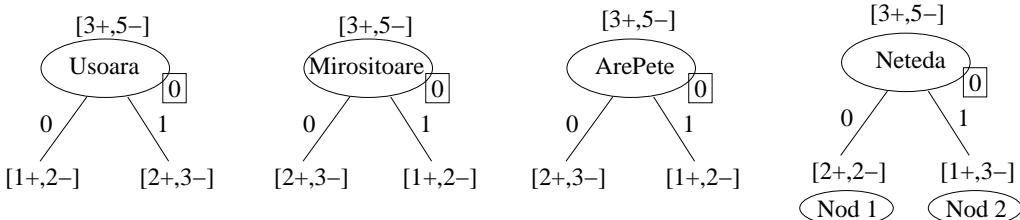
Răspuns:

- Entropia atributului *Comestibilă* este:

$$\begin{aligned} H_{\text{Comestibilă}} &\stackrel{\text{not.}}{=} H[3+, 5-] \stackrel{\text{def.}}{=} -\frac{3}{8} \log_2 \frac{3}{8} - \frac{5}{8} \log_2 \frac{5}{8} = \frac{3}{8} \log_2 \frac{8}{3} + \frac{5}{8} \log_2 \frac{8}{5} = \\ &= \frac{3}{8} \cdot 3 - \frac{3}{8} \log_2 3 + \frac{5}{8} \cdot 3 - \frac{5}{8} \log_2 5 = 3 - \frac{3}{8} \log_2 3 - \frac{5}{8} \log_2 5 \approx \\ &\approx 0.9544 \end{aligned}$$

Observație (1): Notația $[3+, 5-]$ simbolizează o mulțime partionată în 3 exemple pozitive și 5 exemple negative. Vom folosi acest gen de notăție peste tot în continuare, cu mici variații determinate de valorile pe care le poate lua atributul de ieșire. De exemplu, dacă vorbim despre o mulțime cu 5 obiecte roșii, 3 albastre și 4 verzi, am putea nota: $[5R, 3A, 4V]$.

- În rădăcina arborelui de decizie se alege atributul care aduce cel mai mare câștig de informație. Adică, atributul care, intuitiv vorbind, partionează cel mai bine datele de antrenament în raport cu atributul de ieșire. În cazul nostru, variantele pe care le avem la dispoziție pentru rădăcina arborelui (nodul 0) sunt:



Este ușor de observat că atributele *Ușoară*, *Mirositoare* și *ArePete* împart mulțimea exemplelor în mod similar: o submulțime cu 3 elemente, dintre care unul este pozitiv, iar două sunt negative, și o submulțime cu 5 elemente, dintre care două sunt pozitive, iar trei sunt negative.

Dacă am considera un arbore de decizie cu un singur nod de test în care plasăm atributul *Netedă*, atunci numărul minim de erori la antrenare pe care îl putem obține este 3, utilizând următoarea clasificare:

- $Netedă = 0 : Comestibilă = 1 \Rightarrow$ ciupercile E și H sunt clasificate greșit
- $Netedă = 1 : Comestibilă = 0 \Rightarrow$ ciuperca C este clasificată greșit

Dacă, în schimb, vom pune în rădăcina arborelui de decizie unul dintre celelalte trei atribute, spre exemplu atributul $Ușoară$, și dacă vom lua votul majoritar în fiecare nod descendant din nodul rădăcină, eroarea rezultată la antrenare va fi aceeași ca mai sus ($3/8$), însă toate instanțele vor fi clasificate la fel (și anume, negativ). Dacă nu lucrăm cu vot majoritar pentru ambii descendenti, ci doar pentru cel cu entropie mai mică (în vreme ce pentru celălalt nod descendant luăm decizia contrară), se observă că pentru atributul $Ușoară$ vom obține 4 erori pe setul de antrenament, iar pentru atributul $Netedă$ vom obține 3 erori.

Sumarizând, suntem înclinați să credem că ar fi o alegere sensibil mai bună să punem în rădăcină atributul $Netedă$. Pentru o justificare numerică riguroasă a acestei alegeri folosind criteriul maximizării câștigului de informație, vedeți punctul d.

c. Pentru a obține câștigul de informație pentru atributul $Netedă$, se fac calculele:

$$\begin{aligned} H_{0/Netedă} &\stackrel{\text{def.}}{=} \frac{4}{8}H[2+, 2-] + \frac{4}{8}H[1+, 3-] = \frac{1}{2} \cdot 1 + \frac{1}{2} \left(\frac{1}{4} \log_2 \frac{4}{1} + \frac{3}{4} \log_2 \frac{4}{3} \right) \\ &= \frac{1}{2} + \frac{1}{2} \left(\frac{1}{4} \cdot 2 + \frac{3}{4} \cdot 2 - \frac{3}{4} \log_2 3 \right) = \frac{1}{2} + \frac{1}{2} \left(2 - \frac{3}{4} \log_2 3 \right) \\ &= \frac{1}{2} + 1 - \frac{3}{8} \log_2 3 = \frac{3}{2} - \frac{3}{8} \log_2 3 \approx 0.9056 \end{aligned}$$

$$IG_{0/Netedă} \stackrel{\text{def.}}{=} H_{Comestibilă} - H_{0/Netedă} = 0.9544 - 0.9056 = 0.0488$$

Observație (2): În cele de mai sus am notat cu $H_{0/Netedă}$ entropia partitiei [mulțimii de exemple de antrenament] determinate de alegerea atributului $Netedă$ în nodul 0,⁴⁸⁴ iar cu $IG_{0/Netedă}$ câștigul de informație corespunzător acestei alegeri. În general, prin notația $H_{n/A}$ vom înțelege entropia partitiei determinate de alegerea atributului A în nodul n .

d. Arborele de decizie ID3 se construiește pornind din rădăcină și alegând atributul pentru fiecare nod de test în modul următor:

Nodul 0 (rădăcina):

Să verificăm dacă alegerea făcută la punctul b este cea corectă:

$$\begin{aligned} H_{0/Ușoară} &\stackrel{\text{def.}}{=} \frac{3}{8}H[1+, 2-] + \frac{5}{8}H[2+, 3-] \\ &= \frac{3}{8} \left(\frac{1}{3} \log_2 \frac{3}{1} + \frac{2}{3} \log_2 \frac{3}{2} \right) + \frac{5}{8} \left(\frac{2}{5} \log_2 \frac{5}{2} + \frac{3}{5} \log_2 \frac{5}{3} \right) \\ &= \frac{3}{8} \left(\frac{1}{3} \log_2 3 + \frac{2}{3} \log_2 3 - \frac{2}{3} \cdot 1 \right) + \frac{5}{8} \left(\frac{2}{5} \log_2 5 - \frac{2}{5} \cdot 1 + \frac{3}{5} \log_2 5 - \frac{3}{5} \log_2 3 \right) \\ &= \frac{3}{8} \left(\log_2 3 - \frac{2}{3} \right) + \frac{5}{8} \left(\log_2 5 - \frac{3}{5} \log_2 3 - \frac{2}{5} \right) \\ &= \frac{3}{8} \log_2 3 - \frac{2}{8} + \frac{5}{8} \log_2 5 - \frac{3}{8} \log_2 3 - \frac{2}{8} = \frac{5}{8} \log_2 5 - \frac{4}{8} \approx 0.9512 \end{aligned}$$

⁴⁸⁴Mai riguros, folosind terminologia din *Teoria informației*, vom spune că notația $H_{0/Netedă}$ se referă la *entropia condițională medie* a atributului de ieșire $Comestibilă$ în raport cu atributul de intrare $Netedă$.

Urmează că

$$IG_{0/U\text{şoară}} \stackrel{\text{def.}}{=} H_{\text{Comestibilă}} - H_{0/U\text{şoară}} = 0.9544 - 0.9512 = 0.0032,$$

deci

$$IG_{0/U\text{şoară}} = IG_{0/Mirositoare} = IG_{0/ArePete} = 0.0032 < IG_{0/Netedă} = 0.0488$$

Am avut deci dreptate să alegem atributul *Netedă* la punctul *b*.

Observație importantă:

În loc să fi calculat efectiv aceste câștiguri de informație, pentru a determina atributul cel mai „bun“, ar fi fost *suficient* să elaborăm un *raționament* de tip *relational*, bazat pe comparația dintre valorile entropiilor condiționale medii $H_{0/Netedă}$ și $H_{0/U\text{şoară}}$:

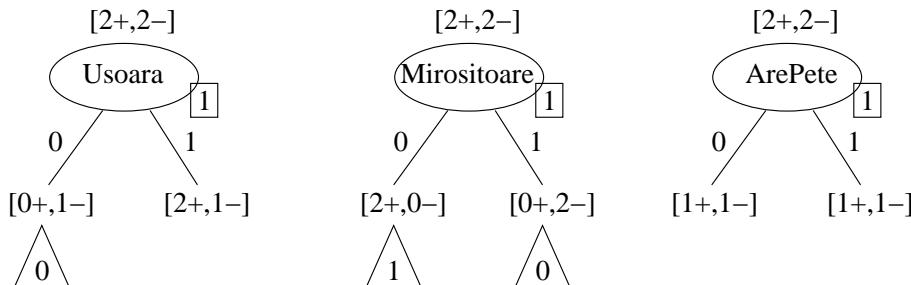
$$\begin{aligned} IG_{0/Netedă} > IG_{0/U\text{şoară}} &\Leftrightarrow H_{0/Netedă} < H_{0/U\text{şoară}} \\ &\Leftrightarrow \frac{3}{2} - \frac{3}{8} \log_2 3 < \frac{5}{8} \log_2 5 - \frac{1}{2} \Leftrightarrow 12 - 3 \log_2 3 < 5 \log_2 5 - 4 \\ &\Leftrightarrow 16 < 5 \log_2 5 + 3 \log_2 3 \Leftrightarrow 16 < 11.6096 + 4.7548 \text{ (adev.)} \end{aligned}$$

În mod *alternativ*, ținând cont de relația (265) de la problema 36, putem proceda chiar *mai simplu* relativ la calcule (nu doar aici, ci ori de câte ori nu avem de-a face cu un *număr mare de instanțe*):

$$\begin{aligned} H_{0/Netedă} < H_{0/U\text{şoară}} &\Leftrightarrow \frac{4^4}{2^4 \cdot 2^4} \cdot \frac{4^4}{3^3} < \frac{3^8}{2^4 \cdot 2^4} \cdot \frac{5^5}{3^3} \Leftrightarrow \frac{4^8}{3^3} < 5^5 \Leftrightarrow 4^8 < 3^3 \cdot 5^5 \\ &\Leftrightarrow 2^{16} < 3^3 \cdot 5^5 \Leftrightarrow 64 \cdot 2^{10} < 27 \cdot 25 \cdot 125 \text{ (adev.)} \end{aligned}$$

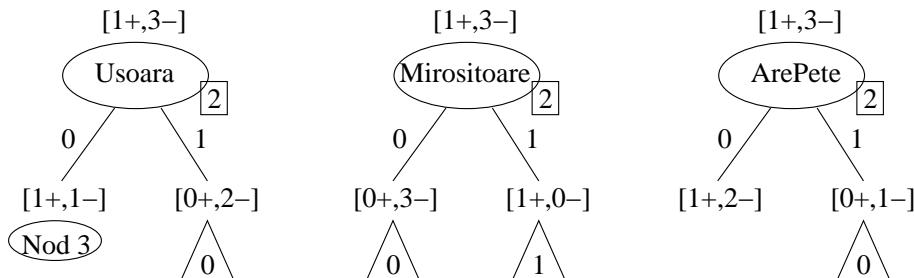
Vă sfătuim să *procedați aşa* la rezolvarea problemelor propuse din acest capitol, acolo unde este cazul.

Nodul 1: Trebuie să clasificăm acele exemple care au *Netedă* = 0; avem de ales între 3 attribute - *Uşoară*, *Mirositoare* și *ArePete*.

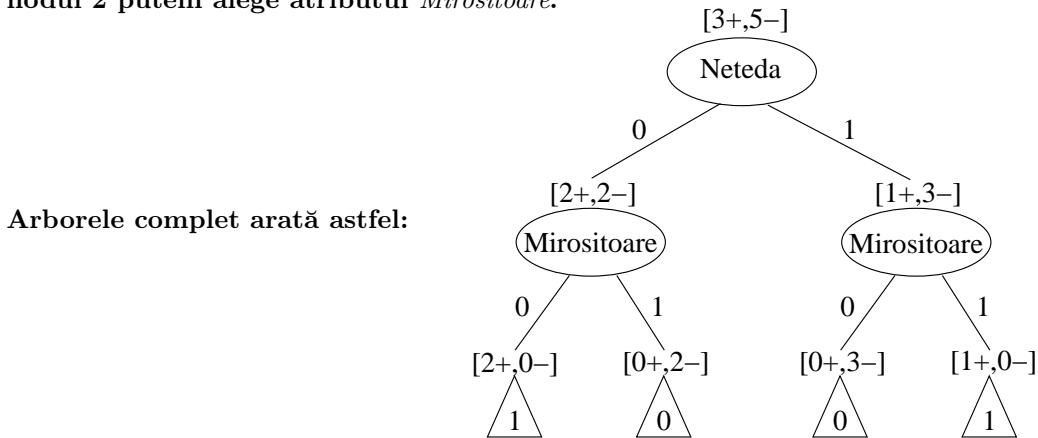


Aveam $H_{1/Miroslitoare} = \frac{2}{4}H[2+, 0-] + \frac{2}{4}H[0+, 2-] = 0$. Oricare ar fi valorile pentru $H_{1/ArePete}$ și $H_{1/U\text{şoară}}$, întrucât știm că entropia are întotdeauna valori nene-gative, rezultă că atributul *Miroslitoare* maximizează în nodul 1 câștigul de informație. În imaginea de mai sus valorile din triunghi reprezintă decizia luată de subarborele construit în nodul-frunză respectiv.

Nodul 2: Avem de clasificat exemplele pentru care *Netedă* = 1. Atributele disponibile sunt: *Uşoară*, *Miroslitoare* și *ArePete*.



Evident, $H_{2/Mirosoitoare} = \frac{3}{4}H[0+,3-] + \frac{1}{4}H[1+,0-] = \frac{3}{4} \cdot 0 + \frac{1}{4} \cdot 0 = 0$ Așadar, pentru nodul 2 putem alege atributul *Mirosoitoare*.



Parcurgând arborele construit, ciupercile *U*, *V* și *W* vor fi clasificate astfel:

<i>U</i>	<i>Netedă</i> = 1, <i>Mirosoitoare</i> = 1 \Rightarrow <i>Comestibilă</i> = 1
<i>V</i>	<i>Netedă</i> = 1, <i>Mirosoitoare</i> = 1 \Rightarrow <i>Comestibilă</i> = 1
<i>W</i>	<i>Netedă</i> = 0, <i>Mirosoitoare</i> = 1 \Rightarrow <i>Comestibilă</i> = 0

e. $\text{Comestibilă} \leftrightarrow (\neg\text{Netedă} \wedge \neg\text{Mirosoitoare}) \vee (\text{Netedă} \wedge \text{Mirosoitoare})$

Același lucru poate fi exprimat și sub forma unui pseudo-cod *if* ... *then* *Comestibilă* *else* \neg *Comestibilă*:

```

IF      (Netedă = 0 AND Mirosoitoare = 0) OR
        (Netedă = 1 AND Mirosoitoare = 1)
THEN   Comestibilă;
ELSE    $\neg$  Comestibilă;
  
```

f. Arborele de decizie produs de către algoritmul ID3 elaborat mai sus este consistent cu datele de antrenament pe care le-am avut la dispoziție (fiindcă aceste date sunt necontradictorii). Întrucât în realitate clasificarea poate depinde și de alte trăsături / informații decât cele de care dispunem noi, nu avem garanția că arborele ID3 face identificarea corectă a etichetei / clasei pentru toate instanțele din setul de test. Așadar, nu putem fi siguri că nu ne vom îmbolnăvi dacă vom consuma ciupercile *U* și *V*, sau că ne vom îmbolnăvi dacă vom consuma ciuperca *W*. În multe aplicații practice, calitatea unui model de învățare automată (în cazul de față, un arbore de decizie) se verifică pe un set de *date de validare*.

3.

(Algoritmul ID3, aplicat pe expresii booleene;
exploatarea simetriilor operațiilor \vee, \wedge în alegerea atributelor;
analiza „optimalității“ arborelui ID3, ca număr de noduri de test)

*prelucrare de Liviu Ciortuz, după
Tom Mitchell, "Machine Learning", 1997, ex. 3.1.b*

Considerăm următoarea funcție booleană: $A \vee (B \wedge C)$. Presupunem că această funcție este deja definită — adică valoarea ei este cea cunoscută din logica propozițiilor —, însă dorim să o reprezentăm ca arbore de decizie.

a. Aplicați algoritmul ID3 [tabelei de adevăr corespunzătoare] acestei funcții.
Observație: Dacă exploatați simetriile, veți avea nevoie doar de puține calcule, altfel vă veți complica inutil.

b. Arborele ID3 obținut la punctul precedent este optimal?

Altfel spus, puteți găsi un alt arbore de decizie, de adâncime mai mică sau cu număr mai mic de noduri (comparativ cu arborele obținut la punctul a), care să reprezinte această funcție? (Țineți cont că în fiecare nod al unui arbore de decizie se poate testa un singur atribut.)

Răspuns:

a. Observăm că funcția dată este simetrică în B și C , datorită comutativității operatorului logic \wedge . O consecință a acestui fapt este că dacă, pe parcursul algoritmului ID3, avem de ales (și) între cele două attribute este nevoie să-l studiem doar pe unul dintre ele, celălalt comportându-se identic.

A	B	C	$Y = A \vee (B \wedge C)$
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	1
1	0	1	1
1	1	0	1
1	1	1	1

Nodul 0 (rădăcină):

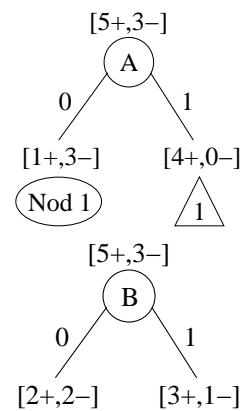
$$\begin{aligned} H_{0/A} &= \frac{4}{8}H[1+, 3-] + \frac{4}{8}H[4+, 0-] = \\ &= \frac{1}{2}H[1+, 3-] + \frac{1}{2} \cdot 0 = \frac{1}{2}H[1+, 3-] \end{aligned}$$

$$\begin{aligned} H_{0/B} &= \frac{4}{8}H[2+, 2-] + \frac{4}{8}H[3+, 1-] = \\ &= \frac{1}{2} \cdot 1 + \frac{1}{2}H[3+, 1-] = \frac{1}{2} + \frac{1}{2}H[1+, 3-] \end{aligned}$$

Este evident că $H_{0/A} < H_{0/B}$, deci vom alege atributul A în rădăcină.

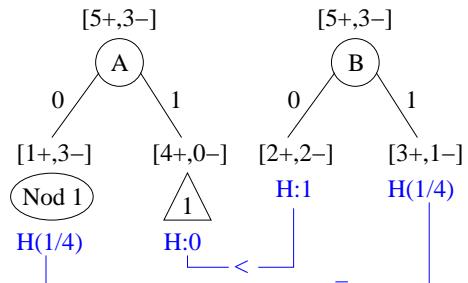
Observații importante:

- La aceeași concluzie se putea ajunge *imediat* pe baza unui *raționament* de tip *calitativ*, și anume, comparând atent cei doi arbori („comparați de decizie“) de mai sus. Mai precis, vom compara (două câte două) entropiile condiționale



specifice din nodurile descendente, precum și ponderile cu care se combină aceste entropii în scrierea entropiilor condiționale medii corespunzătoare atributelor A și B .

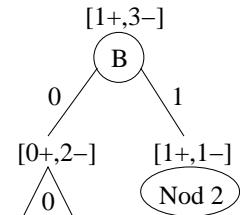
Putem pune în evidență acest fapt în figura alăturată, în care simbolul H , scris uneori însotit de un argument (așadar, ca $H(p)$), se referă la entropia unei variabile Bernoulli de parametru p .



Mai facem *precizarea* că semnele $<$ și $=$ din figura de mai sus se referă de fapt nu [doar] la entropiile condiționale specifice, ci [și] la produsul acestora cu ponderile asociate în mod corespunzător: $\frac{4}{8}H[1+, 3-] = \frac{4}{8}H[3+, 1-]$ și respectiv $\frac{4}{8}H[4+, 0-] < \frac{4}{8}H[2+, 2-]$.

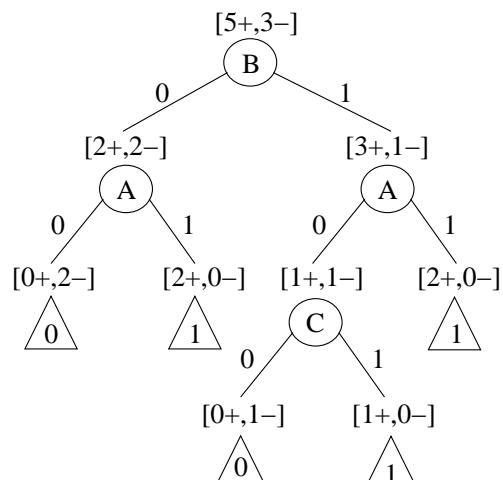
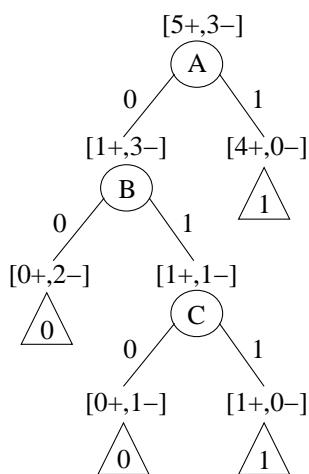
2. Pentru câteva formule de calcul convenabile pentru entropii și câștiguri de informație relative la compași de decizie, atunci când se folosește calculatorul de buzunar, dar numărul de instanțe de antrenament asociate nu este prea mare, vedeti problema 36.

Nodul 1: Avem de clasificat instanțele care au $A = 0$ și putem alege între atributele B și C . Datorită simetriei, îl putem alege pe oricare dintre ele. Pentru fixare, îl alegem pe B .



Nodul 2: La acest punct a mai rămas disponibil doar atributul C .

Arborele construit de ID3 este cel reprezentat mai jos, în partea stângă:⁴⁸⁵



⁴⁸⁵Un alt arbore ID3 este cel obținut din acesta interschimbând atributele B și C . (Vedeți *Observația* din enunț.)

b. Pentru a vedea dacă arborele construit de algoritmul ID3 este cel optim al, trebuie să reconsiderăm toate deciziile pe care le-am luat în construirea acestuia:

– La nodul 1 al arborelui avem de clasificat exemplele pentru care $A = 0$, deci funcția care trebuie reprezentată de subarborele în cauză este $f' = f[A/0] = 0 \vee (B \wedge C) = B \wedge C$, funcție care este reprezentată în mod optimal de subarborele construit de ID3. (Notația $A/0$ semnifică faptul că variabila logică A este instanțiată la valoarea 0.) Prin urmare, nu există un arbore mai bun care să reprezinte funcția dată și să aibă în rădăcină atributul A .

– În rădăcină am ales atributul A în detrimentul celorlalte două attribute deoarece am demonstrat că aduce cel mai mare câștig de informație. Să vedem ce se întâmplă dacă alegem unul dintre attributele B sau C . După cum am discutat mai sus, datorită simetriei, pe oricare dintre cele două l-am alege, arborele rezultat ar avea aceeași formă. Pentru fixare, să-l alegem pe B .

Subarborele stâng și drept vor trebui să reprezinte funcțiile:

$$f'' = f[B/0] = A \vee (0 \wedge C) = A \vee 0 = A$$

și respectiv

$$f''' = f[B/1] = A \vee (1 \wedge C) = A \vee C.$$

Arborele minimal care poate fi construit în aceste circumstanțe este cel reprezentat mai sus în partea dreaptă. Acest arbore are 3 niveluri și 4 noduri, cu un nod în plus față de cel construit de algoritmul ID3.

Putem deci conchide că arborele construit respectând specificațiile algoritmului ID3 este cel optim al.

Observație:

Această problemă pune în evidență două modalități de parcursere a spațiului de versiuni pentru un concept (în cazul de față un concept din logica propozițiilor) care este reprezentat cu ajutorul arborilor de decizie. Pe de o parte avem explorarea (incompletă) făcută de algoritmul ID3 care este de tip “greedy”, iar pe de altă parte avem explorarea exhaustivă. Prima strategie de explorare procedează la o căutare „orientată” a soluției (și din această cauză este mai eficientă, dar se va vedea, ca revers, că nu asigură întotdeauna găsirea optimului), iar cea de-a doua strategie de explorare, deși asigură găsirea optimului, nu este utilizabilă în cazurile (frecvente!) în care spațiul de versiuni este foarte mare.

4.

(ID3, ca algoritm “greedy”; un exemplu când arborele ID3 nu este optim ca număr de noduri și de niveluri)

prelucrare de Liviu Ciortuz, după

■ CMU, 2003 fall, T. Mitchell, A. Moore, midterm exam, pr. 9.a

Fie attributele binare de intrare A, B, C , atributul de ieșire Y și următoarele exemple de antrenament:

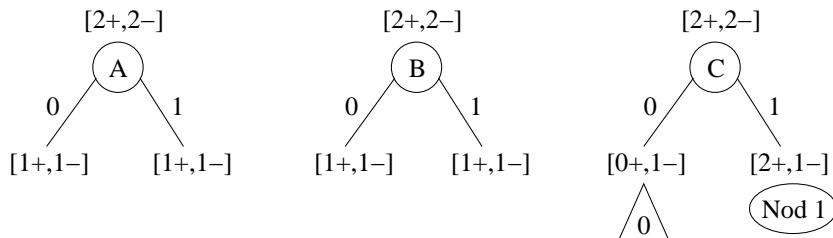
A	B	C	Y
1	1	0	0
1	0	1	1
0	1	1	1
0	0	1	0

- a. Determinați arborele de decizie calculat de algoritmul ID3. Este acest arbore de decizie *consistent* cu datele de antrenament?
- b. Există un arbore de decizie de adâncime mai mică (decât cea a arborelui ID3) consistent cu datele de mai sus? Dacă da, ce concept (logic) reprezintă acest arbore?

Răspuns:

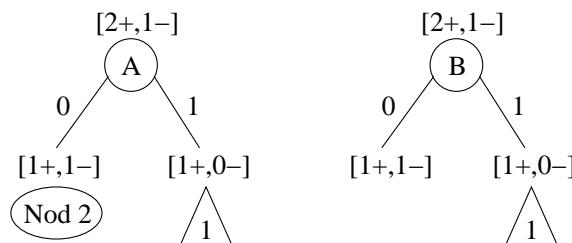
- a. Se construiește arborele de decizie cu algoritmul ID3 astfel:

Nodul 0 (rădăcina):



Observăm că $H_{0/A} = H_{0/B} = \frac{2}{4}H[1+, 1-] + \frac{2}{4}H[1+, 1-] = H[1+, 1-] = 1$, care este valoarea maximă a entropiei condiționale medii a unei variabile booleene. Prin urmare, $H_{0/C}$ nu poate fi decât mai mică sau egală cu $H_{0/A}$ și $H_{0/B}$. Deci vom alege în nodul rădăcină atributul C .

Nodul 1: Avem de clasificat instanțele cu $C = 1$, deci alegerea se face între atrbutele A și B .

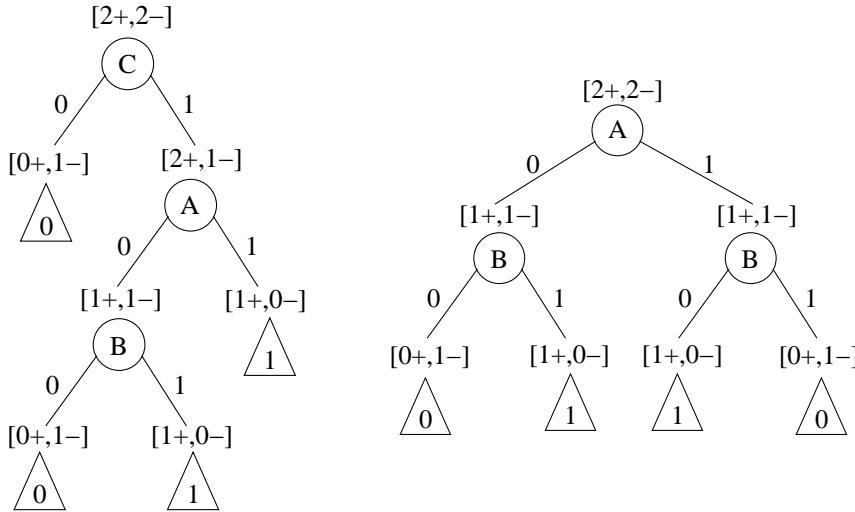


Cele două entropii condiționale medii sunt egale:

$$H_{1/A} = H_{1/B} = \frac{2}{3}H[1+, 1-] + \frac{1}{3}H[1+, 0-]$$

Așadar, putem alege oricare dintre cele două atrbute. Pentru fixare, îl alegem pe A .

Nodul 2: La acest nod nu mai avem decât atributul B , deci îl vom pune pe acesta. Arborele complet este reprezentat în partea stângă:



Prin construcție, arborele ID3 este consistent cu datele de antrenament dacă acestea sunt consistente (i.e., necontradictorii). În cazul nostru, se verifică imediat că datele de antrenament sunt consistente.

b. Se observă că atributul de ieșire Y reprezintă de fapt funcția logică $A \oplus B$. Reprezentând această funcție ca arbore de decizie, vom obține arborele desenat mai sus în partea dreaptă. Acest arbore are cu un nivel mai puțin decât arborele construit cu algoritmul ID3. Prin urmare, arborele obținut de algoritmul ID3 pe datele din enunț *nu este optim* din punctul de vedere al numărului de niveluri. Aceasta este o *consecință* a caracterului “greedy” al algoritmului ID3, datorat faptului că la fiecare iterație alegem „cel mai bun“ atribut în raport cu criteriul câștigului de informație. Se știe că algoritmii de tip “greedy” nu grantează obținerea optimului global.

5.

(Clasificare ternară: “decision stump” produs de ID3, pe date care conțin duplicări și „zgomote“)

Presupunem că se dau şase date de antrenament (precizate în tabel) pentru o problemă de clasificare cu două atrbute binare și trei clase $Y \in \{1, 2, 3\}$. Se va crea un arbore ID3, bazat pe câștigul de informație.

a. Calculați câștigul de informație atât pentru X_1 cât și pentru X_2 . Se va folosi aproximarea $\log_2 3 = 19/12$ și se va scrie câștigul de informație sub formă de fracții.

b. Pe baza rezultatelor anterioare, ce atrbut va fi folosit pentru primul nod al arborelui ID3? Desenați arborele de decizie care rezultă folosind doar acest singur nod. Etichetați corespunzător nodul, ramurile și eticheta prevăzută în fiecare frunză.

c. Cum va clasifica acest arbore instanța determinată de $X_1 = 0$ și $X_2 = 1$?

X_1	X_2	Y
1	1	1
1	1	1
1	1	2
1	0	3
0	0	2
0	0	3

Răspuns:

a. Redăm formula pentru calculul câștigului de informație în varianta folosită de Tom Mitchell:

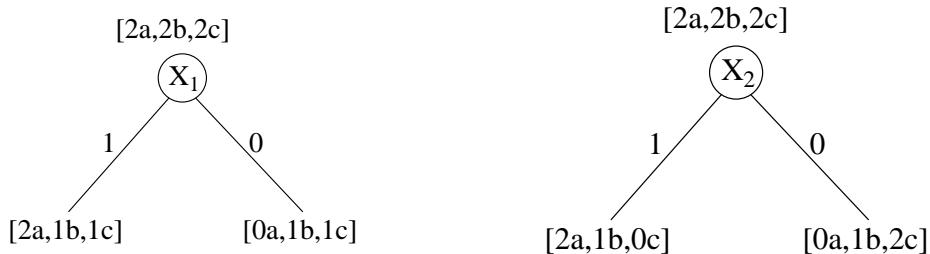
$$Gain(X_i) = Entropy(S) - \sum_{v \in Values(X_i)} \frac{|S_v|}{|S|} Entropy(S_v)$$

unde S este mulțimea celor 6 date de antrenament, iar

$$Entropy(S) = - \sum_{y \in Y} p_y \log_2 p_y$$

Notăm cu a clasa instanțelor având eticheta $Y = 1$, cu b clasa instanțelor cu $Y = 2$ și cu c clasa instanțelor cu $Y = 3$. În mulțimea S există câte 2 elemente din fiecare clasă și atunci putem scrie că $S = [2a, 2b, 2c]$.

Pentru nodul rădăcină, putem alege fie atributul X_1 , fie atributul X_2 , ceea ce determină următoarele împărțiri ale mulțimii S :



Putem calcula câștigurile de informație pentru cele două atrbute:

$$Gain(X_1) = H[2a, 2b, 2c] - \left(\frac{4}{6}H[2a, 1b, 1c] + \frac{2}{6}H[0a, 1b, 1c] \right)$$

unde $H[2a, 2b, 2c]$ este o altă notație pentru entropia mulțimii compuse din două exemple de clasă a , două de clasă b și două de clasă c .

Calculăm entropiile care intervin în formulă:

$$\begin{aligned} H[2a, 2b, 2c] &= -\frac{2}{6} \log_2 \frac{2}{6} - \frac{2}{6} \log_2 \frac{2}{6} - \frac{2}{6} \log_2 \frac{2}{6} \\ &= -3 \cdot \frac{2}{6} \log_2 \frac{2}{6} = -\log_2 \frac{1}{3} = \log_2 3 = \frac{19}{12} \\ H[2a, 1b, 1c] &= -\frac{2}{4} \log_2 \frac{2}{4} - \frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{4} \log_2 \frac{1}{4} = +\frac{1}{2} \cdot \log_2 2 + \frac{1}{4} \log_2 4 + \frac{1}{4} \log_2 4 \\ &= \frac{1}{2} + \frac{1}{4} \cdot 2 + \frac{1}{4} \cdot 2 = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = \frac{3}{2} \\ H[0a, 1b, 1c] &= 0 - \frac{1}{2} \cdot \log_2 \frac{1}{2} - \frac{1}{2} \cdot \log_2 \frac{1}{2} = \log_2 2 = 1 \end{aligned}$$

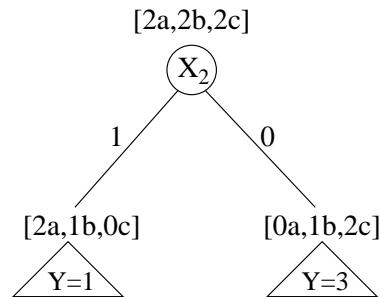
Înlocuind aceste valori numerice în formula câștigului de informație, obținem:

$$Gain(X_1) = \frac{19}{12} - \left(\frac{2}{3} \cdot \frac{3}{2} + \frac{1}{3} \cdot 1 \right) = \frac{19}{12} - \left(1 + \frac{1}{3} \right) = \frac{19}{12} - \frac{4}{3} = \frac{3}{12} = \frac{1}{4}$$

Se aplică aceleiasi formule și pentru atributul X_2 :

$$\begin{aligned}
 Gain(X_2) &= H[2a, 2b, 2c] - \left(\frac{3}{6}H[2a, 1b, 0c] + \frac{3}{6}H[0a, 1b, 2c] \right) \\
 H[2a, 1b, 0c] &= H[0a, 1b, 2c] \\
 &= -\frac{1}{3}\log_2 \frac{1}{3} - \frac{2}{3}\log_2 \frac{2}{3} = -\frac{2}{3}(\log_2 2 - \log_2 3) - \frac{1}{3}(-1)\log_2 3 \\
 &= -\frac{2}{3} \cdot 1 + \frac{2}{3} \cdot \frac{19}{12} + \frac{1}{3} \cdot \frac{19}{12} = \frac{19}{12} - \frac{2}{3} = \frac{11}{12} \\
 \Rightarrow Gain(X_2) &= \frac{19}{12} - \left(\frac{1}{2} \cdot \frac{11}{12} + \frac{1}{2} \cdot \frac{11}{12} \right) = \frac{19}{12} - \frac{11}{12} = \frac{8}{12} = \frac{2}{3}.
 \end{aligned}$$

b. Deoarece $Gain(X_1) = \frac{3}{12}$, $Gain(X_2) = \frac{8}{12}$, și deci $Gain(X_1) < Gain(X_2)$, se va alege atributul X_2 ca rădăcină a arborelui. Arborele de decizie construit din acest singur nod este cel din figura alăturată.



c. O instanță care are $X_1 = 0$ și $X_2 = 1$ va fi clasificată de acest arbore cu $Y = 1$ (cu probabilitate $2/3$).

6.

(Algoritmul ID3: aplicare pe date inconsistente, “decision stumps”, calculul acurateții)

• o CMU, 2012 fall, T. Mitchell, Z. Bar-Joseph, HW1, pr. 2.ab

Tabelul de mai jos summarizează situația celor 2201 de pasageri și membri ai echipajului de la bordul vasului Titanic, în urma naufragiului din data de 15 Aprilie 1912. Pentru fiecare combinație de valori ale celor 3 variabile (Clasă, Sex, Vârstă) am indicat în tabel câți oameni au supraviețuit și câți nu au supraviețuit. (*Observație:* Datele originale au patru valori pentru atributul Clasă; am comasat valorile II, III, și Echipaj într-o singură valoare, denumită „Inferioară“.)

Clasa	Sexul	Vârstă	Supraviețuitori		
			Nu	Da	Total
I	Masculin	Copil	0	5	5
I	Masculin	Adult	118	57	175
I	Feminin	Copil	0	1	1
I	Feminin	Adult	4	140	144
Inferioară	Masculin	Copil	35	24	59
Inferioară	Masculin	Adult	1211	281	1492
Inferioară	Feminin	Copil	17	27	44
Inferioară	Feminin	Adult	105	176	281
Total			1490	711	2201

Pentru a vă ușura calculele pe care va trebui să le faceți, am făcut noi totalurile pentru fiecare variabilă:

Clasa	Supraviețuitori		
	Nu	Da	Total
I	122	203	325
Inferioară	1368	508	1876

Sexul	Supraviețuitori		
	Nu	Da	Total
Masculin	1364	367	1731
Feminin	126	344	470

Vârstă	Supraviețuitori		
	Nu	Da	Total
Copil	52	57	109
Adult	1438	654	2092

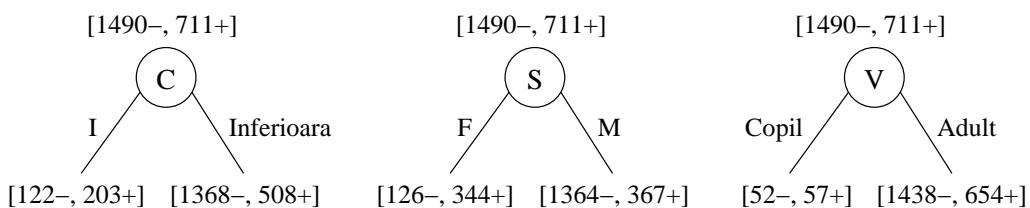
a. Folosind un arbore de decizie, dorim să prezicem variabila de ieșire Y (Supraviețuitor), pornind de la atributele de intrare C (Clasa), S (Sexul), V (Vârstă). Utilizați criteriul câștigului de informație pentru a alege care dintre aceste trei atrbute C , S sau V trebuie să fie folosit în nodul-rădăcină al arborelui de decizie.

De fapt, ce vi se cere este să învățați un arbore de decizie de adâncime 1 care folosește doar atributul din rădăcină pentru a clasifica datele. (Astfel de arbori de decizie de adâncime 1 sunt adesea numiți în terminologia de limbă engleză “decision stumps”.) Parcurgeți toate etapele rezolvării, redând inclusiv calculele pentru câștig de informație al fiecărui atribut.

- b. Care este acuratețea [medie] obținută pe datele de antrenament de către arboarele de decizie cu adâncime 1 de la punctul precedent?
- c. Dacă ați crea un arbore de decizie care folosește toate cele trei variabile, care ar fi acuratețea lui [medie] pe datele de antrenament? (*Observație:* Nu trebuie neapărat să creați arboarele de decizie pentru a afla răspunsul!)

Răspuns:

- a. Totalurile care au fost furnizate în enunț pentru fiecare dintre variabilele C , S și V ne servesc foarte bine pentru a crea rapid cei trei “decision stumps”:



Analizând datele conform figurii de mai sus, se poate „intui“ că atributul S va avea un câștig de informație (în raport cu atributul de ieșire Y – *Supraviețuitor*) mai bun decât al celorlalte două atrbute de intrare (C și V). Intuiția se verifică făcând calculele:

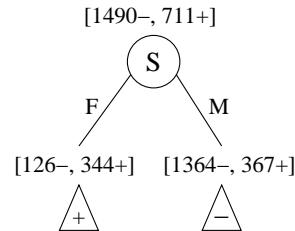
$$\begin{aligned}
 IG(Y, C) &= H[1490-, 711+] - \left(\frac{325}{2201} H[122-, 203+] + \frac{1876}{2201} H[1368-, 508+] \right) \\
 &= 0.048501 \\
 IG(Y, S) &= H[1490-, 711+] - \left(\frac{470}{2201} H[126-, 344+] + \frac{1731}{2201} H[1364-, 367+] \right) \\
 &= 0.142391
 \end{aligned}$$

$$\begin{aligned} IG(Y, V) &= H[1490-, 711+] - \left(\frac{109}{2201} H[52-, 57+] + \frac{2092}{2201} H[1438-, 654+] \right) \\ &= 0.006411. \end{aligned}$$

Deci, într-adevăr, câștigul maxim de informație se obține pentru atributul S .

- b. Arborele de decizie de adâncime 1 care are în nodul rădăcină atributul S este cel din figura alăturată. Acuratețea [medie a] acestui arbore de decizie este:

$$\frac{470}{2201} \cdot \frac{344}{470} + \frac{1731}{2201} \cdot \frac{1364}{1731} = \frac{344 + 1364}{2201} = \frac{1708}{2201} = 0.776.$$



- c. Se poate constata imediat că arborele ID3 produs pe datele din această problemă va avea 8 noduri-frunză, iar în fiecare dintre aceste noduri-frunză se va asigna câte una dintre mulțimile descrise (pe linie) în coloanele 4 și 5 ale tabelului principal din enunț: $[0-, 5+]$, $[118-, 57+]$, ..., $[17-, 27+]$, $[105-, 176+]$. Decizia care va fi luată în fiecare nod-frunză este dictată de votul majoritar, și anume: $+, -, \dots, +$ și respectiv $+$.

Putem calcula acuratețea [medie] astfel:

$$\frac{5 + 118 + 1 + 140 + 35 + 1211 + 27 + 176}{2201} = \frac{1713}{2201} = 0.778.$$

Se observă că se produce (din păcate) o creștere foarte mică în raport cu acuratețea celui mai bun “decision stump”: doar 0.002.

7.

(Algoritmul ID3: cazul când există repetiții și inconsistențe în datele de antrenament; o margine superioară pentru eroarea la antrenare în funcție de numărul de valori ale variabilei de ieșire)

CMU, 2002 fall, Andrew Moore, midterm exam, pr. 1.fg

Presupunem că învățăm un arbore de decizie care să prezică atributul de ieșire Z pornind de la atributele de intrare A, B, C . Se folosesc datele de antrenament din tabelul alăturat.

- a. Care va fi eroarea la antrenare pe acest set de date? Exprimăți răspunsul sub forma fracției de înregistrări care vor fi clasificate eronat ($n/12$).
 b. Considerăm un arbore de decizie construit pe un set arbitrar de date. Dacă atributul de ieșire este cu valori discrete și poate lua k valori distincte, care este maximul erorii la antrenare (exprimată ca fracție)?

A	B	C	Z
0	0	0	0
0	0	1	0
0	0	1	0
0	1	0	0
0	1	1	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	0
1	1	1	1

Răspuns:

- a. Este ușor de observat că datele de antrenament conțin „inconsistențe“ (contradicții, relativ la etichetare), și anume la exemplele $(0, 1, 1)$ și $(1, 1, 1)$.

Fiecare dintre aceste exemple sunt etichetate o dată cu 0 și altă dată cu 1. Prin urmare, jumătate din aceste exemple vor fi clasificate eronat de arborele învățat de către algoritm ID3. Eroarea la antrenare va fi deci $\frac{2}{12}$.

b. Vom analiza pe rând mai multe cazuri, care sunt din ce în ce mai generale.

Cazul i: Mai întâi vom calcula eroarea la antrenare pentru cazul în care setul de date de antrenament este compus din k instanțe care sunt identice ca tupluri de valori pentru atributele ce le caracterizează, dar sunt clasificate pe rând cu fiecare din cele k valori posibile ale atributului de ieșire. Este evident că arborele de decizie învățat va clasifica eronat $k - 1$ instanțe. Așadar, în acest caz, eroarea la antrenare va fi

$$E = \frac{k-1}{k}$$

Cazul ii: Aceeași eroare [la antrenare] ca mai sus se va înregistra dacă în locul fiecărei instanțe dintre cele considerate la cazul precedent vom avea l instanțe identice, inclusiv în ce privește eticheta. (În total sunt kl instanțe de antrenament.)

$$E = \frac{(k-1) \cdot l}{k \cdot l} = \frac{k-1}{k}$$

Cazul iii: Dacă relaxăm condiția de mai sus considerând l_1, l_2, \dots, l_k instanțe identice, iar $l = \max_{i=1}^k l_i$, este imediat că eroarea maximă se va atinge în cazul $l_1 = l_2 = \dots = l_k = l$, și va avea aceeași valoare ca mai sus. Așadar, în continuare vom putea renunța la a considera factorul de multiplicare l , fără ca prin aceasta să restrângem generalitatea raționamentului.

Cazul iv: Fie n exemple de antrenament (instanțe etichetate) dintre care d sunt distințe (ca tupluri de valori ale atributelor de intrare). Fie $k_1, k_2 \dots k_d$ numărul de instanțe etichetate distințe pentru fiecare caz în parte din cele d . Atunci vom avea:

$$k_1 \leq k, k_2 \leq k, \dots, k_d \leq k \Rightarrow n = k_1 + k_2 + \dots + k_d \leq k \cdot d \text{ deci } n \leq k \cdot d$$

Eroarea maximă la antrenare va fi dată de formula

$$E = \frac{(k_1 - 1) + (k_2 - 1) + \dots + (k_d - 1)}{n} = \frac{n - d}{n}$$

Avem:

$$E \leq \frac{k-1}{k} \Leftrightarrow \frac{n-d}{n} \leq \frac{k-1}{k} \Leftrightarrow k \cdot n - k \cdot d \leq k \cdot n - n \Leftrightarrow k \cdot d \geq n \text{ (adev.)}$$

Prin urmare, eroarea maximă la antrenare care poate fi atinsă atunci când atributul de ieșire poate lua k valori distințe este $\frac{k-1}{k}$.

8. (ID3, aspecte computaționale: influența atributelor duplicate, respectiv a instanțelor de antrenament duplicate asupra arborelui ID3 rezultat)

CMU, 2009 spring, Ziv Bar-Joseph, final exam, pr. 3.1-2

Se dorește construirea unui arbore de decizie pentru n vectori, cu m atribute.

a. Să presupunem că există i și j astfel încât pentru TOTI vectorii X din datele de antrenament, aceste atribute au valori egale (adică, $x_i = x_j$ pentru toți vectorii, unde x_i este valoarea atributului i în vectorul X). Să presupunem de asemenea că în cazul în care ambele atribute duc la același câștig de informație vom folosi atributul i . Stergerea atributului j din datele de antrenament poate schimba arborele de decizie obținut? Explicați pe scurt.

b. Să presupunem că există în mulțimea de antrenament doi vectori egali X și Z (adică, toate atributele lui X și Z sunt exact la fel, inclusiv etichetele). Stergerea vectorului Z din datele de antrenament poate schimba arborele de decizie obținut? Explicați pe scurt.

Răspuns:

- a. Nu, îndepărțarea atributului j nu schimbă arborele de decizie, deoarece atributele i și j conduc la valori egale pentru câștigul de informație în fiecare nod al arborelui.
- b. Da, în acest caz arborele de decizie se poate schimba, fiindcă entropia condițională — care se calculează în fiecare nod pentru a determina atributul cu câștigul de informație cel mai mare — depinde de numărul de instanțe de antrenament luate în considerare.

9. (Arbori de decizie: o margine superioară pentru numărul de noduri frunză, în funcție de numărul atributelor și numărul de exemple)

CMU, 2005 fall, T. Mitchell, A. Moore, midterm exam, pr. 2.d

Presupunem că învățăm un arbore de decizie folosind un set de R instanțe de antrenament descrise de M atribute de intrare având valori binare.

Care este numărul maxim posibil de noduri frunză din arborele de decizie, presupunând că fiecărui nod frunză îi este asociat măcar un exemplu de antrenament? Încercuiți unul din răspunsurile de mai jos; justificați alegerea făcută.

$R, \log_2(R), R^2, 2^R, M, \log_2(M), M^2, 2^M,$
 $\min(R, M), \min(R, \log_2(M)), \min(R, M^2), \min(R, 2^M),$
 $\min(\log_2(R), M), \min(\log_2(R), \log_2(M)), \min(\log_2(R), M^2), \min(\log_2(R), 2^M),$
 $\min(R^2, M), \min(R^2, \log_2(M)), \min(R^2, M^2), \min(R^2, 2^M),$
 $\min(2^R, M), \min(2^R, \log_2(M)), \min(2^R, M^2), \min(2^R, 2^M),$
 $\max(R, M), \max(R, \log_2(M)), \max(R, M^2), \max(R, 2^M),$
 $\max(\log_2(R), M), \max(\log_2(R), \log_2(M)), \max(\log_2(R), M^2), \max(\log_2(R), 2^M),$
 $\max(R^2, M), \max(R^2, \log_2(M)), \max(R^2, M^2), \max(R^2, 2^M),$
 $\max(2^R, M), \max(2^R, \log_2(M)), \max(2^R, M^2), \max(2^R, 2^M).$

Răspuns:

Notăm cu \max_{frunze} valoarea căutată. Trebuie luate în considerare două aspecte:

- (a) fiecare nod frunză trebuie să clasifice măcar un exemplu de antrenament
 \Rightarrow nu putem să avem mai multe frunze decât exemple de antrenament
 $\Rightarrow \max_{frunze} \leq R$
- (b) fiecare atribut poate fi testat o singură dată pe un drum de la rădăcină la o frunză oarecare \Rightarrow arborele obținut va avea adâncimea cel mult M
 $\Rightarrow \max_{frunze} \leq 2^M$

$$\left. \begin{array}{l} \max_{frunze} \leq R \\ \max_{frunze} \leq 2^M \end{array} \right\} \Rightarrow \max_{frunze} \leq \min(R, 2^M)$$

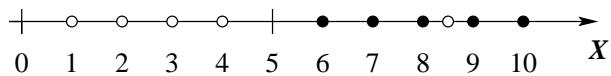
Această valoare poate fi atinsă: putem să luăm, de exemplu, cazul trivial când avem o singură instanță de antrenament. Prin urmare, avem $\max_{frunze} = \min(R, 2^M)$.

10. (Extensii ale algoritmului ID3: variabile de intrare continue; “decision stumps”; eroarea la antrenare, eroarea la CVLOO; overfitting)

■ • CMU, 2002 fall, Andrew Moore, midterm exam, pr. 3

Fie setul de date de mai jos. X este atribut de intrare și ia valori reale, iar Y este variabilă de ieșire cu valori booleene. (*Observație:* Am reprezentat acest set de date sub formă grafică, marcând valoarea / eticheta $Y = 1$ prin bulină neagră, iar valoarea / eticheta $Y = 0$ prin cerculeț alb.) Pe acest set de date se folosește algoritmul ID3 pentru învățare de arbori de decizie.

X	Y
1	0
2	0
3	0
4	0
6	1
7	1
8	1
8.5	0
9	1
10	1



Algoritmul ID3 (extins) va trebui să decidă cum divide (engl., split) intervale de valori asociate variabili reale X . Separarea în intervale diferite va fi stabilită cu ajutorul unor *praguri de separare* (engl., split thresholds), determinate în felul următor:

- Mai întâi se ordonează crescător acele valori ale variabilei X care apar în datele de antrenament.

- Se stabilesc apoi perechi de valori consecutive pentru care există instanțe de antrenament care sunt etichetate în mod diferit pentru o valoare, comparativ cu cealaltă valoare. Pentru fiecare pereche de valori de acest fel, va fi plasat un prag de separare la jumătatea distanței dintre cele două valori.

Inițial, se alege pragul de separare care conduce la un câștig de informație maxim. Apoi, la fiecare nouă execuție a buclei principale a algoritmului ID3 — vă readucem aminte că acest algoritm este recursiv — se va selecta câte un alt prag dintre cele rămase disponibile, aplicând același criteriu: maximizarea câștigului de informație.

De exemplu, pentru $X = 4$ avem o instanță de antrenament negativă, iar pentru $X = 6$ avem o instanță de antrenament pozitivă. Se poate arăta că algoritmul ID3 va decide să splitze mai întâi la valoarea $X = 5$ (care reprezintă jumătatea distanței dintre $X = 4$ și $X = 6$) și apoi la valoarea $X = 8.25$ (care reprezintă jumătatea distanței dintre $X = 8$ și $X = 8.5$).

Fie DT* arborele de decizie complet, obținut de algoritm ID3 fără a face pruning, iar DT2 arborele de decizie produs de ID3 urmat de pruning, care are doar două noduri frunză (deci DT2 face o singură divizare de interval).

- Care este eroarea la antrenare produsă de DT2 respectiv DT* (exprimată ca număr de exemple clasificate eronat din totalul de 10 exemple)?
- Care este eroarea produsă de DT2 respectiv DT* la cross-validation folosind metoda *Leave-One-Out* (CVLOO)?

Răspuns:

- Deoarece DT2 reține doar un singur split, și anume cel de la pragul $X = 5$, regula de decizie pe care o reprezintă este:

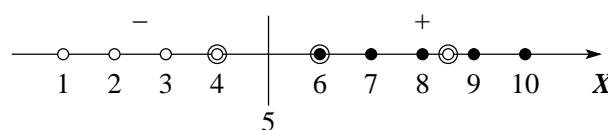
```

IF      X ≤ 5
THEN    Y = 0
ELSE    Y = 1
  
```

Este evident că această regulă produce o clasificare eronată doar pentru una dintre instanțele de antrenament, și anume $X = 8.5$. Avem deci $E_{antren.}(DT2) = 1/10$.

Întrucât exemplele nu conțin inconistențe, arborele de decizie ID3 clasifică corect toate datele de antrenament. Avem deci $E_{antren.}(DT^*) = 0/10 = 0$.

- Figura de mai jos reprezintă împărțirea axei reale în *intervale / zone de decizie* conform arborelui (și *pragului de decizie*) învățat de către algoritmul DT2, folosind întregul set de exemple date. (Am încercuit acele puncte care, după cum se va vedea mai jos, vor constitui cazuri aparte la calcularea erorii de tip CVLOO cu algoritmul DT2.)



În ce privește cross-validationa prin metoda “Leave-One-Out”, se poate demonstra — calculele nu sunt arătate aici⁴⁸⁶ — următorul fapt: pentru fiecare din cele 10 exemple ($X = 1, X = 2, \dots, X = 10$) considerate pe rând, pragul de decizie identificat de algoritm DT2 va fi $X = 5$, cu excepția următoarelor două cazuri:

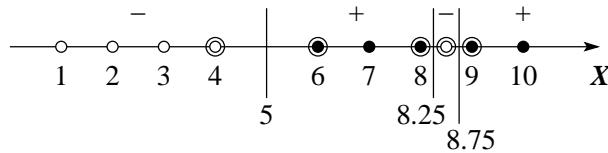
- $X = 4$: în acest caz, splitarea se va face la mijlocul intervalului $[3, 6]$, deci la 4.5. Cum $4 \leq 4.5$, rezultă că exemplul $X = 4$ va fi clasificat corect;
- $X = 6$: splitarea se va face la mijlocul intervalului $[4, 7]$, deci la 5.5. Cum $6 > 5.5$, rezultă că exemplul $X = 6$ va fi clasificat corect.

Atunci când punctul $X = 8.5$ este lăsat deoparte, pragul de decizie selectat fiind $X = 5$, va rezulta că punctul $X = 8.5$ este clasificat pozitiv (deci eronat), întrucât $8.5 > 5$.

Este imediat că pentru restul de 7 cazuri ($X = 1, 2, 3, 7, 8, 9, 10$), arborele determinat de algoritm DT2 va clasifica corect punctul X .

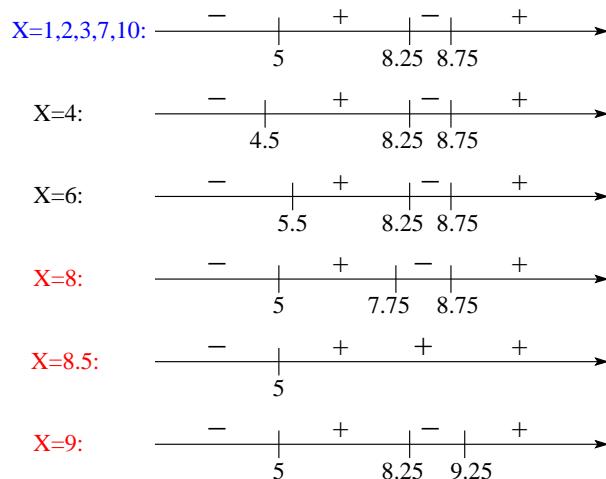
Așadar, pentru DT2 rezultă $E_{CVLOO} = 1/10$.

Figura de mai jos reprezintă împărțirea axei reale în *intervale / zone de decizie* conform *pragurilor de decizie* (și arborelui DT*) determinate de către algoritm ID3 pe întregul set de date de antrenament.⁴⁸⁷



În această figură am încercuit exemplele care ar putea conduce la eroare la cross-validationa de tip “Leave-One-Out”:

- $X = 4$: corect clasificat, explicația este aceeași ca în cazul DT2;
- $X = 6$: idem;
- $X = 8$: split-ul trebuie făcut la mijlocul intervalului $[7, 8.5]$, adică 7.75. Cum $8 > 7.75$, rezultă că punctul $X = 8$ va fi clasificat negativ, ceea ce este eronat;
- $X = 8.5$: nu este nevoie de căt de un singur split, arborele DT* învățat în acest caz fiind identic cu cel construit de algoritmul DT2. Cum $8 > 5$, rezultă că punctul $X = 8.5$ va fi clasificat pozitiv, deci eronat;



⁴⁸⁶ Aceste calcule nu sunt dificile, mai ales dacă se utilizează raționamente de tip „calitativ“, aşa cum am arătat la problema 3.

⁴⁸⁷ Observație: Atunci când ne raportăm doar la zonele de decizie în ansamblu, ordinea în care s-au stabilit testele în arborele ID3 este irelevantă! Acest fapt este valabil și mai jos, unde discutăm despre eroarea la CVLOO cu algoritmul DT*.

- $X = 9$: intervalul de split devine $[8.5, 10]$, split-ul făcându-se la 9.25. Cum $9 \leq 9.25$, rezultă că punctul $X = 9$ va fi clasificat negativ, ceea ce este eronat.

Așadar, pentru DT^* avem $E_{CVLOO} = 3/10$.

În concluzie, se observă că $E_{antren.}(DT2) = 1/10 > 0 = E_{antren.}(DT^*)$, în vreme ce $E_{CVLOO}(DT2) = 1/10 < 3/10 = E_{CVLOO}(DT^*)$. Aceasta este un caz tipic de manifestare a fenomenului de *overfitting (supra-specializare)*.

11.

(Arbore de decizie cu variabile de intrare continue;
o margine superioară pentru adâncimea arborilor
în cazul (ne)separabilității liniare în \mathbb{R}^2)

CMU, 2009 spring, Ziv Bar-Joseph, midterm exam, pr. 5.cd

Se consideră n vectori bidimensionali ($x = \{x_1, x_2\}$) care pot fi clasificați folosind o funcție [de regresie] liniară, adică există $w \in \mathbb{R}^2$ și $b \in \mathbb{R}$ astfel încât:

$$y = \begin{cases} +1 & \text{if } w \cdot x + b > 0 \\ -1 & \text{if } w \cdot x + b \leq 0 \end{cases}$$

- Poate un arbore binar de decizie — atenție!, nu neapărat arboarele ID3 — să clasifice corect acești vectori? Dacă nu, justificați. Dacă da, determinați adâncimea maximă (adică numărul maxim de niveluri de test) ale unui astfel de arbore de decizie, care este optim (ca număr de niveluri de test).
- Acum să presupunem că aceste n date nu sunt separabile liniar (adică nu există $w \in \mathbb{R}^2$ și $b \in \mathbb{R}$ cu proprietatea de mai sus). Poate un arbore de decizie (binar) să clasifice corect acești vectori? Dacă nu, justificați. Dacă da, care este adâncimea maximă a unui arbore de decizie corespunzător, optim ca număr de niveluri?

Răspuns:

- Da. O strategie posibilă pentru a construi un arbore de decizie binar care să clasifice corect aceste puncte este următoarea:

Mai întâi vom construi un arbore de decizie considerând doar atributul x_1 . În particular, făcând „înjumătățiri“ successive ale mulțimii de valori ale acestui atribut, arboarele rezultat va avea o adâncime maximă de $\lceil \log_2 n \rceil$ (adică partea întreagă superioară din $\log_2 n$) niveluri. În fiecare nod frunză al acestui arbore se va găsi o mulțime de puncte, însă nu neapărat cu aceeași clasificare. Totuși, deoarece datele sunt liniar separabile, pentru fiecare dintre aceste mulțimi de puncte se poate găsi o valoare a atributului x_2 care să o împartă în două submulțimi clasificate corect.

Prin urmare, arboarele construit inițial considerând doar valoarea x_1 are nevoie să î se adauge doar cel mult câte un nod în fiecare frunză, nod care să clasifice corect datele luând în considerare valoarea x_2 . Adâncimea totală a arborelui astfel construit este $1 + \lceil \log_2 n \rceil$, adică de ordinul $O(\log n)$.

- Similar punctului anterior, se construiește un arbore de decizie considerând doar atributul x_1 , ceea ce înseamnă o adâncime de $\lceil \log_2 n \rceil$. În fiecare

nod frunză al acestui arbore se va găsi o mulțime de puncte, posibil cu clasificări diferite. Datele de antrenament nu mai sunt liniar separabile, deci nu pot fi clasificate corect printr-un singur nod.

Pentru fiecare astfel de nod frunză în care punctele nu au aceeași clasificare, se aplică algoritmul de determinare a arborelui de decizie, de această dată luând în considerare doar valoarea lui x_2 . Aceasta presupune din nou o adâncime maximă a subarborelui de $\lceil \log_2 n \rceil$. În total, arborele obținut are adâncimea maximă $\lceil \log_2 n \rceil + \lceil \log_2 n \rceil$, deci tot de ordinul $O(\log n)$.

12.

(Algoritmul ID3 cu atrbute discrete, respectiv atrbute discrete și un atrbut continuu: aplicare; predicție)

■ • CMU, 2012 fall, E. Xing, A. Singh, HW1, pr. 1.1

Până în luna septembrie a anului 2012, 800 de planete extrasolare (numite în continuare exoplanete) au fost identificate în galaxia noastră. Niște navete spațiale super-secrete au fost trimise pentru a survola toate aceste exoplanete, cu scopul de a stabili dacă ele sunt locuibile sau nu de către oameni. Evident, a trimite câte o navetă spațială la fiecare dintre aceste exoplanete este extrem de costisitor. De aceea, în această problemă vă propunem să elaborați un arbore de decizie pentru a prezice dacă o exoplanetă este locuibilă sau nu, folosind doar trăsături / caracteristici (engl., features) observabile cu ajutorul telescoapelor terestre.

a. În tabelul de mai jos vi se dău anumite date în legătură cu toate cele 800 de planete survolate până acum. Trăsăturile observate cu ajutorul telescoapelor sunt *Size* (“Big” sau “Small”) și *Orbit* (“Near” sau “Far”).

Fiecare linie din tabel indică valori ale acestor două trăsături, caracterul habitabil (“Yes” sau “No”), precum și de câte ori a fost identificată fiecare combinație de valori [pentru cele trei trăsături]. De exemplu, au fost identificate 20 de planete mari (“Big”), care sunt situate pe orbite apropiate (“Near”) de soarele / steaua lor și sunt locuibile.

Size	Orbit	Habitable	Count
Big	Near	Yes	20
Big	Far	Yes	170
Small	Near	Yes	139
Small	Far	Yes	45
Big	Near	No	130
Big	Far	No	30
Small	Near	No	11
Small	Far	No	255

Elaborați și desenați arborele de decizie învățat de către algoritmul ID3 pe aceste date. (Folosiți criteriul câștigului de informație; nu aplicați pruning-ul.) La fiecare nod din arbore, scrieți numărul de planete locuibile și respectiv nelocuibile din datele de antrenament care sunt asociate cu nodul respectiv.

b. Pentru doar 9 dintre aceste exoplanete, a fost măsurată o a treia trăsătură, *Temperature* (exprimată în grade Kelvin), după cum se arată în tabelul de mai jos.

Refaceti toti pasii de la punctul a , de data aceasta folosind toate cele trei trasaturi de intrare. Pentru trasatura *Temperature* (vazuta ca atribut numeric cu valori continue), la fiecare iteratie va trebui sa faceți maximizarea in raport cu toate pragurile adecvate pentru separare binara (engl., binary thresholding splits). Iata un exemplu de test pentru o astfel de separare binara: $T \leq 250$ vs. $T > 250$.

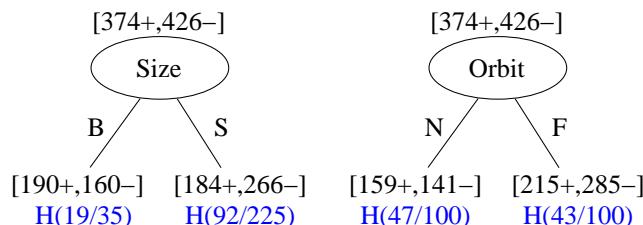
Size	Orbit	Temperature	Habitable
Big	Far	205	No
Big	Near	205	No
Big	Near	260	Yes
Big	Near	380	Yes
Small	Far	205	No
Small	Far	260	Yes
Small	Near	260	Yes
Small	Near	380	No
Small	Near	380	No

- c. Conform arborelui de decizie pe care l-ati obtinut la punctul b , cum va fi clasificata o planetă avand trasaturile (Big, Near, 280), locuibila sau nelocuibilă?

Indicatie: Este posibil sa aveți nevoie de următoarele valori pentru entropia ($H(p)$) unei variabile aleatoare Bernoulli de parametru p : $H(1/3) = 0.9182$, $H(2/5) = 0.9709$, $H(92/225) = 0.9759$, $H(43/100) = 0.9858$, $H(16/35) = 0.9946$, $H(47/100) = 0.9974$.

Răspuns:

- a. „Compașii de decizie“ corespunzători nodului rădăcină sunt infățișați în desenul următor:



Sub fiecare nod descendant am notat entropia nodului respectiv, facand referire la distribuția Bernoulli și dând (de fiecare dată) parametrului acestei distribuții valoarea corespunzătoare. Pentru nodul descendant corespunzător lui *Orbit* = *Near* am ținut cont și de faptul că entropia distribuției Bernoulli, ca funcție de parametru p , este simetrică față de valoarea $1/2$.⁴⁸⁸

Entropiile condiționale medii corespunzătoare acestor doi „compași de decizie“ sunt:

$$\begin{aligned}
 H(Habitable|Size) &= \frac{35}{80} \cdot H\left(\frac{19}{35}\right) + \frac{45}{80} \cdot H\left(\frac{92}{225}\right) \\
 &= \frac{35}{80} \cdot 0.9946 + \frac{45}{80} \cdot 0.9759 = 0.9841
 \end{aligned}$$

⁴⁸⁸Este util să revedeți explicațiile date în *observația importantă* de la pagina 483.

$$\begin{aligned}
 H(Habitable|Orbit) &= \frac{3}{8} \cdot H\left(\frac{47}{100}\right) + \frac{5}{8} \cdot H\left(\frac{43}{100}\right) \\
 &= \frac{3}{8} \cdot 0.9974 + \frac{5}{8} \cdot 0.9858 = 0.9901
 \end{aligned}$$

Suntem acum în măsură să desemnăm atributul care va fi plasat în nodul rădăcină al arborelui care va fi construit de algoritm ID3 pe datele din enunț: este atributul *Size*, întrucât $H(Habitable|Size) < H(Habitable|Orbit)$.

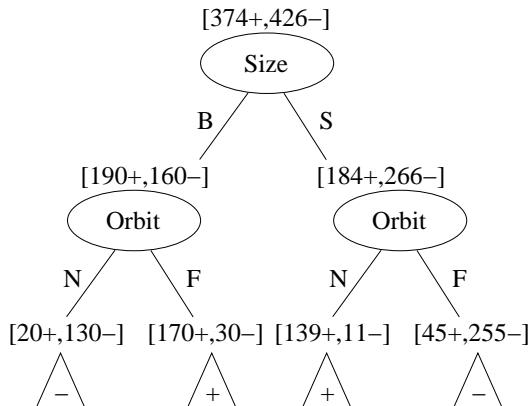
Doar cu titlu de informare, precizăm și câștigurile de informație realizate de cele două atrbute de intrare în raport cu atributul de ieșire:

$$IG(Habitable; Size) = H(Habitable) - H(Habitable|Size) = 0.9969 - 0.9841 = 0.0128$$

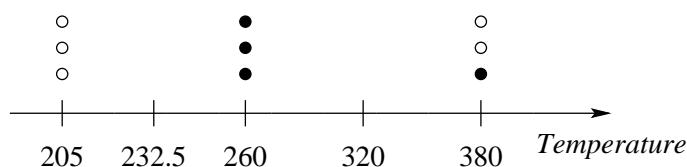
și, similar,

$$IG(Habitable; Orbit) = 0.0067.$$

Întrucât nu avem decât două atrbute de intrare, arborele de decizie va fi:

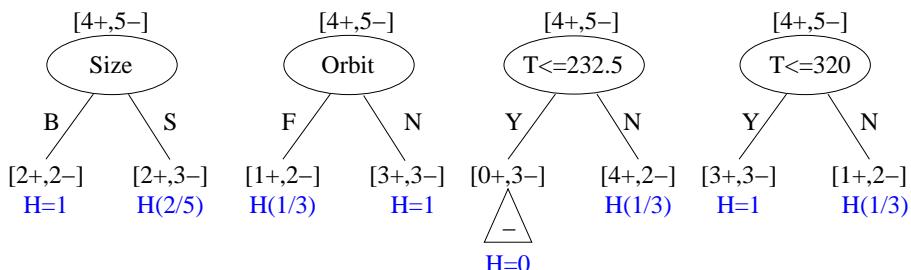


b. Pragurile de separare corespunzătoare atrbutului *Temperature* sunt determinate conform imaginii următoare:



Nivelul 1 (rădăcina):

„Compașii de decizie“ corespunzători acestui nivel sunt:



Observând cu atenție partițiile formate,⁴⁸⁹ vom constata că entropia condițională medie pentru testul $Temperature \leq 232.5$ (în raport cu atributul de ieșire *Habitable*) are o valoare mai mică decât fiecare dintre entropiile condiționale medii $H(Habitable|Orbit)$ și $H(Habitable|Temperature \leq 320)$ (care, de fapt, sunt egale între ele).⁴⁹⁰

Relația dintre $H(Habitable|Temperature \leq 232.5)$ și $H(Habitable|Size)$ se determină cu ajutorul calculelor:

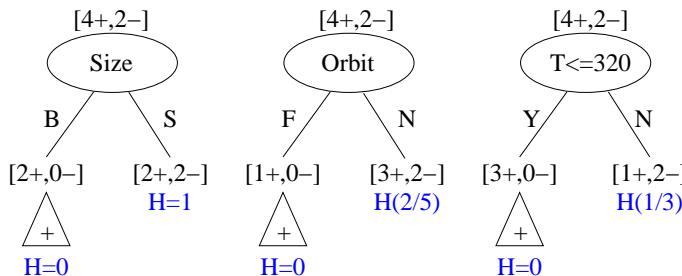
$$\begin{aligned} H(Habitable|Size) &= \frac{4}{9} + \frac{5}{9} \cdot H\left(\frac{2}{5}\right) = \frac{4}{9} + \frac{5}{9} \cdot 0.9709 = 0.9838 \\ H(Habitable|Temp \leq 232.5) &= \frac{2}{3} \cdot H\left(\frac{1}{3}\right) = \frac{2}{3} \cdot 0.9182 = 0.6121. \end{aligned}$$

Deși nu mai este necesar, indicăm și valorile câștigurilor de informație:

$$\begin{aligned} IG(Habitable; Size) &= 0.0072 \\ IG(Habitable; Orbit) &= 0.0183 \\ IG(Habitable; Temp \leq 232.5) &= 0.3788 \\ IG(Habitable; Temp \leq 320) &= 0.0183. \end{aligned}$$

Prin urmare, vom reține pentru nivelul 1 testul $Temperature \leq 232.5$.

Nivelul 2 (mai exact, aici ne limităm la datele cu $Temperature > 232.5$: „Compașii de decizie“ corespunzători [completării] acestui nivel sunt:



Este imediat că, pe aceste date, $H(Habitable|Temp \leq 320) < H(Habitable|Size)$ și, de asemenea, $H(Habitable|Temp \leq 320) < H(Habitable|Orbit)$.⁴⁹¹ Prin urmare, aici va fi ales testul $Temp \leq 320$.

⁴⁸⁹ Se compară două câte două entropiile condiționale specifice, precum și ponderile datelor respective în [raport cu] numărul total de instanțe asociate nodului părinte.

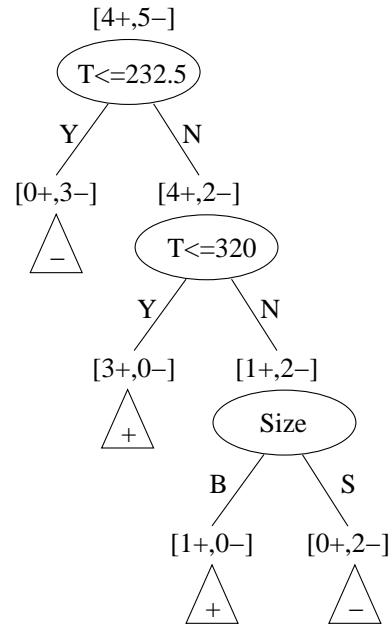
⁴⁹⁰ Observați de exemplu că $H(Habitable|Orbit = F) > H(Habitable|Temperature \leq 232.5)$ și $H(Habitable|Orbit = N) > H(Habitable|Temperature > 232.5)$, iar ponderile corespunzătoare acestor entropii condiționale specifice în calculul entropiilor condiționale medii $H(Habitable|Orbit = N)$ și $Temperature \leq 232.5$ sunt egale două câte două (și anume, cu 3/9 și respectiv 6/9).

⁴⁹¹ Mai exact, ar fi trebuit să scriem: $H(Habitable|Temp > 232.5, Temp \leq 320) < H(Habitable|Temp > 232.5, Size)$ și respectiv $H(Habitable|Temp > 232.5, Temp \leq 320) < H(Habitable|Temp > 232.5, Orbit)$.

Nivelul 3 (mai exact, aici ne limităm la datele cu Temperature > 320):

Se poate observa că, pentru aceste date, atributul *Size* are putere discriminativă maximă. Așadar, arborele de decizie final va fi cel din figura alăturată.

c. Conform arborelui de decizie obținut la punctul anterior, o exoplanetă având trăsăturile (Big, Near, 280) va fi clasificată ca fiind locuibilă.



13.

(Extensiile ale algoritmului ID3:
cazul atributelor discrete cu număr mare de valori)

CMU, 2008(?) spring, HW2, pr. 1

Se dorește antrenarea unui arbore de decizie care să clasifice exemple cu două atrbute de intrare X_1, X_2 și un atrbut de ieșire Y care are valorile 1 și 2. Primul atrbut, X_1 , este binar, pe când al doilea atrbut, X_2 , are 6 valori posibile A, B, C, D, E, F . Se dau următoarele 12 exemple de antrenament, câte 6 din fiecare clasă:

$Y = 1$	(1, A)	(0, E)	(1, B)	(1, B)	(1, F)	(0, D)
$Y = 2$	(0, A)	(0, C)	(1, E)	(0, F)	(0, B)	(1, D)

Vă reamintim formula câștigului de informație la partajarea mulțimii de exemple S în funcție de valorile atrbutului A :

$$Gain(S, A) = H(S) - H(S | A), \text{ cu } H(S | A) = \sum_{v \in \text{valori}(A)} \frac{|S_v|}{|S|} \cdot H(S_v),$$

unde $H(S)$ este entropia mulțimii de exemple S , iar $H(S | A)$ este entropia condițională medie a mulțimii S în raport cu atrbutul A , calculată aşa cum se vede mai sus, ca sumă ponderată a entropiilor submulțimilor lui S determinate de valorile atrbutului A .

- a. Determinați atrbutul ales în rădăcină folosind câștigul de informație. Folosiți aproximarea $\log_2 3 = 1.585$.
- b. Determinați atrbutul ales în nodul rădăcină folosind o măsură numită *gain ratio impurity*, adică alegeti acel atrbut care maximizează raportul

$$\frac{Gain(S, A)}{-\sum_v P(A = v) \cdot \log_2 P(A = v)} = \frac{H(S) - H(S | A)}{H(A)}.$$

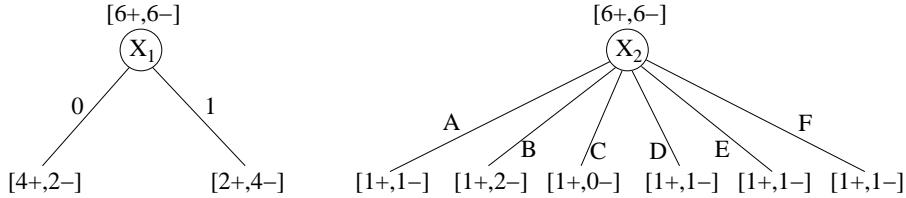
c. Având în vedere rezultatele de la punctele precedente, analizați utilitatea folosirii măsurii *gain ratio impurity* în cazurile în care atributele au numere diferite de valori posibile.

Răspuns:

a. Calculăm în primul rând entropia nodului rădăcină (sau a atributului de ieșire), care are 6 exemple pozitive și 6 negative:

$$H(Y) = H[6+, 6-] = -\frac{6}{12} \log_2 \frac{6}{12} - \frac{6}{12} \log_2 \frac{6}{12} = 2 \cdot \frac{1}{2} \log_2 2 = 1$$

În rădăcină poate fi ales fie atributul X_1 , fie atributul X_2 , obținând următoarele împărțiri:



Dacă vom pune atributul X_1 în nodul rădăcină, câștigul de informație va fi:

$$\begin{aligned} Gain(S, X_1) &= H(Y) - \frac{6}{12} H[4+, 2-] - \frac{6}{12} H[2+, 4-] = \\ &= 1 - \frac{6}{12} \left(-\frac{4}{6} \log_2 \frac{4}{6} - \frac{2}{6} \log_2 \frac{2}{6} \right) - \frac{6}{12} \left(-\frac{2}{6} \log_2 \frac{2}{6} - \frac{4}{6} \log_2 \frac{4}{6} \right) \\ &= 1 - 2 \cdot \frac{1}{2} \left(\frac{1}{3} \log_2 3 + \frac{2}{3} \log_2 \frac{3}{2} \right) = 1 - \left(\log_2 3 - \frac{2}{3} \right) \\ &= \frac{5}{3} - \log_2 3 \approx 0.0817 \end{aligned}$$

Altminteri, punând atributul X_2 în nodul rădăcină, câștigul de informație este:

$$\begin{aligned} Gain(S, X_2) &= H(Y) - 4 \cdot \frac{2}{12} H[1+, 1-] - \frac{3}{12} H[1+, 2-] - \frac{1}{12} H[1+, 0-] \\ &= 1 - \frac{2}{3} \cdot 1 - \frac{3}{12} \left(-\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} \right) - \frac{1}{12} \cdot 0 \\ &= 1 - \frac{2}{3} - \frac{1}{4} \left(\frac{1}{3} \log_2 3 + \frac{2}{3} \log_2 \frac{3}{2} \right) = 1 - \frac{2}{3} - \frac{1}{4} \left(\log_2 3 - \frac{2}{3} \right) \\ &= \frac{1}{2} - \frac{1}{4} \log_2 3 \approx 0.1037 \end{aligned}$$

Folosind drept criteriu de optimizat în fiecare nod câștigul de informație, în rădăcină vom pune atributul X_2 , întrucât $Gain(S, X_1) < Gain(S, X_2)$.

b. Dacă vom pune atributul X_1 în nodul rădăcină, *gain ratio impurity* va fi:

$$\begin{aligned} \frac{Gain(S, X_1)}{-\sum_{i \in \{0,1\}} P(X_1 = i) \cdot \log_2 P(X_1 = i)} &= \frac{Gain(S, X_1)}{-\frac{6}{12} \log_2 \frac{6}{12} - \frac{6}{12} \log_2 \frac{6}{12}} = \frac{Gain(S, X_1)}{1} \\ &\approx 0.0817 \end{aligned}$$

În schimb, plasând atributul X_2 în nodul rădăcină, *gain ratio impurity* va fi:

$$\begin{aligned} \frac{\text{Gain}(S, X_2)}{-\sum_{j \in \{A, \dots, F\}} P(X_2 = j) \cdot \log_2 P(X_2 = j)} &= \frac{\text{Gain}(S, X_2)}{-4 \cdot \frac{2}{12} \log_2 \frac{2}{12} - \frac{3}{12} \log_2 \frac{3}{12} - \frac{1}{12} \log_2 \frac{1}{12}} \\ &= \frac{\text{Gain}(S, X_2)}{\frac{2}{3} \log_2 6 + \frac{1}{4} \log_2 4 + \frac{1}{12} \log_2 12} = \frac{\text{Gain}(S, X_2)}{\frac{4}{3} + \frac{3}{4} \log_2 3} \approx 0.0411 \end{aligned}$$

Prin urmare, în nodul rădăcină vom pune atributul X_1 , pentru care măsura *gain ratio impurity* are valoarea cea mai mare.

c. Câștigul de informație favorizează alegerea atributelor care au un număr mare de valori, indiferent dacă ele determină sau nu partitioarea în mod semnificativ a datelor de antrenament. În schimb, măsura *gain ratio impurity* ia în considerare, prin cantitatea de la numitor (vedeți definiția), numărul de valori ale atributului respectiv, mai exact mărimea mulțimilor de instanțe asignate nodurilor-fii, care au fost generate ca urmare a alegerii respectivelui atribut. Valoarea de la numitor va crește odată cu numărul de noduri-fii, și totodată cu numărul de noduri-fii care au asignate puține exemple. Prin urmare, această măsură penalizează attributele cu multe valori, evitând favorizarea de care se face vinovat câștigul de informație într-o atare situație.

14.

(Extensii ale algoritmului ID3: cazul când se ia în considerare costul calculării atributelor; attribute continue; “decision stumps”)

CMU, 2008 spring, T. Mitchell, W. Cohen, HW1, pr. 1

Se dă următorul set de date. Acesta reprezintă fișele a 12 pacienți ipotetici, ținând cont de sex, vîrstă (peste 60 ani sau nu), dacă suferă sau nu de diabet, dacă au pulsul mărit (sau nu) și EKG-ul anormal (sau nu). Pacienții sunt clasificați în final după cum prezintă (sau nu) aritmie.

Pacient	Sex	Peste60	Diabetic	Puls	EKG	AreAritmie
1	M	1	1	0	0	0
2	M	0	0	1	1	1
3	M	0	1	1	0	0
4	M	1	0	0	1	1
5	M	1	1	1	0	1
6	M	0	1	1	0	1
7	F	0	0	1	0	0
8	F	1	1	1	1	1
9	F	0	1	0	1	1
10	F	1	0	0	0	0
11	F	1	1	0	0	0
12	F	1	0	1	1	1

- a. Calculați entropia condițională specifică $H(\text{AreAritmie} \mid \text{Sex} = F)$.
- b. Dacă pentru selectarea atributelor se folosește măsura $\frac{\text{Gain}^2(S, A)}{\text{Cost}(A)}$ în schimbul câștigului informational, care va fi atributul pus în nodul rădăcină? Se

consideră că $\text{Cost}(\text{Sex}) = \text{Cost}(\text{Peste60}) = 1$, $\text{Cost}(\text{Diabetic}) = 3$, $\text{Cost}(\text{Puls}) = 2$ și $\text{Cost}(\text{EKG}) = 5$.

Observație: În calcule se vor utiliza următoarele aproximări: $\log_2 3 = 1.585$, $\log_2 5 = 2.322$ și $\log_2 7 = 2.807$.

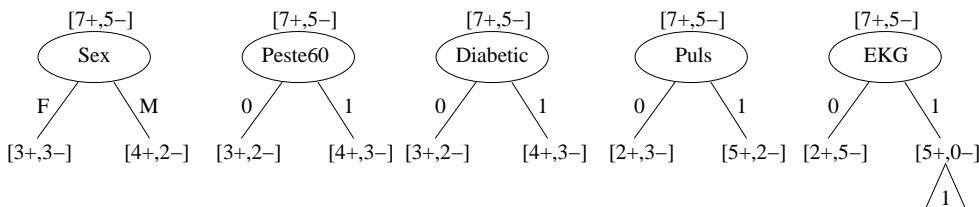
c. Să presupunem că, pentru un alt set de pacienți, se cunoaște vârstă lor exactă. Pentru exemplele pozitive, vârstele sunt: {40, 60, 62, 64, 70, 74, 75, 82}, iar pentru exemplele negative sunt: {33, 35, 42, 45, 49, 52, 58, 59, 80}. Să presupunem că toate celelalte attribute sunt predictori „slabi“, prin urmare dorim ca arborele să aibă un singur nod, rădăcina, care să împartă exemplele cu valori continue ale atributului vârstă în două: $vârstă < k$ și $vârstă \geq k$. Care va fi valoarea aleasă pentru k , bazat pe câștigul de informație?

Răspuns:

a. Entropia condițională cerută este:

$$H(\text{AreAritmie} | \text{Sex} = F) = H[3+, 3-] = 1$$

b. În rădăcina arborelui de decizie se alege atributul A pentru care raportul $\frac{\text{Gain}^2(S, A)}{\text{Cost}(A)}$ este maxim. În cazul nostru, variantele pe care le avem sunt:



Se observă direct că $\text{Gain}(S, \text{EKG})$ este mai mare decât $\text{Gain}(S, A)$ pentru orice atribut $A \neq \text{EKG}$. Însă și $\text{Cost}(\text{EKG})$ este mai mare decât $\text{Cost}(A)$ pentru orice $A \neq \text{EKG}$. Așadar, trebuie să facem calculele în detaliu.

Entropia atributului de ieșire, AreAritmie, este:

$$\begin{aligned} H(\text{AreAritmie}) &= H[7+, 5-] = \frac{7}{12} \cdot \log_2 \frac{12}{7} + \frac{5}{12} \cdot \log_2 \frac{12}{5} \\ &= \frac{7}{12} \cdot \log_2 12 - \frac{7}{12} \cdot \log_2 7 + \frac{5}{12} \cdot \log_2 12 - \frac{5}{12} \cdot \log_2 5 \\ &= \log_2(3 \cdot 4) - \frac{7}{12} \cdot \log_2 7 - \frac{5}{12} \cdot \log_2 5 \\ &= \log_2 3 + \underbrace{\log_2 4}_{=2} - \frac{7}{12} \cdot \log_2 7 - \frac{5}{12} \cdot \log_2 5 \approx 0.98 \end{aligned}$$

Vom calcula câștigul de informație pentru fiecare din cele 5 attribute — se observă că pentru attributele Peste60 și Diabetic, câștigurile de informație sunt egale —, și apoi vom face raportul $\frac{\text{Gain}^2}{\text{Cost}}$:

Pentru atributul Sex:

$$\begin{aligned} \text{Gain}(S, \text{Sex}) &= H(\text{AreAritmie}) - \frac{6}{12}H[3+, 3-] - \frac{6}{12}H[4+, 2-] \\ &= H(\text{AreAritmie}) - \frac{1}{2} \cdot 1 - \frac{1}{2} \left(-\frac{4}{6} \log_2 \frac{4}{6} - \frac{2}{6} \log_2 \frac{2}{6} \right) \end{aligned}$$

$$= 0.98 - \frac{1}{2} - \frac{1}{2} \left(\log_2 3 - \frac{2}{3} \right) = 0.98 - \frac{1}{2} - \frac{1}{2} \log_2 3 + \frac{1}{3} \approx 0.02$$

Prin urmare, $\frac{Gain^2(S, Sex)}{Cost(Sex)} \approx \frac{0.02^2}{1} = 0.0004.$

Pentru atributele Peste60 și Diabetic:

$$\begin{aligned} Gain(S, Peste60) &= Gain(S, Diabetic) \\ &= H(\text{AreAritmie}) - \frac{5}{12}H[3+, 2-] - \frac{7}{12}H[4+, 3-] \\ &= 0.98 - \frac{5}{12} \left(\frac{3}{5} \log_2 \frac{5}{3} + \frac{2}{5} \log_2 \frac{5}{2} \right) - \frac{7}{12} \left(\frac{4}{7} \log_2 \frac{7}{4} + \frac{3}{7} \log_2 \frac{7}{3} \right) \\ &= 0.98 - \frac{5}{12} \left(\log_2 5 - \frac{3}{5} \log_2 3 - \frac{2}{5} \cdot 1 \right) - \frac{7}{12} \left(\log_2 7 - \frac{4}{7} \cdot 2 - \frac{3}{7} \log_2 3 \right) \\ &= 0.98 - \frac{5}{12} \log_2 5 - \frac{7}{12} \log_2 7 + \frac{1}{2} \log_2 3 + \frac{5}{6} \approx 0.0009 \end{aligned}$$

Deci $\frac{Gain^2(S, Peste60)}{Cost(Peste60)} \approx \frac{0.0009^2}{1} = 81 \cdot 10^{-8}$ și $\frac{Gain^2(S, Diabetic)}{Cost(Diabetic)} \approx \frac{0.0009^2}{3} = 27 \cdot 10^{-8}$, ambele fiind niște valori foarte mici.

Pentru atributul Puls:

$$\begin{aligned} Gain(S, Puls) &= H(\text{AreAritmie}) - \frac{5}{12}H[2+, 3-] - \frac{7}{12}H[5+, 2-] \\ &= 0.98 - \frac{5}{12} \left(\frac{2}{5} \log_2 \frac{5}{2} + \frac{3}{5} \log_2 \frac{5}{3} \right) - \frac{7}{12} \left(\frac{5}{7} \log_2 \frac{7}{5} + \frac{2}{7} \log_2 \frac{7}{2} \right) \\ &= 0.98 - \frac{7}{12} \log_2 7 + \frac{1}{4} \log_2 3 + \frac{1}{3} \approx 0.072 \end{aligned}$$

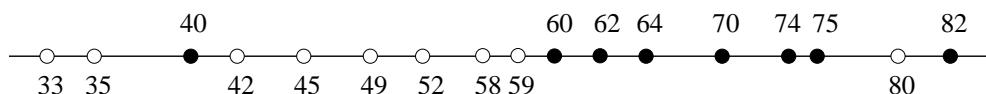
Prin urmare, $\frac{Gain^2(S, Puls)}{Cost(Puls)} \approx \frac{0.072^2}{2} = 0.002592.$

Pentru atributul EKG:

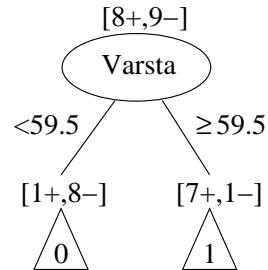
$$\begin{aligned} Gain(S, EKG) &= H(\text{AreAritmie}) - \frac{7}{12}H[2+, 5-] - \frac{5}{12}H[5+, 0-] \\ &= H(\text{AreAritmie}) - \frac{7}{12} \left(\frac{2}{7} \log_2 \frac{7}{2} + \frac{5}{7} \log_2 \frac{7}{5} \right) - \frac{5}{12} \cdot 0 \\ &= 0.98 + \frac{5}{12} \log_2 5 - \frac{7}{12} \log_2 7 + \frac{1}{6} \approx 0.476 \end{aligned}$$

Deci $\frac{Gain^2(S, EKG)}{Cost(EKG)} \approx \frac{0.476^2}{5} = 0.0453152$. Este evident că aceasta este cea mai mare valoare, de aceea în nodul rădăcină va fi ales atributul *EKG*, deși costul acestuia este cel mai mare.

c. Putem reprezenta exemplele astfel:



Se observă că alegerea cea mai bună este $k = 59.5$, arborele de decizie fiind cel din figura alăturată.



15. (Alte criterii posibile pentru selecția atributelor în ID3:
Gini impurity și Misclassification impurity)

■ □ • CMU, 2003 fall, T. Mitchell, A. Moore, HW1, pr. 4

Entropia este o mărime care cuantifică gradul de „impuritate“ (engl., impurity) al unui set de instanțe în raport cu etichetele asignate.⁴⁹² Algoritmul ID3 folosește entropia drept criteriu de partitioare (engl., splitting criterion), calculând *câștigul de informație* pentru a decide care este atributul care trebuie testat în nodul curent. Există, însă, și alte măsuri de impuritate care pot fi folosite, de asemenea, drept criterii de partitioare. În această problemă, vom investiga două astfel de măsuri.

Presupunem că nodul curent (n) din arborele de decizie aflat în curs de elaborare are asignate instanțe din k clase: c_1, c_2, \dots, c_k . Definim⁴⁹³

$$\text{Gini Impurity: } i(n) = 1 - \sum_{i=1}^k P^2(c_i)$$

și

$$\text{Misclassification Impurity: } i(n) = 1 - \max_{i=1}^k P(c_i),$$

unde am notat cu $P(c_i)$ probabilitatea [sau: frecvența de apariție a instanțelor aparținând] clasei c_i în ansamblul instanțelor asignate la nodul curent.

- Presupunem $k = 2$. Așadar, în acest caz nodul n are două clase: c_1 și c_2 . Desenați un grafic în care cele trei măsuri de impuritate — *Entropia*, *Gini Impurity* și *Misclassification Impurity* — sunt reprezentate în funcție de $P(c_1)$.
- Acum putem să definiția unui nou criteriu de partitioare, bazat pe măsurile de impuritate *Gini* și *Misclassification*. În literatura de specialitate, acest nou criteriu este denumit uneori *Drop-of-Impurity* (pentru care propunem ca traducere în limba română termenul de *diminuarea impurității*). El reprezintă diferența dintre impuritatea nodului curent pe de o parte și suma ponderată a impurităților fililor pe de altă parte. În cazul partitioarei atributelor binare, definim *Drop-of-Impurity* ca fiind:

$$\Delta i(n) = i(n) - P(n_l) i(n_l) - P(n_r) i(n_r),$$

⁴⁹²În cazul clasificării binare, unii autori numesc partitioile de forma $[n+, 0-]$ sau $[0+, m-]$ *partitii pure*. Prin opoziție, *partitii impure* sunt de forma $[n+, m-]$, cu $n \neq 0$ și $m \neq 0$. Însă, în contextul termenului de *impuritate* de aici, condiția $n, m \neq 0$ este ignorată.

⁴⁹³Indexul / coeficientul / raportul Gini este denumit după numele creatorului său, Corrado Gini, sociolog și statistician italian (1884-1965).

unde n_l și n_r reprezintă fiul-stânga și, respectiv, fiul-dreapta, care au fost derivați din nodul n după partitioare.

Folosind mai întâi *Gini Impurity* și apoi *Misclassification Impurity*, calculați *Drop-of-Impurity* pentru următorul set de instanțe asignate nodului pentru care se testează atributul A luând valorile a_1 și a_2 . Am notat cu C variabila de ieșire (desemnând clasa) și cu c_1 și c_2 cele două valori ale ei.

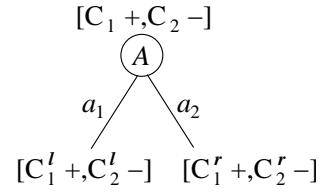
A	a_1	a_1	a_1	a_2	a_2	a_2
C	c_1	c_1	c_2	c_2	c_2	c_2

- c. Se poate crea un set de date de antrenament (sau: se poate modifica setul de date de mai sus) astfel încât, pe noul set, *Drop-of-Impurity* bazat pe *Misclassification* să fie 0 însă bazat pe *Entropy* și, respectiv, *Gini* să fie diferit de 0?

Sugestie: Puteți folosi următoarea proprietate, care este ușor de demonstrat:

Dacă într-un set de date avem C_1 instanțe din clasa (sau: cu eticheta) c_1 și C_2 instanțe din clasa c_2 , cu $C_1 < C_2$, iar după partitioarea în funcție de valoarile atributului A această relație se păstrează, adică $C_1^l < C_2^l$ și $C_1^r < C_2^r$ (evident, cu $C_1 = C_1^l + C_1^r$ și $C_2 = C_2^l + C_2^r$), unde l și r desemnează nodul-fiu stâng și respectiv nodul-fiu drept, atunci *Drop-of-Impurity* pentru *Misclassification* va fi 0. Însă, în aceleasi condiții, *Drop-of-Impurity* bazat pe *Gini* sau pe *Entropy* va avea, în general, valori nenule.

(Evident, proprietatea de mai sus se menține dacă în locul relației $<$ vom considera peste tot relația $>$.)

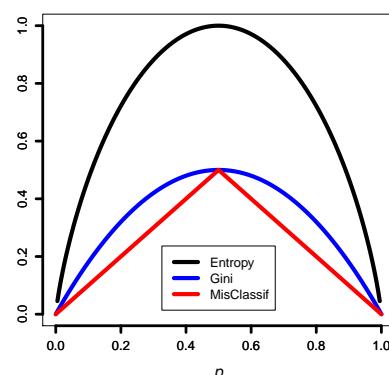


Răspuns:

- a. Vom scrie mai întâi expresiile celor trei funcții, iar apoi vom trasa graficele lor:

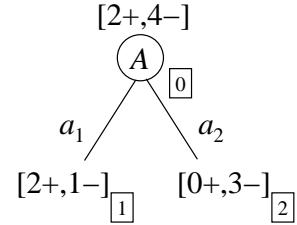
$$\begin{aligned} \text{Entropy}(p) &= -p \log_2 p - (1-p) \log_2(1-p) \\ \text{Gini}(p) &= 1 - p^2 - (1-p)^2 = 2p(1-p) \end{aligned}$$

$$\begin{aligned} \text{MisClassif}(p) &= \\ &= \begin{cases} 1 - (1-p), & \text{pentru } p \in [0; 1/2] \\ 1 - p, & \text{pentru } p \in [1/2; 1] \end{cases} \\ &= \begin{cases} p, & \text{pentru } p \in [0; 1/2] \\ 1 - p, & \text{pentru } p \in [1/2; 1] \end{cases} \end{aligned}$$



Se observă că toate cele trei funcții iau valori în intervalul $[0; 1]$, sunt simetrice în raport cu punctul $p = 1/2$, sunt strict crescătoare pe intervalul $[0; 1/2]$ și strict descrescătoare pe intervalul $[1/2; 1]$, maximul fiecareia dintre ele fiind obținut pentru $p = 1/2$.

- b. Partitionarea datelor de antrenament în funcție de valorile atributului A se face aşa cum se arată în figura alăturată. (În această figură și, de asemenea, în calculele de mai jos, pentru conveniență / simplitate, am asociat semnul + etichetei c_1 și semnul - etichetei c_2 .)



Aplicând formula $\Delta i(n) = i(n) - P(n_l) i(n_l) - P(n_r) i(n_r)$, vom obține pentru Drop-of-Impurity următoarele valori:

Gini: $p = 2/6 = 1/3 \Rightarrow$

$$\left. \begin{array}{l} i(0) = 2 \cdot \frac{1}{3} \left(1 - \frac{1}{3}\right) = \frac{2}{3} \cdot \frac{2}{3} = \frac{4}{9} \\ i(1) = 2 \cdot \frac{2}{3} \left(1 - \frac{2}{3}\right) = \frac{4}{3} \cdot \frac{1}{3} = \frac{4}{9} \\ i(2) = 0 \end{array} \right\} \Rightarrow \Delta i(0) = \frac{4}{9} - \frac{3}{6} \cdot \frac{4}{9} = \frac{4}{9} - \frac{2}{9} = \frac{2}{9}.$$

Misclassification: $p = 1/3 < 1/2 \Rightarrow$

$$\left. \begin{array}{l} i(0) = p = \frac{1}{3} \\ i(1) = 1 - \frac{2}{3} = \frac{1}{3} \\ i(2) = 0 \end{array} \right\} \Rightarrow \Delta i(0) = \frac{1}{3} - \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}.$$

c. Într-adevăr, putem justifica ușor proprietatea din enunț:

$$\begin{aligned} \Delta i(n) &= \frac{C_1}{C_1 + C_2} - \left(\frac{C_1^l + C_2^r}{C_1 + C_2} \cdot \frac{C_1^l}{C_1^l + C_2^r} + \frac{C_1^r + C_2^l}{C_1 + C_2} \cdot \frac{C_1^r}{C_1^r + C_2^l} \right) \\ &= \frac{C_1}{C_1 + C_2} - \frac{C_1^l + C_1^r}{C_1 + C_2} = \frac{C_1}{C_1 + C_2} - \frac{C_1}{C_1 + C_2} = 0. \end{aligned}$$

Dacă în setul de date din enunț una dintre instanțele (a_1, c_1) se modifică în (a_1, c_2) și se adaugă o instanță (a_2, c_1) , atunci Drop-of-Impurity va avea următoarele valori:

Entropy: $\Delta i(0) = H[2+, 5-] - \left(\frac{3}{7}H[1+, 2-] + \frac{4}{7}H[1+, 3-]\right) = 0.006 \neq 0;$

$$\begin{aligned} \text{Gini: } 2 &\left\{ \frac{2}{7} \left(1 - \frac{2}{7}\right) - \left[\frac{3}{7} \cdot \frac{1}{3} \left(1 - \frac{1}{3}\right) + \frac{4}{7} \cdot \frac{1}{4} \left(1 - \frac{1}{4}\right) \right] \right\} = 2 \left\{ \frac{10}{49} - \left[\frac{2}{21} + \frac{3}{28} \right] \right\} \\ &= 2 \left(\frac{10}{49} - \frac{17}{84} \right) \neq 0; \end{aligned}$$

Misclassification: $\Delta i(0) = \frac{2}{7} - \left(\frac{3}{7} \cdot \frac{1}{3} + \frac{4}{7} \cdot \frac{1}{4} \right) = 0.$

16.

(Algoritmul ID3 ca metodă de învățare de tip “eager”: posibilitatea suplimentării datelor de antrenament)

CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh, HW1, pr. 3.4

Presupunem că, pornind de la un set de date de antrenament D , obținem un arbore de decizie ID3, notat cu T . Ulterior, cineva ne mai dă un set suplimentar de date de antrenament, D' . Putem proceda într-unul din următoarele două moduri:

- Putem rula din nou ID3, de această dată pe datele de antrenament $D \cup D'$, obținând arborele T_1 . (Dezavantaj: dacă $|D|$ este foarte mare, această procedură poate fi costisitoare ca timp.)
- Putem extinde T , arborele ID3 obținut pe mulțimea de antrenament D , ținând cont de datele D' . Arborele nou, T_2 , ar putea să nu fie la fel de bun ca T_1 (vedeți cazul de mai sus), dar el este consistent cu datele din $D \cup D'$ în cazul în care aceste mulțimi de antrenament nu conțin zgomote. În special dacă $|D'|$ este mic, această metodă este acceptabilă din punct de vedere practic.

Propuneți o procedură pentru obținerea efectivă a arborelui T_2 .

Răspuns:

Instanțele din D' se atașează nodurilor frunză ale arborelui T , conform procedurii de clasificare de la arbori de decizie. Pentru fiecare dintre aceste noduri frunză, dacă instanțele atașate nodului respectiv nu sunt toate etichetate identic, se aplică algoritmul ID3 (folosind mulțimea de atrbute neutilizate pe drumul care unește nodul rădăcină al arborelui cu acest nod frunză). În acest mod se obține un alt arbore de decizie T_2 care este consistent cu datele din $D \cup D'$.

17.

(Reducerea caracterului “greedy” al algoritmului ID3
prin calcularea câștigului de informație
în maniera “2-step look-ahead”)

CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW4, pr. 1.2-6

Învățarea automată a arborilor de decizie depinde mult de utilizarea unui mecanism “greedy” de selecție a atributelor.

a. Dacă un set de date are A atrbute booleene, calculați (ca expresie în funcție de A) numărul total de apeluri la funcția care calculează câștigul de informație pentru elaborarea întregului arbore de decizie.

Pentru *simplitate*, se va presupune că toate atrbutele sunt necesare pentru clasificarea unei instanțe, iar setul de date de antrenament conține toate instanțele posibile (adică toate combinațiile posibile de perechi atrbut-valoare, inclusiv pentru atrbutul de ieșire).

b. Este posibil să îmbunătățim algoritmul ID3, făcându-l să se comporte mai puțin “greedy”, prin explorarea / prospectarea în avans (engl., look-ahead) a spațiului de căutare. La o explorare cu 2 pași înainte (engl., 2-step look-ahead), calculul câștigului de informație pentru un atrbut a_i pe mulțimea de instanțe D va fi făcut cu ajutorul formulei

$$IG_{2\text{-step}}(D, a_i) = \max_{a_l, a_r} \left\{ \frac{n_l}{n_l + n_r} IG(D_l, a_l) + \frac{n_r}{n_l + n_r} IG(D_r, a_r) \right\},$$

unde

- a_l și a_r sunt atrbutele din nodurile descendente din nodul marcat cu atrbutul a_i ,

- D_l și D_r sunt seturile de instanțe asignate nodurilor descendente din a_i , iar n_l și n_r reprezintă cardinalul mulțimii D_l și respectiv D_r ;
- $IG(D_l, a_l)$ este câștigul de informație calculat (în sens clasic) pentru atributul a_l pe setul D_l ; similar, $IG(D_r, a_r)$ este câștigul de informație pentru atributul a_r pe setul D_r .

b1. Explicați pe scurt de ce sunt necesari factorii $\frac{n_l}{n_l+n_r}$ și $\frac{n_r}{n_l+n_r}$ în formula de mai sus.

b2. Dacă se folosesc A attribute booleane, câte apeluri la funcția $IG(\dots, \dots)$ sunt necesare pentru a stabili atributul din nodul rădăcină?

b3. Vom evalua acum cât de costisitoare este această explorare în avans a spațiului de căutare.

Dacă $A = 10$ și se face aceeași presupozitie ca la punctul a , calculați câte niveluri complete din arborele ID3 standard se pot calcula cu același efort de calcul — exprimat ca număr de apeluri la funcția IG — ca la stabilirea atributului rădăcină în varianta *2-step look-ahead*.

b4. Metoda de învățare a arborilor de decizie folosind *2-step look-ahead* crează o clasă de modele / ipoteze mai largă decât algoritmul ID3 simplu? Altfel spus, putem să calculăm în acest mod funcții de clasificare pe care nu le putem reprezenta cu arborii de decizie standard?

Răspuns:

a. Datorită presupunerii făcute pentru *simplitate*, arborele de decizie final va fi un arbore binar complet cu $A + 1$ niveluri (inclusiv și nodurile de decizie), notate în mod convențional de la 0 la A . Pe fiecare nivel $i = \overline{0, A - 1}$ al arborelui binar se găsesc 2^i noduri. În fiecare dintre aceste noduri poate fi ales un atribut din cele $A - i$ rămase disponibile. Deci pentru determinarea nivelului i se fac $2^i \cdot (A - i)$ apeluri ale funcției IG .

Desigur, pe penultimul nivel, $A - 1$, nu mai există decât un singur atribut rămas disponibil, prin urmare nu este necesar să se calculeze câștigul de informație. Așadar, numărul total de apeluri ale funcției IG pentru elaborarea întregului arbore de decizie este:

$$N_{IG} = A + 2(A - 1) + 4(A - 2) + \dots + 2^{A-2}(A - (A - 2)) = \sum_{i=0}^{A-2} 2^i(A - i)$$

b1. n_l și n_r reprezintă numărul de instanțe asignate nodurilor descendente din a_i , deci factorii $\frac{n_l}{n_l + n_r}$ și $\frac{n_r}{n_l + n_r}$ reprezintă ponderile acestor sub-mulțimi în raport cu reunionea lor. Cei doi factori vor pondera corespunzător câștigurile de informație de pe cele două ramuri din arborele de decizie. Dacă nu facem astfel de ponderări, este posibil să avem un câștig de informație mare pe o mulțime mică sau invers. În consecință, simpla însumare a celor două câștiguri de informație nu ar emula în mod veridic câștigul de informație pe întreg ansamblul lui D .

b2. Pentru stabilirea atributului din nodul rădăcină făcând o explorare cu 2 pași înapoi, se calculează pentru fiecare dintre cele A attribute câștigurile de informație corespunzătoare celor 2 descendenți, adică:

$$N_{IG_{2-step}}(\text{rădăcină}) = A \cdot 2(A - 1) = 2A^2 - 2A$$

b3. Dacă $A = 10$, atunci pentru stabilirea atributului rădăcină în varianta *2-step look-ahead* se apelează funcția IG de $N_{IG_{2-step}}$ (rădăcină) = $200 - 20 = 180$ de ori. Trebuie determinat numărul x de niveluri complete din arborele ID3 standard care se pot calcula folosind maxim 180 de apeluri ale funcției IG , adică argmax_x astfel încât $N_{IG}(x) = \sum_{i=0}^{x-1} 2^i(10 - i) \leq 180$.

$$\begin{aligned} x = 1 &\Rightarrow N_{IG}(1) = 2^0(10 - 0) = 10 \\ x = 2 &\Rightarrow N_{IG}(2) = 10 + 2^1(10 - 1) = 28 \\ x = 3 &\Rightarrow N_{IG}(3) = 28 + 2^2(10 - 2) = 60 \\ x = 4 &\Rightarrow N_{IG}(4) = 60 + 2^3(10 - 3) = 116 \\ x = 5 &\Rightarrow N_{IG}(5) = 116 + 2^4(10 - 4) = 212 > 180 \end{aligned}$$

Așadar, se pot calcula 4 niveluri complete din arborele ID3 standard cu același efort de calcul ca la stabilirea atributului din rădăcină în varianta (mai puțin “greedy”) a algoritmului ID3 cu *2-step look-ahead*.

b4. Nu. Clasa de modele / ipoteze pe care lucrează algoritmul ID3 cu *2-step look-ahead* este aceeași ca la algoritmul ID3 standard. Însă este foarte posibil ca arborele de decizie construit să fie mai bun dacă se face cătarea în maniera *2-step look-ahead*.

18.

(Îmbunătățirea algoritmului ID3, folosind IG cu “2-step look-ahead”)

Liviu Ciortuz, folosind date de la CMU, 2008 fall, Eric Xing, HW2, pr. 3

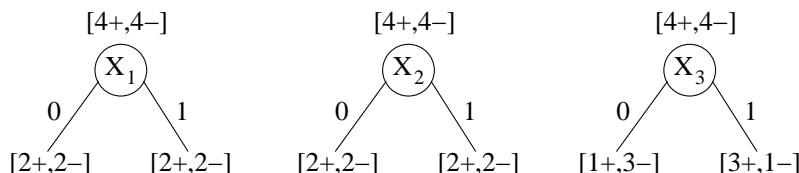
Se consideră variabilele booleene X_1 , X_2 și X_3 , precum și clasificarea $Y = \{0, 1\}$. Fie setul de date de antrenament din tabelul de mai jos.

	X_1	X_2	X_3	Y
1	0	0	0	0
2	0	0	1	0
3	0	1	0	1
4	0	1	1	1
5	1	0	1	1
6	1	0	1	1
7	1	1	0	0
8	1	1	0	0

Răspuns:

a. Elaborați arborele de decizie cu algoritmul ID3 standard.

b. Elaborați arborele de decizie cu algoritmul ID3 folosind câștigul de informație cu *2-step look-ahead*, aşa cum este acesta definit în problema 17.

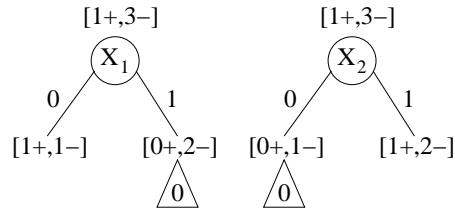


În cele ce urmează vom scrie câștigul de informație (IG) omitând primul argument, întrucât se consideră cunoscut din context. Vom calcula câștigul de informație corespunzător fiecărui atribut:

$$\begin{aligned} IG(X_1) = IG(X_2) &= H[4+, 4-] - \left(\frac{4}{8}H[2+, 2-] + \frac{4}{8}H[2+, 2-] \right) = 1 - \left(\frac{1}{2} + \frac{1}{2} \right) = 0 \\ IG(X_3) &= H[4+, 4-] - \left(\frac{4}{8}H[1+, 3-] + \frac{4}{8}H[3+, 1-] \right) \end{aligned}$$

Însă $H[1+, 3-] = H[3+, 1-]$, iar valoarea corespunzătoare este sub-unitară, deci $IG(X_3) > 0$. Așadar, în nodul rădăcină se alege atributul X_3 .

Pentru nodul corespunzător ramurii $X_3 = 0$ se poate alege dintre atrbutele care au mai rămas, adică X_1 sau X_2 . Se calculează câstigurile de informație corespunzătoare:



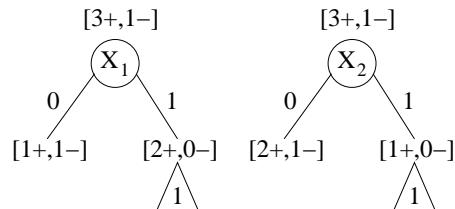
$$\begin{aligned} IG(X_1 | X_3 = 0) &= H[1+, 3-] - \left(\frac{2}{4}H[1+, 1-] + \frac{2}{4}H[0+, 2-] \right) \\ &= 2 - \frac{3}{4}\log_2 3 - \left(\frac{1}{2} + 0 \right) = \frac{3}{2} - \frac{3}{4}\log_2 3 = 0.311 \\ IG(X_2 | X_3 = 0) &= H[1+, 3-] - \left(\frac{1}{4}H[0+, 1-] + \frac{3}{4}H[1+, 2-] \right) \\ &= 2 - \frac{3}{4}\log_2 3 - \left(0 + \frac{3}{4}H[1+, 2-] \right) \\ &= 2 - \frac{3}{4}\log_2 3 - \frac{3}{4} \left(\frac{1}{3}\log_2 3 + \frac{2}{3}\log_2 \frac{3}{2} \right) \\ &= 2 - \frac{3}{4}\log_2 3 - \frac{3}{4} \left(\log_2 3 - \frac{2}{3} \right) = \frac{5}{2} - \frac{3}{2}\log_2 3 = 0.122 \end{aligned}$$

Cum $IG(X_1 | X_3 = 0) > IG(X_2 | X_3 = 0)$, se alege atributul X_1 .

În cele de mai sus, s-au folosit notațiile în maniera condițională $IG(X_1 | X_3 = 0)$ și $IG(X_2 | X_3 = 0)$ doar pentru a identifica în mod neambiguu care este nodul din arbore pentru care se calculează câstigul de informație respectiv.

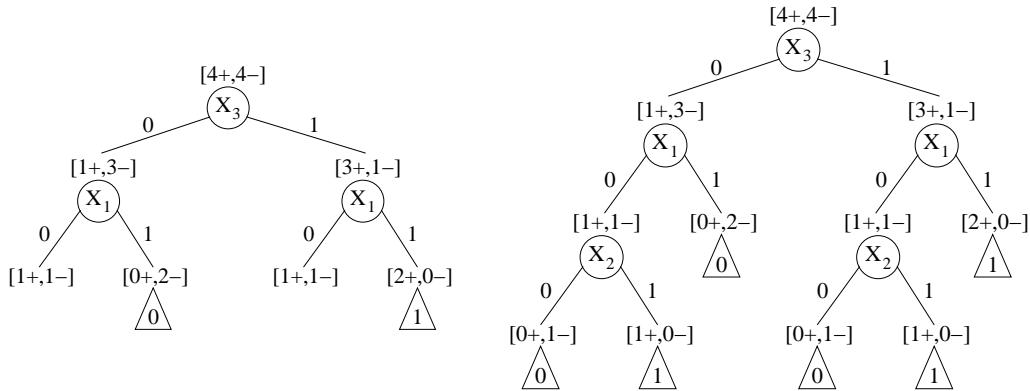
Pentru nodul corespunzător ramurii $X_3 = 1$ ce pornește din nodul rădăcină se poate alege din nou unul dintre atrbutele X_1 și X_2 .

Pentru aceasta, se compară câstigurile de informație corespunzătoare. Se observă că acestea sunt egale cu cele din cazul precedent. Așadar,



$$\begin{aligned} IG(X_1 | X_3 = 1) &= IG(X_1 | X_3 = 0) = \frac{3}{2} - \frac{3}{4}\log_2 3 = 0.311 \\ IG(X_2 | X_3 = 1) &= IG(X_2 | X_3 = 0) = \frac{5}{2} - \frac{3}{2}\log_2 3 = 0.122 \end{aligned}$$

Prin urmare, se alege tot atributul X_1 . Arborele de decizie construit până în acest moment este cel reprezentat mai jos în partea stângă:



În cele două noduri care au rămas, nu poate fi ales decât atributul X_2 . Deci arborele de decizie complet construit de algoritm ID3 este cel reprezentat mai sus în partea dreaptă.

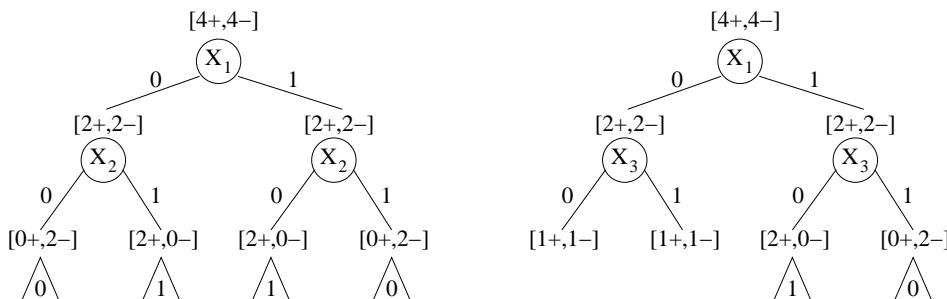
- b. Dacă se folosește câștigul de informație cu *2-step look-ahead* pentru algoritmul ID3, atunci la alegerea fiecărui nod se iau în considerare [și] toate combinațiile posibile de noduri de pe nivelul următor, utilizându-se formula:

$$IG_{2\text{-step}}(D, a_i) = \max_{a_l, a_r} \left\{ \frac{n_l}{n_l + n_r} IG(D_l, a_l) + \frac{n_r}{n_l + n_r} IG(D_r, a_r) \right\}$$

Pentru nodul rădăcină se poate alege, la fel ca la punctul precedent, unul dintre attributele X_1 , X_2 și X_3 . Așadar, câștigul de informație corespunzător atributului X_1 va fi calculat astfel:

$$IG_{2\text{-step}}(X_1) = \max \begin{cases} \frac{4}{8} IG(X_2 \mid X_1 = 0) + \frac{4}{8} IG(X_2 \mid X_1 = 1) \\ \frac{4}{8} IG(X_2 \mid X_1 = 0) + \frac{4}{8} IG(X_3 \mid X_1 = 1) \\ \frac{4}{8} IG(X_3 \mid X_1 = 0) + \frac{4}{8} IG(X_2 \mid X_1 = 1) \\ \frac{4}{8} IG(X_3 \mid X_1 = 0) + \frac{4}{8} IG(X_3 \mid X_1 = 1) \end{cases}$$

Pentru a determina această valoare, va trebui să calculăm cele 4 câștiguri de informație (în sens clasic) implicate în formulă, și este util să reprezentăm două din cele 4 situații posibile:



$$\begin{aligned}
 IG(X_2 | X_1 = 0) &= IG(X_2 | X_1 = 1) = IG(X_3 | X_1 = 1) \\
 &= H[2+, 2-] - \left(\frac{1}{2}H[0+, 2-] + \frac{1}{2}H[2+, 0-] \right) = 1 - 0 = 1 \\
 IG(X_3 | X_1 = 0) &= H[2+, 2-] - \left(\frac{1}{2}H[1+, 1-] + \frac{1}{2}H[1+, 1-] \right) = 1 - \left(\frac{1}{2} + \frac{1}{2} \right) = 0
 \end{aligned}$$

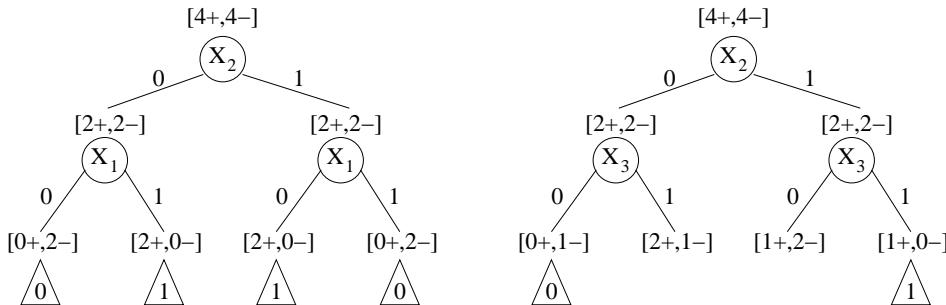
Prin urmare,

$$IG_{2\text{-step}}(X_1) = \max \begin{cases} \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 1 \\ \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 1 \\ \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 1 \\ \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 1 \end{cases} = \max \left\{ 1, 1, \frac{1}{2}, \frac{1}{2} \right\} = 1.$$

Câștigul de informație corespunzător atributului X_2 plasat în nodul rădăcină este:

$$IG_{2\text{-step}}(X_2) = \max \begin{cases} \frac{4}{8}IG(X_1 | X_2 = 0) + \frac{4}{8}IG(X_1 | X_2 = 1) \\ \frac{4}{8}IG(X_1 | X_2 = 0) + \frac{4}{8}IG(X_3 | X_2 = 1) \\ \frac{4}{8}IG(X_3 | X_2 = 0) + \frac{4}{8}IG(X_1 | X_2 = 1) \\ \frac{4}{8}IG(X_3 | X_2 = 0) + \frac{4}{8}IG(X_3 | X_2 = 1) \end{cases}$$

Vom reprezenta din nou două dintre situațiile posibile:



$$\begin{aligned}
 IG(X_1 | X_2 = 0) &= IG(X_1 | X_2 = 1) = H[2+, 2-] - 0 = 1 \\
 IG(X_3 | X_2 = 0) &= IG(X_3 | X_2 = 1) = H[2+, 2-] - \left(\frac{1}{4}H[0+, 1-] + \frac{3}{4}H[2+, 1-] \right) = \\
 &= 1 - \frac{3}{4} \left(\log_2 3 - \frac{2}{3} \right) = \frac{3}{2} - \frac{3}{4} \log_2 3 = 0.311
 \end{aligned}$$

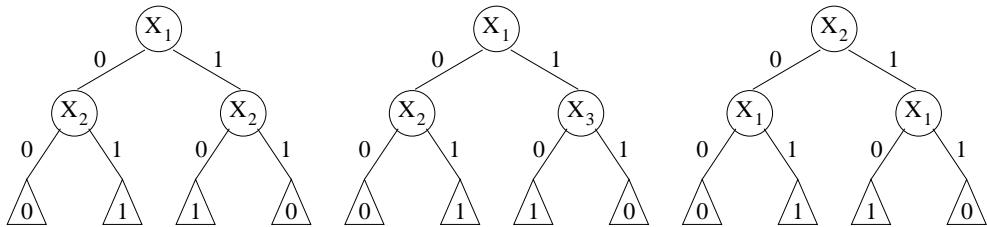
Prin urmare,

$$IG_{2\text{-step}}(X_2) = \max \begin{cases} \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 1 \\ \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 0.311 \\ \frac{1}{2} \cdot 0.311 + \frac{1}{2} \cdot 1 \\ \frac{1}{2} \cdot 0.311 + \frac{1}{2} \cdot 0.311 \end{cases} = \max \{1, 0.655, 0.655, 0.311\} = 1$$

Pentru calculul câștigului de informație corespunzător atributului X_3 plasat în nodul rădăcină, au fost calculate la punctul a cele 4 câștiguri de informație în sens clasic implicate în formulă, deci rezultă:

$$\begin{aligned}
 IG_{2\text{-step}}(X_3) &= \max \left\{ \begin{array}{l} \frac{4}{8}IG(X_1 | X_3 = 0) + \frac{4}{8}IG(X_1 | X_3 = 1) \\ \frac{4}{8}IG(X_1 | X_3 = 0) + \frac{4}{8}IG(X_2 | X_3 = 1) \\ \frac{4}{8}IG(X_2 | X_3 = 0) + \frac{4}{8}IG(X_1 | X_3 = 1) \\ \frac{4}{8}IG(X_2 | X_3 = 0) + \frac{4}{8}IG(X_2 | X_3 = 1) \end{array} \right. \\
 &= \max \left\{ \begin{array}{l} \frac{1}{2}\left(\frac{3}{2} - \frac{3}{4}\log_2 3\right) + \frac{1}{2}\left(\frac{3}{2} - \frac{3}{4}\log_2 3\right) \\ \frac{1}{2}\left(\frac{3}{2} - \frac{3}{4}\log_2 3\right) + \frac{1}{2}\left(\frac{5}{2} - \frac{3}{2}\log_2 3\right) \\ \frac{1}{2}\left(\frac{5}{2} - \frac{3}{2}\log_2 3\right) + \frac{1}{2}\left(\frac{3}{2} - \frac{3}{4}\log_2 3\right) \\ \frac{1}{2}\left(\frac{5}{2} - \frac{3}{2}\log_2 3\right) + \frac{1}{2}\left(\frac{5}{2} - \frac{3}{2}\log_2 3\right) \end{array} \right. = \max \left\{ \begin{array}{l} \frac{3}{2} - \frac{3}{4}\log_2 3 \\ 2 - \frac{9}{8}\log_2 3 \\ 2 - \frac{9}{8}\log_2 3 \\ \frac{5}{2} - \frac{3}{2}\log_2 3 \end{array} \right. \\
 &= \max\{0.311, 0.216, 0.216, 0.122\} = 0.311
 \end{aligned}$$

Comparând $IG_{2\text{-step}}(X_1)$, $IG_{2\text{-step}}(X_2)$ și $IG_{2\text{-step}}(X_3)$, obținem valoarea maximă 1 fie pentru X_1 , fie pentru X_2 în nodul rădăcină. Având în vedere calculele realizate pentru a obține aceste valori, nu mai sunt necesare operații suplimentare pentru determinarea arborelui de decizie construit de ID3 cu metoda *2-step look-ahead*. Există de fapt 3 arbori de decizie optimi (ca număr de niveluri și / sau noduri) pentru aceste date de antrenament, și anume:



Este important de remarcat faptul că algoritmul ID3 cu *2-step look-ahead* identifică toate aceste trei soluții optimale, în vreme ce algoritmul ID3 standard nu identifică niciuna dintre ele.

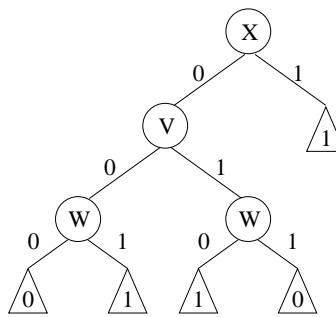
19.

(O strategie de pruning pentru arborele ID3:
eliminarea nodurilor cu $IG < \varepsilon$;
explorare top-down vs. bottom-up)

■ □ • ○ CMU, 2006 spring, Carlos Guestrin, midterm exam, pr. 4

Puteți constata ușor că, aplicând algoritmul ID3 pe datele din tabelul de mai jos, se va obține arborele de decizie alăturat.

V	W	X	Y
0	0	0	0
0	1	0	1
1	0	0	1
1	1	0	0
1	1	1	1



a. Pentru un astfel de arbore de decizie, o strategie simplă de trunchiere (engl., pruning) în vederea contracărării fenomenului de “overfitting” constă în a parcurge arborele de sus în jos, începând deci cu nodul-rădăcină și identificând fiecare nod de test pentru care câștigul de informație (sau un alt criteriu fixat în avans) are o valoare mai mică decât o valoare pozitivă, mică, fixată de la început, ε . Orice astfel de nod de test este imediat înlocuit — împreună cu subarborele corespunzător lui — cu un nod de decizie, conform etichetei majoritară a instanțelor asignate nodului de test. Această strategie se numește “top-down pruning”.

Care este arborele de decizie obținut aplicând această strategie pe arborele de mai sus, dacă se consideră $\varepsilon = 0.0001$? Care este eroarea la antrenare pentru noul arbore?

b. O altă posibilitate de a face pruning este să parcurgem arborele de decizie începând cu părinții nodurilor-frunză și să eliminăm în mod recursiv acele noduri de test pentru care câștigul de informație (sau un alt criteriu ales) este mai mic decât ε . Aceasta este strategia de “bottom-up pruning”.

Observație: Spre deosebire de strategia top-down, în varianta de pruning de tip bottom-up nu vor fi eliminate noduri (cu $IG < \varepsilon$) pentru care există descendenți al căror câștig de informație este mai mare sau egal cu ε .

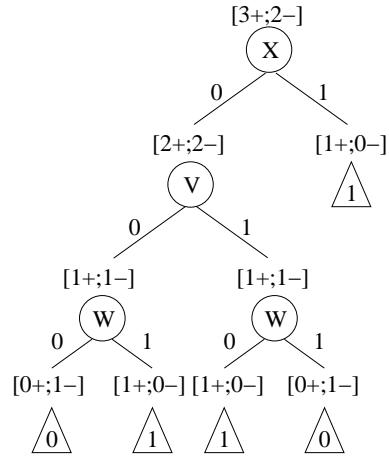
Ce arbore se obține făcând “bottom-up pruning” pe arborele dat mai sus, dacă se consideră $\varepsilon = 0.0001$? Care este eroarea la antrenare pentru arborele rezultat?

c. Stabiliți în ce situații ar fi indicat să alegem strategia “bottom-up pruning” în loc de “top-down pruning” și viceversa. Comparați acuratețea la antrenare și complexitatea computațională a celor două strategii de pruning.

d. Cât este înălțimea — adică, numărul de niveluri de test — pentru arborele returnat de ID3 urmat de “bottom-up pruning”? Puteți găsi un arbore de decizie având o înălțime mai mică, dar care clasifică perfect setul de antrenament? Ce concluzie putem trage despre calitatea [outputului] algoritmului ID3?

Răspuns:

Înainte de a rezolva efectiv punctele a și b , vom augmenta arborele de decizie dat cu informațiile referitoare la numărul de instanțe (pozitive și, respectiv, negative) asignate fiecărui nod de test. Obținem astfel figura alăturată.



a. Câștigul de informație al atributului X plasat în nodul-rădăcină este:

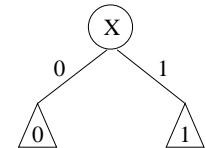
$$H[3+;2-] - 1/5 \cdot 0 - 4/5 \cdot 1 = 0.971 - 0.8 = 0.171 > \varepsilon.$$

Prin urmare, acest nod nu va fi eliminat din arbore.

Câștigul de informație al atributului V este:

$$H[2+;2-] - 1/2 \cdot 1 - 1/2 \cdot 1 = 1 - 1 = 0 < \varepsilon.$$

Așadar, nodul reprezentat de atributul V va fi eliminat și vom obține arborele de decizie [trunchiat], reprezentat în figura alăturată. Menționăm că am fi putut la fel de bine (din punctul de vedere al numărului de erori la antrenare) să alegem decizia $Y = 1$ în nodul-fiu stâng, însă în acel caz arborele s-ar fi redus de fapt la un singur nod de decizie (cu outputul $Y = 1$).



Eroarea la antrenare produsă de acest arbore este $2/5$.

b. Câștigul de informație al celor două noduri marcate cu atributul W în arborele dat în enunț este același, și anume 1 (se poate verifica imediat). Prin urmare, aplicând strategia de “bottom-up pruning”, arborele de decizie rămâne identic cu cel inițial. Evident, eroarea la antrenare pentru acest arbore este 0, întrucât datele de antrenament nu conțin inconsistențe.

c. Din cauza faptului că la top-down pruning, odată cu un nod de test pentru care câștigul de informație (IG) este mai mic decât valoarea ε se elimină întregul subarbore care are ca rădăcină acel nod de test, această strategie este mai rapidă decât (sau, în cel mai rău caz, la fel de rapidă / lentă ca și) pruning-ul de tip bottom-up.

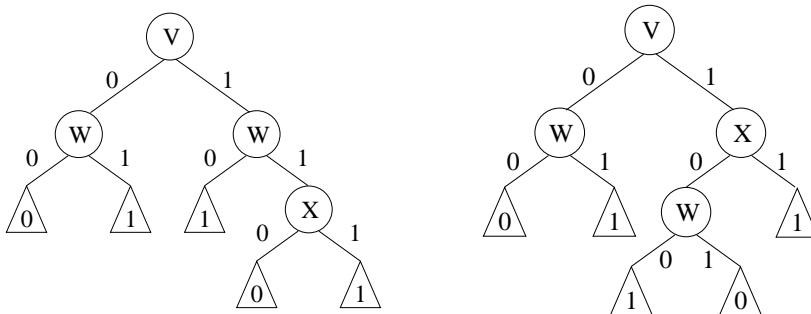
Așa cum s-a menționat în *observația* din enunț și s-a exemplificat apoi la punctele a și b , dezavantajul pruning-ului de tip top-down este că odată ce este eliminat un nod cu IG mai mic decât ε , este posibil ca între descendenții săi (eliminați) să fie și noduri care au câștigul de informație mai mare sau egal cu ε . Pruning-ul bottom-up nu are acest dezavantaj; el este deci mai „conservativ“.

O problemă care poate însă să apară și în cazul pruning-ului de tip bottom-up este faptul că în nodurile apropiate de nodurile de decizie (acestea din urmă fiind nodurile-frunză), câștigul de informație se calculează uneori pe mulțimi mici de exemple, deci testul $IG \geq \varepsilon$ nu este neapărat semnificativ din punct de vedere statistic. Așadar, este *recomandabil* ca în astfel de situații să se folosească un test statistic, de exemplu *testul χ^2* . (A se vedea problemele 20 și 55.)

În ce privește comparația dintre acuratețile arborilor obținuți prin aplicarea celor două variante de pruning: se observă că arborele mai simplu (cel de la punctul *a*) are o eroare la antrenare mai mare decât arborele mai complex (cel de la punctul *b*), însă este mai probabil ca acesta din urmă să producă overfitting.

d. Din rezolvarea dată la punctul *b*, rezultă imediat că înălțimea arborelui obținut prin aplicarea algoritmului ID3 urmat de pruning de tip bottom-up este 3. (Nu se iau în considerare nodurile frunză.) Vom demonstra — exact ca la problema 1 — că nu există un arbore de decizie consistent cu datele de antrenament, care să aibă adâncimea strict mai mică decât 3:

Se observă ușor din tabelul dat în enunț că $Y = (V \text{ XOR } W) \vee X$, așadar variabilele V și W au rol simetric în definirea funcției reprezentate de variabila Y . Punând atributul X în nodul rădăcină, arborele de decizie minimal care se poate obține este cel dat în enunț (sau, echivalent, arborele care se obține din acesta interschimbând V și W). Dacă, în schimb, punem atributul V în nodul rădăcină, se pot obține doi arbori de decizie minimali, aşa cum se arată grafic mai jos. (Și similar, dacă în nodul rădăcină punem atributul W .)



20.

(ID3 cu post-pruning: folosirea testului statistic χ^2 pentru limitarea overfitting-ului)

*prelucrare de Liviu Ciortuz, după
■ • CMU, 2010 fall, Ziv Bar-Joseph, HW2, pr. 2.1*

În acest exercițiu vom face pruning asupra unui arbore ID3 (după ce s-a făcut antrenarea pe toate datele disponibile), folosind o metodă statistică de testare / verificare a ipotezelor.

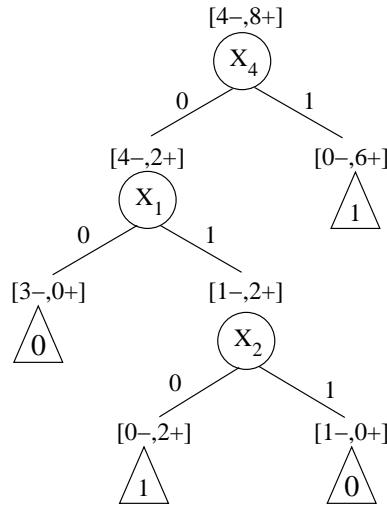
După ce a fost învățat arboarele de decizie, vizităm fiecare nod intern (inclusiv nodul rădăcină) și testăm dacă atributul care a fost pus în nodul respectiv nu este cumva necorelat cu eticheta / clasa specificată de către atributul de ieșire.

Pentru aceasta, mai întâi presupunem că atributul din nodul respectiv este independent de atributul de ieșire (aceasta este aşa-numita „ipoteză nulă“),

iar apoi folosim testul χ^2 al lui Pearson pentru a genera o „statistică“, care poate constitui temeiul pentru respingerea „ipotezei nule“. Dacă ipoteza nulă nu poate fi respinsă, eliminăm sub-arborele din nodul respectiv (de fapt, îl înlocuim cu un nod de decizie).

Pentru a ilustra aceste chestiuni, considerăm arborele de decizie de mai jos (partea dreaptă); el a fost construit pornind de la datele din tabelul alăturat lui, folosind algoritmul ID3.

X_1	X_2	X_3	X_4	Class
1	1	0	0	0
1	0	1	0	1
0	1	0	0	0
1	0	1	1	1
0	1	1	1	1
0	0	1	0	0
1	0	0	0	1
0	1	0	1	1
1	0	0	1	1
1	1	0	1	1
1	1	1	1	1
0	0	0	0	0



a. Pentru fiecare nod intern din arborele de decizie vom crea o *tabelă de contingență* (engl., contingency table) pentru exemplele de antrenare care sunt asignate nodului respectiv. Tabela aceasta va avea coloanele etichetate cu cele c clase / valori ale variabilei de ieșire. Similar, valorile atributului testat în nodul respectiv (având în total r valori) vor fi asignate liniilor talelei. Dacă acceptăm o ușoară simplificare în forma de exprimare, vom putea spune că un element oarecare $O_{i,j}$ din tabela de contingență reprezintă numărul de „observații“ (adică, instanțe de antrenament asignate nodului respectiv) pentru care valoarea atributului testat este i , iar eticheta / clasa este j .

Calculați tabelele de contingență pentru cele trei noduri de test ale arborelui de decizie dat mai sus. Apoi, pornind de la datele conținute în fiecare dintre aceste matrice de contingență, estimați (în sensul verosimilității maxime) probabilitățile pentru valorile variabilei de ieșire (*Class*), precum și pentru valorile atributului din nodul corespunzător.

b. Pentru a aplica testul statistic χ^2 , avem nevoie să calculăm pentru fiecare nod intern al arborelui de decizie încă o tabelă (pe care o vom nota cu E), în care să consemnăm *numărul așteptat* de apariții (engl., expected counts) ale instanțelor de antrenament la nodul respectiv, pentru fiecare pereche de indici i, j având semnificația de mai sus. Acest număr așteptat este numărul (mediu) de instanțe de antrenament pe care le-am „observat“ în nodul respectiv dacă atributul selectat și clasa (variabila de ieșire) ar fi independente.

Derivați o formulă pentru calculul fiecărui element (notat $E_{i,j}$) din această a doua tabelă.

Întrebări ajutătoare: Care este probabilitatea ca exemplele de antrenament asignate nodului respectiv să aibă o anumită etichetă (j)? Tinând cont de

această probabilitate, precum și de presupoziția de independentă formulată prin ipoteza „nulă“, care este numărul de exemple cu o anumită valoare (i) pentru atributul selectat în nodul respectiv, care ar trebui (i.e., „ne așteptăm“) să aibă acea etichetă / clasă (j)?

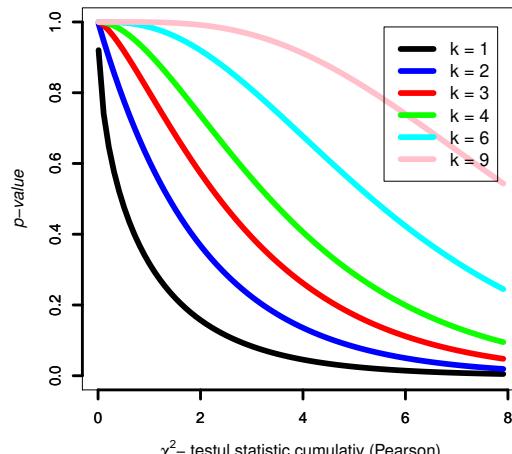
Folosind formula pe care tocmai ați derivat-o, calculați matricea E pentru fiecare dintre cele trei noduri de test ale arborelui de decizie dat mai sus.

c. Date fiind cele două tabele pentru nodul considerat, puteți calcula acum testul statistic χ^2 -pătrat:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

Puteți introduce valoarea calculată χ^2 precum și numărul de grade de libertate $(r-1)(c-1)$ într-un program⁴⁹⁴ sau într-un calculator on-line⁴⁹⁵ pentru a calcula o așa-numită p -valoare (engl., p -value).⁴⁹⁶

În general, dacă $p < 0.05$, se va considera că nu avem suficientă evidență în favoarea ipotezei „nule“, care afirma că atributul selectat și clasa (variabila de ieșire) sunt independente, și o vom respinge. Într-o astfel de situație, spunem că testul [din nodul respectiv] este *semnificativ* din punct de vedere statistic.



Pentru fiecare dintre cele trei noduri interne din arborele de decizie dat mai sus, găsiți p -valoarea corespunzătoare și precizați dacă testul din nodul respectiv este sau nu semnificativ d.p.v. statistic. Cât de multe noduri interne vor fi eliminate din arbore dacă la pruning impunem condiția $p \geq 0.05$ [pentru a elimina un nod de test din arbore și a-l înlocui cu un nod de decizie]?

Răspuns:

a. Pentru fiecare nod din arborele ID3 vom alcătuiri matricea de contingență asociată, pornind de la partiziionările mulțimilor de exemple care au fost asignate (de către algoritmul ID3) descendenților nodului respectiv. Apoi, din

⁴⁹⁴Folosiți 1-chi2cdf(x,df) în MATLAB sau CHIDIST(x,df) în Excel.

⁴⁹⁵<https://www.socscistatistics.com/pvalues/chidistribution.aspx> (accesat la 9.11.2022) este un astfel de calculator.

⁴⁹⁶Statistica χ^2 este aproximată de distribuția χ^2 cu numărul corespunzător (k) de grade de libertate. (Distribuția χ^2 reprezintă suma pătratelor a k variabile gaussiene standard independente.)

p -valoarea despre care este vorba mai sus reprezintă probabilitatea ca distribuția χ^2 să ia valori mai mari sau egale cu valoarea considerată (i.e., valoarea calculată pentru statistică χ^2). Așadar, p -valoarea pentru testul χ^2 se calculează făcând diferența dintre 1 și valoarea funcției de distribuție cumulative (c.d.f.) pentru distribuția χ^2 cu k de grade de libertate. Vedeți site-ul

https://en.m.wikipedia.org/wiki/Chi-square_distribution#Table_of_CF.872_value_vs_p-value (accesat la 5.09.2015).

fiecare matrice de contingență vom estima (în sensul verosimilității maxime) probabilitățile pentru valorile variabilei de ieșire (*Class*), precum și probabilitățile pentru valorile atributului din acel nod. (Atenție la condiționarea probabilităților!)

$$\begin{array}{c|cc} O_{X_4} & \text{Class} = 0 & \text{Class} = 1 \\ \hline X_4 = 0 & 4 & 2 \\ X_4 = 1 & 0 & 6 \end{array} \xrightarrow{N=12} \begin{cases} P(X_4 = 0) = \frac{6}{12} = \frac{1}{2}, P(X_4 = 1) = \frac{1}{2} \\ P(\text{Class} = 0) = \frac{4}{12} = \frac{1}{3}, P(\text{Class} = 1) = \frac{2}{3} \end{cases}$$

$$\begin{array}{c|cc} O_{X_1|X_4=0} & \text{Class} = 0 & \text{Class} = 1 \\ \hline X_1 = 0 & 3 & 0 \\ X_1 = 1 & 1 & 2 \end{array} \xrightarrow{N=6} \begin{cases} P(X_1 = 0 | X_4 = 0) = \frac{3}{6} = \frac{1}{2} \\ P(X_1 = 1 | X_4 = 0) = \frac{1}{2} \\ P(\text{Class} = 0 | X_4 = 0) = \frac{4}{6} = \frac{2}{3} \\ P(\text{Class} = 1 | X_4 = 0) = \frac{1}{3} \end{cases}$$

$$\begin{array}{c|cc} O_{X_2|X_4=0, X_1=1} & \text{Class} = 0 & \text{Class} = 1 \\ \hline X_2 = 0 & 0 & 2 \\ X_2 = 1 & 1 & 0 \end{array} \xrightarrow{N=3} \begin{cases} P(X_2 = 0 | X_4 = 0, X_1 = 1) = \frac{2}{3} \\ P(X_2 = 1 | X_4 = 0, X_1 = 1) = \frac{1}{3} \\ P(\text{Class} = 0 | X_4 = 0, X_1 = 1) = \frac{1}{3} \\ P(\text{Class} = 1 | X_4 = 0, X_1 = 1) = \frac{2}{3} \end{cases}$$

b. Considerăm i o valoare arbitrar aleasă (dar fixată) pentru atributul de intrare A care este testat în nodul curent, iar j o valoare arbitrar aleasă (de asemenea, fixată) pentru atributul de ieșire *Class* (renotat cu C). Înănd cont de presupozitia de independentă stipulată de ipoteza „nulă“, putem scrie:

$$P(A = i, C = j) = P(A = i) \cdot P(C = j)$$

Probabilitățile $P(A = i)$ și $P(C = j)$ pot fi estimate — în sensul verosimilității maxime (MLE) —, cu ajutorul celor N instanțe de antrenament asignate nodului respectiv. Instanțele pentru care atributul A are valoarea i sunt tocmai cele din linia i a matricei. În mod similar, instanțele care au eticheta / clasa j sunt cele din coloana j a matricei de count-uri observate. Așadar,

$$P(A = i) = \frac{\sum_{k=1}^c O_{i,k}}{N} \text{ și } P(C = j) = \frac{\sum_{k=1}^r O_{k,j}}{N}$$

În consecință,

$$P(A = i, C = j) = \frac{(\sum_{k=1}^c O_{i,k})(\sum_{k=1}^r O_{k,j})}{N^2},$$

iar valoarea așteptată — repetăm, în condițiile presupozitiei de independentă — pentru numărul de instanțe având atributul $A = i$ și clasa $C = j$ va fi dată de formula

$$E_{i,j} = N \cdot P(A = i, C = j) = \frac{(\sum_{k=1}^c O_{i,k})(\sum_{k=1}^r O_{k,j})}{N}$$

Folosind probabilitățile calculate la punctul precedent și ținând cont de presupoziția de independentă, calculăm numărul de observații așteptate în fiecare nod, pentru a completa matricele E :

E_{X_4}	$Class = 0$	$Class = 1$	$E_{X_1 X_4=0}$	$Class = 0$	$Class = 1$
$X_4 = 0$	2	4	$X_1 = 0$	2	1
$X_4 = 1$	2	4	$X_1 = 1$	2	1
$E_{X_2 X_4=0, X_1=1}$	$Class = 0$	$Class = 1$			
$X_2 = 0$	$\frac{2}{3}$	$\frac{4}{3}$			
$X_2 = 1$	$\frac{1}{3}$	$\frac{2}{3}$			

Ca să exemplificăm cum am procedat, detaliem mai jos calculul pentru primul element din matricea E_{X_4} :

$$N = 12, P(X_4 = 0) = \frac{1}{2} \text{ și } P(Class = 0) = \frac{1}{3} \Rightarrow$$

$$N \cdot P(X_4 = 0, Class = 0) = N \cdot P(X_4 = 0) \cdot P(Class = 0) = 12 \cdot \frac{1}{2} \cdot \frac{1}{3} = 2$$

c. Aplicăm pentru fiecare nod din arborele de decizie formula de calcul a valorilor / statisticilor χ^2 care ne-a fost dată în enunț:

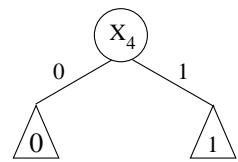
$$\chi^2_{x_4} = \frac{(4-2)^2}{2} + \frac{(2-4)^2}{4} + \frac{(0-2)^2}{2} + \frac{(6-4)^2}{4} = 2+1+2+1=6$$

$$\chi^2_{x_1|x_4=0} = \frac{(3-2)^2}{2} + \frac{(0-1)^2}{1} + \frac{(1-2)^2}{2} + \frac{(2-1)^2}{1} = 3$$

$$\chi^2_{x_2|x_4=0, x_1=1} = \frac{\left(0-\frac{2}{3}\right)^2}{\frac{2}{3}} + \frac{\left(2-\frac{4}{3}\right)^2}{\frac{4}{3}} + \frac{\left(1-\frac{1}{3}\right)^2}{\frac{1}{3}} + \frac{\left(0-\frac{2}{3}\right)^2}{\frac{2}{3}} = \frac{4}{9} \cdot \frac{3}{4} \cdot 9 = 3$$

Accesând pagina web indicată în enunț, am obținut p -valorile următoare: 0.0143, 0.0833 și 0.0833.

În consecință, cu un grad de încredere de cel puțin 95%, nodurile situate pe nivelurile 1 și 2 din arborele ID3 pot fi eliminate. Pentru nodul rădăcină, ipoteza „nulă“ nu se verifică, adică variabilele X_4 și $Class$ nu sunt independente. Arborele obținut în urma pruning-ului este cel din figura alăturată.



Observație: Este de remarcat faptul că arborele ID3 furnizat în enunț are (întâmplător) atât pentru atributul X_4 (din nodul rădăcină) cât și pentru nodul X_1 (de pe primul nivel) același câștig de informație, 0.4591. În urma testului χ^2 , se va elibera însă doar nodul care-l conține pe X_1 . Valoarea statisticii χ^2 asociate nodului X_1 este mult mai mică (3, față de 6, cât este pentru nodul care-l conține pe X_4), deci vom avea suficientă „evidență“ pentru a trage concluzia, cu un grad de încredere de 95%, că variabila de ieșire, $Class$, și $X_1|X_4 = 0$ sunt independente.⁴⁹⁷ Dacă am fi aplicat o metodă de pruning

⁴⁹⁷ Observați și faptul că nodul care conține atributul X_1 are mai puține instanțe asociate (și anume, jumătate) față de cele asociate nodului rădăcină.

bazată pe câștigul de informație, aceasta n-ar fi putut să trateze în mod diferit cele două noduri, adică să-l păstreze pe unul și să-l elimine pe celălalt.

21. (Adevărat sau Fals?)

a.

Liviu Ciortuz

Algoritmul ID3 garantează obținerea arborelui de decizie optimal (ca număr de niveluri sau de noduri).

CMU, 2002 spring, A. Moore, midterm example questions, pr. 1.c

b. Întrucât arborii de decizie pot învăța să clasifice instanțe într-un număr discret de clase (deci nu învăță funcții cu valori reale), este imposibil ca ei să manifeste fenomenul de overfitting.

CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh, midterm, pr. 1.4

c. Fie A și B doi algoritmi de clasificare automată. Algoritmul A este mai bun decât algoritmul B dacă eroarea la antrenare a algoritmului A este mai mică decât eroarea la antrenare a algoritmului B . Justificați.

CMU, 2005 spring, C. Guestrin, T. Mitchell, midterm, pr. 2.c

d. Presupunem că avem m instanțe și că vom folosi jumătate dintre ele pentru antrenarea unui clasificator oarecare (nu neapărat ID3) și jumătate pentru testare. Diferența dintre eroarea la antrenare și eroarea la testare descrește pe măsură ce numărul m crește.

Răspuns:

a. Fals. ID3 nu garantează obținerea arborelui optim (relativ la numărul de niveluri și / sau noduri), ci încearcă să găsească o soluție convenabilă, însă fără să caute în mod exhaustiv în tot spațiul soluțiilor. Mai exact, căutarea soluției se face în manieră “greedy”, maximizând un anumit criteriu (e.g., câștigul de informație) la fiecare iterație. O astfel de căutare nu garantează obținerea optimului.

b. Fals. Arborii de decizie manifestă fenomenul de overfitting. Prima parte a afirmației din enunț este adevarată — arborii de decizie pot învăța să clasifice instanțe într-un număr discret de clase —, însă a doua parte este falsă, deci avem o implicație de forma: $T \rightarrow F \equiv \neg T \vee F \equiv F$.

c. Fals, fiindcă la testare algoritmul B poate să aibă o eroare mai mică decât algoritmul A . Într-un astfel de caz, se spune că algoritmul A este “overfit” (rom., supra-antrenat).

d. Adevărat. Pe măsură ce dispunem de tot mai multe date de antrenament, dacă datele de antrenament sunt inconsistente, eroarea la antrenare va crește, fiindcă va fi din ce în ce mai greu ca modelul învățat să se adapteze la „zgomote“ din date. Similar, eroarea la testare va descrește, fiindcă producem un clasificator care este din ce în ce mai puțin afectat de “overfitting” pe datele de antrenament. Cele două erori vor converge la aşa-numita *eroare adevărată* (engl., true error) fiindcă diferențele statistice dintre datele de antrenament și datele de testare vor dispărea.

4.1.2 Algoritmul AdaBoost

22.

(Algoritmul AdaBoost: formulare; demonstarea unor relații de bază)

prelucrare de Liviu Ciortuz, după

■ • ○ CMU, 2015 fall, Z. Bar-Joseph, E. Xing, HW4, pr. 2.1
CMU, 2009 fall, Carlos Guestrin, HW2, pr. 3.1

Fie o mulțime de m exemple de antrenament, $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$, cu $x_i \in \mathcal{X}$ (unde \mathcal{X} poate fi de exemplu \mathbb{R}^d , cu $d \in \mathbb{N}^*$) și $y_i \in \{-1, +1\}$ pentru $i = 1, \dots, m$.

Presupunem că dispunem de un algoritm A care este un *clasificator* automat „slab“ (engl., weak classifier). A primește ca *input* o distribuție probabilistă oarecare D definită peste mulțimea S de exemple de antrenament și produce ca *output* o ipoteză $h : \mathcal{X} \rightarrow \{-1, +1\}$. Faptul că algoritmul A este un clasificator „slab“ înseamnă că orice astfel de ipoteză h produsă de el este doar [cu puțin] mai bună decât ghicirea / alegerea aleatorie (engl., random guessing).

Algoritmul AdaBoost⁴⁹⁸ este un algoritm iterativ care livrează ca *output* [un clasificator bazat pe] o *combinație liniară de ipoteze* care sunt produse de către clasificatorul „slab“ A , câte una la fiecare iterație. Concret, AdaBoost lucrează astfel:

- Inițial (adică pentru $t = 1$), se folosește distribuția uniformă $D_1(i) = \frac{1}{m}$, $i = 1, \dots, m$.
- La fiecare iterație $t = 1, \dots, T$,
 - folosind distribuția probabilistă D_t , rulează algoritmul „slab“ de clasificare A pe setul de exemple de antrenament S , obținând ipoteza h_t ;
 - calculează eroarea ponderată la antrenare produsă de ipoteza h_t ,

$$\varepsilon_t \stackrel{\text{not.}}{=} \text{err}_{D_t}(h_t) \stackrel{\text{def.}}{=} \Pr_{D_t}(\{x | y \neq h_t(x)\}), \quad (225)$$

iar după aceea,⁴⁹⁹ ponderea / „votul“

$$\alpha_t \stackrel{\text{not.}}{=} \frac{1}{2} \ln \frac{1 - \varepsilon_t}{\varepsilon_t} \quad (226)$$

asociat ipotezei h_t ;⁵⁰⁰

- definește o nouă distribuție probabilistă D_{t+1} , folosind „regula de actualizare“ (engl., update rule)

$$D_{t+1}(i) = \frac{1}{Z_t} D_t(i) e^{-\alpha_t y_i h_t(x_i)}, \text{ pentru } i = 1, \dots, m \text{ și } \forall t \geq 1, \quad (227)$$

unde Z_t este factorul de normalizare.

⁴⁹⁸Yoav Freund, Robert Schapire, “Experiments with a new boosting algorithm”, 1996, “A Decision Theoretic Generalization of On-Line Learning and an Application to Boosting”, 1997 și “A short introduction to boosting”, 1999.

⁴⁹⁹Dacă la o iterație $t \leq T$ clasificatorul slab A nu poate produce nicio ipoteză mai bună decât alegerea aleatorie (altfel spus, $\varepsilon_t = 1/2$) sau dacă, dimpotrivă, A nu poate produce decât o ipoteză perfectă (adică $\varepsilon_t = 0$), atunci algoritmul AdaBoost trebuie oprit.

⁵⁰⁰Am notat cu $\Pr_{D_t}(E)$ probabilitatea lui E (eveniment aleatoriu oarecare) în raport cu distribuția probabilistă D_t . Remarcăți faptul că din $\varepsilon_t \in (0, 1/2)$ rezultă $(1 - \varepsilon_t)/\varepsilon_t > 1$, deci $\alpha_t > 0$.

- În final, algoritmul AdaBoost va livra — ca ipoteză învățată — funcția $H_T \stackrel{\text{def}}{=} \text{sign} \left(\sum_{t=1}^T \alpha_t h_t \right)$; ea va acționa asupra instanțelor de test x conform [principiului] *votului ponderat majoritar* (engl., weighted majority vote).

Observație importantă: Această formulare a algoritmului AdaBoost nu impune nicio restricție asupra ipotezei h_t furnizate de către clasificatorul slab A la iterarea t , cu excepția condiției $\varepsilon_t < 1/2$. Totuși, într-o formulare ulterioară a algoritmului AdaBoost, ca în problema 28, dar și într-un cadru mai extins, ca în problema 29, se poate cere / recomanda în mod explicit⁵⁰¹ ca ipoteza h_t să fie aleasă *minimizând* (eventual aproximativ) criteriul erorii ponderate la antrenare pe o întreagă clasă de ipoteze (separatori decizionali). Astfel de ipoteze pot fi, de exemplu, arborii de decizie de adâncime 1 (compași de decizie; engl., decision stumps). În mod implicit la toate exercițiile unde vom exemplifica aplicarea algoritmului AdaBoost, *recomandarea* aceasta va fi aplicată.

Demonstrați următoarele relații:

- $Z_t = e^{-\alpha_t} \cdot (1 - \varepsilon_t) + e^{\alpha_t} \cdot \varepsilon_t$ (**consecință din relația (227)**)
- $Z_t = 2\sqrt{\varepsilon_t(1 - \varepsilon_t)}$ (**consecință din i și din relația (226)**)
- $0 < Z_t < 1$ (**consecință din ii.**)

$$iv. D_{t+1}(i) = \begin{cases} \frac{D_t(i)}{2\varepsilon_t}, & i \in M \stackrel{\text{not.}}{=} \{i | y_i \neq h_t(x_i)\} \\ \frac{D_t(i)}{2(1 - \varepsilon_t)}, & i \in C \stackrel{\text{not.}}{=} \{i | y_i = h_t(x_i)\} \end{cases}$$

(**consecință din relația (227) și ii.**)

- $\varepsilon_i > \varepsilon_j \Leftrightarrow \alpha_i < \alpha_j$ (**consecință din relația (226)**)
- $\text{err}_{D_{t+1}}(h_t) = 1/2$, unde $\text{err}_{D_{t+1}}(h_t) \stackrel{\text{not.}}{=} \Pr_{D_{t+1}}(\{x_i | h_t(x_i) \neq y_i\})$ (**consecință din relația (227) și ii.**)

Observații:

1. Vom explica acum de ce anume se folosește regula de actualizare a distribuțiilor D_t sub forma dată (227). Dubla inegalitate $0 < \varepsilon_t < 1/2$ implică $\alpha_t > 0$ și, prin urmare, rezultă că $e^{\alpha_t} > 1$, iar $e^{-\alpha_t} < 1$. Pentru instanțe corect clasificate de către ipoteza h_t , vom avea $y_i h_t(x_i) = 1$, iar pentru instanțe incorect clasificate $y_i h_t(x_i) = -1$. Rezultă că la calculul distribuției D_{t+1} , algoritmul AdaBoost mărește probabilitatea alocată instanțelor incorect clasificate și diminuează probabilitatea alocată instanțelor corect clasificate.
2. Faptul că $\text{err}_{D_{t+1}}(h_t) = 1/2$ (egalitate care trebuie demonstrată la punctul vi.) înseamnă că algoritmul ia exact „masa“ de probabilitate $\gamma_t \stackrel{\text{not.}}{=} \frac{1}{2} - \varepsilon_t$ (care reprezintă cu cât este mai bună ipoteza h_t decât alegerea aleatorie) de la multimea de instanțe care au fost corect clasificate de către ipoteza h_t și o distribuie în mod proporțional la multimea de instanțe incorect clasificate de către același h_t , pentru că algoritmul slab A să se centreze ulterior asupra „învățării“ unui model (sau a unei ipoteze) care să clasifice corect și aceste instanțe. Termenul de *boosting adaptiv* — din care a derivat numele algoritmului AdaBoost — corespunde exact acestei strategii, care este pusă de lucru prin *creșterea* probabilităților asociate instanțelor incorect clasificate.
3. Prin ipoteză, clasificatorul slab A produce pentru distribuția D_{t+1} o ipoteză (h_{t+1}) având eroarea ε_{t+1} strict mai mică decât $1/2$. Prin urmare, rezultatul $\text{err}_{D_{t+1}}(h_t) = 1/2$

⁵⁰¹Vedeți de exemplu MIT, 2006 fall, Tommi Jaakkola, HW4, problema 3.

(demonstrat la punctul *vi.*) va avea drept *consecință* faptul că în mod cert ipoteza h_t nu va fi [produsă și, deci, nici] aleasă la iterația $t + 1$.

Răspuns:

i. Întrucât Z_t este factorul de normalizare în scrierea probabilităților $D_{t+1}(i)$, conform relației (227) putem scrie:⁵⁰²

$$\begin{aligned} Z_t &= \sum_{i=1}^m D_t(i) \cdot e^{-\alpha_t y_i h_t(x_i)} = \sum_{i \in C} D_t(i) \cdot e^{-\alpha_t y_i h_t(x_i)} + \sum_{i \in M} D_t(i) \cdot e^{-\alpha_t y_i h_t(x_i)} \\ &= \sum_{i \in C} D_t(i) \cdot e^{-\alpha_t} + \sum_{i \in M} D_t(i) \cdot e^{\alpha_t} = e^{-\alpha_t} \cdot \underbrace{\sum_{i \in C} D_t(i)}_{1-\varepsilon_t} + e^{\alpha_t} \cdot \underbrace{\sum_{i \in M} D_t(i)}_{\varepsilon_t} \\ &= e^{-\alpha_t} \cdot (1 - \varepsilon_t) + e^{\alpha_t} \cdot \varepsilon_t. \end{aligned} \quad (228)$$

Observație: La problema 23.c vom arăta că valoarea $\alpha_t = \frac{1}{2} \ln \frac{1 - \varepsilon_t}{\varepsilon_t}$ este punct de minim pentru expresia $e^{-\alpha_t} \cdot (1 - \varepsilon_t) + e^{\alpha_t} \cdot \varepsilon_t$, dacă aceasta este văzută ca funcție de argumentul α_t .

ii. Tinând cont de relația de definiție dată pentru ponderea α_t în enunț, vom avea:

$$e^{\alpha_t} = e^{\frac{1}{2} \ln \frac{1 - \varepsilon_t}{\varepsilon_t}} = e^{\ln \sqrt{\frac{1 - \varepsilon_t}{\varepsilon_t}}} = \sqrt{\frac{1 - \varepsilon_t}{\varepsilon_t}}, \quad (229)$$

ceea ce implica

$$e^{-\alpha_t} = \frac{1}{e^{\alpha_t}} = \sqrt{\frac{\varepsilon_t}{1 - \varepsilon_t}}. \quad (230)$$

În consecință, folosind relația (228), factorul de normalizare Z_t poate fi exprimat în funcție de ε_t (eroarea ponderată comisă la antrenare de către ipoteza h_t în raport cu distribuția D_t) astfel:

$$Z_t = \sqrt{\frac{\varepsilon_t}{1 - \varepsilon_t}} \cdot (1 - \varepsilon_t) + \sqrt{\frac{1 - \varepsilon_t}{\varepsilon_t}} \cdot \varepsilon_t = 2\sqrt{\varepsilon_t(1 - \varepsilon_t)}. \quad (231)$$

iii. Pe baza cunoștințelor de analiză matematică de liceu, deducem imediat că maximul funcției de gradul al doilea $\varepsilon_t(1 - \varepsilon_t)$ se atinge pentru $\varepsilon_t = 1/2$, iar valoarea acestui maxim este $1/4$. Tinând cont că $\varepsilon_t \in (0, 1/2)$, din relația (231) rezultă că $Z_t > 0$ și $Z_t < 2\sqrt{\frac{1}{4}} = 1$.

iv. Pe baza relației (227), putem scrie imediat:

$$D_{t+1}(i) = \frac{1}{Z_t} \cdot D_t(i) \cdot \begin{cases} e^{\alpha_t}, & \text{pentru } i \in M \\ e^{-\alpha_t}, & \text{pentru } i \in C. \end{cases}$$

⁵⁰²Vom uza de faptul (specificat în enunț la proprietatea *iv.*) că s-a notat cu C mulțimea indicilor celor exemple care sunt corect clasificate la iterația t (adică, $C = \{i : y_i h_t(x_i) \geq 0\}$) și cu M mulțimea indicilor celor exemple care sunt incorect clasificate la aceeași iterație t (adică, $M = \{i : y_i h_t(x_i) < 0\}$).

Prin urmare,

$$\begin{aligned} i \in M \Rightarrow D_{t+1}(i) &= \frac{1}{Z_t} \cdot D_t(i) \cdot e^{\alpha_t} \stackrel{(231),(229)}{=} \frac{1}{2\sqrt{\varepsilon_t(1-\varepsilon_t)}} \cdot D_t(i) \cdot \frac{\sqrt{1-\varepsilon_t}}{\sqrt{\varepsilon_t}} = \frac{D_t(i)}{2\varepsilon_t} \\ i \in C \Rightarrow D_{t+1}(i) &= \frac{1}{Z_t} \cdot D_t(i) \cdot e^{-\alpha_t} \stackrel{(231),(230)}{=} \frac{1}{2\sqrt{\varepsilon_t(1-\varepsilon_t)}} \cdot D_t(i) \cdot \frac{\sqrt{\varepsilon_t}}{\sqrt{1-\varepsilon_t}} = \frac{D_t(i)}{2(1-\varepsilon_t)}. \end{aligned}$$

v. Plecând de la definiția $\alpha_t = \ln \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}}$ care a fost dată în pseudo-codul algoritmului AdaBoost, putem scrie mai întâi

$$\alpha_i < \alpha_j \Leftrightarrow \ln \sqrt{\frac{1-\varepsilon_i}{\varepsilon_i}} < \ln \sqrt{\frac{1-\varepsilon_j}{\varepsilon_j}}.$$

Apoi, întrucât atât funcția \ln cât și funcția rădăcină pătrată sunt strict crescătoare, rezultă că

$$\begin{aligned} \alpha_i < \alpha_j &\Leftrightarrow \frac{1-\varepsilon_i}{\varepsilon_i} < \frac{1-\varepsilon_j}{\varepsilon_j} \stackrel{\varepsilon_i, \varepsilon_j > 0}{\Leftrightarrow} \varepsilon_j(1-\varepsilon_i) < \varepsilon_i(1-\varepsilon_j) \\ &\Leftrightarrow \varepsilon_j - \cancel{\varepsilon_i \varepsilon_j} < \varepsilon_i - \cancel{\varepsilon_i \varepsilon_j} \Leftrightarrow \varepsilon_i > \varepsilon_j \end{aligned}$$

Așadar, are loc echivalența $\alpha_i < \alpha_j \Leftrightarrow \varepsilon_i > \varepsilon_j$.

vi. Putem exprima eroarea ponderată produsă la antrenare de către ipoteza h_t , în raport cu distribuția probabilistă D_{t+1} , în felul următor:

$$\begin{aligned} err_{D_{t+1}}(h_t) &\stackrel{\text{def.}}{=} \Pr_{D_{t+1}}(\{x_i | h_t(x_i) \neq y_i\}) = \sum_{i=1}^m D_{t+1}(i) \cdot 1_{\{y_i \neq h_t(x_i)\}} \\ &\stackrel{(227)}{=} \sum_{i \in M} \frac{1}{Z_t} \cdot D_t(i) \cdot e^{\alpha_t} = \frac{1}{Z_t} \cdot e^{\alpha_t} \cdot \underbrace{\sum_{i \in M} D_t(i)}_{\varepsilon_t} \\ &= \frac{1}{Z_t} \cdot e^{\alpha_t} \cdot \varepsilon_t. \end{aligned} \tag{232}$$

Înlocuind în formula (232) valoarea lui Z_t dată de expresia (231), obținem:

$$err_{D_{t+1}}(h_t) = \frac{1}{Z_t} \cdot e^{\alpha_t} \cdot \varepsilon_t = \frac{1}{2\sqrt{\varepsilon_t(1-\varepsilon_t)}} \cdot \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}} \cdot \varepsilon_t = \frac{1}{2}.$$

23.

(Algoritmul AdaBoost:
analiza teoretică a convergenței erorii la antrenare;
învățabilitate empirică γ -slabă)

prelucrare de Liviu Ciortuz, după

■ • ○ CMU, 2015 fall, Z. Bar-Joseph, E. Xing, HW4, pr. 2.2-5

CMU, 2009 fall, Carlos Guestrin, HW2, pr. 3.1

CMU, 2005 spring, T. Mitchell, C. Guestrin, HW2, pr. 1.1.3

Acest exercițiu vă va ghida pas cu pas ca să demonstrați că $err_S(H_T)$, eroarea la antrenare a algoritmului AdaBoost, prezentat la problema 22 — văzută ca

numărul de instanțe greșit clasificate din totalul de m instanțe —, descrește foarte repede (adică, cu o rată foarte mare) în raport cu numărul de iterații efectuate (T), iar în anumite condiții converge la 0.

Observație: În acest exercițiu se lucrează fără a cere ca la fiecare iterație a algoritmului AdaBoost să se aleagă cea mai bună ipoteză „slabă”.⁵⁰³

a. Notăm cu f_T combinația liniară de ipoteze $\sum_{t=1}^T \alpha_t h_t$.⁵⁰⁴ Arătați că din relația (227) rezultă $D_{T+1}(i) = \frac{1}{m \cdot \prod_{t=1}^T Z_t} e^{-y_i f_T(x_i)}$. (Această expresie reprezintă de fapt o formulă de calcul nerecursivă pentru valorile distribuțiilor probabiliste D_t .)

b. Folosind rezultatul de la punctul a, arătați că $\text{err}_S(H_T) \leq \prod_{t=1}^T Z_t$, unde $\text{err}_S(H_T)$ desemnează eroarea produsă la antrenare de către AdaBoost.⁵⁰⁵

$$\frac{1}{m} \sum_{i=1}^m 1_{\{H_T(x_i) \neq y_i\}}.$$

Notăția $1_{\{H_T(x_i) \neq y_i\}}$ desemnează, ca de obicei, *funcția-indicator* pentru testul $H_T(x_i) \neq y_i$; astădat, $1_{\{H_T(x_i) \neq y_i\}} = 1$ pentru acei i pentru care $H_T(x_i) \neq y_i$ și 0 în caz contrar.⁵⁰⁶

Sugestie (1): Puteți folosi inegalitatea: $1_{\{x < 0\}} \leq e^{-x}$.⁵⁰⁷

c. La acest punct vom vedea că algoritmul AdaBoost — în loc să optimizeze în mod direct eroarea produsă la antrenare, $\text{err}_S(H_T)$ —, se mulțumește să minimizeze în mod *greedy* (deci, secvențial) produsul $\prod_{t=1}^T Z_t$, care reprezintă o *margine superioară* (engl., upper bound) pentru *eroarea la antrenare*, după cum am arătat la punctul b.

Se observă în pseudo-codul algoritmului AdaBoost că factorii de normalizare Z_1, \dots, Z_{t-1} sunt determinați în cadrul primelor $t-1$ iterații și nu pot fi modificați la iterația t . Astădat, a minimiza $\prod_{t=1}^T Z_t$ la iterația t în mod *greedy* revine la a minimiza Z_t .

Dacă facem abstracție de valoarea atribuită ponderii α_t la iterația t (și anume, $\frac{1}{2} \ln \frac{1 - \varepsilon_t}{\varepsilon_t}$), atunci factorul de normalizare Z_t pentru distribuția D_{t+1} va putea fi scris ca o funcție de parametru α_t , pornind de la relația (228). Arătați că valoarea pentru care se atinge minimul acestei funcții (în raport cu toate valorile posibile pentru α_t) este exact valoarea $\frac{1}{2} \ln \frac{1 - \varepsilon_t}{\varepsilon_t}$, pe care am întâlnit-o în formularea algoritmului AdaBoost.

d. Arătați că $\prod_{t=1}^T Z_t \leq e^{-2 \sum_{t=1}^T \gamma_t^2}$, unde $\gamma_t \stackrel{\text{not.}}{=} \frac{1}{2} - \varepsilon_t$.⁵⁰⁸

⁵⁰³Vedeți *Observația importantă* de la problema 22.

⁵⁰⁴Observați că expresia funcției H_T din finalul pseudo-codului algoritmului AdaBoost — vedeți enunțul problemei 22 — se poate scrie ca $H_T(x) = \text{sign}(f_T(x))$.

⁵⁰⁵Vă reamintim că S este setul de date de antrenament.

⁵⁰⁶Observație: $H_T(x_i) \neq y_i \stackrel{\text{def.}}{\Leftrightarrow} \text{sign}(\sum_{t=1}^T \alpha_t h_t(x_i)) \neq y_i \Leftrightarrow \text{sign}(f_T(x_i)) \neq y_i \Leftrightarrow y_i \text{sign}(f_T(x_i)) = -1 \Leftrightarrow y_i f_T(x_i) < 0$. Astădat, funcțiile $1_{\{H_T(x_i) \neq y_i\}}$ și $1_{\{y_i f_T(x_i) < 0\}}$ sunt identice.

⁵⁰⁷Putem interpreta această inegalitate sub forma următoare: funcția de cost / „pierdere“ [negativ] exponențială reprezintă o margine superioară pentru funcția de cost / pierdere 0 – 1 (engl., the 0 – 1 loss function). Pentru o definiție [de lucru] pentru funcțiile de cost / pierdere, vedeți *Explicația* de la problema 29.

⁵⁰⁸LC: Putem numi variabilă γ_t „ecart“, fiindcă ea reprezintă diferența dintre 1/2, eroarea alegerii aleatorii, și ε_t , eroarea ponderată a ipotezei h_t la antrenare.

Sugestie (2): De data aceasta, puteți folosi o altă margine inferioară pentru funcția de pierdere [negativ] exponențială: $1 - x \leq e^{-x}$.

e. Combinând rezultatele de la punctele c și d , observăm că eroarea la antrenare produsă de algoritmul AdaBoost se micșorează (cu o rată exponențială) pe măsură ce T crește. Întrucât rata aceasta este cuprinsă în intervalul $(0, 1)$, convergența ei este asigurată.⁵⁰⁹ La acest punct vom pune în evidență o condiție suficientă pentru ca eroarea la antrenare produsă de AdaBoost să conveargă la 0.

Presupunem că există $\gamma > 0$ astfel încât $\gamma \leq \gamma_t$ pentru orice $t = 1, 2, \dots$.⁵¹⁰ (Această proprietate se numește „învățabilitate empirică”⁵¹¹ γ -slabă“ (engl., empirical γ -weak learnability), iar γ se numește *garanție de învățabilitate empirică slabă*.) Considerând $\varepsilon > 0$ fixat, determinați (în funcție de γ și ε) o margine superioară pentru cât de multe iterații sunt necesare pentru ca eroarea la antrenare produsă de algoritmul AdaBoost să fie mai mică decât ε , adică $err_S(H_T) < \varepsilon$. Vă cerem să exprimați răspunsul sub forma $T = \mathcal{O}(\cdot)$.

Răspuns:

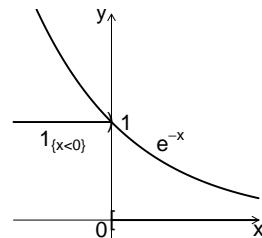
a. Pentru a demonstra egalitatea din enunț, vom porni de la relația (227), exprimând probabilitatea $D_{T+1}(i)$ recursiv în funcție de $D_T(i)$, apoi de $D_{T-1}(i)$, până ajungem la $D_1(i)$:

$$\begin{aligned} D_{T+1}(i) &\stackrel{\text{def.}}{=} \frac{1}{Z_T} D_T(i) e^{-\alpha_T y_i h_T(x_i)} = D_T(i) \frac{1}{Z_T} e^{-\alpha_T y_i h_T(x_i)} \\ &\stackrel{\text{def.}}{=} D_{T-1}(i) \frac{1}{Z_{T-1}} e^{-\alpha_{T-1} y_i h_{T-1}(x_i)} \frac{1}{Z_T} e^{-\alpha_T y_i h_T(x_i)} \\ &= D_{T-1}(i) \frac{1}{Z_{T-1} Z_T} e^{-y_i (\alpha_{T-1} h_{T-1}(x_i) + \alpha_T h_T(x_i))} \\ &\quad \vdots \\ &= D_j(i) \frac{1}{Z_j \dots Z_{T-1} Z_T} e^{-y_i (\alpha_j h_j(x_i) + \dots + \alpha_{T-1} h_{T-1}(x_i) + \alpha_T h_T(x_i))} \\ &\quad \vdots \\ &= D_1(i) \frac{1}{\prod_{t=1}^T Z_t} e^{-\sum_{t=1}^T y_i \alpha_t h_t(x_i)} = \frac{1}{m \cdot \prod_{t=1}^T Z_t} e^{-y_i f_T(x_i)}. \end{aligned}$$

Produsul de forma $y_i f_T(x_i)$ (mai general, $y_i f_t(x_i)$) se numește *magine algebraică* a instanței x_i .

b. După cum s-a precizat în enunț, putem exprima eroarea produsă de ipoteza H_T (outputul algoritmului AdaBoost) pe setul de date de antrenament S cu ajutorul funcției de cost / pierdere $0 - 1$:

$$err_S(H_T) = \frac{1}{m} \sum_{i=1}^m 1_{\{y_i f_T(x_i) < 0\}}.$$



⁵⁰⁹Se știe că orice sir mărginit și monoton este convergent.

⁵¹⁰LC: De fapt, este suficient ca această condiție să fie îndeplinită de la o iterație oarecare t_0 încolo.

⁵¹¹Adjectivul *empiric* se referă la faptul că eroarea analizată este *eroarea la antrenare*.

Folosind inegalitatea menționată în *Sugestia (1)* din enunț, putem scrie:

$$\text{err}_S(H_T) \leq \frac{1}{m} \sum_{i=1}^m e^{-y_i f_T(x_i)}.$$

Conform rezultatului de la punctul *a*, vom putea substitui $e^{-y_i f_T(x_i)}$ cu produsul $D_{T+1}(i) \cdot m \cdot \prod_{t=1}^T Z_t$ și vom obține:

$$\begin{aligned} \text{err}_S(H_T) &\leq \frac{1}{m} \sum_{i=1}^m \left(D_{T+1}(i) \cdot m \cdot \prod_{t=1}^T Z_t \right) = \sum_{i=1}^m \left(D_{T+1}(i) \prod_{t=1}^T Z_t \right) \\ &= \left(\prod_{t=1}^T Z_t \right) \cdot \underbrace{\left(\sum_{i=1}^m D_{T+1}(i) \right)}_1 = \prod_{t=1}^T Z_t. \end{aligned}$$

Egalitatea $\sum_{i=1}^m D_{T+1}(i) = 1$ se justifică prin faptul că D_{T+1} reprezintă o distribuție probabilistă.

c. Vom porni de la relația (228), care a fost demonstrată la problema 22.i, și anume $Z_t = \varepsilon_t \cdot e^{\alpha_t} + (1 - \varepsilon_t) \cdot e^{-\alpha_t}$. Pentru a facilita înțelegerea raționamentului care urmează, întrucât aici α_t este lăsat liber, îl vom renota cu α , deci expresia anterioară va fi văzută ca funcție de acest α , adică $Z(\alpha) = \varepsilon_t \cdot e^\alpha + (1 - \varepsilon_t) \cdot e^{-\alpha}$. (Observați că Z_t dinainte tocmai a fost și el renotat mai simplu: Z .)

Ca de obicei, pentru a găsi minimul expresiei $\varepsilon_t \cdot e^\alpha + (1 - \varepsilon_t) \cdot e^{-\alpha}$, în care ε_t (eroarea produsă de ipoteza care tocmai a fost produsă de clasificatorul „slab“ A) este considerată constantă, vom calcula derivata ei în raport cu α și apoi vom egala cu 0 această derivată:

$$\begin{aligned} \frac{\partial}{\partial \alpha} (\varepsilon_t \cdot e^\alpha + (1 - \varepsilon_t) \cdot e^{-\alpha}) = 0 &\Leftrightarrow \varepsilon_t \cdot e^\alpha - (1 - \varepsilon_t) \cdot e^{-\alpha} = 0 \Leftrightarrow \\ \varepsilon_t \cdot (e^\alpha)^2 = 1 - \varepsilon_t &\Leftrightarrow e^{2\alpha} = \frac{1 - \varepsilon_t}{\varepsilon_t} \Leftrightarrow 2\alpha = \ln \frac{1 - \varepsilon_t}{\varepsilon_t} \Leftrightarrow \alpha = \frac{1}{2} \ln \frac{1 - \varepsilon_t}{\varepsilon_t} = \ln \sqrt{\frac{1 - \varepsilon_t}{\varepsilon_t}}. \end{aligned}$$

Remarcați faptul că fracția $\frac{1 - \varepsilon_t}{\varepsilon_t}$ este pozitivă (deci i se poate aplica logaritmul), fiindcă $\varepsilon_t \in (0, 1/2)$. Chiar mai mult, $\alpha > 0$, fiindcă $\frac{1 - \varepsilon_t}{\varepsilon_t} > 1$, datorită aceluiasi motiv ca mai înainte. Se poate verifica imediat (analizând semnele derivatei de mai sus) că $\alpha = \frac{1}{2} \ln \frac{1 - \varepsilon_t}{\varepsilon_t}$ este într-adevăr punctul în care se atinge *minimul* expresiei $\varepsilon_t \cdot e^\alpha + (1 - \varepsilon_t) \cdot e^{-\alpha}$, deci și al lui Z_t (văzut ca funcție de α):

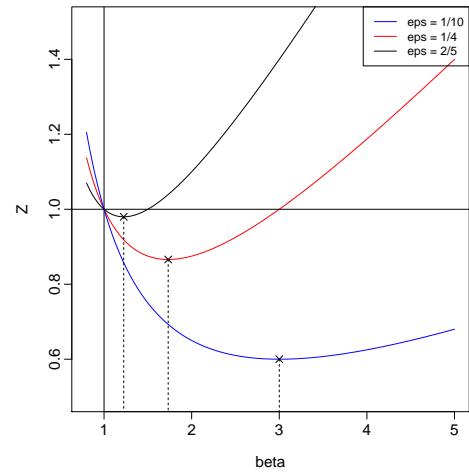
$$\varepsilon_t \cdot e^\alpha - (1 - \varepsilon_t) \cdot e^{-\alpha} > 0 \Leftrightarrow e^{2\alpha} - \frac{1 - \varepsilon_t}{\varepsilon_t} > 0 \Leftrightarrow \alpha > \frac{1}{2} \ln \frac{1 - \varepsilon_t}{\varepsilon_t},$$

ceea ce înseamnă că funcția $Z(\alpha) = \varepsilon_t \cdot e^\alpha + (1 - \varepsilon_t) \cdot e^{-\alpha}$ este strict crescătoare pe intervalul $\left(\frac{1}{2} \ln \frac{1 - \varepsilon_t}{\varepsilon_t}, +\infty \right)$ și strict descrescătoare pe intervalul $\left(-\infty, \frac{1}{2} \ln \frac{1 - \varepsilon_t}{\varepsilon_t} \right]$.

Pentru a ilustra grafic acest rezultat, vom renota e^α cu β , deci vom scrie Z ca o funcție de argumentul β :

$$Z(\beta) = \varepsilon_t \cdot \beta + (1 - \varepsilon_t) \cdot \frac{1}{\beta}.$$

Graficul funcției $Z(\beta)$ pentru trei valori diferite ale lui ε_t este prezentat în figura alăturată. Conform raționamentului de mai sus, minimul funcției $Z(\beta)$ este atins pentru valoarea $\beta_{min} = \sqrt{\frac{1 - \varepsilon_t}{\varepsilon_t}}$. Vechea expresie a lui Z_t o vom regăsi astfel: $Z_t \stackrel{not.}{=} Z(\beta_{min}) \stackrel{calcul}{=} 2\sqrt{\varepsilon_t(1 - \varepsilon_t)}$.⁵¹²

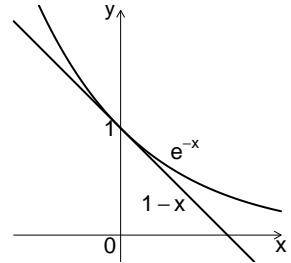


d. Folosind relația (231) care a fost dedusă la problema 22.ii și apoi relația dată în enunț pentru a exprima legătura dintre ε_t și „ecartul” γ_t , vom putea scrie:

$$\begin{aligned} \prod_{t=1}^T Z_t &= \prod_{t=1}^T 2 \cdot \sqrt{\varepsilon_t(1 - \varepsilon_t)} = \prod_{t=1}^T 2 \cdot \sqrt{\left(\frac{1}{2} - \gamma_t\right) \left(1 - \left(\frac{1}{2} - \gamma_t\right)\right)} \\ &= \prod_{t=1}^T 2 \cdot \sqrt{\left(\frac{1}{2} - \gamma_t\right) \left(\frac{1}{2} + \gamma_t\right)} = \prod_{t=1}^T 2 \cdot \sqrt{\left(\frac{1}{4} - \gamma_t^2\right)} = \prod_{t=1}^T \sqrt{1 - 4\gamma_t^2}. \end{aligned}$$

Utilizând inegalitatea din Sugestia (2) din enunț, rezultă $1 - 4\gamma_t^2 \leq e^{-4\gamma_t^2}$ și, mai departe:

$$\begin{aligned} \prod_{t=1}^T Z_t &\leq \prod_{t=1}^T \sqrt{e^{-4\gamma_t^2}} = \prod_{t=1}^T \sqrt{(e^{-2\gamma_t^2})^2} = \prod_{t=1}^T e^{-2\gamma_t^2} \\ &= e^{-2 \sum_{t=1}^T \gamma_t^2}. \end{aligned}$$



⁵¹²Iată alte câteva proprietăți ale funcției $Z(\beta)$, care pot fi demonstrează cu ușurință (și care sunt reflectate în graficul de mai sus):

- $\lim_{\beta \searrow 0} Z(\beta) = +\infty$.
- $\lim_{\beta \rightarrow 1} Z(\beta) = Z(1) = \varepsilon_t + 1 - \varepsilon_t = 1$.
- $\lim_{\beta \rightarrow +\infty} Z(\beta) = +\infty$.
- Dreapta de ecuație $y = \varepsilon_t \beta$ este asimptotă oblică la $+\infty$ pentru funcția $Z(\beta)$, fiindcă $\lim_{\beta \rightarrow +\infty} \frac{Z(\beta)}{\beta} = \varepsilon_t$, iar $\lim_{\beta \rightarrow +\infty} (Z(\beta) - \varepsilon_t \beta) = 0$.
- Atunci când $\varepsilon_t \nearrow \frac{1}{2}$, rezultă că $\sqrt{\frac{1 - \varepsilon_t}{\varepsilon_t}} \searrow 1$, deci $\ln \sqrt{\frac{1 - \varepsilon_t}{\varepsilon_t}} \searrow 0$, adică $\alpha_t \searrow 0$ (și, deci, $\beta_{min} \searrow 1$). Totodată, $Z_t \stackrel{not.}{=} Z(\beta_{min}) = 2\sqrt{\varepsilon_t(1 - \varepsilon_t)} \nearrow 1$.
- Atunci când $\varepsilon_t \searrow 0$, rezultă pe de o parte că $\sqrt{\frac{1 - \varepsilon_t}{\varepsilon_t}} \nearrow +\infty$, deci $\alpha_t \rightarrow +\infty$ (și $\beta_{min} \nearrow +\infty$), iar pe de altă parte că $Z_t \stackrel{not.}{=} Z(\beta_{min}) = 2\sqrt{\varepsilon_t(1 - \varepsilon_t)} \searrow 0$.

e. Combinând rezultatele pe care le-am obținut la punctele c și e , rezultă $\text{err}_S(H_T) \leq e^{-2\sum_{t=1}^T \gamma_t^2}$. Apoi, ținând cont că $\gamma \leq \gamma_t$ pentru $t = 1, 2, \dots$, vom avea:⁵¹³

$$\text{err}_S(H_T) \leq e^{-2T\gamma^2}. \quad (233)$$

Prin urmare, pentru ca inegalitatea $\text{err}_S(H_T) < \varepsilon$ să aibă loc, este suficient să impunem restricția următoare:

$$-2T\gamma^2 < \ln \varepsilon \Leftrightarrow 2T\gamma^2 > -\ln \varepsilon \Leftrightarrow 2T\gamma^2 > \ln \frac{1}{\varepsilon} \Leftrightarrow T > \frac{1}{2\gamma^2} \ln \frac{1}{\varepsilon}.$$

Așadar, $T = \mathcal{O}\left(\frac{1}{\gamma^2} \ln \frac{1}{\varepsilon}\right)$.

În particular, atunci când considerăm $\varepsilon = \frac{1}{m}$, obținem numărul T de iterații începând de la care [cu certitudine] vom avea $\text{err}_S(H_{T'}) = 0$ pentru orice $T' \geq T$: $T = \left\lceil \frac{1}{2\gamma^2} \ln m \right\rceil$.⁵¹⁴

Observație: Din relația (233) rezultă că $\text{err}_S(H_T) \rightarrow 0$ atunci când $T \rightarrow \infty$.

24.

(Algoritmul AdaBoost, folosind “decision stumps”: aplicare pe date din \mathbb{R}^2)

■ • CMU, 2015 fall, Z. Bar-Joseph, E. Xing, HW4, pr. 2.6

Considerăm setul de date de antrenament din tabelul de mai jos.

Rulați $T = 3$ iterații ale algoritmului AdaBoost, folosind drept clasificatori slabii *compași de decizie* (engl., decision stumps), care determină separatoare liniare paralele cu axe de coordonate (engl., axis-aligned separators). Reprezentați aceste date în planul euclidian; pe figura obținută veți trasa dreptele corespunzătoare ipotezelor „slabe“ h_t ,⁵¹⁵ iar la final veți completa tabelul de mai jos. Pentru pseudo-codul algoritmului AdaBoost, vedeti problema 22. Vă rugăm să citiți și să rețineți *Observația importantă* pe care am redactat-o imediat după acel

x_i	X_1	X_2	y_i
x_1	1	2	+1
x_2	2	3	+1
x_3	3	4	-1
x_4	3	2	-1
x_5	3	1	-1
x_6	4	4	-1
x_7	5	4	-1
x_8	5	2	+1
x_9	5	1	+1

⁵¹³Întrucât $\gamma_t \geq \gamma > 0$, au loc următoarele echivalențe: $\gamma_t \geq \gamma \Leftrightarrow \gamma_t^2 \geq \gamma^2 \Leftrightarrow -2\gamma_t^2 \leq -2\gamma^2 \Leftrightarrow e^{-2\gamma_t^2} \leq e^{-2\gamma^2}$. Considerând această ultimă inegalitate pentru fiecare valoare a lui $t \in \{1, 2, \dots, T\}$ și înmulțind membru cu membru, rezultă: $\prod_{t=1}^T e^{-2\gamma_t^2} \leq \prod_{t=1}^T e^{-2\gamma^2} \Leftrightarrow e^{-2\sum_{t=1}^T \gamma_t^2} \leq (e^{-2\gamma^2})^T \Leftrightarrow e^{-2\sum_{t=1}^T \gamma_t^2} \leq e^{-2T\gamma^2}$.

⁵¹⁴Observație: Dacă toate instanțele de antrenament x_i sunt corect clasificate (de către ipoteza combinată $H_S(t)$) la iterația t , asta nu înseamnă că toate instanțele vor fi corect clasificate și la iterația următoare (de către $H_S(t+1)$), fiindcă eroarea la antrenare $\text{err}_S(H_t)$ nu este în mod neapărat descrescătoare (în raport cu t). (Vedeți de exemplu problema de tip implementare de la Stanford, 2016 fall, A. Ng, J. Duchi, HW2, pr. 6.d.) Însă, dacă pragul inferior γ pentru γ_t se menține pentru orice $t = t_0, t_0 + 1, \dots, t_0 + T$, unde $t_0 \in \mathbb{N}$, iar T se calculează ca mai sus, atunci avem certitudinea că eroarea la antrenare $\text{err}_S(H_t)$ va rămâne 0 pentru orice $t \geq t_0 + T$.

⁵¹⁵Însoțiti fiecare dintre dreptele respective cu o pereche de semne + și -, corespunzătoare zonelor de decizie determinate.

t	ε_t	α_t	$D_t(1)$	$D_t(2)$	$D_t(3)$	$D_t(4)$	$D_t(5)$	$D_t(6)$	$D_t(7)$	$D_t(8)$	$D_t(9)$	$errs(H_T)$
1												
2												
3												

Recomandare: Rolul acestui exercițiu este de a vă ajuta să înțelegeți pas cu pas cum anume lucrează în practică algoritmul AdaBoost. Vă sugerăm ca, după ce veți fi înțeleși rezolvarea acestui exercițiu, să implementați mai întâi un program / o funcție, care să calculeze *eroarea ponderată la antrenare* (engl., weighted training error) produsă de un anumit compas de decizie, în raport cu o distribuție de probabilitate (D) definită pe setul de date de antrenament. Ulterior, puteți extinde acest program la o implementare completă a algoritmului AdaBoost, conform pseudo-codului din problema 22.

Răspuns:

Făță de reprezentarea grafică (de tip arbore de adâncime 1) cu care ne-am obișnuit în trecut pentru *compașii de decizie*, aici vom lucra cu *reprezentarea analitică* următoare: pentru un atribut de tip continuu X care ia valori $x \in \mathbb{R}$ și pentru un prag de separare (engl., split threshold) oarecare $s \in \mathbb{R}$ fixat, putem defini doi compași de decizie ca funcții (mai exact, ca *funcții-treaptă*) de variabila x :

$$\text{sign}(x - s) = \begin{cases} 1 & \text{dacă } x \geq s \\ -1 & \text{dacă } x < s \end{cases} \quad \text{și} \quad \text{sign}(s - x) = \begin{cases} -1 & \text{dacă } x \geq s \\ 1 & \text{dacă } x < s. \end{cases}$$

Pentru conveniență, în cele ce urmează — atunci când nu riscăm să creăm ambiguități — vom renota primul compas de decizie cu $X \geq s$ și al doilea cu $X < s$.

Faptul că în acest exercițiu se cere aplicarea algoritmului AdaBoost cu compași de decizie înseamnă că la fiecare iterare (t) clasificatorul „slab“ (notat cu A) din pseudo-codul algoritmului AdaBoost selectează un compas de decizie, și anume — conform *Observației importante* din enunțul problemei 22 — unul care are eroarea ponderată la antrenare minimă în raport cu distribuția probabilistă curentă (D_t).⁵¹⁶

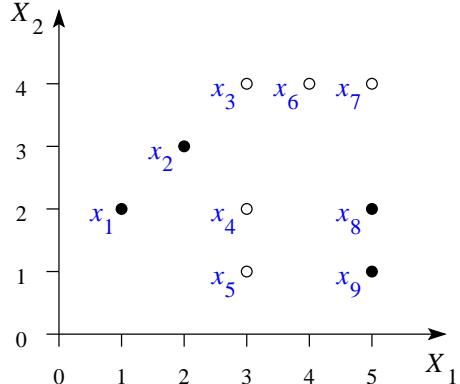
Exact ca atunci când am lucrat cu algoritmul ID3 cu atribute continue, vom considera câte un prag intermediar pentru fiecare pereche de valori succesive [luate de un atribut continuu X] corespunzătoare schimbărilor de etichete. Formal, pentru orice astfel de perechi de valori x_i, x_{i+1} , cu $y_i y_{i+1} < 0$ și $x_i < x_{i+1}$, dar fără să mai existe vreun $x_j \in \text{Val}(X)$ cu proprietatea $x_i < x_j < x_{i+1}$, vom considera pragul $(x_i + x_{i+1})/2$.⁵¹⁷ De asemenea, vom considera și un *prag exterior* intervalului de valori ale atributului X în multimea de instanțe de antrenament.⁵¹⁸

⁵¹⁶Deci algoritmul slab A nu este algoritmul ID3 care produce un compas de decizie cu câștig de informație minim!

⁵¹⁷În cazul algoritmului ID3, există un *rezultat teoretic* (vedeți problema 47) care demonstrează că nu este necesar să considerăm alte praguri pentru un atribut continuu X decât cele situate între perechi de valori succesive având etichete [de semne] contrare, fiindcă valoarea IG-ului celorlalte praguri este situată în mod cert sub valoarea IG-ului maximal pentru acel atribut. LC: În cazul algoritmului AdaBoost, se poate demonstra un rezultat similar care, în consecință, ne va permite să simplificăm aplicarea clasificatorului „slab“ (A).

⁵¹⁸Compașii de decizie corespunzători acestui prag „exterior“ pot fi puși în corespondență cu arborii de decizie de adâncime 0 pe care i-am întâlnit în precedent.

Așadar, la *prima iterație* a algoritmului AdaBoost (adică, pentru $t = 1$), *pragurile* de separare pentru valorile celor două variabile continue (X_1 și X_2) care corespund celor două coordonate ale instanțelor de antrenament (x_1, \dots, x_9) sunt $\frac{1}{2}$, $\frac{5}{2}$ și $\frac{9}{2}$ pentru X_1 , și respectiv $\frac{1}{2}$, $\frac{3}{2}$, $\frac{5}{2}$ și $\frac{7}{2}$ pentru X_2 .



Observăm însă că se poate renunța la pragul „exterior“ $\frac{1}{2}$ pentru X_2 , deoarece compașii de decizie corespunzător lui se comportă identic cu compașii de decizie corespunzător pragului „exterior“ $\frac{1}{2}$ pentru X_1 .⁵¹⁹

Pentru compașii de decizie corespunzători acestei prime iterării, *erorile ponderate la antrenare* sunt cele prezentate centralizat în tabelele de mai jos. Pentru calcule, am folosit egalitățile $err_{D_t}(X_1 \geq s) = 1 - err_{D_t}(X_1 < s)$ și, similar, $err_{D_t}(X_2 \geq s) = 1 - err_{D_t}(X_2 < s)$, pentru orice prag s și orice iterare $t = 1, 2, \dots$. Aceste egalități sunt foarte ușor de demonstrat.

s	$\frac{1}{2}$	$\frac{5}{2}$	$\frac{9}{2}$
$err_{D_1}(X_1 < s)$	$\frac{4}{9}$	$\frac{2}{9}$	$\frac{4}{9} + \frac{2}{9} = \frac{2}{3}$
$err_{D_1}(X_1 \geq s)$	$\frac{5}{9}$	$\frac{7}{9}$	$\frac{1}{3}$

s	$\frac{1}{2}$	$\frac{3}{2}$	$\frac{5}{2}$	$\frac{7}{2}$
$err_{D_1}(X_2 < s)$	$\frac{4}{9}$	$\frac{1}{9} + \frac{3}{9} = \frac{4}{9}$	$\frac{2}{9} + \frac{1}{9} = \frac{1}{3}$	$\frac{2}{9}$
$err_{D_1}(X_2 \geq s)$	$\frac{5}{9}$	$\frac{5}{9}$	$\frac{2}{3}$	$\frac{7}{9}$

Se observă că eroarea ponderată minimă la antrenare ($\varepsilon_1 = 2/9$) este obținută pentru compașii de decizie $X_1 < 5/2$ și $X_2 < 7/2$.⁵²⁰ Alegem drept *cea mai bună ipoteză* la această iterare pe $h_1 = \text{sign}\left(\frac{7}{2} - X_2\right)$, separatorul corespunzător fiind dreapta de ecuație $X_2 = \frac{7}{2}$. Ipoteza h_1 clasifică greșit instanțele x_4 și x_5 . Vom avea:

$$\begin{aligned} \gamma_1 &= \frac{1}{2} - \frac{2}{9} = \frac{5}{18} \\ \alpha_1 &= \frac{1}{2} \ln \frac{1 - \varepsilon_1}{\varepsilon_1} = \ln \sqrt{\left(1 - \frac{2}{9}\right) : \frac{2}{9}} = \ln \sqrt{\frac{7}{2}} \approx 0.626 \end{aligned}$$

⁵¹⁹De aceea în tabelele de mai jos, pentru aceste cazuri se folosește culoarea verde.

⁵²⁰În tabele, pentru erorile minime am folosit culoarea albastră.

Acum algoritmul trebuie să „pregătească“ o nouă distribuție (D_2), pentru iterația următoare. Distribuția D_2 va fi obținută prin modificarea vechii distribuții (D_1), în aşa fel încât algoritmul să se poată concentra cu preponderență asupra instanțelor care au fost greșit clasificate (engl., misclassified). Conform relației (227), vom scrie:

$$D_2(i) = \frac{1}{Z_1} D_1(i) (\underbrace{e^{-\alpha_1}}_{\sqrt{2/7}})^{y_i h_1(x_i)} = \begin{cases} \frac{1}{Z_1} \cdot \frac{1}{9} \cdot \sqrt{\frac{2}{7}} & \text{pentru } i \in \{1, 2, 3, 6, 7, 8, 9\}; \\ \frac{1}{Z_1} \cdot \frac{1}{9} \cdot \sqrt{\frac{7}{2}} & \text{pentru } i \in \{4, 5\}. \end{cases}$$

Vă reamintim că Z_1 este aşa-numitul *factor de normalizare* pentru distribuția probabilistă D_2 . Așadar, din relația $\sum_i D_2(i) = 1$ rezultă:

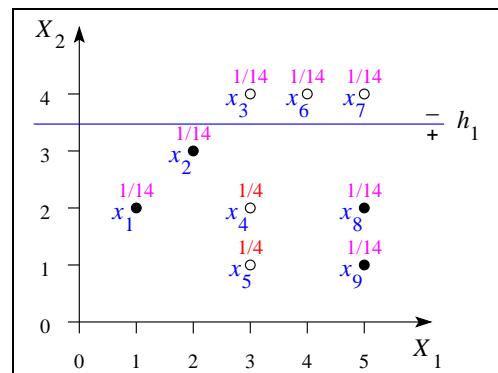
$$Z_1 = \frac{1}{9} \left(7 \cdot \sqrt{\frac{2}{7}} + 2 \cdot \sqrt{\frac{7}{2}} \right) = \frac{2\sqrt{14}}{9} = 0.8315$$

Prin urmare,

$$D_2(i) = \begin{cases} \frac{9}{2\sqrt{14}} \cdot \frac{1}{9} \cdot \sqrt{\frac{2}{7}} = \frac{1}{14} & \text{pentru } i \notin \{4, 5\}; \\ \frac{9}{2\sqrt{14}} \cdot \frac{1}{9} \cdot \sqrt{\frac{7}{2}} = \frac{1}{4} & \text{pentru } i \in \{4, 5\}. \end{cases}$$

Figura următoare prezintă grafic exemplele de antrenament (x_i, y_i) , pentru $i = 1, \dots, 9$, împreună cu probabilitățile care tocmai le-au fost asociate ($D_2(x_i)$), precum și separatorul „învățat“ la această primă iterăție (h_1).

Probabilitățile corespunzătoare instanțelor clasificate eronat de către h_1 au fost scrise cu culoarea roșie. Se observă că aceste probabilități sunt mult mai mari decât probabilitățile celorlalte instanțe și că împreună se sumează la valoarea $1/2$. Am notat cu $+$ și respectiv $-$ cele două zone de decizie determinate de h_1 .



Observație (1): Dacă, drept ipoteză h_1 , în locul lui $\text{sign}\left(\frac{7}{2} - X_2\right)$ am fi ales $\text{sign}\left(\frac{5}{2} - X_1\right)$, atunci cursul rezolvării ulterioare ar fi fost altul (deși ambele ipoteze au aceeași eroare ponderată – minimală – la antrenare): x_8 și x_9 ar fi primit ponderile $1/4$, iar x_4 și x_5 ar fi primit ponderile $1/14$. În consecință, outputul algoritmului AdaBoost nu este în mod neapărat unic determinat!

Iterația $t = 2$:

Vom proceda similar cu iterăția precedentă, doar că, de data aceasta, vom folosi distribuția D_2 .

s	$\frac{1}{2}$	$\frac{5}{2}$	$\frac{9}{2}$
$err_{D_2}(X_1 < s)$	$\frac{4}{14}$	$\frac{2}{14}$	$\frac{2}{14} + \frac{2}{4} + \frac{2}{14} = \frac{11}{14}$
$err_{D_2}(X_1 \geq s)$	$\frac{10}{14}$	$\frac{12}{14}$	$\frac{3}{14}$

s	$\frac{1}{2}$	$\frac{3}{2}$	$\frac{5}{2}$	$\frac{7}{2}$
$err_{D_2}(X_2 < s)$	$\frac{4}{14}$	$\frac{1}{4} + \frac{3}{14} = \frac{13}{28}$	$\frac{2}{4} + \frac{1}{14} = \frac{8}{14}$	$\frac{2}{4} = \frac{1}{2}$
$err_{D_2}(X_2 \geq s)$	$\frac{10}{14}$	$\frac{15}{28}$	$\frac{6}{14}$	$\frac{1}{2}$

Observație (2): Conform rezultatului teoretic de la punctul vi al problemei 22, luarea în calcul a compasului de decizie corespunzător testului $X_2 \geq 7/2$ este aici superfluă, întrucât acest compas de decizie a fost ales ca ipoteză optimală la iterată precedență. L-am pus totuși în tabel (folosind culoarea roșie), de dragul realizării unei prezentări exhaustive.

Cea mai bună ipoteză este acum $h_2 = sign\left(\frac{5}{2} - X_1\right)$; separatorul corespunzător acestei ipoteze este dreapta de ecuație $X_1 = \frac{5}{2}$. Urmează să explicităm cum am calculat eroarea ε_2 și, aferent, să calculăm „ecartul“ γ_2 , precum și ponderea α_2 , după care vom trece la determinarea noii distribuții (D_3).

$$\varepsilon_2 = P_{D_2}(\{x_8, x_9\}) = \frac{2}{14} = \frac{1}{7} = 0.143, \quad \text{deci } \gamma_2 = \frac{1}{2} - \frac{1}{7} = \frac{5}{14}$$

$$\alpha_2 = \ln \sqrt{\frac{1 - \varepsilon_2}{\varepsilon_2}} = \ln \sqrt{\left(1 - \frac{1}{7}\right) : \frac{1}{7}} = \ln \sqrt{6} = 0.896$$

$$\begin{aligned} D_3(i) &= \frac{1}{Z_2} \cdot D_2(i) \cdot (\underbrace{e^{-\alpha_2}}_{1/\sqrt{6}})^{y_i h_2(x_i)} = \begin{cases} \frac{1}{Z_2} \cdot D_2(i) \cdot \frac{1}{\sqrt{6}} & \text{dacă } h_2(x_i) = y_i; \\ \frac{1}{Z_2} \cdot D_2(i) \cdot \sqrt{6} & \text{în caz contrar} \end{cases} \\ &= \begin{cases} \frac{1}{Z_2} \cdot \frac{1}{14} \cdot \frac{1}{\sqrt{6}} & \text{pentru } i \in \{1, 2, 3, 6, 7\}; \\ \frac{1}{Z_2} \cdot \frac{1}{4} \cdot \frac{1}{\sqrt{6}} & \text{pentru } i \in \{4, 5\}; \\ \frac{1}{Z_2} \cdot \frac{1}{14} \cdot \sqrt{6} & \text{pentru } i \in \{8, 9\}. \end{cases} \end{aligned}$$

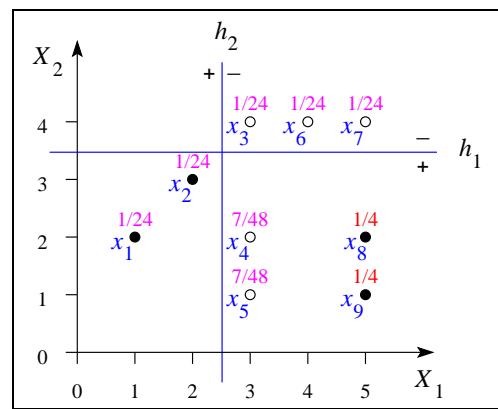
Din relația $\sum_i D_3(i) = 1$ rezultă:

$$\begin{aligned} Z_2 &= 5 \cdot \frac{1}{14} \cdot \frac{1}{\sqrt{6}} + 2 \cdot \frac{1}{4} \cdot \frac{1}{\sqrt{6}} + 2 \cdot \frac{1}{14} \cdot \sqrt{6} = \frac{5}{14\sqrt{6}} + \frac{1}{2\sqrt{6}} + \frac{\sqrt{6}}{7} = \frac{12 + 12}{14\sqrt{6}} \\ &= \frac{24}{14\sqrt{6}} = \frac{2\sqrt{6}}{7} \approx 0.7 \end{aligned}$$

și, prin urmare

$$D_3(i) = \begin{cases} \frac{7}{2\sqrt{6}} \cdot \frac{1}{14} \cdot \frac{1}{\sqrt{6}} = \frac{1}{24} & \text{pentru } i \in \{1, 2, 3, 6, 7\}; \\ \frac{7}{2\sqrt{6}} \cdot \frac{1}{4} \cdot \frac{1}{\sqrt{6}} = \frac{7}{48} & \text{pentru } i \in \{4, 5\}; \\ \frac{7}{2\sqrt{6}} \cdot \frac{1}{14} \cdot \sqrt{6} = \frac{1}{4} & \text{pentru } i \in \{8, 9\}. \end{cases}$$

Facem din nou reprezentarea grafică a datelor de antrenament, împreună cu noile probabilități asociate (D_3) și cei doi separatori care au fost „învățați” până acum (h_1 și h_2); vedeți figura alăturată. Este instructiv să comparăm această figură cu cea dinainte (adică, pentru $t = 1$), pentru a observa evoluția probabilităților de la o iterare la alta. Instanțele x_1, x_2, x_3, x_6 și x_7 au acum ponderile foarte mici, pentru că au fost clasificate corect atât de către h_1 cât și de către h_2 .



Iterația $t = 3$:

Erorile ponderate la antrenare pentru compașii de decizie, în raport cu distribuția D_3 sunt:

s	$\frac{1}{2}$	$\frac{5}{2}$	$\frac{9}{2}$
$err_{D_3}(X_1 < s)$	$\frac{2}{24} + \frac{2}{4} = \frac{7}{12}$	$\frac{2}{4}$	$\frac{2}{24} + 2 \cdot \frac{7}{48} + 2 \cdot \frac{1}{4} = \frac{21}{24}$
$err_{D_3}(X_1 \geq s)$	$\frac{5}{12}$	$\frac{2}{4}$	$\frac{3}{24} = \frac{1}{8}$

s	$\frac{1}{2}$	$\frac{3}{2}$	$\frac{5}{2}$	$\frac{7}{2}$
$err_{D_3}(X_2 < s)$	$\frac{7}{12}$	$\frac{7}{48} + \frac{2}{24} + \frac{1}{4} = \frac{23}{48}$	$2 \cdot \frac{7}{48} + \frac{1}{24} = \frac{1}{3}$	$2 \cdot \frac{7}{48} = \frac{7}{24}$
$err_{D_3}(X_2 \geq s)$	$\frac{5}{12}$	$\frac{25}{48}$	$\frac{2}{3}$	$\frac{17}{24}$

Similar cu *Observația* precedentă (vedeți iterare $t = 2$), facem mențiunea că am fi putut renunța la elaborarea compasului de decizie corespunzător testului $X_1 < 5/2$.

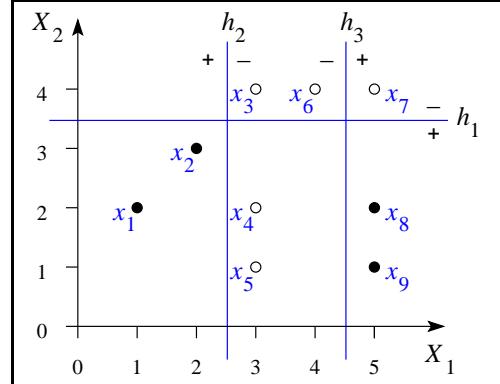
Acum cea mai bună ipoteză este $h_3 = \text{sign}\left(X_1 - \frac{9}{2}\right)$, iar separatorul corespunzător este dreapta de ecuație $X_1 = \frac{9}{2}$. Vom explicita cum am calculat eroarea ε_3 și apoi vom calcula „ecartul” γ_3 , precum și ponderea α_3 :

$$\varepsilon_3 = P_{D_3}(\{x_1, x_2, x_7\}) = 2 \cdot \frac{1}{24} + \frac{1}{24} = \frac{3}{24} = \frac{1}{8}$$

$$\gamma_3 = \frac{1}{2} - \frac{1}{8} = \frac{3}{8}$$

$$\alpha_3 = \ln \sqrt{\frac{1 - \varepsilon_3}{\varepsilon_3}} = \ln \sqrt{\left(1 - \frac{1}{8}\right) : \frac{1}{8}} = \ln \sqrt{7} = 0.973$$

Reprezentarea grafică a datelor de antrenament împreună cu cei trei separatori învățați este furnizată în figura alăturată.



În final, vom pune rezultatele pe care le-am obținut până acum în tabelul care a fost dat în enunț.

t	ε_t	α_t	$D_t(1)$	$D_t(2)$	$D_t(3)$	$D_t(4)$	$D_t(5)$	$D_t(6)$	$D_t(7)$	$D_t(8)$	$D_t(9)$	$err_S(H_t)$
1	$2/9$	$\ln \sqrt{7/2}$	$1/9$	$1/9$	$1/9$	$1/9$	$1/9$	$1/9$	$1/9$	$1/9$	$1/9$	$2/9$
2	$2/14$	$\ln \sqrt{6}$	$1/14$	$1/14$	$1/14$	$1/4$	$1/4$	$1/14$	$1/14$	$1/14$	$1/14$	$2/9$
3	$1/8$	$\ln \sqrt{7}$	$1/24$	$1/24$	$1/24$	$7/48$	$7/48$	$1/24$	$1/24$	$1/4$	$1/4$	0

Observație (3): Tabelul de mai jos vă servește ca să înțelegeți cum anume se calculează $err_S(H_t)$. În acest tabel, am marcat cu culoarea roșie clasificările eronate făcute de ipotezele h_t , pentru $t = 1, 2, 3$.

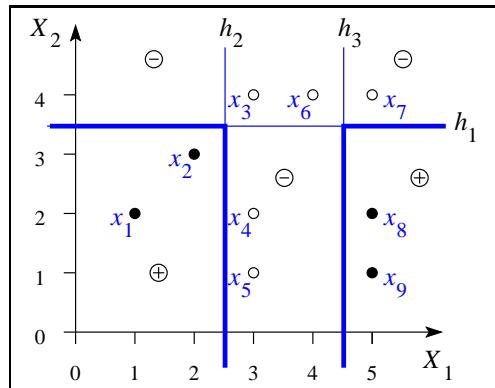
t	α_t	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
1	0.626	+1	+1	-1	+1	+1	-1	-1	+1	+1
2	0.896	+1	+1	-1	-1	-1	-1	-1	-1	-1
3	0.973	-1	-1	-1	-1	-1	+1	+1	+1	+1
	$H_T(x_i)$	+1	+1	-1	-1	-1	-1	-1	+1	+1

Se observă că H_1 (deci, la finalul primei iterații a algoritmului AdaBoost) a clasificat greșit instanțele x_4 și x_5 , iar H_2 (deci, la finalul celei de-a doua iterații) instanțele x_8 și x_9 . La finalul celei de-a treia iterații, eroarea produsă pe datele de antrenament (de către ipoteza H_3) este 0. Vă readucem aminte că $H_T(x_i) \stackrel{\text{def.}}{=} \text{sign} \left(\sum_{t=1}^T \alpha_t h_t(x_i) \right)$, deci notând $x_i = (x_{i,1}, x_{i,2})$, putem scrie

$$H_3(x_i) = \text{sign} \left(\ln \sqrt{\frac{7}{2}} \cdot \text{sign} \left(\frac{7}{2} - x_{i,2} \right) + \ln \sqrt{6} \cdot \text{sign} \left(\frac{5}{2} - x_{i,1} \right) + \ln \sqrt{7} \cdot \text{sign} \left(x_{i,1} - \frac{9}{2} \right) \right).$$

Observație (4): Se poate constata imediat că o instanță nouă aleasă în mod arbitrar în partea din stânga sus a figurii care reprezintă datele de antrenament (de exemplu, instanța $(1, 4)$) va fi clasificată de către ipoteza H_3 — care a fost „învățată“ de către algoritmul AdaBoost după cele trei iterării — ca fiind negativă, fiindcă $-\alpha_1 + \alpha_2 - \alpha_3 = -0.626 + 0.896 - 0.973 < 0$.

Făcând și alte raționamente similare, putem conchide că *zonele de decizie și granițele de decizie* produse de către AdaBoost pentru acest set de date de antrenament vor fi cele indicate în figura alăturată.

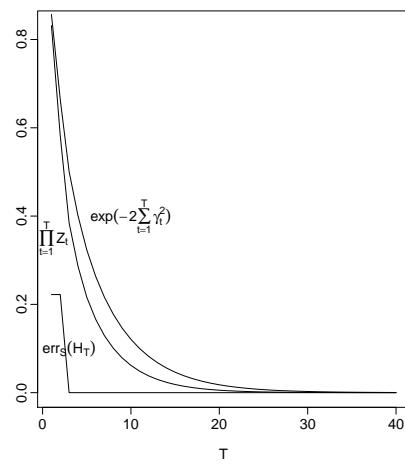
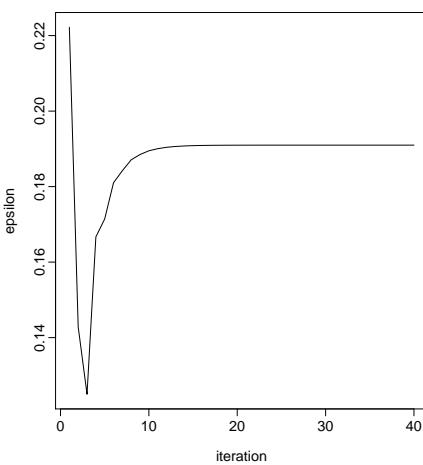


Observație (5): Execuția algoritmului AdaBoost ar putea continua (dacă am fi fixat inițial un $T > 3$), chiar dacă am obținut $err_S(H_t) = 0$ la iterarea $t = 3$. Dacă veți elabora detaliile, veți vedea că pentru $t = 4$ am obține ca ipoteză optimă $X_2 < \frac{7}{2}$ (care a fost selectată și la iterarea $t = 1$). Această ipoteză ar produce acum la antrenare eroarea ponderată $\varepsilon_4 = \frac{1}{6}$, deci ar primi în noul output H_4 factorul $\alpha_4 = \ln \sqrt{5}$ (care s-ar alătura factorului $\alpha_1 = \ln \sqrt{7/2}$). Astfel se va întări încrederea în ipoteza $X_2 < \frac{7}{2}$. Să reținem, deci, că algoritmul AdaBoost poate selecta de mai multe ori o aceeași ipoteză „slabă“ (însă niciodată la iterării consecutive, conform pr. 22.vi).

În final prezentăm două grafice utile, care au fost realizate de către studentul Sebastian Ciobanu:

Evoluția erorii ε_t în raport cu t :

Cele două margini superioare (eng., upper bounds) pentru eroarea empirică a ipotezei combinate H_T (cf. ex. 23.bd):

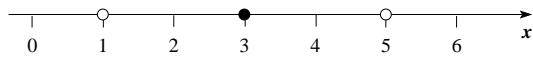


25.

(AdaBoost și non-învățabilitate γ -slabă:
exemplificare pe un set de date din \mathbb{R})

■ • ○ CMU, 2012 fall, T. Mitchell, Z. Bar-Joseph, final, pr. 8.a-e

În această problemă vom studia modul în care se comportă algoritmul AdaBoost pe un set de date foarte simplu din \mathbb{R} , $x_1 = 1, x_2 = 3, x_3 = 5$, ilustrat în figura alăturată.



Vă readucem aminte *convenția* noastră de notare: simbolul • desemnează instanțe pozitive, iar simbolul ○ instanțe negative.

Vom folosi ca ipoteze „slabe” (h_t) compași de decizie. Vă readucem aminte că un compas de decizie alege o valoare reală constantă s și clasifică toate punctele $x > s$ ca fiind într-o clasă, iar toate celelalte puncte ($x \leq s$) în cealaltă clasă.

- i. Care este valoarea ponderilor / probabilităților asignate inițial fiecărei instanțe? ii. Marcați pe desenul de mai sus separatorul decizional (granița) corespunzătoare primului compas de decizie. Indicați zona pozitivă și zona negativă, de o parte și de alta a separatorului decizional. iii. Încercuiți instanța a cărei pondere / probabilitate crește la execuția primei iterații de boosting. iv. Calculați ponderile / probabilitățile asignate fiecărei instanțe după prima iterație.
- b. Poate oare algoritmul AdaBoost să clasifice în mod perfect toate aceste exemple de antrenament? Dacă răspunsul este *nu*, justificați în detaliu. Dacă răspunsul este *da*, care este numărul minim de iterații în care se atinge eroarea la antrenare 0?

Răspuns:

Observație: La stabilirea distribuțiilor probabiliste D_{t+1} (pentru $t = 1, 2, \dots$), vom ține cont

- în primul rând de rezultatul de la exercițiul 22.vi, și anume că $D_{t+1}(h_t) = 1/2$,
- iar apoi de faptul — specificat în relația de definiție (227) pentru distribuția D_{t+1} — că probabilitățile instanțelor de antrenament care au fost corect (respectiv incorect) catalogate de către ipoteza slabă h_t vor fi modificate (pentru iterația $t + 1$) proporțional, și anume cu factorul $e^{-\alpha_t}$ (respectiv e^{α_t}).⁵²¹

- a. La iterația $t = 1$, se lucrează cu distribuția probabilistă $D_1(i) = 1/3$ pentru $i \in \{1, 2, 3\}$ și, indiferent de cum am fixa pragul compasului de decizie h_1 ,⁵²² — în legătură cu care nu precizăm deocamdată decât că trebuie să aibă eroarea ponderată la antrenare minimală, ca de obicei — o [singură] instanță este

⁵²¹Mai precis, folosind notațiile de la ex. 22, $D_{t+1}(i) = \frac{1}{Z_t} D_t(i) e^{-\alpha_t}$, pentru $i \in C$, și $D_{t+1}(i) = \frac{1}{Z_t} D_t(i) e^{\alpha_t}$, pentru $i \in M$.

⁵²²Atât pentru varianta când se folosește prag exterior (engl., outside threshold) cât și pentru varianta când nu se folosește prag exterior. Vedeți precizările de mai jos.

incorrect clasificată. Dacă folosim praguri exterioare, pentru ipoteza slabă h_1 putem alege pragul 0 (vedeți primul desen de mai jos). În acest caz, instanța incorrect clasificată este $x_2 = 3$.

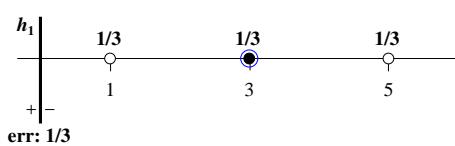
Pentru iterația $t + 1 = 2$, instanța x_2 va primi (cf. rezultatul teoretic de la ex. 22.vi) ponderea / probabilitatea $1/2$, în vreme ce probabilitatea care rămâne, și anume $1 - (1/2) = 1/2$ va fi împărțită în mod egal între instanțele $x_1 = 1$ și $x_3 = 5$; aşadar acestea vor primi acum ponderile $1/4$.

b. La iterația $t = 2$, instanța incorrect clasificată de către ipoteza slabă h_2 (vedeți detaliile din al doilea desen de mai jos) este $x_3 = 5$. În continuare, la iterăția $t + 1 = 3$, instanța x_3 va primi probabilitatea $1/2$, iar restul masei de probabilitate $(1/2)$ va fi împărțit în mod proporțional între instanțele $x_1 = 1$ și $x_2 = 3$, deci acestea vor avea acum probabilitățile $\frac{1}{6} = \frac{1}{2} \cdot \frac{1}{3}$ și respectiv $\frac{1}{3} = \frac{1}{2} \cdot \frac{2}{3}$.

Erorile ponderate ε_t se calculează ușor, la fel și ponderile α_t , să că vom reda acum în mod direct rezultatele obținute la fiecare iterăție (presupunând, ca și mai sus, că folosim praguri exterioare).

Se observă că după $T = 3$ iterății, algoritmul AdaBoost obține eroarea la antrenare $err_S(H_3) = 0$.

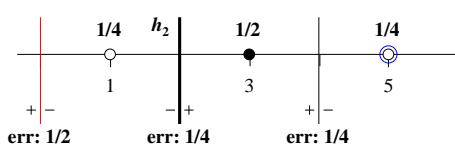
• $t = 1 :$



• $t = 1 :$

$$\varepsilon_1 = 1/3 \Rightarrow \alpha_1 = \ln \sqrt{2} = 0.3465, \\ err_S(H_1) = 1/3$$

• $t = 2 :$



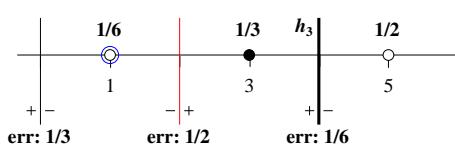
• $t = 2 :$

$$\varepsilon_2 = 1/4 \Rightarrow \alpha_2 = \ln \sqrt{3} = 0.5493$$

	x_1	x_2	x_3
α_1	—	—	—
α_2	—	+	+
$H_2(x_i)$	—	+	+

$$\Rightarrow err_S(H_2) = 1/3$$

• $t = 3 :$



• $t = 3 :$

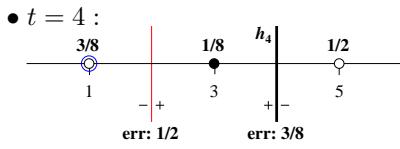
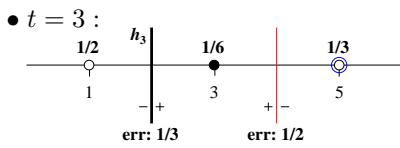
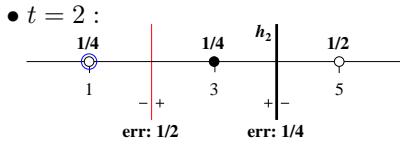
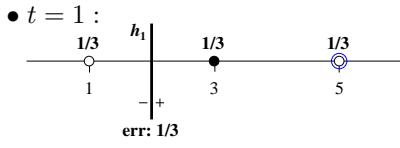
$$\varepsilon_3 = 1/6 \Rightarrow \alpha_3 = \ln \sqrt{5} = 0.8047$$

	x_1	x_2	x_3
α_1	—	—	—
α_2	—	+	+
α_3	+	+	—
$H_3(x_i)$	—	+	—

$$\Rightarrow err_S(H_3) = 0$$

Atenție! Întrucât se va dovedi a fi un fapt foarte instructiv, vom elabora acum o *a doua soluție*, pentru care însă ne vom auto-impune din start să nu folosim praguri exterioare.⁵²³ Iată soluția obținută de data aceasta (mai exact, rezultatul primelor 4 iterății):

⁵²³Aceasta a fost, în esență, soluția dată pentru acest exercițiu la CMU, însă ea este greșită, iar când spunem aceasta ne bazăm pe rezultatul teoretic de la problema 71.d, cu care această soluție (de la CMU) vine în contradicție.



• $t = 1 :$

$$\varepsilon_1 = 1/3 \Rightarrow \alpha_1 = \ln \sqrt{2} = 0.3465, \text{ errs}(H_1) = 1/3$$

• $t = 2 : \varepsilon_2 = 1/4 \Rightarrow \alpha_2 = \ln \sqrt{3} = 0.5493$

	x_1	x_2	x_3
α_1	-	+	+
α_2	+	+	-
$H_2(x_i)$	+	+	-

$$\Rightarrow \text{errs}(H_2) = 1/3$$

• $t = 3 : \varepsilon_3 = 1/3 = \varepsilon_1 \Rightarrow \alpha_3 = \ln \sqrt{2} = 0.3465 = \alpha_1$

	x_1	x_2	x_3
α_1	-	+	+
α_2	+	+	-
α_3	-	+	+
$H_3(x_i)$	-	+	+

$$\Rightarrow \text{errs}(H_3) = 1/3$$

• $t = 4 : \varepsilon_4 = 3/8 \Rightarrow \alpha_4 = \ln \sqrt{5/3} = 0.2554$

	x_1	x_2	x_3
α_1	-	+	+
α_2	+	+	-
α_3	-	+	+
α_4	+	+	-
$H_4(x_i)$	+	+	-

$$\Rightarrow \text{errs}(H_4) = 1/3$$

Se observă că

- la oricare două iterații consecutive, pragurile alese ($s = 2$ și $s = 4$) vor alterna, conform unei consecințe imediate de la relația demonstrată la ex. 22.vi (vedeți *Observația 2* de la pagina 528);
- instanța $x_2 = 3$ este clasificată corect de către oricare ipoteză selectată, h_t , de aceea, conform relației (227) ponderea ei va scădea mereu (observați valorile $1/3, 1/4, 1/6, 1/8$). Mai mult, se poate demonstra că această probabilitate tinde descrescător la 0.⁵²⁴
- conform aceliei relații de la ex. 22.vi, va rezulta că la fiecare iterație una dintre instanțele $x_1 = 1$ și $x_3 = 5$ va avea probabilitatea $1/2$, iar cealaltă instanță va avea probabilitatea $1/2$ minus probabilitatea instanței (corect clasificate) $x_2 = 3$. Așadar, această ultimă probabilitate va tinde crescător la $1/2$. În consecință, în acest caz — adică, atunci când nu se lucrează cu praguri exterioare, pe acest set de date de antrenament — nu există *garanție de învățabilitate empirică slabă* ($\gamma > 0$).

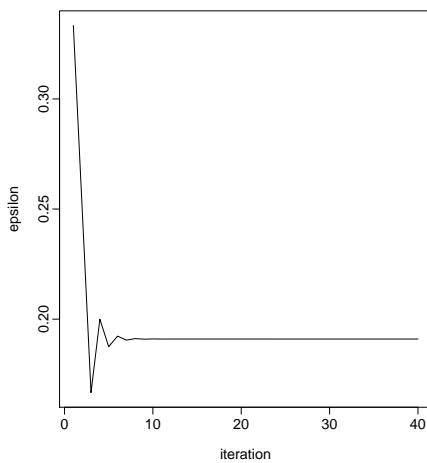
Observație: Se poate constata — din tabelele pentru $\text{errs}(H_T)$ și în special din ultimul tabel — că $H_T(x_1)$ va avea întotdeauna semnul contrar semnului lui $H_T(x_3)$, deoarece la fiecare iterație (t) instanțele x_1 și x_3 vor fi clasificate cu semne contrare de către ipoteza slabă selectată, h_t . Prin urmare, în acest caz — repetăm, atunci când nu se lucrează cu praguri exterioare, pe acest set de date de antrenament — nu este posibil ca AdaBoost să obțină eroare la antrenare 0.

⁵²⁴Vedeți graficul de mai jos. Analitic, la această concluzie se poate ajunge prin reducere la absurd, pornind de la rezultatul teoretic obținut la ex. 23.e, coroborat cu *Observația* de mai jos.

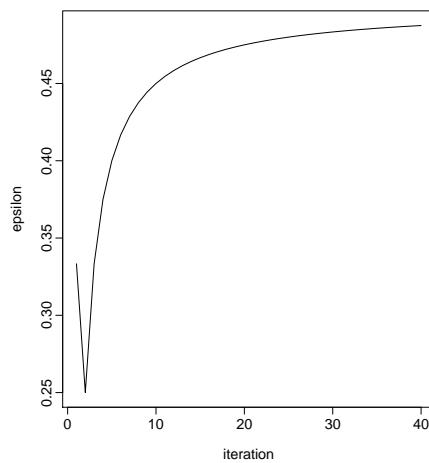
În concluzie, este absolut necesar ca în toate cazurile, la aplicarea algoritmului AdaBoost să fie folosite parguri exterioare (mai precis, este suficient un prag exterior). Altfel, este posibil ca AdaBoost să nu poată să clasifice perfect datele de antrenament.⁵²⁵

Graficele următoare, făcute de Sebastian Ciobanu, arată evoluția valorilor *erorilor ponderate la antrenare* care au fost produse atunci când am rulat 40 de iterații cu algoritmul AdaBoost pe aceste date, în cele două variante: cu și respectiv fără prag exterior. Se observă clar existența unui *prag $\gamma > 0$* de învățabilitate empirică slabă în primul caz, și lipsa unui astfel de prag în cel de-al doilea caz.

cu prag exterior:

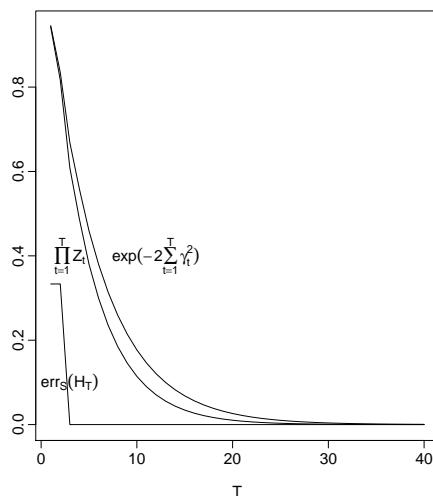


fără prag exterior:

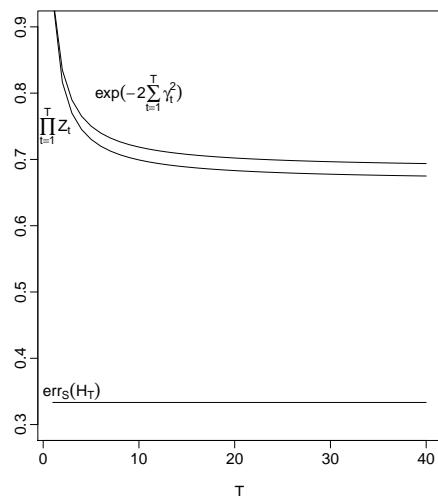


În sfârșit, următoarele grafice, făcute tot de Sebastian Ciobanu, pun în evidență legătura dintre rezultatele obținute mai sus și cele două *margini superioare* pentru *eroarea la antrenare* produsă de *ipoteza combinată* H_T livrată la ieșire de algoritmul AdaBoost pe aceste date (vedeți ex. 23.bd).

cu prag exterior:



fără prag exterior:



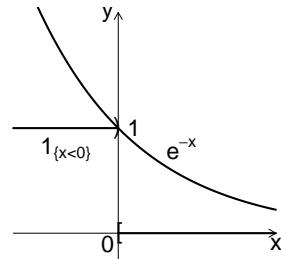
⁵²⁵Exercițiul 66 pune în evidență o situație (adică, un set de date) pentru care AdaBoost nu obține învățabilitate empirică γ -slabă, chiar dacă se folosesc [și] compași de decizie exterioiri.

26. (Algoritmul AdaBoost ca algoritm de *optimizare secvențială*: deducerea regulii de calcul a ponderilor α_t folosind *metoda minimizării secvențiale* a costului [negativ] exponențial; monotonia costurilor [negativ] exponențiale minime: $J_t^* \leq J_{t-1}^*$)

■ • ○ CMU, 2008 fall, Eric Xing, HW3, pr. 4.1.1
CMU, 2012 fall, E. Xing, A. Singh, HW4, pr. 3.ab

La problema 23.c am văzut că în algoritmul AdaBoost se urmărește (în mod indirect) să se minimizeze eroarea la antrenare $err_S(H_T)$ prin minimizarea *secvențială* a marginii sale superioare $\prod_{t=1}^T Z_t$. Aceasta înseamnă că la fiecare iterație t (unde $1 \leq t \leq T$) alegem valoarea ponderii α_t astfel încât să minimizăm Z_t (văzut ca funcție de α_t , conform relației (228) de la problema problema 22.i).

Aici veți vedea că o altă cale de a explica modul în care funcționează algoritmul AdaBoost este determinată de *obiectivul* de a minimiza în mod *greedy* și *secvențial costul / pierdere exponențială* (engl., the exponential loss):⁵²⁶



$$J_T \stackrel{\text{def.}}{=} \frac{1}{m} \sum_{i=1}^m \exp(-y_i f_T(x_i)) \stackrel{\text{not.}}{=} \frac{1}{m} \sum_{i=1}^m \exp(-y_i \sum_{t=1}^T \alpha_t h_t(x_i)). \quad (234)$$

Aceasta revine la a spune că la fiecare iterare $t \in \{1, \dots, T\}$ urmărim să alegem — pe lângă cea mai bună ipoteză „slabă” h_t — o valoare pentru ponderea α_t astfel încât costul / „pierdere“ J_t să fie minimizată.

În acest execițiu se demonstrează că această nouă strategie va conduce la aceeași regulă de actualizare pentru ponderea α_t folosită în algoritmul AdaBoost, adică $\alpha_t = \frac{1}{2} \ln \frac{1 - \varepsilon_t}{\varepsilon_t}$.

a. Arătați că

$$J_t = c_t \cdot \underbrace{\sum_{i=1}^m D_t(i) \cdot \exp(-y_i \alpha_t h_t(x_i))}_{\text{not.: } \tilde{J}_t}$$

unde c_t este un factor constant, care nu depinde de α_t și nici de h_t , dar pe care îl veți preciza, iar

$$\tilde{J}_t = (1 - \varepsilon_t) \cdot e^{-\alpha_t} + \varepsilon_t \cdot e^{\alpha_t},$$

unde ε_t este eroarea ponderată la antrenare pentru ipoteza „slabă” h_t .

⁵²⁶Pentru o definiție [de lucru] pentru funcțiile de cost / pierdere, vedeți *Explicația* de la problema 29.

Veți observa că pentru orice h_t și α_t fixați, $J_t \geq err_S(H_t) \stackrel{\text{def.}}{=} \frac{1}{m} \sum_{i=1}^m 1_{\{y_i \neq \text{sign}(f_t(x_i))\}}$. Pentru aceasta, analizați separat cazurile $y_i f_t(x_i) < 0$ și $y_i f_t(x_i) \geq 0$. Pentru a minimiza $err_S(H_t)$, aici vom urmări (în principal) să minimizăm J_t în funcție de α_t .

Sugestie: Veți putea folosi rezultatul obținut la problema 23.a, și anume că $D_t(i) \propto \exp(-y_i f_{t-1}(x_i))$, adică probabilitatea $D_t(i)$ este proporțională cu $\exp(-y_i f_{t-1}(x_i))$.⁵²⁷

b. Arătați că dacă se consideră $\alpha_t > 0$ fixat, atunci pentru a minimiza expresia / valoarea lui J_t trebuie să alegem dintre toate ipotezele „slabe“ pe aceea (notată cu h_t) pentru care eroarea ponderată la antrenare (notată cu ε_t) este minimă.

c. Arătați că dacă se consideră h_t fixat, atunci minimizarea lui J_t revine la a atribui lui α_t valoarea $\ln \sqrt{\frac{1 - \varepsilon_t}{\varepsilon_t}}$.

d. Rescrieți pseudo-codul algoritmului AdaBoost ca un algoritm de *optimizare secvențială* a costului [negativ] exponențial al clasificării instantelor de antrenament.

e. Dacă notăm cu J_t^* valoarea optimă a lui J_t (văzut ca funcție de ipoteza „slabă“ h_t și votul $\alpha_t > 0$), veți constata că are loc o anumită relație de inegalitate între J_t^* și J_{t-1}^* pentru orice $t > 1$. Care este această relație?

Dar între J_t^* pe de o parte și Z_1, Z_2, \dots, Z_t , pe de altă parte?

Răspuns:

a. La iterația t vom avea:

$$\begin{aligned} J_t &= \frac{1}{m} \sum_{i=1}^m \exp(-y_i f_t(x_i)) = \frac{1}{m} \sum_{i=1}^m \exp \left(-y_i \left(\sum_{t'=1}^{t-1} \alpha_{t'} h_{t'}(x_i) \right) - y_i \alpha_t h_t(x_i) \right) \\ &= \frac{1}{m} \sum_{i=1}^m \exp(-y_i f_{t-1}(x_i)) \cdot \exp(-y_i \alpha_t h_t(x_i)) \\ &= \frac{1}{m} \sum_{i=1}^m \left(m \prod_{t'=1}^{t-1} Z_{t'} \right) \cdot D_t(i) \cdot \exp(-y_i \alpha_t h_t(x_i)) \\ &= \left(\prod_{t'=1}^{t-1} Z_{t'} \right) \cdot \sum_{i=1}^m D_t(i) \cdot \exp(-y_i \alpha_t h_t(x_i)) \end{aligned} \quad (235)$$

$$\propto \sum_{i=1}^m D_t(i) \cdot \exp(-y_i \alpha_t h_t(x_i)) \stackrel{\text{not.}}{=} \tilde{J}_t. \quad (236)$$

Mai departe, ținând cont de faptul că $y_i \in \{-1, +1\}$ și $h_t(x_i) \in \{-1, +1\}$ pentru $i = 1, \dots, m$ și $t = 1, \dots, T$, putem scrie expresia \tilde{J}_t astfel:

$$\begin{aligned} \tilde{J}_t &= \sum_{i=1}^m D_t(i) \cdot \exp(-y_i \alpha_t h_t(x_i)) \\ &= \sum_{i \in C} D_t(i) \exp(-\alpha_t) + \sum_{i \in M} D_t(i) \exp(\alpha_t) \\ &= \exp(-\alpha_t) \underbrace{\sum_{i \in C} D_t(i)}_{1 - \varepsilon_t} + \exp(\alpha_t) \underbrace{\sum_{i \in M} D_t(i)}_{\varepsilon_t} \\ &= (1 - \varepsilon_t) \cdot e^{-\alpha_t} + \varepsilon_t \cdot e^{\alpha_t}, \end{aligned} \quad (237)$$

⁵²⁷Mai exact, $D_t(i) = \frac{1}{mZ_1 \dots Z_{t-1}} \exp(-y_i f_{t-1}(x_i))$.

unde, ca și la problema 22, C este mulțimea [indicilor] exemplelor de antrenament care sunt corect clasificate de către ipoteza h_t , iar M este mulțimea [indicilor] exemplelor de antrenament care sunt incorect clasificate de către h_t .

b. Dacă rescriem expresia (237) sub forma

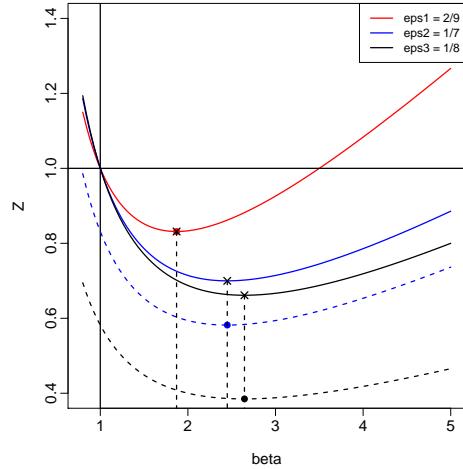
$$e^{-\alpha_t} + \varepsilon_t \underbrace{(e^{\alpha_t} - e^{-\alpha_t})}_{>0},$$

se observă că dacă fixăm α_t atunci a minimiza expresia J_t în raport cu h_t revine la a minimiza ε_t (care nu depinde de α_t).

c. Expresia (237) este identică cu expresia (228) pentru factorul de normalizare Z_t de la problema 22.i.⁵²⁸ La problema 23.c am arătat că expresia (228) își atinge minimul pentru $\alpha_t = \frac{1}{2} \ln \frac{1 - \varepsilon_t}{\varepsilon_t}$. Prin urmare, ținând cont de faptul că

ε_t se consideră fixat, „costul“ J_t va fi minimizat pentru același $\alpha_t = \frac{1}{2} \ln \frac{1 - \varepsilon_t}{\varepsilon_t}$.

Exemplificare: În graficul alăturat, realizat de către Sebastian Ciobanu, sunt reprezentate funcțiile \tilde{J}_t și J_t corespunzătoare celor trei iterații realizate de către algoritmul AdaBoost, care au fost descrise la problema 24. Pentru conveniență, am folosit notația $\beta = e^\alpha$ și am reprezentat funcțiile în raport cu acest argument. În acest grafic, liniile continue corespund funcțiilor \tilde{J}_t , iar liniile discontinue corespund funcțiilor J_t asociate lor: $J_t(\beta) = \left(\prod_{t'=1}^{t-1} Z_{t'} \right) \cdot \tilde{J}_t(\beta)$. Remarcați că din relația (235) rezultă că $J_1(\beta) = \tilde{J}_1(\beta), \forall \beta$.



d. Ca o consecință a punctelor b și c , algoritmul AdaBoost poate fi reformulat ca un algoritm de *optimizare secvențială* a costurilor / „pierderilor“ J_t . La fiecare iterație a algoritmului se adaugă în mod „greedy“ căte o ipoteză „slabă“ $h \in \mathcal{H}$ la ipoteza combinată curentă pentru a minimiza costul determinat de funcția ϕ , și cu obiectivul final de a minimiza eroarea la antrenare $errs(H_T)$. (De aceea această metodă de optimizare se numește *optimizare secvențială*.) Iată care este noua formulare a algoritmului AdaBoost:

Intrare: $\{(x_i, y_i)\}_{i=1,\dots,m}$ – setul de date de antrenament, T – numărul de iterații de executat, \mathcal{H} – setul de ipoteze „slabe“, $\phi(y, y') \stackrel{\text{def.}}{=} \exp(-yy')$ – funcția de cost / pierdere.

Procedură:

Initializare: $f_0(x) = 0$ și $D_1(i) = 1/m$ pentru $i = 1, \dots, m$
pentru t de la 1 la T execută:

begin

1. Calculează

⁵²⁸[S. Ciobanu:] De fapt, se putea vedea chiar din relația (236) că $\tilde{J}_t = Z_t$ (văzut ca funcție de α_t), ținând cont de faptul că Z_t este factor de normalizare în definiția distribuției D_{t+1} (vedeți relația (227)).

$$(h_t, \alpha_t) = \arg \min_{h \in \mathcal{H}, \alpha \in \mathbb{R}_+} \sum_{i=1}^m \phi(y_i, f_{t-1}(x_i) + \alpha h(x_i))$$

$\underbrace{\hspace{10em}}_{J_t(h, \alpha)}$

2. Actualizează clasificatorul $f_t(x) = f_{t-1}(x) + \alpha_t h_t(x)$
și calculează noua distribuție, D_{t+1}
end
returnează clasificatorul $\text{sign}(f_T(x))$

Precizăm faptul că minimizarea efectuată la pasul 1 din corpul iterativ al acestui algoritm este făcută *pe coordonate*. Aceasta înseamnă că mai întâi minimizăm funcția J_t în raport cu argumentul h (care nu depinde de α) și apoi — cu h fixat la valoarea pe care tocmai am găsit-o, h_t — minimizăm funcția J_t în raport cu argumentul α (care depinde de h_t).

e. Valorile minime ale funcțiilor J_t nu cresc niciodată de la o iterație la alta, fiindcă $\min_{h \in \mathcal{H}, \alpha \in \mathbb{R}_+} J_t(h, \alpha) \leq \sum_{i=1}^m \phi(y_i, f_{t-1}(x_i))$. Introducând notația $J_t^* = \min_{h \in \mathcal{H}, \alpha \in \mathbb{R}_+} J_t(h, \alpha)$, putem scrie aceasta inegalitate sub forma $J_t^* \leq J_{t-1}^*, \forall t \geq 1$. Din relația (235) putem deduce imediat că

$$J_t^* = \prod_{t'=1}^t Z_{t'},$$

fiindcă expresia (237) are valoarea Z_t atunci când îi atribuim lui α_t valoarea $\frac{1}{2} \ln \frac{1 - \varepsilon_t}{\varepsilon_t}$.⁵²⁹ Pe graficul precedent se poate vedea că într-adevăr $J_t^* \leq J_{t-1}^*$, pentru $t \in \{2, 3\}$, întrucât $Z_1 = 0.8315$, $Z_2 = 0.7$ și $Z_3 = 0.6614$, cf. problemei 24.

Observație importantă: La problema 29 vom arăta că putem adapta algoritmul nou introdus astfel încât să lucrăm cu orice funcție de cost dorim în locul funcției de cost [negativ] exponentiale ϕ .

27.

(Noțiunea de *margine de votare* în conexiune cu algoritmul AdaBoost; o referire la chestiunea overfitting-ului)

■ □ • · CMU, 2016 spring, W. Cohen, N. Balcan, HW4, pr. 3.3

Deși la aplicarea algoritmului AdaBoost *complexitatea modelului* produs crește la fiecare iterare, *în general nu se produce overfitting*. Motivul este că modelul capătă din ce în ce mai multă „încredere“ pe măsură ce numărul de iterări executate crește. Această „încredere“ (engl., confidence) poate fi exprimată din punct de vedere matematic cu ajutorul noțiunii de *margine de votare* (engl., voting margin).⁵³⁰

⁵²⁹LC: Mulțumesc lui Sebastian Ciobanu pentru dialogul în urma căruia am identificat această proprietate. Din relația $J_t^* = \prod_{t'=1}^t Z_{t'}$, știind că $Z_t \in (0, 1)$ (vedeți problema 22.iii), se obține imediat o a doua demonstrație pentru faptul că $J_t^* \leq J_{t-1}^*, \forall t \geq 1$.

⁵³⁰În mod concret, „încrederea“ în clasificarea lui x_i la iterarea t va fi definită ca $|\text{Margin}_k(x_i)|$, unde $\text{Margin}_k(x_i)$ urmează să fie definită mai jos.

Vă readucem aminte că,⁵³¹ după efectuarea celor T iterații, algoritmul AdaBoost livrează la ieșire clasificatorul H_T , definit astfel:

$$H_T(x) = \text{sign}\left(\sum_{t=1}^T \alpha_t h_t(x)\right).$$

În mod similar, putem defini *clasificatorul ponderat intermedian* (engl., intermediate weighted classifier) după k iterații: $H_k(x) = \text{sign}\left(\sum_{t=1}^k \alpha_t h_t(x)\right)$.

Întrucât outputul acestei funcții este $+1$ sau -1 , aceasta nu ne comunică nimic despre încrederea [pe care o putem avea] în deciziile acestui clasificator. De aceea, fără a afecta regula de decizie în sine, o putem redefini sub forma

$$H_k(x) = \text{sign}\left(\sum_{t=1}^k \bar{\alpha}_t h_t(x)\right),$$

unde $\bar{\alpha}_t = \frac{\alpha_t}{\sum_{t'=1}^k \alpha_{t'}}$, deci ponderile (engl., weights) ipotezelor „slabe“ sunt acum normalizate, în sensul că $\sum_{t=1}^k \bar{\alpha}_t = 1$ și, desigur, $\bar{\alpha}_t > 0$.⁵³²

Marginea de votare a instanței de antrenament x_i după iterația k se definește ca fiind diferența dintre suma ponderilor / voturilor normalize ale acelor h_t — evident, $t \in \{1, \dots, k\}$ — care clasifică corect instanța x_i și suma ponderilor / voturilor normalize ale acelor h_t care îl clasifică incorect pe x_i .⁵³³ Așadar,

$$\text{Margin}_k(x_i) = \sum_{t:h_t(x_i)=y_i} \bar{\alpha}_t - \sum_{t:h_t(x_i) \neq y_i} \bar{\alpha}_t.$$

Observație: Este imediat că $\text{Margin}_k(x_i) \in [-1, 1]$, pentru orice instanță x_i , cu $i = 1, \dots, m$ și pentru orice iterație k a algoritmului AdaBoost.

a. Fie $f_k(x) \stackrel{\text{not.}}{=} \sum_{t=1}^k \bar{\alpha}_t h_t(x)$. Arătați că $\text{Margin}_k(x_i) = y_i f_k(x_i)$ pentru orice instanță de antrenament x_i , cu $i = 1, \dots, m$.

Așadar, avem în acest fel o legătură foarte semnificativă între noțiunea de *margine de votare* nou-introdusă și noțiunea de *margine algebraică* introdusă la rezolvarea problemei 23.a: atunci când toate voturile α_t sunt normalizate, noțiunea de *margine de votare* coincide cu noțiunea de *margine algebraică*. (De aceea, în cele ce urmează, atunci când nu este niciun pericol de confuzie ne vom referi la aceste două noțiuni pur și simplu prin termenul *margine*.)

Consecință imediată: x_i , o instanță de antrenament oarecare, este corect clasificată de către ipoteza combinată H_k produsă de către AdaBoost la iterația k dacă și numai dacă $\text{Margin}_k(x_i) \geq 0$.

b. Dacă $\text{Margin}_k(x_i) > \text{Margin}_k(x_j)$, care dintre probabilitățile asociate celor două instanțe va fi mai mare la iterația $k+1$ (adică, $D_{k+1}(i)$ sau $D_{k+1}(j)$)?

⁵³¹Vedeți enunțul problemei 22.

⁵³²Aceasta normalizare, referitoare la ponderile α_t (asociate ipotezelor „slabe“ h_t), nu trebuie confundată cu normalizarea sau, de fapt, normalizările reprezentate de factorii Z_t din pseudo-codul algoritmului AdaBoost — vedeți enunțul problemei 22 —, care se referă la distribuțiile probabiliste D_t (asociate, la fiecare iterație, setului de instanțe de antrenament).

⁵³³Noțiunea de *margine geometrică* (definită altfel decât cele două tipuri de *margine* de aici) este specifică clasificatorului SVM. Vedeți capitolul *Mașini cu vectori-suport*.

Sugestie: Din relația $D_{k+1}(i) = \frac{1}{m \cdot \prod_{t=1}^k Z_t} \cdot \exp(-y_i \cdot \sum_{t=1}^k \alpha_t \cdot h_t(x_i))$,⁵³⁴ care a fost demonstrată la problema 23.a, putem deduce $D_{k+1}(i) \propto \exp(-y_i f_k(x_i))$.⁵³⁵

Răspuns:

a. Vom demonstra egalitatea cerută pornind de la termenul din partea dreaptă:

$$\begin{aligned} y_i f_k(x_i) &= y_i \sum_{t=1}^k \bar{\alpha}_t h_t(x_i) = \sum_{t=1}^k \bar{\alpha}_t y_i h_t(x_i) = \sum_{t:h_t(x_i)=y_i} \bar{\alpha}_t - \sum_{t:h_t(x_i)\neq y_i} \bar{\alpha}_t \\ &= Margin_k(x_i). \end{aligned} \quad (238)$$

b. Conform relației pe care tocmai am demonstrat-o la punctul precedent, inegalitatea $Margin_k(x_i) > Margin_k(x_j)$ este echivalentă cu $y_i f_k(x_i) > y_j f_k(x_j)$. La rândul ei, aceasta din urmă este echivalentă cu $-y_i f_k(x_i) < -y_j f_k(x_j)$, deci și cu $\exp(-y_i f_k(x_i)) < \exp(-y_j f_k(x_j))$. Înținând cont de Sugestia din enunț, rezultă imediat că $D_{k+1}(i) < D_{k+1}(j)$.

Așadar, se verifică *intuiția* conform căreia este natural ca instanțele care sunt [cel] mai bine clasificate, deci care au o margine [mai] mare la o anumită iterare, să primească la iterarea următoare o probabilitate mai mică, algoritmul AdaBoost concentrându-se asupra instanțelor care sunt incorrect (sau, mai puțin bine) clasificate.

Observație importantă: Se poate constata practic⁵³⁶ că în cursul execuției sale, algoritmul AdaBoost tinde să mărească per ansamblu, de la o iterare la alta, marginile corespunzătoare exemplelor de antrenament,⁵³⁷ iar o margine mai mare [pentru aceste exemple] implică în mod obișnuit o eroare la testare / generalizare mai mică. Aceasta este o explicație pentru faptul că, deși numărul de „parametri“ ai modelului rezultat crește cu 2 la fiecare iterare a algoritmului AdaBoost (și, deci, complexitatea crește),⁵³⁸ în general acest algoritm *nu produce overfitting*.⁵³⁹

28. (AdaBoost: o [altă] condiție suficientă pentru învățabilitate γ -slabă: media [probabilistă a] marginilor de votare să fie de cel puțin 2γ , la fiecare iterare a algoritmului)

■ □ • ○ *prelucrare de Liviu Ciortuz, după CMU, 2016 spring, W. Cohen, N. Balcan, HW4, pr. 3.1.4*

Introducere: La problema 23.e am făcut cunoștință cu noțiunea de *învățabilitate empirică γ -slabă* (engl., empirical γ -weak learnability). Concret, am demonstrat acolo că

⁵³⁴Atenție! Spre deosebire de definiția funcției f_k care a fost dată la punctul a, definiția funcției f_k de la problema 23 nu include normalizarea „voturilor“ α_t .

⁵³⁵Observați că o *sugestie* similară a fost făcută și la problema 26.a.

⁵³⁶Vedeți de exemplu problemele de tip implementare MIT, 2001 fall, Tommi Jaakkola, HW3, pr. 2.4 și MIT, 2009 fall, Tommi Jaakkola, HW3, pr. 1.4.

⁵³⁷Acest fapt se poate deduce din relația (234) de la problema 26, care dă expresia *costului* de minimizat de către algoritmul AdaBoost. Concret, a minimiza $E_T = \sum_{i=1}^m \exp(-y_i f_T(x_i))$ implică, în idee, a maximiza, pe cât posibil, fiecare *margine* $y_i f_T(x_i)$.

⁵³⁸Cei doi parametri adăugați la fiecare iterare a algoritmului AdaBoost sunt unul pentru identificarea ipotezei h_t (de exemplu, în cazul compașilor de decizie, pragul corespunzător, s_t), iar celălalt ponderea α_t .

⁵³⁹Există totuși situații în care algoritmul AdaBoost produce overfitting. Vedeți de exemplu pr. 72.

eroarea la antrenare produsă de algoritmul AdaBoost descrește [rapid] la 0 atunci când există $\gamma > 0$ astfel încât $\gamma \leq \gamma_t$ pentru orice t , unde $\gamma_t \stackrel{\text{def.}}{=} \frac{1}{2} - \varepsilon_t$, iar ε_t este eroarea ponderată la antrenare a ipotezei „slabe“ h_t .⁵⁴⁰ În acest exercițiu vom demonstra o nouă condiție suficientă pentru ca să aibă loc învățabilitatea empirică γ -slabă, făcând apel la noțiunea de *magine de votare* (engl., voting margin), care a fost prezentată la problema 27. Concret, vom arăta că atunci când există $\theta > 0$ astfel încât media [în sens probabilist] a marginilor de votare ale instanțelor de antrenament este mărginită inferior de θ la fiecare iterație a algoritmului AdaBoost,⁵⁴¹ proprietatea de învățabilitate empirică γ -slabă este „garantată“, cu $\gamma = \theta/2$.⁵⁴²

Presupunem că dispunem de setul de date de antrenament $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$ și că există ipotezele „slabe“ h_1, \dots, h_k din spațiul de ipoteze \mathcal{H} , precum și coeficienții pozitivi $\alpha_1, \dots, \alpha_k$ cu proprietatea $\sum_{j=1}^k \alpha_j = 1$. Presupunem, de asemenea, că există $\theta > 0$ astfel încât, pentru o distribuție probabilistă D (oarecare) peste setul de date de antrenament S are loc inegalitatea

$$E_{i \sim D} [\text{Margin}_k(x_i)] \geq \theta. \quad (239)$$

Vom demonstra că de fiecare dată când condiția (239) este satisfăcută există [cel puțin] o ipoteză $h_l \in \{h_1, \dots, h_k\}$ pentru care eroarea ponderată la antrenare în raport cu distribuția D este de cel mult $\frac{1}{2} - \frac{\theta}{2}$, adică $\varepsilon_l \leq \frac{1}{2} - \frac{\theta}{2}$.⁵⁴³ Demonstrația va fi făcută în doi pași, care corespund punctelor a și b de mai jos.

Observație:

Relația (239) implică faptul că există $i \in \{1, \dots, m\}$ cu proprietatea $\text{Margin}_k(x_i) \geq \theta$. Știm din enunțul problemei 27 că $\text{Margin}_k(x_i) \in [-1, 1]$, deci inegalitatea $\text{Margin}_k(x_i) \geq \theta > 0$ implică $\theta \in (0, 1]$ și, în consecință, $\frac{\theta}{2} \in (0, 1/2]$.

a. Fie D este o distribuție probabilistă oarecare definită pe S . Arătați că atunci când

$$E_{i \sim D} [\text{Margin}_k(x_i)] \stackrel{(238)}{=} E_{i \sim D} [y_i f_k(x_i)] \geq \theta,$$

adică media marginilor pentru instanțele x_1, \dots, x_m este mărginită inferior de θ , există cel puțin o ipoteză „slabă“ h_l din mulțimea $\{h_1, \dots, h_k\}$ astfel încât $E_{i \sim D} [y_i h_l(x_i)] \geq \theta$.⁵⁴⁴

b. Arătați că inegalitatea $E_{i \sim D} [y_i h_l(x_i)] \geq \theta$ este echivalentă cu inegalitatea $\text{err}_D(h_l) \leq \frac{1}{2} - \frac{\theta}{2}$, unde, ca și la problema 22, $\text{err}_D(h_l) \stackrel{\text{not.}}{=} \Pr_{i \sim D} [y_i \neq h_l(x_i)]$ desemnează eroarea ponderată la antrenare a ipotezei „slabe“ h_l în raport cu distribuția probabilistă D .

⁵⁴⁰Totuși, această condiție nu este satisfăcută întotdeauna. Vedeți, de exemplu, soluția a doua de la problema 25 (adică soluția care a fost obținută fără a folosi compași de decizie cu „prag“ exterior în raport cu instanțele de antrenament).

⁵⁴¹De fapt, este suficient ca această condiție să fie satisfăcută de la o anumită iterație t_0 încolo.

⁵⁴²În enunțul original al acestui exercițiu se formulează o altă condiție suficientă, care este mai „tare“ decât cea adoptată de noi: există $\theta > 0$ astfel încât marginile de votare ale instanțelor de antrenament sunt mărginite inferior de θ la fiecare iterație a algoritmului AdaBoost.

⁵⁴³În particular, D poate fi considerată ca fiind distribuția calculată de AdaBoost pentru iterarea k .

⁵⁴⁴Puteam vedea produsele $y_i h_l(x_i)$ (care iau valori în mulțimea $\{-1, +1\}$) ca fiind generate de către o variabilă aleatoare, cu distribuția D . Observați că $E_{i \sim D} [y_i h_l(x_i)]$ este media acestei variabile aleatoare.

Răspuns:

a. Tinând cont de relația (238), inegalitatea (239) se rescrie (conform definiției mediei) sub forma

$$\sum_{i=1}^m y_i f_k(x_i) \cdot D(i) \geq \theta. \quad (240)$$

Vom folosi *metoda reducerii la absurd* pentru a demonstra că există cel puțin un indice $l \in \{1, \dots, k\}$ astfel încât $E_{i \sim D}[y_i h_l(x_i)] \geq \theta$. Presupunem că pentru orice $l = 1, \dots, k$ are loc inegalitatea $E_{i \sim D}[y_i h_l(x_i)] < \theta$, adică $\sum_{i=1}^m y_i h_l(x_i) \cdot D(i) < \theta$. Prin înmulțirea ambilor termeni ai acestei inegalități cu $\alpha_l > 0$ rezultă

$$\sum_{i=1}^m y_i h_l(x_i) \cdot D(i) \cdot \alpha_l < \theta \cdot \alpha_l \text{ pentru } l = 1, \dots, k.$$

Însumând membru cu membru aceste inegalități pentru $l = 1, \dots, k$, rezultă

$$\begin{aligned} \sum_{l=1}^k \sum_{i=1}^m y_i h_l(x_i) \cdot D(i) \cdot \alpha_l &< \sum_{l=1}^k \theta \cdot \alpha_l \Leftrightarrow \sum_{i=1}^m y_i D(i) \left(\sum_{l=1}^k h_l(x_i) \alpha_l \right) < \theta \sum_{l=1}^k \alpha_l \Leftrightarrow \\ &\sum_{i=1}^m y_i f_k(x_i) \cdot D(i) < \theta, \end{aligned} \quad (241)$$

fiindcă $\sum_{l=1}^k \alpha_l = 1$ și $f_k(x_i) \stackrel{\text{not.}}{=} \sum_{l=1}^k \alpha_l h_l(x_i)$.

Inegalitatea (241) se rescrie sub forma $E_{i \sim D}[y_i f_k(x_i)] < \theta$. Evident, aceasta contrazice relația (240). Prin urmare, presupunerea făcută anterior este falsă. Rezultă că există $l \in \{1, \dots, k\}$ astfel încât $E_{i \sim D}[y_i h_l(x_i)] \geq \theta$.

b. După cum am menționat deja la rezolvarea punctului a, inegalitatea $E_{i \sim D}[y_i h_l(x_i)] \geq \theta$ se rescrie în mod echivalent sub forma $\sum_{i=1}^m y_i h_l(x_i) \cdot D(i) \geq \theta$. Tinând cont de faptul că $y_i \in \{-1, +1\}$ și $h_l(x_i) \in \{-1, +1\}$ pentru $i = 1, \dots, m$ și $l \in \{1, \dots, k\}$, putem scrie următorul sir de echivalențe:⁵⁴⁵

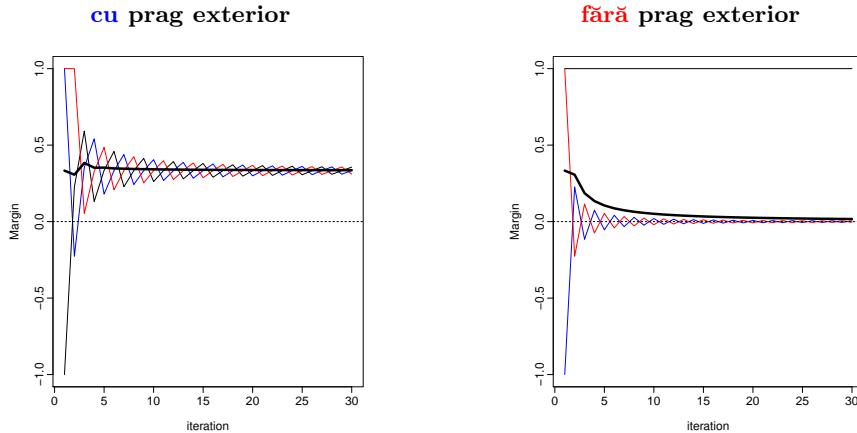
$$\begin{aligned} \sum_{i=1}^m y_i h_l(x_i) \cdot D(i) \geq \theta &\Leftrightarrow \sum_{i:y_i=h_l(x_i)} D(x_i) - \sum_{i:y_i \neq h_l(x_i)} D(x_i) \geq \theta \Leftrightarrow \\ (1 - \varepsilon_l) - \varepsilon_l &\geq \theta \Leftrightarrow 1 - 2\varepsilon_l \geq \theta \Leftrightarrow 2\varepsilon_l \leq 1 - \theta \Leftrightarrow \varepsilon_l \leq \frac{1}{2} - \frac{\theta}{2} \Leftrightarrow \\ &\stackrel{\text{def.}}{\Leftrightarrow} \text{err}_D(h_l) \leq \frac{1}{2} - \frac{\theta}{2}. \end{aligned}$$

În consecință, o condiție suficientă pentru a asigura γ -învățabilitate slabă pe un dataset de antrenament S este ca media [probabilistă a] marginilor de votare să fie de cel puțin $2\gamma \stackrel{\text{not.}}{=} \theta$, pentru orice exemplu de antrenament, la fiecare iterație a algoritmului AdaBoost (de fapt, este destul ca această condiție să fie îndeplinită de la o anumită iterație t_0 încolo).⁵⁴⁶

Exemplificare: Următoarele grafice, făcute de Sebastian Ciobanu, pun în evidență evoluția marginilor de votare, precum și a mediei lor (vedeți linia mai îngroșată), pentru datele de la problema 25, pentru fiecare dintre cele două variante de rezolvare care au fost prezentate acolo:

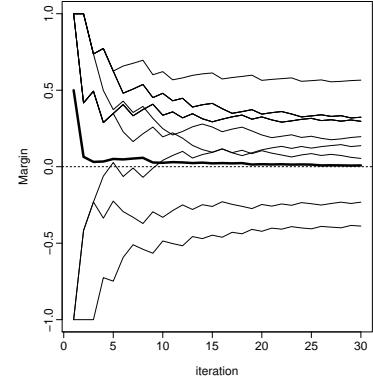
⁵⁴⁵Ca și la problema 22, vom nota cu ε_l eroarea ponderată produsă la antrenare de către ipoteza h_l .

⁵⁴⁶Observați că această condiție suficientă este mai „tare” / restrictivă decât condiția pentru învățabilitate γ -slabă de la problema 23.e, fiindcă prezenta condiție o implică pe cea de la problema 23.e.



Se constată ușor că în prima variantă (și anume, când se folosește prag exterior) există un prag inferior θ pentru media marginilor de votare (și chiar pentru toate marginile), începând chiar cu una din primele iterații. În schimb, în a doua variantă (și anume, când nu se folosește prag exterior), este evident că nu va exista un prag inferior θ pentru media marginilor de votare, întrucât la fiecare ierație una dintre instanțele $x_1 = 1$ și $x_3 = 5$ este clasificată eronat în mod flagrant. (Se observă însă că instanța $x_2 = 3$ este clasificată corect la fiecare ierație, ba chiar cu „maximum“ de voturi, de aceea marginea ei de votare este întotdeauna 1. Probabilitatea asociată acestei instanțe este însă din ce în ce mai mică.)

Graficul alăturat, făcut tot de Sebastian Ciobanu, prezintă evoluția marginilor de votare, precum și a mediei lor (vedeți linia mai îngroșată), pentru datele de la problema 66, pentru care se poate arăta că nu admite [nici măcar direct, adică în maniera de la problema 23.e] garanție de învățabilitate γ -slabă.



Observații:

1. Așa cum am procedat la problema 23.e, putem formula și la această problemă o *condiție suficientă* pentru ca eroarea la antrenare $H_{T'}$, ipoteza combinată produsă de AdaBoost la ierația T' , să rămână 0, după ce s-a constatat că există $\theta > 0$ astfel încât timp de un anumit număr de ierații (T) este satisfăcută inegalitatea $E_{i \sim D_k} [\text{Margin}_k(x_i)] \geq \theta$. Concret, se poate arăta ușor că $T = \left\lceil \frac{2}{\theta^2} \ln m \right\rceil$.

2. În afară de proprietatea de mai sus (1.), se poate constata că proprietatea de bază pe care am demonstrat-o în acest exercițiu

$$E_{i \sim D} [\text{Margin}_k(x_i)] \geq \theta \Rightarrow \exists h_t : \text{err}_D(h_t) \leq \frac{1}{2} - \frac{\theta}{2} \quad (242)$$

poate avea utilitate practică doar în cazul în care în algoritmul AdaBoost nu

impunem condiția să fie aleasă la fiecare iterație cea mai bună ipoteză „slabă“ (în sensul minimizării erorii ponderate la antrenare). În această situație, ipoteza „slabă“ h_t , a cărei existență este stipulată în relația (242), poate fi înlocuită cu ipoteza h'_t dacă, bineînțeles, $\text{err}_{D_t}(h'_t) < \text{err}_{D_t}(h_t)$.

3. Chiar dacă această *condiție suficientă* — există $\theta > 0$ astfel încât media $E_{i \sim D_t}[\text{Margin}_t(x_i)]$ este mai mare sau egală cu θ pentru orice t de la un t_0 încolo timp de T de iterării — nu este satisfăcută, totuși studiul evoluției marginilor de votare are o mare valoare informativă.⁵⁴⁷

4. La capitolul *Mașini cu vectori-suport* (SVM) se arată ca pentru o altă categorie de margini — și anume, *margini geometrice* —, se poate formula în mod riguros o *problemă de optimizare* care se pretează (inclusiv) la tratarea *excepțiilor* de la clasificare. În urma rezolvării acelei probleme de optimizare se obține un *clasificator / separator* care este foarte convenabil, din multe puncte de vedere.

29.

(O generalizare a algoritmului AdaBoost,
pentru diverse funcții de pierdere / cost)

*prelucrare de Liviu Ciortuz, după
■ • ○ MIT, 2003 fall, Tommi Jaakkola, HW4, pr. 2.1-3*

A. [Comentariu — fără „cerințe de rezolvat“]

În această problemă vom deriva un *algoritm de boosting, mai general* decât algoritmul AdaBoost din problema 22.⁵⁴⁸ Noul algoritm va putea fi aplicat la o întreagă clasă de *funcții de cost / pierdere* (engl., loss functions), în particular pentru funcția de cost [negativ] exponențială.

Scopul nostru este să generăm *funcții de discriminare* (engl., discriminant functions) de forma următoare:

$$f_T(x) = \alpha_1 h(x; \theta_1) + \dots + \alpha_T h(x; \theta_T).$$

În această expresie, $x \in \mathbb{R}^d$, iar θ_i este un parametru sau set / vector de parametri (tot dintr-un spațiu de tip \mathbb{R}^n), depinzând de tipul ipotezelor „slabe“ h . În continuare veți putea presupune că aceste ipoteze $h(x; \theta_i)$ sunt compași de decizie (engl., decision stumps), ale căror predicții pot fi $+1$ sau -1 . Orice alte categorii de ipoteze „slabe“ vor putea fi folosite în acest cadru. Vom adăuga în mod *secvențial* componente la funcția generală de discriminare, într-o manieră care va separa (atât cât este posibil) *estimarea* parametrilor θ ai ipotezelor „slabe“ de *setarea* voturilor / ponderilor α .

Explicație: Vom începe prin a defini noțiunea de *funcție de cost / pierdere*, pe care o vom folosi mai jos.⁵⁴⁹ Singurele *restrictii* pe care o astfel de funcție va trebui să le îndeplinească sunt următoarele: *i.* să fie nenegativă, *ii.* să fie *monoton descrescătoare* și *iii.* să fie *derivabilă*.⁵⁵⁰ În contextul nostru, argumentul unei funcții de pierdere va

⁵⁴⁷Vedeți de exemplu problemele de tip implementare MIT, 2001 fall, Tommi Jaakkola, HW3, pr. 1.4 și MIT, 2009 fall, Tommi Jaakkola, HW3, pr. 2.4.

⁵⁴⁸Sămânța pentru acest nou algoritm a fost sădită la *Observația importantă* de la finalul rezolvării problemei 26.

⁵⁴⁹Veți vedea că definiția pe care o dăm aici pentru funcții de cost / pierdere diferă ușor de cea pe care am folosit-o la problema 88 de la capitolul de *Fundamente*.

⁵⁵⁰Într-o accepție mai largă, condiția *ii.* din definiția funcției de pierdere / cost devine: să fie *convexă*. (Această accepție va fi folosită la punctul *c* al acestei probleme.)

fi *marginea algebrică* $y_i f_T(x_i)$, care reprezintă o măsură a acordului dintre eticheta y_i asociată instanței de antrenament x_i pe de o parte și [semnul dat de] valoarea funcției de discriminare f_T pe de altă parte. Vrem ca, pentru fiecare instanță x_i , „pierdere“ să fie cu atât mai mică cu cât valoarea funcției de discriminare este mai mult în acord cu valoarea ± 1 a etichetei y_i (adică, cu cât marginea algebrică $y_i f_T(x_i)$ are o valoare mai mare). Funcția de pierdere [negativ] *exponențială* pe care am întâlnit-o deja la problema 26, adică

$$\text{Loss}(y_i f_T(x_i)) = \exp(-y_i f_T(x_i))$$

satisfac cele două restricții pe care tocmai le-am menționat mai sus.⁵⁵¹

Criteriul de evaluare / „estimare“ pentru combinația de ipoteze „slabe“ este pur și simplu costul / pierderea empirică (engl., empirical loss), definită prin expresia

$$J_T = \frac{1}{m} \sum_{i=1}^m \text{Loss}(y_i f_T(x_i)), \quad (243)$$

unde sumarea se face parcurgând toate exemplele de antrenament disponibile.

Așadar, vrem să *derivăm algoritmul de boosting* în aşa fel încât să putem lucra cu orice funcție de pierdere de tipul discutat mai sus. În acest scop, la *iterația curentă* (t) presupunem că am inclus deja în combinația liniară un număr de $t - 1$ componente (ipoteze „slabe“):

$$f_{t-1}(x) = \alpha_1 h(x; \hat{\theta}_1) + \dots + \alpha_{t-1} h(x; \hat{\theta}_{t-1}) \quad (244)$$

și că dorim să mai adăugăm încă una, $h(x; \theta)$. *Criteriul de evaluare* pentru funcția de discriminare — inclusiv ultima componentă adăugată, al cărei „vot“ asociat este α —, este dat de expresia

$$J_t(\alpha, \theta) = \frac{1}{m} \sum_{i=1}^m \text{Loss}(y_i f_{t-1}(x_i) + y_i \alpha h(x_i; \theta)).$$

Remarcați faptul că urmărim să explicăm / explorăm doar modul în care acest „obiectiv“ / criteriu depinde de alegerea ultimei componente ($h(x_i; \theta)$) și de votul corespunzător (α), fiindcă parametrii precedentelor $t - 1$ componente, împreună cu voturile asociate lor au fost deja setate și nu vor mai fi modificate.

B. Pentru elaborarea ciclului repetitiv al noului algoritm, *mai întâi* vom urmări să identificăm *noua componentă*, mai precis spus valoarea parametrului θ în aşa fel încât să maximizăm „potențialul“ ei (adică, al noii componente) de a reduce pierderea empirică — „potențial“ în sensul că ulterior vom putea să ajustăm „votul“ asociat acestei componente, ca să reducem [și mai mult] pierderea empirică. Concret, la *Pasul 1* al iterației t vom seta θ astfel încât să *minimizăm* — în raport cu diferențele valori ale parametrului θ — valoarea expresiei următoare:⁵⁵²

$$\frac{\partial}{\partial \alpha} J_t(\alpha, \theta)|_{\alpha=0} = \frac{1}{m} \sum_{i=1}^m \frac{\partial}{\partial \alpha} \text{Loss}(y_i f_{t-1}(x_i) + y_i \alpha h(x_i; \theta))|_{\alpha=0}$$

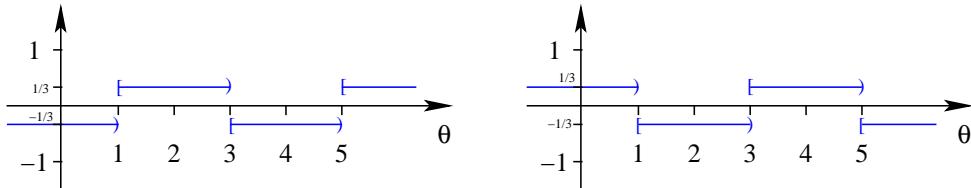
⁵⁵¹La problema 88 de la capitolul *Fundamente* am ilustrat / reprezentat pe un același grafic trei funcții de cost: funcția de cost [negativ] exponențială pentru AdaBoost, funcția de cost logistică pentru regresia logistică și funcția de cost *hinge* pentru mașini cu vectori-suport (SVM).

⁵⁵²Pentru o *justificare informală* [de ce se procedează așa], puteți face analogie cu graficul de la soluția problemei 26.c. Faceți corelația între pantele tangentelor la curbele din grafic pentru valoarea $\alpha = 0$ pe de o parte și valorile minime ale funcțiilor de cost (J_t) pe de altă parte.

$$\begin{aligned}
&= \frac{1}{m} \sum_{i=1}^m [dL(y_i f_{t-1}(x_i) + y_i \alpha h(x_i; \theta)) \cdot y_i h(x_i; \theta)]_{|\alpha=0} \\
&= \frac{1}{m} \sum_{i=1}^m dL(y_i f_{t-1}(x_i)) y_i h(x_i; \theta),
\end{aligned} \tag{245}$$

unde prin $dL(z)$ am notat derivata $\frac{\partial \text{Loss}(z)}{\partial z}$. (Am aplicat formula de derivare a funcțiilor compuse.)

Exemplificare: Graficele următoare, corespunzătoare datelor de la problema 25, reprezintă cele două variante ale expresiei / funcției (de argumentul θ) $\frac{\partial}{\partial \alpha} J_t(\alpha, \theta)_{|\alpha=0}$ calculate pentru iterată $t = 1$ a algoritmului AdaBoost folosind funcția de cost [negativ] exponentială. Aceste două variante corespund semnelor care pot fi alese pentru zonele de decizie determinate de compasul de decizie de prag (generic) θ : $+|-$ și respectiv $-|+$.



Remarcați faptul că expresia $\frac{\partial}{\partial \alpha} J_t(\alpha, \theta)_{|\alpha=0}$ exprimă în mod precis cât de „repede“ (în funcție de θ) vom putea începe să reducem pierderea empirică dacă vom crește în mod gradual „votul“ (α) pentru noua componentă, $h(x; \theta)$. Minimizarea expresiei (245) în raport cu θ pare a fi o modalitate rezonabilă de „estimare“ a parametrului θ al noii componente, $h(x; \theta)$. Acest plan ne permite ca mai întâi să setăm θ la o anumită valoare ($\hat{\theta}_t$) și apoi, la *Pasul 2* al iterării t , să alegem valoarea lui α în aşa fel încât să minimizăm pierderea empirică. (Concret, votul α_t va fi determinat egalând cu 0 derivata parțială a lui $J_t(\alpha, \hat{\theta}_t)$ în raport cu α .)

Acum vom modifica ușor algoritmul schițat mai sus în aşa fel încât să semene mai mult cu un *algoritm de boosting* (pregătind elaborarea *Pasului 3* al iterării curente).⁵⁵³ Vom defini următoarele *ponderi* (engl., weights) și *ponderi normalize* asociate exemplelor de antrenament:⁵⁵⁴

$$\begin{aligned}
W_i^{(t)} &= -dL(y_i f_{t-1}(x_i)), \text{ pentru } i = 1, \dots, m \text{ și} \\
\tilde{W}_i^{(t)} &= \frac{W_i^{(t)}}{\sum_{j=1}^m W_j^{(t)}}, \text{ pentru } i = 1, \dots, m.
\end{aligned} \tag{246}$$

ACESTE PONDERI SUNT, ÎN MOD EVIDENT, NENEGATIVE, ÎNTRUCÂT FUNCȚIA DE COST / PIERDERE ESTE DESCRESCĂTOARE ȘI DERIVABILĂ (DECI DERIVATA EI TREBUIE SĂ FIE NEGATIVĂ SAU ZERO). *Intuiția* este că ponderea unei instanțe x_i este cu atât

⁵⁵³Termenul de *boosting* se referă la faptul că instanțele incorect clasificate la o iterată (de către ipoteza „slabă“ aleasă la acea iterată) vor avea la iterată următoare probabilități mai mari, pentru ca algoritmul să se concentreze cu preponderență asupra lor.

⁵⁵⁴Ponderile normalize $\tilde{W}_i^{(t)}$ de aici le corespund în pseudo-codul algoritmului AdaBoost dat la problema 22 probabilitățile $D_t(i)$. Este îndreptățit să folosim pentru $\tilde{W}_i^{(t)}$ și denumirea de probabilități (*atenție!*, pentru iterată t), fiindcă $\tilde{W}_i^{(t)} \in [0, 1]$ și $\sum_{i=1}^m \tilde{W}_i^{(t)} = 1$.

mai mică (respectiv mai mare) cu cât marginea algebraică $y_i f_{t-1}(x_i)$ este mai mare (respectiv mai mică), adică cu cât x_i este mai bine (respectiv mai prost) clasificat.

a. Arătați că a minimiza (în funcție de θ) expresia $\frac{\partial}{\partial \alpha} J_t(\alpha, \theta) |_{\alpha=0}$ revine la a identifica ipoteza „slabă“ care are cea mai bună eroare ponderată la antrenare, ε_t , unde

$$\varepsilon_t \stackrel{not.}{=} \sum_{i: y_i \neq h(x_i; \hat{\theta}_t)} \tilde{W}_i^{(t)} y_i h(x_i; \theta) \stackrel{calcul}{=} \frac{1}{2} \left(1 - \sum_{i=1}^m \tilde{W}_i^{(t)} y_i h(x_i; \theta) \right). \quad (247)$$

Observație: A minimiza valoarea expresiei $\frac{1}{2} \left(1 - \sum_{i=1}^m \tilde{W}_i^{(t)} y_i h(x_i; \theta) \right)$ revine la a minimiza $-\sum_{i=1}^m \tilde{W}_i^{(t)} y_i h(x_i; \theta)$.

C. Acum putem începe să formulăm pașii corpului iterativ al noului algoritm de boosting într-un pseudo-cod similar cu algoritmul AdaBoost care a fost prezentat la problema 22:⁵⁵⁵

Pasul 1: Identifică o ipoteză / un clasificator $h(x; \hat{\theta}_t)$ care are o eroare ponderată la antrenare (engl., weighted training error) ε_t mai bună decât alegerea aleatorie, adică $\varepsilon_t < 1/2$, unde:

$$\varepsilon_t \stackrel{not.}{=} \sum_{i: y_i \neq h(x_i; \hat{\theta}_t)} \tilde{W}_i^{(t)} y_i h(x_i; \hat{\theta}_t). \quad (248)$$

Pasul 2: Setează „votul“ α_t pentru noua componentă, minimizând pierderea empirică totală (engl., overall empirical loss):

$$J_t(\alpha, \hat{\theta}_t) = \frac{1}{m} \sum_{i=1}^m \text{Loss}(y_i f_{t-1}(x_i) + y_i \alpha h(x_i; \hat{\theta}_t)),$$

adică

$$\alpha_t = \arg \min_{\alpha > 0} J_t(\alpha, \hat{\theta}_t).$$

Se consideră că $f_0(x_i) = 0$ pentru $i = 1, \dots, m$.

Pasul 3: Recalculează ponderile normalizate pentru următoarea iterație, astfel:

$$\tilde{W}_i^{(t+1)} = -c_t \cdot dL(y_i f_{t-1}(x_i) + y_i \alpha_t h(x_i; \hat{\theta}_t)) \text{ pentru } i = 1, \dots, m,$$

unde constanta c_t este aleasă astfel încât $\sum_{i=1}^m \tilde{W}_i^{(t+1)} = 1$.

b. Arătați că, atunci când funcția de pierdere este cea [negativ] exponențială, adică $\text{Loss}(z) = \exp(-z)$, cei trei pași din noul algoritm sunt în corespondență

⁵⁵⁵Pentru partea de inițializare, ca și în cazul algoritmului AdaBoost, vom folosi $\tilde{W}_i^{(1)} = \frac{1}{m}$, pentru $i = 1, \dots, m$. După aceea, corpul iterativ al algoritmului, format din pașii 1, 2 și 3, se execută pentru $t = 1, \dots, T$. Ca și în cazul algoritmului AdaBoost de la problema 22, algoritmul generalizat va fi oprit dacă la o iterație oarecare t nu există nicio ipoteză $h(x; \theta)$ având eroarea ponderată la antrenare $\varepsilon_t \in (0, 1/2)$.

directă cu pașii algoritmului AdaBoost (de la problema 22). Mai concret, arătați că în acest caz setarea lui α_t (la Pasul 2) bazat pe noua ipoteză „slabă“ $h(x; \hat{\theta}_t)$, precum și actualizarea ponderii $\tilde{W}_i^{(t+1)}$ (la Pasul 3) conduc la un rezultat identic cu AdaBoost. (După cum am precizat deja, în problema 22 corespondentul lui $\tilde{W}_i^{(t+1)}$ este $D_{t+1}(i)$.)

MIT, 2003 fall, Tommi Jaakkola, HW4, pr. 2.3

c. Arătați că, pentru orice funcție de pierdere validă de tipul discutat mai sus, componenta $h(x; \hat{\theta}_t)$, care a fost adăugată la iterată t , are eroarea ponderată la antrenare în raport cu ponderile actualizate $\tilde{W}_i^{(t+1)}$ exact $1/2$.⁵⁵⁶

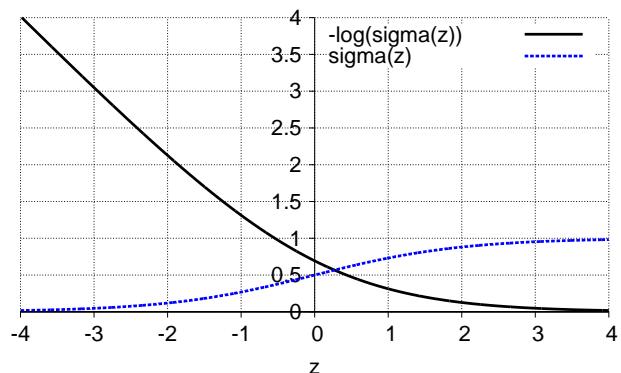
CMU, 2008 fall, Eric Xing, HW3, pr. 4.1.2

CMU, 2008 fall, Eric Xing, midterm, pr. 5.1

d. Să presupunem acum că schimbăm funcția obiectiv J_t în *media pătratelor erorilor*, adică $J_t = \frac{1}{m} \sum_{i=1}^m (y_i - f_t(x_i))^2$ sau, alternativ, $\frac{1}{m} \sum_{i=1}^m [(y_i - f_t(x_i))^2 \cdot 1_{\{y_i f_t(x_i) < 1\}}]$, ca să fie satisfăcută condiția de monotonie a funcției de cost / pierdere Loss pe întreaga axă reală. Urmărim să optimizăm costul J_t , ca și mai înainte, în mod secvențial.⁵⁵⁷ Care este noua regulă de actualizare (engl., update rule) pentru α_t ?

D. *Funcția de cost / pierdere logistică* se definește astfel:

$$\begin{aligned} \text{Loss}(y_i f_T(x_i)) \\ = \ln(1 + \exp(-y_i f_T(x_i))). \end{aligned}$$



Observație (legătura cu regresia logistică): Funcția de cost / pierdere logistică are o interpretare interesantă, și anume [valoarea ei este] negativul unei log-probabilități. Într-adevăr, [vă reamintim că] pentru un model aditiv de *regresie logistică*,⁵⁵⁸ avem

$$-\ln P(y = 1|x, w) = -\ln \frac{1}{1 + \exp(-z)} = \ln(1 + \exp(-z)),$$

unde $z = w_1 \phi_1(x) + \dots + w_T \phi_T(x)$.⁵⁵⁹ Înlocuind combinația aditivă de *funcții de bază* ($\phi_i(x)$) cu combinația aditivă de ipoteze „slabe“ ($h(x; \theta_i)$), vom obține un *model aditiv de regresie logistică*, în care ipotezele „slabe“ operează ca funcții de bază. Diferența este că în

⁵⁵⁶Acest nou rezultat generalizează proprietatea care a fost demonstrată la problema 22.vi.

⁵⁵⁷Observați că $(y_i - f(x_i))^2 = [y_i(1 - y_i f(x_i))]^2 = (1 - y_i f(x_i))^2 = (1 - z_i)^2$, unde am notat $z_i = y_i f(x_i)$. Funcția $(1 - z)^2$ este derivabilă și convexă; este descrescătoare pe intervalul $(-\infty, 1]$ și crescătoare pe intervalul $[1, +\infty)$. Varianta alternativă, adică funcția $(1 - z)^2 \cdot 1_{\{z < 1\}}$, este descrescătoare pe toată axa reală.

Derivata funcției $(1 - z)^2$ este $-2(1 - z)$, iar derivata funcției $(1 - z)^2 \cdot 1_{\{z < 1\}}$ este $-2(1 - z) \cdot 1_{\{z < 1\}}$.

⁵⁵⁸Pentru o introducere succintă în chestiunea regresiei logistice, vedeti problema 13 de la capitolul *Metode de regresie*.

⁵⁵⁹Mentionăm că am omis termenul liber (engl., bias) w_0 — care apare la regresia logistică — din motive care țin de simplitate.

cazul algoritmului AdaBoost [generalizat] vom „estima“ atât [parametrii pentru] funcțiile de bază (ipotezele „slabe“) cât și coeficienții cu care acestea vor fi înmulțite. În modelul [clasic] de regresie logistică avem de-a face în mod tipic cu un set fixat de funcții de bază.

*MIT, 2006 fall, Tommi Jaakkola, HW4, pr. 3.a
MIT, 2009 fall, Tommi Jaakkola, HW3, pr. 2.1*

e. Arătați că, atunci când folosim funcția de *pierdere logistică* [în locul celei [negativ] exponentiale], ponderile nenormalizate $W_i^{(t+1)}$ au valori mărginite superior de 1 (adică, $W_i^{(t+1)} < 1$). (Sugestie: Veți exprima aceste ponderi în funcție de $y_i f_t(x_i)$.)

*MIT, 2003 fall, Tommi Jaakkola, HW4, pr. 2.2
MIT, 2011 fall, L. P. Kaelbling, HW5, pr. 1.1*

f. În cazul funcției de pierdere *logistice*, care este valoarea / expresia ponderilor normalize, $\tilde{W}_i^{(t+1)}$?

Ce puteți spune despre ponderile normalize pentru exemplele care sunt clasificate eronat în mod flagrant, comparativ cu cele care sunt clasificate ușor eronat de către combinația curentă („ansamblul“ curent), $f_T(x) \stackrel{\text{not.}}{=} \sum_{t=1}^T \alpha_t h(x; \hat{\theta}_t)$? Dacă datele de antrenament conțin exemple care au fost etichetate în mod greșit (engl., mislabeled), de ce credeți că se preferă funcția de pierdere logistică în locul celei [negativ] exponentiale, $\text{Loss}(z) = \exp(-z)$?

Liviu Ciortuz, 2020

g. Presupunem că lucrăm cu funcția de *pierdere logistică*. Care este în acest caz regula de actualizare pentru α_t ?

Indicație: În loc să minimizați J_t în raport cu argumentul α (considerând ipoteza h_t deja aleasă / fixată), veți urmări să minimizați în raport cu α o *magine superioară* pentru *diferența* dintre J_t și J_{t-1}^* , unde, aşa cum am definit la problema 26, $J_{t-1}^* \stackrel{\text{not.}}{=} \min_{h \in \mathcal{H}, \alpha' \in \mathbb{R}_+} J_{t-1}(h, \alpha')$.⁵⁶⁰

*MIT, 2006 fall, Tommi Jaakkola, HW4, pr. 3.b
MIT, 2009 fall, Tommi Jaakkola, HW3, pr. 2.2*

h. Presupunem din nou că lucrăm cu funcția de *pierdere logistică*. Considerăm un set de date de antrenament liniar-separabil. Am dori să utilizăm o mașină cu vectori-suport liniară cu margine “hard” — adică, fără variabile / penalizări ecart (engl., slack penalties) — pe post de clasificator „slab“.⁵⁶¹ Presupunând că la *Pasul 1* al algoritmului AdaBoost generalizat se minimizează eroarea ponderată la antrenare ε_t , care va fi valoarea ponderii α_1 la prima iterare de boosting?

Răspuns:

a. Pornind de la relația (245), o vom putea scrie — folosind notația (246) — sub forma următoare:

$$\frac{\partial}{\partial \alpha} J_t(\alpha, \theta) |_{\alpha=0} = -\frac{1}{m} \sum_{i=1}^m W_i^{(t)} y_i h(x_i; \theta)$$

⁵⁶⁰Spre deosebire de J_t care depinde de α , J_{t-1}^* este constant, deci nu depinde de α . Prin urmare, a minimiza J_t în raport cu α este echivalent cu a minimiza $J_t - J_{t-1}^*$ în raport cu α . La finalul problemei 26.e am arătat că $J_t^* \geq J_{t-1}^*$ și $J_t(h, 0) = J_{t-1}^*$.

⁵⁶¹LC: În locul acestui tip de SVM putem considera orice alt separator liniar consistent cu astfel de date de antrenament (adică, liniar separabile).

$$\begin{aligned}
&= -\frac{1}{m} \left(\sum_j W_j^{(t)} \right) \cdot \sum_{i=1}^m \frac{W_i^{(t)}}{\sum_j W_j^{(t)}} y_i h(x_i; \theta) \\
&= -\frac{1}{m} \left(\sum_j W_j^{(t)} \right) \cdot \sum_{i=1}^m \tilde{W}_i^{(t)} y_i h(x_i; \theta).
\end{aligned}$$

Întrucât factorul $\frac{1}{m} \sum_j W_j^{(t)}$ este pozitiv și constant în raport cu θ , rezultă că a minimiza valoarea expresiei $\frac{\partial}{\partial \alpha} J_t(\alpha, \theta)|_{\alpha=0}$ este echivalent cu a minimiza expresia

$$-\sum_{i=1}^m \tilde{W}_i^{(t)} y_i h(x_i; \theta),$$

ceea ce, conform realției (247), este echivalent cu a minimiza eroarea ponderată la antrenare ε_t .

Rămâne doar să demonstrăm egalitatea a doua din realția (247). Într-adevăr, notând — ca la problema 22 — multimile de indici $C = \{i = 1, \dots, m \mid y_i = h(x_i; \hat{\theta}_t)\}$ și $M = \{i = 1, \dots, m \mid y_i \neq h(x_i; \hat{\theta}_t)\}$, este imediat că

$$\begin{aligned}
\sum_{i=1}^m \tilde{W}_i^{(t)} y_i h(x_i; \hat{\theta}_t) &= \sum_{i \in C} \tilde{W}_i^{(t)} y_i h(x_i; \hat{\theta}_t) + \sum_{i \in M} \tilde{W}_i^{(t)} y_i h(x_i; \hat{\theta}_t) = \underbrace{\sum_{i \in C} \tilde{W}_i^{(t)}}_{1 - \varepsilon_t} - \underbrace{\sum_{i \in M} \tilde{W}_i^{(t)}}_{\varepsilon_t} \\
&= 1 - 2\varepsilon_t.
\end{aligned}$$

Așadar,

$$\varepsilon_t = \frac{1}{2} \left(1 - \sum_{i=1}^m \tilde{W}_i^{(t)} y_i h(x_i; \hat{\theta}_t) \right).$$

Observație: Întrucât $\tilde{W}_i^{(t)} \geq 0$ și $\sum_{i=1}^m \tilde{W}_i^{(t)} = 1$, iar $y_i, h(x_i; \theta) \in \{-1, +1\}$, rezultă că $\sum_{i=1}^m \tilde{W}_i^{(t)} y_i h(x_i; \theta) \in [-1, +1]$, deci $1 - \sum_{i=1}^m \tilde{W}_i^{(t)} y_i h(x_i; \hat{\theta}_t) \in [0, +2]$. Prin urmare, $\varepsilon_t = \frac{1}{2} \left(1 - \sum_{i=1}^m \tilde{W}_i^{(t)} y_i h(x_i; \hat{\theta}_t) \right) \in [0, +1]$.

b. Pentru prima parte, vom arăta că atunci când se folosește $\text{Loss}(z) = e^{-z}$, valoarea obținută pentru coeficientul α la minimizarea de la *Pasul 2* din algoritm general (vedeți partea stângă a egalității următoare) este aceeași cu valoarea lui α rezultată prin minimizarea efectuată de către algoritmul Ada-Boost (vedeți partea dreaptă a egalității următoare), adică

$$\arg \min_{\alpha > 0} \sum_{i=1}^m \text{Loss}(y_i f_{t-1}(x_i) + \alpha y_i h(x_i; \hat{\theta}_t)) = \arg \min_{\alpha > 0} \sum_{i=1}^m \tilde{W}_i^{(t)} \exp(-\alpha y_i h(x_i; \hat{\theta}_t)),$$

unde, conform asignării de la execuția *Pasului 3* al iterăției $t-1$,

$$\tilde{W}_i^{(t)} = c_{t-1} \cdot \exp(-y_i f_{t-1}(x_i)),$$

c_{t-1} fiind constanta de normalizare (astfel încât ponderile însumate să dea valoarea 1). Calculând expresia funcției obiectiv din partea stângă a egalității de mai sus, obținem:⁵⁶²

⁵⁶²LC: A se vedea similaritatea cu demonstrația de la problema 26.a.

$$\begin{aligned}
\sum_{i=1}^m \text{Loss}(y_i f_{t-1}(x_i) + \alpha y_i h(x_i; \hat{\theta}_t)) &= \sum_{i=1}^m \exp(-y_i f_{t-1}(x_i) - \alpha y_i h(x_i; \hat{\theta}_t)) \\
&= \sum_{i=1}^m \exp(-y_i f_{t-1}(x_i)) \exp(-\alpha y_i h(x_i; \hat{\theta}_t)) \\
&= \frac{1}{c_{t-1}} \sum_{i=1}^m \tilde{W}_i^{(t)} \exp(-\alpha y_i h(x_i; \hat{\theta}_t)) \\
&= \frac{1}{c_{t-1}} \left[\sum_{i \in \{i | y_i h(x_i; \hat{\theta}_t) = 1\}} \left(\tilde{W}_i^{(t)} \right) e^{-\alpha} + \sum_{i \in \{i | y_i h(x_i; \hat{\theta}_t) = -1\}} \left(\tilde{W}_i^{(t)} \right) e^{\alpha} \right],
\end{aligned}$$

cantitate care este proporțională cu funcția obiectiv minimizată de către algoritmul AdaBoost (vedeți de asemenea problema 26.a). Prin urmare, minimizarea în raport cu α conduce la același rezultat în ambii algoritmi.

Pentru partea a doua, observați faptul că la *Pasul 3* (de la iterată t) din algoritmul general asignarea ponderilor se face astfel:

$$\tilde{W}_i^{(t+1)} = -c_t \cdot dL(y_i f_t(x_i)) = c_t \cdot \exp(-y_i f_t(x_i)),$$

ceea ce coincide cu modul cum procedează algoritmul AdaBoost (vedeți problema 23.a).

c. La iterată t , coeficientul α_t este ales minimizând $J_t(\alpha, \hat{\theta}_t)$, ceea ce implică $\frac{\partial J_t(\alpha, \hat{\theta}_t)}{\partial \alpha} = 0$. Întrucât

$$\begin{aligned}
\frac{\partial}{\partial \alpha} J_t(\alpha, \hat{\theta}_t) &= \frac{1}{m} \sum_{i=1}^m \frac{\partial}{\partial \alpha} \text{Loss}(y_i f_{t-1}(x_i) + y_i \alpha h(x_i; \hat{\theta}_t)) \\
&= \frac{1}{m} \sum_{i=1}^m \underbrace{dL(y_i f_{t-1}(x_i) + y_i \alpha h(x_i; \hat{\theta}_t))}_{-\frac{\tilde{W}_i^{(t+1)}}{c_t}} y_i h(x_i; \hat{\theta}_t) \propto \sum_{i=1}^m \tilde{W}_i^{(t+1)} y_i h(x_i; \hat{\theta}_t),
\end{aligned}$$

rezultă că trebuie să avem

$$\sum_{i=1}^m \tilde{W}_i^{(t+1)} y_i h(x_i; \hat{\theta}_t) = 0.$$

După aceasta, eroarea ponderată la antrenare produsă de $h(x; \hat{\theta}_t)$ (corespunzător ponderilor actualizate $\tilde{W}_i^{(t+1)}$ determinate de α_t) poate fi calculată într-un mod similar cu relația (247):

$$\frac{1}{2} \left(1 - \underbrace{\sum_{i=1}^m \tilde{W}_i^{(t+1)} y_i h(x_i; \hat{\theta}_t)}_0 \right) = \frac{1}{2} (1 - 0) = \frac{1}{2}.$$

d. Vom lucra [mai întâi] cu $\text{Loss}(y_i f_t(x_i)) = (1 - y_i f_t(x_i))^2 = (y_i - f_t(x_i))^2$, deci vom considera $J_t = \frac{1}{m} \sum_{i=1}^m (y_i - f_t(x_i))^2$.⁵⁶³ Pentru a găsi valoarea lui α_t , vom

⁵⁶³Atenție! În acest caz, ponderile $\tilde{W}_i^{(t)}$ trebuie definite astfel: $W_i^{(t)} = |dL(y_i f_{t-1}(x_i))|$, întrucât derivata funcției de cost $(1 - z)^2$ nu este negativă pe toată axa reală.

calcula derivata lui J_t în raport cu α_t și apoi o vom egala cu zero.

$$\frac{\partial J_t}{\partial \alpha_t} = \frac{1}{m} \frac{\partial \sum_{i=1}^m (y_i - f_t(x_i))^2}{\partial \alpha_t} = \frac{1}{m} \sum_{i=1}^m 2(y_i - f_t(x_i)) \frac{\partial(y_i - f_t(x_i))}{\partial \alpha_t}.$$

Stim că⁵⁶⁴

$$f_t(x_i) = f_{t-1}(x_i) + \alpha_t h_t(x_i).$$

În această expresie, f_{t-1} nu depinde de α_t . Înlocuind în expresia derivatei, obținem

$$\frac{\partial J_t}{\partial \alpha_t} = \frac{2}{m} \sum_{i=1}^m (y_i - f_t(x_i)) \frac{\partial(y_i - f_{t-1}(x_i) - \alpha_t h_t(x_i))}{\partial \alpha_t} = \frac{2}{m} \sum_{i=1}^m (y_i - f_t(x_i))(-h_t(x_i)).$$

Egalând această derivată cu zero, vom avea:

$$\begin{aligned} \frac{\partial J_t}{\partial \alpha_t} = 0 &\Leftrightarrow \sum_{i=1}^m (y_i - f_t(x_i))h_t(x_i) = 0 \Leftrightarrow \sum_{i=1}^m (y_i - f_{t-1}(x_i) - \alpha_t h_t(x_i))h_t(x_i) = 0 \\ &\Leftrightarrow \sum_{i=1}^m (y_i - f_{t-1}(x_i))h_t(x_i) = \alpha_t \sum_{i=1}^m h_t^2(x_i) \Leftrightarrow \alpha_t = \frac{\sum_{i=1}^m (y_i - f_{t-1}(x_i))h_t(x_i)}{\sum_{i=1}^m h_t^2(x_i)} \\ &\Leftrightarrow \alpha_t = \frac{1}{m} \sum_{i=1}^m (y_i - f_{t-1}(x_i))h_t(x_i), \end{aligned}$$

cu condiția ca această valoare să fie strict pozitivă.⁵⁶⁵

Observație: Egalitatea $h_t^2(x_i) = 1$ (pe care am folosit-o mai sus) decurge din presupunerea că ipotezele „slabe“ sunt compași de decizie.

Dacă schimbăm funcția obiectiv în $J_t = \frac{1}{m} \sum_{i=1}^m [(y_i - f_t(x_i))^2 \cdot 1_{\{y_i f_t(x_i) < 1\}}]$, se verifică ușor că se obține următoarea expresie pentru votul α_t :

$$\alpha_t = \frac{1}{\sum_{i=1}^m 1_{\{y_i f_{t-1}(x_i) < 1\}}} \sum_{i=1}^m [(y_i - f_{t-1}(x_i))h_t(x_i) \cdot 1_{\{y_i f_{t-1}(x_i) < 1\}}],$$

cu condiția ca această valoare să fie strict pozitivă.

e. Ponderea $W_i^{(t+1)}$ a fost definită ca fiind $-dL(y_i f_t(x_i))$, unde $dL(z) \stackrel{not.}{=} \frac{\partial}{\partial z} (\ln(1 + e^{-z}))$. Așadar,

$$W_i^{(t+1)} = \frac{e^{-z_i}}{1 + e^{-z_i}} = \frac{1}{1 + e^{z_i}} < 1, \text{ unde } z_i = y_i f_t(x_i).$$

f. Ponderile normalize au fost definite astfel:

$$\tilde{W}_i^{(t+1)} = c_t \cdot \frac{\exp(-y_i f_t(x_i))}{1 + \exp(-y_i f_t(x_i))},$$

⁵⁶⁴Atât aici cât și în continuare, vom scrie $h_t(x_i)$ în loc de $h_t(x_i; \hat{\theta}_t)$ dacă se subînțelege din context că h_t a fost deja fixat.

⁵⁶⁵Verificarea condiției de minim este imediată. Dacă valoarea calculată aici pentru α_t este negativă, atunci $\min_{\alpha \geq 0} J_t$ se obține pentru $\alpha = 0$. În acest caz, se poate renunța la h_t și se poate alege în schimb o altă ipoteză „slabă“, cu următoarea cea mai bună eroare ponderată la antrenare.

unde constanta de normalizare este

$$c_t = \left(\sum_{i=1}^m \frac{\exp(-y_i f_t(x_i))}{1 + \exp(-y_i f_t(x_i))} \right)^{-1}.$$

Pentru exemplele de antrenament care sunt clasificate eronat în mod flagrant, $y_i f_t(x_i)$ are o valoare negativă mare în valoare absolută, aşadar $W_i^{(t+1)}$ va fi în acest caz aproape de [dar mai mic decât] 1. Pentru exemplele de antrenament care sunt clasificate ușor eronat, $W_i^{(t+1)}$ va fi aproape de [și mai mare decât] 1/2. Aşadar, ponderile normalize pentru aceste două cazuri se vor afla într-un raport de cel mult 2 : 1. Aceasta înseamnă că un singur outlier care este clasificat eronat în mod flagrant va avea întotdeauna o pondere care este de cel mult de două ori mai mare decât ponderea unei instanțe care este clasificată ușor eronat. Din acest motiv, algoritmul AdaBoost cu funcție de cost logistică este robust la outliers.

În ce privește situația exemplelor de antrenament care au fost greșit etichetate, putem raționa în felul următor. În cazul funcției de pierdere logistice, dacă avem un x_i care ar avea drept etichetă $y_i = +1$, dar se consideră (în mod eronat) $y_i = -1$, pierderea este de aproximativ $f_t(x_i)$ dacă $f_t(x_i) > 0$,⁵⁶⁶ în vreme ce în cazul funcției de pierdere [negativ] exponentiale pierderea este $\exp(f_t(x_i))$ care este în general mult mai mare decât $f_t(x_i)$. Cazurile simetrice ($f_t(x_i) \leq 0$ și apoi $y_i = -1 \rightarrow +1$) se tratează în mod similar.

g.⁵⁶⁷ Calculăm mai întâi valoarea expresiei $J_t - J_{t-1}^*$:

$$\begin{aligned} J_t - J_{t-1}^* &= \frac{1}{m} \sum_{i=1}^m \ln(1 + \exp(-y_i f_t(x_i))) - \sum_{i=1}^m \ln(1 + \exp(-y_i f_{t-1}(x_i))) \\ &= \frac{1}{m} \sum_{i=1}^m \ln \frac{1 + \exp(-y_i f_t(x_i))}{1 + \exp(-y_i f_{t-1}(x_i))} \\ &= \frac{1}{m} \sum_{i=1}^m \ln \frac{1 + \exp(-y_i f_{t-1}(x_i)) - \exp(-y_i f_{t-1}(x_i)) + \exp(-y_i f_t(x_i))}{1 + \exp(-y_i f_{t-1}(x_i))} \\ &= \frac{1}{m} \sum_{i=1}^m \ln \left(1 + \frac{\exp(-y_i f_t(x_i)) - \exp(-y_i f_{t-1}(x_i))}{1 + \exp(-y_i f_{t-1}(x_i))} \right) \\ &= \frac{1}{m} \sum_{i=1}^m \ln \left(1 + \frac{\exp(-y_i f_{t-1}(x_i)) - y_i \alpha h(x_i) - \exp(-y_i f_{t-1}(x_i))}{1 + \exp(-y_i f_{t-1}(x_i))} \right) \\ &= \frac{1}{m} \sum_{i=1}^m \ln \left(1 + \frac{\exp(-y_i f_{t-1}(x_i)) \exp(-y_i \alpha h(x_i)) - \exp(-y_i f_{t-1}(x_i))}{1 + \exp(-y_i f_{t-1}(x_i))} \right) \\ &= \frac{1}{m} \sum_{i=1}^m \ln \left(1 + \frac{\exp(-y_i f_{t-1}(x_i)) [\exp(-y_i \alpha h(x_i)) - 1]}{1 + \exp(-y_i f_{t-1}(x_i))} \right) \\ &= \frac{1}{m} \sum_{i=1}^m \ln \left(1 + \frac{\exp(-y_i \alpha h(x_i)) - 1}{\frac{1}{\exp(-y_i f_{t-1}(x_i))} + 1} \right) \end{aligned}$$

⁵⁶⁶Vedeți graficul funcției de *cost logistic* din enunțul problemei de față.

⁵⁶⁷Soluția care urmează a fost redactată initial de către studentul Ștefan Matcovici, cf. https://mitpress.mit.edu/sites/default/files/titles/content/boosting_foundations_algorithms/chapter007.html.

$$= \frac{1}{m} \sum_{i=1}^m \ln \left(1 + \frac{\exp(-y_i \alpha h(x_i)) - 1}{\exp(y_i f_{t-1}(x_i)) + 1} \right) \quad (249)$$

Remarcați faptul că logaritmul din expresia (249) există, întrucât

$$\begin{aligned} \frac{\exp(-y_i \alpha h(x_i)) - 1}{\exp(y_i f_{t-1}(x_i)) + 1} > -1 &\Leftrightarrow \exp(-y_i \alpha h(x_i)) - 1 > -\exp(y_i f_{t-1}(x_i)) - 1 \\ &\Leftrightarrow \underbrace{\exp(-y_i \alpha h(x_i))}_{>0} > \underbrace{-\exp(y_i f_{t-1}(x_i))}_{<0}, \\ &\text{inegalitate adevărată pentru } \forall \alpha. \end{aligned}$$

Mai departe, întrucât $\ln(1 + z) \leq z$ pentru orice $z > -1$,⁵⁶⁸ rezultă că

$$J_t - J_{t-1}^* \leq \frac{1}{m} \sum_{i=1}^m \frac{\exp(-y_i \alpha h(x_i)) - 1}{\exp(y_i f_{t-1}(x_i)) + 1} \quad (250)$$

Expresia din partea dreaptă a inegalității (250) este *marginea superioară* (engl., upper bound) pentru $J_t - J_{t-1}^*$ pe care o vom minimiza în raport cu α .

Stim din enunț că $W_i^{(t)} = -dL(y_i f_{t-1}(x_i))$. Un calcul simplu ne arată că

$$\frac{\partial}{\partial z} \ln(1 + e^{-z}) = \frac{-e^{-z}}{1 + e^{-z}} = -\frac{1}{\frac{1}{e^{-z}} + 1} = -\frac{1}{e^z + 1}. \quad (251)$$

Rezultă că

$$W_i^{(t)} = -\left(-\frac{1}{\exp(y_i f_{t-1}(x_i)) + 1} \right) = \frac{1}{\exp(y_i f_{t-1}(x_i)) + 1},$$

iar

$$\tilde{W}_i^{(t)} = c_{t-1} \cdot \frac{1}{\exp(y_i f_{t-1}(x_i)) + 1},$$

unde c_{t-1} este constanta de normalizare a ponderilor $W_i^{(t)}$.

Prin urmare,

$$\begin{aligned} J_t - J_{t-1}^* &\leq \frac{1}{m} \sum_{i=1}^m \frac{1}{\exp(y_i f_{t-1}(x_i)) + 1} (\exp(-y_i \alpha h(x_i)) - 1) \\ &= \frac{1}{m c_{t-1}} \sum_{i=1}^m \tilde{W}_i^{(t)} \cdot (\exp(-y_i \alpha h(x_i)) - 1) \\ &\propto \sum_{i=1}^m \tilde{W}_i^{(t)} \cdot \exp(-y_i \alpha h(x_i)) - \underbrace{\sum_{i=1}^m \tilde{W}_i^{(t)}}_1. \end{aligned}$$

Așadar, a minimiza marginea superioară pentru $J_t - J_{t-1}^*$ revine la a minimiza (în raport cu α) expresia $\sum_{i=1}^m \tilde{W}_i^{(t)} \cdot \exp(-y_i \alpha h(x_i))$, pe care am întâlnit-o și la optimizarea costului [negativ] exponențial (vedeți problema 26.a). Putem conchide că și în cazul funcției de cost logistice putem alege $\alpha_t = \frac{1}{2} \ln \frac{1 - \varepsilon_t}{\varepsilon_t}$.

⁵⁶⁸Această inegalitate este foarte ușor de verificat analitic.

Observație: Tinând cont de faptul că aici nu se minimizează J_t ci o margine superioară pentru $J_t - J_{t-1}^*$, rezultă că atunci când se lucrează astfel (adică se folosește $\alpha_t = \frac{1}{2} \ln \frac{1 - \varepsilon_t}{\varepsilon_t}$ dacă funcția de cost este cea logistică), nu neapărat se va respecta proprietatea care a fost enunțată și demonstrată la punctul c.⁵⁶⁹

h. Conform enunțului, setul de date de antrenament este liniar-separabil. O mașină cu vectori-suport cu margine “hard” va produce pe un astfel de set de date un separator (liniar) care are eroarea ponderată la antrenare 0. Ca și în cazul lui AdaBoost, algoritmul trebuie opriit.⁵⁷⁰

30.

(O margine superioară pentru eroarea la generalizare produsă de AdaBoost)

□ • Liviu Ciortuz, Andi Munteanu, după Princeton Univ., 2006 spring, COS 511, Rob Schapire, Lecture #11

În această problemă vom calcula o margine superioară pentru eroarea pe care o obține algoritmul AdaBoost la generalizare / testare, în funcție de eroarea produsă la antrenare pe un set de date oarecare S . Vom folosi următoarele notații:

$\mathcal{H} \stackrel{\text{not.}}{=} \text{spațiul ipotezelor „slabe“; vom presupune că } \mathcal{H} \text{ este finit } (|\mathcal{H}| < +\infty)$

$$co(\mathcal{H}) \stackrel{\text{def.}}{=} \{f(x) = \sum_j \alpha_j h_j(x) | \alpha_j \geq 0, \sum_j \alpha_j = 1, h_j \in \mathcal{H}\}$$

(această mulțime se numește *înfășurătoarea convexă* a lui \mathcal{H})

$$\mathcal{C}_N \stackrel{\text{def.}}{=} \{g(x) = \frac{1}{N} \sum_{j=1}^N h_j(x) | h_j \in \mathcal{H}\}, N \in \mathbb{N}^*$$

(evidenț, \mathcal{C}_N constituie o submulțime al lui $co(\mathcal{H})$, $\forall N \in \mathbb{N}^*$)

$D \stackrel{\text{not.}}{=} \text{o distribuție de probabilitate peste } X \times \{-1, +1\}$

$S \stackrel{\text{not.}}{=} \text{setul de exemple de antrenament}$

$m \stackrel{\text{not.}}{=} \text{numărul de exemple de antrenament}$

$P_D(\cdot) \stackrel{\text{not.}}{=} \text{probabilitatea unui set de exemple } (x, y) \sim D$

⁵⁶⁹Mulțumesc studentului Bogdan Palanici pentru discuția care a dus la formularea acestei observații.

⁵⁷⁰Dacă am gândi în mod miop, am judeca astfel:

La Pasul 1, alegem $\hat{\theta}_1$, astfel încât să minimizăm $\frac{\partial J_t(\alpha, \theta)}{\partial \alpha}|_{\alpha=0}$. Echivalent, putem gândi că acest $\hat{\theta}_1$ este ales astfel încât să minimizăm suma ponderată $2\varepsilon_1 - 1 = -\sum_{i=1}^m \tilde{W}_i^{(1)} y_i h(x_i; \theta)$, unde $\tilde{W}_i^{(1)} = \frac{1}{m}$ pentru $i = 1, 2, \dots, m$. Setul de date de antrenament fiind liniar-separabil, mașina cu vectori-suport cu margine “hard” va produce un separator $h(\cdot; \hat{\theta}_1)$ care satisfacă inegalitatea $y_i h(x_i; \hat{\theta}_1) \geq 1$ pentru $i = 1, 2, \dots, m$.

La Pasul 2, trebuie să alegem α_1 astfel încât să minimizăm $J_t(\alpha_1, \hat{\theta}_1) = \frac{1}{m} \sum_{i=1}^m L(y_i h_0(x_i) + \alpha_1 y_i h(x_i; \hat{\theta}_1)) = \frac{1}{m} \sum_{i=1}^m L(\alpha_1 y_i h(x_i; \hat{\theta}_1))$. Remarcăți faptul că $J_t(\alpha_1, \hat{\theta}_1)$ este o sumă de termeni care sunt strict decrescători în raport cu α_1 (pentru că $y_i h(x_i; \hat{\theta}_1) \geq 1$). Așadar, și suma $J_t(\alpha_1, \hat{\theta}_1)$ este strict decrescătoare în raport cu α_1 . Aceasta ar implica faptul că algoritmul AdaBoost generalizat folosind funcție de pierdere logistică va lua $\alpha_1 = \infty$.

În general, dacă am avea un clasificator „de bază“ care separă în mod perfect datele de antrenament, îi vom asocia un vot oricărput de mare, ca să minimizăm pierderea / costul. Leția este simplă: atunci când facem boosting, trebuie să folosim clasificatori „de bază“ care nu sunt atât de puternici încât să separe în mod perfect datele de antrenament.

$P_S(\cdot) \stackrel{not.}{=} \text{probabilitatea unui set de exemple } (x, y) \text{ selectate în mod uniform din setul } S.$

În cele ce urmează, ne vom concentra atenția asupra unei combinații liniare de ipoteze „slabe“ $f(x) = \sum_{t=1}^T \alpha_t h_t(x)$, care a fost obținută de către algoritmul AdaBoost după ce s-au executat T iterații pe setul de exemple S . Vom considera că $\sum_{t=1}^T \alpha_t = 1$; aşadar, voturile α_t sunt presupuse a fi normalizate.

Observație: În demonstrația care urmează un rol esențial îl va avea următoarea proprietate interesantă: funcția $f \in co(\mathcal{H})$ poate fi aproximată cu ajutorul unor funcții din \mathcal{C}_N . Într-adevăr, putem considera funcțiile $g \in \mathcal{C}_N$ definite prin expresia $g(x) = \frac{1}{N} \sum_{j=1}^N g_j(x)$, unde fiecare componentă g_j este selectată în mod aleatoriu din mulțimea de ipoteze „slabe“ $\{h_1, \dots, h_N\}$, și anume: g_j este desemnată ca fiind ipoteza „slabă“ h_t cu probabilitatea α_t .⁵⁷¹ ⁵⁷² În aceste condiții, este imediat că $E_g[g_j(x)] \stackrel{\text{def}}{=} \sum_{t=1}^T \alpha_t h_t(x) = f(x)$.⁵⁷³

În vederea determinării unei *margini superioare* în raport cu eroarea la generalizare produsă de algoritmul AdaBoost — calculul propriu-zis va fi elaborat la punctul e —, vom avea nevoie să demonstrăm patru rezultate intermediare / „leme“, care vor constitui, câte una pe rând, obiectul punctelor *a-d* care urmează.

a. (*Lema 1.*) Considerând instanța de antrenament x arbitrar aleasă (dar fixată) în mulțimea X , precum și un număr oarecare $\theta > 0$, demonstrați că are loc următoarea inegalitate:

$$P_g \left(|f(x) - g(x)| > \frac{\theta}{2} \right) \leq 2 \exp \left(-N \frac{\theta^2}{8} \right) \stackrel{not.}{=} \beta_\theta. \quad (252)$$

Sugestie: Folosiți inegalitatea lui Hoeffding, enunțată la pr. 22.b, de la capitolul de *Fundamente*.

b. (*Lema 2.*) Fie D' o distribuție de probabilitate oarecare definită pe $X \times \{-1, +1\}$. Arătați că are loc următoarea inegalitate, care este asemănătoare cu cea de la punctul precedent, însă diferă de aceasta prin faptul că instanța etichetată (x, y) este lăsată acum să urmeze distribuția (nespecificată) D' :

$$P_{D',g} \left(|yf(x) - yg(x)| > \frac{\theta}{2} \right) \leq \beta_\theta. \quad (253)$$

⁵⁷¹ Atenție! Este posibil ca pentru doi indici diferenți $j \neq j'$ să existe un același $t \in \{1, \dots, T\}$ astfel încât să avem $g_j = g_{j'} = h_t$.

⁵⁷² În mod alternativ, g poate fi văzut ca un sondaj (engl., survey) efectuat asupra ipotezelor „slabe“ h_t . (Aceste ipoteze „slabe“ ar putea fi văzute ca „votanții“.)

⁵⁷³ Andi Munteanu a implementat un program pentru a „verifica“ această proprietate interesantă pe datele de la problema 25. Pentru cazul în care se lucrează cu compași de decizie exterioiri, au fost obținute următoarele aproximări ale voturilor normalize α_t (și anume, $\alpha_1 = 0.20376359$, $\alpha_2 = 0.32302264$, $\alpha_3 = 0.47321376$):

iterația	α_1	α_2	α_3
1	0.0	0.0	1.0
10	0.0	0.8	0.2
10^2	0.15	0.41	0.44
10^3	0.207	0.327	0.466
10^4	0.2065	0.3209	0.4726
10^5	0.20326	0.32399	0.47275
10^6	0.202648	0.323807	0.473545
10^7	0.2037366	0.3230384	0.473225

Sugestie: Puteti folosi egalitatea $P_{D',g}(\cdot) = E_{D'}[P_g(\cdot)]$, unde P_g , este probabilitatea care a constituit obiectul inegalitatii de la punctul precedent. (Observati ca P_g este probabilitate marginala in raport cu $P_{D',g}$.)

c. (*Lema 3.*) Fie g definit ca mai sus, precum si $\theta > 0$, ambii fixati. Consideram $p_{g,\theta} \stackrel{\text{not.}}{=} P_D\left(yg(x) \leq \frac{\theta}{2}\right)$ si $\hat{p}_{g,\theta} \stackrel{\text{not.}}{=} P_S\left(yg(x) \leq \frac{\theta}{2}\right)$, unde D si S au fost introduse in prima parte a acestei probleme.⁵⁷⁴ Demonstrați că in raport cu alegerea setului de date de antrenament (S), are loc inegalitatea următoare:

$$P_{\text{date}}(p_{g,\theta} - \hat{p}_{g,\theta} > \varepsilon) \leq \exp(-2m\varepsilon^2). \quad (254)$$

Sugestie: Folositi din nou inegalitatea lui Hoeffding.

d. (*Lema 4.*) Fie $\delta > 0$, oarecare. Aratați că are loc inegalitatea

$$P_{\text{date}}(\exists g \in \mathcal{C}_N \text{ și } \theta > 0 : p_{g,\theta} > \hat{p}_{g,\theta} + \varepsilon) \leq \delta \quad (255)$$

dacă se ia

$$\varepsilon = \sqrt{\frac{\ln\left((\frac{N}{2} + 1)|\mathcal{H}|^N/\delta\right)}{2m}}. \quad (256)$$

Sugestie: Folositi Lema 3 și inegalitatea lui Boole (“union bounds”):

$$P\left(\bigcup_{i=1}^n A_i\right) \leq \sum_{i=1}^n P(A_i),$$

unde A_1, \dots, A_n sunt evenimente aleatoare oarecare dintr-un spațiu pe care a fost definită o distribuție de probabilitate P .

Consecință dedusă din relația (255): Prin complementaritate, urmează că pentru orice $\delta \in (0, 1]$, orice $g \in \mathcal{C}_N$ și orice $\theta > 0$ (de fapt, este suficient pentru $\theta \in (0, 2]$), are loc inegalitatea

$$p_{g,\theta} \leq \hat{p}_{g,\theta} + \varepsilon,$$

cu probabilitate de cel puțin $1 - \delta$.

e. Acum vom „asambla“ rezultatele care au fost demonstate la punctele a-d.
i. Folosind Lema 2, aratați că

$$P_D(yf(x) \leq 0) \leq E_g[P_D(yg(x) \leq \theta/2 | g)] + \beta_\theta,$$

iar după aceea, aplicând membrului drept al acestei inegalități Lema 4, va rezulta că pentru orice $\delta \in (0, 1]$, inegalitatea următoare este satisfăcută cu [o] probabilitate de cel puțin $1 - \delta$:

$$P_D(yf(x) \leq 0) \leq E_g[P_S(yg(x) \leq \theta/2 | g)] + \varepsilon + \beta_\theta,$$

unde ε a fost deja definit în relația (256).

ii. Folosind din nou Lema 2, aratați că

$$E_g[P_S(yg(x) \leq \theta/2 | g)] + \varepsilon + \beta_\theta \leq P_S(yf(x) \leq \theta) + 4e^{-N\theta^2/8} + \sqrt{\frac{\ln\left[\left(\frac{N}{2} + 1\right)|\mathcal{H}|^N/\delta\right]}{2m}}.$$

⁵⁷⁴Funcția g fiind construită în mod aleatoriu, rezultă că $p_{g,\theta}$ și $\hat{p}_{g,\theta}$ pot fi văzute ca fiind variabile aleatoare.

iii. Arătați că pentru $N = \left\lceil \frac{4}{\theta^2} \ln \frac{m}{\ln |\mathcal{H}|} \right\rceil$, inegalitatea rezultată de la punctele precedente (i și ii), și anume

$$P_D(yf(x) \leq 0) \leq P_S(yf(x) \leq \theta) + 4e^{-N\theta^2/8} + \sqrt{\frac{\ln \left[\left(\frac{N}{2} + 1 \right) |\mathcal{H}|^N / \delta \right]}{2m}}$$

poate fi exprimată sub forma

$$P_D(yf(x) \leq 0) \leq P_S(yf(x) \leq \theta) + O\left(\frac{1}{\sqrt{m}} \sqrt{\frac{\ln m \cdot \ln |\mathcal{H}|}{\theta^2} - \ln \delta}\right).$$

Răspuns:

a. Fie $Z_j \stackrel{\text{not.}}{=} g_j(x)$. Rezultă că $g(x) = \frac{1}{N} \sum_{j=1}^N g_j(x) = \frac{1}{N} \sum_{j=1}^N Z_j \stackrel{\text{not.}}{=} \bar{Z}$. Întrucât variabilele aleatoare Z_j sunt mărginite de intervalul $[-1, 1]$, aplicând inegalitatea lui Hoeffding, vom obține:

$$\begin{aligned} P_g\left(|f(x) - g(x)| > \frac{\theta}{2}\right) &= P_g\left(|E_g[Z] - \bar{Z}| > \frac{\theta}{2}\right) \stackrel{\text{Hoeffding}}{\leq} 2 \exp\left(-\frac{2N^2\theta^2}{4N}\right) \\ &= 2 \exp\left(-\frac{N\theta^2}{8}\right). \end{aligned}$$

La scrierea primei egalități am folosit *Observația* din enunț, conform căreia $f(x) = E_g[g_j(x)] = E_g[Z]$.

b. Folosind *sugestia* din enunț și apoi inegalitatea care a fost dedusă la punctul a (Lema 1), putem scrie:

$$\begin{aligned} P_{D',g}\left(|yf(x) - yg(x)| > \frac{\theta}{2}\right) &\stackrel{\text{not.}}{=} P_{D',g}\left(\{(x,y) \mid |yf(x) - yg(x)| > \frac{\theta}{2}\}\right) \\ &= E_{D'}\left[P_g\left(|yf(x) - yg(x)| > \frac{\theta}{2} \mid D'\right)\right] \\ &\stackrel{(252)}{\leq} E_{D'}[\beta_\theta] = \beta_\theta. \end{aligned}$$

c. Definim variabilele aleatoare Q_i pentru $i = 1, \dots, m$, unde $m = |S|$, astfel:

$$Q_i = \begin{cases} 1, & \text{dacă } y_i g(x_i) \leq \theta/2; \\ 0, & \text{în cazul contrar.} \end{cases}$$

Rezultă că $p_{g,\theta} \stackrel{\text{not.}}{=} P_D\left(yg(x) \leq \frac{\theta}{2}\right) = E_D[Q_i]$ și $\hat{p}_{g,\theta} \stackrel{\text{not.}}{=} P_S\left(yg(x) \leq \frac{\theta}{2}\right) = \frac{1}{m} \sum_{i=1}^m Q_i \stackrel{\text{not.}}{=}$

\bar{Q} . Tinând cont că variabilele Q_i iau valori în mulțimea $\{0, +1\}$ și aplicând a doua inegalitate a lui Hoeffding (vedeți pr. 22.b), obținem:

$$P_{date}(p_{g,\theta} - \hat{p}_{g,\theta} > \varepsilon) = P_{date}(E_D[Q_i] - \bar{Q} > \varepsilon) \leq \exp\left(-\frac{2m^2\varepsilon^2}{m \cdot 1^2}\right) = \exp(-2m\varepsilon^2).$$

d. Pentru a evalua probabilitatea $P_{date}(\exists g \in \mathcal{C}_N \text{ și } \theta > 0 : p_{g,\theta} > \hat{p}_{g,\theta} + \varepsilon)$, vom face două *remarci*. Prima dintre ele se referă la numărul de valori pe care

este suficient să le analizăm pentru C_N și respectiv pentru θ . Din definiția mulțimii C_N , rezultă $|C_N| = |\mathcal{H}|^N$. Acum vom arăta că din infinitatea de valori pe care θ le poate lua, există doar un număr finit de valori care sunt de interes. Într-adevăr,

$$yg(x) \leq \theta/2 \Leftrightarrow \frac{y}{N} \sum_j g_j(x) \leq \frac{\theta}{2} \Leftrightarrow y \sum_j g_j(x) \leq \frac{N}{2} \theta \Leftrightarrow y \sum_{j=1}^N \underbrace{g_j(x)}_{\in \{-1, +1\}} \leq \left\lfloor \frac{N}{2} \theta \right\rfloor.$$

Din enunțul punctului e se observă că este suficient să considerăm $\theta \leq 1$. Prin urmare, valorile nenegative pe care le poate lua expresia $y \sum_{j=1}^N g_j(x)$ sunt $0, 1, \dots, \left\lfloor \frac{N}{2} \right\rfloor$. Acestea sunt în număr de $\left\lfloor \frac{N}{2} \right\rfloor + 1$.

A doua remarcă este următoarea. Fie $\tilde{\theta} = \frac{2}{N}v$, cu $v \in \{0, 1, \dots, \left\lfloor \frac{N}{2} \right\rfloor\}$. Este imediat că $p_{g,\theta} = p_{g,\tilde{\theta}}$ pentru orice $\theta \in \left[\tilde{\theta}, \tilde{\theta} + \frac{2}{N} \right)$:

$$p_{g,\tilde{\theta}} \stackrel{\text{not.}}{=} P_D \left(yg(x) \leq \frac{1}{2}\tilde{\theta} \right) = P_D \left(yg(x) \leq \frac{1}{2}\theta \right) \stackrel{\text{not.}}{=} p_{g,\theta}.$$

Similar, are loc egalitatea $\hat{p}_{g,\theta} = \hat{p}_{g,\tilde{\theta}}$.

Folosind aceste două remarcă, putem scrie acum:

$$\begin{aligned} P_{date} (\exists g \in \mathcal{C}_N \text{ și } \theta > 0 : p_{g,\theta} > \hat{p}_{g,\theta} + \varepsilon) \\ = P_{date} (\exists g \in \mathcal{C}_N \text{ și } \theta > 0 : p_{g,\tilde{\theta}} > \hat{p}_{g,\tilde{\theta}} + \varepsilon) \\ \stackrel{\text{ineq. Boole}}{\leq} |\mathcal{H}|^N \left(\frac{N}{2} + 1 \right) P_{date} (p_{g,\tilde{\theta}} > \hat{p}_{g,\tilde{\theta}} + \varepsilon) \\ \stackrel{(254)}{\leq} |\mathcal{H}|^N \left(\frac{N}{2} + 1 \right) e^{-2\varepsilon^2 m}. \end{aligned}$$

Egalând cu δ ultima expresie pe care am obținut-o, vom avea:

$$\begin{aligned} |\mathcal{H}|^N \left(\frac{N}{2} + 1 \right) e^{-2\varepsilon^2 m} = \delta \Leftrightarrow \\ |\mathcal{H}|^N \left(\frac{N}{2} + 1 \right) = \delta e^{2\varepsilon^2 m} \Leftrightarrow \\ 2\varepsilon^2 m = \ln \left(|\mathcal{H}|^N \left(\frac{N}{2} + 1 \right) / \delta \right) \Leftrightarrow \\ \varepsilon^2 = \frac{1}{2m} \ln \left(|\mathcal{H}|^N \left(\frac{N}{2} + 1 \right) / \delta \right) \Leftrightarrow \\ \varepsilon = \sqrt{\frac{\ln \left[\left(\frac{N}{2} + 1 \right) |\mathcal{H}|^N / \delta \right]}{2m}}. \end{aligned}$$

e. Folosind lemele pe care le-am demonstrat anterior, vom obține:

$$\begin{aligned} P_D (yf(x) \leq 0) \\ = P_{D,g} (yf(x) \leq 0) \end{aligned} \tag{257}$$

$$\begin{aligned} = P_{D,g} (yf(x) \leq 0 \wedge yg(x) \leq \theta/2) + \\ P_{D,g} (yf(x) \leq 0 \wedge yg(x) > \theta/2) \end{aligned} \tag{258}$$

$$\leq P_{D,g}(yg(x) \leq \theta/2) + \underbrace{P_{D,g}(|yf(x) - yg(x)| > \theta/2)}_{\leq \beta_\theta} \quad (259)$$

$$\stackrel{\text{Lema 2}}{\leq} E_g[P_D(yg(x) \leq \theta/2 | g)] + \beta_\theta \quad (260)$$

Lema 4
 $\leq E_g[P_S(yg(x) \leq \theta/2 | g) + \varepsilon] + \beta_\theta$ cu probabilitate $> 1 - \delta$,
 unde ε a fost definit de relația (256);

$$= E_g[P_S(yg(x) \leq \theta/2 | g)] + \varepsilon + \beta_\theta \quad (261)$$

$$= P_{S,g}(yg(x) \leq \theta/2) + \varepsilon + \beta_\theta \quad (262)$$

$$= P_{S,g}(yg(x) \leq \theta/2 \wedge yf(x) \leq \theta) + P_{S,g}(yg(x) \leq \theta/2 \wedge yf(x) > \theta) + \varepsilon + \beta_\theta \quad (263)$$

$$\leq P_{S,g}(yf(x) \leq \theta) + \underbrace{P_{S,g}(|yf(x) - yg(x)| > \theta/2)}_{\leq \beta_\theta} + \varepsilon + \beta_\theta \quad (264)$$

$$\stackrel{\text{Lema 2}}{\leq} P_S(yf(x) \leq \theta) + 2\beta_\theta + \varepsilon$$

$$\stackrel{(252),(256)}{=} P_S(yf(x) \leq \theta) + 4e^{-N\theta^2/8} + \sqrt{\frac{\ln[(\frac{N}{2}+1)|\mathcal{H}|^N/\delta]}{2m}}.$$

Iată câteva explicații succinte pentru deducerea relațiilor de mai sus:

(257): P_D este distribuție / probabilitate marginală în raport cu $P_{D,g}$, iar $yf(x)$ nu depinde de g ;

(258): se aplică proprietatea de aditivitate numărabilă din definiția funcției de probabilitate;

(259): se aplică proprietatea $A \subseteq B \Rightarrow P(A) \leq P(B)$, respectiv se ține cont de faptul că în condițiile date avem $|yf(x) - yg(x)| = -(yf(x) - yg(x)) > \theta/2$;

(260): P_D este distribuție / probabilitate marginală în raport cu $P_{D,g}$;

(261): ε este constant în raport cu alegerea lui g ;

(262): P_S este distribuție / probabilitate marginală în raport cu $P_{S,g}$;

(263): se aplică proprietatea de aditivitate numărabilă;

(264): se aplică proprietatea $A \subseteq B \Rightarrow P(A) \leq P(B)$, respectiv se ține cont de faptul că în condițiile date avem $yf(x) - yg(x) < -\theta/2 < 0$, deci $|yf(x) - yg(x)| = -(yf(x) - yg(x)) > \theta/2$.

În fine, se poate arăta relativ ușor că atunci când se ia $N = \left\lceil \frac{4}{\theta^2} \ln \frac{m}{\ln |\mathcal{H}|} \right\rceil$, ultima expresie de mai sus poate fi pusă sub forma

$$P_S(yf(x) \leq \theta) + O\left(\frac{1}{\sqrt{m}} \sqrt{\frac{\ln m \cdot \ln |\mathcal{H}|}{\theta^2} - \ln \delta}\right).$$

Într-adevăr,

$$\begin{aligned} N &= \left\lceil \frac{4}{\theta^2} \ln \frac{m}{\ln |\mathcal{H}|} \right\rceil \Rightarrow \frac{4}{\theta^2} \ln \frac{m}{\ln |\mathcal{H}|} \leq N < \frac{4}{\theta^2} \ln \frac{m}{\ln |\mathcal{H}|} + 1 \\ &\Rightarrow -\frac{1}{2} \ln \frac{m}{\ln |\mathcal{H}|} \geq -N \frac{\theta^2}{8} > -\frac{1}{2} \ln \frac{m}{\ln |\mathcal{H}|} - \frac{\theta^2}{8} \\ &\Rightarrow \exp\left(-\frac{1}{2} \ln \frac{m}{\ln |\mathcal{H}|}\right) \geq \exp\left(-N \frac{\theta^2}{8}\right) > \exp\left(\frac{1}{2} \ln \frac{m}{\ln |\mathcal{H}|} - \frac{\theta^2}{8}\right) \end{aligned}$$

$$\Rightarrow \sqrt{\frac{\ln |\mathcal{H}|}{m}} \geq e^{-N} \frac{\theta^2}{8} > \sqrt{\frac{\ln |\mathcal{H}|}{m}} \cdot e^{-\frac{\theta^2}{8}}.$$

Evident, $e^{-\theta^2/8} \leq 1$. Așadar, ordinul expresiei $e^{-N} \frac{\theta^2}{8}$ este $O\left(\frac{\sqrt{\ln |\mathcal{H}|}}{\sqrt{m}}\right)$.

Apoi, întrucât

$$\begin{aligned} \varepsilon &= \sqrt{\frac{\ln \left[\left(\frac{N}{2} + 1 \right) |\mathcal{H}|^N / \delta \right]}{2m}} = \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{m}} \cdot \sqrt{\ln \left[\left(\frac{N}{2} + 1 \right) |\mathcal{H}|^N / \delta \right]} = \\ &= \frac{1}{\sqrt{2}} \cdot \frac{1}{\sqrt{m}} \cdot \sqrt{\ln \left(\frac{N}{2} + 1 \right) + N \ln |\mathcal{H}| - \ln \delta}, \end{aligned}$$

ordinul expresiei $\frac{N}{2} + 1$ este cel mult $O\left(\frac{\ln m}{\theta^2}\right)$ și ordinul expresiei $N \ln |\mathcal{H}|$ este cel mult $O\left(\frac{\ln m}{\theta^2} \cdot \ln |\mathcal{H}|\right)$, rezultă că ordinul expresiei $2\beta_\theta + \varepsilon$ este cel mult $O\left(\frac{1}{\sqrt{m}} \sqrt{\frac{\ln m \cdot \ln |\mathcal{H}|}{\theta^2}} - \ln \delta\right)$.

31.

(Algoritmul AdaBoost: Adevărat sau Fals?)

- MIT, 2003 fall, Tommi Jaakkola, final, pr. 3.1-2
- MIT, 2001 fall, Tommi Jaakkola, midterm, pr. 4.3
- MIT, 2002 fall, Tommi Jaakkola, midterm, pr. 5.4
- CMU, 2011 spring, Eric Xing, HW5, pr. 3.1.b

- a. ε_t , eroarea ponderată produsă la antrenare de către ipoteza h_t / clasificatorul „slab“ A (măsurată relativ la ponderile de la începutul iterăției t) tinde să crească în raport cu t .
- b. În decursul iterățiilor executate de algoritmul AdaBoost, erorile ponderate ε_t (produse la antrenare, pe rând, de către ipotezele „slabe“ h_t , în ocurență, compașii de decizie) pe de o parte, și erorile produse la antrenare de către clasificatorii combinați H_t pe de altă parte, variază aproximativ la fel.
- c. Ponderile / „voturile“ α_t asignate de către algoritmul AdaBoost clasificatorilor „slabi“ h_t asamblați sunt întotdeauna nenegative.
- d. Probabilitățile / ponderile $D_t(i)$ alocate de către algoritmul AdaBoost exemplelor de antrenament care au fost clasificate eronat [de către ipoteza h_t] vor crește cu un același factor multiplicativ.
- e. Întotdeauna după ce algoritmul AdaBoost execută suficient de multe iterății, eroarea la antrenare produsă de ipoteza combinată H_t descrește la o valoare care este oricât [dorim să fie] de apropiată de zero, indiferent de tipul de clasificatori „slabi“ folosiți.

Răspuns:

a. Adevărat (în general). Modul de definire a probabilităților / ponderilor $D_t(i)$ asignate exemplelor de antrenament face ca algoritmul AdaBoost să se concentreze asupra exemplelor care sunt dificil de clasificat corect. După câteva iterații, cea mai mare parte a „masei“ de probabilitate va fi alocată acestor exemple „dificele“, iar eroarea ponderată la antrenare comisă de următoarea ipoteză „slabă“ va fi mai apropiată de $1/2$ (care reprezintă eroarea corespunzătoare alegării aleatorii).

b. Fals. În vreme ce eroarea la antrenare produsă de către clasificatorul combinat H_t în mod tipic descrește ca funcție de t (numărul de iterații executate de AdaBoost), erorile ponderate la antrenare ε_t produse de ipotezele „slabe“ h_t în mod tipic devin din ce în ce mai mari (așa cum am justificat deja la punctul precedent), fiindcă ponderile / probabilitățile $D_t(i)$ se alocă din ce în ce mai mult exemplelor care sunt dificil de clasificat.

c. Adevărat. După cum s-a specificat în pseudo-codul din problema 22, algoritmul AdaBoost alege la fiecare iterație (t) ipoteze „slabe“, care au o eroare ponderată la antrenare ε_t strict mai mică decât $1/2$. Prin urmare, $\ln((1 - \varepsilon_t)/\varepsilon_t) > 0$, ceea ce înseamnă că „votul“ α_t este pozitiv.

d. Adevărat. Puteți verifica acest fapt analizând formula (227) din problema 22, pentru actualizarea probabilităților $D_t(i)$. Întrucât pentru toate exemplele incorrect clasificate avem $y_i \neq h_t(x_i)$, iar y_i și $h_t(x_i)$ pot fi doar ± 1 , rezultă că probabilitățile / ponderile alocate lor vor fi multiplicate cu factorul $\exp(-\alpha_t y_i h_t(x_i)) = \exp(\alpha_t)$. După aceea se face normalizarea, cu un același factor, Z_t .

e. Fals. Dacă la o anumită iterație t a algoritmului AdaBoost clasificatorul „slab“ A nu poate produce nicio ipoteză care să aibă eroarea ponderată la antrenare $\varepsilon_t < 1/2$, atunci algoritmul AdaBoost se oprește.⁵⁷⁵ La momentul respectiv, cea mai mică dintre erorile produse la antrenare de către clasificatorii [combinări] $H_{t'}$ cu $t' = 1, \dots, t$, este fie 0 fie strict pozitivă.

Însă, chiar și în cazul în care clasificatorul „slab“ A produce la orice iterație ipoteze h_t cu eroare $\varepsilon_t < 0.5$ (deci $\gamma_t \stackrel{\text{not.}}{=} \frac{1}{2} - \varepsilon_t > 0$), deși avem certitudinea că *marginea superioară* $\exp(-2 \sum_{t'=1}^t \gamma_t^2)$ pentru $\text{err}_S(H_t)$ descrește ca funcție de t (vedeți pr. 23.d), nu putem avea și certitudinea că ea converge la 0, și nici certitudinea că *eroarea* $\text{err}_S(H_t)$ însăși descrește mereu. Vedetă, de exemplu, soluția a două de la problema 25 (cea care a fost obținută fără a folosi prag exterior).

⁵⁷⁵ Altminteri, adică atunci când $\varepsilon_t = 1/2$, ponderea ipotezei h_t respectivă ar fi $\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \varepsilon_t}{\varepsilon_t} \right) = \frac{1}{2} \ln 1 = 0$, deci h_t nu mai poate îmbunătăți combinația liniară învățată $f(x)$ și, prin urmare, eroarea clasificatorului combinat H_t rămâne neschimbată.

4.2 Arbori de decizie — Probleme propuse

4.2.1 Algoritmul ID3

32.

(Arbori de decizie; optimalitate,
ca număr minim de noduri)

*

Reprezentați arborele / arborii de decizie care are / au numărul minim posibil de noduri (de test) și corespunde / corespund funcției booleene $(A \text{ XOR } B) \wedge C$ definită peste atributele booleene A, B și C .

33.

(Expresivitatea arborilor de decizie:
un rezultat privind funcțiile booleene)

Orice funcție booleană (care primește n argumente din mulțimea $\{0, 1\}$ și întoarce un element din mulțimea $\{0, 1\}$) poate fi reprezentată cu ajutorul unui arbore de decizie. Adevărat sau fals?

În cazul afirmativ, explicați succint cum anume poate fi construit arborele de decizie respectiv.

În cazul negativ, dați un exemplu de funcție booleană pentru care nu se poate construi un arbore de decizie consistent cu funcția respectivă.

34.

(Calcularea câștigului de informație pe “decision stumps”)

■ □ • CMU, 2013 fall, W. Cohen, E. Xing, Sample Questions, pr. 4

Studentul Timmy dorește să știe cum [ar trebui să procedeze cel mai bine ca] să promoveze examenul de învățare automată. Pentru aceasta, a cules informații de la studenții care au urmat acest curs în anii precedenți și apoi a decis să-și construiască un *model* bazat pe arbori de decizie. A colectat în total nouă *instanțe* / exemple, descrise cu ajutorul a două *trăsături* (văzute în cele ce urmează ca două variabile aleatoare, S și A): „este bine să stai și să înveți până noaptea târziu înainte de examen“ (S) și „este bine să mergi la toate cursurile și seminariile“ (A). Timmy dispune acum de următoarele „statistici“ (care sunt de fapt *partitionări* ale datelor sale):

$$\begin{aligned} Set(\text{all}) &= [5+, 4-] \\ Set(S+) &= [3+, 2-], Set(S-) = [2+, 2-] \\ Set(A+) &= [5+, 1-], Set(A-) = [0+, 3-] \end{aligned}$$

Presupunând că se folosește drept criteriu de selecție a celei mai bune trăsături câștigul maxim de informație, ce trăsătură va alege Timmy? Care este valoarea câștigului de informație?

Puteți folosi la calcule următoarele aproximății:

N	3	5	7
$\log_2 N$	1.5850	2.3219	2.8073

35.

(Implementare: compas de decizie,
entropie, entropie condițională specifică,
entropie condițională medie, câștig de informație)

□ *Liviu Ciortuz, 2016*

Folosind limbajul de programare pe care-l preferați, implementați un program care, pornind de la o structură de date de tip compas de decizie (engl., decision stump), calculează entropia, entropiile condiționale specifice, entropia condițională medie, precum și câștigul de informație aferent.

În mod concret, programul va primi ca *input*

- m — numărul de valori posibile ale etichetei / atributului de ieșire (în mod implicit, se va considera $m = 2$);
- n — numărul de valori ale atributului (notat mai jos cu A) în raport cu care se face partitioarea mulțimii de instanțe asociate nodului-rădăcină al compasului de decizie (valoarea implicită: $n = 2$);
- partițiile (de fapt, count-urile) corespunzătoare nodurilor descendente. Pornind de la aceste partiții, programul va calcula partiția asociată nodul-rădăcină al compasului de decizie.

De exemplu, pentru primul compas de decizie de la problema 34, inputul va avea forma $[3, 2]$, $[2, 2]$, în vreme ce pentru al doilea compas de decizie va fi $[5, 1]$, $[0, 3]$.

Programul va calcula și apoi va afișa

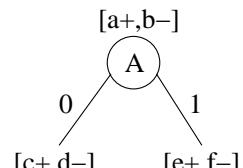
- entropia atributului / variabilei de ieșire (notată aici cu Y);
- entropiile condiționale specifice pentru fiecare descendant din nodul-rădăcină;
- entropia condițională medie a atributului A ;
- câștigul de informație al atributului [de ieșire] Y în raport cu atributul [de intrare] A .

36.

(Compași de decizie;
entropie, entropia condițională medie și câștigul de informație;
formule de calcul adecvate pentru folosirea calculatorului de buzunar)

■ • * *UAIC, Iași, 2017, Sebastian Ciobanu, Liviu Ciortuz*

Fie compasul de decizie din figura alăturată. a, b, c, d, e și f reprezintă count-uri corespunzătoare unui set de date de antrenament. După cum se observă, eticheta (sau, variabilă de ieșire), notată cu Y , este binară, iar atributul (sau, variabilă de intrare) A este de asemenea binar. Evident, $a = c + e$ și $b = d + f$.



a. Arătați că entropia [variabilei de ieșire] corespunzătoare partiției asociate nodului de test este

$$H[a+, b-] = \frac{1}{a+b} \log_2 \frac{(a+b)^{a+b}}{a^a b^b} \text{ dacă } a \neq 0 \text{ și } b \neq 0.$$

b. Cum s-ar scrie formula corespunzătoare entropiei variabilei de ieșire în cazul când ea este ternară, iar partiția din nodul de test [al compasului de decizie] este $[a+, b-, c*]$?

Atenție! Nu există nicio legătură între acest ultim c și count-ul c din compasul de decizie de mai sus.

c. Presupunând că niciunul dintre c, d, e și f nu este nul, arătați că entropia condițională medie corespunzătoare compasului de decizie din desenul de mai sus este

$$H_{nod|atribut} = \frac{1}{a+b} \log_2 \left(\frac{(c+d)^{c+d}}{c^c d^d} \cdot \frac{(e+f)^{e+f}}{e^e f^f} \right).$$

d. Să presupunem acum că unul dintre count-urile c, d, e și f este 0; pentru fixarea ideilor vom considera $c = 0$. Elaborați formula entropiei condiționale medii pentru compasul de decizie în acest caz.

e. Demonstrați următoarea formulă pentru câștigul de informație corespunzătoare compasului de decizie de mai sus, presupunând că a, b, c, d, e și f sunt strict pozitive:

$$IG_{nod;atribut} = \frac{1}{a+b} \log_2 \left(\frac{(a+b)^{a+b}}{a^a b^b} \cdot \frac{c^c d^d}{(c+d)^{c+d}} \cdot \frac{e^e f^f}{(e+f)^{e+f}} \right).$$

Observație (1): Întrucât majoritatea calculatoarelor de buzunar nu au funcția \log_2 ci funcțiile \ln și \lg , în formulele prezentate sau deduse la punctele $a-e$ ar fi de dorit să schimbăm baza logaritmului. Aceasta revine – pe lângă înlocuirea lui \log_2 cu \ln sau \lg – la înmulțirea membrului drept cu $1/\ln 2$, respectiv $1/\lg 2$.

Observație (2): Întrucât, la aplicarea algoritmului ID3, pentru alegerea celui mai bun atribut de pus în nodul curent este suficient să calculăm entropiile condiționale medii, va fi suficient să comparăm produsele de forma

$$\frac{(c+d)^{c+d}}{c^c d^d} \cdot \frac{(e+f)^{e+f}}{e^e f^f} \tag{265}$$

pentru compașii de decizie considerați la nodul respectiv și să alegem minimul dintre aceste produse.

Atenție! O problemă importantă care poate apărea la folosirea acestor formule în lucrul cu calculatorul de buzunar este *depășirea capacitatii de reprezentare* a rezultatelor intermediare. Spre exemplu, la un calculator Sharp EL-531VH, putem lucra cu 56^{56} dar nu și cu 57^{57} . Similar, pe calculatorul disponibil în [meniul *Accessories* din] sistemul de operare Linux Mint putem lucra cu 179^{179} dar nu și cu 180^{180} . Din acest motiv, în cazul depășirii capacitatii de reprezentare pe calculatoare de buzunar, trebuie să utilizați formulele de bază pentru entropii și pentru câștigul de informație, întrucât ele folosesc mult mai mult funcția \log .

Comentariu:

Raționamentele relaționale / „calitative“

în contextul calculării entropiilor condiționale medii

Acet tip de raționament — spre deosebire de *raționamentul cantitativ*, reprezentat de calculul entropiei condiționale (specifice și apoi medii) folosind efectiv *definițiile clasice*⁵⁷⁶ — este justificat în felul următor.

Conform notației folosite de Tom Mitchell în cartea *Machine Learning*, știm că entropia condițională medie a lui S (setul de date de antrenament) în raport cu atributul A este definită astfel:

$$H(S|A) \stackrel{\text{def.}}{=} \sum_{v \in \text{Val}(A)} \frac{|S_v|}{|S|} \cdot H(S|A = v) \stackrel{\text{not.}}{=} \sum_{v \in \text{Val}(A)} P(A = v) \cdot H(S|A = v), \quad (266)$$

unde S_v este submulțimea lui S formată din acele instanțe pentru care $A = v$, iar $H(S|A = v)$ este entropia lui S_v (care este numită *entropie condițională specifică*).

Considerăm acum A' , un alt atribut din descrierea instanțelor de antrenament din setul S . Are loc următoarea proprietate:

Proprietate: Dacă există o corespondență injectivă $v \mapsto v'$,⁵⁷⁷ unde $v \in \text{Val}(A)$ și $v' \in \text{Val}(A')$, astfel încât

$$\frac{|S_v|}{|S|} \cdot H(S|A = v) \leq \frac{|S_{v'}|}{|S|} \cdot H(S|A' = v') \text{ pentru orice } v \in \text{Val}(A), \quad (267)$$

sau, scris echivalent

$$P(A = v) \cdot H(S|A = v) \leq P(A' = v') \cdot H(S|A' = v') \text{ pentru orice } v \in \text{Val}(A),$$

din relația (266) rezultă imediat că

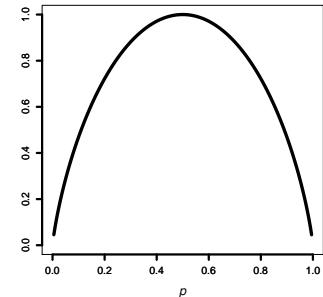
$$H(S|A) \leq H(S|A').$$

Evident, relația (267) este satisfăcută în următorul *caz particular*:

$$\frac{|S_v|}{|S|} \leq \frac{|S_{v'}|}{|S|} \text{ și } H(S|A = v) \leq H(S|A' = v') \text{ pentru orice } v \in \text{Val}(A). \quad (268)$$

Verificarea acestor relații (din varianta (268)) este mult mai ușor de realizat în mod vizual (decât varianta (267)), atunci când — aşa cum procedăm noi de obicei la aplicarea algoritmului ID3 — se lucrează cu compași de decizie pentru a determina „cel mai bun atribut“ de pus în nodul curent al arborelui de decizie aflat în curs de construire.

Atunci când se procedează aşa, dacă A și A' sunt variabile Bernoulli (adică, fiecare dintre ele are câte două valori), pentru a verifica inegalitățile de tipul $H(S|A = v) \leq H(S|A' = v')$, se apelează de obicei la *monotonia* (și *simetria*) funcției entropie pentru distribuția Bernoulli. Graficul acestei funcții este prezentat în figura alăturată.



⁵⁷⁶Vedeți formulele date la problema 55 de la capitolul de *Fundamente*.

⁵⁷⁷Mai exact, ar fi trebuit să notăm corespondența injectivă respectivă cu $f : \text{Val}(A) \rightarrow \text{Val}(A')$ și în loc de v' să scriem $f(v)$ peste tot în cele ce urmează. În text, vom păstra însă notația [mai] simplă: v și corespondentul său v' . Evident, $S_{v'}$ este submulțimea lui S formată din acele instanțe pentru care $A' = v'$.

Observații:

1. În ce privește corespondența injectivă a cărei existență am presupus-o / cerut-o mai sus, este suficient ca ea să fie definită pe acea submulțime a lui $Val(A)$ pentru care este îndeplinită condiția

$$P(A = v) \neq 0 \text{ și, mai ales}(!), H(S|A = v) \neq 0.$$

2. În diagramele din culegere (sau din slide-uri) unde utilizăm astfel de raționamente relaționale / „calitative“, folosim următoarea *convenție*: cu o linie continuă punem în evidență relații de tipul (268), iar printr-o linie punctată relații de tipul $H(S|A = v) \leq H(S|A' = v')$.
3. În exercițiile didactice — mai ales în condiții de test sau examen, când timpul pus la dispoziție este limitat! — este adeseori utilă combinarea raționamentelor calitative cu cele cantitative,⁵⁷⁸ încrucișând adeseori inegalitățile din relațiile (267) și / sau (268) nu sunt satisfăcute pentru toate valorile $v \in Val(A)$.

37.

(Algoritmul ID3: aplicare)

*prelucrare de Liviu Ciortuz, după
CMU, 2020 fall, Aarti Singh, Recitation*

Fie următorul set de date, relativ la când anume un copil iese afară să se joace:

Day	Weather	Temperature	Humidity	Wind	Play
1	Sunny	Hot	High	Weak	No
2	Cloudy	Hot	High	Weak	Yes
3	Sunny	Mild	Normal	Strong	Yes
4	Cloudy	Mild	High	Strong	Yes
5	Rainy	Mild	High	Strong	No
6	Rainy	Cool	Normal	Strong	No
7	Rainy	Mild	High	Weak	Yes
8	Sunny	Hot	High	Strong	No
9	Cloudy	Hot	Normal	Weak	Yes
10	Rainy	Mild	High	Strong	No

În acest exercițiu vă cerem să elaborați arborele învățat de algoritmul ID3 pe acest set de date.

- a. Stabiliți în mod riguros ce atribut va pune algoritmul ID3 în nodul rădăcină. Veți folosi următorul tabel, în care se dau entropiile mai multor variabile Bernoulli, determinate de valoarea parametrului p :

p	1/4	1/3	2/5	3/7
$H(p)$	0.8112	0.9182	0.9709	0.9852

- b. Satisfacăți aceeași cerință, apelând acum la alte două metode: mai întâi raționamentul relațional („calitativ“) care a fost prezentat la *Comentariul* de la pag. 578 și apoi metoda expusă la problema 36 (vedeți în special *Observația* (2) de acolo).

⁵⁷⁸Vedeți și metoda expusă la problema 2.d.

c. Stabiliți ce atrbute pot fi puse în nodurile de test de pe nivelul 2 (adică, în descendenții nodului rădăcină), iar apoi completați arborele ID3.⁵⁷⁹ Este oare arborele [elaborat de algoritm] ID3 unic? Care este eroarea pe setul de date de antrenare?

38.

(Calcularea unor entropii;
aplicarea algoritmului ID3)

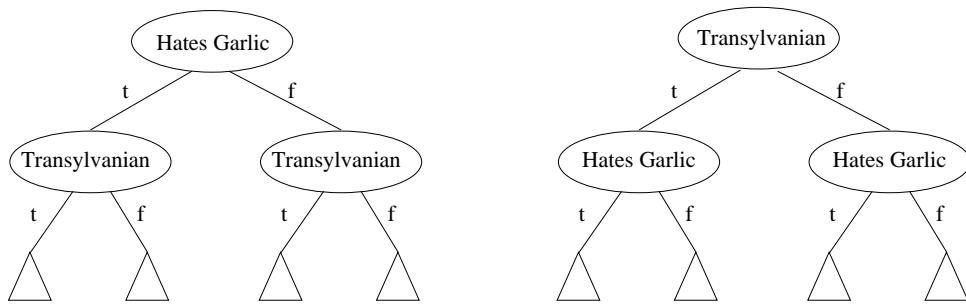
*prelucrare de Liviu Ciortuz, după
• CMU, 2014 spring, Seyoung Kim, HW2, pr. 1.4.1-5*

Centrul [de Medicină] pentru Controlul și Prevenția Maladiilor a fost sesizat în legătură cu o creștere surprinzătoare a aparițiilor de vampiri. Acest centru a colectat date preliminare referitoare la anumite *caracteristici*, atât pentru vampiri [deja] cunoscuți cât și pentru non-vampiri, și acum ar dori să construiească un arbore de decizie ca să-i ajute pe cetăteni să identifice noi vampiri. Datele culese sunt prezentate în tabelul următor:

V (Vampire)	G (Hates Garlic)	T (Transylvanian)	Nr. apariții
+	t	t	9
+	t	f	4
+	f	t	3
+	f	f	0
-	t	t	1
-	t	f	3
-	f	t	1
-	f	f	6

Fiecare linie indică ce caracteristici au fost „observate“, și de câte ori a fost observată fiecare combinație de caracteristici. De exemplu, combinația (+, t, t) a fost observată de 9 ori, pe când combinația (+, f, f) n-a fost observată niciodată. (Simbolii 't' și 'f' au fost folosiți în locul lui True și False, pentru a evita confuzia cu atributul T.)

a. Completați arborii de mai jos cu informații (count-uri) referitoare la partitioanarea datelor (sub forma $[+n, -m]$) în fiecare nod, precum și cu deciziile care trebuie luate (în sens majoritar) în nodurile frunză.



⁵⁷⁹ Atenție! Puteți elabora raționamente perfecte fără să faceți calcule laborioase!

Explorând cu atenție datele de antrenament, veți vedea că puteți să identificați atrbute cu putere de predicție / „discriminare“ maximă [a valorilor variabilei de ieșire]. Cât este entropia condițională medie a acestor atrbute [în raport cu variabila de ieșire]?

- b. Calculați entropia condițională medie $H(V|G)$. (Toate calculele intermediare trebuie făcute cu o precizie de cel puțin 4 zecimale, pentru a ne asigura că răspunsul final are primele 3 zecimale corecte.)
- c. Calculați entropia condițională medie $H(V|T)$. (Din nou, toate calculele intermediare trebuie făcute cu o precizie de cel puțin 4 zecimale.)
- d. Care dintre cei doi arbori de decizie de mai sus reprezintă rezultatul învățării realizate de algoritm ID3 pe aceste date?
- e. Adevărat sau Fals: Arboarele produs de către ID3 va clasifica o persoană căreia-i displace usturoiul (engl., hates garlic) dar nu este transilvănean ca fiind vampir.

Indicație: Este posibil să aveți nevoie de următoarele valori pentru entropia ($H(p)$) unei variabile aleatoare Bernoulli de parametru p : $H(1/7) = 0.5916$, $H(4/17) = 0.7871$, $H(3/10) = 0.8812$, $H(4/13) = 0.8904$, $H(11/27) = 0.9751$.

39.

(Algoritmul ID3: aplicare pe expresii booleene; exploatarea simetriilor operațiilor \vee, \wedge în alegerea atributelor; analiza „optimalității“ arborelui ID3)

* prelucrare de Liviu Ciortuz, după Tom Mitchell, "Machine Learning", 1997, ex. 3.1.d

Considerăm următoarea funcție booleană: $(A \wedge B) \vee (C \wedge D)$. Valorile pe care le ia această funcție, calculate conform diferitelor valori de adevăr atribuite variabilelor / atributelor A, B, C și D sunt cele cunoscute din logica propozițiilor. Dorim însă să reprezentăm această funcție ca arbore de decizie.

- a. Aplicați algoritmul ID3 acestei funcții.

Observație: Dacă exploatați simetriile, este nevoie doar de puține calcule, altfel vă complicați în mod inutil.

- b. Arboarele ID3 obținut la punctul precedent este optimal?

Altfel spus, puteți găsi alt arbore de decizie de adâncime mai mică sau cu număr mai mic de noduri (de test) pentru această funcție? (Țineți cont că în fiecare nod al unui arbore de decizie se poate testa un sigur atribut.)

40.

(Algoritmul ID3: aplicare; analiza „optimalității“ arborelui ID3)

* CMU, 2005 spring, C. Guestrin, T. Mitchell, midterm, pr. 4

Agenția spațială NASA dorește să distingă între marțieni (M) și pământeni (H) folosind următoarele caracteristici: $Green \in \{N, Y\}$, $Legs \in \{2, 3\}$, $Height \in \{S, T\}$, $Smelly \in \{N, Y\}$.

Datele de antrenament de care dispunem sunt prezentate în tabelul alăturat.

a. Învătați un arbore de decizie folosind algoritmul ID3 și trasați arborele respectiv.

	Species	Green	Legs	Height	Smelly
1	M	N	3	S	Y
2	M	Y	2	T	N
3	M	Y	3	T	N
4	M	N	2	S	Y
5	M	Y	3	T	N
6	H	N	2	T	Y
7	H	N	2	S	N
8	H	N	2	T	N
9	H	Y	2	S	N
10	H	N	2	T	Y

b. Descrieți conceptul M (marțian) ca un set de reguli conjunctive din logica propozițiilor. Spre exemplu:

```
if Green = Y and Legs = 2 and Height = T and Smelly = N then M;  
else  
if ... then M; else H.
```

c. Soluția de la punctul b de mai sus folosește cel mult 4 atribute în fiecare conjuncție. Găsiți un set de reguli conjunctive care folosesc doar 2 atribute pentru fiecare conjuncție, păstrând însă eroarea la antrenare zero. Această ipoteză mai simplă poate fi reprezentată ca un arbore de decizie de adâncime 2? Justificați răspunsul.

41.

(Algoritmul ID3: aplicare; cazul instanțelor de antrenament cu multiple apariții)

• CMU, 2010 fall, Aarti Singh, HW2, pr. 5.1

Tabelul de mai jos descrie instanțe (înregistrări) pozitive și instanțe negative pentru persoane cărora banca le-a acordat (sau nu le-a acordat) un card de credit.

Fiecare linie din tabel indică niște combinații de valori observate pentru atributele considerate (*Gender*, *Income* și *Approved*) și de câte ori a fost înregistrată respectiva combinație de valori. De exemplu, (F, Low, +) a apărut de 10 ori, iar (F, Low, -) de 80 de ori.

Gender	Income	Approved	Counts
F	Low	+	10
F	High	+	95
M	Low	+	5
M	High	+	90
F	Low	-	80
F	High	-	20
M	Low	-	120
M	High	-	30

a. Calculați entropia atributului *Approved* pe acest set de date de antrenament (folosind logaritmul cu baza 2).

b. Calculați de asemenea câștigurile de informație $IG(Approved, Gender)$ și $IG(Approved, Income)$.

c. Desenați un arbore de decizie produs de către algoritmul ID3 (fără post-pruning) pe baza acestui set de date de antrenament.

42.

(“Decision stump” produs de ID3:
raționament calitativ pe un exemplu simplu)

- o CMU, 2010 fall, Ziv Bar-Joseph, midterm exam, pr. 5.a

Vrem să construim un arbore de decizie care să ne ajute să prezicem întârzierile avioanelor. Timp de câteva luni am colectat informații, iar un rezumat al acestora este prezentat în tabelul următor:

Atribut	Valoare = Da		Valoare = Nu	
	#Zboruri amâname	neamâname	#Zboruri amâname	neamâname
Ploaie	30	10	10	30
Vânt	25	15	15	25
Vara	5	35	35	5
Iarna	20	10	20	30
Ziua	20	20	20	20
Noaptea	15	10	25	30

- a. Pe baza acestui tabel precizați ce atribut ar trebui să fie pus în rădăcina arborelui de decizie, folosind criteriul câștigului de informație. Justificați riguros; nu este însă necesar să elaborați în detaliu toate calculele.
- b. Pe baza aceluiași tabel, precizați care dintre attribute ar trebui să apară pe al doilea nivel (nivelul de sub rădăcină) al arborelui de decizie.

43.

(O variantă ipotetică a algoritmului ID3)

*enunț formulat de Liviu Ciortuz,
pornind de la un slide al lui A. Moore
(vedeți și CMU, 2011 fall, E. Xing, HW1, pr. 1.2.4)*

La curs am precizat că algoritmul ID3 în varianta standard (cea fără attribute numerice continue) respectă următoarea *restrictie*: niciun atribut de intrare (notat A) nu poate să apară de două sau mai multe ori pe vreun drum care unește nodul rădăcină cu un nod frunză oarecare (dar fixat).

La acest exercițiu ne propunem să examinăm ce se întâmplă dacă eliminăm această restricție.

- a. Cum anume va fi partionată mulțimea de instanțe asignată nodului în care atributul A apare pentru a doua (sau a n -a oară)?
- b. Cât va fi câștigul de informație calculat pentru un astfel de nod?
- c. Înând cont de răspunsurile date la punctele precedente, se justifică *restrictia* formulată mai sus?

44.

(Algoritmul ID3: aplicare; calculul erorii la antrenare, respectiv la validare)

* CMU, 2003 fall, T. Mitchell, A. Moore, midterm exam, pr. 1

Folosiți setul de date alăturat pentru a învăța cu ajutorul unui arbore de decizie dacă o anumită floare este *Iris* (acesta este numele latinesc pentru *stânjenel*) sau nu, utilizând atribuțiile discrete *Formă*, *Culoare* și *Miros*.

<i>Formă</i>	<i>Culoare</i>	<i>Miros</i>	<i>Iris</i>
C	B	1	1
D	B	1	1
D	W	1	1
D	W	2	1
C	B	2	1
D	B	2	0
D	G	2	0
C	U	2	0
C	B	3	0
C	W	3	0
D	W	3	0

- a. Ce atribut va alege algoritmul ID3 ca rădăcină a arborelui de decizie?
- b. Elaborați întregul arbore de decizie care va fi învățat din datele de mai sus (fără pruning).
- c. Exprimăți cu ajutorul unui set de reguli din calculul propozițional clasificarea produsă de arborele de decizie obținut. (IF ... THEN *Iris*; IF ... THEN \neg *Iris*.)
- d. Să presupunem că avem un set de date de validare:

<i>Formă</i>	<i>Culoare</i>	<i>Miros</i>	<i>Iris</i>
C	B	2	0
D	B	2	0
C	W	2	1

Care va fi eroarea produsă de arborele de decizie pe mulțimea de date de antrenare respectiv pe datele de validare? (Exprimăți răspunsul ca număr de exemple clasificate greșit.)

45.

(O aproximare a numărului de instanțe greșit clasificate care au fost asignate la un nod frunză dintr-un arbore ID3)

• CMU, 2003 fall, T. Mitchell, A. Moore, midterm exam, pr. 9.b

Învățăm un arbore de decizie folosind un set de date de antrenament cu atributul de ieșire (*class*) având valorile 0 sau 1.

Presupunând că pentru un nod frunză *l* din acest arbore,

- există *M* instanțe de antrenament asignate la acel nod, iar
- entropia sa este *H*,

schițați un algoritm simplu care ia ca valori de intrare *M* și *H* și furnizează la ieșire numărul de exemple de antrenament clasificate greșit de către nodul frunză *l*.

Sugestie: Folosiți o aproximare simplă (polinomială) pentru funcția entropie *H(p)*.

46.

(Algoritmul ID3: eroarea la antrenare)

• ○ CMU, 2003 fall, T. Mitchell, A. Moore, HW1, pr. 2.1

Un student mi-a spus următoarele:

- el poate să construiască un set de instanțe cu atributele de intrare discrete și atributul de ieșire binar;
- mie îmi dă voie să aleg o parte din acest set de instanțe (dar nu toate!) pentru a antrena un arbore de decizie;
- indiferent de modul cum mi-ăs alege datele de antrenament din setul construit de el, eroarea de clasificare pe care arborele de decizie (obținut în urma antrenării cu algoritmul ID3) o va face pe instanțele care nu au fost incluse în setul de antrenament va fi de cel puțin 50%.

Credeți că studentul are dreptate? Explicați de ce sau dați un exemplu.

47.

(Extensiile ale algoritmului ID3: variabile de intrare continue; determinarea celui mai bun prag de separare: o proprietate)

□ • USC, 2008 fall, Sofus Macskassy, HW2, pr. 3

Atunci când construim arbori de decizie, selecția atributelor se face folosind de obicei criteriul câștigul de informație maxim.

Arătați că în cazul variabilelor de intrare continue, pragurile de separare (engl., split-point) pentru care de o parte și de alta etichetele sunt de același tip nu vor conduce niciodată la câștig de informație maxim.

Sugestie: Vă recomandăm să citiți [secțiunea 4 din] lucrarea *Technical note: On the handling of continuous-valued attributes in decision tree generation*, Usama M. Fayyad, Keki B. Irani, *Journal of Machine Learning*, nr. 8, 1992, pages 87-102.

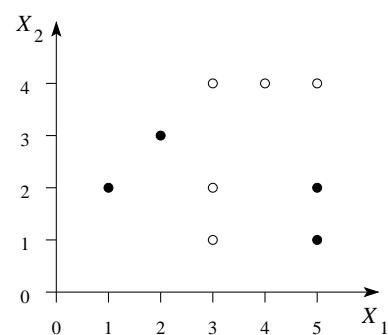
48.

(ID3 cu atrbute continue: zone de decizie și separatori decizionali)

■ □ • ○ Liviu Ciortuz, 2017,
folosind datele de la problema 24

Fie setul de date de antrenament din figura de mai jos (partea dreaptă. X_1 și X_2 sunt considerate atrbute numerice continue. Vă readucem aminte *convenția* noastră de notare: simbolul • desemnează instanțe pozitive, iar simbolul ○ instanțe negative.

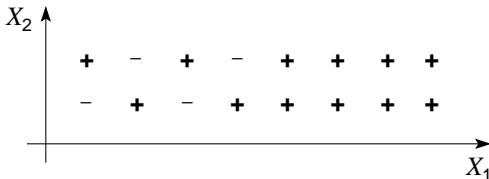
Aplicați algoritmul ID3 pe acest set de date. (Faceți toate calculele necesare, în mod detaliat; precizați la fiecare pas care sunt pragurile de split-are pentru cele două atrbute.) Desenați arborele de decizie rezultat. La final, reprezentați grafic zonele de decizie și separatorii decizionali, marcând clar zona pozitivă (sau zonele pozitive) și zona negativă (sau zonele negative).



49. (Extinderea algoritmului ID3 cu atrbute continue; eroarea la antrenare, eroarea la CVLOO; overfitting)

* CMU, 2005 fall, T. Mitchell, A. Moore, midterm exam, pr. 2

Figura alăturată prezintă un set de date cu două intrări X_1 și X_2 , variabile cu valori reale, și o ieșire Y care poate lua valori pozitive (+) sau negative (-).



Testăm doi algoritmi extremi de învățare de arbori de decizie. Algoritmul **OVERFIT** construiește un arbore de decizie în maniera standard a algoritmului ID3, fără a face pruning. Algoritmul **UNDERFIT** refuză complet să-și asume riscul splitării intervalelor de valori pentru X_1 și X_2 și construiește un arbore de decizie alcătuit doar dintr-un singur nod (care va fi simultan și rădăcină și nod frunză, deci și nod de decizie).

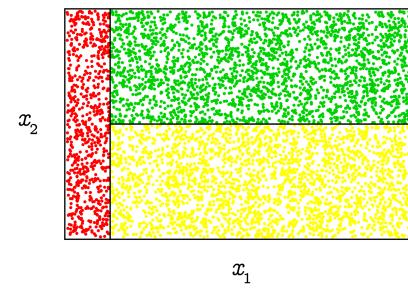
- Câte noduri frunză vor fi în arborele de decizie învățat de **OVERFIT** pe aceste date?
- Care va fi eroarea de clasificare la cross-validation cu metoda “Leave-One-Out” pe acest set de date atunci când folosim algoritmul **OVERFIT**? (Indicați datele incorect clasificate.)
- Similar punctului anterior, pentru cazul în care se folosește algoritmul **UNDERFIT**.

50. (Arborei de decizie cu variabile continue: ID3 ca algoritm “greedy”)

• ○ CMU, 2014 fall, W. Cohen, Z. Bar-Joseph, HW1, pr. 8.abcdef

Fie următoarea problemă de clasificare ternară.

Considerăm că în figura alăturată regiunea dreptunghiulară este populată în mod dens cu puncte caracterizate de două atrbute numerice (continue), x_1 și x_2 . Cele trei subdreptunghiuri (roșu, verde și galben) reprezintă trei clase de puncte, C_1, C_2 , și C_3 .



Dimensiunile $x_1 \times x_2$ ale dreptunghiurilor roșu, verde și galben sunt 1×6 , 7×3 și respectiv 7×3 . Dreptunghiul roșu este populat în mod uniform cu 6000 de puncte din clasa C_1 . Dreptunghiul verde este populat în mod uniform cu 42000 de puncte din clasa C_2 . Dreptunghiul galben este populat în mod uniform cu 42000 de puncte din clasa C_3 . Pentru simplitate, nu vom considera alte puncte decât acestea.

- Care este numărul minim de noduri de test pe care trebuie să le aibă un arbore de decizie pentru a clasifica în mod corect acest set de date?

b. Câte noduri de test are arborele de decizie obținut în urma antrenării algoritmului ID3 pe acest set de date, folosind criteriul maximizării câștigului de informație?

Indicație: Pentru a determina entropiile condiționale minime, puteți folosi proprietatea A3 de la problema 62 de la capitolul de *Fundamente*.

c. Avem același număr de noduri în cele două cazuri de mai sus, sau nu? Care credeți că este explicația?

d. Un arbore de decizie poate să clasifice setul de date din figura de mai sus cu 100% acuratețe (presupunând că nu există zgomote la nivel de etichete). Ce condiții trebuie să satisfacă în general un set de date de acest gen astfel încât arborele de decizie rezultat în urma antrenării să fie cât mai compact și să producă o acuratețe de 100%?

Indicație: Fiecare nod intern al arborelui de decizie corespunde unui test bazat pe o singură trăsătură. Gândiți-vă ce fel de clase de funcții / granițe de separare corespund unui astfel de arbore de decizie.

51.

(Un exemplu de aplicare a algoritmului ID3:
cazul când se folosesc atât variabile discrete
cât și variabile continue)

□ • * CMU, 2010 fall, Ziv Bar-Joseph, HW2, pr. 1

Cursul de Învățare Automată pe care l-am urmat în acest semestru, îmi-a insuflat dorința ca după absolvirea facultății să fondezi pe cont propriu o firmă (engl., start-up company). Pentru a-ți evalua şansele de succes — adică, mai exact, dacă vei deveni milionar sau nu —, ai colectat date de la absolvenții de la CMU, referitoare la foști studenți care și-au înființat propriile lor companii start-up. Pentru fiecare start-up, știi acum ce anume produce compania respectivă, ce fonduri de capital de investiție a obținut (exprimat în milioane de dolari), dacă directorul companiei a studiat la CMU o disciplină din domeniul științelor exacte și, în final, dacă directorul a devenit milionar. Tabelul de mai jos centralizează datele pe care le-am cules.

Produs	Capital atras	Ştiințe exacte	Milionar
SiteDeSocializare	2.9	Da	Nu
SiteDeSocializare	1.7	Da	Nu
SiteDeSocializare	3.4	Da	Nu
SiteDeSocializare	2.3	Nu	Da
MașinaAlimentatăCuCombustibilBio	3.4	Da	Da
MașinaAlimentatăCuCombustibilBio	6.1	Da	Da
MașinaAlimentatăCuCombustibilBio	5.6	Nu	Nu
MașinaAlimentatăCuCombustibilBio	0.6	Nu	Nu
NanoVaccin	1.9	Nu	Nu
NanoVaccin	2.9	Nu	Da
NanoVaccin	3.1	Da	Da
NanoVaccin	0.3	Da	Nu

Acum ai vrea să construiești un arbore de decizie pornind de la aceste date. Referitor la atributul cu valori continue *CapitalAtras*, știm că arborele de decizie poate conține teste (partiționări binare) de forma $\text{CapitalAtras} \leq v$ și $\text{CapitalAtras} > v$ și că pot exista mai multe teste de acest fel în arbore.

- Câte „praguri“ distincte v trebuie să considerăm pentru *CapitalAtras* atunci când căutăm atributul (optim) care trebuie pus în nodul rădăcină?
- Desenați arborele de decizie care va fi învățat de către algoritmul ID3 extins cu atribute cu valori continue, aşa cum a fost prezentat la curs. Ne vom referi ulterior la acest arbore ca fiind *arborele original*. Adnotați fiecare nod intern (i.e., nod de test) din arbore cu câștigul de informație obținut în urma aplicării testului respectiv.

Indicație: Pentru punctele c și d de mai jos, deși nu este neapărat necesar, este recomandabil să folosiți o implementare a algoritmului ID3 (extins cu atribute numerice continue). Vă sugerăm să abordați mai întâi problemele 35 și 58, care vă ghidează cum să construiți propria dumneavoastră implementare.

- Schimbați eticheta (i.e., valoarea clasei binare *Milionar*) pentru o instanță din tabelul de mai sus astfel încât arborele care va fi învățat ulterior să conțină cel puțin încă un nod de test în raport cu arborele de la punctul b .
- Un exemplu de antrenament este consistent cu arborele învățat dacă este clasificat corect de către acel arbore. Întrebarea pe care v-o adresăm acum este următoarea:

Este posibil să adăugăm la setul de date de antrenament de mai sus noi exemple care sunt consistente cu arborele original, dar care fac totuși ca algoritmul ID3 extins, executat pe noul set de date de antrenament, să învețe un arbore cu o rădăcină diferită de cea originală și cu mai multe noduri decât arborele original?

Dacă răspunsul dumneavoastră este afirmativ, indicați noile exemple pe care ați putea să le adăugați, precum și arborele de decizie rezultat. Dacă răspunsul este negativ, explicați de ce este așa.

52.

(Extensiile ale algoritmului ID3:
cazul atributelor care au multe valori
[algoritmul C4.5])

• * prelucrare de Liviu Ciortuz, 2020, după
CMU, 2015 fall, Z. Bar-Joseph, E. Xing, HW1, pr. 4.2

Una dintre limitările algoritmului ID3 [LC: ne referim la varianta de bază, fără extensiile pe care le-am discutat ulterior la curs] este faptul că el este prea sensibil la prezența atributelor care au un număr mare de valori. De exemplu, dacă fiecare instanță de antrenament are un ID unic, atunci câștigul de informație va fi maxim atunci când folosim acest ID ca atribut de intrare, ceea ce nu este deloc bine pentru faza de generalizare / predicție, în care putem întâlni adeseori instanțe având ID-uri care nu se regăsesc în datele de antrenament. Algoritmul C4.5 remediază acest aspect din funcționarea lui ID3 folosind în locul câștigului de informație, pentru a evalua atributele, *raportul câștigului de informație* (engl., information gain ratio).

Vom nota un atribut de intrare oarecare cu X , iar eticheta sa cu Y . Vă rea-ducem aminte că în algoritmul ID3, alegem pentru nodul curent acel atribut X care maximizează câștigul de informație, $IG(X)$:

$$IG(X) = H(Y) - H(Y|X).$$

Acum vom defini noțiunea de [LC: cantitate de] *informație la separare* (engl., split information) după cum urmează. Presupunem că avem $|D|$ instanțe atașate la nodul curent, iar după ce se face testul pe baza valorii atributului X , aceste instanțe sunt repartizate la V noduri descendente din nodul curent. Presupunând că numărul de instanțe care sunt asociate la aceste noduri-fii sunt respectiv $|D_1|, |D_2|, \dots, |D_V|$, vom defini *informația la separarea* valorilor atributului X astfel:⁵⁸⁰

$$SplitInfo(X) = - \sum_{j=1}^V \frac{|D_j|}{|D|} \log_2 \frac{|D_j|}{|D|}.$$

Ca și în cazul entropiei, se va considera, prin convenție, că $0 \cdot \log_2 0 = 0$.

Raportul câștigului de informație se definește în felul următor:⁵⁸¹

$$GainRatio(X) = \frac{IG(X)}{SplitInfo(X)}.$$

Algoritmul C4.5 (succesorul lui ID3) folosește *raportul câștigului de informație* pentru a determina „cel mai bun” atribut de pus în nodul curent. Intuiția spune că $SplitInfo(X)$ actionează ca un „normalizator” (engl., normalizer), care penalizează atributele care au un număr mare de valori. De exemplu, dacă pentru $\forall i, j$ avem $|D_i| = |D_j|$, atunci $SplitInfo(X) = \log_2 V$ (adică, maximum posibil), aşadar atributele care au multimea de valori (V) mai restrânsă vor fi preferate.

În acest exercițiu veți elabora arborele de decizie corespunzător setului de date din tabelul de mai jos, folosind pe de o parte algoritmul ID3 și pe de altă parte algoritmul C4.5. Atributele de intrare sunt *Outlook*, *Temperature*, *Humidity* și *Wind*, iar eticheta este asociată cu variabila de ieșire *EnjoyTennis*. Veți trata *Temperature* ca atribut discret.

⁵⁸⁰Se poate observa din formula de definiție că $SplitInfo(X)$ este *entropia* atributului X calculată pentru cele D instanțe asociate la nodul curent. (Ea nu este însă nicidecum același lucru cu *entropia conditională medie* a atributului X în raport cu variabila de ieșire Y , pe care o folosim la calcularea câștigului de informație.)

⁵⁸¹Cazul când $SplitInfo(X) = 0$ corespunde situației când X are o singură valoare (pentru instanțele din nodul în care se calculează acest $SplitInfo$), deci nu are putere discriminativă în raport cu variabila de ieșire. În consecință, astfel de atrbute nu vor fi luate în considerare atunci când, pentru nodul curent, se va pune problema să alegem atributul cu *GainRatio* maxim.

Day	Outlook	Temperature	Humidity	Wind	EnjoyTennis
D1	Sunny	26	High	Weak	No
D2	Sunny	25	High	Strong	No
D3	Overcast	25	High	Weak	Yes
D4	Rain	24	High	Weak	Yes
D5	Rain	19	Normal	Weak	Yes
D6	Rain	20	Normal	Strong	No
D7	Overcast	20	Normal	Strong	Yes
D8	Sunny	23	High	Weak	No
D9	Sunny	20	Normal	Weak	Yes
D10	Rain	25	Normal	Weak	Yes
D11	Sunny	24	Normal	Strong	Yes
D12	Overcast	22	High	Strong	Yes
D13	Overcast	23	Normal	Weak	Yes
D14	Rain	23	High	Strong	No

a. Pentru fiecare dintre atributele *Outlook*, *Humidity*, *Wind* și *Temperature*, faceți reprezentări grafice pentru compașii de decizie corespunzători, astfel:

- asociați mai întâi la nodului rădăcină al compasului de decizie *partiția* de instanțe corespunzătoare;
- procedați similar pentru fiecare dintre descendenții lui direcți (adică nodurile-fii);
- calculați entropia condițională medie a respectivului atribut și, în fine, câștigul de informație în raport cu atributul de ieșire *EnjoyTennis*.

Indicație: Următoarele valori pentru entropia distribuției Bernoulli vă pot fi de folos:

p	0	$1/3$	$2/5$	$3/7$	$1/2$	1
$H(p)$	0	0.918	0.970	0.985	1	0

b. Ce atribut va selecta algoritmul ID3 pentru nodul rădăcină al arborelui pe care-l „învață“?

c. Calculați valorile $SplitInfo(\cdot)$ pentru toate atributele de intrare, relativ la întregul set de date de antrenament.

Atenție! $SplitInfo(Humidity)$ și $SplitInfo(Wind)$ se calculează foarte ușor folosind informațiile din tabelul dat la *Indicația* de mai sus.

Pentru $SplitInfo(Outlook)$ și $SplitInfo(Temperature)$, vă cerem elaborați calculul cât mai complet posibil. (La calcule, puteți folosi aproximăriile $\log_2 3 = 1.584$, $\log_2 5 = 2.322$ și $\log_2 7 = 2.807$.)

Presupunând că $SplitInfo(Outlook) = 1.577$, și $SplitInfo(Temperature) = 2.646$, ce atribut va selecta algoritmul C4.5 pentru nodul rădăcină al arborelui pe care-l „învață“?

d. Elaborați complet arborele de decizie produs de algoritmul ID3. (*Atenție!* Dacă ați procedat corect la punctele *a* și *b*, atunci aici nu veți avea de făcut calcule aproape deloc!)

e. Elaborați complet arborele de decizie produs de algoritmul C4.5. (*Atenție!* Dacă ați procedat corect la punctul *c*, atunci aici nu veți avea de făcut calcule

aproape deloc!)

Ce observați comparând arborele obținut aici cu arborele care a fost obținut la punctul d ?

53. (Deficiențe ale pruning-ului top-down și respectiv bottom-up)

• CMU, 2009 fall, Carlos Guestrin, HW1, pr. 2.4

a. Știm că algoritmul ID3 alege în fiecare nod de test atributul care are cel mai mare câștig de informație dintre toate atributele de intrare rămase disponibile pentru acel nod. O posibilă metodă de *pruning* în manieră *top-down* este următoarea: dacă toate atributele despre care am vorbit au câștig de informație 0 atunci opriți aplicarea agloritmului ID3 (adică elaborarea arborelui de decizie) pe ramura respectivă. Această strategie este însă problematică, întrucât ea poate conduce la arbori de decizie inconsistenti cu datele, chiar și atunci când datele sunt consistente / necontradictorii. Vă cerem să furnizați un mic set de exemple de antrenament care să demonstreze acest fapt.
Sugestie: Este suficient să folosiți două atrbute de intrare booleene.

b. La problema 19.b, am discutat [și] despre *pruning bottom-up* ca modalitate de control asupra „complexității“ arborelui creat de algoritmul ID3. Totuși, și această metodă are anumite neajunsuri.

Pornind de la soluția pe care ați dat-o la punctul precedent, clonați fiecare exemplu în parte de K ori (unde K este un număr suficient de mare). Apoi adăugați N noi atrbute de intrare care iau valorile True / False cu o probabilitate uniformă. Cum va arăta arborele de decizie care rezultă? Explicați acum de ce poate să eșueze pruning-ul de tip bottom-up.

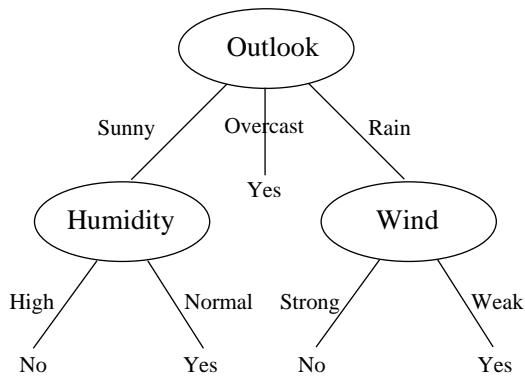
54.

(Algoritmul ID3: întrebări recapitulative;
compararea strategiilor de pruning:
reduced-error pruning vs. rule post-pruning)

• CMU, 2005 fall, T. Mitchell, A. Moore, HW1, pr. 1

Considerăm următorul set de date de antrenament (care diferă ușor de cel de la problema 52; vedeti pag. 590), împreună cu arborele de decizie care a fost învățat pe aceste date de către algoritmul ID3 (fără a se aplica vreo strategie de post-pruning).

Day	Outlook	Temperature	Humidity	Wind	EnjoyTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



a. Demonstrați că alegerea atributului *Wind* pe nivelul secund din arbore este corectă, arătând că în cazul lui câștigul de informație este superior față de cazurile / posibilitățile alternative.

b. Orice arbore de decizie poate fi exprimat sub forma unui set de reguli, scriind căte o regulă pentru fiecare nod-frunză (de fapt, pentru întreaga cale începând cu nodul-rădăcină și continuând până la nodul-frunză respectiv). *Pre-condițiile* dintr-o astfel de regulă corespund secvenței / succesiunii de attribute testate de-a lungul căii respective. De exemplu, nodul-frunză cel mai din stânga din arborele de mai sus corespunde regulii următoare:

IF (Outlook = sunny AND Humidity=High) THEN PlayTennis = No

Scriți regulile corespunzătoare celorlalte noduri-frunză. Observați faptul că acest set de reguli produce clasificări care sunt identice cu cele din arborele de decizie de mai sus, pentru orice instanță posibilă [LC: de antrenament sau de test].

c. Este posibil să „traducem“ orice arbore de decizie într-un set de reguli care reprezintă un clasificător echivalent. Invers, este oare posibil să „traducem“ orice set de reguli într-un arbore echivalent? Explicați succint sau, dacă este cazul, dați un contraexemplu.

d. La curs am discutat despre strategii de post-pruning pentru arbori de decizie, folosind abordarea *reduced-error pruning* relativ la un set de date de validare, pentru a evita overfitting-ul. Aici vom considera o strategie alternativă, care constă în a converti / „traduce“ arborele în setul de reguli echivalent și făcând apoi pruning pe reguli. În particular, vom considera că pruning-ul se execută în mod independent pe fiecare regulă în parte, parcurgând iterativ următorii pași:

- i. Determină în regula respectivă acele pre-condiții care, dacă sunt înlăturate (în mod individual) produc o îmbunătățire a acurateții pe setul de date de validare.⁵⁸²
- ii. Dacă există astfel de pre-condiții, identific-o pe cea care îmbunătățește cel mai mult acuratețea pe setul de date de validare, șterge-o din regulă și apoi *iterează* mai departe; în caz contrar *opreste* pruning-ul pe această regulă.

Considerați cele două strategii de pruning (adică pruning pe arbore, versus pruning pe setul de reguli echivalent). Este oare adevărat faptul că aceste două strategii de pruning produc clasificatori trunchiați (engl., pruned) care sunt echivalenți? Adică, produc oare arborele de decizie trunchiat și setul de reguli trunchiate clasificări identice (engl., equivalent) pentru orice instanță posibilă? Explicați de ce da, sau de ce nu.

55.

(Post-pruning pentru arborele ID3:
folosirea testului statistic χ^2 pentru identificarea
partiționărilor care sunt semnificative d.p.v. statistic)

• CMU, 2007 fall, Carlos Guestrin, HW1, pr. 1.3.1

La curs am menționat că arborii de decizie pot să manifeste fenomenul de “overfitting” și, în consecință, pentru a generaliza cât mai corect, trebuie să limităm „complexitatea“ arborilor pe care îi învățăm.

O modalitate de a face pruning pe arbori de decizie — de obicei, prin parcursul *de jos în sus*, după antrenare — este să analizăm din punct de vedere statistic instanțele asignate fiecărui nod de test din arbore. Mai precis, vom analiza etichetele acestor instanțe, cu *obiectivul* de a vedea cât de probabil este d.p.v. statistic să „observăm“ proporția acestor etichete în cazul în care atributul de intrare ales în nodul respectiv nu se află de fapt în corelație cu valorile atributului-țintă (etichetele).

În acest sens putem folosi *testul χ^2* al lui Pearson, care este un exemplu clasic de test bazat pe o *ipoteză statistică*. Vom pleca de la ipoteza că atributul de intrare asociat nodului respectiv nu este corelat cu atributul de ieșire (repräsentat de etichete), și vom verifica dacă datele „observate“ conduc în mod ferm la respingerea acestei ipoteze.

Mai exact, pornind de la datele de antrenament, vom calcula o anumită *mărimă* statistică, despre care știm că are o distribuție de tipul χ^2 în cazul în care cele două entități (atributul și etichetele) nu sunt corelate. Apoi, vom

⁵⁸²Acuratețea unei reguli este [o fracție, dată de] numărul de *predicții corecte* pe care le face această regulă, raportat la numărul total de *predicții*.

verifica dacă acea *mărime* statistică calculată are (sau nu) o *valoare* prea puțin probabil să fi fost generată de o distribuție χ^2 . În cazul afirmativ, vom respinge ipoteza că cele două entități nu sunt corelate. În cazul negativ, vom înlături nodul de test respectiv, împreună cu întregul subarbore atașat lui, cu un nod de decizie.

Concret, să presupunem că am învățat un arbore de decizie. Notăm cu S setul de exemple de antrenament ale căror drumuri de clasificare trec prin nodul a cărui semnificație statistică dorim să o testăm. Să zicem că în nodul respectiv se testează un atribut discret X care poate lua valorile $1, 2, \dots, k$. Fie p numărul de exemple din S care au eticheta $+$, iar $n = |S| - p$ numărul de exemple din S având eticheta $-$. Fie

S_i submulțimea lui S formată din instanțele care au $X = i$,

p_i numărul de exemple din S_i având eticheta $+$,

$n_i = |S_i| - p_i$ numărul de exemple din S_i având eticheta $-$.

În plus, fie $\bar{p}_i = p \cdot \frac{|S_i|}{|S|}$ și $\bar{n}_i = |S_i| - \bar{p}_i$. Acestea sunt valorile „așteptate“ ale lui p_i și respectiv n_i , în ipoteza că atributul X nu este corelat cu atributul de ieșire (*clasa*).

Folosind datele de mai sus, vom calcula *mărimea numerică corespunzătoare testului statistic* χ^2 :

$$\chi^2 = \sum_{i=1}^k \left(\frac{(p_i - \bar{p}_i)^2}{\bar{p}_i} + \frac{(n_i - \bar{n}_i)^2}{\bar{n}_i} \right)$$

În ipoteza că atributul X nu este corelat cu *clasa*, mărimea calculată mai sus urmează *distribuția* χ^2 .

Un *parametru* al distribuției χ^2 este așa-numitul *grad de libertate*. Pentru X , gradul de libertate este $k - 1$ atunci când se lucrează cu doar două etichete (+ și -).⁵⁸³ În particular, pentru un atribut binar gradul de libertate este 1.

După ce vom calcula valoarea mărimii statistice χ^2 pornind de la date, o vom compara cu o *valoare critică*, care reprezintă un prag astfel încât probabilitatea ca mărimea statistică calculată să depășească acel prag este de cel mult α în cazul în care variabilele nu sunt corelate. α este *nivelul de încredere* (engl., confidence level) și de obicei se consideră $\alpha = 0.05$, fiind astfel siguri în proporție de 95% că o partitioare (engl., split) care trece testul este semnificativă d.p.v. statistic.

De exemplu, atunci când gradul de libertate este 1 iar $\alpha = 0.05$, valoarea critică este 3.841; pentru 2 grade de libertate și același α , ea este 5.991. Spunem că o partitioare este *statistic semnificativă* dacă mărimea calculată χ^2 trece de valoarea critică.

⁵⁸³Dacă în loc de două etichete am fi considerat n etichete [de clasă] diferite, atunci am fi folosit o statistică asemănătoare, cu $(n - 1)(k - 1)$ grade de libertate.

Conform testului χ^2 pentru $\alpha = 0.05$, care dintre cele trei partiționări din arborele ID3 obținut pe datele din tabelul alăturat (Y fiind atributul-țintă) sunt statistic semnificative?

Y	X_1	X_2	Nr. apariții
+	T	T	3
+	T	F	4
+	F	T	4
+	F	F	1
-	T	T	0
-	T	F	1
-	F	T	3
-	F	F	5

56.

(Algoritmul ID3: Adevărat sau Fals?)

* CMU, 2004, fall, final exam, pr. 1.b

Învățăm un arbore de decizie folosind algoritmul ID3 standard, fără pruning. Atributele de intrare (X_1, X_2, \dots, X_m) sunt categoriale, iar atributul de ieșire (Y) este de asemenea categorial.

Marcați cu A (adevărat) sau F (fals) fiecare din afirmațiile de mai jos și dați în fiecare caz o explicație succintă, însotită eventual de un exemplu sau un contraexemplu minimalist.

- Dacă X_i și Y văzute ca variabile aleatoare sunt independente (raportat la distribuția probabilistă care a generat datele de antrenament), atunci X_i nu va apărea în arborele de decizie.
- Dacă $IG(Y, X_i) = 0$, atunci atributul X_i nu va apărea în arborele de decizie.
- Adâncimea maximă a arborelui de decizie este de cel mult m .

Notă: Dacă arborele este format doar din nodul rădăcină, ceea ce corespunde cazului în care toate exemplele de antrenament sunt identic clasificate, atunci se consideră că adâncimea arborelui este 0.

- Dacă sunt R exemple de antrenament, atunci adâncimea maximă a arborelui de decizie este de cel mult $1 + \log_2 R$.
- Dacă sunt R exemple de antrenament, iar unul dintre atrbutele de intrare are R valori distințe și ia valori v_1, \dots, v_R în mod injectiv (pe mulțimea formată de exemplele de antrenament), atunci arborele de decizie va avea adâncimea 0 sau 1.

57.

(Adevărat sau Fals?
ID3 vs. regresia logistică;
Acuratețe la antrenare vs. acuratețe la testare)

• ○ CMU, 2011 fall, T. Mitchell, A. Singh, midterm exam, pr. 1.1-3

Presupunem că avem un set de date de imagini celulare de la pacienți cu și, respectiv, fără cancer.

- Dacă ţi se cere să antrenezi un clasificator care prezice probabilitatea ca pacientul să aibă cancer, ai prefera să folosești mai degrabă arbori de decizie decât regresia logistică.

- b. Să zicem că setul de date conține 900 de imagini de la pacienți fără cancer și 100 de imagini de la pacienți cu cancer. Dacă antrenăm un clasificator și obținem 85% acuratețe pe acest set de date, putem spune că acesta este un clasificator bun.
- c. Un clasificator care atinge 100% acuratețe pe setul de antrenament și 70% acuratețe pe setul de test este mai bun decât un clasificator care atinge 70% acuratețe pe setul de antrenament și 75% acuratețe pe setul de test.

58.

(Implementare: algoritmul ID3)

 Liviu Ciortuz, 2016

Pornind de la pseudo-codul algoritmului ID3 în varianta de bază (adică, folosind doar atrbute cu valori discrete) dată în cartea *Machine Learning* de Tom Mitchell la pagina 56, elaborați o implementare, în limbajul de programare preferat. „Inima“ acestui algoritm este calculul câștigului de informație (a se vedea pr. 35). Programul va conține o funcție de antrenare și una de testare. Ca input (la linia de comandă), programul va primi

- numele unui fișier în format SSV (engl., space-separated values), conținând instanțele de antrenament, câte una pe fiecare linie, variabila de ieșire aflându-se pe ultima poziție;
- numele unui al doilea fișier, tot în format SSV, conținând instanța de test sau instanțele de test (de asemenea, câte una pe fiecare linie).

Optional, pe prima linie a fișierului de antrenament vor fi indicate pentru fiecare atrbut numele lui, precum și numărul de valori pe care le ia respectivul atrbut (drept separatori, se vor folosi tot spații). În acest caz, primul caracter din această linie-comentariu va fi %.

Folosiți o modalitate convenabilă pentru a afișa arborele ID3 obținut, de exemplu ca program în logica propozițiilor cu variabile cu valori multiple.

Pentru a testa programul, puteți folosi de pildă datele de la problema 2.

Ulterior, veți putea adăuga programului una dintre următoarele extensii (sau chiar ambele):

- variabile cu atrbute continue; testați programul pe datele de la problema 10 din prezentul capitol și pe cele de la problema 11.b de la capitolul *Învățare bazată pe memorare*;⁵⁸⁴
- o funcție care, în vederea contracărării fenomenului de overfitting, face trunchierea / pruning-ul arborelui ID3, în varianta “reduced-error”, folosind un al treilea set / fișier de instanțe, pentru validare; a se vedea cartea *Machine Learning* de Tom Mitchell, pag. 69-70.⁵⁸⁵

Indicație: Pentru a lucra în mod convenabil cu atrbutele discrete, mai ales atunci când aceste valori nu sunt numerice, în loc să folosiți (în timpul procesărilor) valorile lor aşa cum apar în fișierele de date, puteți crea un *vocabular*

⁵⁸⁴ Pentru testarea variantei algoritmului ID3 care folosește atât atrbute discrete cât și atrbute continue, puteți folosi datele de la problema 51, precum și cele de la problema 12.

⁵⁸⁵ Pentru testare, vedeti problemele CMU, 2008 spring, T. Mitchell, W. Cohen, HW1, pr. 2, CMU, 2012 spring, Roni Rosenfeld, HW3 și / sau CMU, 2011 fall, T. Mitchell, A. Singh, HW1, pr. 2.

(ca vector în care se memorează valorile tuturor acestor variabile) și apoi să înregistrați în tabelele de date (în locul valorile atributelor discrete) indecsii corespunzători „intrărilor“ din vocabular.

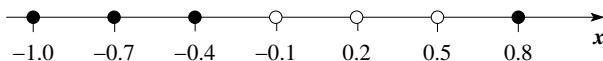
4.2.2 Algoritmul AdaBoost

59.

(Algoritmul AdaBoost: întrebări în legătură cu aplicarea algoritmului pe un set de date din \mathbb{R})

- ○ CMU, 2011 fall, T. Mitchell, A. Singh, HW6, pr. 3.2-8

Considerăm setul de date de antrenament din figura următoare. Simbolurile • și ○ desemnează etichete pozitive și respectiv etichete negative. Vom folosi algoritmul AdaBoost cu compași de decizie în rolul de ipoteze „slabe“.



- Determinați separatorul decizional corespunzător primei ipoteze „slabe“, h_1 . Desenați-l pe figura de mai sus și indicați [eventual printr-o mică săgeată perpendiculară pe acest separator] care este zona clasificată cu +.
- Calculați ε_1 și α_1 . Cât este acuratețea obținută de AdaBoost la antrenare dacă oprim acum algoritmul?
- Cât va fi valoarea noilor probabilități / ponderi $D_2(i)$ pentru fiecare dintre cele șapte exemple de antrenament?
- Identificați în mod riguros separatorul decizional corespunzător celei de-a doua ipoteze „slabe“, h_2 și apoi includeți-l în desen. Indicați iarăși zona de decizie corespunzătoare clasei +.
- Care sunt exemplele de antrenament cărora le va fi asignată cea mai mică pondere / probabilitate după ce algoritmul AdaBoost va fi terminat cea de-a doua sa iterată?
- Se îmbunătățește oare acuratețea la antrenare obținută de AdaBoost la a doua iteratie în raport cu cea obținută la prima iteratie?

60.

(AdaBoost: aplicare pe un set de date din \mathbb{R}^2)

adaptare facută de Liviu Ciortuz, după

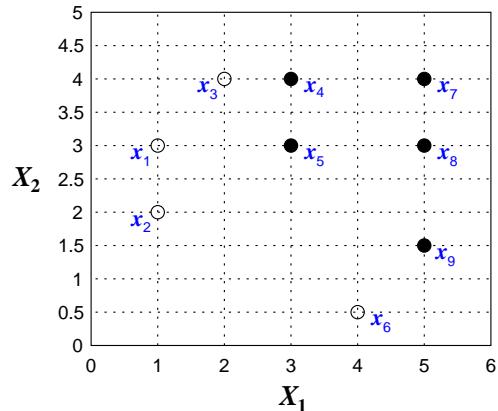
- ○ MIT, 2004 fall, Tommi Jaakkola, final, pr. 1

Fie setul de exemple de antrenament (instanțe etichetate) din figura de mai jos. Simbolurile • și ○ desemnează etichete pozitive și respectiv etichete negative. Pentru a rezolva această problemă de clasificare, vă cerem să aplicați algoritmul AdaBoost folosind drept ipoteze „slabe“ compași de decizie (engl., decision stump). La fiecare iteratie de boosting veți selecta acel compas de decizie care minimizează eroarea ponderată la antrenare (engl., weighted training error). În cazul în care există mai mulți compași de decizie care au [o

aceeași!] cea mai bună eroare ponderată la antrenare, veți putea alege unul dintre ei în mod arbitrar.

a. Pe figura dată, desenați primul compas de decizie și etichetați-l cu h_1 , indicând de asemenea cu $+/ -$ cele două zone de decizie pe care le determină acest compas.

b. Pe aceeași figură, scrieți în apropierea fiecărei instanțe probabilitatea asignată ei după prima iteratăie executată de algoritmul AdaBoost. De asemenea, încercuiți instanțele care au probabilitatea cea mai mare. Justificați răspunsul în mod riguros.



c. Cât este eroarea ponderată la antrenare produsă de primul compas de decizie după prima iteratăie, adică după ce probabilitățile asociate instanțelor au fost recalculate? Justificați.

d. Desenați pe figura dată cel de-al doilea compas de decizie și etichetați-l cu h_2 , indicând de asemenea cu $+/ -$ cele două zone de decizie pe care le determină acest compas.

e. Există oare instanțe de antrenament care sunt clasificate eronat de către H_2 , ipoteza combinată produsă de AdaBoost după două iteratăii? Justificați răspunsul în mod riguros.

f. Dacă răspunsul pe care l-ați dat la punctul e este pozitiv, ce se poate spune dacă se execută încă o iteratăie (adică, a treia)? Desenați pe figura dată cel de-al treilea compas de decizie și etichetați-l cu h_3 , indicând de asemenea cu $+/ -$ cele două zone de decizie pe care le determină acest compas. Dacă eroarea la antrenare este acum 0, trasați separatorul decizional determinat de algoritmul AdaBoost.

Care ar fi rezultatul dacă AdaBoost ar alege ca a treia ipoteză „slabă“ un alt (cel mai bun!) compas de decizie (h'_3) în locul lui h_3 ?

61.

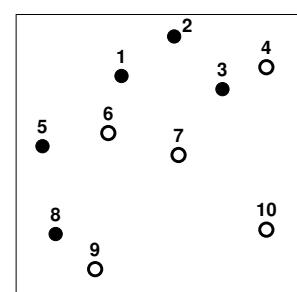
(Algoritmul AdaBoost: aplicare pe un set de date din \mathbb{R}^2)

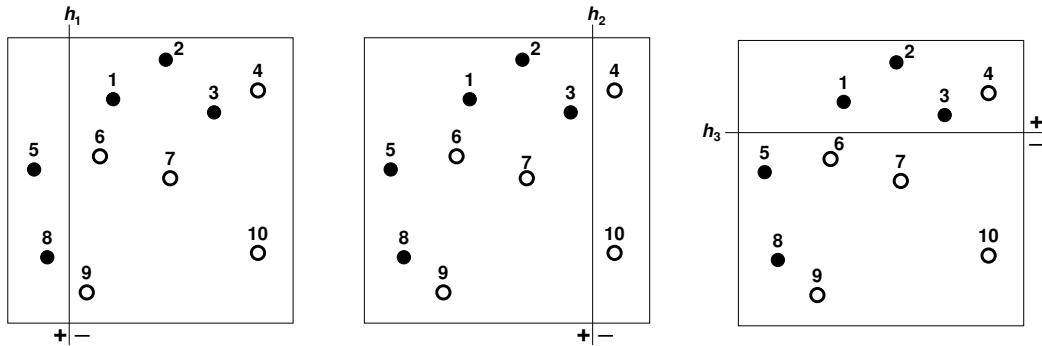
Liviu Ciortuz, 2021, pornind de la un slide al lui
· Virgil Pavlu, Northeastern University, USA

Se consideră că aplicăm algoritmul AdaBoost pe dataset-ul din figura alăturată. (Pentru ușurință exprimării la calcule, am notat pe figură indicii instanțelor de antrenament, în imediata apropiere a acestora.)

Folosim convenția noastră obișnuită de notare: simbolul \bullet desemnează instanțe pozitive, iar simbolul \circ instanțe negative.

La primele trei iteratăii ale algoritmului au fost selectați compașii de decizie h_1 , h_2 și h_3 (în această ordine), așa cum se indică în figurile de mai jos.





Obiectivul acestei probleme este să determinăm dacă la sfârșitul celor trei iterării algoritmul AdaBoost reușește să clasifice perfect toate instanțele de antrenament.

a. Calculați

- distribuțiile probabiliste corespunzătoare celor 3 iterării, adică $D_1(x_i)$, $D_2(x_i)$ și $D_3(x_i)$, pentru $i = 1, \dots, 10$;
- pentru fiecare dintre cele 3 iterării ($t = 1, 2, 3$): eroarea ponderată la antrenare (ε_t) produsă de compasul de decizie h_t , precum și ponderea (α_t) asociată ipotezei / compasului de decizie h_t .

Veți completa tabelele următoare și veți indica succint(!) modul în care ați procedat pentru a ajunge la rezultatele respective.

i	1	2	3	4	5	6	7	8	9	10
$D_1(x_i)$										
$D_2(x_i)$	1/6			1/14						
$D_3(x_i)$	7/66			1/22		1/6				

t	1	2	3
ε_t			
α_t			

Atenție! Pentru a vă ușura munca, am completat noi câteva dintre elementele tabelului precedent. Bazați-vă pe valorile indicate de noi, ca să nu faceți calcule laborioase, păstrând însă rigurozitatea / corectitudinea raționamentelor.

b. Folosind ipoteza combinată obținută de algoritmul AdaBoost la finalul celei de-a treia iterării, stabiliți:

- eroarea la antrenare produsă (pentru calcularea ei, puteți folosi tabelul de mai jos),
- zonele de decizie corespunzătoare acestui clasificator (veți justifica modul în care ați procedat!). Veți indica aceste zone de decizie, precum și granițele de decizie, pe primul desen din enunț.

t	α_t	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1	$\alpha_1 = \dots$										
2	$\alpha_2 = \dots$										
3	$\alpha_3 = \dots$										
	$H_3(x_i)$										

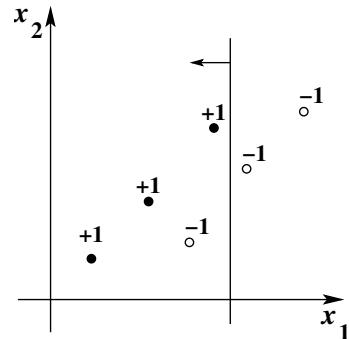
62.

(AdaBoost: întrebări în legătură cu aplicarea algoritmului pe un set de date din \mathbb{R}^2)

• o CMU, 200X spring, midterm, pr. 3
 MIT, 2006 fall, Tommi Jaakkola, final, pr. 2

Folosind algoritmul AdaBoost, vrem să obținem un ansamblu de compași de decizie (engl., decision stumps) h_t , de forma $H(x) = \text{sign}\left(\sum_{t=1}^T \alpha_t h_t(x)\right)$.

În figura alăturată sunt desenate câteva puncte (instanțe) etichetate în planul bidimensional, precum și primul compas de decizie care a fost ales de către algoritmul AdaBoost. Un compas de decizie oarecare produce valori binare ± 1 , ținând cont doar de un anumit prag (engl., the split point). Săgeata mică din figură, care este perpendiculară pe dreapta care reprezintă compasul de decizie indică *zona de decizie* pentru care compasul de decizie va produce valoarea +1.



a. Încercuiți toate acele instanțe din figură pentru care ponderea / probabilitatea [atribuită de către AdaBoost] va crește ca urmare a incorporării [în ipoteza combinată H] primului compas de decizie. Justificați răspunsul în mod riguros.

b. Desenați pe aceeași figură un compas de decizie care va putea fi selectat la următoarea iterație a algoritmului AdaBoost. Veți trasa atât dreapta care reprezintă compasul de decizie cât și [o săgeată care să indice] zona sa de decizie pozitivă. Justificați în mod riguros.

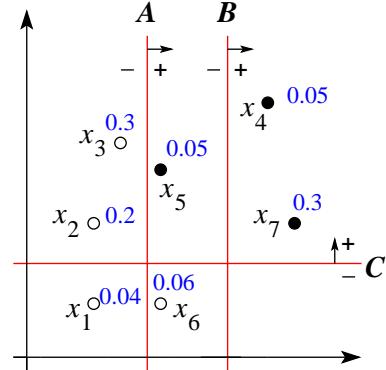
c. Va fi oare coeficientul / votul α_2 , care este asociat celui de-al doilea compas de decizie mai mare decât α_1 , coeficientul [din ansamblul H] pentru primul compas de decizie? Cu alte cuvinte, vom avea oare $\alpha_2 > \alpha_1$? Justificați în mod riguros.

63.

(Algoritmul AdaBoost: exemplu de aplicare pe date din \mathbb{R}^2 ; întrebări calitative)

*prelucrare de Liviu Ciortuz, 2021, după
• o MIT, 2007 fall, Tommi Jaakkola, final ex, pr. 1*

Fie setul de date de antrenament din figura alăturată. Am marcat cu semnul \circ exemplele / instanțele negative ($y_i = -1$) și cu semnul \bullet instanțele pozitive ($y_i = +1$). Figura conține de asemenea ponderile normalizate (adică, probabilitățile) asociate exemplelor de antrenament, aşa cum au rezultat în urma executării unui anumit număr de iterații ale algoritmului AdaBoost. În figură sunt tratați și trei compași de decizie, $h(x; \theta_A)$, $h(x; \theta_B)$ și $h(x; \theta_C)$ sau, pe scurt, A , B și C .



- Care dintre acești trei compași de decizie considerați că a fost folosit la *precedenta iteratie* a algoritmului AdaBoost, în așa fel încât să rezulte ponderile [asociate exemplelor] prezentate în figură? Veți răspunde indicând A , B sau C și veți justifica în mod riguros alegerea pe care ati făcut-o.
- Pe care dintre acești trei compași de decizie considerați că-l va selecta algoritmul AdaBoost la *iterația următoare*? Veți răspunde indicând A , B sau C și veți justifica riguros, prin calcule, alegerea pe care ati făcut-o.
- În figura dată, încercuți instanțele de antrenament (este posibil să nu fie niciuna!) pe care ansamblul (i.e., combinația liniară de compași de decizie) $H_2(x) = \text{sign}(\alpha_A h(x; \theta_A) + \alpha_C h(x; \theta_C))$, cu $\alpha_A = 0.3$ și $\alpha_C = 0.5$ nu le poate clasifica în mod corect.

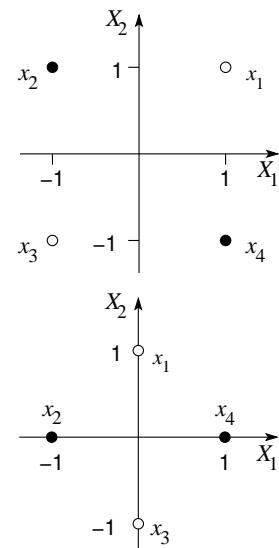
64.

(Algoritmul AdaBoost: aplicare pe dataset-uri de tip XOR)

• o CMU, 2011 fall, T. Mitchell, A. Singh, HW6, pr. 3.1
CMU, 2007 spring, Carlos Guestrin, HW2, pr. 2.2

- Considerați setul de date XOR, reprezentat ca de obicei în planul bidimensional. Presupunând că algoritmul AdaBoost folosește compași de decizie pe post de ipoteze „slabe“, va fi el oare capabil să obțină o acuratețe mai bună de 50% pe acest set de date? Justificați răspunsul dumneavoastră în mod riguros. (Ca de obicei, clasa +1 a fost reprezentată prin simbolul \bullet , iar clasa -1 prin simbolul \circ .)

În cele ce urmează veți aplica același algoritm AdaBoost pe setul de date din figura alăturată (obținut prin rotirea dataset-ului XOR cu 45° la stânga și făcând apoi rescalarea).

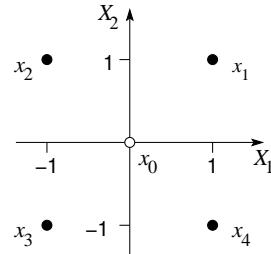


- b. Arătați cum lucrează AdaBoost pe acest [al doilea] set de date, considerând că numărul de iterații de efectuat este $T = 4$. Vă reamintim că pentru fiecare $t = 1, \dots, 4$ va trebui să calculați numerele $\varepsilon_t, \alpha_t, Z_t$, precum și distribuția probabilistă $D_t(i)$ pentru $i = 1, \dots, 4$. De asemenea, la fiecare iterație t veți trasa separatorul decizional corespunzător ipotezei h_t alese de clasificatorul „slab“ (A) și veți indica printr-o mică săgeată desenată perpendicular pe separator zona de decizie corespunzătoare etichetei $+1$.
- c. Cât va fi la finalul celor $T = 4$ iterații eroarea la antrenare produsă de [combinația liniară H livrată la ieșire de către] algoritmul AdaBoost?
- d. Este oare setul de date (cel folosit la punctele b și c) liniar separabil? Explicați de ce algoritmul AdaBoost se comportă mai bine decât un [singur] compas de decizie pe acest set de date.

65.

(Algoritmul AdaBoost: aplicare pe un set de date din \mathbb{R}^2) • CMU, 2010 fall, Aarti Singh, midterm, pr. 6.1-2

În acest exercițiu veți aplica algoritmul AdaBoost folosind drept clasificatori „slabi“ compași de decizie pe setul de date de antrenament prezentat în figura alăturată. (Atenție! Sunt patru instanțe pozitive și una negativă.)



- a. Care dintre aceste exemple de antrenament vor avea probabilitățile ($D_t(i)$) mărite la sfârșitul primei iterații? Încercuiți-le pe desen.
- b. Cât de multe iterații vor fi necesare pentru a atinge eroare zero la antrenare? Justificați elaborând toate detaliile necesare.
- c. Puteți adăuga încă un exemplu (instanță de antrenament) la setul de date de mai sus astfel încât algoritmul AdaBoost să obțină în [doar] două iterații eroare la antrenare zero? Dacă nu, explicați de ce nu este posibil așa ceva.
- d. Care credeți că este motivul (principal) pentru care se folosesc clasificatori „slabi“ (și nu clasificatori mai puternici / „tari“) în conjuncție cu algoritmul AdaBoost?

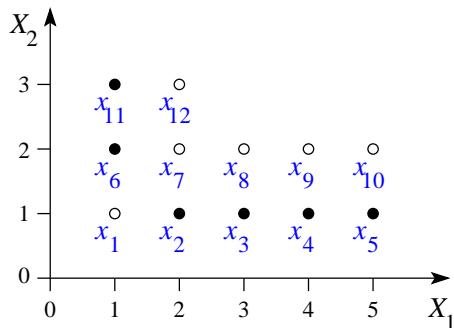
66.

(AdaBoost și non-învățabilitate γ -slabă: exemplificare pe date din \mathbb{R}^2) CMU, 2006 spring, Carlos Guestrin, final, pr. 3

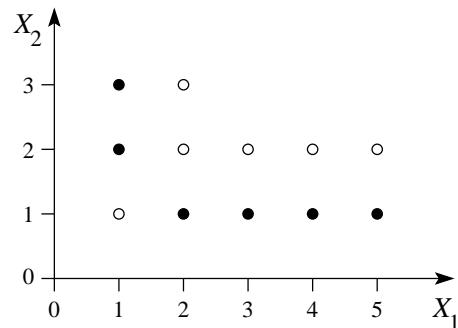
Considerând setul de date de antrenament din figura / figurile de mai jos, am dori să „învățăm“ un clasificator care să separe instanțele pozitive de instanțele negative, folosind algoritmul AdaBoost. Fiecare instanță x_i are o etichetă $y_i \in \{+1, -1\}$; eticheta $+1$ corespunde simbolului \bullet , iar eticheta -1 corespunde simbolului \circ .

Vom folosi ipoteze „slabe“, ale căror granițe de decizie (engl., decision boundaries) sunt paralele cu una din axele de coordonate, adică separatorul este fie vertical, fie orizontal. Puteți gândi aceste ipoteze „slabe“ ca fiind compași de decizie, pentru care pragurile de separare (engl., threshold splits) sunt situate fie pe axa X fie pe axa Y .

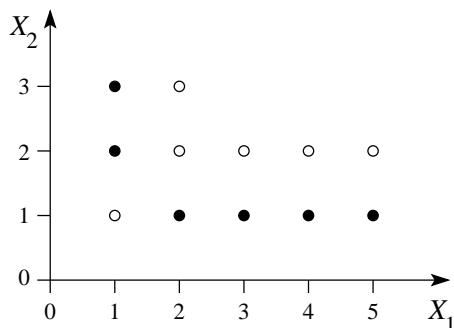
La fiecare iterare t , vom alege acea ipoteză „slabă“ h_t care maximizează acuratețea ponderată la antrenare în raport cu ponderile / probabilitățile curente D_t , adică alegem h_t care maximizează $\sum_i D_t(i) \cdot 1_{\{h_t(x_i)=y_i\}}$. Ca de obicei, se consideră că $h_t(x)$ ia valori doar în mulțimea $\{+1, -1\}$, după cum ipoteza h_t clasifică instanța x ca fiind pozitivă ori negativă.



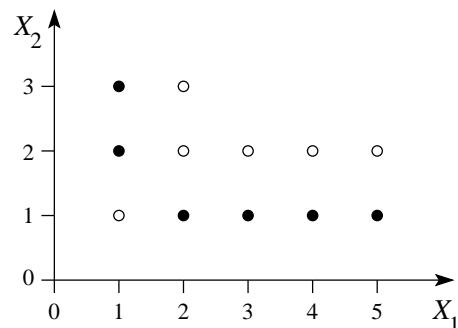
(a) Granița de decizie după prima iterare.



(b) Granița de decizie după a doua iterare.



(c) Exemplul / exemplele cu cea mai mică pondere.



(d) Exemplul / exemplele cu cea mai mare pondere.

- a. Desenați pe figura (a) granița de decizie determinată de algoritmul AdaBoost după prima iterare (adică, după ce a fost aleasă prima ipoteză „slabă“). Nu uitați să indicați ce parte a planului euclidian este clasificată cu '+' și respectiv '-'.

- b. Efectuăm acum cea de-a doua iterare de boosting. Desenați granițele de decizie ale ambelor ipoteze „slabe“ în figura (b). AdaBoost combină deciziile celor două ipoteze „slabe“. Indicați care sunt regiunile din plan [în care sunt situate punctele] clasificate de către boosting-ul cu 2 ipoteze „slabe“ cu '+' și respectiv cu '-'.

- c. La problema 22.v am arătat că $\alpha_i > \alpha_j \Leftrightarrow \varepsilon_i < \varepsilon_j$. Verificați că această

relație este într-adevăr satisfăcută pentru primele două iterații ale algoritmului AdaBoost pe setul de antrenament dat.

- d. Marcați în figura (c) exemplul / exemplele cu cea mai mică pondere / probabilitate (D_{t+1}) după primele două iterații de boosting ($t = 2$).
- e. Marcați în figura (d) exemplul / exemplele cu cea mai mare pondere / probabilitate (D_{t+1}) după primele două iterații de boosting ($t = 2$).
- f. Câte exemple de antrenament sunt clasificate eronat după primele două iterații de boosting?
- g. Folosind acest set de date și ipotezele „slabe“ menționate mai sus, va obține oare AdaBoost vreodată eroare la antrenare zero? Justificați în mod riguros, folosind o *implementare*.

67. (Algoritmul AdaBoost și învățabilitate empirică γ -slabă: o margine superioară pentru numărul de iterații de efectuat până la atingerea erorii empirice 0)

• CMU, 2010 fall, Aarti Singh, midterm, pr. 6.3
CMU, 2011 fall, Eric Xing, HW5, pr. 3.2.ab

Să presupunem că rulăm algoritmul AdaBoost pe m exemple de antrenament și, de asemenea, că la fiecare iterație (t) eroarea ponderată la antrenare (ε_t) produsă de ipoteza „slabă“ h_t este de cel mult $1/2 - \gamma$, unde $\gamma > 0$ nu depinde de t .

Determinați numărul de iterații (T) pe care trebuie să le execute algoritmul AdaBoost pentru ca ipoteza combinată H să fie consistentă cu cele m exemple de antrenament, adică să obțină eroare zero la antrenare. Trebuie să exprimați răspunsul doar în funcție de m și de γ .

Sugestie: Cât este eroarea la antrenare atunci când un singur exemplu este clasificat eronat?

68. (Algoritmul AdaBoost: margini de votare; aplicație)

· Liviu Ciortuz, 2020

În această problemă veți folosi setul de date de la problema 24.

- a. Pentru fiecare dintre iterațiile $t = 1$, $t = 2$ și $t = 3$, calculați $\text{Margin}_t(x_i)$, pentru $i = 1, \dots, 8$. (Pentru definiția noțiunii de margine de votare, vedeti enunțul problemei 27.)
- b. Verificați, utilizând distribuțiile probabiliste D_t care au fost calculate la problema 24, că într-adevăr are loc implicația $\text{Margin}_t(x_i) > \text{Margin}_t(x_j) \Rightarrow D_{t+1}(i) < D_{t+1}(j)$ — care a fost demonstrată pentru cazul general la problema 27.b — pentru cazul particular $t = 1$, $i = 1$ și $j = 4$, precum și pentru cazul $t = 2$, $i' = 4$ și $j' = 8$.

- c. Pentru fiecare dintre iterațiile $t = 1, t = 2$ și $t = 3$ în parte, reprezentați grafic: i. *distribuția empirică* a marginilor de votare care au fost calculate la punctul a , precum și ii. *distribuția empirică cumulativă* a acestor margini.⁵⁸⁶
- d. De asemenea, pentru fiecare dintre iterațiile $t = 1, t = 2$ și $t = 3$ în parte, calculați media marginilor [care au fost calculate la punctul a], în sensul definit/folosit la problema 28.

69.

(Clasificare de documente:
selecție de trăsături folosind algoritmul AdaBoost
cu compași de decizie pentru atrbute booleene)

- o CMU, 2007 spring, midterm, pr. 3
- MIT, 2003 fall, Tommi Jaakkola, final, pr. 3.2-4

Considerăm o problemă de clasificare de texte, în care un document X este reprezentat sub forma unui vector de trăsături binare relativ la cuvintele din dicționarul limbii în care este scris documentul respectiv. Din punct de vedere formal, putem scrie $X = [X_1, X_2, X_3, \dots, X_m]$, unde $X_j = 1$ în cazul în care cuvântul de pe poziția j din dicționar este prezent în documentul X , și 0 în caz contrar.

Vom considera acum că algoritmul AdaBoost folosește niște ipoteze „slabe“ desemnate formal ca perechi (j, y) unde j este indicele unui cuvânt de dicționar, iar y este clasa selectată, cu $y \in \{-1, +1\}$. Din punct de vedere *intuitiv*, acest lucru se poate exprima astfel: fiecare ipoteză slabă este [o pereche formată din] un cuvânt împreună cu o etichetă reprezentând clasa asociată. De exemplu, considerând cuvântul football și clasele {sports, non-sports}, atunci vom avea două ipoteze „slabe“ relative la acest cuvânt, și anume

- ipoteza (football, +): dacă documentul conține cuvântul football, [vom] prezice clasa sports; altfel, [vom] prezice clasa non-sports;
- ipoteza (football, -): dacă documentul conține cuvântul football, [vom] prezice clasa non-sports; altfel, [vom] prezice clasa sports.

a. Cât de multe ipoteze „slabe“ există în acest model?

Acest algoritm de tip boosting poate fi folosit pentru a face *selecția de trăsături*. Rulăm algoritmul și selectăm trăsăturile în *ordinea în care au fost identificate* de către algoritm.

b. Este oare posibil ca acest algoritm de boosting să selecteze o aceeași ipoteză „slabă“ mai mult decât de o singură dată? Justificați.

c. Să zicem că vrem să facem o ordonare (engl., ranking) a trăsăturilor în funcție de informația mutuală [individuală] relativă la variabila care reprezintă clasa (y), adică $IG(y; X_j)$. Va fi oare această ordonare mai „informativă“ / bună decât ordonarea produsă de către algoritmul AdaBoost? Justificați.

⁵⁸⁶Ca aplicație practică, vedeti problema CMU, 2007 spring, Carlos Guestrin, HW2, pr. 2.3, care în Companionul practic al culegerii de exerciții de învățare automată, <https://profs.info.uaic.ro/~ciortuz/ML.ex-book/implementation-exercises/ML.ex-book.Companion.pdf>, este problema 38.c.

70.

(Concepte din \mathbb{R} reprezentabile cu ajutorul combinațiilor liniare de compași de decizie)

*Liviu Ciortuz, 2021, pornind de la
• CMU, 2011 spring, Roni Rosenfeld, HW10, pr. 4.a*

Presupunem că avem o problemă de clasificare a unor instanțe pe axa reală: fiecare instanță x_i este un număr real, iar etichetele pe care urmează să le preziceți sunt binare, $y_i \in \{-1, +1\}$.

Pentru această problemă de clasificare, veți folosi *ansambluri*, adică niște combinații liniare de separatori / ipoteze „slabe“. (Atenție! NU trebuie să folosiți algoritmul AdaBoost!) Vă readucem aminte că un astfel de *clasificator* are forma următoare:

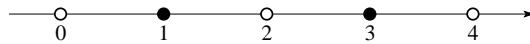
$$\hat{y} = \text{sign}\left(\sum_{t=1}^T \alpha_t h_t(x)\right), \quad (269)$$

unde \hat{y} este eticheta prezisă, $\text{sign}(x)$ este $+1$ dacă $x > 0$ și respectiv -1 în cazul contrar, α_t este o *pondere* (număr real strict pozitiv), iar $h_t(x)$ este predicția făcută de către ipoteza „slabă“ h_t . Fiecare h_t ia una dintre următoarele forme:

$$h_t(x; s, +) = \begin{cases} -1 & \text{dacă } x < s \\ +1 & \text{dacă } x \geq s \end{cases} \quad h_t(x; s, -) = \begin{cases} +1 & \text{dacă } x < s \\ -1 & \text{dacă } x \geq s, \end{cases}$$

pentru un anumit *prag de separare* (engl., split threshold) $s \in \mathbb{R}$.⁵⁸⁷

Considerăm următorul set de date, format din 5 instanțe situate pe axa reală:



a. Arătați că acest set de date (LC: sau, acest *concept*) este *reprezentabil* cu ajutorul unei combinații liniare care este formată din 4 ipoteze „slabe“. Așadar, vă cerem să identificați în mod explicit 4 ipoteze „slabe“ h_1, \dots, h_4 , precum și ponderile lor $\alpha_1, \dots, \alpha_4$, astfel încât ipoteza combinată $\text{sign}\left(\sum_{t=1}^T \alpha_t h_t(x)\right)$ să fie *consistentă* cu datele de mai sus.

Puneți în evidență cele 4 ipoteze „slabe“ h_1, \dots, h_4 pe desenul de mai sus, folosind praguri de separare s_1, \dots, s_4 precum și linii verticale, cărora le veți asocia o etichetare adecvată. Concret, veți desemna cu semnele $+$ și $-$ la stânga și la dreapta fiecărei linii verticale — asociată unui anumit prag s_j — zonele de decizie determinate de această ipoteză „slabă“.

Observație: Înănd cont de modul în care a fost definită funcția *sign* mai sus, *regula de predicție* (269) va trata cazurile de „paritate“ etichetând cu -1 instanțele pentru care suma ponderată (engl., weighted sum) a predicțiilor făcute de ipotezele „slabe“ este 0. Acest mod de tratare a cazurilor de „paritate“ este [foarte] util pentru rezolvarea punctului a.

⁵⁸⁷Remarcați faptul că definiția dată aici pentru noțiunea de *ipoteză „slabă“* corespunde bine-cunoscutului *compas de decizie* (engl., decision stump), pentru cazul particular când datele sunt din \mathbb{R} .

Indicație: Pentru a justifica ușor consistența ansamblului aleas de dumneavoastră cu datele din figura de mai sus, vă cerem să completați un tabel similar cu cel pe care-l folosim la algoritmul AdaBoost atunci când facem calculul erorii la antrenare produse de ipoteza combinată $\text{sign}\left(\sum_{t=1}^T \alpha_t h_t(x)\right)$.

α_t	h_t	0	1	2	3	4
$\alpha_1 =$	$h_1(x_i)$					
$\alpha_2 =$	$h_2(x_i)$					
$\alpha_3 =$	$h_3(x_i)$					
$\alpha_4 =$	$h_4(x_i)$					
	$\text{sign}\left(\sum_{t=1}^4 \alpha_t h_t(x_i)\right)$					

b. Demonstrați că setul de date de mai sus NU este reprezentabil cu mai puțin de 4 ipoteze „slabe“.

c. Generalizați rezultatul obținut la punctul a, referindu-vă la posibilitatea (sau, dimpotrivă, imposibilitatea) de a reprezenta — folosind combinații liniare de ipoteze „slabe“ aşa cum au fost definite mai sus — dataset-uri arbitrarе [formate din instanțe situate] pe axa reală.

71.

(O clasă de concepte învățabile în sens empiric γ -slab cu ajutorul compașilor de decizie:

seturile de instanțe din \mathbb{R} , care sunt etichetate în mod consistent)

Liviu Ciortuz, 2023, pornind de la

■ □ ○ Stanford, 2016 fall, A. Ng, J. Duchi, HW2, pr. 6.abc

Introducere: În această problemă, la punctele *a-c* vom demonstra câteva proprietăți ale compașilor de decizie pentru dataseturi consistente de pe axa reală. Ca o consecință a acestor proprietăți, la punctul *d* vom arăta că dataseturile consistente de pe axa reală constituie o clasă de concepte învățabile în sens empiric γ -slab cu ajutorul compașilor de decizie.

Fie un set de date etichetate pe axa reală, $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$, care este consistent (adică, pentru orice $x_i = x_j$ cu $i \neq j$ rezultă că necesită că $y_i = y_j$). Fără a reduce generalitatea, vom presupune că $x_1 \leq x_2 \leq \dots \leq x_m$. Ba chiar mai mult, pentru a facilita redactarea demonstrațiilor corespunzătoare afirmațiilor de mai jos, vom presupune că $x_1 < x_2 < \dots < x_m$.

Unui prag $s \in \mathbb{R}$ îi putem asocia doi compași de decizie, pe care aici îi vom defini folosind următoarele funcții:

$$\phi_{s,+}(x) = \begin{cases} 1 & \text{dacă } x \geq s \\ -1 & \text{dacă } x < s \end{cases} \quad \text{și respectiv} \quad \phi_{s,-}(x) = \begin{cases} -1 & \text{dacă } x \geq s \\ 1 & \text{dacă } x < s. \end{cases}$$

Prin urmare, $\phi_{s,+}(x) = -\phi_{s,-}(x)$ pentru orice $x \in \mathbb{R}$.

Fie $p \stackrel{\text{not.}}{=} (p_1, \dots, p_m)$ o distribuție de probabilitate discretă asignată celor m instanțe (în acest context considerate fără etichete): $x_1 \mapsto p_1, \dots, x_m \mapsto p_m$.

a. Fie un prag oarecare $s \in \mathbb{R}$, fixat. Vă reamintim faptul că erorile ponderate la antrenare produse de către compașii de decizie $\phi_{s,+}$ și $\phi_{s,-}$ în raport cu distribuția de probabilitate p sunt definite astfel:

$$\text{error}_p(\phi_{s,+}) \stackrel{\text{def.}}{=} \sum_{i=1}^m p_i \cdot 1_{\{y_i \neq \phi_{s,+}(x_i)\}}, \quad \text{respectiv} \quad \text{error}_p(\phi_{s,-}) \stackrel{\text{def.}}{=} \sum_{i=1}^m p_i \cdot 1_{\{y_i \neq \phi_{s,-}(x_i)\}},$$

unde simbolul $1_{\{\cdot\}}$ desemnează binecunoscuta *funcție-indicator*.

Pentru pragul $s \in \mathbb{R}$ considerat mai sus (ales arbitrar, dar fixat) vom defini $m_0(s) \in \{0, 1, \dots, m\}$ astfel:

$$m_0(s) = \begin{cases} 0, & \text{dacă } s < x_1 \\ m, & \text{dacă } s \geq x_m \\ i, & \text{cu proprietatea } s \in [x_i, x_{i+1}) \text{ unde } i \in \{1, \dots, m-1\}, \\ & \text{dacă } x_1 \leq s < x_m. \end{cases}$$

În cele ce urmează vom considera funcția $f : \{0, 1, \dots, m\}$ definită prin expresia următoare:⁵⁸⁸

$$f(j) \stackrel{\text{def.}}{=} \sum_{i=1}^j y_i p_i - \sum_{i=j+1}^m y_i p_i,$$

cu observația că acele sume \sum_i pentru care indicele i ia valori în mulțimea vidă vor fi considerate prin definiție ca fiind nule (adică, egale cu 0). (De exemplu, $\sum_{i=1}^0 a_i = 0$ și $\sum_{i=m+1}^m a_i = 0$.)

Vă cerem să demonstrați că

$$\text{error}_p(\phi_{s,+}) = \frac{1}{2} + \frac{1}{2} f(m_0(s)) \quad \text{și} \quad \text{error}_p(\phi_{s,-}) = \frac{1}{2} - \frac{1}{2} f(m_0(s)). \quad (270)$$

b. Demonstrați că pentru orice $j \in \{1, \dots, m\}$ are loc proprietatea

$$|f(j) - f(j-1)| = 2p_j \quad (271)$$

și, în consecință, există cel puțin un $j^* \in \{1, \dots, m\}$ astfel încât

$$|f(j^*)| \geq \frac{1}{m} \quad \text{sau} \quad |f(j^*-1)| \geq \frac{1}{m}. \quad (272)$$

Observație: Este posibil ca ambele inegalități (272) să aibă loc pentru un același indice j^* .

c. Definim $k \in \{0, 1, \dots, m\}$ astfel:

Dacă în relația (272) prima inegalitate are loc, atunci $k = j^*$. Dacă cea de-a doua inegalitate are loc, atunci $k = j^* - 1$. Dacă ambele inegalități au loc, atunci putem alege k ca fiind fie j^* fie $j^* - 1$, după cum dorim.

Demonstrați că

$$\begin{aligned} &\text{dacă } f(k) \leq -\frac{1}{m}, \text{ atunci } \text{error}_p(\phi_{s,+}) \leq \frac{1}{2} \left(1 - \frac{1}{m}\right), \text{ iar} \\ &\text{dacă } f(k) \geq \frac{1}{m}, \text{ atunci } \text{error}_p(\phi_{s,-}) \leq \frac{1}{2} \left(1 - \frac{1}{m}\right), \end{aligned}$$

pentru orice $s \in [x_k, x_{k+1})$ în cazul în care $k \in \{1, \dots, m-1\}$, respectiv pentru orice $s < x_1$ în cazul în care $k = 0$ și, în sfârșit, pentru orice $s \geq x_m$ în cazul în care $k = m$.⁵⁸⁹

⁵⁸⁸ Pentru a fi și mai pedantă, ar trebui să folosim notația f_p , ca să arătăm că f depinde de distribuția p , însă nu vrem să complicăm prea mult formalismul.

⁵⁸⁹ *Observație:* În cele ce urmează, vom putea ignora cazul $s \geq x_m$, fiindcă $\text{error}_p(\phi_{s,+}) = \text{error}_p(\phi_{s',+})$ și $\text{error}_p(\phi_{s,-}) = \text{error}_p(\phi_{s',-})$ pentru orice $s' < x_1$ și, în consecință, la punctul d ne vom putea limita la a alege doar compași de decizie exterioiri situați la stânga lui x_1 .

d. [Corolar]

Demonstrați că dacă rulăm algoritmul AdaBoost pe un astfel de dataset S , atunci putem alege la fiecare iterare t un compas de decizie h_t astfel încât eroarea sa ponderată la antrenare să fie de cel mult $\frac{1}{2} \left(1 - \frac{1}{m}\right)$. În consecință, datasetul S este *învățabil în sens empiric γ -slab*, cu garanția de învățabilitate $\gamma = \frac{1}{2m}$.

e. Puteți indica o margine superioară (engl., upper bound) pentru numărul de compași de decizie necesari pentru ca algoritmul AdaBoost să obțină eroare la antrenare 0 pe un astfel de set de date de antrenament (S)?

72.

(Algoritmul AdaBoost poate să producă overfitting)

prelucrare de Liviu Ciortuz, 2021, după
 * CMU, 2011 fall, Eric Xing, HW5, pr. 3.2.cd

Spunem că un sistem de clasificare automată este supra-specializat (sau, că produce overfitting) atunci când acuratețea sa este din ce în ce mai bună la antrenare, dar din ce în ce mai slabă la testare. Așadar, el ajunge să modeleze / explice chiar și detalii accidentale („zgomotele“) din datele de antrenament, în loc să generalizeze, făcând abstracție de ele.

Formalizând această noțiune, Tom Mitchell afirmă în cartea sa *Machine Learning* (la pag. 67) că o ipoteză h produsă de către un algoritm de clasificare automată este supra-specializată pe datele de antrenament dacă există o altă ipoteză h' astfel încât

$$\text{error}_{\text{train}}(h) < \text{error}_{\text{train}}(h'), \text{ dar } \text{error}_{\text{test}}(h) > \text{error}_{\text{test}}(h'). \quad (273)$$

În general, algoritmul AdaBoost nu suferă de overfitting. Există totuși situații în care se întâmplă ca el să producă overfitting, și anume atunci când datele de antrenare și de testare sunt foarte mixate (altfel spus, ele conțin mult „zgomot“).⁵⁹⁰

a. Creați un set de date de antrenament și un set de date de test pe care să arătați că într-adevăr algoritmul AdaBoost produce overfitting (conform relației (273)). Faceți graficele funcțiilor de eroare la antrenare și respectiv la testare în raport cu numărul (T) de interații executate de AdaBoost.⁵⁹¹

b. Identificați o strategie prin aplicarea căreia să poată fi contracararat / limitat fenomenul de overfitting la algoritmul AdaBoost.

⁵⁹⁰CMU, 2014 fall, W. Cohen, Z. Bar-Joseph, midterm, pr. 11.3.

⁵⁹¹Indicație: Există un applet interesant scris de Yoav Freund (<http://cseweb.ucsd.edu/~yfreund/adaboost/>), care vă permite să vă creați propriile seturi de date de antrenament și de testare în planul euclidian, să antrenați AdaBoost iar apoi să-l testați. Puteți să faceți apoi un screen-shot care să includă atât datele de antrenament și datele de test cât și graficele curbelor de eroare la antrenare și respectiv la testare.

Alternativ, puteți găsi un set de date pentru care zonele de decizie determinate de către AdaBoost să fie deja cunoscute. (Vedeți de exemplu pr. 24.) Apoi plasați punctele de test în aşa fel încât o parte dintre ele să contravină semnului zonei respective (aşa cum fusese el stabilit de către AdaBoost în urma antrenării).

73. (Algoritmul AdaBoost folosind “confidence rated classifiers” în locul clasificatorilor „slabi“ simpli)

• o MIT, 2001 fall, Tommi Jaakkola, HW3, pr. 1.1-3

Fie $D = \{(x_1, y_1), \dots, (x_m, y_m)\}$ un set de exemple de antrenament, cu $x_i \in \mathbb{R}^d$ și $y_i \in \{-1, +1\}$. Ca și în formularea clasică a algoritmului AdaBoost,⁵⁹² vom nota „ponderile“ / probabilitățile asociate exemplelor de antrenament la începutul iterăției k a cestui algoritm cu $D_k(1), \dots, D_k(m)$, unde $\sum_{i=1}^m D_k(i) = 1$. De asemenea, presupunem că există o metodă [necunoscută nouă] care produce o ipoteză „slabă“ h ca răspuns la un set de antrenament ponderat. Această ipoteză „slabă“ trebuie să fie un clasificator care să se comporte măcar un pic mai bine decât datul cu banul (engl., random guessing) pe setul de antrenament ponderat pe care a fost antrenat. Nu facem nicio presupunere cu privire la cât de bine se comportă acest clasificator „slab“ pe eventuale alte seturi de antrenament ponderate.

Spre deosebire de cazul algoritmului AdaBoost clasic, aici vom considera că ipoteza h produce ca output valori reale. (Din această cauză, astfel de ipoteze se numesc în limba engleză *confidence rated classifiers*.) Semnul output-ului $h(x)$ indică eticheta (± 1) , iar modulul output-ului specifică gradul de „încredere“ (engl., confidence) în decizia luată de clasificator. De exemplu, output-ul $h(x) = 10$ ar putea fi interpretat ca fiind o predicție a etichetei $+1$ cu o încredere destul de mare.

Pornind de la distribuția uniformă $D_1(i) = 1/m$, AdaBoost generează o secvență de ipoteze h_1, \dots, h_t , fiecare ipoteză h_k fiind antrenată utilizând un anumit set de ponderi asociate exemplelor. După ce, în cadrul iterăției k , a fost generată o ipoteză $h_k(x)$ ca „răspuns“ la ponderile $D_k(i)$, ponderile asociate exemplelor vor fi actualizate folosind relația următoare:

$$D_{k+1}(i) = c \cdot D_k(i) \exp(-\hat{\alpha}_k y_i h_k(x_i)), \quad i = 1, \dots, m, \quad (274)$$

unde c este *constanta de normalizare* care asigură faptul că $\sum_{i=1}^n D_{k+1}(i) = 1$, iar $\hat{\alpha}_k$ este *votul* asignat noii componente a clasificatorului, h_k . Vrem să asignăm aceste voturi astfel încât h_k să satisfacă relația următoare:

$$\sum_{i=1}^m D_{k+1}(i) y_i h_k(x_i) = 0, \quad (275)$$

ceea ce înseamnă că la iterăția $k+1$ suma produselor dintre ponderi și încrederi ($D_{k+1}(i)h_k(x_i)$) pentru exemplele corect clasificate (adică, acei x_i pentru care $y_i h_k(x_i) > 0$) de către ipoteza „slabă“ h_k este egală cu suma produselor dintre ponderi și încrederi pentru exemplele incorect clasificate ($y_i h_k(x_i) < 0$).

a. Arătați că dacă ipoteza „slabă“ h_k generează doar output-uri binare (± 1) , atunci *condiția* (275) revine la a cere ca eroarea ponderată la antrenare a lui h_k calculată în raport cu noile ponderi ($D_{k+1}(i)$) să fie exact 0.5.⁵⁹³

Comentariu: Întrucât ipoteza „slabă“ h_t poate genera ca output valori reale, situația este sensibil diferită în cazul de față. Putem vedea *condiția* (275) ca

⁵⁹²Vedeți problema 22.

⁵⁹³LC: Prin urmare, relația (275) reprezintă o generalizare a rezultatului care a fost demonstrat la problema 22.vi.

pe un mod de a *de-corela* predicțiile $h_k(x_i)$ de etichetele y_i , ținând cont de noile ponderi $D_{k+1}(i)$.

b. Arătați că dacă asignăm lui $\hat{\alpha}_k$ acea valoare a argumentului α pentru care se atinge minimul funcției⁵⁹⁴

$$J_k(\alpha) = \ln \left(\sum_{i=1}^m D_k(i) \cdot \exp(-\alpha y_i h_k(x_i)) \right), \quad (276)$$

atunci în mod necesar *condiția de de-corelare* (275) este satisfăcută pentru noile ponderi.

Sugestie: Nu este cazul să încercați să calculați $\hat{\alpha}_k$. În schimb, egalați derivata funcției $J_k(\alpha)$ din relația (276) cu zero și folosiți ecuația care se obține (și pe care $\hat{\alpha}_k$ trebuie să o satisfacă), pentru a arăta că relația (275) este satisfăcută.

c. *Comentariu:* Rezultatul care a fost obținut la punctul b pare să fie un pic suspicios... ca și când ar exista o funcție obiectiv pe care algoritmul nostru de tip boosting ar urmări să o minimizeze la fiecare iterare. La acest punct veți arăta că într-adevăr aşa este.

Fie f_t ipoteza combinată care rezultă în urma efectuării a t iterării de boosting:

$$f_t(x) = \hat{\alpha}_1 h_1(x) + \dots + \hat{\alpha}_t h_t(x).$$

Va trebui să arătați că funcția

$$J(f_t) = \ln \left(\sum_{i=1}^m \exp(-y_i f_t(x_i)) \right)$$

joacă rolul de *funcție obiectiv* pentru noul nostru algoritm de boosting. Cu alte cuvinte, veți arăta că de fiecare dată când adăugăm o nouă componentă (h) la clasificatorul combinat $f_t(x)$, valoarea funcției obiectiv descrește.

Așadar, demonstrați că $J(f_t) \geq J(f_{t+1})$, unde

$$f_{t+1}(x) = f_t(x) + \hat{\alpha}_{t+1} h_{t+1}(x),$$

$h_{t+1}(x)$ este noua ipoteză „slabă“, iar $\hat{\alpha}_{t+1}$ a fost ales (adică, a fost optimizat) după cum am indicat mai sus.

Sugestie: Calculați ponderile $D_t(i)$ într-un mod similar cu cel în care am procedat la problema 23.a și țineți cont de faptul că $\hat{\alpha}_{t+1}$ minimizează funcția $J_k(\alpha)$ care a fost definită prin relația (276).

74.

(Algoritmul AdaBoost, ca instanță a unui algoritm mai general)

• o CMU, 2012 fall, E. Xing, A. Singh, HW4, pr. 3.ab

Considerăm următorul algoritm de clasificare *etapizat, aditiv* (engl., [forward] *stagewise additive modelling*), A:

⁵⁹⁴LC: Observați că $J_k(\alpha)$ este un log-cost ponderat cu ajutorul probabilităților D_k . Funcția logaritm natural (\ln) este strict crescătoare, deci a minimiza un cost este echivalent cu a minimiza log-costul corespunzător.

Intrare: $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$, T , \mathcal{H} , ϕ , unde

T este numărul de iterații de executat,

\mathcal{H} este o mulțime de „ipoteze“,

$\phi(y, y')$ este o funcție de „cost“ / „risc“;

Procedură:

Initialize the classifier by taking $f_0(x) = 0$ (the constant function 0)
for $t = 1$ to T do:

1. Compute

$$(h_t, \alpha_t) = \arg \min_{\alpha \in \mathbb{R}, h \in \mathcal{H}} J_t(\alpha, h)$$

$$\text{where } J_t(\alpha, h) \stackrel{\text{not.}}{=} \frac{1}{m} \sum_{i=1}^m \phi(y_i, f_{t-1}(x_i) + \alpha h(x_i))$$

2. Update the classifier

$$f_t(x) = f_{t-1}(x) + \alpha_t h_t(x)$$

end for

return the classifier $\text{sign}(f_T(x))$

Intuitiv, la fiecare pas algoritmul A adaugă în mod *greedy* câte o ipoteză $h \in \mathcal{H}$ la ipoteza curentă pentru a minimiza costul / riscul definit de funcția ϕ .

a. Cum ar trebui să definim $\phi(y, y')$ pentru ca algoritmul A să fie echivalent cu algoritmul AdaBoost?

b. Demonstrați că într-adevăr, atunci când ϕ este funcția pe care ați indicat-o la punctul a, algoritmul AdaBoost este echivalent cu algoritmul A.

Sugestie: Considerând α fixat (la o valoare pozitivă, arbitrar aleasă), determinați acel $h_t \in \mathcal{H}$ care va minimiza costul / riscul $J_t(\alpha, h)$. (Atenție: h_t nu este o funcție de α .) Apoi, considerând această ipoteză h_t fixată, găsiți acea valoare α_t (a lui α) care va minimiza costul / riscul $J_t(\alpha, h)$. Gândiți-vă de asemenea care este relația dintre probabilitățile $D_t(i)$ din algoritmul AdaBoost și algoritmul A.

75. (Algoritmul AdaBoost generalizat [în raport cu funcția de cost]: aplicare; ilustrarea echivalenței care a fost demonstrată la pr. 29.a)

□ • · Liviu Ciortuz, 2020

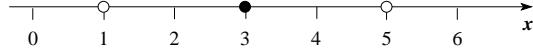
La problema 29 am prezentat algoritmul AdaBoost generalizat, care poate fi parametrizat cu orice funcție de cost / „loss“ / pierdere. La punctul a de la acea problemă am demonstrat că, la fiecare iteratie t a algoritmului AdaBoost generalizat, identificarea celui mai bun compas de decizie (adică, minimizarea erorii ponderate la antrenare, ε_t) este echivalentă cu minimizarea criteriului

$$J_t(\alpha, \theta) = \frac{1}{m} \sum_{i=1}^m \text{Loss}(y_i f_{t-1}(x_i) + y_i \alpha h(x_i; \theta))$$

în raport cu parametrul θ , unde Loss este funcția de cost cu care lucrăm.

Ne propunem ca — în cea mai mare parte a acestui exercițiu — să ilustrăm în mod *practic* acest lucru, pe date concrete. Așadar, vom verifica faptul că într-adevăr are loc echivalența despre care am vorbit mai sus atunci când i. aplicăm algoritmul AdaBoost generalizat folosind funcția de cost [negativ] exponențială pe datele de la problema 25 (ne referim la prima soluție dată

acolo),⁵⁹⁵ respectiv *ii.* atunci când aplicăm algoritmul AdaBoost generalizat folosind funcția de cost logistică (la punctul *c*) pe același set de date:



Atenție! Vom folosi compași de decizie, *inclusiv* compași cu prag *exterior* (engl., outside threshold) în raport cu datele de antrenament.

Prin urmare, ne vom uita la *valorile* pe care le ia *expresia* (245) în cursul rulării acestui algoritm pe datele pe care tocmai le-am menționat, adică derivata parțială a funcției $J_t(\alpha, \theta)$ calculată pentru $\alpha = 0$ și pentru θ (deci o funcție de parametrul θ):

$$\begin{aligned}\frac{\partial}{\partial \alpha} J_t(\alpha, \theta)|_{\alpha=0} &= \frac{1}{m} \sum_{i=1}^m \frac{\partial}{\partial \alpha} \text{Loss}(y_i f_{t-1}(x_i) + y_i \alpha h(x_i; \theta))|_{\alpha=0} \\ &= \frac{1}{m} \sum_{i=1}^m dL(y_i f_{t-1}(x_i)) y_i h(x_i; \theta),\end{aligned}$$

unde $dL(z) \stackrel{\text{not.}}{=} \frac{\partial}{\partial z} \text{Loss}(z)$.

a. Arătați că $\frac{\partial}{\partial \alpha} J_t(\alpha, \theta)|_{\alpha=0}$ are la iterată $t = 1$ expresia

$\frac{\partial}{\partial \alpha} J_1(\alpha, \theta) _{\alpha=0} = \begin{cases} -1/3 & \forall \theta \in (-\infty, 1) \cup [3, 5) \\ +1/3 & \forall \theta \in [1, 3) \cup [5, +\infty) \end{cases}$ <p style="margin-top: 10px;">dacă semnele asociate compasului de decizie (h_1) sunt $+ -$;</p>	$\frac{\partial}{\partial \alpha} J_1(\alpha, \theta) _{\alpha=0} = \begin{cases} +1/3 & \forall \theta \in (-\infty, 1) \cup [3, 5) \\ -1/3 & \forall \theta \in [1, 3) \cup [5, +\infty) \end{cases}$ <p style="margin-top: 10px;">dacă semnele asociate compasului de decizie (h_1) sunt $- +$.</p>
--	--

Notă: Exact acestei expresii îi corespunde graficul care a fost prezentat la pagina 558.

Verificați apoi că $D_2(i) = \tilde{W}_i^{(2)}$ pentru $i \in \{1, 2, 3\}$.⁵⁹⁶

Observații:

1. Vă readucem aminte că parametrul θ desemează *pe lângă* pragul compasului de decizie și semnele asociate acestui compas de decizie. (Aceste semne sunt $+|-$ pentru graficul din stânga și respectiv $-|+$ pentru graficul din dreapta.)
2. Valorile lui θ pentru care se obține minimul lui $\frac{\partial}{\partial \alpha} J_t(\alpha, \theta)|_{\alpha=0}$ sunt eligibile de către algoritmul AdaBoost la iterată curentă.⁵⁹⁷ Astfel, la iterată $t = 1$, după cum se observă în graficul de la pagina 558, putem alege $\theta \in (-\infty, 1) \cup [3, 5)$ cu semnele asociate $+|-$ sau, alternativ, $\theta \in [1, 3) \cup [5, +\infty)$ cu semnele asociate $-|+$. Observați că una dintre aceste posibilități, și anume, pragul $\theta = 0$ cu semnele asociate $+|-$, a fost ales la rezolvarea problemei 25, și anume prima soluție.

⁵⁹⁵ La punctul *b* al problemei 29 am demonstrat (din punct de vedere *teoretic*) că atunci când se lucrează folosind funcția de cost [negativ] exponențială, respectiva „instanță“ a algoritmului AdaBoost generalizat coincide cu algoritmul AdaBoost prezentat la problema 22.

⁵⁹⁶ Vedeți rezolvarea problemei 25, prima soluție.

⁵⁹⁷ Mai general, se poate alege pentru θ o valoare pentru care expresia $\frac{\partial}{\partial \alpha} J_t(\alpha, \theta)|_{\alpha=0}$ este negativă. Aici însă, ca și la problema 25, nu folosim această posibilitate.

b. Procedând similar, calculați expresiile $\frac{\partial}{\partial \alpha} J_t(\alpha, \theta)|_{\alpha=0}$ pentru iterațiile $t = 2$ și $t = 3$ și apoi obțineți graficele corespunzătoare.

Notă: În fiecare din cele două cazuri veți presupune că la iterația / iterațiile precedente s-au selectat compașii de decizie exact ca la rezolvarea problemei 25, prima soluție.

c. Aplicați algoritmul AdaBoost generalizat folosind funcția de cost logistică pe datele de la problema 25; executați 3 iterații, utilizând inclusiv compași de decizie exteriori. Pentru calcularea voturilor α_t , veți folosi formula care a fost dedusă la pr. 29.g, adică $\alpha_t = \ln \sqrt{\frac{1 - \varepsilon_t}{\varepsilon_t}}$. La final veți compara rezultatele obținute aici cu cele care au fost prezentate la rezolvarea problemei 25.

76. **(Algoritmul AdaBoost multi-class: fundamentare teoretică)**

□ · *Liviu Ciortuz, 2023, pornind de la articolul Multi-class AdaBoost, by J. Zhu, H. Zou, S. Rosset, T. Hastie, 2006*⁵⁹⁸

În această problemă vom arăta cum este posibil să generalizăm algoritmul AdaBoost clasic, în aşa fel încât să realizăm clasificare n -ară⁵⁹⁹. Această generalizare va implica adaptarea funcției de cost / pierdere negativ exponențiale⁶⁰⁰ — pe care algoritmul AdaBoost clasic o folosește pentru a realiza clasificare binară — la cazul clasificării n -are:

$$L(y, f) = \exp \left(-\frac{1}{K} (y^{(1)} f^{(1)} + \dots + y^{(K)} f^{(K)}) \right) = \exp \left(-\frac{1}{K} y^\top f \right),$$

unde $f \stackrel{\text{not.}}{=} (f^{(1)}, \dots, f^{(K)})^\top$, $f^{(k)}$ corespunde clasei k , iar $f^{(1)} + \dots + f^{(K)} = 0$ (aceasta este numită *restricția de simetrie*),⁶⁰¹ și $y \stackrel{\text{not.}}{=} (y^{(1)}, \dots, y^{(K)})^\top$, cu

$$y^{(k)} = \begin{cases} 1, & \text{când } y \text{ codifică clasa } k, \\ -\frac{1}{K-1}, & \text{în cazul contrar.} \end{cases}$$

Noul algoritm va fi formulat în maniera optimizării secvențiale.⁶⁰²

0. Input $(x_1, y_1), \dots, (x_n, y_n)$, with $x_i \in \mathbb{R}^p$ and $y_i \in \mathcal{Y}$, for $i = 1, \dots, n$.
1. Initialize $f_0(x) = 0$.
2. For $m = 1$ to M :
 - (a) Compute $(\beta_m, g_m) = \arg \min_{\beta, g} \sum_{i=1}^n L(y_i, f_{m-1}(x_i) + \beta g(x_i))$;
 - (b) Set $f_m(x) = f_{m-1}(x_i) + \beta_m g_m(x)$.
3. Output $\operatorname{argmax}_{k \in \{1, \dots, K\}} (f_m^{(1)}(x), \dots, f_m^{(K)}(x))$.

⁵⁹⁸Acest articol a fost publicat într-o versiune revizuită și extinsă în revista *Statistics and its Interface*, 2009, vol. 2, pag. 349-360.

⁵⁹⁹Engl., multi-class classification.

⁶⁰⁰Vedeți problemele 23 și 26.

⁶⁰¹Observați că în cazul $K = 2$ acest cost negativ exponențial coincide cu costul negativ exponențial de la clasificare binară.

⁶⁰²Acestă strategie este numită “stagewise additive modeling” în terminologia [de limbă engleză] folosită de autorii acestui algoritm, care sunt statisticieni.

Mulțimea $\mathcal{Y} \subset \mathbb{R}^K$ este constituită din „codificările“ claselor / etichetelor $1, \dots, K$:

$$\begin{aligned}\mathcal{Y} = & \left\{ \left(1, -\frac{1}{K-1}, -\frac{1}{K-1}, \dots, -\frac{1}{K-1} \right)^\top, \right. \\ & \left(-\frac{1}{K-1}, 1, -\frac{1}{K-1}, \dots, -\frac{1}{K-1} \right)^\top, \\ & \dots \\ & \left. \left(-\frac{1}{K-1}, -\frac{1}{K-1}, \dots, -\frac{1}{K-1}, 1 \right)^\top \right\}.\end{aligned}$$

Funcțiile g_m (deci și g) sunt definite pe \mathbb{R}^p și iau valori în mulțimea \mathcal{Y} .

Observație: Pasul (2a) din acest algoritm se poate scrie în mod echivalent astfel:

$$(\beta_m, g_m) = \arg \min_{\beta, g} \sum_{i=1}^n w_i \exp \left(-\frac{1}{K} \beta y_i^\top g(x_i) \right), \quad (277)$$

unde

$$w_i = \exp \left(-\frac{1}{K} y_i^\top f_{m-1}(x_i) \right)$$

sunt ponderile nenormalizate care au fost calculate la iterată $m-1$ pentru instanțele / „observațiile“ x_i .

a. Arătați că⁶⁰³

i. fixând β , a minimiza funcția obiectiv din partea dreaptă a relației (277) revine la a afla funcția g care minimizează eroarea ponderată $\sum_{i=1}^n w_i \cdot 1_{\{y_i \neq g(x_i)\}}$.⁶⁰⁴

ii. fixând g , a minimiza funcția obiectiv din partea dreaptă a relației (277) revine la a seta β la valoarea

$$\beta_m = \frac{(K-1)^2}{K} \underbrace{\left[\ln \frac{1-\varepsilon_m}{\varepsilon_m} + \ln(K-1) \right]}_{\text{not.: } \alpha_m}, \quad (278)$$

unde

$$\varepsilon_m = \sum_{i=1}^n w_i \cdot \underbrace{1_{\{y_i \neq g_m(x_i)\}}}_{1_{\{c_i \neq T_m(x_i)\}}} / \sum_{i=1}^n w_i.$$

b. Vom (re)nota cu $w_{i,m}$ ponderea nenormalizată calculată pentru instanța / „observația“ x_i la iterată m , Arătați că regula de actualizare a acestor ponderi nenormalizate w se poate scrie astfel:

$$w_{i,m} = w_{i,m-1} \cdot \exp \left(-\frac{1}{K} \beta_m y_i^\top g_m(x_i) \right) \quad (279)$$

$$= \begin{cases} w_{i,m-1} e^{-\frac{K-1}{K} \alpha_m} & \text{dacă } y_i = g_m(x_i), \\ w_{i,m-1} e^{\frac{1}{K} \alpha_m} & \text{dacă } y_i \neq g_m(x_i) \end{cases} \quad (280)$$

$$= w_{i,m-1} \cdot \exp \left(-\frac{(K-1)^2}{K^2} \alpha_m y_i^\top g_m(x_i) \right). \quad (281)$$

⁶⁰³Punctele i și ii care urmează să fie formulate [în enunț] constituie *Lema 1* din articolul *Multi-class AdaBoost* de J. Zhu, H. Zou, S. Rosset, T. Hastie, 2009.

⁶⁰⁴Corespondentul funcției g în algoritmul SAMME care va fi introdus la punctul c este ipoteza „slabă“ / arborele T . Erorii ponderate $\sum_{i=1}^n w_i \cdot 1_{\{y_i \neq g(x_i)\}}$ îi va corespunde în SAMME expresia $\sum_{i=1}^n w_i \cdot 1_{\{c_i \neq T(x_i)\}}$.

c. Considerăm următorul pseudo-cod (algoritmul SAMME):

0. Input $(x_1, c_1), \dots, (x_n, c_n)$, with $x_i \in \mathbb{R}^p$ and $c_i \in \{1, \dots, K\}$, for $i = 1, \dots, n$.
1. Initialize the observation weights $w_i = 1/n$, $i = 1, 2, \dots, n$.
2. For $m = 1$ to M :
 - (a) Using weights w_i , find the „best“ classifier $T_m(x)$ of the training data w.r.t. the weighted training error, which is defined as follows:
 - (b) $\varepsilon_m = \sum_{i=1}^n w_i \cdot 1_{\{c_i \neq T_m(x_i)\}}$;
 - (c) Compute $\alpha_m = \ln \frac{1 - \varepsilon_m}{\varepsilon_m} + \ln(K - 1)$;
 - (d) Set $w_i \leftarrow w_i \cdot \exp(\alpha_m \cdot 1_{\{c_i \neq T_m(x_i)\}})$
for $i = 1, \dots, n$;
 - (e) Re-normalize w_i .
3. Output $C(x) = \operatorname{argmax}_{k \in \{1, \dots, K\}} \sum_{j=1}^m \alpha_j \cdot 1_{\{T_j(x)=k\}}$.

Veți arăta că algoritmul SAMME este echivalent cu precedentul algoritm (de tip *forward stagewise additive modeling*) atunci când acesta utilizează funcția de cost / pierdere negativ exponentială multi-class.⁶⁰⁵

Concret, arătați că

- i. după normalizare, ponderile (281) coincid cu ponderile obținute la pasul (2e) din algoritm SAMME;
- ii. regulile de decizie ale celor doi algoritmi coincid, adică are loc următoarea egalitate:

$$\operatorname{argmax}_k(f_m^{(1)}(x), \dots, f_m^{(K)}(x)) = \operatorname{argmax}_k \sum_{m=1}^M \alpha_m \cdot 1_{\{T_m(x)=k\}}.$$

Sugestie: Pentru a demonstra cerințele de la acest punct al problemei, țineți cont de faptul că există o corespondență bijectivă între funcțiile $g : \mathbb{R}^p \rightarrow \mathcal{Y}$ de la primul algoritm din enunț și clasificatorii multi-class T de la cel de-al doilea algoritm (SAMME):⁶⁰⁶

$$T(x) = k \text{ dacă } g^{(k)}(x) = 1$$

și invers:

$$g^{(i)}(x) = \begin{cases} 1 & \text{dacă } T(x) = k, \\ -\frac{1}{K-1} & \text{dacă } T(x) \neq k. \end{cases}$$

77.

(Algoritmul AdaBoost: Adevărat sau Fals?)

□ • ○ CMU, 2012 fall, T. Mitchell, Z. Bar-Joseph, final, pr. 8.4

- a. Eroarea la antrenare produsă de clasificatorul H_T obținut de AdaBoost — bazat pe combinația liniară determinată de ipotezele „slabe“ — descrește monoton pe măsură ce crește numărul de iterări executate de algoritm.

MIT, 2002 fall, Tommi Jaakkola, midterm, pr. 5.3

- b. Ponderile / „voturile“ α_t asignate de către AdaBoost ipotezelor „slabe“

⁶⁰⁵Numele SAMME este acronim pentru *Stagewise Additive Modeling with Multi-class Exponential loss*.

⁶⁰⁶Notăm $g(x)$ cu $(g^{(1)}(x), \dots, g^{(K)}(x))$

h_t au tendința să scadă în timpul execuției algoritmului, pentru că eroarea ponderată produsă la antrenare (engl., weighted training error) de ipotezele „slabe“ în general crește.

MIT, 2001 fall, Tommi Jaakkola, midterm, pr. 4.2

- c. Ponderile / „voturile“ α_t pe care algoritmul AdaBoost le asignează ipotezelor „slabe“ h_t sunt optimale, în sensul că ele asigură o *acuratețe la antrenare* mai bună decât în cazul oricărora altor valori asignate acestor voturi / ponderi α_t .

CMU, 2014 spring, B. Poczos, A. Singh, midterm, pr. 1.8

- d. Folosind în algoritmul AdaBoost compași de decizie (cu rol de ipoteze „slabe“), este posibil să obținem granițe de decizie sub formă de parabolă (determinate, deci, de polinoame de ordinul al doilea).

MIT, 2001 fall, Tommi Jaakkola, final, pr. 1.4

- e. Unul dintre avantajele algoritmului AdaBoost este că nu produce niciodată supra-specializare (engl., overfitting).

CMU, 2010 fall, Aarti Singh, midterm, pr. 1.7

- f. Învățăm un clasificator f folosind algoritmul AdaBoost cu ipoteze „slabe“ h . Forma funcțională a separatorului decizional determinat de f este aceeași cu forma funcțională a ipotezelor h , însă cu parametri diferiți. De exemplu, dacă ipotezele h au fost clasificatori liniari, atunci și f este un clasificator liniar.



© M. Romanică

5 Mașini cu vectori-suport

Sumar

Noțiuni preliminare

- elemente [simple] de *calcul vectorial*; proprietăți elementare ale produsului scalar al vectorilor din \mathbb{R}^n , norma euclidiană (L_2) și norma L_1 în \mathbb{R}^n : ex. 1, ex. 35;
- elemente [simple] de *geometrie analitică*: ecuația unei drepte din planul euclidian, ecuația unui plan din \mathbb{R}^3 , ecuația unui hiper-plan din \mathbb{R}^n ; ecuația dreptei care trece prin două puncte date în planul euclidian: ex. 5.c; panta unei drepte perpendiculare pe o dreaptă dată: ex. 8.d, ex. 36, ex. 38;
- distanța (cu sau fără semn) de la un punct la o dreaptă (respectiv la un plan, sau mai general la un hiperplan): ex. 1, ex. 5, ex. 6.c;
- proprietăți de bază din *calculul matriceal*;
- calculul *derivatelor parțiale* [pentru funcții obținute prin compuneri de funcții elementare];
- *metoda lui Lagrange* pentru rezolvarea problemelor de *optimizare convexă cu restricții*:
ex. 35,
ex. 82, ex. 83, ex. 84, ex. 85, ex. 86, ex. 87, ex. 171, ex. 172, ex. 173, ex. 174, ex. 175, ex. 176 de la capitolul de *Fundamente*;
- *separabilitate liniară*, separator optimal, *margine geometrică*: ex. 38, ex. 6.abc, ex. 8.

SVM cu margine “hard”

- (•) deducerea *formei primale* pentru problema SVM (cu margine “hard”), pornind de la principiul maximizării marginii geometrice: ex. 2;⁶⁰⁷
- (P0) o formă [simplă] echivalentă cu *forma primală a problemei de optimizare SVM*: ex. 7;
- *exemplificarea* identificării *separatorului optimal* și a *vectorilor-suport*, pornind de la condițiile din forma primală: ex. 3-5, ex. 36, ex. 38, ex. 39 și CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW4, ex. 4.1-4;
- calcularea *erorii* la cross-validation “leave-one-out” atunci când se folosește o SVM liniară cu margine “hard”: ex. 4.d, ex. 37;
- exemple de [funcții de] *mapare a atributelor*, cu scopul de a obține separabilitate liniară în spațiul de trăsături: ex. 6.d, ex. 8, ex. 40.bd, ex. 41.a; rezolvarea directă a problemei SVM primale în [noul] spațiu de trăsături; identificarea separatorului neliniar din spațiul inițial [de trăsături]: ex. 8.de, ex. 40.ce, ex. 41.b-e;
- (•) deducerea *formei duale* pentru problema SVM cu margine “hard”: ex. 9;

⁶⁰⁷Vedeți și Andrew Ng (Stanford), Lecture Notes, part V, section 3.

- exemplificarea a două modalități de găsire a soluției formei duale a problemei SVM cu margine “hard” pentru învățarea unui concept [reprezentat de un set de date de antrenament neseparabil în spațiul „inițial“ de trăsături] folosind o funcție de mapare Φ dată: prin optimizare directă (ex. 10, ex. 42), respectiv prin folosirea relațiilor de legătură cu soluția problemei primale: ex. 11;
- (P1) dacă datele de antrenament sunt liniar separabile, atunci problema de optimizare SVM în formă primală are soluție unică, întrucât funcția obiectiv a acestei probleme este strict convexă: ex. 2.a (vedeți *Observația*).
- (P2) efectul *multiplicării valorilor atributelor* cu o constantă pozitivă asupra separatorului obținut de SVM), atunci când datele de antrenament sunt liniar separabile cu ajutorul unui hiperplan care trece prin originea sistemului de coordonate: ex. 27.a;⁶⁰⁸
- (P3) efectul unui *atribut irrelevant* — în sensul că nu afectează satisfacerea *restricțiilor* de separabilitate liniară a datelor de antrenament și, în plus, nu mărește marginea de separare — asupra rezultatelor clasificatorului SVM (și respectiv C-SVM): ex. 49, ex. 58;
- vedeți proprietățile (P6) și (P8) enunțate mai jos (la secțiunea despre C-SVM), care sunt valabile și în cazul SVM [cu margine “hard”];
- *comparații* între SVM [având, eventual, diferite funcții-nucleu] și alți clasificatori: ex. 45, ex. 46, ex. 58; vedeți și ex. 12 de la capitolul *Învățare bazată pe memorare*.

SVM cu margine “soft” (C-SVM):

- (•) deducerea *formei duale* pentru problema SVM cu margine “soft” (C-SVM): ex. 12.a-d;
- (P4) legătura dintre valorile (și intervalele de valori) pentru multiplicatorii Lagrange α_i pe de o parte, și valorile (și intervalele de valori) pentru *marginile funcționale* $y_i(\bar{w} \cdot x_i + \bar{w}_0)$ pe de altă parte: ex. 12.e;
- (P5) exprimarea / calcularea distanței geometrice de la un vectori-suport x_i pentru care $\bar{\alpha}_i = C$ la hiperplanul-margine corespunzător etichetei y_i , cu ajutorul variabilei de „destindere“ ξ_i : ex. 13.a;
- exemplificarea noțiunilor de bază: ex. 47, ex. 48 și CMU, 2008 fall, Eric Xing, final, ex. 2.2; un exemplu de calculare a valorii optime pentru funcția obiectiv a problemei de optimizare C-SVM: ex. 13.b;
- exemplificarea poziționării separatorului optimal determinat de C-SVM (pentru diferite valori ale parametrului C), în prezența unui outlier: ex. 15;
- exemplificarea efectului pe care îl are creșterea valorii parametrului de „destindere“ C [asupra marginii și asupra excepțiilor la clasificare]: ex. 16, CMU, 2010 fall, Aarti Singh, HW3, ex. 3.2;
- (P1') spre deosebire de SVM (vedeți proprietatea (P1)), în cazul C-SVM *unicitatea* soluției problemei de optimizare C-SVM în forma primală *nu este garantată* [chiar și] atunci când datele sunt liniar separabile: ex. 17 furnizează un exemplu de situație în care forma duală a problemei de optimizare C-SVM are soluție unică, dar forma sa primală *nu* are soluție unică;

⁶⁰⁸Rezultatul *nu* se menține și în cazul C-SVM: ex. 27.b.

- (P6) o proprietate pentru C-SVM (dar și pentru SVM): ex. 14.
Dacă în setul de date de antrenament două trăsături (engl., features) sunt duplicate ($x_{ij} = x_{ik}$ pentru $i = 1, \dots, m$), atunci ele vor primi ponderi identice ($\bar{w}_j = \bar{w}_k$) în soluția optimală calculată de clasificatorul [C-]SVM;
- (P7) o *margine superioară* (engl., upper bound) pentru numărul de *erori* comise la *antrenare* de către C-SVM: ex. 18;

$$\text{err}_{\text{train}}(\text{C-SVM}) \leq \frac{1}{m} \sum_i \xi_i$$

- (P8) o proprietate pentru C-SVM (dar și pentru SVM):
La CVLOO numai vectorii-suport pot fi (eventual!) clasificați eronat: ex. 19; astăzi avem [și] o *margine superioară* pentru numărul de *erori* comise la *CVLOO* de către C-SVM:

$$\text{err}_{\text{CVLOO}}(\text{C-SVM}) \leq \frac{\#SV_s}{m}$$

- (P9) chestiuni legate de *complexitatea computațională* privind clasificatorul C-SVM: ex. 55;
- (•) deducerea *formei duale* pentru problema SVM cu margine “soft” (C-SVM) de normă \mathcal{L}_2 : ex. 50;
- (•) o formă echivalentă a problemei de optimizare C-SVM, în care nu apar deloc restricții asupra variabilelor, dar în care se folosește funcția de pierdere / cost *hinge*: ex. 20;
exemplificare / aplicare (și comparare cu regresia logistică): ex. 51; exemplu de calculare a costurilor *hinge*: ex. 56.B; algoritmul Pegasos: ex. 21 ;
- (•) algoritmul SMO (Sequential Minimal Optimization):
deducerea relațiilor de actualizare a variabilelor Lagrange: ex. 22;
exemple de aplicare a algoritmului SMO simplificat: ex. 23, ex. 52;
[SMO pentru *one-class*, *Max Margin* SVM: ex. 31.c.];
- o *comparație* asupra efectului *atributelor irelevante* (aici, în sensul că odată eliminate / adăugate, n-ar trebui să afecteze rezultatele clasificării) asupra clasificatorilor 1-NN și C-SVM: ex. 58.

SVM / C-SVM și funcțiile-nucleu — câteva proprietăți

- *exemplificarea* corespondenței dintre forma (primală sau duală) a problemei C-SVM și alegerea valorii parametrului de „destindere“ C și a *funcției-nucleu* pe de o parte și alura și poziționarea separatorului optimal pe de altă parte: ex. 24, ex. 53, ex. 54;
- *exemplificarea* efectului pe care îl are translatarea datelor în raport cu o axă (Oy) asupra poziției separatorului optimal pentru SVM (în raport cu *funcția-nucleu* folosită): ex. 43;
- C-SVM: condiții suficiente asupra parametrului de „destindere“ C și asupra valorilor *funcției-nucleu* pentru ca toate instanțele de antrenament să fie vectori-suport: ex. 25;
- SVM cu nucleu RBF: câteva proprietăți remarcabile
 - pentru SVM pe un set de date [separabil liniar în spațiul de trăsături] instanțe foarte depărtate de separatorul optimal pot fi vectori-suport: ex. 44;

- (P10) pentru orice set de instanțe distincte și pentru orice etichetare a acestora, există o valoare a hiper-parametrului nucleului RBF (σ) astfel încât SVM obține la antrenare eroare 0: ex. 26. Rezultatul *nu* este valabil și pentru C-SVM;
- (P11) pentru orice set de instanțe distincte, pentru orice etichetare a acestora și pentru *orice* valoare a hiper-parametrului nucleului RBF (σ), problema de tip SVM care impune ca toate instanțele să fie corect clasificate și la distanța $1/\|w\|$ față de separatorul optimal are soluție: ex. 29;
- avantaje și dezavantaje ale folosirii metodelor de clasificare liniară de tipul SVM, Perceptron etc. și versiunile lor kernel-izate: ex. 57;
- chestiuni recapitulative: ex. 27, ex. 28, ex. 59, ex. 67.

Alte probleme [de optimizare] de tip SVM

- probleme de tip C-SVM cu valori distincte pentru hiperparametrul C , corespunzător exemplelor negative și respectiv exemplelor pozitive: ex. 60.de;
- SVM pentru clasificare n -ară (SVM multiclass): ex. 30, ex. 61;
- deducerea *formei duale* pentru problema *one-class SVM*, versiunea *Max Margin*: ex. 31 (varianta cu margine “hard”), ex. 63 (varianta cu margine “soft”, folosind ν -SVM);
- legătura dintre soluțiile problemei *one-class SVM*, versiunea *Max Margin*, cu margine “hard” și respectiv cele ale problemei SVM (cu și respectiv fără termen liber (engl., bias)), tot cu margine “hard”: ex. 62;
- deducerea *formei duale* pentru problema *one-class SVM*, versiunea *minimum enclosing ball* (MEB): ex. 32 (varianta cu margine “hard”), ex. 64 (varianta cu margine “soft”, folosind ν -SVM);
- o condiție suficientă pentru ca variantele cu margine “hard” pentru cele două tipuri de probleme de optimizare *one-class SVM*, și anume *Max Margin* și *minimum enclosing ball* (MEB), în forma kernel-izată, să fie echivalente: ex. 32;
- deducerea *formei duale* pentru problema ν -SVM: ex. 33;
- deducerea *formei duale* pentru problema SVR (*Support Vector Regression*), folosind funcție de cost / pierdere ε -senzitivă: ex. 34 (cu margine “hard”), ex. 66 (cu margine “soft” și (echivalent) cu funcție de cost ε -senzitivă); exemplificare / aplicare: ex. 65.

5.1 Maşini cu vectori-suport — Probleme rezolvate

5.1.1 SVM cu margine “hard”

1. (O proprietate a vectorului w care apare în ecuația unui hiperplan; distanța de la un hiperplan din \mathbb{R}^d la un punct oarecare din același spațiu: deducerea formulei)

*prelucrare de Liviu Ciortuz, după
■ □ • ○ CMU, 2010 fall, Ziv Bar-Joseph, HW4, pr. 1.1*

Fie un punct oarecare x_0 din \mathbb{R}^d și un hiperplan de ecuație $w \cdot x + b = 0$, unde w și x sunt de asemenea din \mathbb{R}^d , în vreme ce $b \in \mathbb{R}$, iar simbolul \cdot indică produsul scalar al vectorilor din \mathbb{R}^d .

- Demonstrați că vectorul w este ortogonal pe acest hiperplan.
(*Sugestie:* Fie x_1 și x_2 două puncte situate pe acest hiperplan. Ce puteți spune despre valoarea produsului scalar $w \cdot (x_1 - x_2)$?)
- Demonstrați că în cazul în care $w \neq 0 \in \mathbb{R}^d$, distanța [măsurată pe perpendiculară] de la x_0 la hiperplanul $w \cdot x + b = 0$ este

$$\frac{|w \cdot x_0 + b|}{\|w\|},$$

unde simbolul $\|\cdot\|$ desemnează norma euclidiană.⁶⁰⁹

Răspuns:

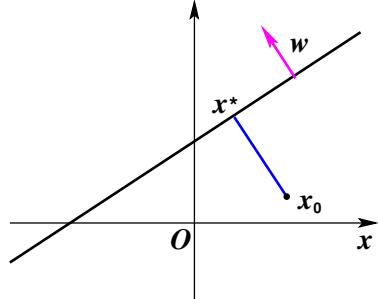
- a. Fie H hiperplanul de ecuație $w \cdot x + b = 0$ (în scriere vectorială, $w^\top x + b = 0$, unde am considerat w și x vectori-colonă din \mathbb{R}^d). Considerăm x_1 și x_2 două puncte oarecare distincte situate pe hiperplanul H . Așadar, vom avea

$$w \cdot x_1 + b = 0 \text{ și } w \cdot x_2 + b = 0,$$

de unde, prin scădere rezultă $w \cdot (x_1 - x_2) = 0$, ceea ce înseamnă că vectorii w și $x_1 - x_2$ sunt perpendiculari. Întrucât vectorul $x_1 - x_2$ are direcția hiperplanului H , rezultă că w este perpendicular pe H .

- b. Notăm cu x^* „piciorul“ perpendiculariei coborâte din punctul x_0 pe hiperplanul H . Prin urmare, vectorul $x^* - x_0$ are aceeași direcție cu vectorul w . Stim că pentru oricare doi vectori paraleli u și u' este satisfăcută proprietatea următoare: există o constantă $\lambda \in \mathbb{R}$ astfel încât $u' = \lambda u$. Așadar, există $t \in \mathbb{R}$ cu proprietatea $x^* - x_0 = tw$. Altfel spus,

$$x^* = x_0 + tw. \quad (282)$$



⁶⁰⁹ $\|w\|^2 = w \cdot w$ sau $\|w\|^2 = w^\top w$, unde simbolul \top indică operația de transpunere de vectori / matrice. Precizăm că norma euclidiană se desemnează de fapt prin $\|\cdot\|_2$, însă putem renunța la indice atunci când nu există posibilitatea de a se face confuzie cu vreo altă normă în contextul respectiv.

Stim de asemenea că punctul x^* aparține hiperplanului H . Prin urmare,

$$w \cdot x^* + b = 0. \quad (283)$$

Din relațiile (282) și (283) rezultă

$$\begin{aligned} w \cdot (x_0 + tw) + b = 0 &\Leftrightarrow w \cdot x_0 + tw^2 + b = 0 \Leftrightarrow w \cdot x_0 + t\|w\|^2 + b = 0 \Leftrightarrow \\ t\|w\|^2 &= -(w \cdot x_0 + b) \Leftrightarrow t = -\frac{w \cdot x_0 + b}{\|w\|^2}. \end{aligned}$$

În concluzie, distanța de la x_0 la hiperplanul H este

$$\|x^* - x_0\| = \|tw\| = |t| \|w\| = \frac{|w \cdot x_0 + b|}{\|w\|^2} \|w\| = \frac{|w \cdot x_0 + b|}{\|w\|}.$$

2.

(Exercițiu teoretic: deducerea formei primale a problemei de optimizare SVM (cu margine “hard”, pornind de la principiul maximizării marginii geometrice)

■ □ • CMU, 2016 fall, N. Balcan. M. Gormley, HW4, pr. 1.1
CMU, 2006 spring, Carlos Guestrin, midterm, pr. 1.4

În acest exercițiu veți vedea cum anume se deduce problema de optimizare SVM cu margine “hard” pornind de la principiul marginii [geometrice] maxime.

Presupunem că avem setul de date de antrenament $D = (X, y)$, unde $X \in \mathbb{R}^{d \times m}$, iar $y \in \{-1, 1\}^m$. Coloana i a matricei X este x_i , vectorul de trăsături al celui de-al i -lea exemplu de antrenament, iar y_i este eticheta acestui exemplu.

Pentru clasificare, vom folosi o funcție liniară, de forma

$$f(x) = w \cdot x, \text{ unde } w \in \mathbb{R}^d,$$

iar operatorul \cdot reprezintă produsul scalar al vectorilor.

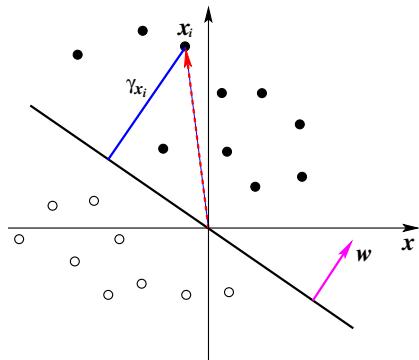
Observăm că, pentru simplitate, în expresia lui f nu a fost adăugat termenul liber (notat îndeobște cu w_0 sau cu b , ultimul provenind de la termenul englezesc *bias*). Așadar, vom trata aici doar cazul separabilității liniare prin originea sistemului de coordonate.⁶¹⁰

O instanță oarecare x va fi clasificată în clasa 1 dacă $f(x) > 0$, respectiv în clasa -1 în caz contrar. Hiperplanul de ecuație $f(x) = 0$ va funcționa ca *separator* al instanțelor de antrenament, sau *graniță de decizie* (engl., decision boundary). Așadar, f este funcția pe care vrem s-o învățăm (adică funcția “target”).

⁶¹⁰Cazul separabilității liniare cu termenul liber w_0 luând o valoare oarecare se tratează extinzând în mod corespunzător demonstrația de la punctul a.

Presupunem că datele de antrenament sunt liniar separabile. Atunci pentru orice exemplu de antrenament (x, y) vom avea $yf(x) > 0$. În consecință, *distanța geometrică* de la x la granița de decizie — distanță despre care, conform problemei 1, știm că este egală cu $\frac{|f(x)|}{\|w\|}$ — poate fi scrisă ca

$$\gamma_x = \frac{yf(x)}{\|w\|}.$$



În contextul învățării automate, această distanță se numește *marginea geometrică* sau, pe scurt, *marginea* [dintre instanța x și separatorul determinat de ecuația $f(x) = 0$].

În vederea reducerii *riscului de clasificare eronată* la faza de *generalizare*, este natural să determinăm f (deci, de fapt, vectorul de ponderi w) astfel încât minimul acestor margini γ_x să fie cât mai mare. Altfel spus, vrem să maximizăm distanța [de la *separatorul optimal*] până la instanțele de antrenament cele mai apropiate. Așadar, *funcția noastră obiectiv* va fi:

$$\max_w \min_{i=1,\dots,m} \frac{y_i f(x_i)}{\|w\|}. \quad (284)$$

a. Arătați că problema de optimizare fără restricții (284) este *echivalentă* cu următoarea problemă de optimizare convexă cu restricții liniare, pe care o vom numi *problema SVM* (în aşa-nimita *formă primală*):⁶¹¹

$$\min_w \frac{1}{2} \|w\|^2 \quad (285)$$

a. i. $y_i(w \cdot x_i) \geq 1$, pentru $i = 1, \dots, m$.

Sugestie: Are oare vreun efect scalarea vectorului de ponderi $w \rightarrow kw$, cu $k \in \mathbb{R}$, $k > 0$?

Observație: Are loc următoarea *proprietate*:⁶¹² dacă multimea S este liniar separabilă, atunci problema de optimizare SVM în formă primală are soluție unică, întrucât funcția obiectiv a acestei probleme este strict convexă.

b. Presupunând că datele de antrenament sunt liniar separabile, precizați ce se va întâmpla cu separatorul obținut prin rezolvarea problemei SVM dacă se renunță la unul dintre exemplele de antrenament: se va deplasa oare separatorul *însprij* exemplul / punctul respectiv, se va retrage *dinspre* punctul respectiv, ori poziția lui va rămâne neschimbată? Justificați.

Răspuns:

⁶¹¹LC: Este vorba de o echivalență *slabă* (engl., soft): optimul — adică, valoarea funcției obiectiv — pentru cele două probleme este același, iar orice soluție a primei probleme poate fi pusă în corespondență cu o soluție a celei de-a doua probleme și invers. (Echivalența ar fi *tare* (engl., hard) dacă valoarea funcției obiectiv pentru cele două probleme ar fi aceeași, iar orice soluție optimă a primei probleme ar fi soluție optimă și pentru cea de-a două problemă și invers.)

⁶¹²Vedeți articolul *Uniqueness of the SVM solution*, de Christopher Burges și David Crisp, 1998.

a. Vom demonstra echivalența dintre problemele de optimizare date în enunț (284) și (285) construind o succesiune de câteva probleme de optimizare echivalente, obținute prin aplicarea unor transformări succesive, pornind de la problema (284) și ajungând în final la problema (285).

Fie, aşadar, următoarele probleme de optimizare:

$$\max_w \min_{i=1,\dots,m} \frac{y_i w \cdot x_i}{\|w\|} \quad (286)$$

a. i. $y_i(w \cdot x_i) > 0$, pentru $i = 1, \dots, m$.

$$\max_w \frac{1}{\|w\|} \min_{i=1,\dots,m} y_i w \cdot x_i \quad (287)$$

a. i. $y_i(w \cdot x_i) > 0$, pentru $i = 1, \dots, m$.

$$\max_w \frac{1}{\|w\|} \min_{i=1,\dots,m} y_i w \cdot x_i \quad (288)$$

a. i. $y_i(w \cdot x_i) \geq 1$, pentru $i = 1, \dots, m$.

$$\max_w \frac{1}{\|w\|} \quad (289)$$

a. i. $y_i(w \cdot x_i) \geq 1$, pentru $i = 1, \dots, m$.

- Se observă că problema de optimizare (286) diferă de problema inițială (284) prin introducerea unor restricții liniare. Mai mult, aceste restricții corespund hiperplanelor care sunt separatori liniari pentru mulțimea de antrenament dată (care este, conform enunțului, separabilă liniar). Așadar, aceste restricții nu schimbă cu nimic soluția problemei care a fost dată inițial.⁶¹³

- Problema de optimizare (287) a fost obținută din problema (286) modificând doar funcția obiectiv: $\frac{1}{\|w\|}$ a fost scos în fața operatorului $\min_{i=1,\dots,m}$ fiindcă w nu depinde de i . (Observați că expresia care urmează acestui operator depinde de w , dar acolo acest w este considerat fixat și variază doar i .)

- Problema de optimizare (288) a fost obținută din problema (287) schimbând în restricții relația > 0 cu ≥ 1 . În sine, această modificare restrâne mulțimea acelor valori ale lui w peste care se aplică operatorul \max_w .

⁶¹³Pentru un separator liniar oarecare al mulțimii de antrenament, expresia $\min_{i=1,\dots,m} \frac{y_i f(x_i)}{\|w\|}$ are o valoare strict pozitivă. Pentru hiperplanele care nu separă mulțimea de antrenament, expresia $\min_{i=1,\dots,m} \frac{y_i f(x_i)}{\|w\|}$ are valoare negativă sau 0.

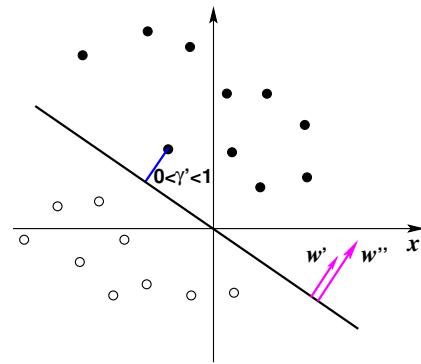
Să considerăm însă un w' arbitrar care este eliminat la trecerea de la forma (287) la forma (288). Notăm $\gamma' = \min_{i=1,\dots,m} y_i \underbrace{w' \cdot x_i}_{f(x_i)}$; stim că $\gamma' \in (0, 1)$. Rezultă că

$$\frac{y_i w' \cdot x_i}{\gamma'} = y_i \frac{w'}{\gamma'} \cdot x_i \geq 1 \text{ pentru } i = 1, \dots, m.$$

Observați că egalitatea chiar se produce, și anume atunci când indexul i ia valoarea pentru care se atinge minimul expresiei $y_i w' \cdot x_i$.

Așadar, $w'' \stackrel{\text{not.}}{=} \frac{w'}{\gamma'}$ satisfacă restricțiile problemei (288). În plus,

$$\frac{1}{\|w''\|} = \frac{1}{\|w'\|} \gamma' \text{ și, echivalent, } \frac{1}{\|w''\|} \underbrace{\min_{i=1,\dots,m} y_i w'' \cdot x_i}_{1} = \frac{1}{\|w'\|} \underbrace{\min_{i=1,\dots,m} y_i w' \cdot x_i}_{\gamma'}.$$



Prin urmare, putem spune că rolul lui w' în relația (287) este jucat de w'' în relația (288). În consecință, la trecerea de la o formă la cealaltă optimul rămâne același.⁶¹⁴

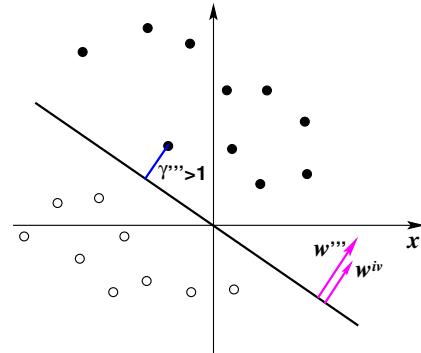
- Problema de optimizare (289) a fost obținută din problema (288) renunțând în funcția obiectiv la componenta $\min_{i=1,\dots,m} y_i w \cdot x_i$.

Aceasta revine de fapt la a renunța la acele valori ale lui w pentru care $\min_{i=1,\dots,m} y_i w \cdot x_i$ este strict mai mare decât 1. (Am văzut mai sus că există valori ale lui w (și anume, w'') pentru care minimul respectiv este chiar 1.) Să considerăm un w''' cu proprietatea $\gamma''' \stackrel{\text{not.}}{=} \min_{i=1,\dots,m} y_i \underbrace{w''' \cdot x_i}_{f(x_i)} > 1$.

Rezultă că

$$\min_{i=1,\dots,m} y_i \frac{w'''}{\gamma'''} \cdot x_i = 1.$$

În continuare, notând $w^{iv} = \frac{w'''}{\gamma'''}$, vom avea:



$$\frac{1}{\|w^{iv}\|} \underbrace{\min_{i=1,\dots,m} y_i w^{iv} \cdot x_i}_{1} = \frac{1}{\|w^{iv}\|} = \frac{\gamma'''}{\|w'''\|} = \frac{1}{\|w'''\|} \underbrace{\min_{i=1,\dots,m} y_i w''' \cdot x_i}_{\gamma'''}$$

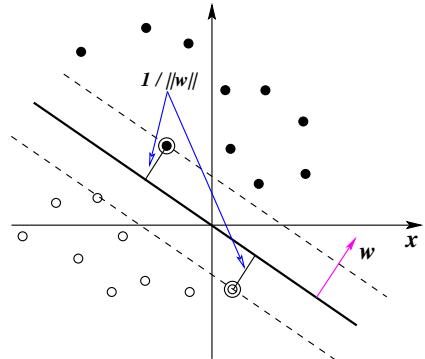
Prin urmare, [ca și mai sus,] putem spune că rolul lui w''' în relația (288) este jucat de w^{iv} în relația (289). În consecință, [ca și mai sus,] nu se modifică

⁶¹⁴Din punct de vedere geometric, trecerea de la (287) la (288) poate fi asociată cu înmulțirea cu o constantă pozitivă (chiar supra-unitară) a ecuațiilor acestor separatori liniari [ai mulțimii de antrenament] pentru care minimul distanțelor geometrice până la instanțele de antrenament este strict *mai mic* decât 1, astfel încât în urma înmulțirii această distanță minimă să devină 1.

valoarea optimă a funcției obiectiv dacă se renunță [și] la acești w''' .⁶¹⁵

- Se constată imediat că problema (289) este echivalentă cu problema de optimizare SVM care a fost dată în enunț (285). Forma problemei de optimizare SVM ne spune că separatorul optimal se alege dintre acei separatori $w \cdot x$ ai mulțimii de antrenament care au exact valoarea 1 pentru $\min_{i=1,\dots,m} y_i w \cdot x_i$.

Concret, alegerea se face maximizând $1/\|w\|$, adică distanța geometrică de la separator(i) până la cele mai apropiate instanțe de antrenament. Intuitiv, maximizarea aceasta va implica faptul că distanțele de la separatorul optimal până la *toate* instanțele de antrenament cele mai apropiate (adică, *vectorii-suport*, cei pozitivi și respectiv cei negativi) vor fi $1/\|w\|$. (În figura alăturată, în care am pus în evidență acest fapt, vectorii-suport sunt încercuți.)



Altfel spus, valoarea funcției $w \cdot x$ pentru w optim va fi +1 pentru vectorii-suport pozitivi și -1 pentru vectorii-suport negativi.

- b. La eliminarea unei instanțe de antrenament x_j , marginea $\max_w \min_i \frac{y_i f(x_i)}{\|w\|}$ nu poate să scadă; poate doar să crească sau, eventual, să rămână neschimbătă. Într-adevăr, pentru un w oarecare (fixat), cantitatea $\min_i \frac{y_i f(x_i)}{\|w\|}$ fie crește fie rămâne aceeași la eliminarea unui x_j , fiindcă acum operatorul min este aplicat pe o mulțime mai puțin amplă. În consecință, lăsându-l apoi pe w să varieze, cantitatea $\max_w \min_i \frac{y_i f(x_i)}{\|w\|}$ crește și ea sau cel puțin rămâne aceeași.

Tinând cont de rezultatul de echivalență de la punctul a, rezultă că la eliminarea unei instanțe de antrenament x_j marginea de separare optimală ($1/\|w\|$) crește sau rămâne, eventual, neschimbătă. Așadar, poziția separatorului optimal (i) va fi deplasată înspre instanța eliminată sau (ii) va rămâne neschimbătă. Cazul (i) este posibil să apară doar dacă x_j este vector-suport, adică se află la distanță minimală față de separatorul optimal.⁶¹⁶ Cazul (ii) se produce atunci când x_j este vector-suport și există mai mulți vectori-suport care au etichete identice cu eticheta lui x_j sau atunci când x_j nu este vector-suport.

Observație importantă:

Până la problema 9 — unde vom prezenta deducerea *formeи duală* a problemei de optimizare SVM —, vom lucra cu o *definiție în sens larg* (geometric) pentru noțiunea de *vector-suport*: spunem că un vector-suport este o instanță x_i care se află la distanță geometrică de $1/\|w\|$ față de separatorul optimal (w, w_0) . Odată cu introducerea formei duală a problemei de optimizare SVM, vom folosi pentru noțiunea de vector-suport *definiția clasică* (în sens (analitic) restrâns în raport cu cel precedent): o instanță x_i este vector-suport dacă în

⁶¹⁵Din punct de vedere geometric, trecerea de la (288) la (289) corespunde înmulțirii cu o constantă pozitivă (sub-unitară) a ecuațiilor acelor separatori liniari [ai mulțimii de antrenament] pentru care minimul distanțelor geometrice până la instanțele de antrenament este strict *mai mare* decât 1, astfel încât în urma înmulțirii această distanță minimă să devină 1.

⁶¹⁶Totuși, nu este *obligatoriu* ca în această situație poziția separatorului optimal să se modifice. Vedeți cazul (ii).

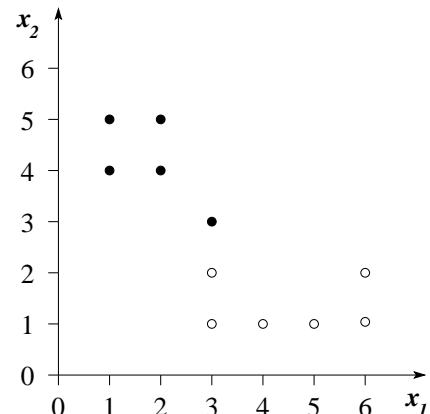
soluția optimă a problemei duale, variabila / multiplicatorul Lagrange $\bar{\alpha}_i$ are valoare nenulă (deci strict pozitivă, pentru că toți multiplicatorii Lagrange sunt acolo nenegativi). În această a doua accepțiune, unele instanțe situate la distanță (geometrică) de $1/\|w\|$ față de separatorul optimal (w, w_0) pot să nu fie vectori-suport. Pentru SVM cu margine “soft” (C-SVM), va fi util să vedeați *Rezumatul* de la rezolvarea punctului e de la problema 12. Se va constata acolo că este posibil să avem vectori-suport aflați la distanță (geometrică) mai mică de $1/\|w\|$ față de separatorul optimal (w, w_0) ; aceștia sunt așa-numiții (în engl.) *non-bound support vectors*.

3.

(Separabilitate în \mathbb{R}^2 ; SVM liniară, forma primală)*CMU, 2003 fall, T. Mitchell, A. Moore, final exam, pr. 5*

Desenul alăturat prezintă un set de date cu două atrbute de intrare x_1 și x_2 , și un atrbut de ieșire y , ale cărui valori sunt reprezentate prin culoarea punctului (alb/- și negru/+).

a. Presupunând că aceste date sunt corecte (adică fără ‘zgomote’) și că folosim o SVM liniară, trasați pe acest desen trei drepte care să indice linia de *separare optimală* și cele două „margini“ (paralelele la separatorul optimal care trec prin vectorii-suport). Încercuiți vectorii-suport.

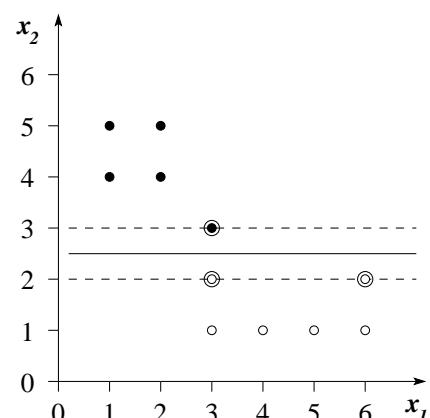


b. Plecând de la forma primală a unei SVM liniare (având regula de decizie $y = \text{sign}(w \cdot x + w_0)$),⁶¹⁷ calculați valorile corespunzătoare parametrilor w și w_0 care determină hiperplanul de separare optimală de la punctul a.

Răspuns:

a. În figura alăturată, linia continuă reprezintă dreapta de separare, iar cele două linii punctate reprezintă marginile maxime. Vectorii-suport sunt punctele $(3, 3)$, $(3, 2)$ și $(6, 2)$.

b. În forma primală a problemei SVM liniare se cere maximizarea valorii $1/2 \cdot \|w\|^2$, simultan cu satisfacerea restricțiilor $y_i(w \cdot x + w_0) \geq 1$, unde $w \stackrel{\text{not.}}{=} (w_1, w_2)$ și $x \stackrel{\text{not.}}{=} (x_1, x_2)$. Tinând cont de faptul că aceste inegalități sunt îndeplinite cu egalitate [doar] în cazul vectorilor-suport, obținem următorul sistem de ecuații:



⁶¹⁷Vedeți la problema 2.a (sau la problema 9) cum se definește *forma primală* a acestei SVM liniare.

$$\begin{cases} -((w_1, w_2) \cdot (3, 2) + w_0) = 1 \\ -((w_1, w_2) \cdot (6, 2) + w_0) = 1 \\ (w_1, w_2) \cdot (3, 3) + w_0 = 1 \end{cases} \Rightarrow \begin{cases} 3w_1 + 2w_2 + w_0 = -1 \\ 6w_1 + 2w_2 + w_0 = -1 \\ 3w_1 + 3w_2 + w_0 = 1 \end{cases} \Rightarrow$$

$$\begin{cases} w_1 = 0 \\ 2w_2 + w_0 = -1 \\ 3w_2 + w_0 = 1 \end{cases} \Rightarrow \begin{cases} w_1 = 0 \\ w_2 = 2 \\ w_0 = -5 \end{cases}$$

Am obținut deci $w = (w_1, w_2) = (0, 2)$ și $w_0 = -5$, ceea ce conduce la ecuația $2x_2 - 5 = 0 \Leftrightarrow x_2 - \frac{5}{2} = 0$ pentru separatorul optimal.

Alternativ, valorile parametrilor w_0, w_1 și w_2 pot fi determinate folosind cunoștințe de geometrie analitică. Ecuația dreptei din figura de mai sus este

$$x_2 = \frac{5}{2} \Leftrightarrow x_2 - \frac{5}{2} = 0$$

sau, mai general, $\alpha(x_2 - \frac{5}{2}) = 0$, cu $\alpha \in \mathbb{R}$, $\alpha \neq 0$. Valoarea lui α corespunzătoare separatorului optimal se obține impunând restricțiile specifice vectorilor-suport:

$$\alpha(x_2 - \frac{5}{2}) = 1 \text{ pentru } x_2 = 3 \text{ sau / și, respectiv, } \alpha(x_2 - \frac{5}{2}) = -1 \text{ pentru } x_2 = 2.$$

Rezultă că $\alpha = 2$, deci ecuația separatorului optimal este $2(x_2 - \frac{5}{2}) = 0 \Leftrightarrow 0 \cdot x_1 + 2 \cdot x_2 - 5 = 0$, de unde, identificând pe componente, rezultă: $w_1 = 0, w_2 = 2$ și $w_0 = -5$.

4.

(Separabilitate în \mathbb{R} ; SVM liniară, forma primală; calculul erorilor la antrenare și CVLOO)

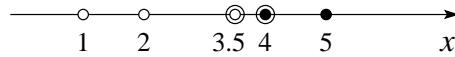
CMU, 2001 fall, Andrew Moore, final exam, pr. 11

Fie următorul set de date:

x_i	1	2	3.5	4	5
y_i	-1	-1	-1	1	1

- Care sunt valorile parametrilor w și w_0 învățate de o SVM liniară pornind de la aceste date? (Vă readucem aminte că regula de decizie a unei astfel de SVM este $y = \text{sign}(w \cdot x + w_0)$.)
- Care sunt vectorii-suport?
- Care este eroarea de clasificare pe mulțimea de date de antrenament de mai sus?
- Care este eroarea de clasificare pe același set de date, folosind metoda de cross-validation "Leave-One-Out"? (Veți exprima eroarea sub forma procentajului de instanțe clasificate corect.)

Răspuns:



În mod *direct*, reprezentarea grafică ne arată că separatorul optimal va fi plasat în dreptul punctului 3.75, folosind vectorii-suport $x_3 = 3.5$ și $x_4 = 4$, și conducând la eroare de antrenare 0 (deoarece datele de antrenament sunt liniar separabile). Este imediat că instanțele $x_3 = 3.5$ și $x_4 = 4$ vor fi clasificate eronat la CVLOO.

a. *Analitic*, din forma primală a problemei SVM rezultă că trebuie satisfăcute următoarele restricții:

$$\left\{ \begin{array}{l} -1(w \cdot 1 + w_0) \geq 1 \\ -1(w \cdot 2 + w_0) \geq 1 \\ -1(w \cdot 3.5 + w_0) \geq 1 \\ 1(w \cdot 4 + w_0) \geq 1 \\ 1(w \cdot 5 + w_0) \geq 1 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} w + w_0 \leq -1 \\ 2w + w_0 \leq -1 \\ 3.5w + w_0 \leq -1 \\ 4w + w_0 \geq 1 \\ 5w + w_0 \geq 1 \end{array} \right.$$

Impunând ca restricțiile corespunzătoare vectorilor-suport ($x_3 = 3.5$ și $x_4 = 4$) să fie satisfăcute cu egalitate — ceea ce va conduce în mod implicit și la satisfacerea celorlalte inegalități din sistem —, rezultă:

$$\left\{ \begin{array}{l} 3.5w + w_0 = -1 \\ 4w + w_0 = 1 \end{array} \right. \Rightarrow \left\{ \begin{array}{l} -3.5w - w_0 = 1 \\ 4w + w_0 = 1 \end{array} \right. \Rightarrow \left\{ \begin{array}{l} 0.5w = 2 \\ 4w + w_0 = 1 \end{array} \right. \Rightarrow \left\{ \begin{array}{l} w = 4 \\ w_0 = -15 \end{array} \right.$$

Am obținut deci soluția $w = 4$, $w_0 = -15$.

b. După cum am menționat mai sus, vectorii-suport sunt $x_3 = 3.5$ și $x_4 = 4$. *Alternativ*, însă tot *analitic*, punctele a și b de mai sus se pot rezolva astfel:

Se observă în mod direct că inegalitatea $x \geq 3.75$ este satisfăcută pentru — deci, funcția $\text{sign}(x - 3.75)$ clasifică corect — toate instanțele, atât cele pozitive cât și cele negative. Mai mult, instanțele $x_3 = 3.5$ (negativă) și $x_4 = 4$ (pozitivă) sunt cele mai apropiate față de „pragul” $x = 3.75$ și sunt situate la distanțe egale față de acest prag, deci sunt vectorii-suport. Ecuatărea $x = 3.75$ se scrie mai general $\alpha(x = 3.75)$, cu $\alpha \neq 0$. Impunând condiția $\alpha(x - 3.75) = -1$ pentru $x = 3.5$ și respectiv $\alpha(x - 3.75) = 1$ pentru $x = 4$, rezultă $\alpha = 4$ și deci, prin identificare pe componente, $w = 4$ și $w_0 = -15$.

c. Judecând *calitativ*, eroarea de clasificare pe mulțimea datelor de antrenament este 0, fiindcă datele sunt liniar separabile. *Analitic*, am arătat deja că $\text{sign}(w \cdot x_i + w_0) = y_i$, pentru $i \in \{1, 2, 3, 4, 5\}$.

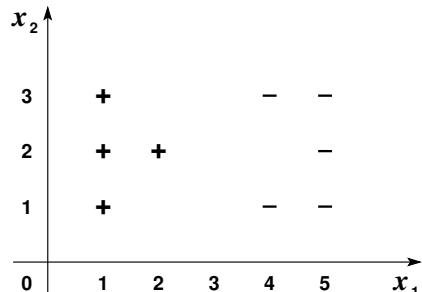
d. La cross-validation cu metoda “Leave-One-Out”, instanțele 3.5 și 4 vor fi clasificate eronat pentru că pragul de separare optimală devine $x = 3$ și respectiv $x = 4.25$. Eroarea este deci de 40%. De notat, ca un *principiu general*: atunci când se lucrează cu SVM, erorile la cross-validation se *pot* produce doar la eliminarea vectorilor-suport.⁶¹⁸

⁶¹⁸Vedeți legătura cu problema 2.b.

5.

(SVM liniară: exemplificare pe date din \mathbb{R}^2 ; modificarea eventuală a poziției separatorului optimal la eliminarea unei instanțe de antrenament)

În imaginea alăturată, încercuiți fiecare punct care are proprietatea că odată ce este eliminat din setul de date de antrenament, la re-antrenarea mașinii cu vectori-suport vom obține un alt separator optimal decât cel rezultat în cazul antrenării pe întreaga mulțime de date. Justificați răspunsul.



Răspuns:

În figura alăturată am reprezentat *grafic* separatorul liniar învățat de SVM din datele de antrenament și am încercuit vectorii-suport. După cum știm deja din rezolvarea problemei 4.d, eliminarea datelor care nu sunt vectorii-suport nu modifică rezultatul procesului de învățare. Așadar, rămâne de verificat comportamentul sistemului la eliminarea câte unui vector-suport.

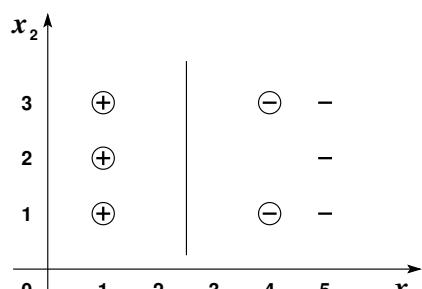
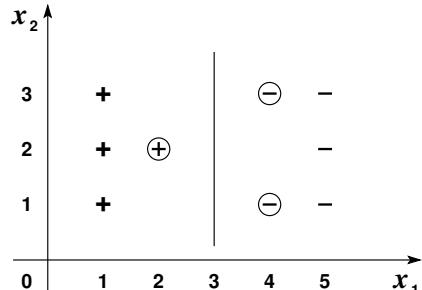
Analitic, putem vedea că separatorul învățat este unic determinat de sistemul:

$$\begin{cases} (w_1, w_2) \cdot (2, 2) + w_0 = 1 \\ -1((w_1, w_2) \cdot (4, 1) + w_0) = 1 \\ -1((w_1, w_2) \cdot (4, 3) + w_0) = 1 \end{cases} \Rightarrow \begin{cases} 2w_1 + 2w_2 + w_0 = 1 \\ 4w_1 + w_2 + w_0 = -1 \\ 4w_1 + 3w_2 + w_0 = -1 \end{cases} \Rightarrow \begin{cases} w_1 = -1 \\ w_2 = 0 \\ w_0 = 3 \end{cases}$$

În consecință, *ecuația separatorului optimal* este $-x_1 + 3 = 0 \Leftrightarrow x_1 = 3$, iar *maginea geometrică* (distanța de la vectorii-suport până la hiperplanul de separare) este $d = \frac{1}{\|w\|} = \frac{1}{\sqrt{(-1)^2 + 0^2}} = 1$, unde după cum știm, $w = (w_1, w_2)$.

La eliminarea punctului (2,2), hiperplanul de separare optimal este altul, vectorii-suport devenind (1,1), (1,2), (1,3), (4,1) și (4,3). Noile valori ale parametrilor (notează w'_0, w'_1, w'_2) se calculează rezolvând sistemul:

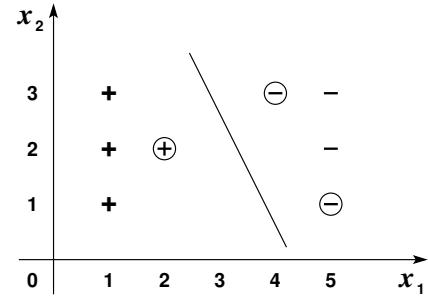
$$\begin{cases} w'_1 + w'_2 + w'_0 = 1 \\ w'_1 + 2w'_2 + w'_0 = 1 \\ w'_1 + 3w'_2 + w'_0 = 1 \\ 4w'_1 + w'_2 + w'_0 = -1 \\ 4w'_1 + 3w'_2 + w'_0 = -1 \end{cases} \Rightarrow \begin{cases} w'_1 = -\frac{2}{3} \\ w'_2 = 0 \\ w'_0 = \frac{5}{3} \end{cases}$$



Așadar, *ecuația separatorului optimal* este acum $-\frac{2}{3}x_1 + \frac{5}{3} = 0 \Leftrightarrow x_1 = \frac{5}{2}$, iar *maginea geometrică* este $d' = \frac{1}{\|w'\|} = \frac{1}{\sqrt{(-2/3)^2 + 0^2}} = \frac{3}{2}$.

La eliminarea punctului (4,1), vectorii-suport devin (2,2), (5,1) și (4,3). Determinarea noilor valori ale parametrilor se face prin rezolvarea sistemului:

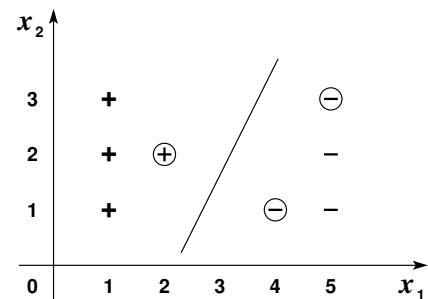
$$\begin{cases} 2w''_1 + 2w''_2 + w''_0 = 1 \\ 5w''_1 + w''_2 + w''_0 = -1 \\ 4w''_1 + 3w''_2 + w''_0 = -1 \end{cases} \Rightarrow \begin{cases} w''_1 = -\frac{4}{5} \\ w''_2 = -\frac{2}{5} \\ w''_0 = \frac{17}{5} \end{cases}$$



În consecință, în acest caz ecuația separatorului optimal este $-\frac{4}{5}x_1 - \frac{2}{5}x_2 + \frac{17}{5} = 0 \Leftrightarrow 4x_1 + 2x_2 - 17 = 0 \Leftrightarrow x_2 = -2x_1 + \frac{17}{2}$, iar marginea geometrică până la hiperplan este $d'' = \frac{1}{\|w''\|} = \frac{1}{\sqrt{(-4/5)^2 + (-2/5)^2}} = \frac{5}{2\sqrt{5}} = \frac{\sqrt{5}}{2}$.

La eliminarea punctului (4,3), vectorii-suport devin (2,2), (4,1) și (5,3). Vom obține următoarele valori ale parametrilor:

$$\begin{cases} 2w'''_1 + 2w'''_2 + w'''_0 = 1 \\ 4w'''_1 + w'''_2 + w'''_0 = -1 \\ 5w'''_1 + 3w'''_2 + w'''_0 = -1 \end{cases} \Rightarrow \begin{cases} w'''_1 = -\frac{4}{5} \\ w'''_2 = \frac{2}{5} \\ w'''_0 = \frac{9}{5} \end{cases}$$



Prin urmare, aici ecuația separatorului optimal este $-\frac{4}{5}x_1 + \frac{2}{5}x_2 + \frac{9}{5} = 0 \Leftrightarrow -4x_1 + 2x_2 + 9 = 0 \Leftrightarrow x_2 = 2x_1 - \frac{9}{2}$, iar marginea geometrică este $d''' = \frac{1}{\|w'''\|} = \frac{1}{\sqrt{(-4/5)^2 + (2/5)^2}} = \frac{5}{2\sqrt{5}} = \frac{\sqrt{5}}{2}$.

6.

(Separabilitate liniară în \mathbb{R}^2 : tratare în raport cu un parametru dat; Separabilitate neliniară în \mathbb{R} : folosirea unei funcții simple pentru maparea atributelor)

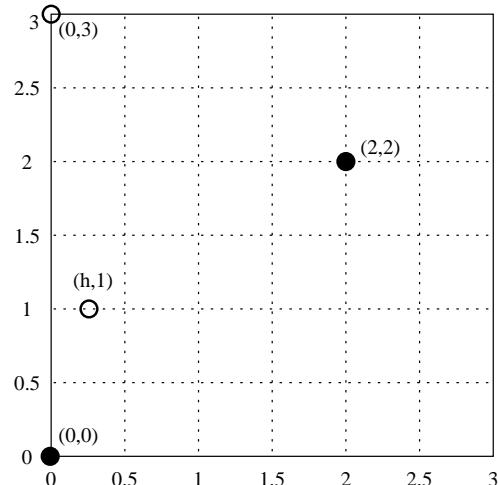
CMU, 2009 spring, Ziv Bar-Joseph, final exam, pr. 4

Presupunem că avem doar patru exemple de antrenament în spațiul euclidian bidimensional, și anume: $x_1 = (0, 0)$, $x_2 = (2, 2)$ sunt exemple pozitive, iar $x_3 = (h, 1)$, $x_4 = (0, 3)$ sunt exemple negative, h fiind un parametru cu proprietatea $0 \leq h \leq 3$.

a. Cât de mare poate fi valoarea lui $h \geq 0$ cu condiția ca punctele de antrenament să rămână liniar separabile?

b. Se schimbă direcția⁶¹⁹ separatorului optimal ca funcție de h atunci când punctele sunt separabile? (Da / Nu; justificare.)

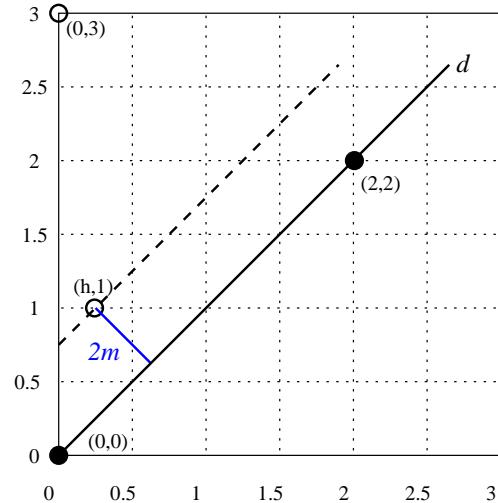
c. Cât este *marginea [geometrică]* corespunzătoare separatorului optimal (adică, distanța de la separator până la vectorii-suport) ca funcție de h ?



d. Presupunem că putem observa doar a doua componentă a instanțelor de antrenament. Fără cealaltă componentă, datele de antrenament etichetate se reduc la $(0,+)$, $(1,-)$, $(2,+)$ și $(3,-)$. Care este gradul minim p al unei funcții polinomiale care, aplicată acestor date, ne permite să le clasificăm corect?

Răspuns:

a. Dreapta d , care este determinată de instanțele pozitive $x_1 = (0,0)$ și $x_2 = (2,2)$ are ecuația $y = x$. Pentru ca punctele de antrenament să rămână liniar separabile trebuie ca punctul x_3 să fie situat de aceeași parte a dreptei d cu instanța negativă $x_4 = (0,3)$. Această condiție se traduce analitic prin inegalitatea $h < 1$.



b. Hiperplanul de separare optimală va fi o dreaptă paralelă cu dreapta d (de ecuație $y = x$), situată la jumătatea distanței dintre aceasta și punctul $x_3 = (h,1)$. Dreapta d este determinată de vectorii-suport pozitivi $x_1 = (0,0)$ și $x_2 = (2,2)$ pe de o parte și de vectorul-suport negativ $x_3 = (h,1)$ de cealaltă parte. Panta dreptei d este complet determinată de cei doi vectori-suport pozitivi, deci este 1; ea nu depinde de valoarea parametrului h . În concluzie, direcția separatorului optimal nu se schimbă în funcție de parametrul h , atât timp cât cele patru instanțe rămân separabile.

⁶¹⁹ Direcția unei drepte din planul euclidian este dată de *panta ei*, adică tangenta unghiului format de dreapta cu axa Ox . O definiție alternativă este următoarea: direcția unei drepte este reprezentată de mulțimea tuturor dreptelor paralele cu ea.

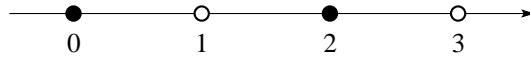
c. Marginea separatorului optimal este $m = \frac{1}{2} \text{dist}(x_3, d)$. Dreapta d are ecuația $-x + y = 0$, iar punctul x_3 este $(h, 1)$. Înțând cont de formula care ne dă distanța de la un punct oarecare la o dreaptă de ecuație cunoscută,⁶²⁰ rezultă că

$$m = \frac{1}{2} \cdot \frac{|(-1)h + 1|}{\sqrt{(-1)^2 + 1^2}} = \frac{1}{2} \cdot \frac{|1 - h|}{\sqrt{2}}$$

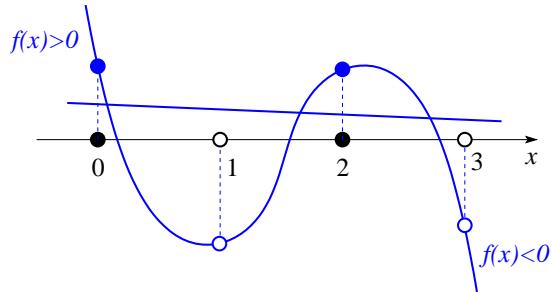
Ca urmare a combinării restricției $h \in [0, 3]$ din enunț cu restricția $h < 1$ rezultată la punctul a , vom avea $h \in [0, 1)$, deci expresia lui m de mai sus devine:

$$m = \frac{1-h}{2\sqrt{2}} = \frac{(1-h)\sqrt{2}}{4}$$

d. Punctele de antrenare $(0, +)$, $(1, -)$, $(2, +)$ și $(3, -)$ se reprezintă pe axa reală după cum urmează:



Pentru a clasifica corect aceste puncte este nevoie de o funcție polinomială de forma reprezentată în figura alăturată. Se observă că „proiecțiile“ instanțelor de antrenament pe curba polinomială sunt separabile liniar. Așadar, gradul minim p al unei funcții polinomiale care ne permite să clasificăm corect aceste puncte este 3. De pildă, putem considera funcția $-(x-0.5)(x-1.5)(x-2.5)$.



7.

(O formă echivalentă cu forma primală a problemei de optimizare SVM)

University of Utah, 2008 spring, Hal Daumé III, HW1C, pr. 3

Forma primală a problemei de optimizare SVM cu margine “hard” este definită astfel:⁶²¹

$$\begin{aligned} \min_{w, w_0} \quad & \frac{1}{2} \|w\|^2 \\ \text{s.t.} \quad & y_i(w \cdot x_i + w_0) \geq 1 \quad (\forall 1 \leq i \leq m) \end{aligned}$$

Arătați că soluția acestei probleme rămâne efectiv neschimbată atunci când numărul 1 din partea dreaptă a restricției este înlocuit cu o constantă oarecare pozitivă a .

Răspuns:

⁶²⁰Vedeți problema 1.

⁶²¹Vedeți problema .a.

Setul de inegalități $y_i(w \cdot x_i + w_0) \geq 1$, cu $i = 1, \dots, m$ constituie o condiție suficientă pentru separabilitatea liniară a instanțelor de antrenament $x_i \in \mathbb{R}^d$. Această inegalitate devine egalitate exclusiv pentru punctele care sunt vectori-suport. Hiperplanul de separare optimală are ecuația $w \cdot x + w_0 = 0$. Distanța dintre acest hiperplan și oricare dintre vectorii-suport este deci $1/\|w\|$.

Dacă numărul 1 din partea dreaptă a restricției este înlocuit cu o constantă oarecare pozitivă a , vom avea următoarea echivalentă între probleme:

$$\begin{aligned} \min_{w,w_0} \frac{1}{2} \|w\|^2 &\Leftrightarrow \min_{w,w_0} \frac{1}{2} \|w\|^2 \\ y_i(w \cdot x_i + w_0) \geq a \quad (i = 1, \dots, m) &\Leftrightarrow y_i\left(\frac{1}{a}w \cdot x_i + \frac{1}{a}w_0\right) \geq 1 \quad (i = 1, \dots, m) \\ \min_{w,w_0} \frac{1}{2} \left\| \frac{1}{a}w \right\|^2 &\Leftrightarrow \min_{w',w'_0} \frac{1}{2} \|w'\|^2 \\ y_i\left(\frac{1}{a}w \cdot x_i + \frac{1}{a}w_0\right) \geq 1 \quad (i = 1, \dots, m) &\Leftrightarrow y_i(w' \cdot x_i + w'_0) \geq 1 \quad (i = 1, \dots, m) \end{aligned}$$

unde $w' \stackrel{\text{not.}}{=} \frac{1}{a}w$ și $w'_0 \stackrel{\text{not.}}{=} \frac{1}{a}w_0$. Echivalentă se datorează pozitivitatea constantei a . Așadar, până la un factor pozitiv ($1/a$), soluția (w, w_0) rămâne neschimbată.

Evident, hiperplanul de separare optimală rămâne același. La fel, mulțimea vectorilor-suport este neschimbată. Distanța dintre hiperplan și vectorii-suport devine $a/\|w\|$, iar egalitatea $y_i(w \cdot x_i + w_0) = a$ va fi adevărată doar pentru punctele x_i care sunt vectori-suport.

8.

(Exemplu de folosire a unei funcții de mapare a atributelor
cu scopul de a obține separabilitate liniară;
rezolvarea directă a problemei SVM primale în spațiul [nou] de trăsături;
identificarea separatorului neliniar din spațiul inițial [de trăsături])

• CMU, 2009 fall, Carlos Guestrin, HW3, pr. 2.1

Se dă un set de date de antrenament D , caracterizate de atributul X care ia valori în \mathbb{R} și de eticheta corespunzătoare $y \in \{+1, -1\}$. Acest set de date este constituit din trei exemple pozitive și anume pentru $X \in \{-3, -2, 3\}$ și trei exemple negative pentru $X \in \{-1, 0, 1\}$.

- Putem separa acest set de date (în spațiul de trăsături specificat mai sus) folosind un separator liniar? De ce da, sau de ce nu?
- Definim funcția $\Phi(x) = (x, x^2)$ care transformă / „mapează“ punctele din \mathbb{R} în \mathbb{R}^2 . Aplicați funcția Φ datelor din D și trasați imaginea lor în \mathbb{R}^2 , nou spațiu de trăsături. Poate un separator liniar să separe în mod perfect punctele din nou spațiu de trăsături, obținut prin funcția Φ ? De ce da, sau de ce nu?
- Găsiți forma analitică a funcției-nucleu $K(x, x')$ care corespunde transformării Φ . Vă reamintim că $K(x, x') \stackrel{\text{def.}}{=} \Phi(x) \cdot \Phi(x')$.
- Identificați un hiperplan de separare optimală pentru mulțimea $\Phi(D)$. Acest hiperplan este o dreaptă în planul euclidian, caracterizată de o ecuație de forma $w_1Y_1 + w_2Y_2 + w_0 = 0$, unde $\Phi(X) = (Y_1, Y_2)$. Găsiți valorile lui w_0, w_1 și w_2 ; conform problemei de optimizare SVM, pentru fiecare vector-suport X având eticheta y trebuie să avem satisfăcută restricția $w_1Y_1 + w_2Y_2 + w_0 = y$.

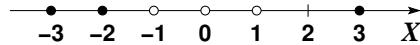
Desenați hiperplanul de separare optimală și încercuiți vectorii-suport. De asemenea, calculați *marginea* hiperplanului, adică distanța de la hiperplan la vectorii-suport.

e. Desenați în \mathbb{R} corespondentul hiperplanului de separare optimală din \mathbb{R}^2 .

f. Dacă adăugăm încă un punct pozitiv ($y=+1$) la mulțimea de antrenament, și anume $X = 5$, se va schimba hiperplanul de separare optimală sau marginea lui? De ce da, sau de ce nu?

Răspuns:

a. Figura de mai jos reprezintă datele de antrenament în spațiul original, \mathbb{R} , folosind puncte negre pentru exemplele pozitive și puncte albe pentru exemplele negative.

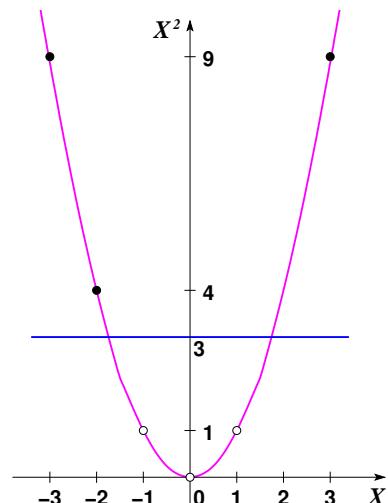


Conform *definiției*, un separator liniar în mulțimea numerelor reale este un punct de pe axă care are de o parte a sa toate exemplele pozitive și de celalătă parte toate exemplele negative. Se observă foarte ușor că nu există niciun punct pe axa reală care să satisfacă această condiție. Prin urmare, mulțimea datelor de antrenament nu este separabilă liniar în spațiul de trăsături inițial.

b. Figura alăturată reprezintă imaginea setului de antrenament în noul spațiu de trăsături, \mathbb{R}^2 . Se observă ușor că aici datele sunt separabile liniar. Putem lua, de exemplu, dreapta de ecuație $y - 3 = 0$, reprezentată în figura alăturată. Avem deci $w_1 = 0, w_2 = 1, w_0 = -3$.

Analitic, se observă imediat că sunt îndeplinite inegalitățile de forma $\text{sign}(w_1 u_i + w_2 u_i^2 + w_0) = y_i$:

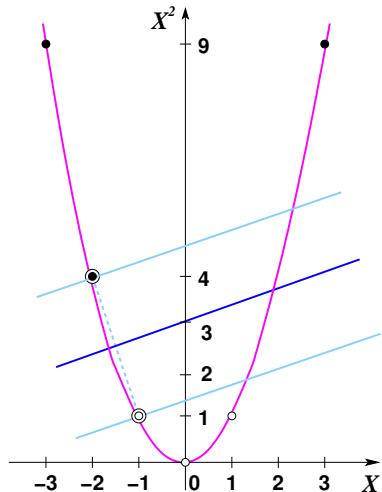
$$\begin{cases} 0 \cdot (-3) + 1 \cdot 9 - 3 = 6 > 0 \\ 0 \cdot (-2) + 1 \cdot 4 - 3 = 1 > 0 \\ 0 \cdot (-1) + 1 \cdot 1 - 3 = -2 < 0 \\ 0 \cdot 0 + 1 \cdot 0 - 3 = -3 < 0 \\ 0 \cdot 1 + 1 \cdot 1 - 3 = -2 < 0 \\ 0 \cdot 3 + 1 \cdot 9 - 3 = 6 > 0 \end{cases}$$



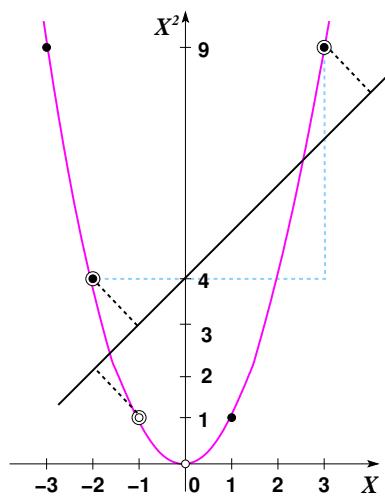
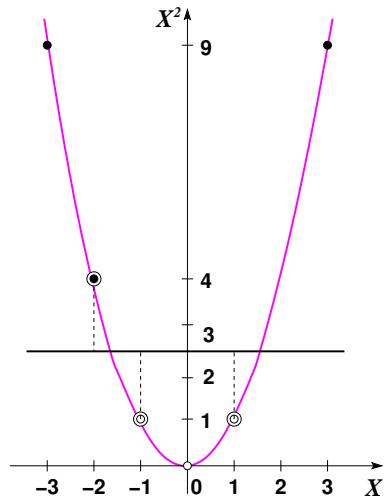
c. Conform *definiției* funcției-nucleu,

$$K(x, x') = \Phi(x) \cdot \Phi(x') = (x, x^2) \cdot (x', x'^2) \Rightarrow K(x, x') = xx' + x^2x'^2.$$

d. Hiperplanul de separare optimală pentru setul de date de antrenament D din enunț este determinat de punctele $(-2, 4)$ și $(-1, 1)$. Se observă că aceste două puncte constituie perechea de puncte de semne / etichete diferite aflate la distanță minimă, între toate perechile de puncte de semne contrare din D . Mediatoarea punctelor $(-2, 4)$ și $(-1, 1)$ este hiperplanul de separare maximală pentru mulțimea D , iar cele două puncte sunt vectorii-supt. Marginea, adică distanța dintre orice vector-supt și hiperplanul de separare optimă este $\frac{\sqrt{10}}{2} = \sqrt{\frac{5}{2}}$.



Observație: Se poate verifica [și] în mod *analytic* că, dintre toți separatorii liniari ai mulțimii D , dreapta de mai sus maximizează distanța până la cele mai apropiate instanțe pozitive și respectiv negative. În particular, comparând marginea corespunzătoare mediatoarei punctelor $(-2, 4)$ și $(-1, 1)$ cu marginile determinate respectiv de către cele două drepte din figurile de mai jos, vom obține: $\sqrt{\frac{5}{2}} > \frac{3}{2}$ pentru figura din partea stângă și $\sqrt{\frac{5}{2}} > \sqrt{2}$ pentru figura din partea dreaptă.⁶²²



Ecuatăia separatorului optimal se determină astfel:

⁶²²Mai mult, cele două drepte menționate au poziții extreme în raport cu toate dreptele care

- separă instanțele din D ,
- trec prin punctul de la jumătatea segmentului determinat de punctele $(-2, 4)$ și $(-1, 1)$, care sunt cele mai apropiate puncte de semne / etichete contrare din D ,
- nu se apropiă de punctul $(3, 9)$ și respectiv $(1, 1)$ mai mult decât este distanța până la fiecare din punctele $(-2, 4)$ și $(-1, 1)$.

- ecuația dreptei determinată de punctele $(-2, 4)$ și $(-1, 1)$ este

$$\frac{x - (-1)}{-2 - (-1)} = \frac{y - 1}{4 - 1} \Leftrightarrow 3(x + 1) = -(y - 1) \Leftrightarrow y = -(3x + 2),$$

așadar, panta acestei drepte este -3 ;

- orice dreaptă perpendiculară pe dreapta de mai sus are panta $-\frac{1}{-3} = \frac{1}{3}$ și deci este determinată de o ecuație de forma $y = \frac{1}{3}x + c$, unde c este o constantă reală;
- mijlocul segmentului determinat de punctele $(-2, 4)$ și $(-1, 1)$ este $\left(-\frac{3}{2}, \frac{5}{2}\right)$;
- mediatoarea segmentului determinat de punctele $(-2, 4)$ și $(-1, 1)$ va determina valoarea corespunzătoare pentru constanta c , prin faptul că punctul $\left(-\frac{3}{2}, \frac{5}{2}\right)$ aparține acestei mediatoare, deci:

$$\frac{5}{2} = \frac{1}{3}\left(-\frac{3}{2}\right) + c \Leftrightarrow c = \frac{5}{2} + \frac{1}{2} = 3;$$

- așadar, ecuația separatorului optimal este:

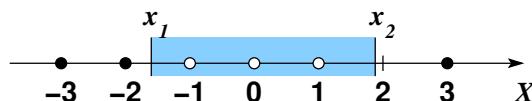
$$y = \frac{1}{3}x + 3 \Leftrightarrow x - 3y + 9 = 0 \Leftrightarrow -\frac{1}{5}x + \frac{3}{5}y - \frac{9}{5} = 0.$$

În legătură cu ultima formă a ecuației de mai sus se poate observa că expresia $-\frac{1}{5}x + \frac{3}{5}y - \frac{9}{5}$ satisfacă condiția — exprimată în definiția problemei SVM, forma primală — de a produce valorile 1 și respectiv -1 pentru punctele $(-2, 4)$ și $(-1, 1)$. Așadar, funcția care definește separatorul optimal este $-\frac{1}{5}x + \frac{3}{5}y - \frac{9}{5}$, iar soluțiile sunt $\bar{w} = \left(-\frac{1}{5}, \frac{3}{5}\right)$ și $\bar{w}_0 = -\frac{9}{5}$. Rezultă că marginea este $\frac{1}{\|\bar{w}\|} = \frac{5}{\sqrt{1+9}} = \sqrt{\frac{5}{2}}$.

e. Clasificarea unui punct $x \in \mathbb{R}$ cu separatorul determinat la punctul d se face în felul următor:

$$y = sign(\bar{w} \cdot \Phi(x) + \bar{w}_0) = sign\left(\left(-\frac{1}{5}, \frac{3}{5}\right) \cdot (x, x^2) - \frac{9}{5}\right) = sign\left(\frac{3}{5}x^2 - \frac{1}{5}x - \frac{9}{5}\right)$$

Prin urmare, vom avea eticheta $y = -1$ dacă și numai dacă $3x^2 - x - 9 < 0 \Leftrightarrow x \in (x_1, x_2)$, unde $x_1 = \frac{1 - \sqrt{1 + 12 \cdot 9}}{6} \approx -1.57$ și $x_2 = \frac{1 + \sqrt{1 + 12 \cdot 9}}{6} \approx 1.90$. Corespondentul hiperplanului de separare din \mathbb{R}^2 în spațiul inițial de trăsături este ilustrat în figura de mai jos, sub forma „tăieturilor“ reprezentate de x_1 și x_2 :



Observație: Rezultatul de mai sus corespunde (intuitiv) cu rezultatul de la punctul precedent (d): instanțele (x, x^2) situate de partea „pozitivă“ a dreptei

de ecuație $-\frac{1}{5}x + \frac{3}{5}y - \frac{9}{5} = 0 \Leftrightarrow y = \frac{1}{3}x + 3$ satisfac condiția $x^2 > y \Leftrightarrow x^2 > \frac{1}{3}x + 3 \Leftrightarrow 3x^2 - x - 9 > 0$. Similar, instanțele (x, x^2) situate de partea „negativă“ a dreptei de ecuație $-\frac{1}{5}x + \frac{3}{5}y - \frac{9}{5} \Leftrightarrow y = \frac{1}{3}x + 3$ satisfac condiția $x^2 < y \Leftrightarrow x^2 < \frac{1}{3}x + 3 \Leftrightarrow 3x^2 - x - 9 < 0$.

f. Se observă ușor (vedeți prima figură de la punctul d) că imaginea noului exemplu de antrenament $(5, 25)$ cade de partea „pozitivă“ a hiperplanului de separare și, mai mult, cade în afara marginii. Prin urmare, separatorul optimal rămâne același.

9.

(Exercițiu teoretic: deducerea formei duale pentru problema SVM cu margine “hard”)

*prelucrare de L. Ciortuz, după
■ CMU, 2010 fall, Ziv Bar-Joseph, HW4, pr. 1.3-5*

Se consideră vectorii de intrare $x_1, \dots, x_m \in \mathbb{R}^d$ și etichetele corespunzătoare $y_1, \dots, y_m \in \{-1, 1\}$. Problema SVM cu margine “hard” — termen care desemnează cazul în care instanțele $(x_1, y_1), \dots, (x_n, y_n)$ se presupune că sunt liniar separabile — este o problemă de optimizare convexă, exprimată sub forma primală astfel:

$$\begin{aligned} \min_{w, w_0} \quad & \frac{1}{2} \|w\|^2 \\ \text{a. i.} \quad & (w \cdot x_i + w_0)y_i \geq 1, \text{ pentru } i = 1, \dots, m, \end{aligned} \tag{P}$$

unde $w \in \mathbb{R}^d$ și $w_0 \in \mathbb{R}$. În urma rezolvării acestei probleme se va obține un model liniar, de forma $y(x) = w \cdot x + w_0$, ce va servi ulterior pentru clasificare, conform funcției de decizie $\text{sign}(y(x))$.⁶²³

La curs am prezentat metoda dualității Lagrange,⁶²⁴ care ne permite în anumite condiții să rezolvăm probleme de optimizare convexă cu restricții de asemenea convexe, transpunând aceste probleme într-o formă mai convenabilă numită *forma duală*. Pentru început, vom defini o altă funcție, numită *lagrangeanul generalizat*,⁶²⁵ care combină funcția obiectiv din (P) cu expresiile

⁶²³ Este ușor de constatat faptul că problema (P) este un caz particular de problemă de optimizare convexă cu restricții (vedeți *Introducerea* de la pr. 82 de la capitolul de *Fundamente*) și, conform teoremei Karush-Kuhn-Tucker (vedeți *Comentariul* de la pr. 83 tot de la capitolul de *Fundamente*), ea are soluție atunci când *regiunea fezabilă* (engl., feasible region), adică mulțimea convexă formată din valorile lui w și w_0 pentru care restricțiile $(w \cdot x_i + w_0)y_i \geq 1$ cu $i = 1, m$ sunt satisfăcute este nevidă.

Din punct de vedere *computațional*, rezolvarea acestei probleme devine neficientă atunci când m , numărul de instanțe de antrenament, este foarte mare. Dificultatea ține de satisfacerea restricțiilor. Ca să depășim această dificultate, *ideea* de bază este să punem problema noastră sub o altă formă, cu restricții mai simple. „Prețul“ pe care va trebui să-l plătim în schimb este „complicarea“ funcției obiectiv; se va dovedi ulterior că acest schimb este convenabil.

⁶²⁴Vedeți problemele 82 și 83 de la capitolul de *Fundamente*.

⁶²⁵ Unii autori numesc funcția L_P lagrangeanul *primal*. Pentru problema de optimizare convexă (cazul general), el se definește astfel:

$$L_P(x, \alpha, \beta) \stackrel{\text{def.}}{=} f(x) + \sum_i \alpha_i g_i(x) + \sum_j \beta_j h_j(x).$$

Vedeți *Introducerea* de la pr. 82 de la capitolul de *Fundamente*.

care intervin în partea stângă a restricțiilor:⁶²⁶

$$L_P(w, w_0, \alpha) \stackrel{\text{def.}}{=} \frac{1}{2} \|w\|^2 - \sum_{i=1}^m \alpha_i ((w \cdot x_i + w_0)y_i - 1),$$

unde $\alpha_i \geq 0$ pentru $i = \overline{1, m}$ sunt așa-numiții multiplicatori Lagrange sau *variabilele duale*. Variabilele *primale* sunt w și w_0 .

Comentariu: Vă readucem aminte următoarele *puncte principale* [ale teoriei dualității Lagrange], care justifică de ce anume procedăm aşa cum procedăm în rezolvarea problemei noastre.

- [P→P1] Este relativ ușor de arătat⁶²⁷ că problema (P) este echivalentă cu următoarea problemă de optimizare

$$\begin{aligned} & \min_x \max_{\alpha, \beta} L_P(x, \alpha, \beta) \\ & \text{a. i. } \alpha_i \geq 0 \text{ pentru } i = 1, \dots, m. \end{aligned} \quad (\text{P1})$$

Se observă că în această problemă restricțiile sunt mult mai simple decât în problema (P).

- [P1→D1] Inversând operatorii min și max în problema (P1), obținem problema de optimizare convexă

$$\begin{aligned} & \max_{\alpha, \beta} \min_x L_P(x, \alpha, \beta) \\ & \text{a. i. } \alpha_i \geq 0 \text{ pentru } i = 1, \dots, m. \end{aligned} \quad (\text{D1})$$

Dacă notăm cu p^* optimul problemei (P1) (deci și al problemei (P)) și cu d^* optimul problemei (D1), este relativ ușor de demonstrat⁶²⁸ proprietatea de *dualitate slabă*:

$$p^* \geq d^*. \quad (290)$$

Proprietatea de dualitate slabă este valabilă și în cazul general al problemelor de optimizare convexă [cu restricții].

- Este imediat că pentru problema (P1) este satisfăcută așa-numita *condiție a lui Slater*:⁶²⁹

$$\exists w \in \mathbb{R}^d \text{ și } w_0 \in \mathbb{R} \text{ astfel încât } (w \cdot x_i + w_0)y_i - 1 > 0, \text{ pentru } i = 1, \dots, m. \quad (291)$$

Într-adevăr, faptul că instanțele $(x_1, y_1), \dots, (x_n, y_n)$ sunt liniar separabile implică imediat satisfacerea condiției de mai sus. Se demonstrează că în general, adică pentru orice problemă de optimizare convexă cu restricții, dacă este îndeplinită condiția lui Slater (sub forma $\exists x$ a. i. $g_i(x) < 0$ pentru $i = 1, \dots, n$ și $h_j(x) = 0$ pentru $j = 1, \dots, p$), atunci are loc egalitatea

$$d^* = p^*,$$

care reprezintă proprietatea de *dualitate tare*. Așadar, optimul problemei (D1) este optim și pentru problema (P1) și, corespunzător, pentru problema (P).

⁶²⁶Semnul minus ($-$) din fața simbolului de sumare (\sum) din funcția L_P corespunzătoare problemei de optimizare SVM în forma primală (P) apare din cauza faptului că restricțiile din cadrul acestei probleme sunt de tip \geq , în timp ce restricțiile din cadrul formulării generale a problemei de optimizare convexă (vedeți *Introducerea* de la problema 82 de la capitolul de *Fundamente*) sunt de tip \leq .

⁶²⁷Vedeți documentul *Convex Optimization Overview (cont'd)*, de Chuong B. Do, 2009, pag. 4-5 sau documentul *Support Vector Machines* de Andrew Ng, (Stanford University, CS229 Lecture Notes, Part V), pag. 8-9.

⁶²⁸Vedeți problema 82 de la capitolul de *Fundamente* și / sau documentul *Convex Optimization Overview (cont'd)*, de Chuong B. Do, 2009, pag. 5-6.

⁶²⁹Vedeți *Learning with Kernels*, B. Schölkopf, A. Smola, MIT Press, 2002, pag 167.

O teoremă importantă din teoria problemelor de optimizare convexă⁶³⁰ demonstrează următoarea implicație: în cazul în care proprietatea de dualitate tare este satisfăcută, notând cu $\bar{\alpha}$, $\bar{\beta}$ o soluție a problemei (D1), urmează că

$$\bar{\alpha}_i g_i(\bar{x}) = 0 \text{ pentru } i = 1, \dots, n.$$

Aceste egalități se numesc *condițiile de complementaritate duală Karush-Kuhn-Tucker*.

Pentru problema SVM cu margine “hard” din enunț, considerând \bar{w} , \bar{w}_0 și $\bar{\alpha}$ soluțiile problemei (D1) (deci și ale problemei (P1)), condițiile de complementaritate duală Karush-Kuhn-Tucker se vor exprima astfel:

$$\bar{\alpha}_i ((\bar{w} \cdot x_i + \bar{w}_0) y_i - 1) = 0 \text{ pentru } i = 1, \dots, m, \quad (292)$$

adică, pentru fiecare valoare admisibilă a lui i , avem fie $\bar{\alpha}_i = 0$, fie $\bar{\alpha}_i > 0$ și, în consecință, $(\bar{w} \cdot x_i + \bar{w}_0) y_i - 1 = 0$.

- [D1→D] În fine, pentru că funcția L_P este derivabilă în raport cu w și respectiv w_0 , vom putea rescrie expresia lui $\min_{w,w_0} L_P(w, w_0, \alpha)$ din cadrul funcției obiectiv a problemei (D1) calculând în prealabil rădăcinile derivatele parțiale ale lui L_P . (Se poate arăta⁶³¹ că funcția $\min_{w,w_0} L_P(w, w_0, \alpha)$ este concavă.) Se va ajunge astfel la a asocia problemei (P) forma duală (D) din enunț; veДЕti punctul b.

Raționamentul acesta este valabil în general pentru problemele de optimizare convexă în care toate funcțiile f , g_i și h_j sunt convexe și derivabile.

Condițiile ca derivatele parțiale ale lui $L_P(w, w_0, \alpha)$ în raport cu w și respectiv w_0 să se anuleze se numesc *condițiile de staționaritate* (sau: *optimalitate*) *Karush-Kuhn-Tucker*.

- Calculând derivatele parțiale ale funcției L_P în raport cu variabilele primale w și w_0 , arătați că între valorile \bar{w} , \bar{w}_0 și $\bar{\alpha}$ pentru care aceste derivate parțiale se anulează există relațiile:**

$$\bar{w} = \sum_{i=1}^m \bar{\alpha}_i x_i y_i, \quad (293)$$

$$\sum_{i=1}^m \bar{\alpha}_i y_i = 0. \quad (294)$$

De asemenea, arătați că din *condiția de complementaritate Karush-Kuhn-Tucker* se poate deduce relația:

$$\bar{w}_0 = y_i - \bar{w} \cdot x_i \text{ pentru orice } i \text{ astfel încât } \bar{\alpha}_i > 0. \quad (295)$$

Notă: Instantele x_i pentru care $\bar{\alpha}_i > 0$ sunt numite vectori-suport [în sens clasic, analitic].⁶³² Aceasta este definiția pe care o vom folosi de acum încolo pentru această noțiune. Observați că în relația (293) soluția \bar{w} se scrie ca o combinație liniară de vectorii-suport!

- Calculați funcția $L_D(\alpha)$ — numită *lagrangeanul dual* — care se obține din expresia lagrangeanului generalizat $L_P(w, w_0, \alpha)$ substituind variabila w cu**

⁶³⁰Vedeți documentul *Convex Optimization Overview (cont'd)*, de Chuong B. Do, 2009, pag. 6 sau *Learning with Kernels*, B. Schölkopf, A. Smola, MIT Press, 2002, pag. 165, Teorema 6.21.

⁶³¹Vedeți documentul *Convex Optimization Overview (cont'd)*, de Chuong B. Do, 2009, pag. 5.

⁶³²Vedeți *Observația importantă* de la pag. 629.

$\sum_{i=1}^m \alpha_i x_i y_i$ și folosind egalitatea $\sum_{i=1}^m \alpha_i y_i = 0$, conform relațiilor (293) și (294) de la punctul precedent.

Observație (1): Este evident acum că problemei (P) îi poate fi asociată următoarea formă (numită „duală“):

$$\begin{aligned} & \max_{\alpha} L_D(\alpha) \\ \text{a. i. } & \alpha_i \geq 0, \text{ pentru } i = 1, \dots, m \\ & \sum_{i=1}^m \alpha_i y_i = 0 \end{aligned} \tag{D}$$

în care *restrictiile* sunt mult mai *simple* decât erau în forma primală (P). Este de reținut faptul că relațiile (293) și (295) de la punctul precedent constituie legătura dintre soluția problemei (D) și soluția problemei (P).

c. Dacă se dă o instanță nouă (de test) x_{new} , cum veți decide clasa ei?

Răspuns:

a. Calculăm mai întâi derivatele parțiale ale funcției L_P în raport cu w și respectiv w_0 :⁶³³

$$\begin{aligned} \frac{\partial}{\partial w} L_P(w, w_0, \alpha) &= w - \sum_{i=1}^m \alpha_i x_i y_i \\ \frac{\partial}{\partial w_0} L_P(w, w_0, \alpha) &= - \sum_{i=1}^m \alpha_i y_i. \end{aligned}$$

Atunci când se atinge optimul funcției L_P (considerând argumentul său α fixat), aceste derivate parțiale devin egale cu 0. Din $\frac{\partial}{\partial w} L_P(\bar{w}, \bar{w}_0, \bar{\alpha}) = 0$ rezultă că

$$\bar{w} = \sum_{i=1}^m \bar{\alpha}_i x_i y_i.$$

Similar, $\frac{\partial}{\partial w_0} L_P(\bar{w}, \bar{w}_0, \bar{\alpha}) = 0$ implică relația $\sum_{i=1}^m \bar{\alpha}_i y_i = 0$.

În fine, din condiția de complementaritate Karush-Kuhn-Tucker, care se exprimă aici sub forma

$$\bar{\alpha}_i [(\bar{w} \cdot x_i + \bar{w}_0) y_i - 1] = 0 \text{ pentru } i = 1, \dots, m$$

⁶³³În notația matriceală, considerând w și x_i pentru $i = 1, \dots, m$ vectori-colonă, putem scrie lagrangeanul L_P astfel:

$$L_P(w, w_0, \alpha) = \frac{1}{2} w^\top w - \sum_{i=1}^m \alpha_i ((w^\top x_i + w_0) y_i - 1).$$

Pentru derivarea lui L_P în raport cu vectorul w , se folosesc regulile (5a) și (5b) din documentul *Matrix Identities* de Sam Roweis (New York University, June 1999), pe care le-am folosit (de exemplu) la problema 3.d de la capitolul *Metode de regresie*, precum și la problema 24 de la capitolul *Clusterizare*.

Revenind la formula inițială, care folosește notația vectorială și produsul scalar din \mathbb{R}^d , se poate constata că într-un astfel de caz (simplu!), regulile de derivare sunt similare cu cele din \mathbb{R} .

rezultă că pentru orice $i \in \{1, \dots, m\}$ cu $\bar{\alpha}_i > 0$ avem $(\bar{w} \cdot x_i + \bar{w}_0)y_i - 1 = 0$.⁶³⁴ Această relație este echivalentă cu $\bar{w} \cdot x_i + \bar{w}_0 = y_i$ fiindcă $y_i \in \{-1, 1\}$. Din această ultimă egalitate rezultă:

$$\bar{w}_0 = y_i - \bar{w} \cdot x_i.$$

b. Substituind $w = \sum_{i=1}^m \alpha_i x_i y_i$ în expresia lui L_P și ținând cont că $\sum_{i=1}^m \alpha_i y_i = 0$, vom obține:

$$\begin{aligned} L_D(\alpha) &= \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j - \sum_{i=1}^m \alpha_i \sum_{j=1}^m [(\alpha_j y_j x_j \cdot x_i + w_0)y_i - 1] \\ &= \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \end{aligned} \quad (296)$$

c. Mai întâi vom calcula

$$\begin{aligned} f(x_{new}) &= \bar{w} \cdot x_{new} + \bar{w}_0 \\ &= \left(\sum_{i=1}^m \bar{\alpha}_i y_i x_i \right) \cdot x_{new} + \bar{w}_0 = \sum_{i=1}^m \bar{\alpha}_i y_i x_i \cdot x_{new} + \bar{w}_0, \end{aligned} \quad (297)$$

unde $\bar{\alpha}$ este soluția problemei duale (D), iar \bar{w} și \bar{w}_0 , soluțiile problemei primale (P) sunt calculate conform relațiilor (293) și (295) de la punctul a.

După aceea, dacă $f(x_{new}) \geq 0$ atunci x_{new} va fi clasificat pozitiv, iar în caz contrar va fi clasificat negativ.

Observație (2): Remarcăm faptul că atât în funcția obiectiv a problemei de optimizare SVM în formă duală (D) cât și în funcția f care servește la clasificarea instanțelor noi, operațiile care se execută asupra instanțelor sunt doar de tip produs scalar: $x_i \cdot x_j$ și respectiv $x_i \cdot x_{new}$. Acest fapt face posibilă folosirea *funcțiilor-nucleu* în contextul SVM (așa cum vom exemplifica la problema 10), ceea ce este convenabil atât din punctul de vedere al obținerii (eventuale) a separabilității, cât și din punctul de vedere al executării eficiente a calculelor.

10.

(Învățarea conceptului \neg XOR
folosind forma duală a problemei SVM
și o mapare particulară a trăsăturilor)

CMU, 2006 fall, E. Xing, T. Mitchell, midterm exam, pr. 5

Fie o problemă de învățare supervizată în care exemplele de antrenare se află în spațiul euclidian bidimensional. Exemplele pozitive sunt $x_1 = (1, 1)$ și $x_3 = (-1, -1)$ iar exemplele negative sunt $x_2 = (-1, 1)$ și $x_4 = (1, -1)$.

⁶³⁴Dacă $\bar{\alpha}_i = 0$ pentru $i = \overline{1, m}$, din relația $\bar{w} = \sum_{i=1}^m \bar{\alpha}_i y_i x_i$ rezultă că $\bar{w} = 0$. În consecință, funcția $f(x) = \bar{w} \cdot x + \bar{w}_0$ care dă ecuația separatorului optimal ($f(x) = 0$) nu va avea putere de discriminare între instanțele pozitive și instanțele negative (indiferent de valoarea atribuită lui \bar{w}_0 , care este o constantă).

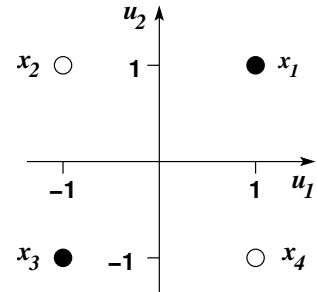
De fapt, atunci când s-a introdus problema SVM ca o problemă de optimizare a marginii / distanței $\frac{1}{\|w\|}$, s-a considerat în mod implicit că se caută soluții $w \neq 0$.

Așadar, în cazul în care în urma rezolvării problemei primale, respectiv a celei duale — urmată în ultimul caz de aplicarea relațiilor de legătură între cele două tipuri / seturi de soluții — obținem $\bar{w} = 0$, căutarea lui \bar{w}_0 (valoarea optimă pentru w_0) pur și simplu nu are sens.

- a. Sunt exemplele pozitive separabile liniar față de exemplele negative?
- b. Considerăm transformarea $\Phi : \mathbb{R}^2 \rightarrow \mathbb{R}^4$, definită astfel: $\Phi(u) = (1, u_1, u_2, u_1u_2)$, unde $u = (u_1, u_2)$. Funcția de predicție pe care vrem să o obținem este de forma $y(x) = w \cdot \Phi(x) + w_0$, unde $x \in \mathbb{R}^4$, $w \in \mathbb{R}^4$, $w_0 \in \mathbb{R}$ (cu x variabil și w și w_0 fixați), iar \cdot reprezintă produsul scalar. Indicați coeficienții w și w_0 corespunzători unui hiperplan de separare cu margine maximă pentru imaginea punctelor de antrenament din enunț în „spațiul de trăsături“ (\mathbb{R}^4).
- Indicație:* Vă recomandăm să rezolvați problema SVM în formă duală,⁶³⁵ iar apoi să exprimați soluția formei primale folosind relațiile dintre cele două tipuri de soluții, care au fost date la curs.⁶³⁶ (*Observație:* Se poate constata în prealabil că mulțimea $\{\Phi(x_1), \Phi(x_2), \Phi(x_3), \Phi(x_4)\}$ este liniar separabilă și (ca atare) în spațiul de trăsături determinat de maparea Φ este satisfăcută condiția lui Slater.)
- c. Adăugați un nou exemplu la mulțimea de date de antrenament, astfel încât această nouă mulțime să nu mai fie separabilă liniar în spațiul corespunzător transformării Φ de la punctul anterior.
- d. Cărei funcții-nucleu (engl., kernel function) îi corespunde transformarea Φ ?

Răspuns:

- a. Reprezentând grafic datele într-un sistem de coordinate bidimensional, se poate observa imediat că ele nu sunt separabile liniar. *Analitic*, putem demonstra acest lucru după cum urmează:
Să presupunem prin reducere la absurd că exemplele negative sunt separabile liniar de cele pozitive în \mathbb{R}^2 . Atunci ar exista $\bar{w} = (w_1, w_2) \in \mathbb{R}^2$ și $w_0 \in \mathbb{R}$ astfel încât:



$$\left. \begin{array}{l} \left. \begin{array}{l} w_1 + w_2 + w_0 > 0 \\ -w_1 - w_2 + w_0 > 0 \\ -w_1 + w_2 + w_0 < 0 \\ w_1 - w_2 + w_0 < 0 \end{array} \right\} \Rightarrow 2w_0 > 0 \Rightarrow w_0 > 0 \\ \left. \begin{array}{l} \\ \\ \\ \end{array} \right\} \Rightarrow 2w_0 < 0 \Rightarrow w_0 < 0 \end{array} \right\} \Rightarrow \text{contradicție!}$$

Așadar, presupunerea făcută este falsă. În consecință, cele 4 puncte nu sunt liniar separabile în \mathbb{R}^2 .

- b. Conform formei duale a problemei SVM, avem de maximizat funcția „lagrangeană“ L_D , unde:

$$L_D(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = \sum_{i=1}^4 \alpha_i - \frac{1}{2} \sum_{i,j=1}^4 \alpha_i \alpha_j y_i y_j \Phi(x_i) \cdot \Phi(x_j) ,$$

ținând de asemenea cont de restricțiile $\sum_{i=1}^4 y_i \alpha_i = \alpha_1 - \alpha_2 + \alpha_3 - \alpha_4 = 0$ și $\alpha_i \geq 0$ pentru $i = \overline{1, 4}$.

⁶³⁵Vedeți problema de optimizare (D) de la pagina 644 și relația (296).

⁶³⁶Adică relațiile (293) și (295).

Conform definiției funcției de mapare Φ , se observă că $\Phi(x_i) \cdot \Phi(x_i) = 4$ pentru $i = \overline{1, 4}$ și $\Phi(x_i) \cdot \Phi(x_j) = 0$ pentru orice $i, j = \overline{1, 4}$ cu $i \neq j$.

Înlocuind aceste produse în expresia lui L_D de mai sus, obținem:

$$\begin{aligned} L_D(\alpha_1, \alpha_2, \alpha_3, \alpha_4) &= \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - \frac{1}{2}(4\alpha_1^2 + 4\alpha_2^2 + 4\alpha_3^2 + 4\alpha_4^2) = \\ &= \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - 2\alpha_1^2 - 2\alpha_2^2 - 2\alpha_3^2 - 2\alpha_4^2. \end{aligned}$$

Vom încerca să rezolvăm problema duală căutând punctul de optim al lui L_D ,⁶³⁷ adică determinând $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ astfel încât derivatele parțiale ale lui L_D în funcție de aceste variabile să se anuleze:

$$\left. \begin{array}{lcl} \frac{\partial}{\partial \alpha_1} L_D(\alpha_1, \alpha_2, \alpha_3, \alpha_4) &=& 1 - 4\alpha_1 = 0 \\ \frac{\partial}{\partial \alpha_2} L_D(\alpha_1, \alpha_2, \alpha_3, \alpha_4) &=& 1 - 4\alpha_2 = 0 \\ \frac{\partial}{\partial \alpha_3} L_D(\alpha_1, \alpha_2, \alpha_3, \alpha_4) &=& 1 - 4\alpha_3 = 0 \\ \frac{\partial}{\partial \alpha_4} L_D(\alpha_1, \alpha_2, \alpha_3, \alpha_4) &=& 1 - 4\alpha_4 = 0 \end{array} \right\} \Rightarrow \bar{\alpha}_1 = \bar{\alpha}_2 = \bar{\alpha}_3 = \bar{\alpha}_4 = \frac{1}{4}.$$

Se observă că aceste soluții satisfac restricțiile din forma duală a problemei SVM: $\bar{\alpha}_i \geq 0$ pentru $i = \overline{1, 4}$ și $\sum_{i=1}^4 y_i \bar{\alpha}_i = \bar{\alpha}_1 - \bar{\alpha}_2 + \bar{\alpha}_3 - \bar{\alpha}_4 = 0$.

Pentru a determina vectorul \bar{w} din soluția formei primale a problemei SVM, vom folosi formula (293), care leagă soluția formei duale de soluția formei primale:

$$\begin{aligned} \bar{w} &= \sum_{i=1}^4 y_i \bar{\alpha}_i \Phi(x_i) \\ &= \frac{1}{4}(1, 1, 1, 1) - \frac{1}{4}(1, -1, 1, -1) + \frac{1}{4}(1, -1, -1, 1) - \frac{1}{4}(1, 1, -1, -1) = (0, 0, 0, 1). \end{aligned}$$

Aflarea valorii lui \bar{w}_0 se face folosind una dintre ecuațiile (295), adică $\alpha_i(y_i(\bar{w} \cdot \Phi(x_i) + \bar{w}_0) - 1) = 0$ cu $i = \overline{1, 4}$. Luând, spre exemplu, $i = 1$ obținem:

$$\begin{aligned} \frac{1}{4}(1((0, 0, 0, 1) \cdot \Phi((1, 1)) + \bar{w}_0) - 1) &= 0 \Leftrightarrow \\ (0, 0, 0, 1) \cdot (1, 1, 1, 1) + \bar{w}_0 - 1 &= 0 \Leftrightarrow 1 + \bar{w}_0 = 1 \Leftrightarrow \bar{w}_0 = 0. \end{aligned}$$

Așadar, ecuația separatorului optimal este

$$\bar{w} \cdot \Phi(u) + \bar{w}_0 = 0 \Leftrightarrow (0, 0, 0, 1) \cdot (1, u_1, u_2, u_1 u_2) + 0 = 0 \Leftrightarrow u_1 u_2 = 0,$$

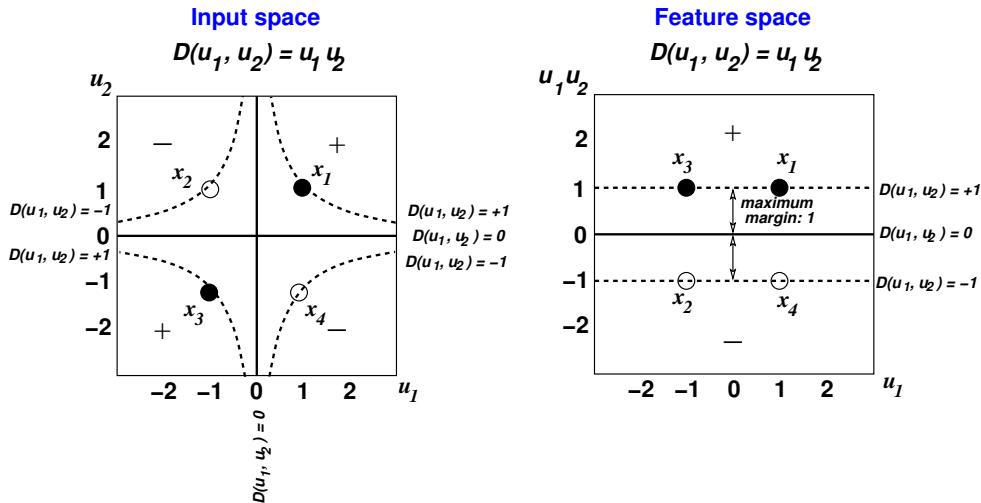
ceea ce corespunde mulțimii de puncte (u_1, u_2) formate din cele două axe de coordonate. Cele două „margini“ au ecuațiile $u_1 u_2 = -1$ și respectiv $u_1 u_2 = 1$ și sunt reprezentate grafic sub forma celor două hiperbole din figura de mai

⁶³⁷ Observație importantă: Se poate constata imediat că hessianul lui $L_D(\alpha)$ este o matrice negativ definită, prin urmare lagrangeanul $L_D(\alpha)$ este funcție concavă și, deci, admite un maxim. A priori, este posibil ca acest punct de maxim să fie în afara intervalului de fezabilitate, adică să nu satisfacă restricțiile $\alpha_i \geq 0$, și / sau să nu satisfacă condiția Karush-Kuhn-Tucker $\sum_i y_i \alpha_i = 0$. Vom vedea că, din fericire, în problema de față punctul de optim al lui $L_D(\alpha)$ satisfac aceste două restricții.

jos, partea stângă. Clasificarea unei instanțe oarecare de test $u = (u_1, u_2)$ se va face conform expresiei:

$$y = \text{sign}(\bar{w} \cdot \Phi(u) + \bar{w}_0) = \text{sign}(u_1 u_2).$$

Punctele din primul și din al treilea cadran vor fi clasificate pozitiv, iar punctele din al doilea și al patrulea cadran vor fi clasificate negativ.



Se constată ușor acum că în „spațiul de trăsături“ rezultat prin mapare avem separabilitate liniară, după cum indică figura de mai sus, partea dreaptă. (Pentru conveniență, am reținut doar coordonatele a două și a patra, adică u_1 și $u_1 u_2$. Deși nu era necesar, am fi putut adăuga și coordonata u_2 , dar atunci ar fi trebuit să facem graficul în 3D.) În acest spațiu, separatorul liniar ($u_1 u_2 = 0$) reprezintă o dreaptă; la fel și marginile ($u_1 u_2 = -1$ și $u_1 u_2 = 1$). Semiplanul superior corespunde punctelor clasificate pozitiv, iar semiplanul inferior corespunde punctelor clasificate negativ.

c. Se poate observa ușor că, dacă se consideră adițional instanța de antrenament $x_5 = (0.5, 4)$ clasificată negativ, mulțimea $\{\Phi(x_1), \Phi(x_2), \Phi(x_3), \Phi(x_4), \Phi(x_5)\}$ din „spațiul de trăsături“ este liniar neseparabilă.

d. Considerând $u = (u_1, u_2)$ și $u' = (u'_1, u'_2)$, urmează că

$$\Phi(u) \cdot \Phi(u') = (1, u_1, u_2, u_1 u_2) \cdot (1, u'_1, u'_2, u'_1 u'_2) = 1 + u_1 u'_1 + u_2 u'_2 + u_1 u_2 u'_1 u'_2 = K(u, u').$$

11. (Un [alt] exemplu de rezolvare a problemei duale SVM)

- CMU, 2009 fall, Carlos Guestrin, HW3, pr. 2.1.8
- CMU, 2011 fall, T. Mitchell, A. Singh, HW6, pr. 2.1

Calculați soluția \bar{a} pentru forma duală a problemei de optimizare SVM pentru setul de date D și transformarea Φ (așadar, în spațiul de „trăsături“ determinat de Φ) de la problema 8.

Răspuns:

Putem rezolva această problemă fie (i.) în mod de sine stătător, adică rezolvând problema de optimizare în forma duală, făcând abstracție de rezolvarea problemei în forma primală (obținută la pr. 8), fie (ii.) folosind soluțiile problemei primale (din rezolvarea problemei 8) și relațiile dintre soluțiile celor două forme ale problemei de optimizare SVM (fără însă a folosi efectiv lagrangeanul dual), fie (iii.) în mod combinat, adică rezolvând problema în forma duală după ce în prealabil am simplificat expresia lagrangeanului dual ținând cont de soluția problemei în forma primală.

i. Doar schițând calculele, veți vedea că această variantă de rezolvare nu este „fezabilă“ în mod analitic pe aceste date (deși, de exemplu pe datele de la problema 10 ea a funcționat) din cauza restricțiilor care însotesc multiplicatorii Lagrange.

Lagrangeanul dual este

$$\begin{aligned} L_D(\alpha) &= \sum_{i=1}^6 \alpha_i - \frac{1}{2} \sum_{i=1}^6 \sum_{j=1}^6 \alpha_i \alpha_j y_i y_j \Phi(x_i) \cdot \Phi(x_j) \\ &= \sum_{i=1}^6 \alpha_i - \frac{1}{2} (90\alpha_1^2 + 20\alpha_2^2 + 2\alpha_3^2 + 2\alpha_5^2 + 90\alpha_6^2 + \\ &\quad 2(42\alpha_1\alpha_2 - 12\alpha_1\alpha_3 - 6\alpha_1\alpha_5 + 72\alpha_1\alpha_6 \\ &\quad - 6\alpha_2\alpha_3 - 2\alpha_2\alpha_5 + 30\alpha_2\alpha_6 - 6\alpha_3\alpha_6 - 12\alpha_5\alpha_6)), \end{aligned}$$

iar condiția de optimalitate Karush-Kuhn-Tucker asociată este $\sum_{i=1}^6 y_i \alpha_i = 0$,⁶³⁸ adică $\alpha_1 + \alpha_2 - \alpha_3 - \alpha_4 - \alpha_5 + \alpha_6 = 0$.

Pentru a găsi soluția problemei duale am putea încerca să rezolvăm sistemul obținut prin egalarea derivatelor parțiale ale lui $L_D(\alpha)$ cu 0, după care să verificăm dacă soluția $\bar{\alpha}$ satisfac restricțiile $\alpha_i \geq 0$ pentru $i = 1, \dots, 6$, precum și condiția de optimalitate Karush-Kuhn-Tucker.⁶³⁹

Vă puteți convinge, făcând calculele, că soluția acestui sistem $\bar{\alpha} = (\bar{\alpha}_1, \dots, \bar{\alpha}_6)$ nu satisfac restricțiile $\alpha_i \geq 0$. Prin urmare, soluția problemei duale ar trebui căutată altfel (folosind eventual o metodă numerică). Din fericire, vom vedea mai jos că metodele ii și iii funcționează (pe aceste date) foarte bine.⁶⁴⁰

ii. Vom arăta că soluția $\bar{\alpha} = (\bar{\alpha}_1, \dots, \bar{\alpha}_6)$ a problemei duale se poate determina folosind relațiile dintre soluțiile celor două forme ale problemei SVM:

$$\bar{w} = \sum_{i=1}^6 \bar{\alpha}_i y_i \Phi(x_i) \quad \text{și} \quad \bar{\alpha}_i (y_i (\bar{w} \cdot \Phi(x_i) + \bar{w}_0) - 1) = 0 \text{ pentru } i = \overline{1, 6}.$$

De la rezolvarea problemei în forma primală (vedeți pr. 8) stim că $\bar{\alpha}_1 = \bar{\alpha}_4 = \bar{\alpha}_5 = \bar{\alpha}_6 = 0$, fiindcă vectorii-suport sunt (doar) instanțele x_2 și x_3 . În consecință, relația $\bar{w} = \sum_{i=1}^6 \bar{\alpha}_i y_i \Phi(x_i)$ devine:

⁶³⁸Vedeți relația (294) de la problema 9.

⁶³⁹Se poate observa că $\frac{\partial L}{\partial \alpha_4} = 1 \neq 0!$ De aceea este convenabil ca mai întâi să înlocuim în L_D variabila α_4 cu $\alpha_1 + \alpha_2 - \alpha_3 - \alpha_5 + \alpha_6$ și apoi să rescriem sistemul de ecuații formate cu ajutorul derivatelor parțiale ale noului L_D .

⁶⁴⁰De fapt, în general / practică, alternativa cea mai bună este aplicarea unei metode „numerice“ de optimizare, care să ia în calcul restricțiile de tip $\alpha_i \geq 0$ pentru $i = 1, \dots, m$. O astfel de metodă este algoritmul SMO (pentru rezolvarea problemei de optimizare C-SVM), a cărui aplicare o vom exemplifica la problema 23.

$$\begin{pmatrix} -1/5 \\ 3/5 \end{pmatrix} = \bar{\alpha}_2 \begin{pmatrix} -2 \\ 4 \end{pmatrix} - \alpha_3 \begin{pmatrix} -1 \\ 1 \end{pmatrix} \Leftrightarrow \begin{cases} -2\bar{\alpha}_2 + \bar{\alpha}_3 = -\frac{1}{5} \\ 4\bar{\alpha}_2 - \bar{\alpha}_3 = \frac{3}{5} \end{cases} \Leftrightarrow \bar{\alpha}_2 = \bar{\alpha}_3 = \frac{1}{5}.$$

Soluția astfel obținută, $\bar{\alpha} = \left(0, \frac{1}{5}, \frac{1}{5}, 0, 0, 0\right)$, verifică și restricția Karush-Kuhn-Tucker $\sum_{i=1}^6 \bar{\alpha}_i y_i = 0$ din forma duală a problemei SVM.

În consecință, funcția de decizie obținută de SVM se poate exprima ca

$$\begin{aligned} y(x) &= \text{sign}\left(\sum_i \bar{\alpha}_i y_i \Phi(x_i) \cdot \Phi(x) + \bar{w}_0\right) = \text{sign}\left(\frac{1}{5}K(x, -2) - \frac{1}{5}K(x, -1) - \frac{9}{5}\right) = \\ &= \text{sign}(-2x + 4x^2 + x - x^2 - 9) = \text{sign}(3x^2 - x - 9). \end{aligned}$$

Așadar, regăsim rezultatul de la punctul e de la problema 8.

iii. Tinând cont de relațiile $\alpha_2 = \alpha_3 \stackrel{\text{not.}}{=} \alpha > 0$ și $\alpha_1 = \alpha_4 = \alpha_5 = \alpha_6 = 0$ care rezultă (ca o consecință) din soluția problemei în forma primală (așa cum am arătat la ii), urmează că putem scrie lagrangeanul dual astfel:

$$L_D(\alpha) = 2\alpha - \frac{1}{2}(20\alpha^2 + 2\alpha^2 - 12\alpha^2) = 2\alpha - 5\alpha^2 = \alpha(2 - 5\alpha).$$

Maximul funcției $L_D(\alpha)$ se obține pentru $\alpha = -\frac{2}{2 \cdot (-5)} = \frac{1}{5} > 0$. Așadar, regăsim rezultatul de la ii.

5.1.2 SVM cu margine “soft”

12. (Exercițiu teoretic: deducerea formei duale pentru problema de optimizare SVM cu margine “soft” (C-SVM))

■ *prelucrare de Liviu Ciortuz, după CMU, 2012 spring, Ziv Bar-Joseph, HW3, pr. 3.2*

Antrenăm o SVM pe un set de date de intrare $\{x_i\}$ cu $i = 1, \dots, n$, considerate împreună cu setul de etichete asociate $\{y_i\}$, cu $y_i \in \{-1, 1\}$. Se [poate] presupune că aceste date sunt neseparabile liniar. Permitând ca la finalul antrenării unele (în genere puține) dintre datele de antrenament să fie clasificate eronat, *obiectivul* pe care ni-l propunem aici este ca, simultan cu maximizarea *marginii geometrice*, să „controlăm“ efectul cumulativ determinat de aceste excepții.

Așadar, vom considera problema de optimizare

$$\begin{aligned} \min_{w, w_0, \xi} & \left(\frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i \right) \\ \text{a. i. } & (w \cdot x_i + w_0) y_i \geq 1 - \xi_i, \text{ pentru } i = 1, \dots, m \\ & \xi_i \geq 0, \text{ pentru } i = 1, \dots, m, \end{aligned} \tag{P'}$$

unde $\xi \stackrel{\text{not.}}{=} (\xi_1, \dots, \xi_m)$, iar $C > 0$ este un parametru („cost“) care controlează compromisul (engl., trade-off) pe care urmărim să-l facem între mărimea *marginii* pe de o parte,⁶⁴¹ și *penalizările pentru „destindere“* (engl., slack penalty) reprezentate prin variabilele $\xi_i \geq 0$ pe de altă parte.

a. Verificați faptul că pentru problema (P') este satisfăcută *condiția lui Slater*.⁶⁴²

b. Folosind variabile duale (adică, multiplicatori Lagrange), scrieți expresia *lagrangeanului generalizat* care corespunde problemei (P') . Specificați pentru fiecare multiplicator în parte care este restricția care-i corespunde lui în problema (P') .

c. Scrieți *condițiile de complementaritate Karush-Kuhn-Tucker* corespunzătoare problemei (P') .

d. Calculând derivatele parțiale ale lagrangeanului generalizat $L_P(w, w_0, \xi, \alpha, \beta)$ în raport cu variabilele w, w_0 și respectiv ξ , arătați că forma duală a problemei (P') este:

$$\begin{aligned} \max_{\alpha} & \left(\sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \right) \\ \text{a. i. } & 0 \leq \alpha_i \leq C \text{ pentru } i = 1, \dots, m \\ & \sum_{i=1}^m \alpha_i y_i = 0. \end{aligned} \tag{D'}$$

Faceți toate calculele în mod detaliat. Analizați deosebirile dintre (D') și forma duală a problemei SVM cu margine “hard” (a se vedea definiția (D) de la problema 9, pag. 644).

e. Dată fiind o soluție a problemei duale (D') , cum identificăm vectorii-suport?

f. Cum va fi clasificată o instanță nouă x' ?

Răspuns:

a. Condiția lui Slater pentru problema (P') se formulează astfel:

$$\exists w \in \mathbb{R}^d, w_0 \in \mathbb{R} \text{ și } \xi \in \mathbb{R}^m \text{ a. i. } (w \cdot x_i + w_0)y_i - 1 + \xi_i > 0 \text{ și } \xi_i > 0, \text{ pentru } i = 1, \dots, m.$$

Luând $w = 0, w_0 = 0$ și $\xi_i = 2$ pentru $i = 1, \dots, m$,⁶⁴³ se constată că această condiție se verifică imediat. În consecință, optimul problemei primale (P') va coincide cu optimul problemei duale (de la punctul d).

b. Lagrangeanul generalizat care corespunde problemei de optimizare (P') este:

$$L_P(w, w_0, \xi, \alpha, \beta) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i - \sum_{i=1}^m \alpha_i ((w \cdot x_i + w_0)y_i - 1 + \xi_i) - \sum_{i=1}^m \beta_i \xi_i.$$

⁶⁴¹Prin *margine* aici vom înțelege distanța dintre hiperplanul de separare optimală definit de soluția (w, w_0) a problemei (P') și oricare dintre instanțele x_i pentru care $y_i(w \cdot x_i + w_0) = 1$. Această distanță este egală cu $1/\|w\|$.

⁶⁴²Pentru formalizarea acestei condiții — ca și pentru diverse noțiuni implicate la punctele următoare —, a se vedea *Comentariul* de la problema 9 (pag. 642), care tratează cazul formei duale a problemei SVM cu margine “hard”.

⁶⁴³De fapt, este suficient ca — pe lângă $w = 0, w_0 = 0$ — să considerăm $\xi_i > 1$ pentru $i = 1, \dots, m$.

Multiplicatorii $\alpha_i \geq 0$ corespund restricțiilor $(w \cdot x_i + w_0)y_i \geq 1 - \xi_i$, în vreme ce multiplicatorii $\beta_i \geq 0$ corespund restricțiilor $\xi_i \geq 0$.

c. $\alpha_i((w \cdot x_i + w_0)y_i - 1 + \xi_i) = 0$ și $\beta_i\xi_i = 0$, pentru $i = 1, \dots, m$.

d. Calculând derivatele parțiale indicate în enunț, vom avea:

$$\begin{aligned} \frac{\partial}{\partial w} L_P(w, w_0, \xi, \alpha, \beta) &= 0 \Leftrightarrow w - \sum_{i=1}^m \alpha_i y_i x_i = 0 \Leftrightarrow w = \sum_{i=1}^m \alpha_i y_i x_i \\ \frac{\partial}{\partial w_0} L_P(w, w_0, \xi, \alpha, \beta) &= 0 \Leftrightarrow - \sum_{i=1}^m \alpha_i y_i = 0 \Leftrightarrow \sum_{i=1}^m \alpha_i y_i = 0 \\ \frac{\partial}{\partial \xi_i} L_P(w, w_0, \xi, \alpha, \beta) &= 0 \Leftrightarrow C - \alpha_i - \beta_i = 0 \Leftrightarrow \alpha_i + \beta_i = C \text{ pentru } i = 1, \dots, m. \end{aligned}$$

Se poate constata că primele două egalități care tocmai au fost obținute, și anume $w = \sum_{i=1}^m \alpha_i y_i x_i$ și $\sum_{i=1}^m \alpha_i y_i = 0$, coincid cu relațiile (293) și (294) derivate din *condiția de staționaritate / optimizare* corespunzătoare problemei de optimizare SVM cu margine “hard” (vedeți problema 9).

Substituind aceste două rezultate în expresia de definiție a lui L_P vom obține:

$$\begin{aligned} L_D(\alpha, \beta) &\stackrel{\text{def.}}{=} \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j + C \sum_{i=1}^m \xi_i - \\ &\quad - \sum_{i=1}^m \alpha_i \left(\left(\sum_{j=1}^m \alpha_j y_j x_j \right) \cdot x_i + w_0 \right) y_i - 1 + \xi_i - \sum_{i=1}^m \beta_i \xi_i \\ &= \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j + C \sum_{i=1}^m \xi_i - \\ &\quad - \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j - w_0 \underbrace{\sum_{i=1}^m \alpha_i y_i}_{0} + \sum_{i=1}^m \alpha_i - \sum_{i=1}^m \alpha_i \xi_i - \sum_{i=1}^m \beta_i \xi_i \\ &= -\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j + C \sum_{i=1}^m \xi_i + \sum_{i=1}^m \alpha_i - \sum_{i=1}^m \underbrace{(\alpha_i + \beta_i)}_C \xi_i \\ &= \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \stackrel{\text{not.}}{=} L_D(\alpha). \end{aligned}$$

Observați că la scrierea argumentelor lagrangeanului L_D , am renunțat în final la β , întrucât β_i (pentru $i = 1, \dots, m$) a fost eliminat în timpul efectuării calculului. Se constată de asemenea că acest lagrangean este identic cu cel care a fost obținut în cazul SVM cu margine „hard“ (vedeți relația (296) de la problema 9).

Tinând cont că $\beta_i = C - \alpha_i \geq 0$ implică $\alpha_i \leq C$, rezultă imediat că forma duală asociată problemei (P') este cea indicată în enunț.

Remarcați faptul că forma duală (D') pentru problema C-SVM (cu margine “soft”) diferă de forma duală (D) pentru problema SVM cu margine “hard” doar prin restricția suplimentară $\alpha_i \leq C$. Aceasta înseamnă că „importanța“ care revine fiecărei instanțe etichetate (x_i, y_i) cu privire la determinarea separatorului optimal devine limitată.

Facem aici mențiunea că legătura dintre soluțiile problemelor (P') și (D') este dată de relațiile care au fost obținute la punctele c și d : mai întâi

$$\bar{w} = \sum_{i=1}^m \bar{\alpha}_i y_i x_i, \quad (298)$$

iar apoi \bar{w}_0 se obține din relația $(\bar{w} \cdot x_i + \bar{w}_0)y_i = 1 - \bar{\xi}_i$ pentru un $i \in \{1, \dots, m\}$ astfel încât $\bar{\alpha}_i > 0$.⁶⁴⁴ Așadar,

$$\bar{w}_0 = -\bar{w} \cdot x_i + y_i(1 - \bar{\xi}_i), \text{ cu } \bar{\xi}_i = 0 \text{ dacă } \bar{\alpha}_i < C. \quad (299)$$

Această ultimă egalitate este implicată de relația $\alpha_i + \beta_i = C$ dedusă mai sus și de condiția de complementaritate Karush-Kuhn-Tucker $\beta_i \xi_i = 0$.⁶⁴⁵

e. Fie $\bar{\alpha}$ o soluție a problemei (D') , iar \bar{w} și \bar{w}_0 stabiliți ca mai sus (vedeți relațiile (298) și (299)). Vectorii-suport sunt instanțele x_i pentru care $\bar{\alpha}_i > 0$. Mai sus (vedeți rezolvarea de la punctul d , la final) am arătat că $\bar{\alpha}_i \in (0, C) \Rightarrow \bar{\xi}_i = 0$, deci $y_i(\bar{w} \cdot x_i + \bar{w}_0) = 1$. Alternativ, pentru $\bar{\alpha}_i = C$ rezultă $\bar{\beta}_i = C - \bar{\alpha}_i = 0$, deci $\bar{\xi}_i \geq 0$ și, ținând cont de egalitatea $y_i(\bar{w} \cdot x_i + \bar{w}_0) = 1 - \bar{\xi}_i$, rezultă că $y_i(\bar{w} \cdot x_i + \bar{w}_0) \leq 1$.⁶⁴⁶ Similar, pentru instanțele care nu sunt vectori-suport, $\bar{\alpha}_i = 0 \Rightarrow \bar{\beta}_i = C \Rightarrow \bar{\xi}_i = 0$, deci $y_i(\bar{w} \cdot x_i + \bar{w}_0) \geq 1$.⁶⁴⁷

Observație: Cele trei relații deduse mai sus, rescrise sintetizat ca

$$\begin{cases} \bar{\alpha}_i \in (0, C) \Rightarrow y_i(\bar{w} \cdot x_i + \bar{w}_0) = 1 \\ \bar{\alpha}_i = C \Rightarrow y_i(\bar{w} \cdot x_i + \bar{w}_0) \leq 1 \\ \bar{\alpha}_i = 0 \Rightarrow y_i(\bar{w} \cdot x_i + \bar{w}_0) \geq 1 \end{cases} \quad (300)$$

⁶⁴⁴Vedeți condițiile de complementaritate Karush-Kuhn-Tucker: $\bar{\alpha}_i[(\bar{w} \cdot x_i + \bar{w}_0)y_i - 1 + \bar{\xi}_i] = 0$ și $\bar{\beta}_i \bar{\xi}_i = 0$ pentru $i = 1, \dots, m$. Dacă $\bar{\alpha}_i > 0$, atunci $(\bar{w} \cdot x_i + \bar{w}_0)y_i - 1 + \bar{\xi}_i = 0$, deci $(\bar{w} \cdot x_i + \bar{w}_0)y_i = 1 - \bar{\xi}_i$. Dacă $\bar{\alpha}_i < C$, atunci $\bar{\beta}_i = C - \bar{\alpha}_i > 0$ și deci $\bar{\xi}_i = 0$.

⁶⁴⁵Dacă $\bar{\alpha}_i = 0$ pentru $i = \overline{1, m}$, obținem $\bar{w} = 0$, ceea ce reprezintă o soluție inadmisibilă.

Rămâne de tratat cazul în care $\exists \bar{\alpha}_i = C$ și pentru orice altă $\bar{\alpha}_j$ avem fie $\bar{\alpha}_j = 0$ fie $\bar{\alpha}_j = C$. Este de notat mai întâi faptul că din relația $\sum_{i=1}^m \bar{\alpha}_i y_i = 0$ va rezulta că jumătate dintre vectorii-suport sunt instanțe pozitive, iar restul (cealaltă jumătate) sunt instanțe negative. Apoi, $\bar{\alpha}_j = C > 0$ implică $y_j(\bar{w} \cdot x_j + w_0) = 1 - \bar{\xi}_j$ și $\bar{\beta}_j = 0$ și deci $\bar{\xi}_j \geq 0$. Pe de altă parte, $\bar{\alpha}_j = 0$ implică $\bar{\beta}_j = C$ și deci $\bar{\xi}_j = 0$, și $y_j(\bar{w} \cdot x_j + w_0) \geq 1$.

În consecință,

$$\begin{aligned} \sum_i \xi_i &= \sum_{i:\bar{\alpha}_i=C} \xi_i = \sum_{i:\bar{\alpha}_i=C} (1 - y_i(\bar{w} \cdot x_i + \bar{w}_0)) = \sum_{i:\bar{\alpha}_i=C} (1 - y_i \bar{w} \cdot x_i) \\ &= |\{i : \bar{\alpha}_i = C\}| - \sum_{i:\bar{\alpha}_i=C} y_i \bar{w} \cdot x_i = |\{i : \bar{\alpha}_i = C\}| - \bar{w} \cdot \sum_{i:\bar{\alpha}_i=C} y_i x_i \\ &= |\{i : \bar{\alpha}_i = C\}| - \frac{1}{C} \bar{w} \cdot \sum_{i:\bar{\alpha}_i=C} \bar{\alpha}_i y_i x_i = |\{i : \bar{\alpha}_i = C\}| - \frac{1}{C} \bar{w}^2 \end{aligned}$$

Aceasta ne arată că optimul problemei primale (P') depinde doar de \bar{w} (nu și de \bar{w}_0).

Notând în mod generic prin x_+ instanțele pozitive (și cu $\bar{\xi}_+$ valoarea variabilei de „destindere“ corespunzătoare), iar prin x_- instanțele negative (și, corespunzător, $\bar{\xi}_-$), vom avea de satisfăcut restricțiile $\bar{w} \cdot x_+ + \bar{w}_0 \geq 1 - \bar{\xi}_+$ și $\bar{w} \cdot x_- + \bar{w}_0 \leq -1 + \bar{\xi}_-$. Tinând cont că $\bar{\xi}_+ \geq 0$ și $\bar{\xi}_- \geq 0$, considerăm că din punct de vedere practic se poate lua pentru \bar{w}_0 valoarea $-\frac{1}{2}(\min_{x_+} \{\bar{w} \cdot x_+\} + \max_{x_-} \{\bar{w} \cdot x_-\})$. De remarcat că într-o astfel de situație, marginea va reprezenta distanța dintre separatorul optimal și cele mai depărtate instanțe pozitive / negative clasificate corect. Evident, aceasta este o situație (destul de) extremă în raport cu ideea cu care s-a plecat la drum în formalizarea clasificării cu margine maximală.

⁶⁴⁶Dacă $0 < \bar{\xi}_i \leq 1$, atunci instanța x_i este situată în interiorul „marginii“ de separare, iar dacă $\bar{\xi}_i > 1$, instanța x_i este clasificată eronat.

⁶⁴⁷Vectorii pentru care are loc egalitatea $y_i(\bar{w} \cdot x_i + \bar{w}_0) = 1$ se mai numesc *vectori-margine* (engl., margin vectors).

se întâlnesc uneori în literatura de specialitate sub numele de *condiții Karush-Kuhn-Tucker*, dar acesta este impropriu. Ele sunt *condiții necesare implicate de condițiile de optimizare și complementaritate Karush-Kuhn-Tucker*. Însă ceea ce este important de reținut este faptul că relațiile de mai sus sunt folosite ca o *condiție de oprire* pentru algoritmul SMO,⁶⁴⁸ care rezolvă în practică problema de optimizare cu margine “soft” (desemnată în continuare prin C-SVM).

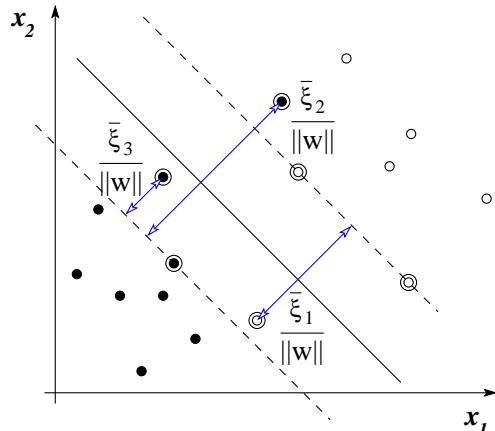
Rezumat: Pentru SVM cu margine “soft” (C-SVM), *vectorii-suport* sunt instanțele pentru care $\bar{\alpha}_i > 0$ (pentru acestea, avem $y_i(\bar{w} \cdot x_i + \bar{w}_0) \leq 1$), astădat inclusiv instanțele pentru care $y_i(\bar{w} \cdot x_i + \bar{w}_0) < 1$ (deci cu $\bar{\alpha}_i = C$).⁶⁴⁹

f. Păstrând notațiile de mai sus, instanța de test x' va fi clasificată pozitiv dacă $\bar{w} \cdot x' + \bar{w}_0 \geq 0$, și negativ în caz contrar.

13. (C-SVM: calculul distanței față de hiperplanul-margine corespunzător, pentru acei vectori-suport x_i pentru care $\bar{\alpha}_i = C$; calculul valorii optime pentru funcția obiectiv)

UAIC Iași, 2018 spring, Sebastian Ciobanu

a. Folosind notațiile pentru forma primală și forma duală de la problema de optimizare C-SVM, — vedeti problema 12 — demonstrați următoarea afirmație: pentru toate acele instanțe de antrenament x_i pentru care $\bar{\alpha}_i = C$ (adică, pentru acei x_i pentru care, conform rezolvării problemei 12.e, $y_i(\bar{w} \cdot x_i + \bar{w}_0) \leq 1$ și, deci $\bar{\xi}_i \geq 0$), *distanța geometrică* de la x_i la *hiperplanul-margine*, care este paralel cu hiperplanul de separare optimă și are ecuația $\bar{w} \cdot x + \bar{w}_0 = y_i$, este $\frac{\bar{\xi}_i}{\|\bar{w}\|}$.



Credit: cf. <http://efavdb.com/svm-classification/>.

Altfel spus,

$$\bar{\alpha}_i = C \Rightarrow \frac{\bar{\xi}_i}{\|\bar{w}\|} = d(x_i, \bar{w} \cdot x + \bar{w}_0 = y_i), \forall i \in \{1, \dots, m\}.$$

⁶⁴⁸Vedeți problemele 22 și 23.

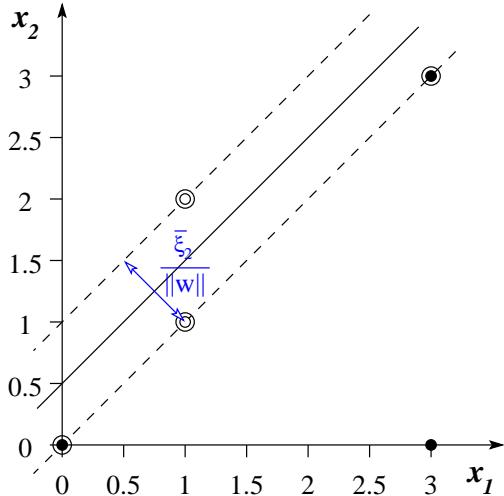
⁶⁴⁹Din punct de vedere *geometric*, vectorii-suport care nu sunt vectori-magine, adică aceia pentru care $y_i(\bar{w} \cdot x_i + \bar{w}_0) < 1$ sunt fie instanțe etichetate (x_i, y_i) incorrect clasificate de către separatorul optimal (adică, ele satisfac inegalitatea $y_i(\bar{w} \cdot x_i + \bar{w}_0) \leq 0$), fie instanțe situate între separatorul optimal și hiperplanul „margine“ de ecuație $\bar{w} \cdot x_i + \bar{w}_0 = y_i$.

b. Fie următorul set de date de antrenament:

i	1	2	3	4	5
x_i	(0, 0)	(1, 1)	(3, 3)	(1, 2)	(3, 0)
y_i	+1	-1	+1	-1	+1

Valorile multiplicatorilor Lagrange învățate de către o C-SVM liniară pentru $C = 10$, pornind de la aceste date sunt:

i	1	2	3	4	5
α_i	8.66	10	5.33	3.99	0



i. Verificați numeric relația $\frac{\bar{\xi}_2}{\|\bar{w}\|} = d(x_2, \bar{w} \cdot x + \bar{w}_0 = y_2)$.

ii. Calculați valoarea funcției obiectiv de la problema primală C-SVM pentru același $C = 10$.

Sugestii:

1. Folosiți condițiile [de complementaritate] Karush-Kuhn-Tucker pentru problema de optimizare C-SVM, precum și relațiile de legătură dintre soluțiile formei primale și soluțiile formei duale ale problemei C-SVM. (Vedeți rezolvarea problemei 12, subpunctele c și d.)
2. Pentru ușurință calculelor, se pot face aproximări. (De exemplu, 1.99 poate fi aproximat cu 2.)

Răspuns:

a. Întrucât $\bar{\alpha}_i = C > 0$, din condiția de complementaritate Karush-Kuhn-Tucker $\bar{\alpha}_i[y_i(\bar{w} \cdot x_i + \bar{w}_0) - (1 - \bar{\xi}_i)] = 0$ (vedeți rezolvarea problemei 12.c), rezultă

$$y_i(\bar{w} \cdot x_i + \bar{w}_0) = 1 - \bar{\xi}_i \Rightarrow \bar{w} \cdot x_i + \bar{w}_0 = y_i(1 - \bar{\xi}_i) \Rightarrow \bar{w} \cdot x_i + \bar{w}_0 - y_i = -y_i \bar{\xi}_i.$$

Prin urmare, ținând cont și de formula dată la problema 1, distanța de la punctul x_i la hiperplanul de ecuație $\bar{w} \cdot x_i + \bar{w}_0 - y_i = 0$ este

$$\frac{|-y_i \bar{\xi}_i|}{\|\bar{w}\|} \stackrel{\bar{\xi}_i \geq 0}{=} \frac{\bar{\xi}_i}{\|\bar{w}\|}.$$

b.i. Conform problemei 12.d, mai precis conform relației (298), $\bar{w} = \sum_{i=1}^5 \bar{\alpha}_i y_i x_i$. Deci, în cazul nostru,

$$\bar{w} = 8.66 \begin{pmatrix} 0 \\ 0 \end{pmatrix} - 10 \begin{pmatrix} 1 \\ 1 \end{pmatrix} + 5.33 \begin{pmatrix} 3 \\ 3 \end{pmatrix} - 3.99 \begin{pmatrix} 1 \\ 2 \end{pmatrix} + 0 = \begin{pmatrix} -10 + 15.99 - 3.99 \\ -10 + 15.99 - 7.98 \end{pmatrix} = \begin{pmatrix} 2 \\ -1.99 \end{pmatrix}.$$

Pentru ușurință calculelor vom aproxima:

$$\bar{w} \approx \begin{pmatrix} 2 \\ -2 \end{pmatrix}.$$

Aşadar,

$$\|\bar{w}\| = \sqrt{2^2 + (-2)^2} = \sqrt{4+4} = \sqrt{8} = 2\sqrt{2}.$$

Conform problemei 12.d, putem aproxima termenul liber \bar{w}_0 folosind ecuaţiile / proprietăţile corespunzătoare unui vector-suport x_i pentru care $\bar{\xi}_i = 0$:

$$\bar{\alpha}_1 = 8.66 \in (0, C) \Rightarrow \bar{\beta}_1 > 0 \xrightarrow{\text{KKT}} \bar{\xi}_1 = 0 \Rightarrow x_1 \text{ este vector-suport.}$$

$$y_1(\bar{w} \cdot x_1 + \bar{w}_0) = 1 - \bar{\xi}_1 \Rightarrow (+1) \left(\begin{pmatrix} 2 \\ -2 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \end{pmatrix} + \bar{w}_0 \right) = 1 - 0 \Rightarrow \bar{w}_0 = 1.$$

Acum putem calcula $\bar{\xi}_2$ folosind prima parte a raţionamentului de la punctul a:

$$\begin{aligned} \bar{\alpha}_2 = 10 = C \Rightarrow y_2(\bar{w} \cdot x_2 + \bar{w}_0) &= 1 - \bar{\xi}_2 \\ \Rightarrow \bar{\xi}_2 &= 1 - y_2(\bar{w} \cdot x_2 + \bar{w}_0) = 1 - (-1) \left(\begin{pmatrix} 2 \\ -2 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} + 1 \right) = 2. \end{aligned}$$

În consecinţă,

$$\begin{aligned} d(x_2, \bar{w} \cdot x + \bar{w}_0 = y_2) &= d\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ -2 \end{pmatrix} \cdot x + 1 = -1\right) \\ &= \frac{\left| \begin{pmatrix} 2 \\ -2 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} + 1 - (-1) \right|}{\|\bar{w}\|} = \frac{|2|}{2\sqrt{2}} = \frac{1}{\sqrt{2}}. \end{aligned}$$

Aşadar, se verifică egalitatea

$$\frac{\bar{\xi}_2}{\|\bar{w}\|} = d(x_2, \bar{w} \cdot x + \bar{w}_0 = y_2).$$

Observaţie: Se poate verifica şi din punct de vedere pur geometric, lucrând pe figura din enunţ, că distanţa dintre punctul x_2 şi hiperplanul $\bar{w} \cdot x + \bar{w}_0 = y_2$ este de (aproximativ⁶⁵⁰) $1/\sqrt{2}$.

b.ii. Trebuie să calculăm valoarea

$$\frac{1}{2} \|\bar{w}\|^2 + C \sum_{i=1}^5 \bar{\xi}_i.$$

De la subpunctul precedent ştim că $\|\bar{w}\| \approx 2\sqrt{2}$ şi $\bar{\xi}_2 \approx 2$. Pentru a determina celelalte valori ale variabilelor de „destindere“ $\bar{\xi}_i$, vom face din nou apel la *condiţiile Karush-Kuhn-Tucker*:

$$0 \leq \bar{\alpha}_1, \bar{\alpha}_3, \bar{\alpha}_4, \bar{\alpha}_5 < C \xrightarrow{\alpha_i + \beta_i = C} 0 < \bar{\beta}_1, \bar{\beta}_3, \bar{\beta}_4, \bar{\beta}_5 \leq C \xrightarrow{\beta_i \bar{\xi}_i = 0} \bar{\xi}_1, \bar{\xi}_3, \bar{\xi}_4, \bar{\xi}_5 = 0.$$

În consecinţă,

$$\frac{1}{2} \|\bar{w}\|^2 + C \sum_{i=1}^5 \bar{\xi}_i \approx \frac{1}{2} (2\sqrt{2})^2 + 10 \cdot (0 + 2 + 0 + 0 + 0) = \frac{1}{2} \cdot 8 + 10 \cdot 2 = 24.$$

⁶⁵⁰Din cauza „rotunjirii“ aplicate mai sus asupra lui \bar{w} .

14. (SVM și C-SVM: o proprietate simplă a soluției (\bar{w}) , indusă de o caracteristică particulară a datelor de antrenament)

prelucrare de Liviu Ciortuz, după

- ○ *MIT, 2008 fall, Tommi Jaakkola, midterm exam, pr. 1.1*
- CMU, 2010 fall, Aarti Singh, HW3, pr. 3.3.a*

Considerăm instanțele de antrenament $x_1 = (1, 1)$, $x_2 = (2, 2)$, $x_3 = (-1.5, -1.5)$ și $x_4 = (4, 4)$. Antrenăm o mașină cu vectori-suport pe aceste date.

Arătați că

- indiferent care este etichetarea celor patru instanțe de antrenament, și
- indiferent dacă mașina cu vectori-suport lucrează cu margine “hard” ori cu margine “soft”,

vectorul $\bar{w} = (\bar{w}_1, \bar{w}_2)$ care constituie soluția problemei de optimizare [C-]SVM pe aceste date are următoarea proprietate: $\bar{w}_1 = \bar{w}_2$.

Răspuns:

Stim că soluția problemei [C-]SVM în forma primală este $\bar{w} = \sum_{i=1}^m \bar{\alpha}_i y_i x_i$, unde $\bar{\alpha}_i$, cu $i = 1, \dots, m$, constituie soluția problemei în forma duală.⁶⁵¹ Întrucât toate instanțele de antrenament au proprietatea $x_{i1} = x_{i2}$, rezultă imediat că $\bar{w}_1 = \bar{w}_2$.

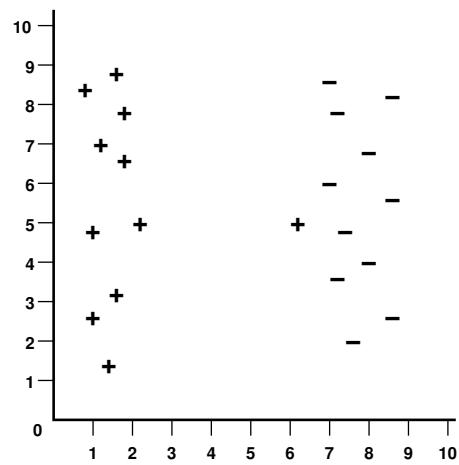
Observație: Exercițiul acesta a ilustrat de fapt următoarea proprietate importantă: dacă în setul de date de antrenament două trăsături (engl., features) sunt duplicate ($x_{ij} = x_{ik}$ pentru $i = 1, \dots, m$), atunci ele vor primi ponderi identice ($\bar{w}_j = \bar{w}_k$) în soluția optimală calculată de clasificatorul [C-]SVM.

15. (C-SVM, cazul [datelor separabile] liniar: aplicare în prezența unui “outlier”)

CMU, 2005 spring, C. Guestrin, T. Mitchell, HW3, pr. 2.2

Se dau datele prezentate în figura alăturată. Lucrând cu clasificatorul C-SVM, parametrul C — pentru penalizarea aferentă „destinderii“; engl., slack penalty — va determina poziția hiperplanului de separare. Presupunem că se lucrează în varianta liniară, adică fără funcții-nucleu. Răspundeți succint, în manieră *calitativă*, la întrebările de mai jos.

- Unde va fi situat separatorul optimal în cazul în care C ia valori foarte mari (adică, $C \rightarrow \infty$)? Indicați pe figura dată.
- Pentru $C \approx 0$, indicați pe figura dată unde va fi situat separatorul optimal.



⁶⁵¹Vedeți relațiile (293) și (298) deduse la problemele 9 și respectiv 12.

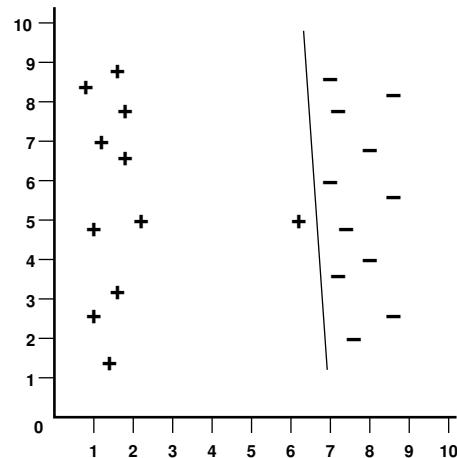
c. Care din cele două situații de mai sus credeți că este mai adekvată pentru clasificare? De ce?

Răspuns:

Putem face mai întâi două *observații* în legătură cu acest set de date de antrenament:

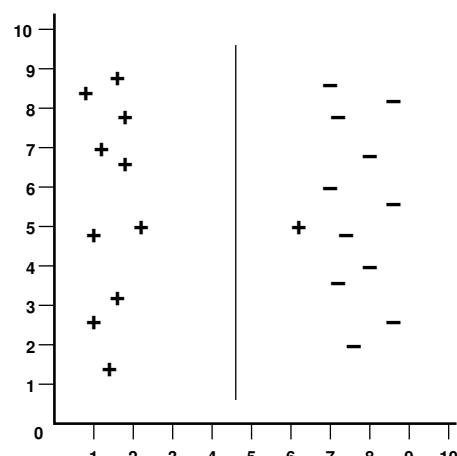
- grupul de date clasificate negativ, precum și grupul de date clasificate pozitiv (din partea stângă) sunt foarte compacte;
- există un singur exemplu pozitiv care este foarte apropiat de grupul exemplelor negative. Acest exemplu-excepție constituie o „anomalie“ (engl., outlier).

a. O valoare mare a parametrului C semnifică o penalizare mai mare a erorilor de clasificare ξ_i . Așadar, pentru valori ale lui C tînzând la $+\infty$ hiperplanul de separare optimală va corespunde celei mai mici valori posibile pentru suma $\sum_i \xi_i$, care reprezintă eroarea totală în raport cu marginile. Întrucât exemplele date sunt separabile liniar, separatorul optimal învățat este cel reprezentat în figura alăturată. Poziția sa este foarte mult influențată de poziția outlierului (+), care este situat în imediata proximitate a instanțelor negative.



Observație: În general, dacă avem încredere în corectitudinea datelor de antrenament, putem alege $C \rightarrow \infty$ pentru a obține o separare cât mai precisă a celor două clase.

b. În cazul $C \approx 0$, erorile ξ_i sunt penalizate foarte puțin. În consecință, separatorul optimal trebuie să maximizeze marginea — adică distanța de la hiperplanul de separare $w \cdot x + w_0 = 0$ la vectorii suport x_i pentru care $y_i(w \cdot x_i + w_0) = 1$ —, chiar dacă asta înseamnă că unele instanțe vor fi clasificate eronat. Separatorul determinat în acest caz va fi poziționat ca în figura alăturată.



Observație: În general, este indicat să alegem o valoare mică pentru parametrul C atunci când datele de antrenament pe care le avem sunt afectate de „zgomote“ / perturbații / anomalii. Într-o astfel de situație am putea avea

de-a face cu un set de date neseparabile liniar, în care câteva exemple pozitive să fie amestecate printre cele negative și invers.

c. Separatorul determinat la punctul b este alegerea cea mai bună pentru generalizare / testare în prezența outlier-elor.

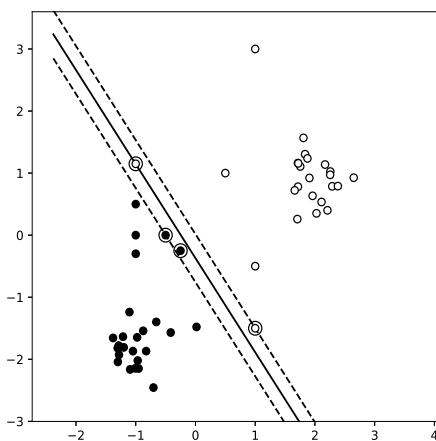
16.

(C-SVM, cazul [datelor separabile] liniar: efectul alegerii diverselor valori pentru parametrul C)
CMU, 2010 fall, Ziv Bar-Joseph, midterm exam, pr. 7.c

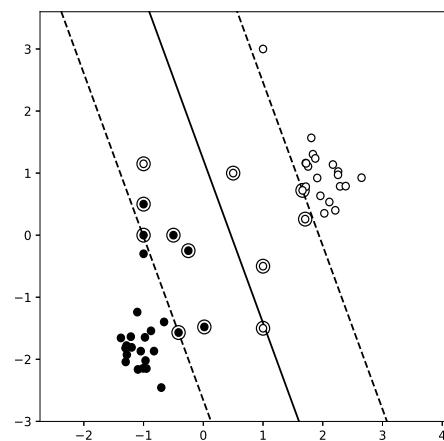
În următoarele patru figuri, se consideră că s-a aplicat o C-SVM liniară — adică o mașină cu vectori-suport cu margine “soft” și cu parametru de „destindere” C , fără funcție-nucleu — pe un set de date din \mathbb{R}^2 , folosind de fiecare dată una din următoarele valori pentru parametrul C : 0.1, 1, 10 sau 100.

Sub fiecare figură scrieți valoarea corespunzătoare pentru parametrul C .

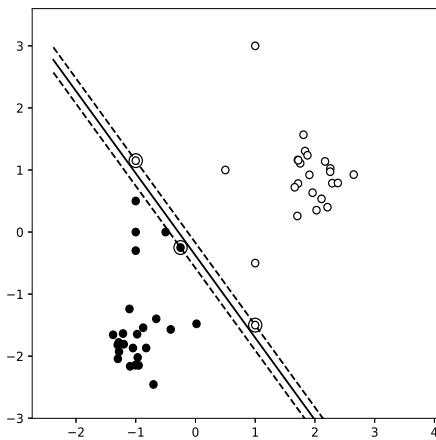
Justificați alegerea pe care ati făcut-o.



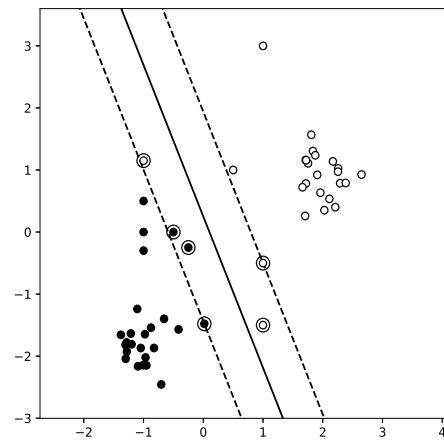
A.



B.



C.



D.

Răspuns:

Considerând datele de antrenament (x_i, y_i) cu $i = \overline{1, m}$, funcția obiectiv pentru forma primală a problemei C-SVM este

$$\min_{w \in \mathbb{R}^d, \xi} \left(\frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i \right),$$

unde $\xi_i \geq 0$ sunt variabile de „destindere“ ($y_i(w \cdot x_i + w_0) \geq 1 - \xi_i$ pentru $i = 1, \dots, m$). Din expresia funcției obiectiv decurge imediat că la valori mari ale parametrului C — vă reamintim, $C > 0$ — sunt permise eventuale erori ξ_i mici, și invers: la valori mici ale parametrului C sunt permise erori ξ_i mari. De asemenea, ne așteptăm ca, pe măsură ce valoarea parametrului C crește, marginea⁶⁵² să scadă.

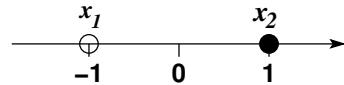
Dintre cele patru grafice din enunț, în cazul celui de-al treilea (C) avem eroare la antrenare 0; în toate celelalte trei cazuri avem în mod evident $\sum_i \xi_i > 0$.⁶⁵³ Așadar, graficul C corespunde celei mai mari valori dintre cele patru valori propuse pentru parametrul C : 100.

În cazul graficului B sunt mult mai mulți vectori-suport decât în cazul graficelor A sau D, iar suma $\sum_i \xi_i$ este cea mai mare dintre toate aceste trei cazuri rămase în discuție (A, B și D). Așadar, graficului B îi corespunde cea mai mică dintre valorile lui C rămase: 0.1. În fine, marginea este mai mare în cazul D decât în cazul A. Prin urmare, lui A îi corespunde $C = 10$, iar lui D valoarea rămasă: $C = 1$.

17. (C-SVM: un set simplu de date pentru care forma duală a problemei de optimizare C-SVM are soluție unică, dar forma primală nu are soluție unică)

□ S. Ciobanu, L. Ciortuz, 2019, după “Uniqueness of the SVM solution”, C. Burges, D. Crisp, 1998

Fie instanțele de antrenament $(x_1 = -1, y_1 = -1)$ și $(x_2 = 1, y_2 = 1)$. Vrem să antrenăm o mașină cu vectori-suport cu margine “soft” (deci, C-SVM) pe acest set de date. Să se arate că



- a. în cazul $C \geq \frac{1}{2}$, soluția problemei de optimizare C-SVM este [unică, și anume]: $\bar{w} = 1$, $\bar{\xi}_1 = \bar{\xi}_2 = \bar{w}_0 = 0$.
- b. în cazul $C < \frac{1}{2}$, rezultă că $\bar{w} = 2C$, $\bar{\xi}_1 = 1 - 2C + \bar{w}_0$, $\bar{\xi}_2 = 1 - 2C - \bar{w}_0$ și orice(!) $\bar{w}_0 \in [2C - 1, 1 - 2C]$ este soluție a formei primale pentru problema de optimizare C-SVM. Așadar, în acest caz soluția problemei de optimizare C-SVM (în forma primală) nu este unică!

Observație: Pentru ambele cazuri, soluția problemei de optimizare C-SVM în forma duală este unică.

⁶⁵²Definită ca $1/\|w\|$, adică distanța de la hiperplanul de separare optimală la instanțele (vectorii-suport) x_i pentru care $(w \cdot x_i + w_0)y_i = 1$.

⁶⁵³Inegalitatea $\sum_i \xi_i > 0$ se datorează faptului că avem (cel puțin!) câte o excepție la clasificare în raport cu cele două margini (hiperplanele de ecuație $w \cdot x_i + w_0)y_i = \pm 1$), în fiecare dintre aceste ultime trei cazuri.

Răspuns:

Conform restricției $\sum_i y_i \bar{\alpha}_i = 0$ din problema duală C-SVM (vedeți ex. 12.d), rezultă $\bar{\alpha}_1 = \bar{\alpha}_2$ și, în consecință, $\bar{w} = \sum_i y_i \alpha_i x_i = 2\bar{\alpha}_1$ (conform relației (298)).

Datorită egalității $\alpha_1 = \alpha_2$, funcția obiectiv din problema duală, $L_D(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i \cdot x_j$, devine:

$$\begin{aligned} L_D(\alpha_1, \alpha_2) &= \alpha_1 + \alpha_2 - \frac{1}{2} [\alpha_1^2 (-1)^2 (-1)^2 + \alpha_2^2 \cdot 1^2 \cdot 1^2 + 2\alpha_1 \alpha_2 (-1) \cdot 1 \cdot (-1) \cdot 1] \\ &= 2\alpha_1 - \frac{1}{2} (2\alpha_1^2 + 2\alpha_1^2) = 2\alpha_1 - 2\alpha_1^2 = 2\alpha_1(1 - \alpha_1). \end{aligned}$$

Evident, maximul acestei funcții polinomiale de gradul al doilea se atinge pentru $\alpha_1 = 1/2$.

Soluțiile optime pentru problema duală C-SVM sunt supuse restricțiilor $0 \leq \bar{\alpha}_i \leq C$.⁶⁵⁴ Așadar, vom avea două cazuri: $\bar{\alpha}_1 = \bar{\alpha}_2 = 1/2$ atunci când $C \geq 1/2$ și, respectiv, $\bar{\alpha}_1 = \bar{\alpha}_2 = C$ atunci când $C < 1/2$.

Cazul 1 ($C \geq 1/2$): $\bar{\alpha}_1 = \bar{\alpha}_2 = 1/2$, $\bar{w} = 2\bar{\alpha}_1 = 1$.

Va trebui să mai calculăm $\bar{\xi}_1$, $\bar{\xi}_2$ și \bar{w}_0 . Analizând aşa-numitele *condiții de tip Karush-Kuhn-Tucker* (300),⁶⁵⁵ constatăm că pentru cazul acesta ($C \geq 1/2$) este convenabil să considerăm două subcazuri: $C > 1/2$ și, respectiv, $C = 1/2$, întrucât calculul valorilor $\bar{\xi}_i$ se va face diferit pentru cele două subcazuri.

Cazul 1.1 ($C > 1/2$): $\bar{\xi}_1 = \bar{\xi}_2 = 0$, fiindcă i. valorile variabilelor lagrangeene $\bar{\beta}_1 = C - \bar{\alpha}_1$ și $\bar{\beta}_2 = C - \bar{\alpha}_2$ sunt strict pozitive (vedeți relația $\alpha_i + \beta_i = C$ din ex. 12.d) și ii. avem condiția de complementaritate Karush-Kuhn-Tucker, $\bar{\beta}_i \bar{\xi}_i = 0$ pentru $i \in \{1, 2\}$. Putem calcula acum \bar{w}_0 :

$$y_1(\bar{w} \cdot x_1 + \bar{w}_0) = 1 \Rightarrow -(-1 + \bar{w}_0) = 1 \Rightarrow \bar{w}_0 = 0.$$

Cazul 1.2 ($C = 1/2$): Calculăm $\bar{\xi}_i$ folosind celelalte relații de complementaritate Karush-Kuhn-Tucker:

$$\begin{cases} \bar{\alpha}_1[y_1(\bar{w} \cdot x_1 + \bar{w}_0) - (1 - \bar{\xi}_1)] = 0 & \bar{\alpha}_i \neq 0 \\ \bar{\alpha}_2[y_2(\bar{w} \cdot x_2 + \bar{w}_0) - (1 - \bar{\xi}_2)] = 0 & \end{cases} \Rightarrow \begin{cases} y_1(\bar{w} \cdot x_1 + \bar{w}_0) = 1 - \bar{\xi}_1 \\ y_2(\bar{w} \cdot x_2 + \bar{w}_0) = 1 - \bar{\xi}_2 \\ -1(-1 + \bar{w}_0) = 1 - \bar{\xi}_1 \\ 1 + \bar{w}_0 = 1 - \bar{\xi}_2 \end{cases} \Rightarrow \bar{w}_0 = \bar{\xi}_1 = -\bar{\xi}_2.$$

Tinând cont de faptul că $\bar{\xi}_1 \geq 0$ și $\bar{\xi}_2 \geq 0$, rezultă $\bar{w}_0 = \bar{\xi}_1 = \bar{\xi}_2 = 0$.

Așadar, în ambele subcazuri a rezultat aceeași soluție ($\bar{w} = 1$, $\bar{w}_0 = 0$), iar separatorul optimal are ecuația $\bar{w} \cdot x + \bar{w}_0 = 0$, deci $x = 0$.

Cazul 2 ($C < 1/2$): $\bar{\alpha}_1 = \bar{\alpha}_2 = C$, $\bar{w} = 2\bar{\alpha}_1 = 2C$.

Aplicând (ca și mai sus) relațiile de complementaritate Karush-Kuhn-Tucker, vom obține:

⁶⁵⁴Cf., din nou, ex. 12.d.

⁶⁵⁵Le reprodusem aici:

$$\begin{cases} \bar{\alpha}_i \in (0, C) \Rightarrow y_i(\bar{w} \cdot x_i + \bar{w}_0) = 1 \\ \bar{\alpha}_i = C \Rightarrow y_i(\bar{w} \cdot x_i + \bar{w}_0) \leq 1 \\ \bar{\alpha}_i = 0 \Rightarrow y_i(\bar{w} \cdot x_i + \bar{w}_0) \geq 1 \end{cases}$$

$$\begin{cases} \bar{\alpha}_1[y_1(\bar{w} \cdot x_1 + \bar{w}_0) - (1 - \bar{\xi}_1)] = 0 \\ \bar{\alpha}_2[y_2(\bar{w} \cdot x_2 + \bar{w}_0) - (1 - \bar{\xi}_2)] = 0 \end{cases} \xrightarrow{\bar{\alpha}_i \neq 0} \begin{cases} y_1(\bar{w} \cdot x_1 + \bar{w}_0) = 1 - \bar{\xi}_1 \\ y_2(\bar{w} \cdot x_2 + \bar{w}_0) = 1 - \bar{\xi}_2 \end{cases}$$

$$\Rightarrow \begin{cases} -(-2C + \bar{w}_0) = 1 - \bar{\xi}_1 \\ 2C + \bar{w}_0 = 1 - \bar{\xi}_2 \end{cases} \Rightarrow \begin{cases} \bar{\xi}_1 = 1 - 2C + \bar{w}_0 \\ \bar{\xi}_2 = 1 - 2C - \bar{w}_0 \end{cases}$$

Ultimul sistem obținut mai sus este un sistem liniar de două ecuații cu trei necunoscute. Soluțiile $\bar{\xi}_1$ și $\bar{\xi}_2$ se scriu în funcție de \bar{w}_0 . În final, ținând cont de restricțiile $\bar{\xi}_1 \geq 0$ și $\bar{\xi}_2 \geq 0$, rezultă $\bar{w}_0 \geq 2C - 1$ și $\bar{w}_0 \leq 1 - 2C$, deci în concluzie $\bar{w}_0 \in [2C - 1, 1 - 2C]$.⁶⁵⁶

Prin urmare, în acest caz, forma primală a problemei C-SVM are o infinitate de soluții — deși forma duală are o singură soluție —, și anume câte una pentru fiecare valoare posibilă a lui \bar{w}_0 în intervalul $[2C - 1, 1 - 2C]$. Ecuația separatorului optimal este $\bar{w} \cdot x + \bar{w}_0 = 0 \Rightarrow 2Cx + \bar{w}_0 = 0 \Rightarrow x = -\bar{w}_0/2C$.

Observație: Este imediat că

$\bar{w}_0 \in [2C - 1, 1 - 2C] \Rightarrow -\bar{w}_0 \in [2C - 1, 1 - 2C] \Rightarrow -\bar{w}_0/2C \in \left[-\frac{1}{2C} + 1, \frac{1}{2C} - 1\right]$, iar $0 < C < 1/2 \Rightarrow 0 < 2C < 1 \Rightarrow \frac{1}{2C} > 1 \Rightarrow -\frac{1}{2C} < -1 \Rightarrow -\frac{1}{2C} + 1 < 0$, iar $\frac{1}{2C} - 1 > 0$. Remarcăți faptul că separatorul optimal $(-\bar{w}_0/2C)$ se poate afla oriunde pe axa reală dacă parametrul C este lăsat să varieze în intervalul $(0, 1/2)$!

18.

(C-SVM: o margine superioară pentru numărul de erori la antrenare)

□ • CMU, 2017 fall, Nina Balcan, HW4, pr. 4.Q12

Fie $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$ un set de m instanțe etichetate din \mathbb{R}^d , cu etichete din mulțimea $\{1, -1\}$. Vă readucem aminte că forma primală a problemei C-SVM este

$$\min_{w, w_0, \xi} \left(\|w\|^2 + C \sum_{i=1}^n \xi_i \right),$$

a. i. $\forall i, y_i(w \cdot x_i + w_0) \geq 1 - \xi_i$, cu $\xi_i \geq 0$.

Fie $\bar{w}, \bar{w}_0, \bar{\xi}$ soluția acestei probleme. Stabiliți o margine superioară (engl., upper bound) pentru numărul de instanțe din S care sunt clasificate în mod eronat la antrenare, folosind $\bar{\xi}_i$ pentru $i = 1, \dots, m$.

Răspuns:

Ținând cont de semnificația [analitică a] variabilelor de destindere ξ_i , rezultă că o instanță x_i este clasificată eronat la antrenare dacă $\bar{\xi}_i > 1$. Așadar, numărul total de erori comise de C-SVM la antrenare poate fi scris ca $\sum_i 1_{\{\bar{\xi}_i > 1\}}$, unde am folosit variabila-indicator

$$1_{\{\bar{\xi}_i > 1\}} = \begin{cases} 1 & \text{dacă } \bar{\xi}_i > 1, \\ 0 & \text{în caz contrar.} \end{cases}$$

⁶⁵⁶Observați că $2C - 1 < 0$, iar $1 - 2C > 0$.

Prin urmare,

$$\sum_i 1_{\{\bar{\xi}_i > 1\}} < \sum_{i: \bar{\xi}_i > 1} \bar{\xi}_i \leq \sum_i \bar{\xi}_i.$$

Dacă definim eroarea comisă de C-SVM la antrenare ca raportul dintre numărul de erori și m , numărul total de instanțe de antrenament, atunci marginea superioară cerută este $\frac{1}{m} \sum_i \bar{\xi}_i$.

19.

(C-SVM: o margine superioară pentru eroarea de tip CVLOO)

*prelucrare de Liviu Ciortuz, după
• o CMU, 2010 fall, Aarti Singh, midterm exam, pr. 5.2*

Considerăm un C-SVM, adică o mașină cu vectori-suport cu margine “soft”, care folosește parametrul de „destindere” C . Arătați că — pentru orice valoare a lui C fixată (în mod arbitrar) — eroarea produsă de acest clasificator la cross-validation cu metoda “Leave-One-Out” este mai mică sau egală cu

$$\frac{\#\text{SVs}}{m},$$

unde m este numărul exemplelor de antrenament, iar $\#\text{SVs}$ este numărul de vectori-suport obținuți (pentru respectiva valoare a lui C) la antrenarea acestui clasificator pe întreg setul de exemple.

Observații:

1. Vom demonstra în mod riguros (i.e., analitic) următoarea *proprietate importantă*: La CVLOO cu C-SVM, numai vectorii-suport pot fi (eventual!) clasificați eronat. Altfel spus, orice instanță x_k care nu este vector-suport nu produce eroare la CVLOO.
2. Vă readucem aminte că în contextul clasificatorului C-SVM, vectorii-suport sunt acei x_i pentru care $\bar{\alpha}_i > 0$. Așadar, ei sunt fie acei x_i pentru care $y_i(\bar{\alpha}_i \cdot x_i + \bar{w}_0) = 1$ (pentru aceștia, $\bar{\alpha}_i \in (0, C)$), fie acei x_i pentru care $y_i(\bar{\alpha}_i \cdot x_i + \bar{w}_0) < 1$ (pentru aceștia, $\bar{\alpha}_i = C$; ei constituie erori la antrenare dacă $\xi_i > 1$). Vedeti *Rezumatul* de la rezolvarea problemei 12.e, pag. 654.
3. Se poate demonstra că rezultatul din enunț este valabil și pentru SVM (adică mașina cu vectori-suport cu margine “hard”).
4. Eroarea de tip CVLOO este un bun indicator pentru capacitatea de generalizare a unui clasificator automat. Numărul vectorilor-suport fiind în general relativ mic în raport cu numărul total de instanțe de antrenament, marginea superioară indicată mai sus pentru eroarea CVLOO produsă de C-SVM „atestă“ într-un anumit sens calitatea acestui clasificator.

Răspuns:

Folosind notațiile de la problema 12, vom considera problemele de optimizare

$$\min_{w, w_0, \xi} \left(\frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i \right) \quad (\mathbf{P}')$$

a. i. $(w \cdot x_i + w_0)y_i \geq 1 - \xi_i$, pentru $i = 1, \dots, m$
 $\xi_i \geq 0$, pentru $i = 1, \dots, m$

și

$$\begin{aligned} \min_{w, w_0, \xi} & \left(\frac{1}{2} \|w\|^2 + C \sum_{i=1, i \neq k}^m \xi_i \right) \\ \text{a. i. } & (w \cdot x_i + w_0) y_i \geq 1 - \xi_i, \text{ pentru } i = 1, \dots, m, i \neq k \\ & \xi_i \geq 0, \text{ pentru } i = 1, \dots, m, i \neq k. \end{aligned} \quad (\mathbf{P}'_k)$$

unde $k \in \{1, \dots, m\}$.

Stim (tot de la problema 12) că problemele (\mathbf{P}') și (\mathbf{P}'_k) sunt în relație de *dualitate tare* cu cu dualele (\mathbf{D}') și respectiv (\mathbf{D}'_k) .⁶⁵⁷

$$\begin{aligned} \max_{\alpha} & \left(\sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \right) \\ \text{a. i. } & 0 \leq \alpha_i \leq C \text{ pentru } i = 1, \dots, m \end{aligned} \quad (\mathbf{D}')$$

$$\sum_{i=1}^m \alpha_i y_i = 0$$

și

$$\begin{aligned} \max_{\alpha} & \left(\sum_{i=1, i \neq k}^m \alpha_i - \frac{1}{2} \sum_{i,j; i,j \neq k} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \right) \\ \text{a. i. } & 0 \leq \alpha_i \leq C \text{ pentru } i = 1, \dots, m, i \neq k \\ & \sum_{i=1, i \neq k}^m \alpha_i y_i = 0. \end{aligned} \quad (\mathbf{D}'_k)$$

În plus, relația dintre soluțiile problemelor (\mathbf{P}') și (\mathbf{D}') este

$$\bar{w} = \sum_{i=1}^m \bar{\alpha}_i y_i x_i,$$

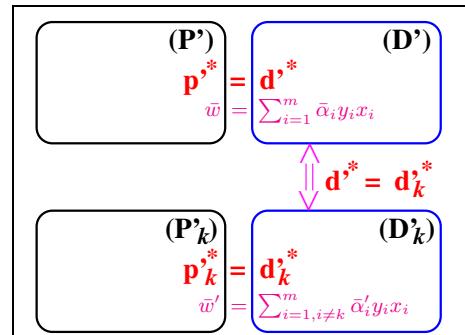
iar \bar{w}_0 se obține din relația $(\bar{w} \cdot x_i + \bar{w}_0) y_i = 1 - \bar{\xi}_i$ pentru un $i \in \{1, \dots, m\}$ cu proprietatea $\bar{\alpha}_i > 0$. Relația dintre soluțiile problemelor (\mathbf{P}'_k) și (\mathbf{D}'_k) se exprimă în mod similar.

Dacă în urma rezolvării problemei (\mathbf{P}') , instanța x_k nu este vector-suport, adică $\bar{\alpha}_k = 0$, atunci urmează că $\bar{\beta}_k = C - \bar{\alpha}_k = C \neq 0$, deci din condiția de complementaritate Karush-Kuhn-Tucker $\bar{\beta}_k \bar{\xi}_k = 0$ rezultă $\bar{\xi}_k = 0$ și, prin urmare

$$y_k(\bar{w} \cdot x_k + \bar{w}_0) \geq 1 - \bar{\xi}_k = 1. \quad (301)$$

Lucrând încălzită cu presupozitia că x_k nu este vector-suport, se poate arăta — vedeți mai jos — că soluțiile optime (\bar{w}, \bar{w}_0) pentru problemele (\mathbf{P}') și (\mathbf{P}'_k) sunt aceleași. În consecință, folosind relația (301), va rezulta că x_k este corect clasificat la CVLOO.

Demonstrația faptului că soluțiile optime ale problemelor (\mathbf{P}') și (\mathbf{P}'_k) sunt aceleași se poate face astfel:



- se ține cont de relațiile de dualitate tare menționate mai sus, desemnate prin notațiile $(\mathbf{P}') \equiv (\mathbf{D}')$ și $(\mathbf{P}'_k) \equiv (\mathbf{D}'_k)$. Referindu-ne la valorile optime ale funcțiilor-obiectiv respective, putem scrie — folosind notații inspirate de proprietatea de *dualitate tare* — $p^* = d'^*$ și $p'^*_k = d'_k^*$;

⁶⁵⁷Pentru definiția dualității tari, vedeți *Comentariul* de la problema 9 (pag. 642).

- se arată că soluția optimă pentru problema (D') coincide cu soluția optimă pentru problema (D'_k) , cu singura „diferență“ că $\bar{\alpha}_k = 0$. (Așadar, $d'^* = d_k'^*$.) Pe scurt, putem justifica aceasta în modul următor:

Valoarea optimă pentru funcția obiectiv a problemei (D') este mai mare sau cel puțin egală cu valoarea optimă pentru funcția obiectiv a problemei (D'_k) , pentru că spațiul de *valori admisibile* pentru variabilele α_i (i.e., valorile care satisfac restricțiile) este mai amplu. Mai departe, știind că $\bar{\alpha}_k = 0$, rezultă că valorile optime ale celor două probleme coincid, iar soluția optimă $\bar{\alpha}$ pentru problema (D') este (abstracție făcând de $\bar{\alpha}_k$) și soluție optimă a problemei (D'_k) .⁶⁵⁸

- se utilizează în final relațiile de legătură de forma $\bar{w} = \sum_{i=1}^m \bar{\alpha}_i y_i x_i$ (cu \bar{w}_0 determinat corespunzător).

Observație: Demonstrația proprietății din enunț pentru SVM [cu margine “hard”] se poate face fie urmând „tipicul“ demonstrației de mai sus — adică, făcând apel la legătura dintre forma primală și forma duală a problemei de optimizare SVM —, fie (considerabil mai simplu!) folosind o formă ușor generalizată a proprietății pe care am demonstrat-o la problema 2, care face legătura dintre problema de optimizare SVM și *interpretarea ei geometrică*.

⁶⁵⁸În manieră riguroasă, considerând (în principal de dragul simplificării exprimării) problema

$$\begin{aligned} \max_{\alpha} & \left(\sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \right) \\ \text{a. i. } & 0 \leq \alpha_i \leq C \text{ pentru } i = 1, \dots, m \text{ cu } i \neq k \text{ și } \alpha_k = 0 \\ & \sum_{i=1}^m \alpha_i y_i = 0 \end{aligned} \tag{D''}$$

veți observa mai întâi că problemele (D') și (D'') sunt *echivalente* — adică orice soluție optimă a lui (D') este și soluție optimă a lui (D'') — și apoi că problemele (D'') și (D'_k) , deși diferite din punct de vedere „sintactic“, sunt practic identice, întrucât pe de o parte α_k (componenta k a vectorului generic α) este 0 în problema (D'') și lipsește în problema (D'_k) , iar pe de altă parte substituind $\alpha_k = 0$ în funcția obiectivă a problemei (D'') obținem funcția obiectivă a problemei (D'_k) .

Pentru a demonstra că problemele (D') și (D'') sunt echivalente, observați mai întâi că orice soluție admisibilă a problemei (D'') este și soluție admisibilă pentru problema (D') . Așadar, mulțimea de soluții admisibile pentru problema (D') este mai amplă decât mulțimea de soluții admisibile pentru problema (D'') . Notăm cu F funcția obiectivă a problemei (D') , adică $F(\alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j$. Fie $\bar{\alpha}$ soluție optimă a problemei (D') . Componența sa de pe poziția k este nulă (adică, $\bar{\alpha}_k = 0$), fiindcă x_k nu este vector-suport. Urmează că $F(\bar{\alpha}) \geq F(\alpha)$ pentru orice soluție admisibilă α a problemei (D') , deci

$$F(\bar{\alpha}) = \max\{F(\alpha) \mid \alpha \text{ este soluție admisibilă a problemei } (D')\}.$$

Întrucât mulțimea de soluții admisibile pentru problema (D') este mai amplă decât mulțimea de soluții admisibile pentru problema (D'') , rezultă că

$$F(\bar{\alpha}) \geq \max\{F(\alpha) \mid \alpha \text{ este soluție admisibilă a problemei } (D'')\}.$$

Însă ținând cont pe de o parte că $\bar{\alpha}$ este și soluție admisibilă pentru problema (D'') , iar pe de altă parte că $\bar{\alpha}_k = 0$, rezultă că inegalitatea precedentă devine

$$F(\bar{\alpha}) = \max\{F(\alpha) \mid \alpha \text{ este soluție admisibilă a problemei } (D'')\}.$$

Așadar, $\bar{\alpha}$ este soluție optimă și pentru problema (D'') .

20. (Problema de optimizare C-SVM: o formulare echivalentă, dar fără restricții, folosind în schimb funcția de cost / pierdere *hinge*)

*prelucrare de Liviu Ciortuz, după
■ • CMU, 2008 fall, Eric Xing, HW2, pr. 1.2
CMU, 2017 fall, Nina Balcan, HW4, pr. 4.1*

Considerăm m instanțe de antrenament $\{x_i, y_i\}_{i=1}^m$. Vă readucem aminte că problema SVM cu margine “soft” și parametru de „destindere” $C > 0$ poate fi formulată ca o problemă de optimizare (pătratică) cu restricții:

$$\begin{aligned} \min_{w, w_0, \xi} & \left(\frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i \right) \\ \text{a. i. } & (w \cdot x_i + w_0) y_i \geq 1 - \xi_i, \text{ pentru } i = 1, \dots, m \\ & \xi_i \geq 0, \text{ pentru } i = 1, \dots, m \end{aligned} \tag{P'}$$

- a. Demonstrați că formularea de mai sus este *echivalentă* cu o problemă de optimizare (tot pătratică) *fără* restricții, de forma:

$$\min_{w, w_0} \left(\|w\|^2 + \lambda \sum_{i=1}^m \max(1 - y_i(w \cdot x_i + w_0), 0) \right), \tag{L}$$

unde λ este un parametru real pozitiv fixat.

- b. Exprimăți valoarea noului parametru λ în funcție de parametrul de „destindere” C .

- c. Cum apreciați din punct de vedere „calitativ“ această nouă formulare a problemei de optimizare C-SVM?

Răspuns:

- a. Din restricțiile din forma primală a problemei C-SVM, avem

$$\xi_i \geq 1 - y_i(w \cdot x_i + w_0) \text{ și } \xi_i \geq 0,$$

ceea ce echivalează cu $\xi_i \geq \max(1 - y_i(w \cdot x_i + w_0), 0)$ pentru $i = 1, \dots, m$.

Operatorul \min din formularea *obiectivului* problemei de optimizare C-SVM (P') implică faptul că $\bar{\xi}_i$ din soluția optimă a acestei probleme $(\bar{w}, \bar{w}_0, \bar{\xi})$ va fi setată chiar la valoarea $\max(1 - y_i(\bar{w} \cdot x_i + \bar{w}_0), 0)$.

Prin urmare, este naturală transformarea funcției obiectiv a problemei (P') în funcția obiectiv a problemei (L). Ceea ce nu apare a fi la fel de ușor de explicat este renunțarea la restricțiile formulate în cadrul problemei (P') atunci când se face trecerea la problema (L). De aceea, în cele ce urmează este necesar să arătam că aceste două probleme sunt *echivalente*, dovedind că orice soluție optimă a problemei (P') corespunde unei soluții optime a problemei (L) și invers, orice soluție optimă a problemei (L) corespunde unei soluții optime a problemei (P').

Considerăm deci mai întâi $\bar{w}, \bar{w}_0, \bar{\xi}$ o soluție optimă a problemei C-SVM (P'). Luând $\lambda = 2C$, vom arăta că \bar{w}, \bar{w}_0 este și soluție optimă a noii probleme de optimizare (L). Presupunem prin *prin reducere la absurd* că \bar{w}, \bar{w}_0 nu este soluție optimă a noii probleme de optimizare (L). Rezultă că există w^*, w_0^* o soluție optimă a problemei (L), care este mai bună decât soluția \bar{w}, \bar{w}_0 , adică

$$L(w^*, w_0^*) < L(\bar{w}, \bar{w}_0),$$

unde prin L am notat funcția obiectiv a problemei de optimizare (L). Notând $\xi_i^* \stackrel{not.}{=} \max(1 - y_i(w^* \cdot x_i + w_0^*), 0)$, rezultă că w^*, w_0^*, ξ^* este pentru problema C-SVM (P') o soluție fezabilă, adică satisfac sistemul de restricții $y_i(w \cdot x_i + w_0) \geq 1 - \xi_i$ și $\xi_i \geq 0$, pentru $i = 1, \dots, m$. Pe lângă aceasta, notând cu P' funcția obiectiv a problemei de optimizare C-SVM (P'), rezultă

$$\underbrace{P'(w^*, w_0^*, \xi^*)}_{= \frac{1}{2} L(w^*, w_0^*)} < \underbrace{P'(\bar{w}, \bar{w}_0, \bar{\xi})}_{= \frac{1}{2} L(\bar{w}, \bar{w}_0)}.$$

În această relație, egalitatea $P'(w^*, w_0^*, \xi^*) = \frac{1}{2} L(w^*, w_0^*)$ are loc datorită modului în care am definit L , P' și ξ^* , iar egalitatea $P'(\bar{w}, \bar{w}_0, \bar{\xi}) = 2L(\bar{w}, \bar{w}_0)$ are loc datorită operatorului min din cadrul obiectivului problemei (P'). Inegalitatea $P'(w^*, w_0^*, \xi^*) < P'(\bar{w}, \bar{w}_0, \bar{\xi})$ implică faptul că soluția w^*, w_0^*, ξ^* constituie pentru problema (P') o soluție mai bună decât soluția optimă $\bar{w}, \bar{w}_0, \bar{\xi}$, ceea ce este absurd. Prin urmare, \bar{w}, \bar{w}_0 este soluție optimă a problemei (L).

Invers, trebuie să arătăm acum că luând $\lambda = 2C$, orice soluție optimă \bar{w}, \bar{w}_0 a problemei (L) este – printr-o extensie naturală – soluție optimă a problemei C-SVM (P'). Este imediat (vedeți raționamentul de mai sus) că orice soluție optimă \bar{w}, \bar{w}_0 a problemei (L), augmentată cu $\xi_i \stackrel{not.}{=} \max(1 - y_i(\bar{w} \cdot x_i + \bar{w}_0), 0)$ satisfac restricțiile problemei (P'). Presupunem, ca și mai sus, prin reducere la absurd că $\bar{w}, \bar{w}_0, \bar{\xi}$ nu este soluție optimă a problemei (P'). Considerând w^*, w_0^*, ξ^* o soluție optimă a problemei (P'), rezultă:

$$\underbrace{P'(w^*, w_0^*, \xi^*)}_{= \frac{1}{2} L(w^*, w_0^*)} < \underbrace{P'(\bar{w}, \bar{w}_0, \bar{\xi})}_{= \frac{1}{2} L(\bar{w}, \bar{w}_0)}.$$

Așadar, $L(w^*, w_0^*) < L(\bar{w}, \bar{w}_0)$, ceea ce contrazice faptul că \bar{w}, \bar{w}_0 este soluție optimă pentru problema (L). Conchidem că presupunerea făcută mai sus este falsă, deci $\bar{w}, \bar{w}_0, \bar{\xi}$ este soluție optimă pentru problema (P').

b. $\lambda = 2C$, conform rezolvării de la punctul a.

c. Forma nou-obținută pentru problema de optimizare C-SVM este caracterizată de i. lipsa restricțiilor asupra variabilelor și ii. de realizarea unui echilibru (engl., trade-off) între

- simplitate,⁶⁵⁹ reflectată de termenul $\|w\|^2$;
- o bună capacitate de predicție / generalizare în cazul neseparabilității liniare a datelor de antrenament, grație termenului $\lambda \sum_{i=1}^m \max(1 - y_i(w \cdot x_i + w_0), 0)$.

Observație importantă: [Relația cu regresia logistică]

Este imediat că problema de optimizare (L) este echivalentă cu problema

$$\min_{w, w_0} \left(\theta \|w\|^2 + \sum_{i=1}^m \max(1 - y_i(w \cdot x_i + w_0), 0) \right), \quad (302)$$

dacă se consideră parametrul (fixat) $\theta = \frac{1}{\lambda} > 0$. Comparând problema de optimizare (302) cu problema de optimizare din definiția *regresiei logistice* cu

⁶⁵⁹În raport cu alți clasificatori, de exemplu rețelele neuronale artificiale.

termen de regularizare L_2 ,⁶⁶⁰ putem observa o mare similaritate. Diferența constă (doar!) în funcția de cost / pierdere folosită: SVM folosește *funcția de cost hinge*, definită prin $f(z) = \max(1 - z, 0)$, pe când regresia logistică folosește *funcția de cost logistică* $\ln(1 + e^{-z})$.⁶⁶¹

Observație: Pentru detalii de implementare (și aplicare) a acestei forme a problemei de optimizare SVM folosind *metoda subgradientului* (pentru că funcția *hinge* nu este derivabilă pe tot domeniul de definiție),⁶⁶² vedeti articolul *Pegasos: Primal estimated sub-gradient solver for SVM*, de Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, Andrew Cotter, publicat în revista *Mathematical programming*, 127(1):3–30, 2011.⁶⁶³

21.

(Algoritmul Pegasos [simplificat]: justificarea regulii de actualizare; kernelizare)

- formulare de Liviu Ciortuz, pornind de la CMU, 2017 fall, Nina Balcan, HW4, pr. 4.1⁶⁶⁴

Fie $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$ o mulțime de n exemple de antrenament, cu $x_i \in \mathbb{R}^d$ și $y_i \in \{-1, +1\}$ pentru $i = 1, \dots, n$.

Considerăm următoarea formulare a problemei de optimizare C-SVM fără restricții, folosind funcția de cost / pierdere (engl., loss) *hinge*:⁶⁶⁵

$$\min_{w \in \mathbb{R}^d} \left(\frac{\lambda}{2} \|w\|^2 + \frac{1}{n} \sum_{i=1}^n \max(1 - y_i(w \cdot x_i), 0) \right), \quad (303)$$

unde λ este un parametru real pozitiv fixat.⁶⁶⁶ Remarcați faptul că în această versiune nu folosim termen liber (w_0).

Scopul acestui exercițiu este să arătăm că funcția obiectiv din relația (303) poate fi optimizată folosind *metoda subgradientului descendant stochastic* (engl., Stochastic Sub-Gradient Descent, SSGD).⁶⁶⁷ Această abordare este foarte simplă și se scalează ușor la seturi mari de date de antrenament. În cadrul metodei / algoritmului SSGD selectăm în mod aleatoriu un exemplu de antrenament

⁶⁶⁰ Puteți vedea problemele 13.c, 15.a și 16 de la capitolul *Metode de regresie*, mai precis relațiile (178), (182), (191) și (192).

⁶⁶¹ În mod similar, regresia liniară folosește ca funcție de cost suma pătratelor erorilor, iar algoritmul AdaBoost folosește funcția de cost [negativ]-exponențială e^{-z} . Pentru un cadru unitar de prezentare a acestor metode de învățare (foarte diferite!), bazat pe minimizarea costurilor / pierderilor, vă recomandăm să citiți documentul *Supplemental lecture notes*, de John Duchi, de la Universitatea Stanford.

⁶⁶² Pentru definiția noțiunii de *subgradient*, vedeti pr. 81 de la capitolul de *Fundamente*.

⁶⁶³ Pentru implementarea unei versiuni simplificate a acestui algoritm și aplicarea lui, vedeti problema 21 și CMU, 2017 fall, Nina Balcan, HW4, pr. 4.1.

Pentru aplicarea metodei subgradientului în cazul regresiei liniare cu regularizare de normă L_1 , vedeti ex. 27 de la capitolul *Metode de regresie*.

⁶⁶⁴ În versiunea originală — CMU, 2017 fall, Nina Balcan, HW4, pr. 4.1 —, această problemă este de tip implementare; acolo se cere să se implementeze algoritmul Pegasos simplificat (prezentat aici) și apoi să se aplice pe un subset al setului de date MNIST.

⁶⁶⁵ Vedeti problema 20.

⁶⁶⁶ Corespondența dintre parametrul θ din relația (302) și parametrul λ din relația (303) este următoarea:

$$\theta = \frac{n\lambda}{2}.$$

⁶⁶⁷ Pentru o variantă (nestochastică) a algoritmului subgradientului descendant, vedeti problema 168 de la capitolul de *Fundamente*. Acolo se arată că algoritmul Perceptron (Rosenblatt) poate fi interpretat ca fiind o instanță a algoritmului subgradientului descendant.

la fiecare iterație și actualizăm vectorul de ponderi w făcând un pas (engl., step) în direcția opusă *subgradientului* funcției de cost / pierdere.⁶⁶⁸ Forma algoritmului SSGD pentru problema dată este următoarea:

```

Input:  $S, \lambda, \eta, T$ ;
initialize the weight vector  $w = 0$ ;
for  $t = 1, \dots, T$ 
{
    choose  $i_t \in \{1, \dots, n\}$  uniformly at random;
    set  $\eta_t = \frac{1}{\lambda t}$ ;
    if  $y_{i_t} w \cdot x_{i_t} < 1$  then
        set  $w \leftarrow (1 - \lambda \eta_t)w + \eta_t y_{i_t} x_{i_t}$ ;
    else
        set  $w \leftarrow (1 - \lambda \eta_t)w$ ;
    }
end for;
return  $w$ ;

```

a. Arătați că regula de actualizare a ponderilor w din acest algoritm corespunde într-adevăr subgradientului descendenter stochastic.

b. Arătați că acest algoritm poate fi kernelizat.

Sugestie: Veți justifica mai întâi faptul că soluția w poate fi scrisă ca o combinație liniară de instanțe de antrenament x_i , iar apoi că în regula de decizie $sign(w \cdot x)$, instanțele de antrenament x_i și instanța de test x apar doar ca argumente ale produsului scalar (și, în consecință, doar ca argumente ale funcției-nucleu considerate, K , atunci când se lucrează în spațiul de trăsături).

Comentariu: Algoritmul care a fost dat în enunț este o versiune simplificată / particulară în raport cu algoritmul Pegasos:⁶⁶⁹

```

Input:  $S, \lambda, T, k$ ;
choose  $w_1$  s. t.  $\|w_1\| \leq 1/\sqrt{\lambda}$ ;
for  $t = 1, \dots, T$ 
{
    choose  $A_t \subseteq S$ , where  $|A_t| = k$ ;
    set  $A_t^+ = \{(x, y) \in A_t : y(w_t \cdot x) < 1\}$ ;
    set  $\eta_t = \frac{1}{\lambda t}$ ;
    set  $w_{t+\frac{1}{2}} = (1 - \lambda \eta_t)w_t + \frac{\eta_t}{k} \sum_{(x, y) \in A_t^+} yx$ ;
    set  $w_{t+1} = \min \left\{ 1, \frac{1/\sqrt{\lambda}}{\|w_{t+\frac{1}{2}}\|} \right\} w_{t+\frac{1}{2}}$ ;
}
end for;
return  $w_{T+1}$ ;

```

Mulțimea A_t este formată din k exemple selectate i.i.d. din setul de date de antrenament S .

Observații:

1. Nici în acest algoritm nu folosim termen liber (w_0). (Pentru varianta în care se folosește

⁶⁶⁸ Notiunea de subgradient generalizează notiunea de gradient în cazul funcțiilor care sunt nederivabile pe tot domeniul de definiție. Vedeți problema 81 de la capitolul de *Fundamente*.

⁶⁶⁹ Vedeți articolul *Pegasos: Primal Estimated sub-GrAdient SOLver for SVM*, de Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, Andrew Cotter, *Mathematical programming*, 127(1):3–30, 2011.

termen liber, vedeți articolul indicat la nota de subsol 669.)

2. La elaborarea acestui algoritm, mai exact la iterația t , funcția obiectiv din relația (303) este înlocuită cu o aproximare a ei:

$$\frac{\lambda}{2} \|w\|^2 + \frac{1}{k} \sum_{(x,y) \in A_t} \max(1 - y(w_t \cdot x), 0).$$

3. w_{t+1} este proiecția lui $w_{t+\frac{1}{2}}$ pe suprafața hipersferei $B = \{w : \|w\| \leq 1/\sqrt{\lambda}\}$, întrucât se poate demonstra că soluția optimă a problemei (303) se află în această hipersferă (vedeți articolul indicat la nota de subsol 669).

c. Arătați că dacă în algoritm Pegasos se setează w_1 la vectorul 0 din \mathbb{R}^d , atunci acest algoritm este kernelizabil. (Atenție! Raționamentul este similar cu cel de la punctul b. Însă, în plus, va trebui să arătați că $\|w_{t+\frac{1}{2}}\|$ se calculează folosind instanțele x_i doar ca argumente pentru funcția-nucleu K .)

Răspuns:

a. Justificarea faptului că regula de actualizare a ponderilor din algoritm prezentat în enunț corespunde subgradientului descendente stochastic a fost deja făcută la rezolvarea problemei 81.b de la capitolul de *Fundamente*. Trebuie să mai precizăm doar că

i. „stochasticitatea“ este dată de alegerea aleatorie a instanței de antrenament care este tratată la iterația t ;

ii. faptul că înainte de a se adăuga la vectorul w componenta $\Delta \stackrel{\text{not.}}{=} \eta_t y_{i-t} x_{i_t}$ se face mișorarea mărimei vectorului w (și anume, prin înmulțirea sa cu constanta $1 - \lambda \eta_t$) în vederea reducerii riscului de overfitting, ține de algoritm propriu-zis nu de metoda subgradientului descendente.

b. Pentru antrenare, vom arăta mai întâi prin inducție completă că în „spațiul de intrare“ la orice iterație t vectorul w este o combinație liniară de instanțele de antrenament x_i .

Pentru iterația $t = 0$ (adică la inițializare), avem $w = 0$, deci este verificată ipoteza inductivă. Pentru pasul inductiv, presupunem că la iterația t vectorul w se scrie ca o combinație liniară de instanțe de antrenament x_i . Înțând cont că regula de actualizare a vectorului w se poate scrie sub forma

$$w \leftarrow (1 - \lambda \eta_t)w + \begin{cases} \eta_t y_{i+t} x_{i_t}, & \text{dacă } y_{i_t} w \cdot x_{i_t} < 1 \\ 0, & \text{altfel,} \end{cases}$$

rezultă că noua valoare a vectorului w care este calculată la iterația t este tot o combinație liniară de instanțe de antrenament x_i .

Așadar, conform principiului inducției complete, la orice iterație t vectorul w poate fi scris ca o combinație liniară de instanțe de antrenament x_i .

În consecință, scriind w calculat în „spațiul de trăsături“ sub forma $\sum_{j=1}^m \alpha_j \phi(x_j)$ unde $\alpha_j \in \mathbb{R}$ pentru $j = 1, \dots, m$, devine imediat faptul că produsele $w \cdot \phi(x_{i_t})$ se pot calcula cu ajutorul funcției-nucleu K :

$$\left(\sum_{j=1}^m \alpha_j \phi(x_j) \right) \cdot \phi(x_{i_t}) = \sum_{j=1}^m \alpha_j \phi(x_j) \cdot \phi(x_{i_t}) = \sum_{j=1}^m \alpha_j K(x_j, x_{i_t}).$$

În sfârșit, pentru testare / generalizare facem un raționament similar:

$$\left(\sum_{j=1}^m \alpha_j \phi(x_j) \right) \cdot \phi(x) = \sum_{j=1}^m \alpha_j \phi(x_j) \cdot \phi(x) = \sum_{j=1}^m \alpha_j K(x_j, x).$$

c. După cum s-a precizat deja în enunț, justificarea kernelizării algoritmului Pegasos [simplificat] este similară cu demonstrația de la punctul b; doar constantele din fața instanțelor x_i care apar în scrierea vectorului w se modifică. Rămâne doar să arătăm că în „spațiul de trăsături“ $\|w_{t+\frac{1}{2}}\|$ se poate calcula în aşa fel încât [în acest calcul] instanțele x_i să apară doar ca argumente ale funcției-nucleu K . Într-adevăr,

$$\begin{aligned} \|w_{t+\frac{1}{2}}\|^2 &= w_{t+\frac{1}{2}}^2 = w_{t+\frac{1}{2}} \cdot w_{t+\frac{1}{2}} \\ &= \left((1 - \lambda \eta_t) w_t + \frac{\eta_t}{k} \sum_{(x,y) \in A_t^+} y \phi(x) \right)^2 \\ &= (1 - \lambda \eta_t)^2 w_t^2 + 2(1 - \lambda \eta_t) \frac{\eta_t}{k} w_t \cdot \left(\sum_{(x,y) \in A_t^+} y \phi(x) \right) + \frac{\eta_t^2}{k^2} \left(\sum_{(x,y) \in A_t^+} y \phi(x) \right)^2. \end{aligned}$$

Justificarea este practic finalizată dacă folosim faptul că în „spațiul de trăsături“ vectorul w_t este o combinație liniară de vectorii $\phi(x_i)$ (deci w_t^2 se scrie ca o combinație liniară de „imagini“ ale funcției-nucleu, $K(x_i, x_j)$), iar

$$\begin{aligned} \left(\sum_{(x,y) \in A_t^+} y \phi(x) \right)^2 &= \left(\sum_{(x,y) \in A_t^+} y \phi(x) \right) \cdot \left(\sum_{(x',y') \in A_t^+} y' \phi(x') \right) \\ &= \sum_{(x,y) \in A_t^+} \sum_{(x',y') \in A_t^+} yy' \phi(x) \cdot \phi(x') = \sum_{(x,y) \in A_t^+} \sum_{(x',y') \in A_t^+} yy' K(x, x'). \end{aligned}$$

22.

(C-SVM: deducerea relațiilor folosite în algoritmul SMO)

□ • ○ *Liviu Ciortuz, 2018, după
■ Nello Cristianini, John Shawe-Taylor,
An Introduction to Support Vector Machines,
Cambridge University Press, 2000, pp. 139-140*

În această problemă veți face deducerea relațiilor (pentru *actualizare* și respectiv pentru *oprire*) folosite în algoritmul SMO,⁶⁷⁰ care rezolvă forma duală a problemei de optimizare C-SVM. Acest algoritm folosește ca strategie de optimizare *metoda creșterii pe coordonate* (engl., *coordinate ascent*).⁶⁷¹ Aceasta

⁶⁷⁰ John C. Platt, *Sequential Minimal Optimization: A fast algorithm for training Support Vector Machines*. Microsoft Research, Technical report MSR-TR-98-14, 1998. O variantă simplificată a algoritmului SMO este prezentată de Andrew Ng în *Machine Learning, Lecture Notes*, Part V, section 9.

⁶⁷¹ Pentru folosirea *metodei creșterii pe coordonate* (engl., *coordinate descent*) la rezolvarea regresiei liniare cu regularizare de normă L_1 , vedeti ex. 11 de la capitolul *Metode de regresie*. Algoritmii AdaBoost și AdaBoost generalizat (ex. 26 și 29 de la cap. *Arbore de decizie*) folosesc și ei metoda creșterii pe coordonate, iar una dintre cele două coordonate corespunde așa-numitelor ipoteze „slabe“ (care nu neapărat trebuie optimizate). De asemenea, metoda creșterii pe coordonate este folosită în mod indirect la algoritmul *K-means* (văzut ca algoritm de minimizare a criteriului sumei celor mai mici pătrate (J)) și în mod direct la algoritmul EM; vedeti problema 12.a de la capitolul de *Clusterizare* și respectiv problema 1 de la capitolul *Schema algoritmică EM*.

este o metodă iterativă, pentru optimizarea unei funcții de mai multe variabile — în cazul nostru funcția concavă $L_D(\alpha_1, \dots, \alpha_m)$. În *cazul clasic* al aplicării acestei metode, la fiecare iterație se alege câte o variabilă care este lăsată liberă, iar celelalte variabile sunt fixate; la iterația respectivă, optimizarea funcției obiectiv se face în raport cu variabila liberă. Însă, spre deosebire de cazul clasic, la aplicarea metodei creșterii pe coordonate pentru rezolvarea problemei duale C-SVM, datorită restricției $\sum_{i=1}^m y_i \alpha_i = 0$ va trebui ca la fiecare iterație să alegem două variabile pe care să le lăsăm libere, iar pe restul să le fixăm:⁶⁷²

Repeat until convergence{

1. Select some pair α_i and α_j to update next (using a heuristic that tries to pick the two that will allow us to make the biggest progress towards the global maximum).
2. Reoptimize $L_D(\alpha)$ with respect to α_i and α_j , while holding all the other α_k 's ($k \neq i, j$) fixed.

}

a. Folosind notațiile standard de la problema de optimizare C-SVM (în forma primală și respectiv forma duală; vedeți pr. 12) și presupunând că *variabilele libere* la iterația curentă sunt α_1 și α_2 , demonstrați că lagrangeanul dual L_D are la această iterație soluțiile optime

$$\begin{aligned}\alpha_2^{new, unclipped} &= \alpha_2^{old} + \frac{y_2(E_2 - E_1)}{\eta} \\ \alpha_1^{new, unclipped} &= \alpha_1^{old} + y_1 y_2 (\alpha_2^{old} - \alpha_2^{new, unclipped}),\end{aligned}$$

unde⁶⁷³

$$\begin{aligned}w &= \sum_{i=1}^m y_i \alpha_i x_i \\ E_k &= \underbrace{w \cdot x_k + w_0}_{\text{not. : } f(x_k)} - y_k \text{ pentru } k \in \{1, 2\} \\ \eta &= -\|x_1 - x_2\|^2.\end{aligned}$$

b. Vă readucem aminte că $0 \leq \alpha_j \leq C$ pentru $j \in \{1, 2\}$ (vedeți problema 12.d). La fiecare iterație a algoritmului SMO trebuie calculate două *margini* (engl., bounds), L și H , astfel încât să restricționăm și mai mult variabila α_2 : $0 \leq L \leq \alpha_2 \leq H \leq C$.

Demonstrați că (în contextul problemei noastre) aceste margini sunt conform următoarelor relații:

- dacă $y_1 \neq y_2$, atunci $L = \max(0, \alpha_2 - \alpha_1)$, $H = \min(C, C + \alpha_2 - \alpha_1)$;
- dacă $y_1 = y_2$, atunci $L = \max(0, \alpha_1 + \alpha_2 - C)$, $H = \min(\alpha_1 + \alpha_2, C)$.

În consecință, regulile de actualizare vor fi:

$$\begin{aligned}\alpha_2^{new, clipped} &= \begin{cases} H & \text{dacă } \alpha_2^{new, unclipped} > H \\ \alpha_2^{new, unclipped} & \text{dacă } L \leq \alpha_2^{new, unclipped} \leq H \\ L & \text{dacă } \alpha_2^{new, unclipped} < L \end{cases} \\ \alpha_1^{new, clipped} &= \alpha_1^{old} + y_1 y_2 (\alpha_2^{old} - \alpha_2^{new, clipped}),\end{aligned}\tag{304}$$

⁶⁷²În prezentă problemă nu includem detalii despre modul în care se aleg aceste două variabile.

⁶⁷³În scrierea lui w sub forma $\sum_{i=1}^m y_i \alpha_i x_i$ se va subînțelege că α_1 și α_2 sunt de fapt α_1^{old} și respectiv α_2^{old} .

unde calificativele *new* și *old* desemnează noile și respectiv vechile valori ale variabilelor α_1 și α_2 , iar calificativul *clipped* (spre deosebire de *unclipped*) corespunde „ajustării“ valorii lui α_2 ca urmare a aplicării restricției $L \leq \alpha_2 \leq H$.

c. Algoritmul SMO folosește următoarea *condiție de oprire*:

$$\text{NOT } [(\alpha_i < C \text{ AND } y_i E_i < -tol) \text{ OR } (\alpha_i > 0 \text{ AND } y_i E_i > tol)], i = 1, \dots, m,$$

unde tol este un număr pozitiv (mic, fixat în prealabil), numit *parametru de toleranță*. Demonstrați că această condiție este implicată în mod natural de relațiile de tip Karush-Kuhn-Tucker (300) care au fost deduse pentru problema de optimizare C-SVM (vedeți problema 12.e).

Răspuns:

a. Știm — conform problemei 12.d — că forma duală a problemei de optimizare C-SVM are funcția obiectiv $L_D(\alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m y_i y_j \alpha_i \alpha_j x_i \cdot x_j$, cu restricțiile $\sum_{i=1}^m y_i \alpha_i = 0$ (atât pentru noile cât și pentru vechile valori ale variabilelor) și $0 \leq \alpha_i \leq C$ pentru $i = 1, \dots, m$. Notând

$$v_i = \left(\underbrace{\sum_{j=3}^m y_j \alpha_j x_j}_{w - (y_1 \alpha_1 x_1 + y_2 \alpha_2 x_2)} \right) \cdot x_i \text{ pentru } i = 1, 2, \quad (305)$$

va rezulta că

$$v_i = w \cdot x_i - (y_1 \alpha_1 x_1 \cdot x_i + y_2 \alpha_2 x_2 \cdot x_i) \text{ pentru } i \in \{1, 2\} \quad (306)$$

și

$$\begin{aligned} L_D(\alpha_1, \alpha_2) &= \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m y_i y_j \alpha_i \alpha_j x_i \cdot x_j \\ &= \alpha_1 + \alpha_2 + \sum_{j=3}^m \alpha_j - \frac{1}{2} \alpha_1^2 x_1^2 - \frac{1}{2} \alpha_2^2 x_2^2 - \frac{1}{2} \sum_{j=3}^m \alpha_j^2 x_j^2 - y_1 y_2 \alpha_1 \alpha_2 x_1 \cdot x_2 \\ &\quad - \underbrace{y_1 \alpha_1 x_1 \cdot \sum_{j=3}^m y_j \alpha_j x_j}_{y_1 \alpha_1 v_1} - \underbrace{y_2 \alpha_2 x_2 \cdot \sum_{j=3}^m y_j \alpha_j x_j}_{y_2 \alpha_2 v_2} - \frac{1}{2} \sum_{j=3}^m \sum_{j'=3}^m y_j y_{j'} \alpha_j \alpha_{j'} x_j \cdot x_{j'} \\ &= \alpha_1 + \alpha_2 - \frac{1}{2} \alpha_1^2 x_1^2 - \frac{1}{2} \alpha_2^2 x_2^2 - \underbrace{y_1 y_2}_{\text{not.: } s} \alpha_1 \alpha_2 x_1 \cdot x_2 - y_1 \alpha_1 v_1 - y_2 \alpha_2 v_2 + \text{const}_1. \end{aligned}$$

Întrucât $y_1 \alpha_1 + y_2 \alpha_2 = -\sum_{j=3}^m y_j \alpha_j$, înmulțind această egalitate cu y_1 — despre care știm că aparține mulțimii $\{-1, 1\}$ —, vom obține

$$\alpha_1 + \underbrace{y_1 y_2}_{s} \alpha_2 = -y_1 \underbrace{\sum_{j=3}^m y_j \alpha_j}_{\text{const}_2}.$$

Aşadar,

$$\alpha_1^{old} + s\alpha_2^{old} = \underbrace{const_2}_{not.: \gamma} = \alpha_1^{new, unclipped} + s\alpha_2^{new, unclipped}. \quad (307)$$

Substituind $\alpha_1 = \gamma - s\alpha_2$ în expresia lagrangeanului $L_D(\alpha_1, \alpha_2)$, vom obține — renunțând totodată la argumentul α_1 — următoarea expresie:

$$\begin{aligned} L_D(\alpha_2) &= \gamma - s\alpha_2 + \alpha_2 - \frac{1}{2}(\gamma - s\alpha_2)^2 x_1^2 - \frac{1}{2}\alpha_2^2 x_2^2 - s(\gamma - s\alpha_2)\alpha_2 x_1 \cdot x_2 \\ &\quad - y_1(\gamma - s\alpha_2)v_1 - y_2\alpha_2 v_2 + const_1. \end{aligned}$$

Pentru a maximiza $L_D(\alpha_2)$, mai întâi vom calcula derivata sa în raport cu α_2 , după care vom egala această derivată cu 0:

$$\begin{aligned} \frac{\partial L_D(\alpha_2)}{\partial \alpha_2} &= -s + 1 + \frac{1}{2}2s(\gamma - s\alpha_2)x_1^2 - \frac{1}{2}2\alpha_2 x_2^2 - s\gamma x_1 \cdot x_2 + 2\alpha_2 x_1 \cdot x_2 + y_1 sv_1 - y_2 v_2 \\ &= -s + 1 + (s\gamma - \alpha_2)x_1^2 - \alpha_2 x_2^2 - s\gamma x_1 \cdot x_2 + 2\alpha_2 x_1 \cdot x_2 + y_1 sv_1 - y_2 v_2 \\ &= -\alpha_2(x_1^2 + x_2^2 - 2x_1 \cdot x_2) - s + 1 + s\gamma x_1^2 - s\gamma x_1 \cdot x_2 + y_1 \underbrace{s}_{y_1 y_2} v_1 - y_2 v_2 \\ &= -\alpha_2(x_1 - x_2)^2 - s + 1 + s\gamma x_1^2 - s\gamma x_1 \cdot x_2 + y_2 v_1 - y_2 v_2 = 0. \end{aligned}$$

Întrucât $(x_1 - x_2)^2 > 0$ pentru orice $x_1 \neq x_2$, rezultă că semnele acestei derive corespund existenței maximului pentru $L_D(\alpha_2)$. Așadar, soluția este:

$$\begin{aligned} \alpha_2^{new, unclipped} &= \frac{-s + 1 + s\gamma(x_1^2 - x_1 \cdot x_2) + y_2(v_1 - v_2)}{\|x_1 - x_2\|^2} \\ &= \frac{y_2(-y_1 + y_2 + y_1\gamma(x_1^2 - x_1 \cdot x_2) + v_1 - v_2)}{\|x_1 - x_2\|^2} \\ &\stackrel{(309)}{=} \frac{y_2(f(x_1) - y_1 - f(x_2) + y_2) + \alpha_2 \|x_1 - x_2\|^2}{\|x_1 - x_2\|^2} \quad (308) \\ &= \alpha_2 + \frac{y_2(E_1 - E_2)}{\|x_1 - x_2\|^2} = \alpha_2 + \frac{y_2(E_2 - E_1)}{-\|x_1 - x_2\|^2}. \end{aligned}$$

Observații:

1. Atât în relația (308) și pe linia care o succedă, α_2 trebuie înțeles ca fiind α_2^{old} .
2. Conform relației (305), valoarea expresiei $v_2 - v_1$ nu depinde de α_1 sau de α_2 . Dacă în scrierea lui $w = \sum_j y_j \alpha_j x_j$ vom considera că α_1 și α_2 apar în accepțiunea α_1^{old} și respectiv α_2^{old} , nimic nu ne împiedică să facem rationamentul următor, obținând în final relația (309). De fapt, relația (309) și a fost invocată mai sus la deducerea relației (308).

$$\begin{aligned} v_1 - v_2 &\stackrel{(306)}{=} f(x_1) - y_1 \alpha_1 x_1^2 - y_2 \alpha_2 x_1 \cdot x_2 - (f(x_2) - y_1 \alpha_1 x_1 \cdot x_2 - y_2 \alpha_2 x_2^2) \\ &= f(x_1) - y_1 \underbrace{\alpha_1 x_1^2}_{\gamma - s\alpha_2} - y_2 \alpha_2 x_1 \cdot x_2 - f(x_2) + y_1 \underbrace{\alpha_1 x_1 \cdot x_2}_{\gamma - s\alpha_2} + y_2 \alpha_2 x_2^2 \\ &= f(x_1) - f(x_2) - y_1 \gamma x_1^2 + \underbrace{s y_1 \alpha_2 x_1^2}_{y_2} - y_2 \alpha_2 x_1 \cdot x_2 + y_1 \gamma x_1 \cdot x_2 - \underbrace{s y_1 \alpha_2 x_1 \cdot x_2}_{y_2} + \\ &\quad y_2 \alpha_2 x_2^2 \\ &= f(x_1) - f(x_2) - y_1 \gamma x_1^2 + y_2 \alpha_2 x_1^2 - y_2 \alpha_2 x_1 \cdot x_2 + y_1 \gamma x_1 \cdot x_2 - y_2 \alpha_2 x_1 \cdot x_2 + y_2 \alpha_2 x_2^2 \\ &= f(x_1) - f(x_2) - y_1 \gamma(x_1^2 - x_1 \cdot x_2) + y_2 \alpha_2(x_1^2 - 2x_1 \cdot x_2 + x_2^2) \\ &= f(x_1) - f(x_2) - y_1 \gamma(x_1^2 - x_1 \cdot x_2) + y_2 \alpha_2 \|x_1 - x_2\|^2. \quad (309) \end{aligned}$$

În final, pentru a deduce $\alpha_1^{new, unclipped}$, vom folosi egalitatea dublă

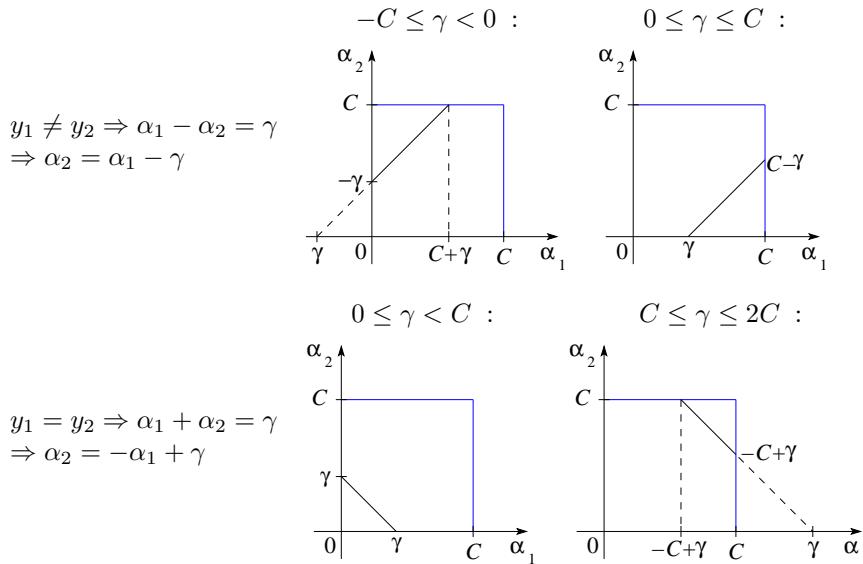
$$\alpha_1^{old} + s\alpha_2^{old} = \gamma = \alpha_1^{new, unclipped} + s\alpha_2^{new, unclipped}.$$

Așadar,

$$\begin{aligned}\alpha_1^{new, unclipped} &= \gamma - s\alpha_2^{new, unclipped} = \alpha_1^{old} + s\alpha_2^{old} - s\alpha_2^{new, unclipped} \\ &= \alpha_1^{old} + y_1 y_2 (\alpha_2^{old} - \alpha_2^{new, unclipped}).\end{aligned}$$

Observație: Egalitatea aceasta corespunde cu ultima relație care a fost inclusă [în vederea rezolvării] în enunțul de la punctul b.

b. Vom analiza pe rând cele două cazuri: $y_1 \neq y_2$ (care implică $\gamma = \alpha_1 - \alpha_2$, conform relației (307)) și respectiv $y_1 = y_2$ (care implică $\gamma = \alpha_1 + \alpha_2$). Pentru fiecare din aceste două cazuri, în funcție de semnul lui γ $\stackrel{not.}{=} \alpha_1 + s\alpha_2$ și respectiv în funcție de semnul lui $\gamma - C$, vom avea câte două subcazuri, reprezentate grafic în partea dreaptă.



Pentru cazul $y_1 \neq y_2$, se observă din cele două grafice că $\alpha_2 \geq -\gamma$ și respectiv $\alpha_2 \geq 0$, deci $\alpha_2 \geq L = \max(-\gamma, 0)$. De asemenea, $\alpha_2 \leq C$ și respectiv $\alpha_2 \leq C - \gamma$, de unde rezultă $H = \min(C, C - \gamma)$. Similar, pentru cazul $y_1 = y_2$ vom obține $L = \max(0, -C + \gamma)$ și $H = \min(\gamma, C)$.

În fine, regula de actualizare (304) decurge în mod natural din proprietățile funcției de gradul al doilea aplicate pentru $L_D(\alpha_2)$, care trebuie maximizată ținând cont de restricțiile $L \leq \alpha_2 \leq H$.⁶⁷⁴

c. Vom folosi egalitatea $y_i E_i = y_i(f(x_i) - y_i) = y_i f(x_i) - 1$. Conform relațiilor (300), cazurile $\alpha_i < C$ și $\alpha_i > 0$ vor avea câte două subcazuri:

⁶⁷⁴ Pentru a vă convinge, vă recomandăm să faceți câte un grafic generic pentru funcția $L_d(\alpha_2)$ pentru fiecare din cele trei cazuri ale regulii (304). (Pentru două exemple concrete, puteți vedea graficele de la rezolvarea problemei 23.b.)

$$\alpha_i < C : \left\{ \begin{array}{l} \alpha_i = 0 \Rightarrow y_i \underbrace{(w \cdot x_i + w_0)}_{f(x_i)} \geq 1 \Rightarrow y_i E_i \geq 0 \\ \alpha_i \in (0, C) \Rightarrow y_i f(x_i) = 1 \Rightarrow y_i E_i = 0 \end{array} \right\} \Rightarrow y_i E_i \geq 0 \quad (310)$$

$$\alpha_i > 0 : \left\{ \begin{array}{l} \alpha_i \in (0, C) \Rightarrow y_i f(x_i) = 1 \Rightarrow y_i E_i = 0 \\ \alpha_i = C \Rightarrow y_i f(x_i) \leq 1 \Rightarrow y_i E_i \leq 0 \end{array} \right\} \Rightarrow y_i E_i \leq 0. \quad (311)$$

Notând premisele și respectiv consecințele relațiilor (310) și (311) cu niște variabile din logica propozițiilor, și anume $p_1 \stackrel{\text{not.}}{=} (\alpha_i < C)$, $q_1 \stackrel{\text{not.}}{=} (y_i E_i \geq 0)$, $p_2 \stackrel{\text{not.}}{=} (\alpha_i > 0)$, și $q_2 \stackrel{\text{not.}}{=} (y_i E_i \leq 0)$, vom putea scrie următoarele echivalențe, folosind mai întâi legea dublei negații și apoi legile lui De Morgan:

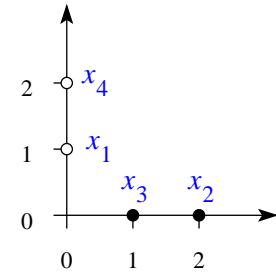
$$(p_1 \rightarrow q_1) \wedge (p_2 \rightarrow q_2) \equiv \neg\neg[(p_1 \rightarrow q_1) \wedge (p_2 \rightarrow q_2)] \equiv \neg[\neg(p_1 \rightarrow q_1) \vee \neg(p_2 \rightarrow q_2)] \equiv \neg[\neg(\neg p_1 \vee q_1) \vee \neg(\neg p_2 \vee q_2)] \equiv \neg[(p_1 \wedge \neg q_1) \vee (p_2 \wedge \neg q_2)].$$

Expresia finală — $\neg[(p_1 \wedge \neg q_1) \vee (p_2 \wedge \neg q_2)]$ — coincide cu *condiția de oprire* din enunț, doar că $\neg q_1$ și $\neg q_2$ au fost recrise (din motive de „stabilitate“ numerică la efectuarea calculelor) sub forma $y_i E_i < -tol$, și respectiv $y_i E_i > tol$, unde tol este constanta pozitivă considerată în enunț. Atunci când această condiție este satisfăcută pentru $i = 1, \dots, m$, algoritmul SMO poate fi oprit, întrucât [se consideră că] sunt satisfăcute condițiile de tip Karush-Kuhn-Tucker (300).

23. (C-SVM: exemplu de aplicare a algoritmului SMO)

■ • ○ CMU, 2008 fall, Eric Xing, HW2, pr. 1.3

Se dau patru instanțe în spațiul euclidian bidimensional, împreună cu etichetele asignate lor: $x_1 = (0, 1)$, $y_1 = -1$; $x_2 = (2, 0)$, $y_2 = +1$; $x_3 = (1, 0)$, $y_3 = +1$ și $x_4 = (0, 2)$, $y_4 = -1$. Vom folosi aceste exemple pentru a antrena un C-SVM (SVM liniar cu margine „soft“ și parametru de „destindere“ C). Fie $\alpha_1, \alpha_2, \alpha_3$ și α_4 multiplicatorii Lagrange asociati instanțelor x_1, x_2, x_3 și respectiv x_4 . Fixăm valoarea parametrului C la 100.



- a. Scrieți forma duală a problemei de optimizare C-SVM în acest caz.
- b. Vă cerem să executați două iterații ale algoritmului SMO (Sequential Minimal Optimization) pe acest set de date.⁶⁷⁵ Presupunem că facem inițializările $\alpha_1 = 5$, $\alpha_2 = 4$, $\alpha_3 = 8$, $\alpha_4 = 7$. Veți proceda astfel:
 - i. La prima iterație veți actualiza multiplicatorii α_1 și α_4 (păstrând α_2 și α_3 fixați). Stabiliti relațiile de actualizare a valorilor pentru α_1 și α_4 în funcție de valorile multiplicatorilor α_2 și α_3 . Ce valori vor avea α_1 și α_4 după actualizare?

⁶⁷⁵Vedeți problema 22 din prezentul capitol.

- ii. La a doua iterație veți fixa α_1 și α_4 și veți stabili relațiile de actualizare pentru α_2 și α_3 în funcție de α_1 și α_4 . Ce valori vor avea α_2 și α_3 după actualizare?

Răspuns:

- a. Particularizând forma duală a problemei de optimizare C-SVM (vedeți problema 12) pentru acest set de date, după efectuarea tuturor produselor scalare $x_i \cdot x_j$ vom obține:⁶⁷⁶

$$\max_{\alpha_1, \alpha_2, \alpha_3, \alpha_4} [\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - \frac{1}{2}(\alpha_1^2 + 4\alpha_2^2 + \alpha_3^2 + 4\alpha_4^2 + 4\alpha_1\alpha_4 + 4\alpha_2\alpha_3)]$$

- a. i. $0 \leq \alpha_i \leq 100$, pentru $i = 1, \dots, 4$
 $-\alpha_1 + \alpha_2 + \alpha_3 - \alpha_4 = 0$

- b. Facem *observația* că, la acest punct, exercițiul cere să se execute două iterări ale algoritmului SMO pe datele furnizate, fără a aplica criteriul de selecție a variabilelor libere și condițiile de oprire formulate de John Platt, autorul acestui algoritm.

Vom da mai jos două rezolvări. Prima va fi mai simplă. Ea va urma *ideea* algoritmului SMO — care aplică o metodă de optimizare numită *creștere pe coordonate* (engl., coordinate ascent) — fără a recurge efectiv la formulele stabilite de John Platt. În schimb, vom proceda direct la optimizarea funcțiilor obiectiv determinate de alegerea (impusă, conform enunțului) a celor două perechi de variabile duale specificate. La a doua rezolvare, vom aplica direct formulele generale pentru „actualizarea“ valorilor libere din algoritm SMO. Facem *observația* că și această rezolvare va fi utilă cititorului, pentru că vom scoate în evidență anumite detalii / modalități de calcul care nu sunt chiar „imediate“ pentru cineva care nu este încă obișnuit cu algoritmul SMO.

Prima soluție:

- i. La prima iterăție, inițial avem $\alpha_1 = 5$, $\alpha_2 = 4$, $\alpha_3 = 8$, $\alpha_4 = 7$, iar apoi se lasă „libere“ variabilele α_1 și α_4 . Datorită restricției $\sum_{i=1}^4 y_i \alpha_i = 0$ din forma duală a problemei de optimizare C-SVM (a se vedea (D') la problema 12), vom avea următoarea relație de legătură dintre valorile variabilelor libere: $\alpha_1^{new} + \alpha_4^{new} = \alpha_2 + \alpha_3 = 12$.

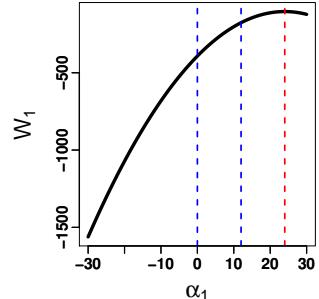
Din restricția $\alpha_i \geq 0$ — dar și datorită faptului că y_1 și y_4 au același semn — va rezulta că valorile posibile („fezabile“) pentru α_1 și α_4 vor fi limitate la intervalul $[0, 12]$, inclus în intervalul $[0, C] = [0, 100]$.

Înlocuind $\alpha_2 = 4$, $\alpha_3 = 8$ și $\alpha_4 = 12 - \alpha_1$ în funcția obiectiv a problemei de optimizare de la punctul a, vom obține expresia funcției pe care va trebui să o maximizăm la această iterăție:

$$\begin{aligned} W_1(\alpha_1) &\stackrel{not.}{=} 24 - \frac{1}{2}(\alpha_1^2 + 4 \cdot 4^2 + 8^2 + 4(12 - \alpha_1)^2 + 4\alpha_1(12 - \alpha_1) + 4 \cdot 4 \cdot 8) \\ &= 24 - \frac{1}{2}(\alpha_1^2 - 48\alpha_1 + 832). \end{aligned}$$

⁶⁷⁶Concret, $x_1^2 = x_3^2 = 1$, $x_2^2 = x_4^2 = 4$, $x_1 \cdot x_4 = x_2 \cdot x_3 = 2$, iar restul produselor $x_i \cdot x_j$ cu $i \neq j$ au valoarea 0.

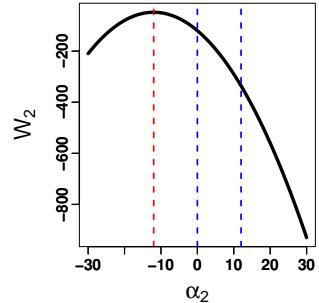
Valoarea lui α_1 pentru care se atinge optimul acestei funcții este notată cu $\alpha_1^{new, unclipped}$ și, evident, este $\frac{48}{2} = 24$. Această valoare se află în afara intervalului $[0, 12]$. Este imediat că maximul funcției $W_1(\alpha_1)$ pe intervalul $[0, 12]$ se atinge în punctul $\alpha_1^{new, clipped} = 12$. În consecință, α_4 va primi valoarea $\alpha_4^{new, unclipped} = 12 - \alpha_1^{new, clipped} = 0$.



ii. La a doua iterație, vom avea $\alpha_1 = 12$ și $\alpha_4 = 0$ fixați, iar $\alpha_2 = 4$ și $\alpha_3 = 8$ liberi. Din relația $\sum_{i=1}^4 y_i \alpha_i = 0$ rezultă $\alpha_2^{new} + \alpha_3^{new} = \alpha_1 + \alpha_4 = 12$. Ca și mai sus, intervalul în care vor fi permise noile valori ale variabilelor α_2 și α_3 este $[0, 12]$. Funcția obiectiv din problema de optimizare convexă devine:

$$\begin{aligned} W_2(\alpha_2) &\stackrel{not.}{=} 24 - \frac{1}{2}(12^2 + 4\alpha_2^2 + (12 - \alpha_2)^2 + 4 \cdot 0^2 + 4 \cdot 12 \cdot 0 + 4\alpha_2(12 - \alpha_2)) \\ &= -\frac{1}{2}\alpha_2^2 - 12\alpha_2 - 120 = -\frac{1}{2}(\alpha_2^2 + 24\alpha_2 + 240) \end{aligned}$$

Maximul global al acestei funcții se atinge în punctul $\alpha_2^{new, unclipped} = -\frac{24}{2} = -12$. Acest punct se situează în exteriorul intervalului de fezabilitate $[0, 12]$. Maximul funcției $W_2(\alpha_2)$ pe intervalul $[0, 12]$ se atinge în punctul $\alpha_2^{new, clipped} = 0$. În consecință, $\alpha_3^{new, unclipped} = 12 - \alpha_2^{new, clipped} = 12$.



A doua soluție:

Formulele date de John Platt pentru actualizarea variabilei libere α_i (la o iterare oarecare a algoritmului SMO) sunt:⁶⁷⁷

$$\begin{aligned} \alpha_i^{new, unclipped} &= \alpha_i + \frac{y_i(E_i - E_j)}{\eta} \\ \alpha_i^{new, clipped} &= \begin{cases} H & \text{dacă } \alpha_i^{new, unclipped} > H \\ \alpha_i^{new, unclipped} & \text{dacă } L \leq \alpha_i^{new, unclipped} \leq H \\ L & \text{dacă } \alpha_i^{new, unclipped} < L, \end{cases} \end{aligned}$$

unde

$$w = \sum_{i=1}^4 y_i \alpha_i x_i$$

$$E_k = w \cdot x_k + w_0 - y_k$$

$$\eta = -\|x_i - x_j\|^2$$

$$L = \max(0, \alpha_i - \alpha_j) \text{ și } H = \min(C, C + \alpha_i - \alpha_j) \text{ dacă } y_i \neq y_j$$

$$L = \max(0, \alpha_i + \alpha_j - C) \text{ și } H = \min(C, \alpha_i + \alpha_j) \text{ dacă } y_i = y_j.$$

⁶⁷⁷În comparație cu problema 22, aici nu vom mai scrie α_i^{old} , ci direct α_i , fiindcă se poate subînțelege fără nicio dificultate, datorită contextului mult mai simplu.

În consecință, vom avea:

i. La prima iterație, $\alpha_1 = 5$, $\alpha_2 = 4$, $\alpha_3 = 8$, $\alpha_4 = 7$, iar $\eta = -\|x_1 - x_4\|^2 = -1$. Fără a face deocamdată calculele, explicităm erorile $E_1 = f(x_1) - y_1 = w \cdot x_1 + w_0 - y_1$ și $E_4 = f(x_4) - y_4 = w \cdot x_4 + w_0 - y_4$, deci rezultă $E_1 - E_4 = w \cdot (x_1 - x_4)$. Din relația $w = \sum_{i=1}^4 y_i \alpha_i x_i$, urmează:

$$\begin{aligned} w &= -\alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 - \alpha_4 x_4 \\ &= -5 \begin{bmatrix} 0 \\ 1 \end{bmatrix} + 4 \begin{bmatrix} 2 \\ 0 \end{bmatrix} + 8 \begin{bmatrix} 1 \\ 0 \end{bmatrix} - 7 \begin{bmatrix} 0 \\ 2 \end{bmatrix} = \begin{bmatrix} 16 \\ -19 \end{bmatrix}, \end{aligned}$$

de unde rezultă că $E_1 - E_4 = w \cdot (x_1 - x_4) = (16, -19) \cdot (0, -1) = 19$ și vom putea calcula $\alpha_1^{new, unclipped} = \alpha_1 + \frac{y_1(E_1 - E_4)}{\eta} = 5 + 19 = 24$.

Acum verificăm dacă $\alpha_1^{new, unclipped}$ este în intervalul de „fezabilitate“: deoarece $y_1 = y_4$, vom avea $L = \max(0, \alpha_1 + \alpha_4 - 100) = 0$, pentru că $\alpha_1 + \alpha_4 = 12$. Similar, $H = \min(100, 12) = 12$.

Întrucât $\alpha_1^{new, unclipped} > H = 12$, vom avea $\alpha_1^{new, clipped} = H = 12$ și, în consecință, $\alpha_4^{new, clipped} = 12 - \alpha_1^{new, clipped} = 0$.

ii. La a doua iterație, $\alpha_1 = 12$, $\alpha_2 = 4$, $\alpha_3 = 8$, $\alpha_4 = 0$ și $\alpha_2^{new} + \alpha_3^{new} = \alpha_2 + \alpha_3 = \alpha_1 + \alpha_4 = 12$. De asemenea, $\eta = -\|x_2 - x_3\|^2 = -1$ și $E_2 = f(x_2) - y_2 = w \cdot x_2 + w_0 - y_2$, iar $E_3 = f(x_3) - y_3 = w \cdot x_3 + w_0 - y_3$, deci $E_2 - E_3 = w \cdot (x_2 - x_3)$. Calculăm w astfel:

$$\begin{aligned} w &= -\alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 - \alpha_4 x_4 \\ &= -12 \begin{bmatrix} 0 \\ 1 \end{bmatrix} + 4 \begin{bmatrix} 2 \\ 0 \end{bmatrix} + 8 \begin{bmatrix} 1 \\ 0 \end{bmatrix} - 0 \begin{bmatrix} 0 \\ 2 \end{bmatrix} = \begin{bmatrix} 16 \\ -12 \end{bmatrix}, \end{aligned}$$

deci $E_2 - E_3 = (16, -12) \cdot (1, 0) = 16$. Așadar, $\alpha_2^{new, unclipped} = \alpha_2 + \frac{y_2(E_2 - E_3)}{\eta} = 4 - 16 = -12$. Deoarece $y_2 = y_3$, vom avea $L = \max(0, \alpha_2 + \alpha_3 - C) = 0$ și $H = \min(C, \alpha_2 + \alpha_3) = 12$. În consecință,

$$\alpha_2^{new, unclipped} = -12 < L = 0,$$

deci în final vom avea $\alpha_2^{new, clipped} = L = 0$ și $\alpha_3^{new, clipped} = 12 - \alpha_2^{new, clipped} = 12$. Se constată imediat că am regăsit rezultatele de la prima soluție.

Observație: Calculând valoarea funcției obiectiv $W(\alpha)$, folosind mai întâi valoare inițiale ale parametrilor α_i , apoi valorile rezultate la fiecare din cele două iterații, obținem valorile: -268.5 , -258 , -129 . Așa cum era de așteptat, aceste numere sunt în ordine crescătoare. Dacă am fi calculat [și valoarea parametrului w_0 care apare în forma primală a problemei de optimizare date, precum] și valorile variabilelor de „destindere“ ξ_i , am fi putut calcula și valorile funcției obiectiv $\frac{1}{2} \|w\|^2 + C \sum_i \xi_i$. Aceste valori trebuie să fie în ordine descrescătoare și mai mari decât valorile determinate pentru funcția $W(\alpha)$ mai sus (conform relației de *dualitate slabă* $p^* \geq d^*$; vedeti relația (290)).

24.

(SVM și C-SVM, în forma primală sau forma duală: alegerea funcției-nucleu și a valorii parametrului C)

CMU, 2010 fall, Aarti Singh, HW3, pr. 3

Figurile date mai jos indică suprafețele de decizie obținute de clasificatorul SVM pe un același set de date de antrenament fie în varianta marginii “hard” (folosind diferite funcții-nucleu), fie în varianta marginii “soft” (cu diferențe de valori pentru parametrul C). Exemplile pozitive sunt reprezentate prin simbolul \circ , iar cele negative prin simbolul \bullet . Acele care sunt încercuite desemnează vectori-suport.

Indicați care dintre cele 6 figuri a fost generată de următoarele probleme de optimizare de tip SVM sau C-SVM (atenție, sunt 6 figuri și doar 5 probleme, deci o figură nu corespunde niciunei probleme):

a. $\min \left(\frac{1}{2} w \cdot w + C \sum_{i=1}^n \xi_i \right)$

a.î. pentru $\forall i = 1, \dots, n$:

$$\xi_i \geq 0$$

$$(w \cdot x_i + w_0)y_i \geq 1 - \xi_i$$

și $C = 0.1$.

b. $\min \left(\frac{1}{2} w \cdot w + C \sum_{i=1}^n \xi_i \right)$

a.î. pentru $\forall i = 1, \dots, n$:

$$\xi_i \geq 0$$

$$(w \cdot x_i + w_0)y_i \geq 1 - \xi_i$$

și $C = 1$.

c. $\max \left(\sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \right)$

a.î. $\sum_{i=1}^n \alpha_i y_i = 0$;

$$\alpha_i \geq 0, \forall i = 1, \dots, n;$$

unde $K(u, v) = u \cdot v + (u \cdot v)^2$.

d. $\max \left(\sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \right)$

a.î. $\sum_{i=1}^n \alpha_i y_i = 0$;

$$\alpha_i \geq 0, \forall i = 1, \dots, n;$$

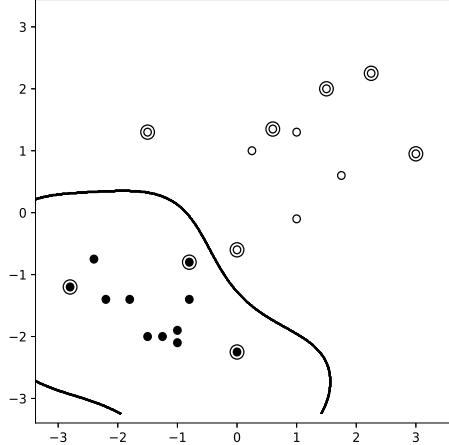
unde $K(u, v) = \exp \left(-\frac{\|u - v\|^2}{2} \right)$.

e. $\max \left(\sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \right)$

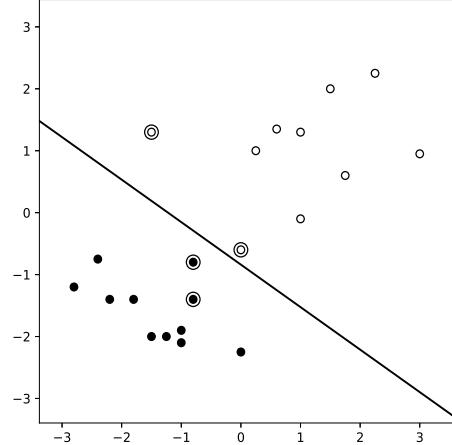
a.î. $\sum_{i=1}^n \alpha_i y_i = 0$;

$$\alpha_i \geq 0, \forall i = 1, \dots, n;$$

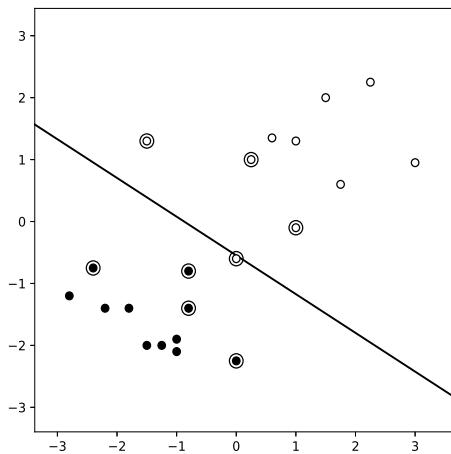
unde $K(u, v) = \exp(-\|u - v\|^2)$.



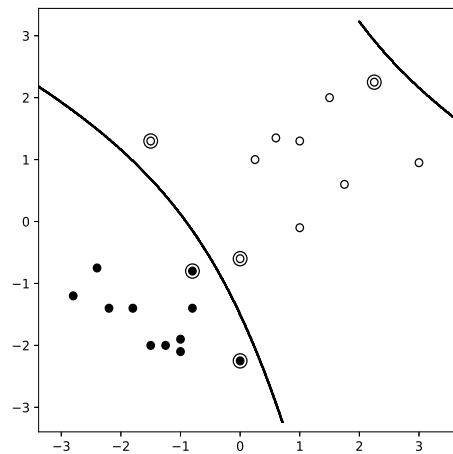
A.



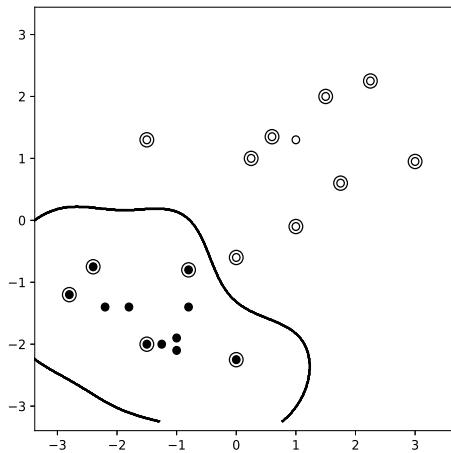
B.



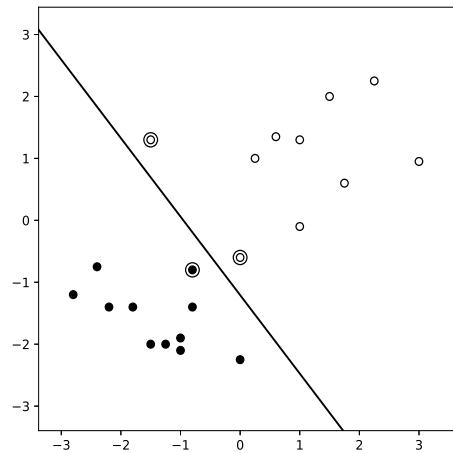
C.



D.



E.



F.

Răspuns:

Mai întâi vom „indica“ în dreptul fiecăreia dintre cele cinci SVM-uri date (a-e) care este tipul mașinii respective (adică, SVM cu margine “hard”, respectiv SVM cu margine “soft”, deci, mai precis, C-SVM), iar în primul caz ce nucleu (liniar ori neliniar, de un anumit tip) îi corespunde.

- C-SVM liniar (i.e., fără funcție nucleu), cu erori mari;
- C-SVM liniar (i.e., fără funcție nucleu), cu erori [mai] mici [decât în cazul a];
- SVM cu margine “hard” (forma duală) cu funcție nucleu pătratică;
- SVM cu margine “hard” (forma duală) cu nucleu RBF (Radial Basis Function) de parametru $\sigma^2 = 1$;
- SVM cu margine “hard” (forma duală) cu nucleu RBF de parametru $\sigma^2 = \frac{1}{2}$.

Evident, este mai ușor de tratat cazul marginii “hard”. Avem trei astfel de probleme / SVM-uri: c, d și e. Dintre acestea, prima (c) lucrează cu nucleu

pătratic, iar celelalte două cu nucleu de tip RBF. Din grafice se observă că doar separatorul din figura D are alură / formă parabolică. Așadar, graficul D corespunde problemei c. Dintre celelalte două grafice cu separatori neliniari (A și E), cel care se „mulează“ mai mult pe datele de antrenament este graficul E, deci lui îi va corespunde funcția RBF cu varianță mai mică, $\sigma^2 = \frac{1}{2}$. Așadar, mașinii e îi corespunde graficul E, iar mașinii d îi corespunde graficul A.

În sfârșit, pentru mașinile a și b — ambele folosind separator liniar, fără funcție nucleu —, cea care are valoarea parametrului C mai mică (0.1) permite erori mai multe / mari în raport cu marginea. Între cele trei grafice cu separator liniar (B, C și F), se observă că B și F nu au erori la antrenare.⁶⁷⁸ Dintre acestea două, se observă că în graficul F există cei mai puțini (doar trei) vectori-suport și, aparent, eroarea totală în raport cu marginea, $\sum_i \xi_i$ este 0. Așadar, lui F îi corespunde $C = \infty$, pentru care nu avem corespondent (apropiat) între cele cinci mașini din enunț. Relativ la cele două grafice rămase în discuție (B și C), se observă că marginea — și, la fel, suma totală a erorilor în raport cu marginea — este mai mare în cazul lui C. Așadar, graficului C îi corespunde cea mai mică dintre valorile parametrului C rămase (0.1), deci mașina a , iar graficului B îi corespunde valoarea $C = 1$, deci mașina b .

25.

(C-SVM cu funcție-nucleu:
două condiții asupra parametrului C
și respectiv asupra funcției-nucleu, suficiente ca
toate instanțele de antrenament să fie vectori-suport)

- prelucrare de Liviu Ciortuz, după
 • o MIT, 2008 fall, Tommi Jaakkola, midterm exam, pr. 1.3
 CMU, 2010 fall, Aarti Singh, HW3, pr. 3.3.b

Considerăm că antrenăm o mașină cu vectori-suport cu variabile de „destindere“ (engl., slack variables), care nu folosește variabila liberă w_0 (engl., bias variable). Se utilizează un nucleu $K(x, z)$ care are proprietatea $|K(x_i, x_j)| < 1$ dacă x_i și x_j (din setul de antrenament) sunt în relația $x_i \neq x_j$ și $|K(x_i, x_i)| \leq 1$ în caz contrar.⁶⁷⁹ Sunt în total $m \geq 2$ puncte în setul de date de antrenament, dintre care cel puțin două sunt distințe.

Arătați că alegând pentru parametrul de „destindere“ C o valoare astfel încât $C < \frac{1}{m-1}$, toate variabilele α_i din soluția problemei duale vor fi nenule. (Așadar, toate instanțele de antrenament devin vectori-suport.)

Răspuns:

Fie $i \in \{1, \dots, m\}$, fixat. Prin definiție, instanța de antrenament x_i este vector-suport dacă $\alpha_i > 0$.

Dacă prin reducere la absurd am avea un $\alpha_i = 0$, ar rezulta

$$y_i w \cdot \Phi(x_i) = y_i \left(\sum_{j=1}^m \alpha_j y_j \Phi(x_j) \right) \cdot \Phi(x_i) = y_i \left(\sum_{j=1}^m \alpha_j y_j K(x_j, x_i) \right)$$

⁶⁷⁸ Atenție: noțiunea de eroare în raport cu marginea diferă de noțiunea de eroare la clasificare. Instanța x_i constituie eroare în raport cu marginea dacă $\xi_i > 0$.

⁶⁷⁹ Nucleul RBF îndeplinește aceste condiții.

$$\begin{aligned}
&= y_i \left(\sum_{j \neq i}^m \alpha_j y_j K(x_j, x_i) \right) \leq \sum_{j \neq i}^m \alpha_j |K(x_j, x_i)| \\
&< \sum_{j \neq i}^m \alpha_j \leq (m-1)C.
\end{aligned}$$

Ultima inegalitate are loc fiindcă $\alpha_j \leq C$ pentru orice j , conform demonstrației de la problema 12.d.

Întrucât $C < 1/(m-1)$, va rezulta $y_i w \cdot \Phi(x_i) < 1$.

Însă știm că din *condițiile de complementaritate duală Karush-Kuhn-Tucker* rezultă cu necesitate că dacă $\alpha_i = 0$ atunci $y_i w \cdot \Phi(x_i) \geq 1$, pentru $i = 1, \dots, m$. (A se vedea problema 12, *observația* de la rezolvarea punctului e.)

Am obținut deci o contradicție!

Prin urmare, singura posibilitate este ca α_i să fie nul, ceea ce înseamnă că x_i este vector-suport.

26.

(Condiții suficiente pentru ca o SVM cu nucleu RBF să producă eroare la antrenare 0)

■ • ○ Stanford, 2007 fall, Andrew Ng, HW2, pr. 3

Considerăm o mașină cu vectori-suport care folosește nucleul de tip gaussian (RBF) $K(x, z) = \exp(-\|x - z\|^2 / \tau^2)$, unde am notat cu $\exp()$ funcția exponentială. Parametrul τ determină mărimea deschiderii „clopotului“ gaussian.

La punctele a și b de mai jos vă vom ghida pas cu pas ca să demonstrați următoarea proprietate: în ipoteza că setul de date de antrenament este consistent etichetat, se poate fixa o valoare a parametrului τ astfel încât eroarea la antrenare produsă de această SVM să fie zero.⁶⁸⁰ La punctul c se va arăta că această proprietate nu este valabilă și în cazul folosirii clasificatorului C-SVM.

a. Presupunem că setul de date de antrenament $\{(x_1, y_1), \dots, (x_m, y_m)\}$ este alcătuit din puncte care sunt separate unele de altele de o distanță de cel puțin ε , adică $\|x_j - x_i\| \geq \varepsilon$ pentru orice $i \neq j$.⁶⁸¹

Vă readucem aminte că funcția de decizie învățată de către SVM cu funcție-nucleu K se poate scrie sub forma:⁶⁸²

$$f(x) = \sum_{i=1}^m \alpha_i y_i K(x_i, x) + w_0, \quad (312)$$

⁶⁸⁰ Problema 29.A arată că separabilitatea aceasta se poate obține pentru orice valoare a parametrului nucleului RBF, dacă în problema SVM se impune condiția suplimentară ca toate instanțele de antrenament să fie vectori-suport.

⁶⁸¹ Evident, dacă $x_i \neq x_j$ pentru orice $i \neq j$ și, bineînțeles, dacă m este finit, atunci putem lua $\varepsilon = \min_{i \neq j} \|x_j - x_i\|$.

⁶⁸² Vedeți relația (297) de la problema 9.

Observație importantă: Această formulă presupunea *acolo* că α_i sunt soluțiile problemei SVM în forma duală. Însă *aici nu* lucrăm cu această presupozitie. Practic, aici vom alege α_i , w_0 și τ astfel încât să rezulte că f (de această formă) determină *separabilitate liniară* în spațiul de „trăsături“ în care sunt $\Phi(x_1), \dots, \Phi(x_m)$, deci și *separabilitate* (simplă, deci în general neliniară) în spațiul original (în care sunt instanțele x_1, \dots, x_m). Ulterior (la punctul b), folosind același f , vom arăta că problema SVM admite cel puțin o soluție strict „fezabilă“ — deci, conform condiției lui Slater și o soluție optimă — pentru forma primală (și, de fapt, și pentru cea duală, dar asta pur și simplu nu are relevanță aici) și, în consecință, soluția optimă va satisface și ea această proprietate de separabilitate, care ne asigură că eroarea la antrenare este 0.

unde $\alpha_1, \dots, \alpha_m \in \mathbb{R}_+$, w_0 este termenul liber / bias-ul din forma primală a problemei de optimizare [C]-SVM, iar $K(x_i, x) \stackrel{\text{def.}}{=} \Phi(x_i) \cdot \Phi(x)$, unde Φ este „maparea“ corespunzătoare funcției-nucleu K .

Găsiți valori pentru $\alpha_1, \dots, \alpha_m$, w_0 , precum și pentru parametrul gaussian τ , astfel încât toate punctele x_i să fie clasificate corect de către clasificatorul $\text{sign}(f(x))$.

Sugestii:

1. Verificați faptul că, lucrând cu $y_i \in \{-1, +1\}$, predicția făcută pentru x_i de către $\text{sign}(f(x))$ va fi corectă dacă $|f(x_i) - y_i| < 1$. Altfel spus, verificați că are loc implicația $|f(x_i) - y_i| < 1 \Rightarrow y_i f(x_i) > 0$.⁶⁸³
2. Fixând $\alpha_i = 1$ pentru $i = 1, \dots, m$ și $w_0 = 0$, găsiți o valoare a lui τ pentru care inegalitatea $|f(x_i) - y_i| < 1$ să fie satisfăcută pentru $i = 1, \dots, m$.
- b. Presupunem că rulăm o SVM fără variabile de „destindere“ (engl., slack variables), folosind pentru parametrul τ valoarea pe care ați găsit-o la punctul precedent. Va obține oare acest clasificator (în mod necesar) eroare de antrenare zero? De ce da, sau de ce nu?
- c. Presupunem că antrenăm un C-SVM (adică o SVM cu variabile de „destindere“) pe datele specificate mai sus, folosind pentru parametrul τ valoarea pe care ați ales-o la punctul a, iar pentru parametrul C o valoare fixată în mod arbitrar, dar pe care nu o cunoaștem dinainte. Va obține oare acest clasificator (în mod necesar) eroare de antrenare zero? De ce da, sau de ce nu?

Răspuns:

- a. Sunt imediate următoarele echivalențe:

$$|f(x_i) - y_i| < 1 \Leftrightarrow -1 < f(x_i) - y_i < 1 \Leftrightarrow -1 + y_i < f(x_i) < 1 + y_i.$$

Pentru $y_i = -1$, partea dreaptă a ultimei inegalități duble de mai sus devine $f(x_i) < 0$. Pentru $y_i = 1$, partea stângă a aceleiași inegalități duble devine $f(x_i) > 0$. Așadar, dacă inegalitatea $|f(x_i) - y_i| < 1$ este adevărată, atunci instanța x_i este corect clasificată de către funcția $\text{sign}(f(x))$.

Conform sugestiei din enunț, vom considera $\alpha_i = 1$ pentru $i = 1, \dots, m$ și $w_0 = 0$. Pentru un exemplu de antrenament oarecare (x_i, y_i) , vom avea:

$$\begin{aligned} |f(x_i) - y_i| &= \left| \sum_{j=1}^m y_j K(x_j, x_i) - y_i \right| = \left| \sum_{j=1}^m y_j \exp(-\|x_j - x_i\|^2 / \tau^2) - y_i \right| \\ &= \left| y_i + \sum_{j \neq i} y_j \exp(-\|x_j - x_i\|^2 / \tau^2) - y_i \right| \\ &= \left| \sum_{j \neq i} y_j \exp(-\|x_j - x_i\|^2 / \tau^2) \right| \\ &\leq \sum_{j \neq i} |y_j \exp(-\|x_j - x_i\|^2 / \tau^2)| = \sum_{j \neq i} |y_j| \exp(-\|x_j - x_i\|^2 / \tau^2) \\ &= \sum_{j \neq i} \exp(-\|x_j - x_i\|^2 / \tau^2) \end{aligned}$$

⁶⁸³Proprietatea / implicația aceasta este adevărată nu doar în cazul funcțiilor de forma specificată în relația (312), ci pentru orice funcție $f : \mathbb{R}^d \rightarrow \mathbb{R}$, unde \mathbb{R}^d este spațiul din care au fost selectate instanțele x_1, \dots, x_m .

$$\leq \sum_{j \neq i} \exp(-\varepsilon^2/\tau^2) = (m-1) \exp(-\varepsilon^2/\tau^2).$$

Prima dintre inegalitățile de mai sus este datorată aplicării repetate a inegalității triunghiului ($|a+b| \leq |a| + |b|$), iar a doua inegalitate decurge din presupunerea că $\|x_j - x_i\| \geq \varepsilon$ pentru orice $i \neq j$. Așadar, pentru a avea $|f(x_i) - y_i| < 1$ pentru $i = 1, \dots, m$ este suficient să-l alegem pe τ astfel încât

$$(m-1) \exp(-\varepsilon^2/\tau^2) < 1,$$

sau, echivalent,⁶⁸⁴

$$\tau < \frac{\varepsilon}{\sqrt{\ln(m-1)}}.$$

De exemplu, putem lua $\tau = \varepsilon/\sqrt{\ln m}$.

Rezumând, am arătat până acum că există o instanțiere pentru variabilele duale ($\alpha_i = 1$) și pentru variabila primală w_0 (și anume, $w_0 = 0$), pentru care funcția $f(x) = \sum_{i=1}^m \alpha_i y_i K(x_i, x) + w_0 = \left(\sum_{i=1}^m \alpha_i y_i \Phi(x_i) \right) \cdot \Phi(x) + w_0$ separă perfect exemplele de antrenament date. Punctele b și c de mai jos vor analiza dacă soluțiile optime produse de SVM și respectiv C-SVM pe același set de antrenament vor avea și ele această proprietate. Vă reamintim că soluțiile optime pentru problema de optimizare [C]-SVM există, atât pentru forma primală cât și pentru forma duală — și ele sunt în relația $\bar{w} = \sum_{i=1}^m \bar{\alpha}_i y_i K(x_i, x)$ — dacă, spre exemplu, este satisfăcută condiția lui Slater (a se vedea *Comentariul* de la pr. 9).

b. Datorită faptului că $f(x) = \sum_{i=1}^m \alpha_i y_i K(x_i, x) + w_0 = \left(\sum_{i=1}^m \alpha_i y_i \Phi(x_i) \right) \cdot \Phi(x) + w_0$, răționamentul prin care am făcut alegerea valorii parametrului τ de la punctul a este în sine o demonstrație a faptului că mulțimea $\{\Phi(x_i)\}_{i=1}^m$, unde Φ este maparea corespunzătoare nucleului RBF este *liniar separabilă*.

Acum vom arăta că putem pune în corespondență funcția f din proprietatea de separabilitate liniară obținută în spațiul de „trăsături“ (în care sunt $\Phi(x_1), \dots, \Phi(x_m)$) cu o anumită pereche w' , w'_0 care satisface condiția lui Slater în spațiul de trăsături (în care sunt $\Phi(x_1), \dots, \Phi(x_m)$).

Condiția lui Slater, relativă la problema de optimizare SVM este următoarea: există o soluție strict „fezabilă“, adică o asignare pentru w și w_0 , astfel încât restricțiile din problema primală SVM sunt satisfăcute cu inegalitate strictă: $y_i(w \cdot \Phi(x_i) + w_0) > 1$ pentru $i = 1, \dots, m$.

Fie $i \in \{1, \dots, m\}$, fixat. Luând $\alpha_1 = 1, \dots, \alpha_m = 1, w_0 = 0$, $f(x) = \sum_{j=1}^m \alpha_j y_j K(x_j, x) + w_0$ și τ ca la punctul a, vom avea $|f(x_i) - y_i| < 1$, deci $y_i f(x_i) > 0$. Notăm $w = \sum_{j=1}^m \alpha_j y_j \Phi(x_j)$. În consecință,

$$\begin{aligned} y_i(w \cdot \Phi(x_i) + w_0) &= y_i w \cdot \Phi(x_i) = y_i \sum_{j=1}^m \alpha_j y_j \Phi(x_j) \cdot \Phi(x_i) = y_i \sum_{j=1}^m \alpha_j y_j K(x_j, x_i) \\ &= y_i f(x_i) > 0. \end{aligned}$$

Așadar,

$$y_i(w \cdot \Phi(x_i) + w_0) = y_i \sum_{j=1}^m \alpha_j y_j K(x_j, x_i) > 0 \text{ pentru } i = 1, \dots, m.$$

⁶⁸⁴Presupunând $m > 1$.

Evident, putem multiplica toți α_i cu o constantă pozitivă astfel încât relația precedentă să devină

$$y_i \sum_{j=1}^m \alpha_j y_j K(x_j, x_i) > 1 \text{ pentru } i = 1, \dots, m.$$

Așadar, am găsit o soluție strict „fezabilă“ (w', w'_0) pentru problema noastră de optimizare.

Prin urmare, condiția lui Slater este îndeplinită. În concluzie, soluția optimă (\bar{w}, \bar{w}_0) a acestei probleme va fi într-adevăr găsită de către SVM cu nucleu RBF, în vreme ce separabilitatea liniară a mulțimii $\{\Phi(x_i)\}_{i=1}^m$ din „spațiul de trăsături“ garantează că această soluție optimă produce eroare nulă la antrenare.

c. Clasificatorul C-SVM cu nucleu RBF nu va obține în mod neapărat eroare nulă la antrenare pe setul de date considerat, chiar dacă asignăm parametrului τ valoarea găsită la punctul a .

[În general, pentru problema de optimizare convexă C-SVM soluția optimă obținută $(\bar{w}, \bar{w}_0, \bar{\xi})$ pentru o valoare fixată a parametrului de „destindere“ C nu produce în mod necesar eroare la antrenare 0, chiar dacă datele de antrenament sunt liniar separabile. Se poate să existe un alt triplet (w', w'_0, ξ') , pentru care eroarea la antrenare rezultată să fie 0, dar pentru care $\frac{1}{2} \|w'\|^2 + C \sum_i \xi'_i > \frac{1}{2} \|\bar{w}\|^2 + C \sum_i \bar{\xi}_i$.]

De exemplu, putem considera cazul extrem în care $C = 0$. În acest caz, funcția obiectiv este $\frac{1}{2} \|w\|^2$ și, evident, $w = 0$ este soluție [optimă] a problemei de optimizare C-SVM, indiferent de alegerea valorii parametrului τ . Însă în acest caz eroarea la antrenare este 0 doar dacă toate instanțele au etichetele de același semn, și anume semnul lui w_0 .

27. (SVM și C-SVM, o proprietate: Adevărat sau Fals?)

CMU, 2017 fall, Nina Balcan, midterm, pr. 1.2.bc

Observație importantă: În contextul formulărilor legate de [C-]SVM din această problemă, nu vom folosi termen liber (engl., bias term).

Fie $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$ o mulțime de m instanțe care sunt separabile liniar [cu ajutorul unui separator care trece] prin originea sistemului de coordinate din \mathbb{R}^d . Considerăm de asemenea mulțimea S' , obținută din mulțimea S astfel: $S' = \{(cx_1, y_1), \dots, (cx_m, y_m)\}$, unde $c > 0$ este o constantă.⁶⁸⁵

Care dintre afirmațiile următoare sunt adevărate?

- a. Atunci când SVM (se subîntâlege, cu margine “hard”) ia ca input S și respectiv S' , se obține același separator decizional, modulo un factor constant. Adică, dacă \bar{w} și \bar{w}' sunt soluțiile produse de către SVM pentru inputul S și respectiv S' , atunci $\bar{w} = c_1 \bar{w}'$, unde c_1 este o anumită constantă.

⁶⁸⁵LC: În textul de la CMU apărea $c > 1$.

b. Atunci când C-SVM (adică, SVM cu margine “soft”) ia ca input S și respectiv S' , se obține același separator decizional, modulo un factor constant.

Răspuns:

a. Fie \bar{w} soluția [optimă a] problemei

$$\min_w \frac{1}{2} \|w\|^2 \quad (\text{P})$$

a. i. $y_i(w \cdot x_i) \geq 1$, pentru $i = 1, \dots, m$,

și \bar{w}' soluția [optimă a] problemei

$$\min_{w'} \frac{1}{2} \|w'\|^2 \quad (\text{P}')$$

a. i. $y_i(w' \cdot cx_i) \geq 1$, pentru $i = 1, \dots, m$.

Subliniem (din nou!) faptul că în formularea problemelor (P) și (P') — contrar formulării generale a problemei de optimizare SVM de la ex. 9 — nu am folosit termen liber.

Întrucât \bar{w}' este soluție a problemei (P'), rezultă că $y_i \bar{w}' \cdot (cx_i) \geq 1 \Leftrightarrow y_i(c \bar{w}') \cdot x_i \geq 1$ pentru $i = 1, \dots, m$. Prin urmare, $c \bar{w}'$ satisfac restricțiile problemei (P), pentru care soluția optimă este \bar{w} , ceea ce implică mai departe faptul că

$$\frac{1}{2} \|\bar{w}\|^2 \leq \frac{1}{2} \|c \bar{w}'\|^2 \Leftrightarrow \|\bar{w}\|^2 \leq \|c \bar{w}'\|^2 \Leftrightarrow \|\bar{w}\|^2 \leq c^2 \|\bar{w}'\|^2. \quad (313)$$

Invers, \bar{w} fiind soluție a problemei (P), implică $y_i \bar{w} \cdot x_i \geq 1 \Leftrightarrow y_i \left(\frac{1}{c} \bar{w} \right) \cdot (cx_i) \geq 1$ pentru $i = 1, \dots, m$. Prin urmare, $\frac{1}{c} \bar{w}$ satisfac restricțiile problemei (P'), pentru care soluția optimă este \bar{w}' , ceea ce implică mai departe faptul că

$$\frac{1}{2} \left\| \frac{1}{c} \bar{w} \right\|^2 \geq \frac{1}{2} \|\bar{w}'\|^2 \Leftrightarrow \frac{1}{c^2} \|\bar{w}\|^2 \geq \|\bar{w}'\|^2 \Leftrightarrow \|\bar{w}\|^2 \geq c^2 \|\bar{w}'\|^2. \quad (314)$$

Din relațiile (313) și (314) rezultă că $\|\bar{w}\|^2 = c^2 \|\bar{w}'\|^2$, deci $\|\bar{w}\| = c \|\bar{w}'\|$. În continuare vom arăta că $\bar{w} = c \bar{w}'$ (o egalitate mai „tare“ decât egalitatea pe care tocmai am demonstrat-o).

Am precizat mai sus că din faptul că \bar{w}' este soluție a problemei (P') rezultă că $c \bar{w}'$ satisfac restricțiile problemei (P), despre a cărei funcție obiectiv știm acum că are valoarea optimă $\frac{1}{2} \|\bar{w}\|^2 = \frac{1}{2} c^2 \|\bar{w}'\|^2 = \frac{1}{2} \|c \bar{w}'\|^2$. Așadar, $c \bar{w}'$ este soluție optimă a problemei (P). Înținând cont de faptul că soluția problemei (P) este unică⁶⁸⁶ — ea existând în cazul în care setul de date de antrenament este liniar separabil și nedegenerat⁶⁸⁷ —, rezultă că are loc egalitatea $c \bar{w}' = \bar{w}$. Așadar, afirmația din enunț este *adevărată*.

Observație: Separatorul ca atare, este unul și același: $\bar{w} \cdot x = 0 \Leftrightarrow c \bar{w}' \cdot x = 0 \Leftrightarrow \bar{w}' \cdot (cx) = 0$.

⁶⁸⁶Motivația ține de faptul că funcția obiectiv a problemei (P) este strict convexă. Vedeti teorema 1 din articolul *Uniqueness of the SVM solution*, de Christopher Burges și David Crisp, 1998.

⁶⁸⁷Aceasta se traduce prin: $\exists w$ a. i. $y_i w \cdot x_i > 0$ (în condițiile acestei probleme) pentru $i = 1, \dots, m$ și respectiv $\exists i, j$ a. i. $y_i = 1$ și $y_j = -1$.

b. Răspunsul în acest caz este *negativ*, din cauza faptului că separatorii depend de valoarea parametrului C . Vom ilustra acest fapt folosind un *exemplu simplu*.

Vom considera setul de date de antrenament din enunțul problemei 17, și anume $S = \{(x_1 = -1, y_1 = -1), (x_2 = 1, y_2 = 1)\}$; el este separabil prin originea sistemului de coordonate.

Observație importantă: Se poate arăta (făcând un raționament similar cu cel de la problema 12) că forma duală a problemei C-SVM în cazul în care nu se folosește termen liber w_0 este similară cu forma duală (D') de la problema 12.d, cu singura diferență că acum nu vom mai avea restricția $\sum_i \alpha_i y_i = 0$.

Procedând într-o manieră similară cu cea din prima parte a rezolvării problemei 17, se constată că

- pentru S , avem $L_D(\alpha) = \alpha_1 + \alpha_2 - \frac{1}{2}(\alpha_1 + \alpha_2)^2 = s - \frac{1}{2}s^2 = \frac{1}{2}s(2-s)$,⁶⁸⁸ unde $s \stackrel{\text{not.}}{=} \alpha_1 + \alpha_2$. Această funcție de gradul al doilea își atinge maximul pentru $s = 1$. Așadar, $\max_{0 \leq \alpha_i \leq C} L_D(\alpha)$ se obține pentru $s = 1$ dacă $C \geq 1/2$, și respectiv pentru $s = C$ dacă $C \in (0, 1/2)$;
- pentru $S' = \{(x_1 = -2, y_1 = -1), (x_2 = 2, y_2 = 1)\}$, avem $L_D(\alpha') = \alpha'_1 + \alpha'_2 - \frac{1}{2}4(\alpha'_1 + \alpha'_2)^2 = s' - 2s'^2 = s'(1 - 2s')$, unde $s' \stackrel{\text{not.}}{=} \alpha'_1 + \alpha'_2$. Această funcție de gradul al doilea își atinge maximul pentru $s' = 1/4$. Așadar, $\max_{0 \leq \alpha'_i \leq C} L_D(\alpha')$ se obține pentru $s' = 1/4$ dacă $C \geq 1/8$ (și respectiv pentru $s' = C$ dacă $C \in (0, 1/8)$).

Notând cu $\bar{\alpha}_i$, $i \in \{1, 2\}$ soluțiile optime pentru problema duală în cazul S și cu $\bar{\alpha}'_i$, $i \in \{1, 2\}$ soluțiile optime pentru problema duală în cazul S' , urmează că în funcție de valorile parametrului C avem următoarele corespondențe între \bar{w} și \bar{w}' :

- $C > 1/2$: pentru S avem $\bar{\alpha}_1 + \bar{\alpha}_2 = 1 \Rightarrow \bar{w} = \bar{\alpha}_1 y_1 x_1 + \bar{\alpha}_2 y_2 x_2 = \bar{\alpha}_1 + \bar{\alpha}_2 = 1$, iar pentru S' avem $\bar{\alpha}'_1 + \bar{\alpha}'_2 = 1/4 \Rightarrow \bar{w}' = \bar{\alpha}'_1 y_1 x_1 + \bar{\alpha}'_2 y_2 x_2 = 2(\bar{\alpha}'_1 + \bar{\alpha}'_2) = \frac{1}{2} = \frac{1}{2}\bar{w}$;
- $C = 1/2$: pentru S avem $\bar{\alpha}_1 = \bar{\alpha}_2 = 1/2 \Rightarrow \bar{w} = \bar{\alpha}_1 + \bar{\alpha}_2 = 1$, iar pentru S' avem $\bar{\alpha}'_1 + \bar{\alpha}'_2 = 1/4 \Rightarrow \bar{w}' = 2(\bar{\alpha}'_1 + \bar{\alpha}'_2) = 1/2 = \frac{1}{2}\bar{w}$;
- $C \in (1/8, 1/2)$: pentru S avem $\bar{\alpha}_1 + \bar{\alpha}_2 = C \Rightarrow \bar{w} = \bar{\alpha}_1 + \bar{\alpha}_2 = C$, iar pentru S' avem $\bar{\alpha}'_1 + \bar{\alpha}'_2 = 1/4 \Rightarrow \bar{w}' = 2(\bar{\alpha}'_1 + \bar{\alpha}'_2) = 1/2$;
- $C = 1/8$: pentru S avem $\bar{\alpha}_1 + \bar{\alpha}_2 = C = 1/8 \Rightarrow \bar{w} = \bar{\alpha}_1 + \bar{\alpha}_2 = 1/4$, iar pentru S' avem $\bar{\alpha}'_1 = \bar{\alpha}'_2 = 1/8 \Rightarrow \bar{w}' = 2(\bar{\alpha}'_1 + \bar{\alpha}'_2) = 1/2 = 2\bar{w}$;
- $C \in (0, 1/8)$: pentru S avem $\bar{\alpha}_1 + \bar{\alpha}_2 = C \Rightarrow \bar{w} = \bar{\alpha}_1 + \bar{\alpha}_2 = C$, iar pentru S' avem $\bar{\alpha}'_1 + \bar{\alpha}'_2 = C \Rightarrow \bar{w}' = 2(\bar{\alpha}'_1 + \bar{\alpha}'_2) = 2C = 2\bar{w}$.

În concluzie,

- nu avem o relație de tip $\bar{w}' = \frac{1}{2}\bar{w}$ (ca la punctul a) decât pentru cazurile $C \geq 1/2$;
- pentru $C \in (0, 1/8]$ avem chiar o relație „inversată“: $\bar{w}' = 2\bar{w}$;

⁶⁸⁸Vedeți expresia obținută pentru $L_D(\alpha)$ în demonstrația de la problema 12.d.

- pentru $C \in (1/8, 1/2)$ nu avem practic o legătură de genul celei sugerate în enunț între \bar{w} și \bar{w}' , însăcumă $\bar{w} = C$ și $\bar{w}' = 1/2$.

28.

(SVM și C-SVM: Adevărat ori Fals?)

CMU, 2003 fall, T. Mitchell, A. Moore, final exam, pr. 7.c

- a. Mașinile cu vectori-suport (ne referim în speță la SVM și C-SVM) identifică întotdeauna optimul global pentru funcția obiectiv din problema de optimizare asociată.

CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, final exam, pr. 3.b

- b. La clasificare cu ajutorul SVM, atunci când transformăm instanțele de antrenament folosind mapări (Φ) ce corespund unor funcții-nucleu polinomiale (K) având gradul $1, 2, 3, \dots$, ne așteptăm ca vectorii-suport să rămână în general aceiași.

Stanford, 2007 fall, Andrew Ng, practice midterm exam, pr. 6.f

- c. Presupunem că lucrăm cu o mașină cu vectori-suport de normă L_1 folosind parametrul de „destindere“ $C > 0$ și că folosim un set de date de antrenament liniar separabile. Urmărind să minimizeze funcția obiectiv, C-SVM-ul va asigna variabilelor de „destindere“ ξ_i valoarea 0, însăcumă datele sunt liniar separabile. În consecință, soluția obținută (w, w_0) va fi aceeași, indiferent de valoarea folosită pentru parametrul C (presupunând, bineînțeles, că această valoare este strict pozitivă). Adevărat sau fals?

Răspuns:

- a. Adevărat. Învățarea unei funcții de clasificare prin intermediul unei mașini cu vectori-suport constă în rezolvarea unei probleme de optimizare convexă, cu funcția obiectiv de ordin pătratic și restricții liniare. Algoritmii / metodele de rezolvare computațională a acestui tip de probleme permit găsirea soluției optime, cu o eroare situată sub un prag oarecare, fixat $\varepsilon > 0$.

- b. Fals. Vectorii de trăsături ($\Phi(x_i)$) corespunzători nucleelor polinomiale sunt rezultatul aplicării unor funcții neliniare asupra vectorilor de intrare (x_i). Din această cauză, vectorii-suport pentru hiperplanul de separare optimală în spațiul de trăsături pot fi foarte diferenți de la o funcție-nucleu la alta. (Pentru *exemplificare*, vedeti problema 24, graficele B, C, D, și F.)

- c. Fals. Chiar dacă datele de antrenament sunt liniar separabile, totuși poziția separatorului optimal poate fi afectată de outlier-e. Astfel, în funcție de valoarea parametrului C , mașina cu vectori-suport cu margine “soft” (C-SVM) poate decide să clasifice [chiar] în mod eronat unu sau mai multe exemple, dacă în acest fel se obține o margine mai mare. Valoarea parametrului C va influența modul în care se face acest compromis (engl., trade-off) între mărimea marginii ($1/\|w\|$) pe de o parte și suma variabilelor ξ_i pe de altă parte. (Pentru *exemplificare*, vedeti problemele 15 și 16, iar la problema 24 graficele B, C și F.)

5.1.3 Alte probleme de optimizare de tip SVM

29.

(Functia-nucleu RBF — o proprietate remarcabilă: pentru orice set de instanțe de antrenament distințe și pentru orice etichetare a acestora și, de asemenea, pentru orice valoare a parametrului funcției-nucleu RBF, problema de optimizare de tip SVM care impune ca toate instanțele de antrenament să fie corect clasificate și la distanța $1/\|w\|$ de separatorul optimal are soluție)

■ □ • ○ * MIT, 2009 fall, Tommi Jaakkola, HW2, pr. 1

A. Expresia de definiție a funcției-nucleu cu bază radială (RBF) poate fi scrisă sub forma următoare:

$$K(x, x') = \exp\left(-\frac{1}{2\sigma^2}\|x - x'\|^2\right), \quad (315)$$

unde x și x' sunt elemente din \mathbb{R}^d , iar σ este parametrul care arată cât de repede se micșorează valoarea funcției-nucleu pe măsură ce punctele x și x' se situează la distanțe din ce în ce mai mari unul față de celălalt.

În acest exercițiu ne propunem să arătăm că funcția-nucleu RBF are câteva proprietăți remarcabile. În primul rând, ea poate separa în mod perfect *orice* mulțime finită de instanțe de antrenament *distințe*,⁶⁸⁹ iar acest rezultat este valabil pentru *orice* valoare pozitivă finită a parametrului σ .⁶⁹⁰ (Trebuie să menționăm totuși că valoarea lui σ afectează calitatea generalizării / predicției pe instanțe de test.)

a. Vom începe să demonstrăm că problema de optimizare

$$\min_w \frac{1}{2}\|w\|^2 \text{ cu restricțiile } y_i w \cdot \phi(x_i) = 1 \text{ pentru } i = 1, \dots, n \quad (316)$$

admete soluție, indiferent de ce valori (± 1) ar avea etichetele y_i asignate instanțelor x_i , cu $i = 1, \dots, n$. Am notat cu $\phi(x_i)$ vectorul de trăsături (engl., feature vector) determinat de funcția de „mapare“ care corespunde funcției K (vedeți definiția dată mai sus).

Comentariu: Formularea problemei de optimizare (316) diferă de formularea problemei SVM], din două motive. Mai întâi, se observă că încercăm să găsim o soluție a acestei probleme [SVM de un tip particular] astfel încât toate instanțele etichetate date să fie vectori-suport. Evident, această proprietate nu poate fi satisfăcută pentru orice funcție-nucleu validă, însă vom păstra pentru început o generalitate mai mare decât este nevoie — la punctele a și b vom considera K funcție-nucleu oarecare și doar începând cu punctul c vom restricționa K la funcția-nucleu RBF, definită în relația (315) — și vom vedea că astfel ne va fi mai ușor să atingem obiectivul pe care ni l-am propus inițial. În al doilea rând, observăm că formularea noastră omite termenul liber (bias-ul, notat în general cu b sau cu w_0), din cauză că el nu este necesar pentru demonstrarea proprietății enunțate la acest punct (a).

⁶⁸⁹La problema 26 am demonstrat un astfel de rezultat folosind SVM și alegând în mod convenabil o anumită valoare pentru σ . Aici vom formula o problema de optimizare convexă similară, însă sensibil diferită de problema SVM.

⁶⁹⁰O altă variantă a acestui rezultat poate fi obținută folosind regresia liniară kernel-izată, cu funcție-nucleu de tip RBF. A se vedea problema 10 de la capitolul *Metode de regresie*.

Introduceți multiplicatori Lagrange pentru restricțiile problemei de optimizare date — similar cu cele de la problema SVM duală, vedeți problema 9 — și indicați forma generală a soluției, w^* . Pentru simplificarea calculelor / raționamentului, vă recomandăm / permitem să lucrați ca și cum vectorii w și $\phi(x_i)$ ar avea întotdeauna dimensiune finită (deși nu acesta este cazul funcțiilor-nucleu RBF; vedeți problema 75 de la capitolul de *Fundamente*).

Observație: Întrucât încercăm să satisfacem restricții de tip egalitate, multiplicatorii Lagrange pot lua orice valori reale. Așadar, spre deosebire de problema de optimizare SVM standard, aici multiplicatorii Lagrange nu [mai] sunt constrânsi să ia valori pozitive.

Vă cerem să exprimați w^* , soluția problemei de optimizare (316), în funcție de acești multiplicatori Lagrange. (Aceasta n-ar trebui să implice calcule prea elaborate.)

b. Introduceți soluția w^* pe care ați obținut-o la punctul precedent în *restricțiile* de clasificare (corespunzătoare „marginilor“) din problema de optimizare în formă primală (316) și exprimați rezultatul sub formă unei combinații liniare de aplicări ale funcției-nucleu K .

c. Arătați în manieră succintă cum anume se poate folosi *teorema lui Michelli* — redată mai jos — pentru a demonstra că pentru orice funcție-nucleu RBF K , matricea pătratică definită prin $K_{ij} = K(x_i, x_j)$ conform relației (315) pentru i și $j \in \{1, \dots, n\}$ este inversabilă.

Teoremă (Michelli, 1986): Dacă $\rho : [0, \infty) \rightarrow \mathbb{R}$ este o funcție monotonă pe intervalul de definiție, atunci pentru orice set de puncte distincte x_i dintr-un spațiu \mathbb{R}^l , cu $i = 1, \dots, n$, matricea pătratică — la care ne vom referi, prin abuz de notație, tot cu ρ — ale cărei elemente sunt $\rho_{ij} \stackrel{\text{def.}}{=} \rho(\|x_i - x_j\|)$, este inversabilă.

d. Coroborând rezultatele obținute la punctele precedente, demonstrați că într-adevăr se poate găsi o soluție a problemei de optimizare date în formă primală (316) [și, în consecință, toate instanțele x_i vor fi vectori-suport].

B. Desigur, faptul că putem separa, în principiu, orice multime de exemple de antrenament nu implică în mod neapărat că un clasificator oarecare antrenat pe aceste date se comportă bine pe date de test. (Din contra!) Așadar, cum se justifică faptul că folosim [pentru clasificare, în particular cu SVM] nucleu RBF? Răspunsul ține de „marginea“ maximă [LC: de separare] pe care o putem atinge atunci când variem parametrul σ .

Observație: Variind valoarea lui σ , efectul asupra marginii [LC: geometrice] nu este pur și simplu [dat de] scalarea vectorilor de trăsături $\phi(x_i)$. Într-adevăr, este ușor să observăm că pentru [orice] nucleu RBF, adică pentru orice valoare a lui σ , avem

$$\|\phi(x)\|^2 = \phi(x) \cdot \phi(x) = K(x, x) = 1 \quad (\text{valoarea maximă posibilă pentru orice RBF}).$$

e. Să începem prin a atribui lui σ o valoare pozitivă foarte mică. Cât este — în acest caz — valoarea marginii $1/\|w^*\|$ pe care o obținem ca răspuns [la rezolvarea problemei de optimizare (316)] pentru orice set de n instanțe de antrenament distincte?

f. Alegeti un set de instanțe din spațiul unidimensional (\mathbb{R}) în aşa fel încât să arătați că marginea [LC: geometrică, văzută ca distanța până la hiperplanul $w^* \cdot x = 0$] poate fi mai mare decât cea obținută ca răspuns la punctul e (pentru $\sigma \rightarrow 0$). Puteți da orice valori pentru σ și puteți seta instanțele cum doriti, astfel încât să puneti în evidență modul în care ele pot determina „marginea” / distanța maximă dintre ele.

Răspuns:

a. Funcția lagrangeană corespunzătoare acestei probleme de optimizare este următoarea:

$$\begin{aligned} L(w, \alpha) &= \frac{1}{2} \|w\|^2 - \sum_{i=1}^n \alpha_i (y_i w \cdot \phi(x_i) - 1) \\ &= \frac{1}{2} \|w\|^2 - w \cdot \left(\sum_{i=1}^n \alpha_i y_i \phi(x_i) \right) + \sum_{i=1}^n \alpha_i. \end{aligned}$$

Așa cum s-a precizat deja, variabilele Lagrange α_i sunt *nerestrictionate*, fiindcă lucrăm doar cu constrângeri de tip egalitate.

Conform metodei lui Lagrange, forma *duală* a problemei de optimizare (316) este:

$$\max_{\alpha} \underbrace{\min_w L(w, \alpha)}_{\text{not.: } g(\alpha)}.$$

Observăm că atunci când fixăm [în mod arbitrar] valori pentru parametrii α_i , expresia $L(w, \alpha)$ va fi un polinom de gradul doi în raport cu variabilele w_i , iar coeficientul dominant al lui w_i este pozitiv, pentru orice $i = 1, \dots, n$. Prin urmare, putem obține valoarea optimă w^* folosind (de exemplu) prima derivată: $\frac{\partial L(w, \alpha)}{\partial w} = 0$. În final, procedând ca la problema 9, vom obține soluția următoare, scrisă sub formă vectorială (nu pe componente):

$$w^* = \sum_{i=1}^n \alpha_i^* y_i \phi(x_i).$$

Convenție de notație: Pentru a simplifica redactarea, vom folosi în continuare notația $w^* = \Phi[y \bullet \alpha^*]$.⁶⁹¹ Operatorul \bullet reprezintă produsul pe componente al vectorilor y și α^* , iar Φ este o matrice de dimensiune $m \times n$, în care coloana i este vectorul $\phi(x_i)$. (Desigur, $m = \infty$ în cazul funcției-nucleu RBF; vedeti problema 75 de la capitolul *Fundamente*.)

b. Se observă ușor că restricțiile $y_i w \cdot \phi(x_i) = 1$ pentru $i = 1, \dots, n$ din problema primală (316) pot fi scrise în mod echivalent (folosind notația matriceală) astfel:

$$\phi(x_i)^\top w = y_i, \text{ pentru } i = 1, \dots, n. \quad (317)$$

Aceste egalități sunt satisfăcute pentru w^* , soluția problemei de optimizare (316), a cărei expresie a fost dedusă la punctul precedent, $w^* = \Phi[y \bullet \alpha^*]$. Prin

⁶⁹¹Folosind operatorul \times pentru înmulțirea matricelor, egalitatea aceasta se scrie: $w^* = \Phi \times [y \bullet \alpha^*]$. Sugerați cititorului să folosească și mai jos această notație la fiecare apariție a vectorul-colonă $[y \bullet \alpha^*]$, dacă astfel sporește înțelegerea textului nostru.

urmăre, relațiile (317) pot fi scrise în manieră unitară (adică, simultan pentru $i = 1, \dots, n$) astfel:

$$\begin{aligned}\Phi^\top w^* &= y \Leftrightarrow \\ \Phi^\top \Phi[y \bullet \alpha^*] &= y \Leftrightarrow \\ K[y \bullet \alpha^*] &= y,\end{aligned}\tag{318}$$

unde prin K am notat matricea $\Phi^\top \Phi$, care este matricea-nucleu (și care se mai numește matricea Gram).⁶⁹²

Observație: Există acum o suprapunere la nivel de notație pentru funcția-nucleu K și matricea-nucleu corespunzătoare acestei funcții, în raport cu o mulțime de instanțe x_1, \dots, x_n . Considerăm însă că dezambiguizarea / distingerea între semnificațiile celor două notații se poate face ușor, ținând cont de context.

c. În cazul în care K este funcție-nucleu cu bază radială, $K(x_i, x_j)$ are prin definiție forma $\exp\left(-\frac{1}{2\sigma^2}\|x_i - x_j\|^2\right)$ pentru orice x_i și x_j din domeniul de definiție. Funcția $\exp\left(-\frac{1}{2\sigma^2}t^2\right)$ este monotonă (descrescătoare) pentru $t \in [0, \infty)$. Aplicând teorema lui Michelli, rezultă că pentru orice set de puncte distincte x_i , cu $i = 1, \dots, n$, matricea K ale cărei elemente sunt $K_{ij} \stackrel{\text{def.}}{=} \exp\left(-\frac{1}{2\sigma^2}\|x_i - x_j\|^2\right)$ este inversabilă.

d. Prin ipoteză, instanțele x_i , cu $i = 1, \dots, n$ sunt distincte, iar K este nucleu RBF. În consecință, conform punctului c, matricea-nucleu K este inversabilă. Vom arăta că în aceste condiții sistemul de ecuații liniare (317) are soluție. Într-adevăr, conform relației (318) de la punctul b,

$$K[y \bullet \alpha^*] = y \Rightarrow [y \bullet \alpha^*] = K^{-1}y.$$

Stim însă de la punctul a că $w^* = \Phi[y \bullet \alpha^*]$. Prin urmare, $w^* = \Phi K^{-1}y$. (Aceasta este soluția unică a sistemului liniar (317), în necunoscutele w_i .)

e. Pe măsură ce $\sigma \rightarrow 0$, matricea-nucleu K va tinde la matricea unitate I , fiindcă

$$\lim_{\sigma \rightarrow 0} \exp\left(-\frac{1}{2\sigma^2}\|x_i - x_j\|^2\right) \rightarrow 0, \text{ pentru orice } i \neq j,$$

iar

$$\exp\left(-\frac{1}{2\sigma^2}\|x_i - x_i\|^2\right) = 1.$$

Conform relației (318), $K[y \bullet \alpha^*] = y$, deci $K \rightarrow I$ implică la rândul său $y \bullet \alpha^* \rightarrow y$ și, deci, $\alpha^* \rightarrow \bar{1}$, vectorul-colonă care are toate elementele egale cu 1.

Pentru a vedea cât este „marginea“ $\frac{1}{\|w\|}$ atunci când $\sigma \rightarrow 0$, vom calcula mai întâi $\|w\|^2$, pornind de la relația $w^* = \Phi[y \bullet \alpha^*]$:

$$\begin{aligned}\|w^*\|^2 &= w^{*\top} w^* = (\Phi[y \bullet \alpha^*])^\top \Phi[y \bullet \alpha^*] = [y \bullet \alpha^*]^\top \underbrace{\Phi^\top \Phi}_{K} [y \bullet \alpha^*] \\ &= [y \bullet \alpha^*]^\top K[y \bullet \alpha^*]\end{aligned}\tag{319}$$

⁶⁹²Această matrice este pătratică, iar elementele ei sunt $\phi(x_i) \cdot \phi(x_j) = K(x_i, x_j)$, unde K desemnează funcția-nucleu de la care am pornit.

În consecință, atunci când $\sigma \rightarrow 0$, vom avea

$$\|w^*\|^2 = [y \bullet \alpha^*]^\top K[y \bullet \alpha^*] \rightarrow y^\top I y = n,$$

deci „marginea“ [LC: de separare] care se obține este $\frac{1}{\|w^*\|} = \sqrt{\frac{1}{n}}$.

f. Considerăm două instanțe distincte x și x' din \mathbb{R} , ambele cu eticheta +1. Vom nota $k = K(x, x') = \exp\left(-\frac{1}{2\sigma^2}\|x - x'\|^2\right)$. Se vede imediat că matricea-nucleu este:

$$K = \begin{bmatrix} 1 & k \\ k & 1 \end{bmatrix}$$

Rezolvând sistemul de ecuații liniare $K[\bar{1} \bullet \alpha] = \bar{1}$ implicit de relația (318), obținem soluția

$$\alpha^* = \left[\frac{1}{k+1}, \frac{1}{k+1} \right]^\top.$$

În continuare, folosind formula (319) dedusă la punctul e , obținem $\|w^*\|^2 = \alpha^{*\top} K \alpha^* = \frac{2}{k+1}$. Prin urmare, „marginea“ [LC: geometrică] este $\frac{1}{\|w^*\|} = \sqrt{\frac{k+1}{2}}$. Evident, $k > 0$, deci „marginea“ obținută aici este strict mai mare decât $\sqrt{\frac{1}{2}}$, valoarea marginii care a fost obținută în cazul $\sigma \rightarrow 0$ (vedeți punctul e).

De asemenea, atunci când apropiem cele două instanțe x și x' (sau, la fel, când $\sigma \rightarrow +\infty$), rezultă $k \rightarrow 1$ și, în consecință, obținem „marginea“ 1.

Observație:

Pornind de la relația $\frac{1}{\|w^*\|} = \frac{1}{\sqrt{n}}$ stabilită la punctul e și / sau de la relația $\frac{1}{\|w^*\|} = \sqrt{\frac{k+1}{2}}$ obținută la acest punct, ne putem pune acum întrebarea următoare: „marginea“ maximă pe care o putem obține în cazul nucleului RBF este oare [întotdeauna] 1?

Ca să răspundem la această întrebare, observăm că pentru orice instanță x_i avem $\|\phi(x_i)\|^2 = K(x_i, x_i) = 1$. În consecință, ne putem gândi la vectorii de trăsături $\phi(x_i)$ (care sunt infinit-dimensionali!) ca fiind poziționați pe sfera de rază 1 (engl., unit ball) având centrul în originea sistemului de coordonate. Rezultă că, într-adevăr, valoarea maximă posibilă pentru marginea de separare este 1.

Intuitiv, ne putem gândi astfel: pe măsură ce $\sigma \rightarrow +\infty$, „distanțele“ $K(x, x')$ vor tinde la 1, iar nucleele centrate pe instanțe distincte vor deveni indiscernabile.⁶⁹³ Prin urmare, vectorii de trăsături vor tinde să coincidă, identificându-se la limită cu un singur punct pe sfera de rază 1 și cu centrul în originea sistemului de coordonate. La limită „marginea“ [LC: geometrică] va avea mărimea 1 (ceea ce este contra-intuitiv!).

⁶⁹³ $\sigma \rightarrow +\infty \Rightarrow K(x, x') \rightarrow 1$ și $K(x, x'') \rightarrow 1$ pentru orice x și $x' \neq x''$. Deci $\phi(x) \cdot \phi(x') \approx \phi(x) \cdot \phi(x'')$ pentru orice x, x' și x'' .

30.

(Clasificare n -ară cu SVM: SVM multiclass, cu margine “hard”)

*prelucrare de Liviu Ciortuz, după
□ • MIT, 2009 fall, Tommi Jaakkola, lecture notes 5*

În practică, majoritatea problemelor de clasificare lucrează cu mai mult de două clase (de exemplu, identificarea persoanelor dintr-o imagine, prezicerea fonemelor în procesarea vorbirii, asocierea genelor la anumite procese biologice, ori prezicerea tipului de cancer pe baza analizei datelor preluate la nivel celular). În acest capitol, până aici ne-am ocupat doar de [task-uri de] clasificare binară. Totuși, metodele de tip SVM cu care ați făcut cunoștință — dar de asemenea și alte metode, precum Perceptronul —, pot fi folosite și pentru clasificare n -ară (engl., multiclass).

Una dintre modalitățile de a face clasificare n -ară, și anume metoda “one versus all”, constă în a reduce problema de clasificare multiclass la un set de task-uri de clasificare binară. În această abordare, fiecare dintre clasificatorii binari este antrenat în mod independent față de ceilalți.⁶⁹⁴ O altă modalitate — cea care ne va interesa pe noi aici — este să încercăm să formulăm o *problemă de optimizare* per ansamblu, în aşa fel încât să putem antrena [în mod direct] un clasificator n -ar.

Vom folosi drept *componente* constitutive ale clasificatorului nostru n -ar mai mulți clasificatori liniari simpli care trec prin originea sistemului de coordinate,⁶⁹⁵ adică având forma analitică $f_j(x) = w_j \cdot x$, cu $x \in \mathbb{R}^d$ și $w_j \in \mathbb{R}^d$, pentru $j = 1, \dots, K$.⁶⁹⁶ Apoi vom impune cerința ca [la clasificarea fiecărei instanțe x] să „câștige“ acel clasificator care corespunde etichetei corecte, în sensul că valoarea *funcției* asociate aceluia clasificator [pentru x] să fie cea mai mare dintre toate valorile funcțiilor asociate diversilor clasificatori liniari [pentru același x], lăsând în plus un mic „spațiu de manevră“. Așadar, problema pe care vom încerca să o rezolvăm este următoarea:

$$\begin{aligned} & \min_{w^{(1)}, \dots, w^{(K)}} \frac{1}{2} \sum_{k=1}^K \|w^{(k)}\|^2 \\ \text{a. i. } & w^{(y_i)} \cdot x_i \geq w^{(y)} \cdot x_i + 1, \quad \text{pentru } i = 1, \dots, m \\ & \text{și orice } y \in \{1, \dots, K\}, \text{ cu } y \neq y_i, \end{aligned} \tag{320}$$

unde $w^{(1)}, \dots, w^{(K)} \in \mathbb{R}^d$, $x_i \in \mathbb{R}^d$, iar $y_i \in \{1, \dots, K\}$ pentru $i = 1, \dots, m$.

Observați că în această abordare componentele clasificatorului n -ar sunt antrenate împreună. Predicția pentru o instanță oarecare x este determinată de cât de mult „susține“ fiecare componentă-clasificator etichetarea corespunzătoare: $\hat{y} = \arg \max_y \{w^{(y)} \cdot x\}$.

⁶⁹⁴Vedeți MIT, 2009 fall, Tommi Jaakkola, lecture notes 5, pag. 7–11, precum și articolul *Reducing multiclass to binary* de Erin Allwein et al, din revista Journal of Machine Learning Research, 2000.

⁶⁹⁵La problema 61, care introduce versiunea SVM multiclass cu margine “soft”, această cerință va fi relaxată, în sensul că vom permite ca acești clasificatori liniari să aibă termeni liberi (b_k , pentru $k = 1, \dots, K$) eventual diferiți între ei. Pe lângă aceasta, vom cere ca ei să se intersecteze într-un același punct (ca și în cazul de față), dar nu neapărat în originea sistemului de coordinate.

⁶⁹⁶Pentru conveniență, deși am folosit mai sus termenul de clasificare n -ară, în continuare în formalismul matematic vom desemna numărul de clase cu simbolul K . (Așadar, se poate considera $K = n$.)

Remarcați faptul că *zonele de decizie* care rezultă au o *interpretare geometrică* simplă. Putem obține fiecare dintre aceste zone ca *intersectie* a două *regiuni*, corespunzător mulțimilor de exemple pentru care o anumită etichetă este preferată în defavoarea celorlalte etichete. Aceste regiuni sunt determinate în mod simplu prin linii (în general, hiperplane) care trec prin originea sistemului de coordonate: $w^{(y)} \cdot x = w^{(y')} \cdot x$, ca în figura alăturată, unde $y, y' \in \{1, 2, 3\}$. Linile punctate arată cum anume se extind aceste hiperplane în regiunile cărora le sunt asociate alte etichete. Regulile de decizie sunt desemnate prin linii continue.

Înainte de a trece la rezolvarea acestei probleme de optimizare în forma duală, ar fi util (pentru mai târziu) să scriem problema într-o formă ușor mai generală. Pentru aceasta, vom nota

$$w = (w^{(1)}, \dots, w^{(y)}, \dots, w^{(K)})^\top, \phi(x, y) = (0, \dots, x, \dots, 0)^\top,$$

unde locația lui x din cadrul vectorului $\phi(x, y)$ garantează faptul că $w \cdot \phi(x, y) = w^{(y)} \cdot x$. În consecință, forma primală a problemei de optimizare *SVM multi-class* se poate scrie în mod echivalent (în raport cu forma anterioară (320)) astfel:

$$\min_w \frac{1}{2} \|w\|^2 \quad (321)$$

a. i. $w \cdot \phi(x_i, y_i) \geq w \cdot \phi(x_i, y) + 1_{\{y \neq y_i\}}$, pentru $i = 1, \dots, m$ și $y \in \{1, \dots, K\}$.

a. Deducreți forma duală a problemei de optimizare (321). Veți proceda similar cu modul în care s-a lucrat pentru clasificatorul binar SVM (vedeți pr. 9).

Sugestie: Introduceți multiplicatorii Lagrange $\alpha_{y,i} \geq 0$ pentru restricțiile de tip inegalitate, obțineți expresia lagrangeanului generalizat, găsiți valoarea lui w (ca funcție de multiplicatorii Lagrange $\alpha_{y,i}$) care minimizează acest lagrangean și apoi substituiți această valoare (pe care o veți nota cu $\hat{w}(\alpha)$) în expresia lagrangeanului, pentru a obține în sfârșit varianta duală.

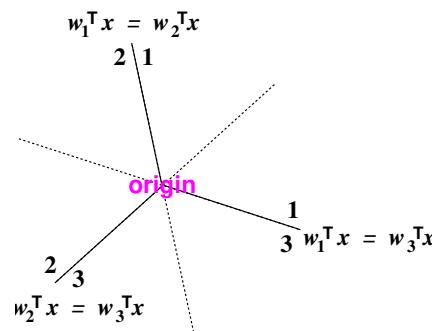
b. Folosind \hat{w} , expresia vectorului de ponderi care reprezintă soluția optimă a problemei SVM multiclass, care a fost dedusă la punctul precedent în funcție de multiplicatorii Lagrange $\alpha_{y,i}$, scrieți *regula de predicție* pentru eticheta unei instanțe noi x , ținând cont că

$$\hat{y} = \arg \max_y \{\hat{w}^{(y)} \cdot x\}.$$

Răspuns:

a. Scriem mai întâi expresia lagrangeanului generalizat:

$$L_P(w, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^m \sum_{y=1}^K \alpha_{y,i} [w \cdot \phi(x_i, y_i) - w \cdot \phi(x_i, y) - 1_{\{y \neq y_i\}}].$$



Apoi, calculăm derivata parțială a acestei funcții în raport cu vectorul de ponderi w :

$$\frac{\partial}{\partial w} L_P(w, \alpha) = w - \sum_{i=1}^m \sum_{y=1}^K \alpha_{y;i} [\phi(x_i, y_i) - \phi(x_i, y)].$$

Conform condiției de staționaritate / optimalitate Karush-Kuhn-Tucker care se referă la atingerea minimului lui L_P în raport cu w , vom egala această derivată parțială cu vectorul 0 și vom obține soluția

$$\hat{w}(\alpha) = \sum_{i=1}^m \sum_{y=1}^K \alpha_{y;i} [\phi(x_i, y_i) - \phi(x_i, y)].$$

Acum vom substitui $\hat{w}(\alpha)$ în expresia lagrangeanului de mai sus, iar după ce vom simplifica termenii asemenea, vom obține imediat următoarea expresie pentru lagrangeanul dual:

$$\begin{aligned} L_D(\hat{w}, \alpha) &= \sum_{i=1}^m \sum_{y=1}^K \alpha_{y;i} 1_{\{y \neq y_i\}} - \\ &\quad \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \sum_{y=1}^K \sum_{y'=1}^K \alpha_{y;i} \alpha_{y';j} [\phi(x_i, y_i) - \phi(x_i, y)] \cdot [\phi(x_j, y_j) - \phi(x_j, y')]. \end{aligned}$$

Observați că această expresie poate fi simplificată și mai mult, dacă ținem cont că notația introdusă mai sus pentru $\phi(x, y)$ implică $\phi(x, y) \cdot \phi(x', y') = 1_{\{y=y'\}} (x \cdot x')$. Forma duală a problemei de optimizare SVM multiclass are ca funcție obiectiv lagrangeanul L_D , iar ca restricții $\alpha_{y;i} \geq 0$ pentru $y = 1, \dots, K$ și $i = 1, \dots, m$.

b. Regula de decizie pentru clasificatorul SVM multiclass se poate scrie astfel:

$$\hat{y} = \arg \max_y \{\hat{w}^{(y)} \cdot x\} = \arg \max_y \left\{ \sum_{i=1}^m \sum_{y'=1}^K \hat{\alpha}_{y';i} [\phi(x_i, y_i) - \phi(x_i, y')] \cdot \phi(x, y) \right\},$$

unde $\hat{\alpha}_{y';i}$ sunt soluțiile problemei duale SVM multiclass. (Remarcați faptul că și în acest caz se pot face simplificări de genul celor indicate la punctul precedent.)

31.

(Problema SVM *one-class* (varianta “Max Margin”), cazul marginii “hard”)

■ Stanford, 2007 fall, Andrew Ng, practice midterm exam, pr. 4

Considerăm un set de instanțe neetichetate $\{x_1, \dots, x_m\} \subset \mathbb{R}^d$.

Un algoritm SVM de tip *one-class* caută să identifice (dacă este posibil) o direcție $w \in \mathbb{R}^d$ care separă în sens maximal datele de originea sistemului de coordinate, ca în figura de mai jos.⁶⁹⁷

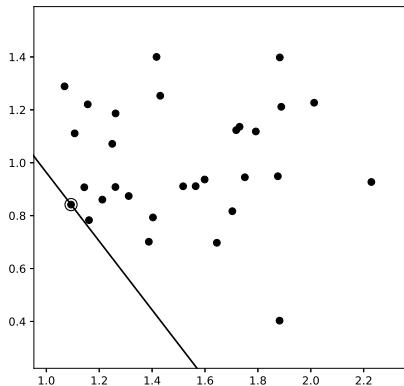
⁶⁹⁷Este de remarcat faptul că *one-class SVM* este o tehnica de învățare nesupervizată (așadar, nu de clasificare), întrucât ea nu necesită ca instanțele să fie etichetate.

Mai precis, acest algoritm rezolvă problema de optimizare (dată în formă primală):⁶⁹⁸

$$\min_w \frac{1}{2} \|w\|^2$$

a. i. $w \cdot x_i \geq 1$, pentru $i = 1, \dots, m$.

O instanță nouă (de test) va fi etichetată cu + dacă $w \cdot x \geq 1$, și cu - în caz contrar.



Observație (1): Un astfel de algoritm de tip SVM este util pentru detectia anomalilor (engl., anomaly detection). În astfel de situații, ni se dă mai întâi un set de date care se consideră „normale“. Apoi ni se cere să decidem pentru alte instanțe dacă sunt (sau nu sunt) „anomalii“ (engl., outliers).

a. Scrieți forma duală corespunzătoare problemei de optimizare *SVM one-class* (de tip *Max Margin*) de mai sus. Simplificați răspunsul cât mai mult posibil; evident, vectorul w nu trebuie să apară în rezultatul final. Verificați în prealabil dacă forma primală satisface condiția lui Slater.

b. Am putea „kerneliza“ algoritmul *SVM one-class* (de tip *Max Margin*) atât la antrenare cât și la testare? Adică: dată fiind o funcție-nucleu K , este posibil ca după maparea corespunzătoare, variabilele x să apară

- atât în expresia lagrangeanului (L_D) care reprezintă funcția obiectiv a formei duale ale problemei SVM one-class,
- cât și în funcția de decizie / clasificare pentru o instanță nouă x' , doar ca argumentele ale funcției-nucleu K ?

c. Concepți un algoritm de tip SMO⁶⁹⁹ care să rezolve problema duală obținută la punctul a. Ideea de bază a unui astfel de algoritm este ca la fiecare pas să se obțină soluția optimă pe cea mai mică dintre toate submulțimile posibile de variabile. Dați formulele analitice (engl., closed form formulas) pentru actualizarea variabilelor din această submulțime. Trebuie să justificați / explicați de ce este suficient să considerăm simultan respectivul număr de variabile la fiecare pas.

Răspuns:

a. Verificăm mai întâi faptul că problema de optimizare care a fost dată în enunț în formă primală satisface condiția lui Slater.⁷⁰⁰ Dacă există un

În altă ordine de idei, putem formula problema *one-class SVM* într-o formă *mai generală*:

$$\min_w \frac{1}{2} \|w\|^2$$

a. i. $w \cdot x_i + w_0 \geq 1$, pentru $i = 1, \dots, m$.

Însă, dacă extindem (cum se face de obicei la rețele neuronale artificiale) orice instanță x_i cu o componentă $x_{i,0} = 1$, ajungem la forma (mai simplă!) din enunț.

⁶⁹⁸La problema 64 vom da o altă variantă a problemei *one-class SVM*, definită cu ajutorul *sferei de inclusiune minimală* (engl., minimum enclosing ball, MEB). Pentru a putea distinge mai ușor cele două versiuni (una de cealaltă), vom numi varianta de aici *Max Margin*.

⁶⁹⁹Vedeți problemele 22 și 23, precum și referințele bibliografice indicate acolo.

⁷⁰⁰Pentru formalizarea acestei condiții, vedeți *Comentariul* de la problema 9 (pag. 642).

hiperplan care trece prin originea sistemului de coordonate și „lasă“ toate instanțele x_i cu $i = 1, \dots, m$ de o aceeași parte a sa, aceasta înseamnă că există $w \in \mathbb{R}^d$ astfel încât $w \cdot x_i > 0$ pentru $i = 1, \dots, m$. Întrucât m este finit, înmulțind acest w cu o anumită constantă pozitivă obținem $w \cdot x_i > 1$ pentru $i = 1, \dots, m$, deci condiția lui Slater este satisfăcută.

Lagrangeanul generalizat care corespunde problemei primale date este:

$$L_P(w, \alpha) = \frac{1}{2}w \cdot w + \sum_{i=1}^m \alpha_i(1 - w \cdot x_i)$$

cu $\alpha_i \geq 0$ pentru $i = 1, \dots, m$.

Egalând cu 0 derivata parțială a lui L_P în raport cu w , obținem $w = \sum_{i=1}^m \alpha_i x_i$. Substituind această egalitate în expresia lui L_P , vom avea:

$$\begin{aligned} L_D(\alpha) &= \frac{1}{2} \left(\sum_{i=1}^m \alpha_i x_i \right) \cdot \left(\sum_{j=1}^m \alpha_j x_j \right) + \sum_{i=1}^m \alpha_i \left(1 - \left(\sum_{j=1}^m \alpha_j x_j \right) \cdot x_i \right) \\ &= \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j x_i \cdot x_j \end{aligned}$$

Așadar, forma lui L_D este similară cu cea de la problema SVM cu margine “hard” (vedeți problema 9.c), iar din cauză că aici nu se folosește termen liber problema duală este acum mai simplă: $\max_{\alpha \geq 0} L_D(\alpha)$.

b. În ce privește kernel-izarea, este imediat că relativ la antrenare răspunsul este afirmativ, deoarece în expresia lagrangeanului L_D care a fost dedusă la punctul precedent x_i și x_j apar doar ca [perechi de] factori în produsele scalare. Concret, după „mapare“ folosind funcția Φ corespunzătoare nucleului K , vom avea:

$$\begin{aligned} L_D(\alpha) &= \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j \Phi(x_i) \cdot \Phi(x_j) \\ &= \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j K(x_i, x_j). \end{aligned}$$

Relativ la testare, răspunsul este de asemenea afirmativ: dată fiind o instanță de test x' , ea va fi clasificată în funcție de semnul expresiei

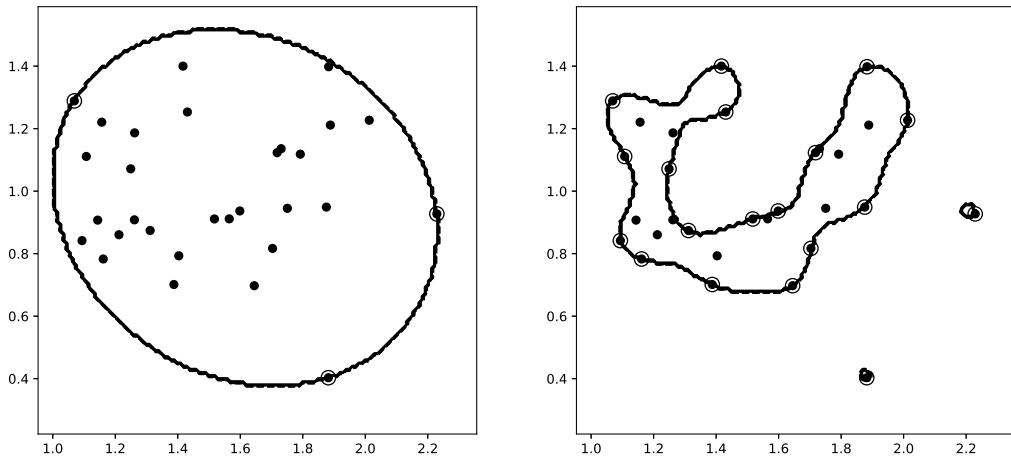
- $w \cdot x' = (\sum_{i=1}^m \alpha_i x_i) \cdot x' = \sum_{i=1}^m \alpha_i x_i \cdot x'$, în cazul în care nu se face mapare
- $w \cdot x' = \sum_{i=1}^m \alpha_i \Phi(x_i) \cdot \Phi(x') = \sum_{i=1}^m \alpha_i K(x_i, x')$, când se face mapare.

Observație (2): Imaginele de mai jos ilustrează rezultatul folosirii a două funcții-nucleu de tip RBF în conjuncție cu SVM *one-class* de tip *Max Margin* pe setul de date din enunț. Pentru rezultatul ilustrat în partea dreaptă, s-a lucrat cu o valoare [mai] mică pentru parametrul σ^2 din definiția nucleului RBF. Facem mențiunea că aici s-a folosit o formă ușor mai generală pentru problema SVM *one-class* (de tip Max Margin):⁷⁰¹

$$\min_{w, \rho} \left(\frac{1}{2} \|w\|^2 - \rho \right)$$

a. i. $w \cdot x_i \geq \rho$, pentru $i = 1, \dots, m$.

⁷⁰¹LC: Această formă este foarte utilă pentru a înțelege modul în care se definește problema de optimizare ν -SVM de la pr. 33. Vedeți și pr. 63, unde se elaborează varianta “soft margin” pentru problema *one-class* de tip Max Margin.



c. Întrucât în formularea dată în enunț pentru problema *one-class SVM* nu se folosește termenul liber (engl., “bias”) w_0 , nu avem în forma duală a problemei de optimizare o restricție de tipul $\sum_{i=1}^n y_i \alpha_i = 0$ (cum este cazul la problemele 22 și 23). Așadar, vom folosi tot metoda creșterii pe coordonate (engl., coordinate ascent) ca în algoritmul clasic SMO, însă vom alege la fiecare iterație doar (câte) o variabilă Lagrange (α_i).

Din expresia lagrangeanului dual $L_D(\alpha)$ pe care l-am calculat la punctul a , obținem funcția pe care trebuie să-o optimizăm la iteratărea curentă:

$$L(\alpha_i) = \alpha_i - \sum_{j \neq i} \alpha_i \alpha_j x_i \cdot x_j - \frac{1}{2} \alpha_i^2 x_i \cdot x_i + const.$$

Punctul de maxim este dat de soluția derivatei de ordinul întâi a acestei funcții:

$$\frac{\partial L(\alpha_i)}{\partial \alpha_i} = 0 \Leftrightarrow 1 - \sum_{j \neq i} \alpha_j x_j \cdot x_i - \alpha_i x_i \cdot x_i = 0 \Leftrightarrow \alpha_i^{new, unclipped} = \frac{1 - \sum_{j \neq i} \alpha_j x_j \cdot x_i}{x_i \cdot x_i}.$$

Tinând cont de restricția $\alpha_i \geq 0$, rezultă că noua valoare pe care o atribuim variabilei alese este $\alpha_i^{new, clipped} = \max \left\{ 0, \frac{1 - \sum_{j \neq i} \alpha_j x_j \cdot x_i}{x_i \cdot x_i} \right\}$.

Mai rămân de specificat:

- criteriul de selecție a variabilei „libere“; de preferință aceasta se va face în aşa fel încât să se obțină o creștere cât mai mare a valorii funcției obiectiv de la o iteratăie la alta;
- criteriul de oprire a algoritmului; spre exemplu, atunci când creșterea funcției obiectiv de la o iteratăie la alta devine nesemnificativă este inutil să mai executăm noi iteratăii.

32.

(O condiție suficientă pentru ca cele două tipuri de probleme de optimizare *one-class SVM*, și anume *Max Margin* și *minimum enclosing ball* (MEB), în varianta kernel-izată, să fie echivalente)

MIT, 2009 fall, Tommi Jaakkola, HW2, pr. 2.a

La exercițiul 31 am prezentat o metodă de detecție a „anomalilor“ dintr-un set de instanțe numită *one-class SVM*. Metoda respectivă se bazează pe separarea vectorilor de „trăsături“ (engl., feature vectors) față de originea sistemului de coordonate. Concret, am pornit de la un separator liniar care trece prin origine și are „magine“ (adică distanță) maximă în raport cu instanțele date. De aceea am numit această metodă problema *one-class Max Margin SVM*. Din punct de vedere matematic, varianta kernel-izată a acestei probleme a fost formulată — ca problemă de optimizare convexă cu restricții — astfel:⁷⁰²

$$\min_{w,\rho} \left(\frac{1}{2} \|w\|^2 - \rho \right) \quad (\text{Max Margin})$$

a. i. $w \cdot \phi(x_i) \geq \rho$, pentru $i = 1, \dots, m$.

unde x_1, \dots, x_m sunt instanțe de antrenament, iar ϕ este o așa-numită funcție de „mapare“ a atributelor.

O altă modalitate prin care putem detecta „anomalii“ este să identificăm o hiper-sferă care să includă în sens minimal (engl., minimum enclosing ball, MEB) instanțele date sau, dacă folosim o funcție de „mapare“, imaginile instanțelor noastre în spațiul de „trăsături“ corespunzător. Matematic, putem formula această metodă în maniera următoare: date fiind instanțele x_1, \dots, x_m și o „mapare“ a trăsăturilor $\phi(x)$, considerăm problema de optimizare

$$\min_{R,w} R^2 \quad (\text{MEB})$$

a. i. $\|w - \phi(x_i)\|^2 \leq R^2$, pentru $i = 1, \dots, m$.

Vă cerem să demonstrați că aceste două probleme devin identice în ipoteza că

$$\|\phi(x_i)\| = c \text{ pentru } i = 1, \dots, m,$$

unde c este o constantă oarecare (reală, pozitivă).⁷⁰³ Cu alte cuvinte, soluția optimă \hat{w} obținută de una (oricare) dintre cele două probleme este soluție optimă și pentru cealaltă problemă.

Sugestie: Deducreți forma duală pentru fiecare dintre cele două probleme de optimizare de mai sus (*Max Margin* și *MEB*). Apoi comparați cele două forme duale obținute, ținând cont de ipoteza $\|\phi(x_i)\| = c$ pentru c fixat și $i = 1, \dots, m$.

Observație: Ca o consecință directă a proprietății demonstrează în această problemă, putem afirma că graficele care au fost obținute la rezolvarea problemei 31.b (Max Margin) sunt identice cu cele care corespund soluțiilor problemei de optimizare MEB pe datele respective.

⁷⁰²Vedeți *Observația* (2) de la pr. 31. Veți constata că varianta dată inițial în enunțul acelei probleme este mai simplă; este suficient de simplă pentru introducerea modelului *Max Margin*, însă este prea simplă pentru a o putea folosi în formularea proprietății pe care o vom da aici.

⁷⁰³*Observație importantă:* Această condiție este satisfăcută în cazul nucleului RBF, fiindcă $\phi(x)^2 = K(x, x) = \exp\left(-\frac{\|x - x\|^2}{2\sigma^2}\right) = 1$, pentru orice x și orice σ .

Răspuns:

Concret, strategia de rezolvare este următoarea: Se va constata ușor că atât în cazul problemei *Max Margin* (pentru detecția anomalilor) cât și în cazul problemei sferei de incluziune minimală (MEB), relația dintre soluția formei primale (\hat{w}) și soluția formei duale ($\hat{\alpha}$) este aceeași: $\hat{w} = \sum_{i=1}^m \hat{\alpha}_i \phi(x_i)$. Prin urmare, dacă presupunând adevărată relația $\|\phi(x_i)\| = c$ pentru $i = 1, \dots, m$ va rezulta că pentru ambele probleme de optimizare forma duală are exact aceeași soluție $\hat{\alpha}$, atunci este imediat că și soluțiile formelor primale ale celor două probleme coincid.⁷⁰⁴

Făcând calculele în maniera în care deja ne-am obișnuit, vom obține pentru problema *Max Margin* următoarea formă duală:⁷⁰⁵

$$\max_{\alpha} \left(- \sum_{i,j=1}^m \alpha_i \alpha_j K(x_i, x_j) \right)$$

a. i. $\alpha_i \geq 0$ pentru $i = 1, \dots, m$ și $\sum_{i=1}^m \alpha_i = 1$.

⁷⁰⁴ Ca și în cazul problemei *Max Margin* — pentru varianta ei [mai] simplă, am văzut deja justificarea la ex. 31.a —, se constată ușor că pentru problema MEB este satisfăcută condiția lui Slater.

⁷⁰⁵ Pe scurt, se procedează astfel:

$$\begin{aligned}
 L_P(w, \rho, \alpha) &= \frac{1}{2} w^2 - \rho - \sum_{i=1}^m \alpha_i (w \cdot \phi(x_i) - \rho) \\
 \frac{\partial}{\partial w} L_P(w, \rho, \alpha) = 0 &\Leftrightarrow w - \sum_{i=1}^m \alpha_i \phi(x_i) = 0 \Leftrightarrow w = \sum_{i=1}^m \alpha_i \phi(x_i) \\
 \frac{\partial}{\partial \rho} L_P(w, \rho, \alpha) = 0 &\Leftrightarrow -1 + \sum_{i=1}^m \alpha_i = 0 \Leftrightarrow \sum_{i=1}^m \alpha_i = 1 \\
 L_D(\alpha) &= \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j \phi(x_i) \cdot \phi(x_j) - \rho - \sum_{i=1}^m \alpha_i \left(\left(\sum_{j=1}^m \alpha_j \phi(x_j) \right) \cdot \phi(x_i) - \rho \right) = \\
 &= -\frac{1}{2} \underbrace{\sum_{i,j=1}^m \alpha_i \alpha_j \phi(x_i) \cdot \phi(x_j)}_1 - \rho + \rho \sum_{i=1}^m \alpha_i = -\frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j \underbrace{\phi(x_i) \cdot \phi(x_j)}_{K(x_i, x_j)} = -\frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j K(x_i, x_j).
 \end{aligned}$$

În mod similar, pentru problema de optimizare MEB, forma duală va fi:⁷⁰⁶

$$\max_{\alpha} \left(\sum_{i=1}^m \alpha_i K(x_i, x_i) - \sum_{i,j=1}^m \alpha_i \alpha_j K(x_i, x_j) \right)$$

a. i. $\alpha_i \geq 0$ pentru $i = 1, \dots, m$ și $\sum_{i=1}^m \alpha_i = 1$.

Întrucât $K(x_i, x_i) = \phi(x_i)^2 = \|\phi(x_i)\|^2 = c^2$ și $\sum_{i=1}^m \alpha_i = 1$, rezultă că prima sumă ($\sum_{i=1}^m \alpha_i K(x_i, x_i) = c^2$) din cadrul funcției obiectiv a problemei duale MEB nu are niciun efect asupra soluției optime a acestei probleme. În consecință, forma duală a problemei MEB este — în cazul $\|\phi(x_i)\| = c$ pentru $i = 1, \dots, m$ — echivalentă cu forma duală a problemei *Max Margin*. Ambele probleme vor produce același \hat{w} (pentru forma duală), și deci aceeași soluție optimă \hat{w} (pentru forma primală).

33.

(Problema ν -SVM)

*prelucrare de Liviu Ciortuz, după
■ B. Schölkopf, A. Smola, "Learning with Kernels",
MIT Press, 2002, pag. 206-209*

Parametrul de „destindere“ $C > 0$, din forma problemei SVM cu margine “soft” (introdus de catre Cortes și Vapnik, în *Support Vector Networks*, 1995),⁷⁰⁷ permite realizarea unui compromis între două obiective antagoniste: maximizarea marginii⁷⁰⁸ și minimizarea erorii la antrenare. La valori mari ale lui C rezultă un nivel scăzut al sumei erorilor în raport cu marginea (și anume, $\sum_{i=1}^m \xi_i$, în notația uzuală). Invers, la valori mici ale lui C rezultă un nivel ridicat al sumei erorilor. Totuși, semnificația parametrului C este prea puțin intuitivă, iar valoarea sa nu poate fi determinată a priori.

⁷⁰⁶Din nou, pe scurt,

$$\begin{aligned} L_P(w, R, \alpha) &= R^2 + \sum_{i=1}^m \alpha_i [(w - \phi(x_i))^2 - R^2] \\ \frac{\partial}{\partial w} L_P(w, R, \alpha) = 0 &\Leftrightarrow \sum_{i=1}^m \alpha_i 2(w - \phi(x_i)) = 0 \Leftrightarrow \sum_{i=1}^m \alpha_i w = \sum_{i=1}^m \alpha_i \phi(x_i) \Leftrightarrow w = \frac{\sum_{i=1}^m \alpha_i \phi(x_i)}{\sum_{i=1}^m \alpha_i} \\ \frac{\partial}{\partial R} L_P(w, \rho, \alpha) = 0 &\Leftrightarrow 2R - \sum_{i=1}^m \alpha_i 2R = 0 \stackrel{R \neq 0}{\Leftrightarrow} \sum_{i=1}^m \alpha_i = 1 \\ &\Rightarrow w = \sum_{i=1}^m \alpha_i \phi(x_i) \text{ și} \\ L_D(\alpha) &= R^2 + \sum_{i=1}^m \alpha_i \left[\underbrace{\left(\sum_{j=1}^m \alpha_j \phi(x_j) - \phi(x_i) \right)^2}_{w} - R^2 \right] \\ &= R^2 - R^2 \underbrace{\sum_{i=1}^m \alpha_i}_{1} + w^2 \underbrace{\sum_{i=1}^m \alpha_i}_{1} - 2 \sum_{i,j=1}^m \alpha_i \alpha_j \underbrace{\phi(x_i) \cdot \phi(x_j)}_{K(x_i, x_j)} + \sum_{i=1}^m \alpha_i \underbrace{\phi(x_i)^2}_{K(x_i, x_i)} \\ &= \sum_{i,j=1}^m \alpha_i \alpha_j \underbrace{\phi(x_i) \cdot \phi(x_j)}_{K(x_i, x_j)} - 2 \sum_{i,j=1}^m \alpha_i \alpha_j K(x_i, x_j) + \sum_{i=1}^m \alpha_i K(x_i, x_i) = - \sum_{i,j=1}^m \alpha_i \alpha_j K(x_i, x_j) + \sum_{i=1}^m \alpha_i K(x_i, x_i). \end{aligned}$$

⁷⁰⁷Vedeți problema 12.

⁷⁰⁸Definită ca distanța de la hiperplanul de separare optimală $w \cdot x + w_0 = 0$ până la vectorii-suport x_i pentru care $(w \cdot x_i + w_0)y_i = 1$.

De aceea, în locul acestei variante de SVM cu margine “soft”, B. Schölkopf, A. Smola, R. Williamson și P. Bartlett au propus în articolul *New Support Vector Machines*⁷⁰⁹ o abordare diferită, în care se folosește un alt parametru numeric, ν , astfel încât dacă m reprezintă numărul instanțelor de antrenament, atunci νm va limita superior numărul de erori produse la antrenare.⁷¹⁰ Se poate demonstra că νm este totodată o margine inferioară pentru numărul de vectori-suport.⁷¹¹

Varianta aceasta este cunoscută sub numele de ν -SVM și este caracterizată de forma primală următoare:

$$\begin{aligned} & \min_{w, w_0, \xi, \rho} \left(\frac{1}{2} \|w\|^2 - \nu \rho + \frac{1}{m} \sum_{i=1}^m \xi_i \right) \\ \text{a. i. } & y_i(w \cdot x_i + w_0) \geq \rho - \xi_i, \text{ pentru } i = 1, \dots, m \\ & \xi_i \geq 0 \text{ pentru } i = 1, \dots, m \\ & \rho \geq 0. \end{aligned} \quad (\mathbf{P}'')$$

De remarcat prezența variabilei suplimentare ρ , care va trebui să fie supusă procesului de optimizare la fel ca și variabilele w , w_0 și $\xi \stackrel{\text{not.}}{=} (\xi_1, \dots, \xi_m)$.⁷¹²

- Derivați forma duală corespunzătoare problemei ν -SVM. Simplificați rezultatul cât mai mult posibil.
- Stabiliti relațiile de legătură între \bar{w} , \bar{w}_0 , $\bar{\rho}$, prin care am notat soluțiile problemei (\mathbf{P}'') , și soluțiile problemei duale de la punctul a.
- Care este regula de clasificare a unei instanțe de test oarecare x' în acest model?

Răspuns:

- Vom urma liniile „metodologice“ care au fost folosite la problemele 9 și 12. Mai întâi, este ușor de verificat faptul că problema de optimizare (\mathbf{P}'') satisfacă condiția lui Slater:⁷¹³ luând $w = 0$, $w_0 = 0$, $\rho = 0$ și $\xi_i = 1$, restricțiile $y_i(w \cdot x_i + w_0) > \rho - \xi_i$ sunt îndeplinite, pentru $i = 1, \dots, m$. Prin urmare, vom putea opta ca în loc să rezolvăm problema (\mathbf{P}'') să rezolvăm duala ei, după care vom obține soluția problemei (\mathbf{P}'') folosind condițiile Karush-Kuhn-Tucker de optimalitate / staționaritate și respectiv complementaritate.

⁷⁰⁹ Publicat în revista Neural Computation, 12:1207-1245, 2000. (Vedeți și și *A Tutorial on ν -Support Vector Machines* de Pai-Hsuen Chen, Chih-Jen Lin, și Bernhard Schoelkopf, în vol. Applied Stochastic Models in Business and Industry, ed. Wiley InterScience, 2005.)

⁷¹⁰ LC: La problema 18 am arătat că în contextul problemei de optimizare C-SVM numărul de erori comise la antrenare este mărginit superior de către suma variabilelor de „destindere“ ($\sum_i \xi_i$). Similar, pentru problema (\mathbf{P}'') de mai jos se poate arăta imediat, analizând doar(I) forma restricțiilor, că numărul de erori comise la antrenare este aici mărginit superior de $\frac{1}{\rho} \sum_i \xi_i$. Prin urmare, dacă impunem condiția $\nu m \leq \frac{1}{\rho} \sum_i \xi_i$ rezultă

$\nu \rho \leq \frac{1}{m} \sum_i \xi_i$. Aceasta explică forma funcției obiectiv din problema (\mathbf{P}'') de mai jos.

⁷¹¹ Vedeți *Observația* de la finalul rezolvării punctului a de mai jos.

⁷¹² Din forma restricțiilor liniare rezultă că distanța de la separatorul optimal la vectorii-suport de pe margine (engl., bound support vectors) — adică acei vectori-suport pentru care multiplicatorii Lagrange corespunzători vor apărea intervalului $(0, \frac{1}{m})$, conform problemei duale (\mathbf{D}'') de mai jos — va fi $\frac{\rho}{\|w\|}$. (Vedeți relația (300) de la problema 12.)

⁷¹³ Pentru formalizarea acestei condiții, vedeți *Comentariul* de la problema 9 (pag. 642).

Apoi, lagrangeanul generalizat pentru problema (P'') este

$$\begin{aligned} L_P(w, w_0, \xi, \rho, \alpha, \beta, \delta) &= \frac{1}{2} \|w\|^2 - \nu\rho + \frac{1}{m} \sum_{i=1}^m \xi_i \\ &\quad - \sum_{i=1}^m \alpha_i (y_i(w \cdot x_i + w_0) - \rho + \xi_i) - \sum_{i=1}^m \beta_i \xi_i - \delta\rho, \end{aligned}$$

unde α_i, β_i și δ sunt multiplicatori Lagrange (corespunzători celor trei tipuri de inegalități din problema (P'')), toți trebuind să satisfacă restricția de ne-negativitate.

Calculând derivatele parțiale ale lui L_P în raport cu variabilele primale w, w_0, ξ și ρ și apoi egalându-le cu 0, vom obține imediat:

$$\begin{aligned} w &= \sum_{i=1}^m \alpha_i y_i x_i \\ \sum_{i=1}^m \alpha_i y_i &= 0 \\ \alpha_i + \beta_i &= \frac{1}{m} \text{ pentru } i = 1, \dots, m \\ \sum_{i=1}^m \alpha_i - \delta &= \nu. \end{aligned}$$

Din relația $\alpha_i + \beta_i = \frac{1}{m}$, ținând cont că $\alpha_i \geq 0$ și $\beta_i \geq 0$, rezultă $\alpha_i \in \left[0, \frac{1}{m}\right]$ și $\beta_i \in \left[0, \frac{1}{m}\right]$ pentru $i = 1, \dots, m$. De asemenea, știind că $\delta \geq 0$, din relația $\sum_{i=1}^m \alpha_i - \delta = \nu$ rezultă că $\sum_{i=1}^m \alpha_i \geq \nu$.

Condițiile de complementaritate Karush-Kuhn-Tucker sunt: $\alpha_i(y_i(w \cdot x_i + w_0) - \rho + \xi_i) = 0$, $\beta_i \xi_i = 0$ și $\delta\rho = 0$.

Substituind $w = \sum_{i=1}^m \alpha_i y_i x_i$ în expresia lui L_P și apoi făcând diversele simplificări posibile, va rezulta că forma duală corespunzătoare problemei (P'') este:

$$\begin{aligned} \max_{\alpha} & \left(-\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \right) \\ \text{a. i. } & 0 \leq \alpha_i \leq \frac{1}{m} \text{ pentru } i = 1, \dots, m \\ & \sum_{i=1}^m \alpha_i y_i = 0 \\ & \sum_{i=1}^m \alpha_i \geq \nu. \end{aligned} \tag{D''}$$

De remarcat că

- multiplicatorii Lagrange β_i și δ nu apar în (D'') ;
- în funcția obiectiv a problemei duale (D'') nu apare termenul $\sum_{i=1}^m \alpha_i$, care era prezent în funcția obiectiv a problemei duale atât pentru SVM cu margine “hard” cât și pentru C-SVM;
- în schimb, în partea de restricții apare condiția suplimentară $\sum_{i=1}^m \alpha_i \geq \nu$.

Observație: Din relația $0 \leq \alpha_i \leq \frac{1}{m}$ rezultă imediat că $\frac{\#SV}{m} \geq \sum_{i=1}^m \alpha_i \geq \nu$, deci $\#SV \geq \nu m$, unde prin $\#SV$ am notat numărul vectorilor-suport.

b. În ce privește legătura dintre soluțiile formei primale (P'') și cele ale formei duale (D'') , avem mai întâi $\bar{w} = \sum_{i=1}^m \bar{\alpha}_i y_i x_i$. Apoi, din condițiile de

complementaritate Karush-Kuhn-Tucker deducem că dacă există un $\bar{\alpha}_i$ astfel încât $0 < \bar{\alpha}_i < \frac{1}{m}$, atunci $\bar{\beta}_i > 0$ și deci $\bar{\xi}_i = 0$. De asemenea, $\bar{\alpha}_i > 0$ implică $y_i(\bar{w} \cdot x_i + \bar{w}_0) - \bar{\rho} + \bar{\xi}_i = 0$, de unde rezultă $y_i(\bar{w} \cdot x_i + \bar{w}_0) = \bar{\rho}$.

Folosind această ultimă relație, valorile optime \bar{w}_0 și $\bar{\rho}$ se vor determina astfel: dacă avem x_+ o instanță pozitivă și x_- o instanță negativă astfel încât multipli-catorii Lagrange corespunzători sunt în intervalul $(0, \frac{1}{m})$,⁷¹⁴ atunci $\bar{w} \cdot x_+ + \bar{w}_0 = \bar{\rho}$ și $\bar{w} \cdot x_- + \bar{w}_0 = -\bar{\rho}$. Aceste ultime două ecuații formează un sistem din care se obțin imediat \bar{w}_0 și $\bar{\rho}$:⁷¹⁵

$$\begin{aligned}\bar{w}_0 &= -\frac{1}{2}\bar{w} \cdot (x_+ + x_-) \\ \bar{\rho} &= \frac{1}{2}\bar{w} \cdot (x_+ - x_-)\end{aligned}$$

Observație: În articolul despre ν -SVM citat în enunț, autorii extind acest procedeu de calcul pentru valorile \bar{w}_0 și $\bar{\rho}$ la mai multe perechi de instanțe pozitive și respectiv negative, pentru a obține un rezultat cât mai robust.

c. Dată fiind o instanță nouă (de test) x' , ea va fi clasificată conform expresiei

$$\text{sign}\left(\sum_{i=1}^m \bar{\alpha}_i y_i x_i \cdot x' + \bar{w}_0\right).$$

34.

(SVR — Regresie cu vectori-suport, varianta “hard-margin”, adică, fără variabile de „destindere”)

*prelucrare de Liviu Ciortuz, după
■ • ○ Stanford, 2014 fall, Andrew Ng, midterm, pr. 4*

Până acum, am văzut cum anume putem face clasificare cu algoritmul [C]-SVM. În acest exercițiu vom studia o variantă a acestui algoritm, care servește pentru a face regresie. Așadar, etichetele asociate aici instanțelor de antrenament vor avea valori continue, $y \in \mathbb{R}$. În mod natural, noul algoritm poartă numele de Regresie cu Vectori-Suport (engl., Support Vector Regression, SVR).⁷¹⁶

Presupunem că avem un set de date de antrenament $\{(x_1, y_1), \dots, (x_m, y_m)\}$, cu $x_i \in \mathbb{R}^{n+1}$ și $y_i \in \mathbb{R}$ pentru $i = 1, \dots, m$. Urmărind să găsim o ipoteză [de regresie] de forma $h_{w,b}(x) = w \cdot x + b$, vom scrie următoarea problemă de optimizare (convexă):

$$\min_{w,b} \frac{1}{2} \|w\|^2 \quad (322)$$

$$\text{a. i. } y_i - (w \cdot x_i + b) \leq \varepsilon \text{ pentru } i = 1, \dots, m \quad (322)$$

$$(w \cdot x_i + b) - y_i \leq \varepsilon \text{ pentru } i = 1, \dots, m, \quad (323)$$

⁷¹⁴De remarcat că dacă există una din cele două instanțe, atunci în mod necesar există și cea de-a două, fiindcă $\sum_i \bar{\alpha}_i y_i = 0$.

⁷¹⁵Dacă toți $\bar{\alpha}_i$ sunt 0, atunci $\bar{w} = 0$, ceea ce este în afara discuției, după cum am justificat la problema 9.

Rămâne de tratat cazul în care pentru orice $i \in \{1, \dots, m\}$ avem fie $\bar{\alpha}_i = 0$ fie $\bar{\alpha}_i = \frac{1}{m}$, știind că există cel puțin un $\bar{\alpha}_i$ care are valoarea $\frac{1}{m}$, fiindcă $\sum_{i=1}^m \bar{\alpha}_i \geq \nu$. Acest caz se tratează similar cu cazul corespunzător de la problema 12.

⁷¹⁶Puteți consulta articolul *A tutorial on support vector regression* de Alex Smola și Bernhard Schoelkopf, publicat în revista *Statistics and Computing*, 14:199–222, 2004.

unde $\varepsilon > 0$ are o valoare dată, fixată.

Observații:

1. Inegalitățile (322) și (323) se pot scrie în mod combinat sub forma

$$-\varepsilon \leq y_i - (w \cdot x_i + b) \leq \varepsilon \Leftrightarrow |y_i - (w \cdot x_i + b)| \leq \varepsilon.$$

Semnificația analitică a acestei inegalități este imediată. Așadar, restricția asupra *marginii funcționale* de la problema de optimizare SVM a fost modificată aici în aşa fel încât să se refere la diferența dintre valorile reale ale lui y și outputul produs de ipoteza pe care urmărim să o învățăm folosind SVR.

2. În mod similar cu problema [C-]SVM, în noua problemă de optimizare se cere ca valoarea lui $\|w\|$ să fie [cât mai] mică. Minimizarea lui $\|w\|$ corespunde maximizării distanței $(\varepsilon/\|w\|)$ dintre hiperplanul reprezentat de ipoteza $w \cdot x + b$ și marginile $w \cdot x + b = \varepsilon$ și $w \cdot x + b = -\varepsilon$ între care trebuie să se situeze toate valorile y_i .⁷¹⁷

a. Scrieți lagrangeanul corespunzător problemei de optimizare (SVR) de mai sus. Vă recomandăm să folosiți două seturi de variabile / multiplicatori Lagrange, α_i și α_i^* , cu $i = 1, \dots, m$, corespunzător restricțiilor reprezentate de cele două inegalități (care au fost etichetate cu (322) și (323)). Prin urmare, funcția lagrangeană va fi de forma $L(w, b, \alpha, \alpha^*)$.

b. Obțineți forma duală a problemei de optimizare. (Va trebui să calculați derivatele parțiale ale lagrangeanului în raport cu w și respectiv cu b .)

c. Demonstrați că acest algoritm poate fi kernel-izat. Pentru aceasta, va trebui să arătați că (i) funcția obiectiv pentru forma duală a problemei de optimizare poate fi scrisă folosind (doar) produse scalare între instanțele de antrenament, și (ii) la faza de testare, dată fiind o instanță nouă x , ipoteza $h_{w,b}(x)$ poate fi calculată folosind de asemenea (doar) produse scalare.

Răspuns:

- Fie $\alpha_i, \alpha_i^* \geq 0$ ($i = 1, \dots, m$) multiplicatorii Lagrange pentru restricțiile (322) și respectiv (323). Atunci, lagrangeanul poate fi scris astfel:

$$L(w, b, \alpha, \alpha^*) = \frac{1}{2} \|w\|^2 + \sum_{i=1}^m \alpha_i(y_i - w \cdot x_i - b - \varepsilon) + \sum_{i=1}^m \alpha_i^*(-y_i + w \cdot x_i + b - \varepsilon).$$

- Mai întâi, precizăm că funcția obiectiv pentru forma duală a problemei de optimizare SVR este definită astfel:

$$L_D(\alpha, \alpha^*) = \min_{w, b} L(w, b, \alpha, \alpha^*).$$

Acum, calculând derivatele lagrangeanului în raport cu variabilele primale și egalându-le apoi cu 0, vom avea:

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^m (\alpha_i - \alpha_i^*) x_i = 0 \Rightarrow w = \sum_{i=1}^m (\alpha_i - \alpha_i^*) x_i \quad (324)$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^m (\alpha_i^* - \alpha_i) = 0. \quad (325)$$

⁷¹⁷De asemenea, minimizarea lui $\|w\|$ corespunde obiectivului de a reduce overfitting-ul, aşa cum ştim și de la celelalte metode de regresie studiate.

Substituind aceste două relații în $L(w, b, \alpha, \alpha^*)$, vom obține:

$$\begin{aligned}
L_D(\alpha, \alpha^*) &= \frac{1}{2} \|w\|^2 + \sum_{i=1}^m \alpha_i (y_i - w \cdot x_i - b - \varepsilon) + \sum_{i=1}^m \alpha_i^* (-y_i + w \cdot x_i + b - \varepsilon) \\
&= \frac{1}{2} \|w\|^2 - \varepsilon \sum_{i=1}^m (\alpha_i + \alpha_i^*) + \sum_{i=1}^m y_i (\alpha_i - \alpha_i^*) \\
&\quad - \sum_{i=1}^m (\alpha_i - \alpha_i^*) w \cdot x_i - b \underbrace{\sum_{i=1}^m (\alpha_i - \alpha_i^*)}_0 \\
&\stackrel{(324)}{=} \frac{1}{2} \left\| \sum_{i=1}^m (\alpha_i - \alpha_i^*) x_i \right\|^2 - \varepsilon \sum_{i=1}^m (\alpha_i + \alpha_i^*) + \sum_{i=1}^m y_i (\alpha_i - \alpha_i^*) \\
&\quad - \sum_{i=1}^m (\alpha_i - \alpha_i^*) \sum_{j=1}^m (\alpha_j - \alpha_j^*) x_j \cdot x_i \\
&= -\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m (\alpha_i - \alpha_i^*) (\alpha_j - \alpha_j^*) x_i \cdot x_j - \varepsilon \sum_{i=1}^m (\alpha_i + \alpha_i^*) + \sum_{i=1}^m y_i (\alpha_i - \alpha_i^*).
\end{aligned}$$

Așadar, problema duală SVR poate fi formulată astfel:

$$\begin{aligned}
&\max_{\alpha, \alpha^*} \left(-\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m (\alpha_i - \alpha_i^*) (\alpha_j - \alpha_j^*) x_i \cdot x_j - \varepsilon \sum_{i=1}^m (\alpha_i + \alpha_i^*) + \sum_{i=1}^m y_i (\alpha_i - \alpha_i^*) \right) \\
&\text{a. i. } \sum_{i=1}^m (\alpha_i - \alpha_i^*) = 0 \\
&\quad \alpha_i, \alpha_i^* \geq 0 \text{ pentru } i = 1, \dots, m.
\end{aligned}$$

c. Se observă că în expresia lagrangeanului dual L_D instanțele de antrenament apar întotdeauna în produse scalare de forma $x_i \cdot x_j$. În mod similar, atunci când vrem să facem predicție pentru o instanță oarecare x , vom avea:

$$w \cdot x + b \stackrel{(324)}{=} \sum_{i=1}^m (\alpha_i - \alpha_i^*) x_i \cdot x + b.$$

Așadar, algoritmul SVR poate fi kernel-izat.

5.2 Maşini cu vectori-suport — Probleme propuse

5.2.1 SVM cu margine “hard”

35.

(Calcularea distanței de la un hiperplan la originea sistemului de coordonate; rezolvare cu metoda vectorială și cu metoda multiplicatorilor lui Lagrange)

*prelucrare de Liviu Ciortuz, după
 CMU, 2018 spring, Nina Balcan, HW0, pr. “Geometry”*

Considerăm hiperplanul de ecuație $w \cdot x + b = 0$, unde w și x sunt din \mathbb{R}^d , iar $b \in \mathbb{R}$.

Demonstrați că distanța geometrică [deci fără semn] de la originea sistemului de coordonate până la [cel mai apropiat punct de pe] hiperplanul de ecuație $w \cdot x + b = 0$ este $\frac{|b|}{\|w\|}$.

Veți face demonstrația atât în manieră vectorială⁷¹⁸ cât și cu ajutorul metodei multiplicatorilor lui Lagrange.⁷¹⁹

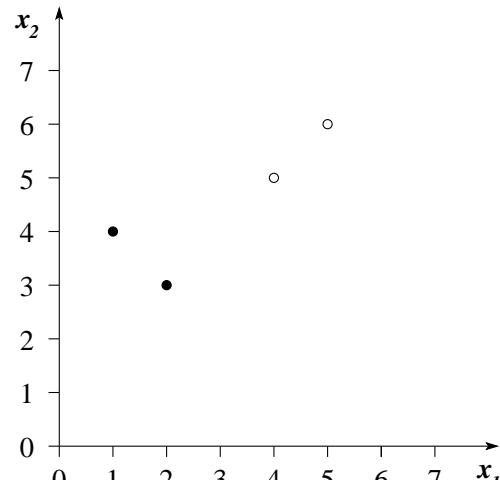
36.

(SVM liniară: aplicare pe date din \mathbb{R}^2)

CMU, 2012 fall, T. Mitchell, Z. Bar-Joseph, final exam, pr. 7.b

Presupunem că antrenați o mașină cu vectori-suport pe setul de date din figura alăturată. Acest set de date constă din două exemple cu eticheta $+1$ (aceste exemple sunt marcate cu semnul \bullet) și două exemple cu eticheta -1 (marcate cu \circ).

- a. Care este ecuația corespunzătoare separatorului optimal?
- b. Trasați separatorul optimal și încercuiți vectorii-suport.



⁷¹⁸Puteți prelua ideea rezolvării de la problema 1, însă vă cerem să nu preluăți pur și simplu formula care a fost obținută acolo și s-o aplicați la cazul de față. Dacă doriti să lucrați în *cazul particular* $d = 2$, puteți folosi mijloace obișnuite de analiză matematică, în special proprietățile funcției polinomiale de gradul al doilea.

⁷¹⁹Pentru o scurtă introducere la metoda multiplicatorilor lui Lagrange, vedeți *Introducerea* de la problema 82, *Comentariul* de la problema 83, precum și problemele 85 și 173 din secțiunea *Metode de optimizare* de la capitolul de *Fundamente*.

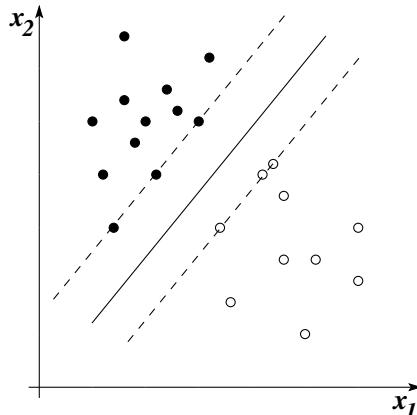
Observație: În problema de optimizare care corespunde formulării din enunț nu avem restricții de tip inegalitate, deci veți putea rezolva ușor problema de optimizare în *forma sa primală*, făcând apel [nu neapărat în mod explicit] la condițiile Karush-Kuhn-Tucker.

37.

(Separabilitate în \mathbb{R}^2 ; SVM liniară; calculul erorii la CVLOO)

* CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, final exam, pr. 3.f

Care este eroarea la cross-validation cu metoda “Leave-One-Out” atunci când folosim o mașină cu vectori-suport ca în figura următoare?



38.

(SVM liniară în \mathbb{R}^2 , forma primală: determinarea hiperplanului de separare optimală)

• CMU, 2009 spring, Ziv Bar-Joseph, HW3, pr. 4.1

Considerăm că în \mathbb{R}^2 sunt date două puncte: (x_1, y_1) cu eticheta +1 și (x_2, y_2) cu eticheta -1. Care este separatorul decizional pe care îl vom obține rulând o SVM liniară pe setul de date de antrenament compus din aceste două puncte? Calculați expresia analitică a acestui separator.

39.

(SVM — problema de optimizare în forma primală: familiarizare cu noțiunile de margine și separator optimal în spațiul de „trăsături“)

• CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW4, pr. 4.5
MIT, 2004 fall, Tommi Jaakkola, HW3, pr. 1.3

Considerăm un set de date din \mathbb{R} , foarte simplu, constând din numai două exemple de antrenament:

$$(x_1 = 0, y_1 = -1) \text{ și } (x_2 = \sqrt{2}, y_2 = 1).$$

Vom folosi o funcție-nucleu polinomială de ordinul al doilea, mai precis vom pune în corespondență fiecare instanță x cu vectorul

$$\Phi(x) = (1, \sqrt{2}x, x^2).$$

Vrem să găsim *soluția* $\hat{w} = (\hat{w}_1, \hat{w}_2, \hat{w}_3)$ și \hat{w}_0 a problemei de optimizare SVM [în așa-numitul spațiu de „trăsături“ determinat de transformarea Φ]:

$$\min_{w,w_0} \frac{1}{2} \|w\|^2$$

astfel încât $y_1(w \cdot \Phi(x_1) + w_0) \geq 1$
 $y_2(w \cdot \Phi(x_2) + w_0) \geq 1.$

- a. Folosind ceea ce știți despre [graniță de] separare cu margine maximală, indicați un vector care are aceeași direcție cu vectorul-soluție \hat{w} .⁷²⁰
- b. Cât este valoarea *marginii de separare* pe care o obținem pentru aceste date (evidenț, în spațiul de „trăsături“ indicat mai sus)?
- c. Făcând legătura dintre marginea de separare și $\|\hat{w}\|$, indicați soluția \hat{w} și deduceți valoarea lui \hat{w}_0 .

40.

(Neseparabilitate liniară:
 două exemple de mapare a trăsăturilor;
 rezolvarea — în manieră directă — a problemei SVM
 în spațiul / spațiile de trăsături)

• CMU, 2011 fall, Eric Xing, HW4, pr. 3.1.1-7

Se dau 6 puncte din \mathbb{R} : $x_1 = -1, x_2 = 0, x_3 = 1$ au etichete negative, iar $x_4 = -2, x_5 = 2, x_6 = 3$ au etichete pozitive.

- a. Desenați cele 6 puncte pe axa reală, folosind simbolul \circ pentru a reprezenta etichetele negative și simbolul \bullet pentru etichetele pozitive.
- b. Aceste 6 puncte nu sunt liniar-separabile. Definiți $f : \mathbb{R} \rightarrow \mathbb{R}$, o funcție de transformare a trăsăturilor astfel încât punctele $f(x_1), f(x_2), \dots, f(x_6)$ să fie liniar-separabile. Desenați pe axa reală cele 6 instanțe astfel obținute. Apoi indicați poziția separatorului optimal calculat de mașina cu vectori-suport liniară cu margine “hard”, marcând instanțele care sunt vectori-suport.
- c. Separatorul de la punctul precedent are forma analitică $w_0 + w_1 f(x) = 0$. Indicați valorile lui w_0 și w_1 .
- d. Să presupunem acum că mapăm cele 6 puncte în spațiul de trăsături $(x, f(x))$, unde $f(x)$ este funcția de transformare a trăsăturilor de la punctul b. Cu alte cuvinte, acum vom avea 6 puncte în plan, $(x_1, f(x_1)), (x_2, f(x_2)), \dots, (x_6, f(x_6))$. Desenați aceste 6 puncte în planul bidimensional, împreună cu separatorul optimal definit de către mașina cu vectori-suport liniară cu margine “hard”. Indicați apoi vectorii-suport.
- e. Separatorul de la punctul precedent are forma analitică $w_0 + w_1 x + w_2 f(x) = 0$. Calculați valorile parametrilor w_0, w_1 și w_2 .
- f. Funcția de mapare a trăsăturilor $x \rightarrow (x, f(x))$ de la punctele d și e este asociată cu o funcție-nucleu $K(x, x')$, unde x și x' sunt puncte din spațiul original (de trăsături) unidimensional. Scrieți expresia acestei funcții-nucleu.

⁷²⁰Veti ține cont că

1. într-un spațiu de tip \mathbb{R}^d , egalitatea $w \cdot x = 0$ implică faptul că vectorii w și x sunt ortogonali / perpendiculari;
 2. ecuația unui hiperplan din \mathbb{R}^d este de forma $w \cdot x + w_0 = 0$, ceea ce implică faptul că vectorul w este perpendicular pe acest hiperplan (întrucât este perpendicular pe orice vector din hiperplan, văzând un astfel de vector ca diferență de doi vectori de poziție $x_1 - x_2$, unde punctele x_1 și x_2 aparțin hiperplanului).

Așadar, vectorul \hat{w} din problema noastră este perpendicular pe hiperplanul de separare optimală.

41.

(Găsirea separatorului liniar optimal,
după maparea într-un „spațiu de trăsături“,
conform unei funcții-nucleu date)

* o CMU, 2008 spring, Eric Xing, midterm exam, pr. 5
MIT, 2006 fall, Tommi Jaakkola, midterm exam, pr. 4

Presupunem că avem instanțele pozitive $x_1 = (0.4, 0.2)$, $x_2 = (0.8, 0.4)$, $x_3 = (0.2, 0.4)$, $x_4 = (0.4, 0.8)$ și instanțele negative $x_5 = (0.4, 0.4)$ și $x_6 = (0.8, 0.8)$. Se poate constata ușor că aceste date nu sunt separabile liniar.

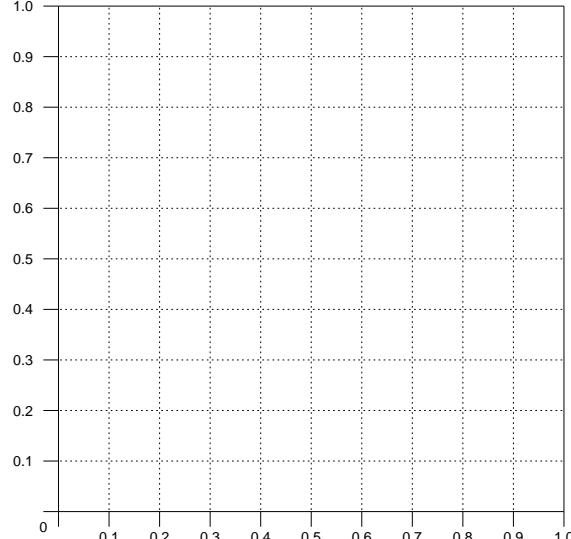
Vom folosi o mapare Φ astfel încât

$$\Phi(x) \cdot \Phi(x') \stackrel{\text{not.}}{=} K(x, x') = \frac{x \cdot x'}{\|x\| \|x'\|}.$$

a. Care este vectorul $\Phi(x)$ corespunzător acestei funcții-nucleu?

b. Pe gridul alăturat, reprezentați atât instanțele de antrenament x_i cât și „imaginile“ $\Phi(x_i)$, pentru $i = 1, \dots, 6$. Fiecare „imagină“ $\Phi(x_i)$ va avea asociată aceeași etichetă ca și x_i . Veți folosi *convenția* noastră de notare: simbolul \bullet desemnează instanțe pozitive, iar simbolul \circ instanțe negative.

Sugestie: În loc să faceți efectiv calcularea coordonatelor pentru „imaginile“ $\Phi(x_i)$, ar fi mai util / interesant să vedeți cât este $\|\Phi(x_i)\|$ pentru fiecare x_i , iar apoi să tragăți o *concluzie generală* cu privire la poziția relativă a lui $\phi(x_i)$ în raport cu x_i , pentru orice i .



c. Veți observa că instanțele $\Phi(x_i)$ sunt liniar-separabile, deci există o dreaptă determinată de parametrii w_1, w_2 și w_0 , care le separă în mod optimal. Cât este raportul w_1/w_2 ?

Indicație: Nu este necesar să calculați efectiv w_1 și w_2 . Este suficient să exploatați simetriile.

d. Pe desenul de la punctul b, încercuiți vectorii-suport $\Phi(x_i)$.

e. Indicați în spațiul de origine zonele de decizie [și separatorul decizional] care corespund separatorului optimal găsit la punctul c.

42.

(Învățarea conceputului \neg XOR,
folosind forma duală a problemei SVM
și o funcție-nucleu polinomială de ordin 2)

* CMU, 2006 fall, E. Xing, T. Mitchell, midterm exam, pr. 5

Fie o problemă de învățare supervizată în care exemplele de antrenament se află în spațiul euclidian bidimensional. Exemplele pozitive sunt $x_1 = (1, 1)$ și $x_3 = (-1, -1)$, iar exemplele negative sunt $x_2 = (-1, 1)$ și $x_4 = (1, -1)$.

Considerăm transformarea $\Phi(x)$ corespunzătoare funcției-nucleu $K(x, x') = (x \cdot x' + 1)^2$, unde x și x' sunt din \mathbb{R}^2 .

Funcția de predicție pe care vrem să-o obținem este de forma $y(x) = w \cdot \Phi(x) + w_0$, unde $\Phi(x), w \in \mathbb{R}^n$ (cu n ales convenabil), $w_0 \in \mathbb{R}$, iar \cdot reprezintă produsul scalar.

Indicați coeficienții w, w_0 corespunzători suprafeței de separare cu margine maximă pentru punctele de antrenament date. Cum va fi clasificată o instanță oarecare de test $x \in \mathbb{R}^2$?

Sugestie de lucru:

- Mai întâi reprezentați datele. Determinați $\Phi(x)$, n , și $\Phi(x_1), \Phi(x_2), \Phi(x_3), \Phi(x_4)$.
- Calculați apoi lagrangeanul dual

$$L_D(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j y_i y_j \alpha_i \alpha_j \Phi(x_i) \cdot \Phi(x_j)$$

- Determinați soluția problemei duale $\bar{\alpha} = \operatorname{argmax}_{\alpha} L_D(\alpha)$ cu restricțiile $\alpha_i \geq 0$ pentru $i = 1, \dots, 4$ și $\sum_i \alpha_i y_i = 0$ și, în final, găsiți soluția problemei primale $\bar{w} = \sum_i \bar{\alpha}_i y_i \Phi(x_i)$ și \bar{w}_0 , știind că \bar{w}_0 poate fi obținut dintr-una din restricțiile $y_i(\bar{w} \cdot \Phi(x_i) + \bar{w}_0) = 1$.

43.

(SVM cu diferite funcții-nucleu:
efectul unei translatări a datelor de antrenament
asupra poziției separatorului optimal)

CMU, 2009 fall, Carlos Guestrin, HW3, pr. 2.10

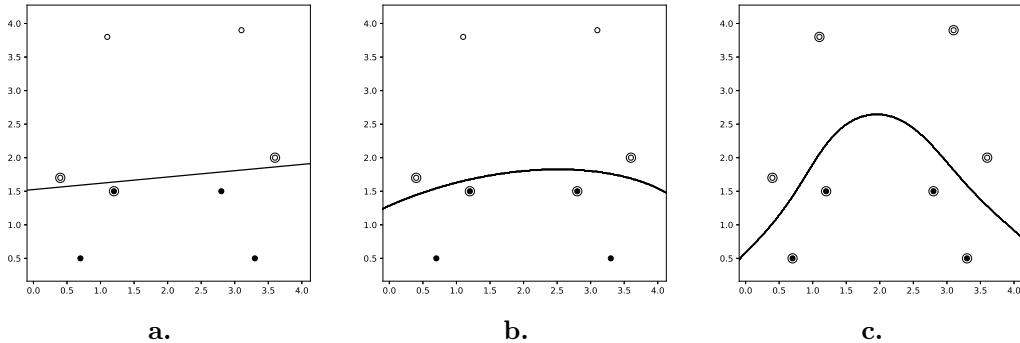
Trei mașini cu vectori-suport au fost antrenate pe un set de date din \mathbb{R}^2 folosind

- o funcție-nucleu liniară $K(x, y) = x \cdot y$ (a se vedea figura a);
- o funcție-nucleu polinomială de gradul al doilea $K(x, y) = (x \cdot y + 1)^2$ (a se vedea figura b);
- o funcție-nucleu cu baza radială $K(x, y) = e^{-\frac{\|x - y\|^2}{2\sigma^2}}$ (a se vedea figura c).

Presupunem că translatăm datele adăugând o constantă mare (de exemplu 10) la coordonata de pe axa verticală, pentru fiecare din datele de antrenament, adică (x, y) devine $(x, y + 10)$.

Dacă reantrenăm SVM-urile de mai sus pe noile date de antrenament, se schimbă și poziția separatorului optimal în raport cu datele? Tratați pe rând cazurile a, b și c.

Explicați pe scurt de ce se schimbă (sau de ce nu se schimbă) poziția separatorului optimal în raport cu datele, pentru fiecare din cazurile a, b și c. Trasați pe desen poziția noului separator optimal, acolo unde este cazul.⁷²¹



44. (SVM cu nucleu RBF — o caracteristică surprinzătoare: instanțe foarte distanțate față de separatorul optimal pot fi vectori-suport)

- MIT, 2002 fall, Tommi Jaakkola, midterm exam, pr. 3.1-3
CMU, 2011 spring, Tom Mitchell, HW6, pr. 1.2

[Remember:] Folosind o funcție-nucleu oarecare, SVM caută într-un anumit „spațiu de trăsături” Q un hiperplan care să maximizeze distanța dintre cele două clase. Clasificarea unei instanțe oarecare de test x se face determinând semnul expresiei

$$\bar{w} \cdot \Phi(x) + \bar{w}_0 = \left(\sum_{i \in SV} y_i \bar{\alpha}_i \Phi(x_i) \right) \cdot \Phi(x) + \bar{w}_0 = \sum_{i \in SV} y_i \bar{\alpha}_i K(x_i, x) + \bar{w}_0 \stackrel{not.}{=} f(x; \bar{\alpha}, \bar{w}_0),$$

unde

\bar{w} și \bar{w}_0 sunt parametrii pentru hiperplanul de clasificare în spațiu Q ,

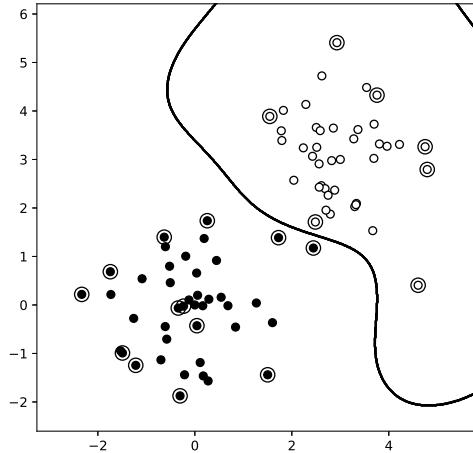
SV este setul de vectori-suport,

$\bar{\alpha}_i$ este coeficientul (multiplicatorul Lagrange) corespunzător vectorului-suport i .

În acest exercițiu vom folosi drept nucleu funcția cu baza radială $K(x_i, x_j) = e^{-\frac{1}{2}||x_i - x_j||^2}$. Vom presupune că instanțele de antrenament sunt liniar separabile în spațiu Q .

⁷²¹Sugestie: Puteți consulta articolul *On In variance of Support Vector Machines*, de Shigeo Abe, <http://www2.kobe-u.ac.jp/~abe/pdf/ideal2003.pdf> accesat la data 12.06.2018. LC: Mulțumesc lui Sebastian Cionanu pentru această sugestie.

Observație: Figura alăturată prezintă grafic rezultatul obținut de SVM folosind nucleu RBF pe un set de date din planul euclidian. Vectorii-suport sunt încercuiți. Ceea ce este curios la această clasificare este că unii vectori-suport sunt destul de departe de separatorul optimal. Și, totuși, ei sunt vectori-suport! Întrebările de mai jos vor încerca să contribuie la elucidarea acestei chestiuni.



- a. Arătați că ori de câte ori alegem un punct de test $x_{\text{far}} \in \mathbb{R}^d$ foarte distanțat de orice instanță de antrenament x_i , rezultă că $f(x_{\text{far}}; \bar{\alpha}, \bar{w}_0) \approx \bar{w}_0$.

În continuare, vom presupune, pentru simplitate, că $\bar{w}_0 = 0$.

- b. Din ipoteză, întrucât setul de date de antrenament este liniar separabil în spațiul de trăsături Q , rezultă — în urma antrenării — că orice instanță de antrenament x_i satisface inegalitatea $y_i \bar{w} \cdot \Phi(x_i) \geq 1$, unde Φ este funcția de mapare corespunzătoare nucleului RBF specificat mai sus. Satisfac și x_{far} această inegalitate?

Observație: Pentru RBF, $\Phi(x)$ este un vector infinit, de aceea de drept întrebarea aceasta nu are sens aici. De fapt, puteți răspunde totuși la această întrebare ținând cont că $\bar{w} \cdot \Phi(x) = \sum_{i \in SV} y_i \bar{\alpha}_i K(x_i, x)$, iar valorile funcției-nucleu $K(x_i, x)$ pot fi calculate fără nicio dificultate.

- c. Dacă am include punctul x_{far} în mulțimea de date de antrenament, iar aceasta ramâne liniar separabilă în spațiul de trăsături Q (în raport cu același nucleu RBF), ar deveni oare x_{far} un vector-suport?

45.

(Comparație între SVM și alți clasificatori)

* o CMU, 2010 fall, Aarti Singh, midterm, pr. 1.2.1

Fie setul de date din \mathbb{R}^2 din tabelul alăturat. Identificați dintre clasificatorii următori pe aceia care obțin eroare de antrenare 0 pe acest set de date.

X_1	X_2	Y
0	0	+
1	0	-
0	1	-
1	1	+

- SVM cu nucleu polinomial de ordin 2, adică $(c + x \cdot x')^2$, în care metaparametrul c este la libera alegere;
- regresia logistică
- arbori ID3 de adâncime 2, adică având două nivele de test, dintre care unul este nivelul-rădăcină;
- 3-NN.

46.

(O legătură (simplă) între rețele neuronale și SVM cu funcție-nucleu polinomială)

*prelucrare de Liviu Ciortuz, după
• CMU, 2010 fall, Aarti Singh, HW5, pr. 4.2*

Se consideră datele de antrenament $\bar{x}_1, \dots, \bar{x}_m \in \mathbb{R}^d$, etichetate respectiv cu $y_1, \dots, y_m \in \{-1, +1\}$. Se consideră de asemenea o funcție-nucleu polinomială de gradul $p \in \mathbb{N}$

$$K(\bar{x}, \bar{x}') = (\gamma \bar{x} \cdot \bar{x}' + r)^p = \Phi(\bar{x}) \cdot \Phi(\bar{x}')$$

unde γ și r sunt numere reale.

Stim că o SVM care folosește această funcție-nucleu și care a fost antrenată pe datele de mai sus va clasifica o instanță de test $\bar{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$ conform expresiei:

$$\text{sign}(w \cdot \Phi(\bar{x}) + w_0) = \text{sign}\left(\sum_{i=1}^m y_i \alpha_i \Phi(\bar{x}) \cdot \Phi(\bar{x}_i) + w_0\right) = \text{sign}\left(\sum_{i=1}^m y_i \alpha_i K(\bar{x}, \bar{x}_i) + w_0\right),$$

unde $w \in \mathbb{R}^n$ (n fiind dimensiunea spațiului asociat cu „maparea” Φ) și $w_0 \in \mathbb{R}$ desemnează ponderile învățate de SVM, iar α_i sunt valorile multiplicatorilor Lagrange (corespunzători instanțelor de antrenament) care constituie soluțiile formei duale a problemei SVM.

Se consideră o rețea neuronală artificială de tip feed-forward cu un singur nivel ascuns, descrisă astfel:

- intrările în rețea sunt x_1, \dots, x_d (a se vedea \bar{x} de mai sus);
- rețeaua are m unități ascunse, și anume câte una pentru fiecare instanță de antrenament $\bar{x}_i = (x_{i1}, \dots, x_{id})$, cu $i \in \{1, \dots, m\}$;
- ponderile de pe conexiunile dintre intrări și unitățile ascunse sunt fixate astfel:
 - pentru unitatea ascunsă i , ponderile de la intrările x_1, \dots, x_d sunt respectiv x_{i1}, \dots, x_{id} (a se vedea \bar{x}_i de mai sus); aceasta are loc pentru fiecare $i \in \{1, \dots, m\}$;
 - nu se folosește termen liber / constant ($x_0 = 1$) pentru aceste unități ascunse;
- ieșirea unității ascunse i (cu $i \in \{1, \dots, m\}$) este definită într-o manieră ne-standard, și anume:

$$o_i = (\gamma \text{net}_i + r)^p, \text{ unde } \text{net}_i \text{ este } \sum_{j=1}^d x_j x_{ij}$$

- rețeaua are o singură unitate pe stratul de ieșire, pe care o vom nota cu $m+1$; ea are funcția de activare de tip sign ;
- ponderile hidden-output sunt de forma $y_i \alpha_i$ pentru fiecare $i \in \{1, \dots, m\}$;
- pentru unitatea de ieșire, ponderea intrării libere ($x_0 = 1$) este w_0 .

a. Să se deseneze această rețea neuronală sub forma unui graf în care intrările și unitățile neuronale sunt noduri, iar ponderile sunt marcate corespunzător pe arce.

b. Arătați că acești doi clasificatori — rețeaua neuronală și SVM-ul de mai sus — sunt echivalenți. Adică, pentru fiecare instanță de test $\bar{x} \in \mathbb{R}^d$ rezultatele

produse de către cei doi clasificatori (văzuți simplu ca funcții de clasificare) sunt identice. (Calculați funcția reprezentată de outputul rețelei neuronale.)

c. O rețea neuronală similară cu cea de mai sus poate chiar să învețe parametrii γ și r . Indicați ce modificări trebuie făcute în rețea și, corespunzător, în algoritm (general) de retro-propagare, pentru a învăța acești doi parametri.

5.2.2 SVM cu margine “soft”

47. (C-SVM cu separator prin originea sistemului de coordonate: exemplificarea noțiunilor de bază)

CMU, 2017 fall, Nina Balcan, midterm, pr. 1.3

Fie următoarea afirmație / „propoziție“ (în sens matematic):

Atunci când *nu* folosim termen liber (engl., bias), problema de optimizare C-SVM are

forma primală:

$$\min_{w, \xi} \left(\frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i \right)$$

a. i. $y_i w \cdot x_i \geq 1 - \xi_i$ și $\xi_i \geq 0$ ($i = 1, \dots, m$)

forma duală:

$$\max_{\alpha \in \mathbb{R}^m} \left(\sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m y_i y_j \alpha_i \alpha_j x_i \cdot x_j \right)$$

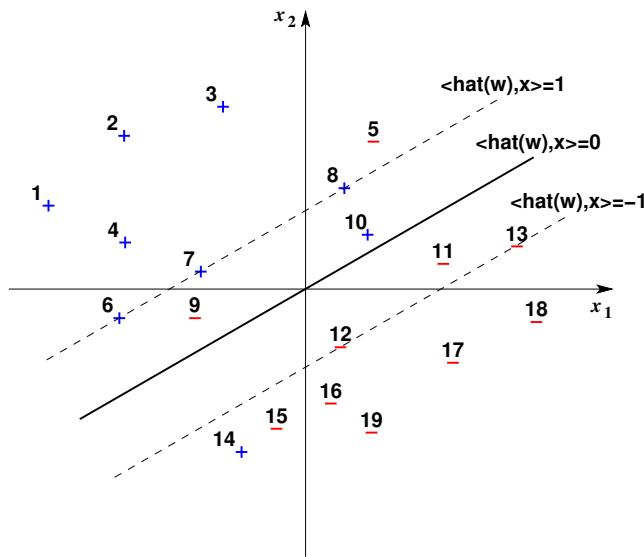
a. i. $0 \leq \alpha_i \leq C$ ($i = 1, \dots, m$).

Notăm cu $\hat{w}, \hat{\xi}$ soluția [optimă a] problemei primale și cu $\hat{\alpha}$ soluția [optimă a] problemei duale. (Vă readucem aminte că $\hat{\beta}_i \stackrel{not.}{=} C - \hat{\alpha}_i \geq 0$ este multiplicatorul Lagrange corespunzător inegalității $\xi_i \geq 0$.)

- a. Demonstrați că într-adevăr forma primală a problemei de optimizare C-SVM din enunț conduce la forma duală care a fost specificată. Puneți în evidență diferențele față de forma duală de la pr. 12.

În figura alăturată vi se dă un set de date de antrenament din \mathbb{R}^2 , precum și separatorul optimal învățat de către clasificatorul C-SVM pe acest set de date. În această figură, notația \hat{w} corespunde lui \hat{w} , iar $\langle \hat{w}, x \rangle$ desemnează produsul scalar $\hat{w} \cdot x$. Numărul marcat în dreptul fiecărei instanțe reprezintă indexul (sau, ID-ul) respectivului punct.

- b. Indicați toate instanțele pentru care $\xi_i \neq 0$.
- c. Indicați toți vectorii-suport. Vă readucem aminte că instanța x_i este vector-suport dacă $\hat{\alpha}_i \neq 0$.
- d. Considerăm următoarea propoziție: $\hat{\alpha}_i = C$ pentru punctele cu indicii 10 și 14. Adevărat sau Fals? Justificați.



48. (SVM vs. C-SVM: comparație între soluțiile celor două probleme pe seturi de date liniar separabile)

CMU, 2017 fall, Nina Balcan, HW4, pr. 4.Q11

Fie $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$ o mulțime de instanțe din \mathbb{R}^d , cu etichetele lăud valori în mulțimea $\{1, -1\}$. Presupunem că exemplele din S sunt liniar separabile și că rulăm C-SVM cu $C > 0$ pe setul de date S .

- a. Este oare separatorul decizional obținut de către clasificatorul C-SVM identic cu separatorul de margine maximă (adică, separatorul obținut de către SVM)? Justificați.
- b. Este oare adevărată afirmația că separatorul decizional obținut de către clasificatorul C-SVM separă întotdeauna în mod corect cele două clase (adică, obține eroare la antrenare 0)? Justificați.

49. (SVM și C-SVM: efectul adăugării unui atribut irelevant (i.e., care nu mărește marginea de separare) asupra vectorului de ponderi w)

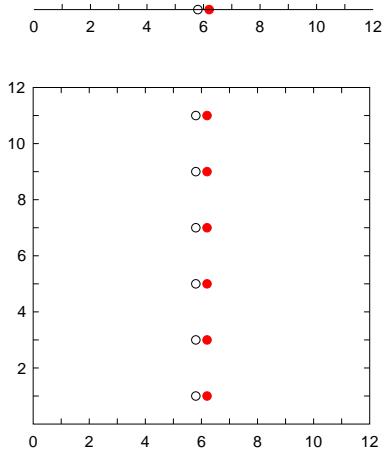
CMU, 2017 fall, Nina Balcan, midterm, pr. 1.4
CMU, 2007 spring, Carlos Guestrin, midterm exam, pr. 6.3

- a. Cazul SVM:⁷²²

⁷²²Se subînțelege, SVM liniară, cu margine "hard".

Fie $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$ un set de date din \mathbb{R}^d care sunt liniar separabile.⁷²³ Presupunem că adăugăm la fiecare instanță x_i un atribut $\tilde{x}_i \in \mathbb{R}$,⁷²⁴ care nu conduce la mărirea marginii [geometrice, de separare a] lui S . În acest context, spunem că atributul \tilde{x} este *irrelevant*. Va ignora oare SVM în mod automat acest atribut (adică, în vectorul soluție \hat{w} , componenta corespunzătoare noului atribut va fi 0)? Justificați răspunsul în mod riguros.

Sugestie: Pentru demonstrație, veți putea folosi proprietatea menționată la Observația de la pr. 2.a.



b. Cazul C-SVM:⁷²⁵

Presupunem că rulăm o C-SVM pe [un dataset S cu] attributele X_1, \dots, X_d , iar apoi adăugăm [la instanțele din S] un atribut X_{d+1} care este *irrelevant*, în sensul că [presupunând că păstrăm aceeași valoare pentru parametrul C] el nu poate crește marginea de separare. Va ignora oare C-SVM în mod automat acest atribut? Justificați răspunsul în mod riguros.

50. (C-SVM cu termeni de „destindere“ în norma L_2)

* University of Utah, 2008 fall, Hal Daumé III, HW5, pr. 3
CMU, 2010 fall, Aarti Singh, midterm exam, pr. 5.1

La curs am arătat că atunci când datele noastre de antrenament nu sunt liniar separabile, putem modifica forma problemei SVM introducând în fiecare restricție câte o variabilă de „destindere“ (engl., slack variable) în raport cu marginea.⁷²⁶ Mai precis, formularea pe care am dat-o este cunoscută sub denumirea *C-SVM de normă L_2* .

Acum vom considera o variantă cunoscută sub numele de *C-SVM de normă L_2* . Se pleacă de la următoarea problemă de optimizare:

$$\min_{w, w_0, \xi} \left(\frac{1}{2} \|w\|^2 + \frac{C}{2} \sum_{i=1}^m \xi_i^2 \right)$$

a.î. $y_i(w \cdot x_i + w_0) \geq 1 - \xi_i, i = 1, \dots, m.$

Remarcați faptul că variabilele de „destindere“ ξ_i apar la puterea a două în funcția obiectiv, cu scopul de a limita posibilitatea ca valorile lor să fie excesiv de mari.⁷²⁷

⁷²³LC: În enunțul original se precizează — în paranteze! — că separarea se [poate] face cu un separator care trece prin originea sistemului de coordinate. Restricția aceasta nu este neapărat necesară, deoarece [veți constata că] demonstrația poate fi făcută în cazul general al separabilității liniare.

⁷²⁴Se poate observa imediat că adăugarea acestui atribut nu afectează proprietatea de separabilitate liniară a datelor de antrenament S .

⁷²⁵Adică, SVM liniară cu margine “soft”.

⁷²⁶Am prezentat această chestiune la problema 12.

⁷²⁷Pentru conveniență calculelor, în locul lui C — factorul care precede suma $\sum_i \xi_i$ în problema 12 —, aici lucrăm cu $\frac{C}{2}$ ca factor pentru $\sum_i \xi_i^2$. A se vedea calculele de la punctele c și d.

a. În forma problemei de mai sus, nu apare condiția $\xi_i \geq 0$. Arătați că aceste restricții de nenegativitate pot fi într-adevăr eliminate. Așadar, vă cerem să demonstrați că valoarea optimă a funcției obiectiv va fi aceeași, indiferent dacă aceste restricții ($\xi_i \geq 0$) sunt sau nu incluse în formularea problemei.

Indicație: Puteți proceda prin reducere la absurd. Dacă ar exista o soluție optimă $(\bar{w}, \bar{w}_0, \bar{\xi})$ a problemei de optimizare C-SVM de normă L_2 astfel încât pentru un $i_0 \in \{1, \dots, m\}$ să avem $\xi_{i_0} < 0$, atunci se poate arăta (în mod concret / constructiv) că există o soluție mai bună decât soluția optimă pe care am considerat-o mai sus.

b. Care este *funcția lagrangeană generalizată* (notație: $L_P(w, w_0, \xi, \alpha)$) asociată problemei de optimizare C-SVM de normă L_2 ?

c. Deducreți condițiile de staționaritate / optimalitate Karush-Kuhn-Tucker pentru această problemă de optimizare convexă, calculând mai întâi derivatele parțiale $\frac{\partial L_P}{\partial w}$, $\frac{\partial L_P}{\partial w_0}$ și $\frac{\partial L_P}{\partial \xi}$, și egalându-le apoi cu 0. Am folosit notația $\xi = (\xi_1, \xi_2, \dots, \xi_m)$.

d. Care este *forma duală* a problemei de optimizare C-SVM de normă L_2 ? Prin ce diferă această formă duală de forma duală a problemei C-SVM (vedeți pr. 12)?

51. (C-SVM folosind funcție de cost *hinge*: exemplu de aplicare (adică, rezolvarea problemei de optimizare corespunzătoare); comparație cu regresia logistică, relativ la efectul outlier-elor asupra poziției separatorului optimal)
 • CMU, 2012 fall, E. Xing, A. Singh, HW2, pr. 4.3

La problema 20 am arătat că putem reformula problema de optimizare C-SVM liniară (adică, nekernel-izată) astfel încât să minimizăm suma costurilor *hinge* pe setul de date de antrenament, cu termen de regularizare $\|w\|^2$:

$$\min_{w, w_0} \left(\|w\|^2 + C \sum_i Loss_{SVM}(f(x_i), y_i) \right),$$

unde (x_i, y_i) sunt instanțe etichetate, cu $y_i \in \{-1, +1\}$, iar costurile *hinge* sunt definite în modul următor:

$$Loss_{SVM}(f(x_i), y_i) \stackrel{\text{def.}}{=} \max(1 - (w \cdot x_i + w_0)y_i, 0).$$

Vom considera setul de date $(x_1, y_1), \dots, (x_n, y_n)$ cu $n = 2000000$, unde pentru $i = 1, \dots, n/2$ avem $x_i = 0$ și $y_i = -1$, iar pentru $i = n/2 + 1, \dots, n$ avem $x_i = 2$ și $y_i = +1$. Cu alte cuvinte, se dă un milion de căpii ale instanței 0 (în spațiul unidimensional, desigur), toate etichetate cu -1 , și un milion de căpii ale instanței 2, toate etichetate cu $+1$.

a. Găsiți valorile lui w și w_0 care minimizează funcția obiectiv a problemei de optimizare C-SVM de mai sus pe aceste date, atunci când se lucrează cu $C = 1$.

Care este granița de decizie și cât este „marginea“ corespunzătoare? Cum va

diferi răspunsul dacă vom considera valori din ce în ce mai mici pentru C , tînzând la 0?

b. Acum vom presupune că pe lângă cele $n = 2000000$ de instanțe de mai sus ni se mai dă încă o instanță:

$$x_{n+1} = 100, y_{n+1} = -1.$$

Care sunt noile valori optimale pentru w și w_0 (folosind din nou $C = 1$)?

c. Considerând că la punctul b în locul algoritmului C-SVM folosim regresia logistică, precizați (raționând în mod intuitiv) care vor fi noile rezultate.

52.

(Algoritmul SMO pentru C-SVM:
o restricție la alegerea variabilelor libere, suficientă ca
algoritmul să nu poată îmbunătăți
soluția (α) fixată inițial)

• ○ * MIT, 2008 fall, Tommi Jaakkola, midterm exam, pr. 1.2

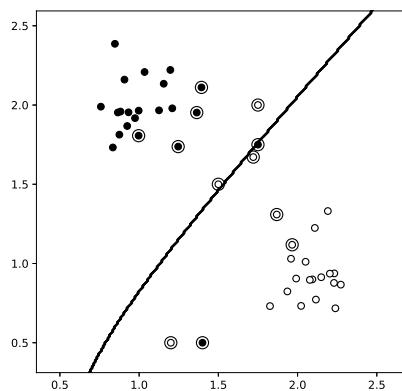
Considerăm algoritmul SMO pentru antrenarea unui C-SVM (adică, SVM cu variabile de „destindere“). Presupunem că inițial toate variabilele duale α_i ($i = 1, \dots, m$) sunt setate la valoarea 0. Apoi, la fiecare iterație a algoritmului alegem două variabile α_i și α_j impunând spre deosebire de [sau: pe lângă criteriul de selecție din] forma clasică a algoritmului SMO un criteriu simplu: cele două variabile trebuie să satisfacă restricția $y_i = y_j$. După selectare, optimizarea celor două variabile se face ca în algoritmul classic SMO.

Ce soluție va calcula această versiune a algoritmului SMO? Justificați răspunsul.

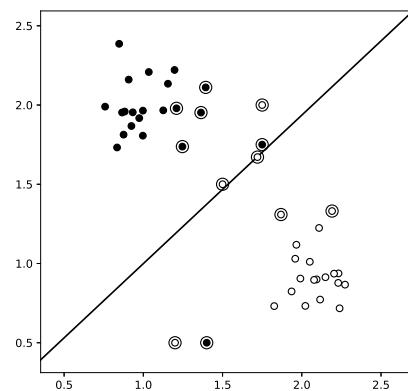
53.

(C-SVM: efectul alegerii
valorii parametrului C și / sau a funcției-nucleu)
○ CMU, 2012 spring, Ziv Bar-Joseph, HW3, pr. 3.1

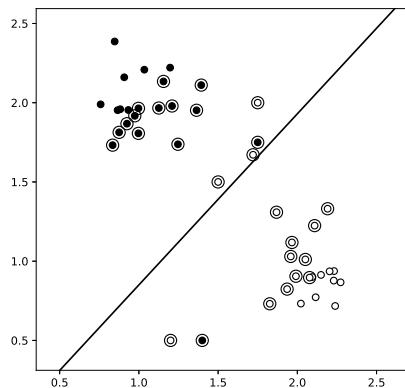
În figura de mai jos sunt ilustrate suprafețele de decizie pentru patru mașini cu vectori-suport cu margine „soft“ (adică, C-SVM-uri), care folosesc diferite funcții-nucleu și diferite valori pentru parametrul C pentru „destindere“.



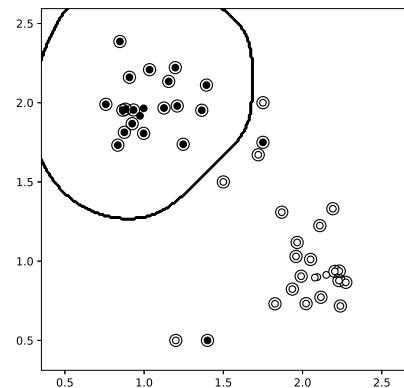
A.



B.



C.



D.

Pentru fiecare dintre aceste exemple, specificați ce setare credeți că a fost folosită.

- $C = 1$ și nicio funcție-nucleu;
- $C = 0.1$ și nicio funcție-nucleu;
- $C = 0.1$ și funcția-nucleu $K(x_i, x_j) = e^{-10\|x_i - x_j\|^2}$;
- $C = 0.1$ și funcția-nucleu $K(x_i, x_j) = x_i \cdot x_j + (x_i \cdot x_j)^2$.

54.

(C-SVM cu funcție-nucleu pătratică: determinarea graniței de decizie în funcție de valoarea parametrului pentru „destindere“ C)

◦ CMU, 2007 spring, Carlos Guestrin, midterm exam, pr. 2

Obiectivul acestei probleme este să clasificăm corect date de test, pornind de la un anumit set de date de antrenament.

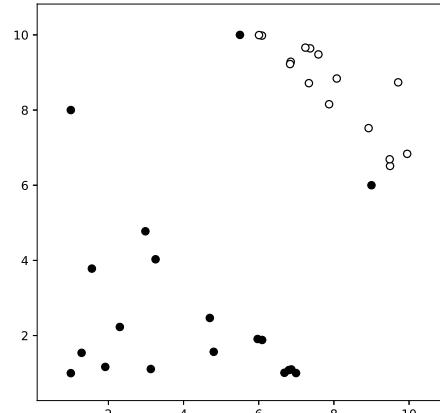
Vom presupune că antrenăm o C-SVM, folosind ca funcție-nucleu o funcție polinomială de gradul al doilea. Vi se dă setul de date de antrenament din figura alăturată.

Avertisment: Se consideră că datele de antrenament provin de la niște senzori care pot fi afectați de „zgomote“ / perturbații, deci ar trebui să evitați să vă încredeți prea mult în vreun punct anume.

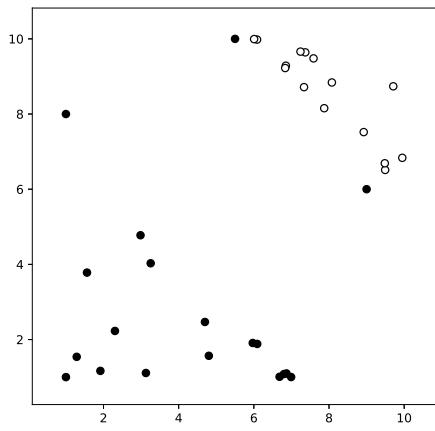
Penalizarea de „destindere“ (engl., slack penalty) C va determina situația hiperplanului de separare optimală.

Vi se cere să răspundeți la întrebările de mai jos în manieră *calitativă*. Pentru fiecare dintre aceste puncte, formulați răspunsul sub forma unei singure fraze. Veți reprezenta grafic soluțiile folosind în ordinea corespunzătoare figurile de la sfârșitul acestui enunț.

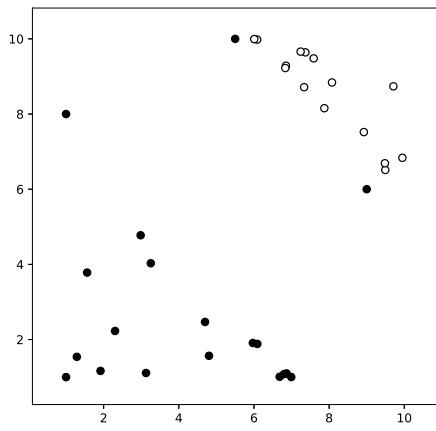
- Unde va fi situată granița de separare atunci când parametrul C ia valori mari (adică $C \rightarrow \infty$)? Desenați răspunsul în figura a.



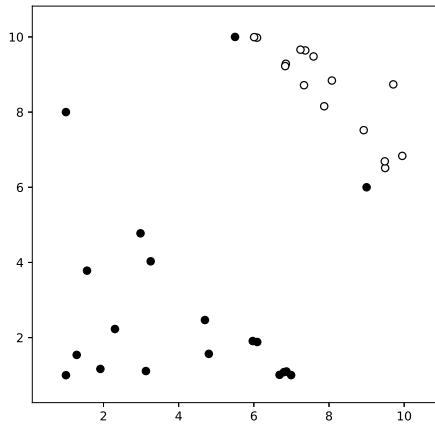
- b. Pentru $C \approx 0$, indicați în figura b unde anume credeți că va fi situată granița de decizie. Justificați răspunsul.
- c. Care dintre cele două cazuri de mai sus credeți că este mai adekvat pentru task-ul de generalizare / predicție? De ce?
- d. Desenați în figura d un punct care nu va schimba granița de decizie care a fost învățată pentru valori foarte mari ale lui C . Justificați răspunsul.
- e. Desenați în figura e un punct care va schimba în mod considerabil granița de decizie care a fost învățată pentru valori foarte mari ale lui C . Justificați răspunsul.



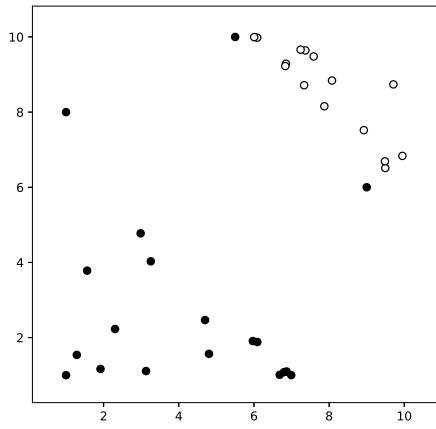
a.



b.



d.



e.

55.

(C-SVM cu funcții-nucleu: *complexitatea computațională la antrenare și respectiv la testare*)

• ○ CMU, 2014 fall, E. Xing, B. Poczos, HW2, pr. 3.6

Presupunem că avem m exemple de antrenament $(x_i, y_i)_{i=1}^m$ din $\mathbb{R}^d \times \{-1, +1\}$ și că dorim să antrenăm un C-SVM (adică SVM cu margine “soft”) care folosește funcție-nucleu. În multe cazuri practice, numărul m este foarte mare (de ordinul sutelor de milioane).

- a. Cât este *complexitatea de spațiu* dacă implementăm C-SVM cu funcție nucleu în manieră naivă? Veți da răspunsul presupunând că $m \gg d$.
- b. Cât este costul calculării *funcției de decizie* pentru o instanță oarecare x , dacă presupunem că există $\mathcal{O}(m)$ vectori-support, iar complexitatea de timp pentru calculul valorii funcției-nucleu $k(x, x')$ este $\mathcal{O}(d)$?
- c. Există mai multe modalități de a aproxima *funcția de decizie*. Citiți secțiunea de introducere din articolul *Fastfood – approximating kernel expansions in loglinear time*, de Quoc Le, Tamás Sarlós și Alex Smola⁷²⁸ și scrieți cât este complexitatea computațională a funcției de decizie în respectiva abordare.

56.

(Perceptronul Rosenblatt stochastic și [C-]SVM:
aplicare pe un dataset din \mathbb{R}^2 ; calcularea costului *hinge*)

prelucrare de Liviu Ciortuz, după

□ • · MIT, 2018 spring, Tommi Jaakkola, midterm review, pr. 1

A. Fie setul de date de antrenament care este prezentat în tabelul și în figura de mai jos:⁷²⁹

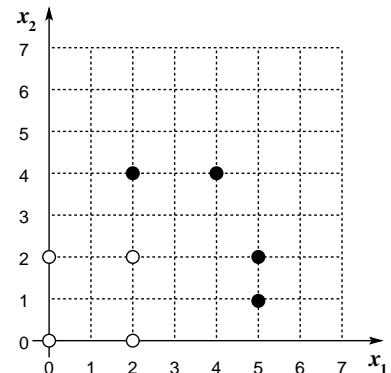
eticheta	-1	-1	-1	-1	+1	+1	+1	+1
coordonatele	(0, 0)	(2, 0)	(0, 2)	(2, 2)	(5, 1)	(5, 2)	(2, 4)	(4, 4)
nr. de greșeli făcute de Perceptronul Rosenblatt	1	6	1	5	3	1	2	0

Algoritmul Perceptron stochastic

```

initialize  $w \leftarrow \bar{0}$ ,  $w_0 \leftarrow 0$ 
for  $t=1, \dots, T$ 
  (or: do until a certain stopping condition is met)
    get the example  $(x_t, y_t)$ 
    if  $y_t (w \cdot x_t + w_0) \leq 0$  then
       $w \leftarrow w + y_t x_t$ ;
       $w_0 \leftarrow w_0 + y_t$ 
    end if
  end for
end do

```



- a. Rulăm Perceptronul Rosenblatt stochastic pe acest dataset, parcurgând exemplele de antrenament într-o ordine aleatoare până când condiția de oprire este satisfăcută.⁷³⁰ Numărul de greșeli făcute de Perceptron pe fiecare dintre aceste exemple este menționat pe ultima linie din tabel. Cât va fi valoarea rezultată pentru termenul liber (engl., offset parameter) w_0 ?
- b. Greșelile făcute de Perceptron depind de ordinea în care sunt parcurse exemplele de antrenament. Este oare posibil ca instanța (4, 4), etichetată cu '+ 1', să fi fost primul exemplu procesat de către Perceptron?

⁷²⁸Proceedings of the 30th International Conference on Machine Learning (ICML), 2013, pp 244–252.

⁷²⁹Vă reamintim convenția noastră de notare: simbolul \bullet desemnează instanțe pozitive, iar simbolul \circ instanțe negative.

⁷³⁰Pentru o prezentare a algoritmului Perceptron [neciclic, nestochastic], vedeți problema 16 de la capitolul *Rețele neuronale artificiale*. Pentru o versiune kernelizată a variantei stochastice a algoritmului Perceptron, vedeți problema 87 de la capitolul de *Fundamente*.

B. Presupunem acum că în loc să rulăm algoritmul Perceptron determinăm separatorul liniar care maximizează *marginea geometrică*.

c. Cât sunt valorile parametrilor \bar{w} și \bar{w}_0 corespunzătoare separatorului optim?

d. Cât este valoarea marginii geometrice?

e. Cât este suma costurilor de tip *hinge* (engl., hinge losses) determinate de aceste exemple?

Indicație (1): Vă readucem aminte că funcția de cost *hinge* este definită astfel: $hinge(z) = \max(0, 1 - z)$, iar la calculul costului hinge determinat de un exemplu oarecare (x_i, y_i) în raport cu separatorul optimal $\bar{w} \cdot x + \bar{w}_0 = 0$, acest z va fi înlocuit cu $y_i(\bar{w} \cdot x_i + \bar{w}_0)$.

f. Presupunem că modificăm soluția corespunzătoare marginii geometrice maxime, împărțind atât \bar{w} cât și \bar{w}_0 la 2. Cât devine acum suma costurilor *hinge* determinate de exemplele de antrenament în raport cu acest nou separator?

g. Dați o interpretare geometrică pentru suma costurilor *hinge* pe care ați calculat-o la punctul f (adică, suma costurilor *hinge* determinate de exemplele de antrenament în raport nou separator).

Indicație (2): La problema 20 am stabilit o corespondență între costurile *hinge* și variabilele ecart ξ_i , iar la problema 13 am demonstrat următorul rezultat:

$$\bar{\xi}_i > 0 \Rightarrow \frac{\bar{\xi}_i}{\|\bar{w}\|} = d(x_i, \bar{w} \cdot x + \bar{w}_0 = y_i).$$

57. (Clasificatori liniari — [C-]SVM, Perceptronul etc. — și versiunile lor kernel-izate: avantaje și dezavantaje)

• CMU, 2017 fall, Nina Balcan, HW3, pr. 3.3.2/Q17

Care credeți că sunt avantajele și dezavantajele (engl., pros and cons) folosirii modelelor liniare precum SVM, Perceptronul etc., precum și versiunile lor kernel-izate?

58. (O comparație între algoritmii 1-NN și C-SVM: efectul atributelor irelevante pentru clasificare)

* CMU, 2007 spring, Carlos Guestrin, midterm exam, pr. 6.1-2

Convenție: În tot acest exercițiu, prin SVM se va înțelege o mașină cu vectori-suport cu margine “soft” (adică, C-SVM), fără funcție-nucleu (cea ce înseamnă că lucrăm cu separator liniar).

a. Alcătuiți un set de date din \mathbb{R}^2 astfel încât pe acest set algoritmul 1-NN să producă eroare la cross-validation de tip “leave one out” (CVLOO) mai mică decât SVM.

b. Alcătuiți un alt set de date din \mathbb{R}^2 astfel încât pe noul dataset eroarea de tip CVLOO produsă de algoritmul 1-NN să fie mai mare decât cea produsă de SVM.

c. Acum veți genera două dataset-uri care să pună în evidență caracterul robust al algoritmului SVM în raport cu *atributele irelevante*.⁷³¹ Veți crea un set de date din \mathbb{R}^2 (câte unul pentru fiecare dintre cele două probleme date mai jos) cu atributele X_1 și X_2 — dintre care X_2 va fi atributul irelevant —, astfel încât:

- dacă se folosește doar atributul X_1 , eroarea la CVLOO produsă de algoritmul 1-NN este mai mică decât cea produsă de SVM,
- dacă se folosesc atributele X_1 și X_2 , eroarea CVLOO produsă de SVM nu se schimbă (în raport cu cazul de mai sus), însă eroarea CVLOO produsă de 1-NN crește în mod semnificativ.

59.

(C-SVM: Adevărat sau Fals?)

• o * CMU, 2010 fall, Aarti Singh, HW3, pr. 3.2

Presupunem că se lucrează cu o mașină cu vectori-suport cu margine “soft” (C-SVM), pe un anumit set de exemple, fără a folosi vreo funcție de „mapare“ a trăsăturilor. Pe măsură ce valoarea parametrului de „destindere“ C crește (pornind de la o anumită valoare de start),

- a. mai multe instanțe de antrenament vor fi clasificate eronat;
- b. marginea — adică $\frac{1}{\|w\|}$, distanța de la hiperplanul de separare optimală la instanțele (vectorii-suport) x_i pentru care $(w \cdot x_i + w_0)y_i = 1$, unde y_i este eticheta asociată instanței x_i — nu va crește (adică fie descrește fie rămâne aceeași).

5.2.3 Alte probleme de optimizare de tip SVM

60.

(Câteva probleme simple de optimizare de tip [C-]SVM)

• o * MIT, 2016 fall, R. Barzilay, S. Sra, Weekly Exercises, week 4, pr. 10

După cum ati observat, profesorului de la cursul de Învățare automată ii place să pună în evidență diverse variante de metode de clasificare automată de tip SVM.

În figurile următoare (notate cu 1., 2., 3. și 4.) sunt înfățișate atât granițele de decizie (engl., decision boundaries) cât și vectorii-suport (care, după cum vedeti, sunt încercuiți) pentru câteva metode de tip SVM aplicate pe un același set de date din \mathbb{R}^2 . În toate cazurile, granițele de decizie corespund unei ecuații de forma $\hat{w} \cdot x + \hat{w}_0 = 0$, unde se consideră în mod implicit că $\hat{w}_0 = 0$, cu excepția cazurilor când w_0 este inclus în mod explicit în formularea analitică a metodei de antrenare. Simbolii J_+ și J_- din aceste formulări analitice desemnează indicii instanțelor de antrenament pozitive (reprezentate cu •) și respectiv negative (reprezentate cu ○).

⁷³¹Adică, acele atrbute care, atunci când sunt eliminate / adăugate, n-ar trebui să schimbe rezultatul clasificării.

În total sunt date aici cinci metode de clasificare de tip SVM și patru figuri/grafice. În aceste figuri, dreptele corespunzătoare separatorilor optimali $\hat{w} \cdot x + \hat{w}_0 = 0$ sunt reprezentate prin linii îngroșate. „Marginile“ / dreptele de ecuație $\hat{w} \cdot x + \hat{w}_0 = -1$ și respectiv $\hat{w} \cdot x + \hat{w}_0 = +1$ sunt reprezentate prin linii simple (neîngroșate).

Pentru fiecare dintre aceste cinci metode de clasificare date mai jos, indicați toate(!) figurile care constituie o posibilă soluție pentru metoda respectivă. Justificați riguros.

Atenție! Este posibil ca anumitor metode de clasificare [din cele cinci] să le corespundă mai mult decât o [singură] figură.

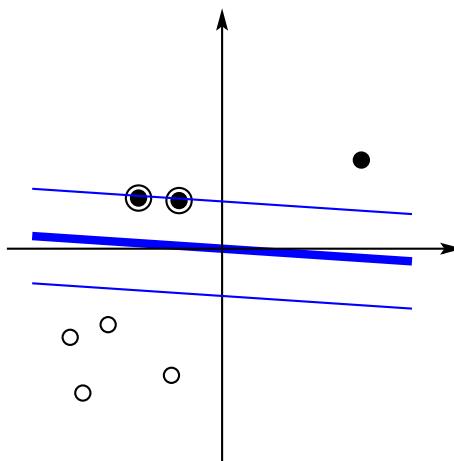
a. $\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i$, a.i. $\xi_i \geq 0$, $y_i(w \cdot x_i + w_0) \geq 1 - \xi_i$, $i = 1, \dots, n$, cu $C = \infty$.

b. $\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i$, a.i. $\xi_i \geq 0$, $y_i(w \cdot x_i) \geq 1 - \xi_i$, $i = 1, \dots, n$, cu $C = \infty$.

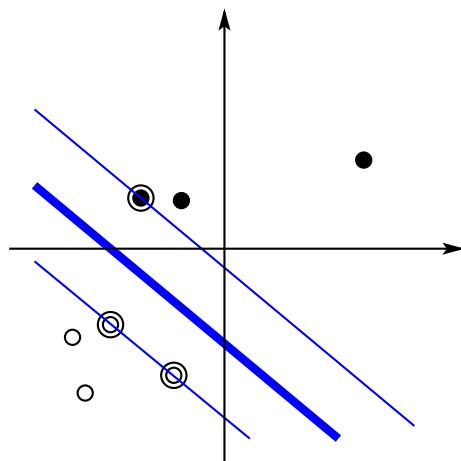
c. $\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i$, a.i. $\xi_i \geq 0$, $y_i(w \cdot x_i) \geq 1 - \xi_i$, $i = 1, \dots, n$, cu $C = 1$.

d. $\min \frac{1}{2} \|w\|^2 + C_+ \sum_{i \in J_+} \xi_i + C_- \sum_{i \in J_-} \xi_i$, a.i. $\xi_i \geq 0$, $y_i(w \cdot x_i) \geq 1 - \xi_i$, $i = 1, \dots, n$, cu $C_+ = 1$ și $C_- = 0$.

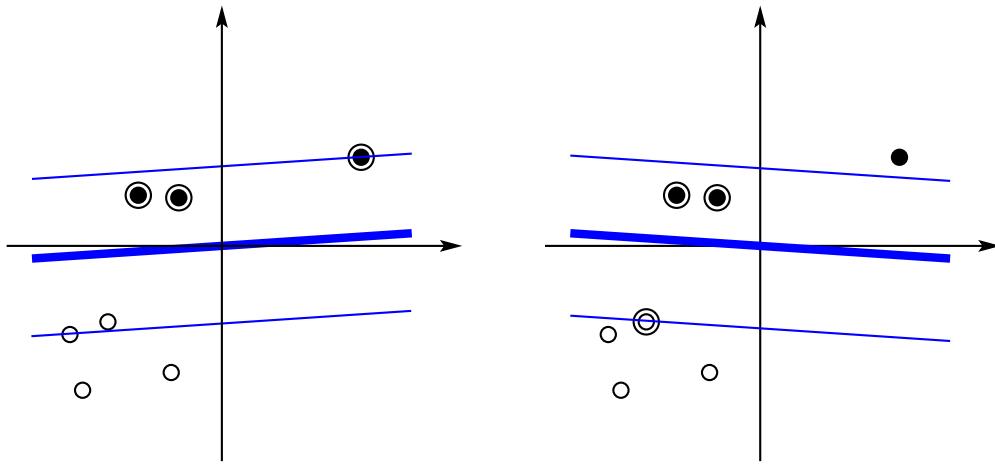
e. $\min \frac{1}{2} \|w\|^2 + C_+ \sum_{i \in J_+} \xi_i + C_- \sum_{i \in J_-} \xi_i$, a.i. $\xi_i \geq 0$, $y_i(w \cdot x_i) \geq 1 - \xi_i$, $i = 1, \dots, n$, cu $C_+ = \infty$ și $C_- = 0$.



1.



2.



61. (SVM multiclass cu margine “soft”: verificarea echivalenței cu C-SVM în cazul clasificării binare)
 • MIT, 2016 spring, David Sontag, HW2, pr. 4

Clasificatorul *SVM multiclass* constituie o generalizare a clasificatorului binar SVM de la două clase la un număr oarecare de clase, $K \geq 2$. După cum am arătat la problema 30 (unde am prezentat varianta SVM-multiclass cu margine “hard”), aceasta implică introducerea unui vector de ponderi $w^{(k)}$ și a unui termen liber $b^{(k)}$ pentru fiecare clasă $k \in \{1, \dots, K\}$.⁷³²

Ca și la alte variante ale clasificatorului [C-]SVM, și în cazul de față învățarea / antrenarea constă în rezolvarea unei *probleme de optimizare*:

$$\min_{w^{(k)}, b^{(k)}, \xi} \left(\frac{1}{2} \|w^{(k)}\|^2 + C \sum_i \xi_i \right)$$

a. i. $w^{(y_i)} \cdot x_i + b^{(y_i)} \geq w^{(k)} \cdot x_i + b^{(k)} + 1 - \xi_i$, pentru $i = 1, \dots, m$ și $k \neq y_i$
 $\xi_i \geq 0, \forall i = 1, \dots, m$.

Observați că în această problemă, pentru fiecare instanță de antrenament x_i apare — ca și în cazul problemei C-SVM, veДЕti pr. 12 — câte o variabilă de destindere (engl., slack variable) ξ_i , însă aici asociem la fiecare instanță [un număr de] $K - 1$ restricții.

Pentru o instanță nouă x , predicția se va face folosind regula următoare:

$$y = \arg \max_k (w^{(k)} \cdot x + b^{(k)}). \quad (326)$$

În acest exercițiu veți compara această regulă de predicție (de tip multiclass) în cazul particular în care se lucrează cu $K = 2$ cu regula de predicție pe care am folosit-o pentru clasificatorul [binar] C-SVM (veДЕti pr. 12.f):

$$\text{sign}(w \cdot x + b). \quad (327)$$

⁷³²La problema 30, din considerente care țin (doar) de simplitatea formulării, nu s-au folosit termeni liberi. Adăugarea lor se poate face în mod natural.

Concret, veți arăta că fiecare dintre cele două reguli se reduce la [adică, este echivalentă cu] cealaltă.

a. Calculați w și w_0 în funcție de $w^{(1)}, b^{(1)}, w^{(2)}$ și $b^{(2)}$ astfel încât pentru fiecare instanță x rezultatul obținut folosind noua regulă de predicție binară (327) să fie același cu rezultatul care se obține dacă se aplică regula de predicție multiclass (326) care folosește $w^{(1)}, b^{(1)}, w^{(2)}$ și $b^{(2)}$.

b. La acest punct veți proceda exact invers față de punctul precedent: Considerând w și w_0 ca fiind date, calculați $w^{(1)}, b^{(1)}, w^{(2)}$ și $b^{(2)}$ (în funcție de w și w_0) astfel încât pentru fiecare instanță x predicția făcută cu ajutorul regulii de predicție multiclass (cu $K = 2$) să fie același cu rezultatul care s-ar obține dacă am folosi regula de predicție binară cu respectivele valori (date) pentru w și w_0 .

62.

(O legătură între *one-class SVM* (versiunea *Max Margin*) și SVM (cu și respectiv fără termen liber (engl., bias))

*prelucrare de Liviu Ciortuz, după
□ • ○ MIT, 2009 fall, Tommi Jaakkola, midterm, pr. 3*

Un prieten ne-a spus că el poate să emuleze clasificatorii de tip SVM folosind doar cod care a fost dezvoltat pentru detecția „anomalilor“, mai precis algoritmul *one-class SVM*, versiunea *Max Margin*, pe care l-am prezentat la pr. 31. Rutinele de antrenare și respectiv testare despre care el afirmează că sunt suficiente pentru acest scop sunt următoarele:⁷³³

$$(\hat{w}, \hat{\rho}) = \text{train}(\phi_1, \dots, \phi_m) : \begin{aligned} &\text{Minimize}_{w, \rho} \left(\frac{1}{2} \|w\|^2 - \rho \right) \\ &\text{subject to } w \cdot \phi_i \geq \rho, \quad i = 1, \dots, m \\ &\text{Return } \hat{w}, \hat{\rho} \end{aligned}$$

$$y = \text{test}(w, \rho, \phi) : \text{Return } +1 \text{ if } w \cdot \phi \geq \rho \text{ else return } -1$$

unde ϕ_1, \dots, ϕ_m și ϕ sunt vectori de trăsături.

Noi n-am fost siguri dacă el are dreptate sau nu, așa că ne-am decis să verificăm.

Să începem prin a găsi clasificatorul SVM fără termen liber (engl., bias) de forma

$$\hat{y} = \text{sign}(w \cdot \phi(x))$$

corespunzător unui set de m instanțe de antrenament $\phi(x_1), \dots, \phi(x_m)$ având etichetele $y_1, \dots, y_m \in \{+1, -1\}$. Presupunem că setul de date de antrenament este liniar separabil.

- a. Precizați care sunt vectorii de trăsături pe care ar trebui să-i transmitem [pentru antrenare] rutinei train?
- b. Fie \hat{w} și $\hat{\rho}$ parametrii returnați de către rutina train în urma alegerii făcute la punctul precedent pentru vectorii de trăsături [pentru antrenare]. Care

⁷³³Vedeți *Observația* (2) de la pr. 31.

sunt argumentele pe care ar trebui să le transmitem rutinei test astfel încât ea să clasifice instanța de test $\phi(x)$ identic cu clasificatorul de margine maximă?

c. Cât este — în funcție de \hat{w} și $\hat{\rho}$ — marginea geometrică pe care o obține clasificatorul SVM pe setul de date de antrenament?

d. Încurajați de aceste rezultate, ne punem întrebarea dacă nu cumva ar fi posibil să antrenăm și clasificatorul SVM care include și termenul liber, adică $\hat{y} = \text{sign}(w \cdot \phi(x) + w_0)$, folosind [doar] cele două rutine. Este într-adevăr posibil? Justificați riguros.

63. (Rezolvarea problemei *one-class*, varianta *Max Margin*, cu margine “soft”, folosind abordarea de la ν -SVM)

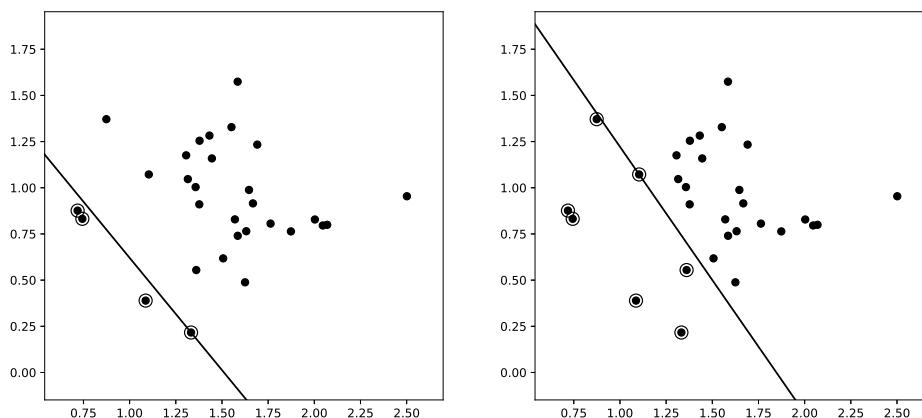
*prelucrare de Liviu Ciortuz, după
■ * MIT, 2009 fall, Tommi Jaakkola, ML course, lecture notes 5*

Vom reveni aici asupra problemei de optimizare *SVM one-class* varianta *Max Margin* (vedeți pr. 31). Mai întâi precizăm că, în ceea ce privește versiunea cu margine “hard”, față de forma primală pe care am considerat-o în problema menționată, acum vom lucra cu o versiune mai generală:⁷³⁴

$$\min_{w,\rho} \left(\frac{1}{2} \|w\|^2 - \rho \right)$$

a. i. $w \cdot x_i \geq \rho$, pentru $i = 1, \dots, m$.

Forma aceasta este mai convenabilă pentru *obiectivul* pe care ni-l fixăm aici, acela de a elabora cazul marginii “soft” pentru problema *one-class SVM* (*Max Margin*) urmând abordarea de tip ν -SVM. (A se vedea problema 33.)



Vă reamintim că ν -SVM folosește un parametru numeric (ν) care va funcționa ca margine superioară pentru proporția de erori la antrenare din totalul instanțelor de antrenament.

⁷³⁴Distanța de la hiperplanul de separare optimală la vectorii-suport care nu produc erori în raport cu marginea va fi $\frac{\rho}{\|w\|}$. Din punctul de vedere al „marginii“ geometrice, noi am vrea ca [și] ρ să fie maximizat. Aceasta echivalează cu a minimiza $-\rho$. Așa se justifică (în mod intuitiv) expresia funcției obiectiv din problema de optimizare pe care urmează să o formulăm.

Forma primală pe care o vom considera aici pentru problema ν -SVM *one-class* (Max Margin) este de asemenea ușor schimbăț în raport cu formularea originală a problemei ν -SVM (B. Schölkopf, A. Smola, R. Williamson și P. Bartlett, *New Support Vector Machines*, 2000):⁷³⁵

$$\begin{aligned} \min_{w, \xi, \rho} & \left(\frac{1}{2} \|w\|^2 - \rho + \frac{1}{\nu m} \sum_{i=1}^m \xi_i \right) \\ \text{a. i. } & w \cdot x_i \geq \rho - \xi_i, \text{ pentru } i = 1, \dots, m \\ & \xi_i \geq 0 \text{ pentru } i = 1, \dots, m. \end{aligned} \quad (\mathbf{P}'')$$

Remarcați faptul că în ambele probleme de optimizare date mai sus n-au fost impuse restricții asupra variabilei ρ .⁷³⁶

a. Demonstrați că forma duală corespunzătoare formei primale (\mathbf{P}'') a problemei ν -SVM *one-class* (Max Margin) este următoarea:

$$\begin{aligned} \max_{\alpha} & \left(-\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j x_i \cdot x_j \right) \\ \text{a. i. } & 0 \leq \alpha_i \leq \frac{1}{\nu m} \text{ pentru } i = 1, \dots, m \\ & \sum_{i=1}^m \alpha_i = 1. \end{aligned} \quad (\mathbf{D}'')$$

b. Scrieți condițiile de complementaritate Karush-Kuhn-Tucker pentru problema primală (\mathbf{P}''), apoi arătați cum anume se poate calcula $\bar{\rho}$, valoarea rezultată pentru variabila ρ la rezolvarea problemei duale (\mathbf{D}''), în funcție de valorile celorlalte variabile ($\bar{w}, \bar{\alpha}_i, \bar{\xi}_i$, etc).

64. (Problema [*one-class SVM*, varianta] *sferei de incluziune minimală*
(engl., minimum enclosing ball, MEB)
în varianta cu margine “soft”, folosind ν -SVM)

■ * MIT, 2009 fall, Tommi Jaakkola, ML course, lecture notes 5

Date fiind instanțele $x_1, \dots, x_m \in \mathbb{R}^d$, ne propunem să găsim o sferă care să includă toate punctele x_i și să aibă cea mai mică rază posibilă. Pentru aceasta, formulăm următoarea problemă de optimizare convexă:

$$\min_{R, w} R^2 \text{ astfel încât } \|w - x_i\|^2 \leq R^2 \text{ pentru } i = 1, \dots, m.$$

De remarcat faptul că restricțiile din formularea acestei probleme sunt de ordin pătratic (spre deosebire de cazul problemei SVM clasice, în care restricțiile sunt liniare). La finalul rezolvării acestei probleme,

- valoarea găsită pentru w va reprezenta centrul sferei;
- vectorii-suport vor fi punctele x_i de pe suprafața sferei; restricțiile corespunzătoare lor vor fi satisfăcute cu egalitate ($\|w - x_i\|^2 = R^2$).

Ca și în cazul problemei 63, putem impune condiția ca maximum νm puncte (din totalul celor m) să fie lăsate în afara sferei. Corespunzător, problema de

⁷³⁵A se vedea și problema 33.

⁷³⁶La problema 33 aveam $\rho \geq 0$.

optimizare convexă de tip ν -SVM va fi:

$$\begin{aligned} \min_{R,w,\xi} & \left(R^2 + \frac{1}{\nu m} \sum_{i=1}^m \xi_i \right) \\ \text{astfel încât} \quad & \|w - x_i\|^2 \leq R^2 + \xi_i \text{ pentru } i = 1, \dots, m \\ & \xi_i \geq 0 \text{ pentru } i = 1, \dots, m. \end{aligned} \tag{P^{iv}}$$

a. Demonstrați că forma duală a problemei de tip ν -SVM de mai sus este:

$$\begin{aligned} \max_{\alpha} & \left(- \sum_i \sum_j \alpha_i \alpha_j x_i \cdot x_j + \sum_{i=1}^m \alpha_i x_i \cdot x_i \right) \\ \text{a. i.} \quad & 0 \leq \alpha_i \leq \frac{1}{\nu m} \text{ pentru } i = 1, \dots, m \\ & \sum_{i=1}^m \alpha_i = 1 \end{aligned} \tag{D^{iv}}$$

unde α_i este multiplicatorul Lagrange corespunzător restricției $i \in \{1, \dots, m\}$.

b. Indicați relația de legătură dintre soluția \bar{w} a problemei primale și soluția $\bar{\alpha}$ a problemei duale. (Remarcați semnificația geometrică a rezultatului!)

c. Cum se poate calcula valoarea optimă \bar{R} (pentru variabila R din forma primală (P^{iv})) pornind de la soluția problemei duale?

Sugestie: Pentru aceasta, este util ca, pentru problema de tip ν -SVM de mai sus (D^{iv}), să enunțați condițiile de complementaritate Karush-Kuhn-Tucker.

65. (SVR cu margine “hard”, cazul liniar: exemplu de aplicare)

□ • ○ CMU, 2008 fall, Eric Xing, midterm, pr. 3

Pentru *regresie*, ni se dau m exemple de antrenament $(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)$, unde fiecare x_i este un vector de trăsături, iar $y_i \in \mathbb{R}$ este outputul pentru exemplul de pe poziția i ($i = 1, 2, \dots, m$).

Vă readucem aminte că în cazul *regresiei liniare cu vectori-suport* (engl., support vector regression, SVR),⁷³⁷ obiectivul este ca, pornind de la setul de exemple de antrenament, să învățăm o funcție liniară de forma $f(x) = w \cdot x + b$, unde w și b sunt parametrii care trebuie „învățați“, iar w este vectorul de ponderi. Problema de *optimizare convexă* pentru SVR liniară poate fi formulată în felul următor:

$$\min_w \frac{1}{2} \|w\|^2 \tag{328}$$

$$\text{a. i. } y_i - (w^\top x_i + b) \leq \varepsilon \text{ și } (w \cdot x_i + b) - y_i \leq \varepsilon \quad (i = 1, \dots, m). \tag{329}$$

Ideea din spatele relațiilor (328) și (329) este următoarea: dorim să „învățăm“ parametrii w și b astfel încât (i) funcția f obținută să fie cât mai „netedă“ (engl., smooth) cu putință (așadar, urmărim ca $\|w\|$ să fie cât mai mic), iar (ii) pentru fiecare exemplu de antrenament, eroarea în raport cu predicția făcută de către funcția f să fie de cel mult ε , unde $\varepsilon \geq 0$ este un *parametru* dat / fixat în acest algoritm.⁷³⁸

⁷³⁷Vedeți problema 34.

⁷³⁸Remarcați faptul că pe măsură ce valorile lui ε cresc, funcția f pe care o căutăm devine tot mai „netedă“.

Acum, date fiind 3 exemple de antrenament, (1, 1), (2, 2) și (3, 3), vrem să folosim relațiile (328) și (329) pentru a învăța un model liniar SVR.

- Cât este w atunci când $\varepsilon = 0$?
- Cât este w atunci când $\varepsilon = 0.5$?
- Cât este w atunci când $\varepsilon = 1$?

66.

(SVR cu margine “soft”: varianta care folosește funcție de cost / pierdere ε -senzitivă)

■ □ • ○ CMU, 2014 spring, B. Poczos, A. Singh, HW2, pr. 1
CMU, 2015 fall, Z. Bar-Joseph, E. Xing, HW3, pr. 2.1-4

În acest exercițiu veți deduce forma duală a problemei de optimizare pentru regresie cu vectori-suport (SVR) cu margine “soft”, care folosește o funcție de cost ε -sensibilă (engl., epsilon-sensitive loss), dată de expresia

$$L_\varepsilon(x, y, f) = |y - f(x)|_\varepsilon \stackrel{\text{not.}}{=} \max(0, |y - f(x)| - \varepsilon), \quad (330)$$

unde x este inputul, y este outputul, iar $f(x) \stackrel{\text{def.}}{=} w \cdot x$ este funcția folosită pentru a predicție.⁷³⁹ Setul de date de antrenament este $(x_1, y_1), \dots, (x_n, y_n)$, unde $x_i \in \mathbb{R}^m$, $y_i \in \mathbb{R}$.⁷⁴⁰

Folosind această notație, funcția obiectiv pentru problema SVR este definită astfel:

$$\frac{1}{2} \|w\|^2 + C \sum_{i=1}^n L_\varepsilon(x_i, y_i, f),$$

unde $f(x) = w \cdot x$, iar $C > 0$ și $\varepsilon > 0$ sunt parametri.

a. Scrieți această problemă ca o problemă de optimizare pătratică cu restricții liniare, folosind variabile de „destindere“ (engl., slack variables). Această formă este numită *forma primală* a problemei de optimizare SVR cu margine “soft”.

Sugestie: Expresia funcției de cost (330) nu trebuie să vă jeneze. Practic va trebui să procedați în manieră inversă față de cum s-a procedat la problema 20, unde, pornind de la forma primală a problemei C-SVM, se obține o formă echivalentă care folosește funcția de cost *hinge*. Alternativ, puteți pleca de la forma primală a problemei SVR cu margine “hard” (vedeți pr. 34) și introduceți variabile de destindere ξ_i similar modului în care am procedat pentru a obține clasificatorul C-SVM din clasificatorul SVM (vedeți pr. 12).

b. Scrieți *funcția lagrangeană* pentru *forma primală* a problemei SVR pe care ati obținut-o la punctul precedent.

c. Folosind *condițiile Karush-Kuhn-Tucker*, deduceți *forma duală* a acestei probleme SVR.

⁷³⁹Notând $z = y - f(x)$, este foarte util — pentru a înțelege mai bine ceea ce urmează — să faceți graficul funcției $\max(0, |z| - \varepsilon)$.

⁷⁴⁰Remarcați faptul că funcția de cost / pierdere *hinge* pe care am folosit-o la problema 20 (în legătură cu C-SVM) este adecvată doar pentru clasificare, nu și pentru regresie.

- d. Am putea folosi solver-e [LC: adică programe de analiză numerică] de probleme de *optimizare pătratică* pentru a rezolva problema duală SVR dedusă mai sus?
- e. Cum ați putea defini *vectorii-suport* pentru această problemă SVR?
- f. Scrieți expresia care poate fi folosită pentru a face *predicția* etichetei pentru o instanță nouă, x .
- g. Este oare posibil să kernel-izăm acest algoritm?
- h. Formulați motivul pentru care în general rezolvăm problemele SVR și SVM în varianta duală, nu în cea primală.
- i. Ce se întâmplă atunci când modificăm valoarea parametrului ε ?
- j. Ce se întâmplă atunci când modificăm valoarea parametrului C ?

67. (Exercițiu recapitulativ: corespondența dintre diverse funcții de cost (respectiv diverse tipuri de granițe de decizie) și diferite metode de învățare automată)

• CMU, 2014 spring, B. Poczos, A. Singh, midterm, pr. 2.1-2

- a. Puneți în corespondență fiecare dintre metodele de învățare automată de mai jos (din coloana din stânga) cu una dintre funcțiile de cost / pierdere (engl., loss functions) din coloana din dreapta, și anume acea funcție de cost pe care respectiva metodă de învățare o minimizează în cazul cel mai frecvent:

regresia liniară	funcția de cost pătratică
regresia logistică	funcția de cost logistică
AdaBoost	funcția de cost exponențială
k -NN	funcția de cost 0 – 1
SVM	funcția de cost hinge

- b. Pentru fiecare dintre metodele de învățare automată de mai jos (din coloana din stânga), încercuiți tipul / tipurile de separator decizional (graniță de decizie) pe care respectiva metodă îl poate produce la efectuarea unui task de clasificare binară. În undele cazuri, este posibil ca mai multe opțiuni să fie corecte. Veți încercui toate opțiunile pe care le considerați corecte.

regresia logistică:	liniar
ID3 (cu atribute numerice continue):	liniar, combinație de separatori liniari ⁷⁴¹
AdaBoost:	liniar, combinație de separatori liniari, pătratic
clasificatorul Bayes Naiv gaussian:	liniar, pătratic
SVM (fără funcție-nucleu):	liniar

⁷⁴¹ Engl., piecewise linear.



© M. Romanică

6 Rețele neuronale artificiale

Sumar

Notiuni preliminare

- funcție matematică; compunere de funcții reale; calculul valorii unei funcții pentru anumite valori specificate pentru argumentele / variabilele ei;
- funcție prag (sau, treaptă), funcție liniară, funcție sigmoidală (sau, logistică), funcție sigmoidală generalizată; separabilitate liniară pentru o mulțime de puncte din \mathbb{R}^d ;
- ecuații asociate dreptelor în plan / planelor în spațiu / hiper-planelor în spațiul \mathbb{R}^d ; ecuația dreptei în plan care trece prin două puncte date; semnele asociate punctelor din semiplanele determinate de o dreaptă dată în plan;
- derivate ale funcțiilor elementare de variabilă reală; derivate parțiale
- vectori; operații cu vectori, în particular produsul scalar al vectorilor;
- metoda gradientului descendente (ca metoda de optimizare); avantaje și dezavantaje; ex. 80, ex. 165 și ex. 166 de la cap. de *Fundamente*; ex. 23;
- funcții de cost / pierdere: ex. 163 de la cap. de *Fundamente*.

Câteva notiuni specifice

- *unități* neuronale artificiale (sau, *neuroni* artificiali, *perceptroni*); tipuri de neuroni artificiali: neuroni-prag, liniari, sigmoidali; *componente* ale unui neuron artificial: input, componenta de sumare, componentă / funcția de activare, output; funcția matematică reprezentată / calculată de un neuron artificial;
- *rețea* neuronală artificială; rețele de tip feed-forward; *niveluri* / straturi de neuroni, niveluri ascunse, niveluri de ieșire; *ponderi* asociate conexiunilor dintr-o rețea neuronală artificială; funcția matematică reprezentată / calculată de o rețea neuronală artificială; *granițe și zone de decizie* determine de o rețea neuronală artificială; funcția de eroare / cost (engl., loss function).

Câteva proprietăți relative la *expresivitatea* rețelelor neuronale artificiale

- (P0) Toate cele trei tipuri de neuroni artificiali (prag, liniar, sigmoidal) produc *separatori liniari*. Consecință: Conceptul XOR nu poate fi reprezentat / învățat cu astfel de „dispozitive“ simple de clasificare.

- (P0') Rețelele neuronale artificiale pot determina granițe de decizie neliniare (și, în consecință, pot reprezenta concepte precum XOR).
- Observație:** Rețele de unități sigmoidale pot determina granițe de decizie curbilinii: ex. 8.
- (P1) Rețele de neuroni diferite (ca structură și / sau tipuri de unități) pot să calculeze o aceeași funcție: ex. 3 și ex. 1.c vs. ex. 2.
- (P1') Dată o topologie de rețea neuronală (i.e., graf de unități neuronale al căror tip este lăsat nespecificat), este posibil ca plasând în noduri unități de un anumit tip să putem reprezenta / calcula o anumită funcție, iar schimbând tipul unora dintre unități (sau al tuturor unităților), funcția respectivă să nu mai potă fi calculată: ex. 4 vs. ex. 33.⁷⁴²
- (P2) Orice unitate liniară situată pe un nivel ascuns poate fi „absorbită“ pe nivelul următor: ex. 32.
 - (P3) Orice funcție booleană poate fi reprezentată cu ajutorul unei rețele neuronale artificiale având doar două niveluri de perceptriони-prag: ex. 5.
 - (P4) Orice funcție definită pe un interval mărginit din \mathbb{R} , care este continuă în sens Lipschitz, poate fi aproximată oricără de bine cu ajutorul unei rețele neuronale care are un singur nivel ascuns: ex. 7.
 - Corespondențe cu regresia liniară, regresia logistică și boosting-ul (ex. 34), respectiv cu arborii de decizie (ex. 35).

Algoritmi de antrenare a neuronilor artificiali folosind metoda gradientului descendente

- algoritmul de antrenare a unității liniare: ex. 36; vedeti T. Mitchell, *Machine Learning*, p. 93, justificare: p. 91-92; convergența: p. 95; exemplu de aplicare: ex. 10; varianta incrementală a algoritmului de antrenare a unității liniare: cartea ML, p. 93-94; despre convergența acestei variante (ca aproximare a variantei precedente (“batch”)): cartea ML, p. 93 jos;
- algoritmul de antrenare a perceptronului-prag și convergența: cartea ML, p. 88-89; exemplu de aplicare: ex. 11;
- algoritmul de antrenare a perceptronului sigmoidal și justificarea sa teoretică: cartea ML, p. 95-97;
- algoritmul *Perceptron* al lui Rosenblatt; exemplu de aplicare: ex. 16, ex. 38;
- deducerea regulii de actualizare a ponderilor pentru tipuri particulare de perceptriони: ex. 12, ex. 25.a, ex. 37, ex. 13.a;
- o justificare probabilistă (gen ipoteză de tip *maximum likelihood*) pentru minimizarea sumei pătratelor erorilor [la deducerea regulii de antrenare] pentru perceptronul liniar: ex. 13.b;
- exemple de [folosire a unei] alte funcții de cost / pierdere / penalizare (engl., loss function) decât semisuma pătratelor erorilor: suma costurilor de tip sigmoidal, ex. 14 (pentru perceptronul liniar), o funcție de tip cross-entropie, ex. 15 (pentru perceptronul sigmoidal).

⁷⁴²Problemele 1.d și ex. 31 au în vedere o chestiune similară, însă pentru rețele cu topologii diferite: o anumită extensie a funcției XOR nu poate fi reprezentată pe rețele de neuroni-prag care au un singur nivel ascuns.

Perceptronul Rosenblatt și rezultate de *convergență*

- exemplu de aplicare [adică, învățare cu perceptronul Rosenblatt]: ex. 16.
- câteva *proprietați* simple ale perceptronului Rosenblatt: ex. 17.
- rezultate de convergență de tip “mistake bound” pentru [algoritmul de antrenare pentru] perceptronul-prag [în varianta] Rosenblatt: ex. 18; pentru perceptronul-prag (clasic): ex. 40; învățare online cu perceptronul-prag de tip Rosenblatt: ex. 39;
- *Perceptronul kernel-izat* [dual]: ex. 19; particularizare pentru cazul nucleului RBF: ex. 41. Perceptronul Rosenblatt, cu termen liber (engl., offset), kernel-izare: ex. 42. Clasificare ternară cu perceptronul Rosenblatt, varianta kernelizată: ex. 43. (Vedeți și ex. 40.c.)

Antrenarea rețelelor neuronale artificiale: algoritmul de *retro-propagare* pentru rețele feed-forward

- T. Mitchell, *Machine Learning*, p. 98: pseudo-cod pentru rețele cu unități de tip sigmoidal, cu 2 niveluri, dintre care unul ascuns; pentru deducerea regulilor de actualizare a ponderilor; în cazul mai general al rețelelor feed-forward (de unități sigmoidale) cu oricâte niveluri, vedeți p. 101-103; ex. 20: deducerea regulilor de actualizare a ponderilor în cazul rețelelor cu 2 niveluri, având însă unități cu funcție de activare oarecare (derivabilă);
- aplicare: ex. 21, ex. 45, ex. 46;
- prevenirea overfitting-ului:
folosirea unei componente de tip „moment“ în expresia regulilor de actualizare a ponderilor: ex. 48;
regularizare: introducerea unei componente suplimentare în funcția de optimizat: ex. 22;
- cazul folosirii unei funcții de activare de tip tangentă hiperbolică: ex. 47;
- cazul folosirii unei funcții de cost / penalizare / eroare de tip cross-entropie: ex. 50;
- execuția manuală a unei iterării a algoritmului de retro-propagare în cazul unei rețele neuronale simple, având un singur nivel ascuns, cu unități ce folosesc funcția de activare ReL: ex. 51.

Rețele neuronale profunde — câteva chestiuni introductive

- fenomenul de „dispariție“ a gradientului [în cazul aplicării algoritmului de retro-propagare] pentru rețele neuronale profunde (engl., deep neural networks) care folosesc funcția de activare sigmoidală: ex. 26;
- determinarea numărului de parametri și de conexiuni din rețeaua neuronală convolutivă LeNet: ex. 27;
- determinarea mărimii hărții de trăsături de pe un anumit nivel, precum și a numărului de operații în virgulă mobilă (FLOPs) executate la procesarea forward într-o rețea neuronală convolutivă: ex. 54.

6.1 Rețele neuronale artificiale — Probleme rezolvate

6.1.1 Chestiuni introductive

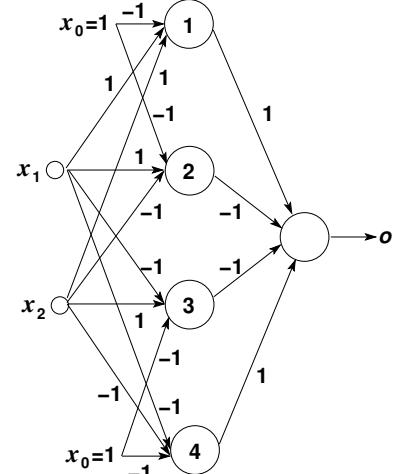
1.

(Rețele de perceptroni-prag, exemplificare: calculul outputului)

*prelucrare de Liviu Ciortuz, după
■ CMU, 2010 fall, Aarti Singh, HW5, pr. 4.1.1*

Considerăm rețeaua neuronală din figura alăturată. Toate unitățile acestei rețele neuronale sunt de tip prag, adică folosesc pentru activare funcția $sign$ definită prin $sign(z) = 1$ dacă $z \geq 0$ și -1 în rest. Pentru unitatea de pe nivelul de ieșire (lăsată nenumerotată), ponderea corespunzătoare termenului liber ($x_0 = 1$) este 0.

- Scriți funcția matematică calculată de fiecare dintre unitățile rețelei, în raport cu intrările x_1 și x_2 . Veți nota cu o_i (unde $i = 1, \dots, 4$) ieșirile unităților de pe nivelul ascuns și cu o ieșirea rețelei, adică valoarea produsă de către unitatea de pe nivelul de ieșire.
- Calculați outputul rețelei atunci când intrările x_1 și x_2 iau valori în mulțimea $\{-1, 1\}$.
- Indicați funcțiile booleene reprezentate de către uitățile de pe nivelul ascuns (1, 2, 3 și 4) atunci când $x_1, x_2 \in \{-1, 1\}$. Procedați similar pentru ieșirea o .
- Specificați cum anume ar putea fi modificată rețeaua dată astfel încât noua variantă să calculeze funcția de variabile reale (nu booleene ca mai înainte!) x_1 și x_2 , a cărei relație de definiție este: $f(x_1, x_2) = 1$ dacă $x_1, x_2 \geq 0$ sau $x_1, x_2 < 0$, și -1 în caz contrar.



Răspuns:

- Sunt imediate următoarele relații:

$$\begin{aligned}
 o_1(x_1, x_2) &= sign(x_1 + x_2 - 1), \\
 o_2(x_1, x_2) &= sign(x_1 - x_2 - 1), \\
 o_3(x_1, x_2) &= sign(-x_1 + x_2 - 1), \\
 o_4(x_1, x_2) &= sign(-x_1 - x_2 - 1), \\
 o(x_1, x_2) &= sign(o_1(x_1, x_2) - o_2(x_1, x_2) - o_3(x_1, x_2) + o_4(x_1, x_2)).
 \end{aligned}$$

- Date fiind formulele de la punctul precedent, calculele cerute sunt simple; centralizăm rezultatele sub forma tabelului următor:

x_1	x_2	o_1	o_2	o_3	o_4	o
1	1	1	-1	-1	-1	1
1	-1	-1	1	-1	-1	-1
-1	1	-1	-1	1	-1	-1
-1	-1	-1	-1	-1	1	1

c. Din tabelul obținut la punctul precedent este imediat că, atunci când $x_1, x_2 \in \{-1, 1\}$, ieșirile calculate de către unitățile 1, 2, 3 și 4 corespund conceptelor / funcțiilor $x_1 \wedge x_2$, $x_1 \wedge \neg x_2$, $\neg x_1 \wedge x_2$ și respectiv $\neg x_1 \wedge \neg x_2$. Ieșirea rețelei, o , corespunde conceptului $\neg(x_1 \text{ XOR } x_2)$.

Observație: Se poate remarcă faptul că în această rețea un neuron (și doar unul!) de pe nivelul ascuns este activat la fiecare combinație de valori (din cele 4 posibile) pentru $x_1, x_2 \in \{-1, 1\}$.

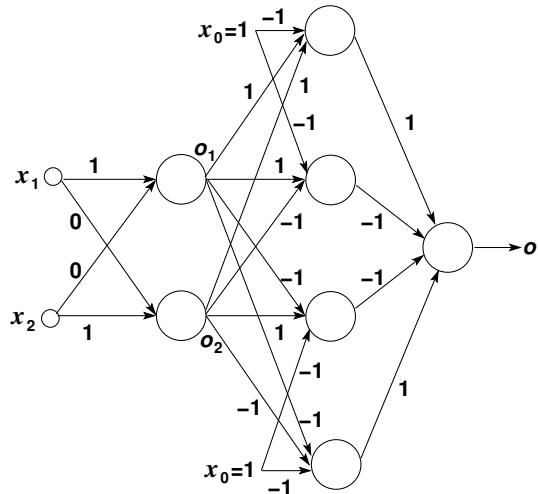
d. Este imediat că funcția indicată în enunț se poate scrie sub forma

$$f(x_1, x_2) = \begin{cases} 1 & \text{dacă } \text{sign}(x_1) \cdot \text{sign}(x_2) \geq 0 \\ -1 & \text{dacă } \text{sign}(x_1) \cdot \text{sign}(x_2) < 0. \end{cases}$$

Se observă imediat că aceasta coincide cu funcția $\neg(\text{sign}(x_1) \text{ XOR } \text{sign}(x_2))$.

Prin urmare, este suficient să adăugăm la rețeaua dată în enunț un nivel ascuns suplimentar, format din două unități cu funcție de activare de tip prag, care să transforme intrările x_1 și x_2 în $\text{sign}(x_1)$ și respectiv $\text{sign}(x_2)$. Vom obține ca rezultat rețeaua din figura alăturată.

Observație: La problema 31 vi se va cere să demonstrați că această funcție de variabile reale nu poate fi calculată de nicio rețea neuronală care are un singur nivel ascuns și este compusă doar din unități [cu funcție de activare] de tip prag.



2. (Reprezentarea unor funcții booleene cu ajutorul perceptronilor-prag sau al rețelelor de perceptoni-prag)

- Tom Mitchell, "Machine Learning", 1997, pr. 4.2
CMU, 1995 fall, Tom Mitchell, HW4, pr. 6

a. Concepți un perceptron care are

- funcția de activare de tip prag, cu valori -1 și +1,
- două intrări

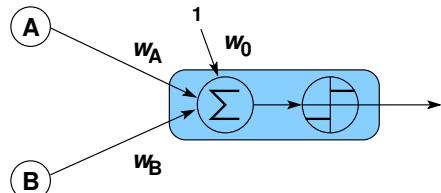
și implementează funcția booleană $A \wedge (\neg B)$.

b. Concepți o rețea neuronală formată din perceptri cu funcția de activare de tip prag dispusi pe două niveluri, care implementează funcția $A \text{ XOR } B$.

Răspuns:

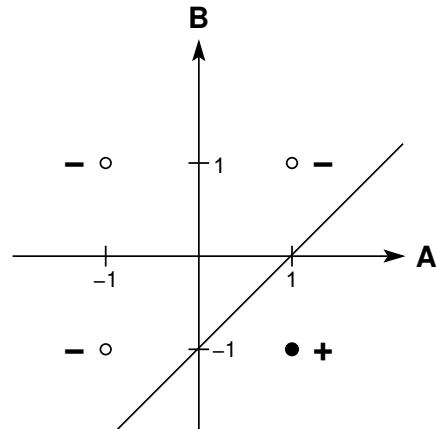
Observație: Deoarece mulțimea de valori de ieșire a perceptronului este $\{-1, 1\}$, vom alege această codificare a valorilor fals / adevărat în locul celei clasice 0/1.

a. Perceptronul-prag are structura din figura alăturată. A îndeplini cerința din enunț revine la a alege în mod convenabil valori pentru ponderile w_0 , w_A și w_B .⁷⁴³



Un perceptron consistent cu funcția booleană $A \wedge (\neg B)$ poate fi reprezentat ca o dreaptă ce separă instanțele pozitive de cele negative, ca în desenul alăturat. O astfel de dreaptă are ecuația $d(A, B) = 0$, unde $d(A, B) = w_0 + w_A A + w_B B$.

Este evident că sunt o infinitate de funcții care îndeplinesc proprietatea de separator liniar pentru cele două mulțimi, însă noi avem nevoie doar de una singură. Putem alege în mod convenabil / preferențial două puncte prin care să treacă dreapta, de pildă ca în figura alăturată. Așadar, analitic, considerăm d astfel încât $(1, 0) \in d$ și $(0, -1) \in d$.



O astfel de alegere va impune anumite *restricții* asupra valorilor w_0 , w_A și w_B . Într-adevăr, ținând cont de ecuația dreptei d scrisă mai sus (în sens generic), va rezulta următorul sistem:

$$\begin{cases} (1, 0) \in d \\ (0, -1) \in d \end{cases} \Rightarrow \begin{cases} w_0 + w_A \cdot 1 + w_B \cdot 0 = 0 \\ w_0 + w_A \cdot 0 + w_B \cdot (-1) = 0 \end{cases} \Rightarrow \begin{cases} w_0 = -w_A \\ w_0 = w_B \end{cases}$$

Audem deci $w_0 = w_B =^{\text{not.}} \alpha$, $w_A = -\alpha$, unde $\alpha \in \mathbb{R}^*$. Prin urmare, $d(A, B) = \alpha - \alpha A + \alpha B$.

Rămâne să mai analizăm în ce condiții dreapta d clasifică pozitiv instanța $(1, -1)$ și negativ instanțele $(1, 1)$ și $(-1, -1)$. De fapt, ținând cont de o proprietate din geometria analitică,⁷⁴⁴ este suficient să impunem una (de exemplu, prima) dintre aceste condiții. Aceasta revine la *restricția* ca expresia $\alpha - \alpha A + \alpha B$ să fie pozitivă pentru $A = 1$ și $B = -1$:

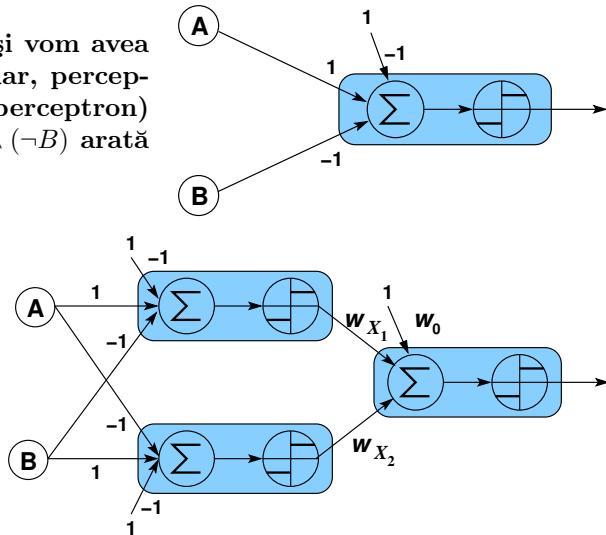
$$d(1, -1) > 0 \Leftrightarrow \alpha - \alpha - \alpha > 0 \Leftrightarrow \alpha < 0$$

⁷⁴³Am putea satisface direct această cerință dacă ne raportăm la problema 1.c, în care $o_2(x_1, x_2) = x_1 \wedge \neg x_2$, deci am putea seta $w_0 = -1$, $w_A = 1$ și $w_B = -1$. Totuși, aici vom proceda independent de problema 1, arătând cum anume se rezolvă [în general] un exercițiu de acest tip.

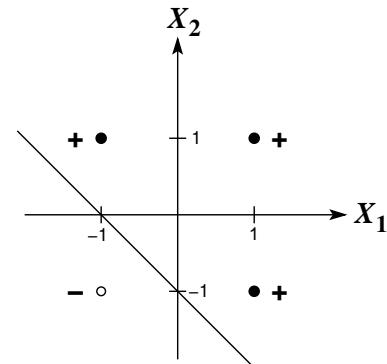
⁷⁴⁴Este vorba despre următoarea proprietate din geometria plană: Toate punctele situate de o parte a dreptei de ecuație $w_0 + w_A A + w_B B$ au același semn, în vreme ce punctele situate de pe celălaltă parte a dreptei au semn contrar.

Pentru fixare, alegem $\alpha = -1$, și vom avea $w_0 = -1, w_A = 1, w_B = -1$. Așadar, perceptronul (sau mai bine spus: un perceptron) care implementează funcția $A \wedge (\neg B)$ arată ca în figura alăturată.

b.⁷⁴⁵ Stim că $A \text{ XOR } B = (A \wedge (\neg B)) \vee (\neg A \wedge B)$. Această formulă poate fi reprezentată printr-o rețea de perceptri cu un nivel ascuns, ca în figura alăturată.

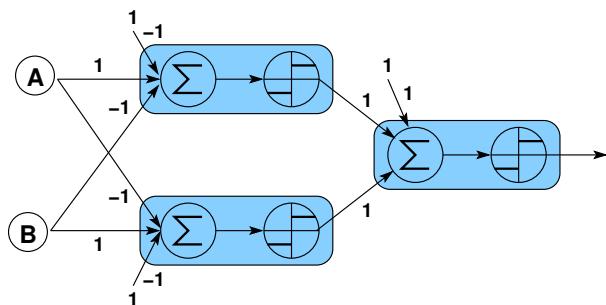


Perceptriile de pe nivelul de ascuns codifică funcția $A \wedge (\neg B)$ (vedeți punctul anterior) și respectiv $(\neg A) \wedge B$, după cum se poate vedea după ponderile de pe muchii, iar perceptronul de ieșire trebuie să descrie funcția logică \vee . Suprafața de decizie pentru aceasta din urmă este precum cea din figura alăturată. Avem $d'(X_1, X_2) = w_0 + w_{X_1}X_1 + w_{X_2}X_2 = 0$. Fie d' dreapta care trece prin punctele $(0, -1)$ și $(-1, 0)$. Atunci:



$$\begin{cases} (0, -1) \in d' \\ (-1, 0) \in d' \end{cases} \Rightarrow \begin{cases} w_0 - w_{X_2} = 0 \\ w_0 - w_{X_1} = 0 \end{cases} \Rightarrow w_0 = w_{X_1} = w_{X_2} \stackrel{\text{not.}}{=} \alpha' \in \mathbb{R}^* \Rightarrow d'(X_1, X_2) = \alpha' + \alpha'X_1 + \alpha'X_2 = 0.$$

Impunând condiția referitoare la semne, vom avea $d'(-1, -1) < 0 \Leftrightarrow \alpha' - \alpha' - \alpha' < 0 \Leftrightarrow \alpha' > 0$. Pentru fixare, alegem $\alpha' = 1$, ceea ce duce la $w_0 = w_{X_1} = w_{X_2} = 1$, iar rețeaua neuronală va arăta ca în figura alăturată.



⁷⁴⁵ Observație: Am putea obține o soluție imediată pentru acest punct al problemei noastre dacă preluăm rețeaua neuronală dată în enunțul problemei 1 — despre care stim (vedeți rezolvarea respectivei probleme) că are outputul $\neg(x_1 \text{XOR } x_2)$ — și schimbăm semnele tuturor ponderilor de pe arcele hidden-to-output din rețea. Vom arăta însă aici cum se poate construi „de la zero” o astfel de rețea, pornind de la cerințele specificate în enunț. În plus, se va vedea la final că noua rețea este mai simplă decât cea de la problema 1. (Așadar, rețele neuronale diferite pot codifica / reprezenta o aceeași funcție reală. Evident, este de dorit ca, pentru o funcție dată, rețeaua neuronală care o codifică să fie cât mai simplă.)

3.

(Rețele neuronale: exemplificare)

• CMU, 2011 spring, Roni Rosenfeld, HW4, pr. 1.c

Să se reprezinte expresia booleană $(A \vee \neg B) \text{ XOR } (\neg C \vee D)$ printr-o rețea neuronală cu două niveluri care este formată din neuroni cu funcție de activare de tip prag.

Răspuns:

Observație: Este imediat — vedeti rezolvările problemelor 1 și / sau 2 — că se poate defini câte un perceptron cu funcție de activare de tip prag (engl., threshold perceptron) care să reprezinte funcțiile $(A \vee \neg B)$ și respectiv $(\neg C \vee D)$. Dacă ieșirile acestor doi perceptri vor fi cuplate la intrările rețelei care reprezintă funcția $A \text{ XOR } B$ (a se vedea exercițiul 2, punctul b), atunci se va obține o rețea cu *trei* niveluri care reprezintă funcția $(A \vee \neg B) \text{ XOR } (\neg C \vee D)$. Exercițiul nostru cere însă să identificăm o rețea cu (doar!) *două* niveluri care să reprezinte această funcție.

Explicitând definiția funcției logice XOR și apoi aplicând regulile lui DeMorgan, expresia booleană din enunț devine:

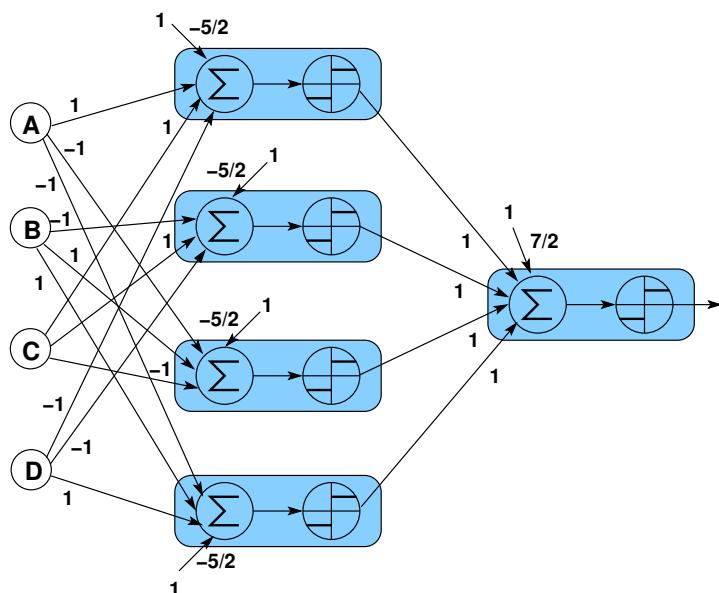
$$\begin{aligned}(A \vee \neg B) \text{ XOR } (\neg C \vee D) &= [(A \vee \neg B) \wedge \neg(\neg C \vee D)] \vee [\neg(A \vee \neg B) \wedge (\neg C \vee D)] \\ &= [(A \vee \neg B) \wedge (C \wedge \neg D)] \vee [(\neg A \wedge B) \wedge (\neg C \vee D)]\end{aligned}$$

Întrucât operatorii logici \vee și \wedge sunt distributivi unul față de celălalt, rezultă că:

$$(A \vee \neg B) \text{ XOR } (\neg C \vee D) = (A \wedge C \wedge \neg D) \vee (\neg B \wedge C \wedge \neg D) \vee (\neg A \wedge B \wedge \neg C) \vee (\neg A \wedge B \wedge D)$$

Fiecare din cele patru paranteze din partea dreaptă a egalității de mai sus poate fi reprezentată cu ajutorul unui perceptron-prag cu patru intrări, și anume câte o intrare pentru fiecare din cele trei variabile booleene din paranteză, plus o intrare pentru termenul liber. (De exemplu, conjuncției $A \wedge C \wedge \neg D$ îi putem asocia inegalitatea $x + z - t > 5/2$, deci ponderile intrărilor perceptronului corespunzător vor fi $-5/2$, 1, 1 și -1 .) Cei patru perceptri vor fi plasati pe primul nivel ascuns al rețelei pe care o construim.

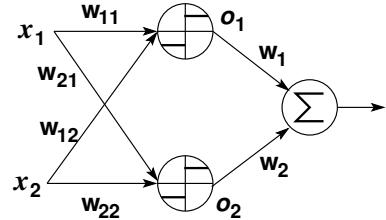
Apoi, ieșirile acestor patru perceptri vor constitui intrări pentru un alt perceptron-prag, care să reprezinte disjuncția a patru variabile booleene. (Acestui perceptron, situat pe nivelul de ieșire, îi putem asocia, de exemplu, inegalitatea $x + y + z + t > -7/2$.) Rețeaua neuronală rezultată este cea din figura alăturată.



4.

(Verificarea (im)posibilității de a reprezenta
funcția booleană XOR
cu o rețea de structură / compozitie specificată)
CMU, 2001 fall, Andrew Moore, final exam, pr. 10.a

Presupunem că lucrăm cu rețeaua neuronală din figura alăturată,⁷⁴⁶ care are un nivel ascuns format din doi perceptroni cu funcția de activare de tip prag: $sign(z) = 1$ dacă $z \geq 0$ și $sign(z) = -1$ dacă $z < 0$.



Folosim setul de date de antrenament din tabelul alăturat, care reprezintă funcția booleană *sau-exclusiv* dacă interpretăm -1 ca desemnând valoarea logică *fals*, și 1 ca *adevărat*.

X_1	X_2	Y
1	1	-1
1	-1	1
-1	1	1
-1	-1	-1

Vă cerem să indicați ce valori putem atribui ponderilor rețelei date, astfel încât să obținem eroare nulă la antrenare. Dacă este imposibil să se găsească astfel de valori pentru ponderile rețelei, dați răspunsul *Imposibil*.

Observație: În raport cu problema 2.b, remarcăți faptul că aici s-a eliminat termenul liber $x_0 = 1$ de la toți neuronii, precum și funcția de activare (de tip prag) de la neuronul de pe nivelul de ieșire.

Răspuns:

Funcțiile calculate de către cele trei unități neuronale sunt:

$$\begin{aligned} o_1(x_1, x_2) &= sign(w_{11}x_1 + w_{12}x_2) \\ o_2(x_1, x_2) &= sign(w_{21}x_1 + w_{22}x_2) \\ o(x_1, x_2) &= w_1 o_1(x_1, x_2) + w_2 o_2(x_1, x_2) \\ &= w_1 sign(w_{11}x_1 + w_{12}x_2) + w_2 sign(w_{21}x_1 + w_{22}x_2) \end{aligned}$$

Așadar,

$$o(1, 1) = w_1 sign(w_{11} + w_{12}) + w_2 sign(w_{21} + w_{22}) = -1 \quad (331)$$

$$o(1, -1) = w_1 sign(w_{11} - w_{12}) + w_2 sign(w_{21} - w_{22}) = 1 \quad (332)$$

$$o(-1, 1) = w_1 sign(-w_{11} + w_{12}) + w_2 sign(-w_{21} + w_{22}) = 1 \quad (333)$$

$$o(-1, -1) = w_1 sign(-w_{11} - w_{12}) + w_2 sign(-w_{21} - w_{22}) = -1 \quad (334)$$

Se poate arăta ușor că $sign(x) = -sign(-x)$ pentru orice $x \in \mathbb{R} \setminus \{0\}$. Așadar, dacă $-w_{11} + w_{12} \neq 0$ (ceea ce este echivalent cu $w_{11} \neq w_{12}$) și $-w_{21} + w_{22} \neq 0$ (echivalent cu $w_{21} \neq w_{22}$), relația (333) implica

$$w_1 sign(w_{11} - w_{12}) + w_2 sign(w_{21} - w_{22}) = -1,$$

⁷⁴⁶ *Observație importantă:* Pentru comoditate, vom opta aici — dar și în [multe dintre] problemele care urmează — pentru o reprezentare simplificată a unităților neuronale (comparativ cu reprezentările din problemele precedente). În figura dată, se va subînțelege că neuronii de pe nivelul ascuns au (ca întotdeauna) o componentă de sumare (care aici nu este pusă în evidență), iar unitatea de pe nivelul de ieșire nu are o componentă de activare (deci este unitate liniară).

ceea ce intră în contradicție evidentă cu relația (332) scrisă mai sus.

Similar, dacă $w_{11} \neq -w_{12}$ și $w_{21} \neq -w_{22}$, relația (334) implică

$$w_1 \text{sign}(w_{11} + w_{12}) + w_2 \text{sign}(w_{21} + w_{22}) = 1,$$

ceea ce contravine relației (331).

Sumarizând cele scrise mai sus, dacă ponderile w_{ij} , cu $i, j \in \{1, 2\}$ satisfac condiția multiplă

$$\begin{aligned} w_{11} &\neq w_{12} \text{ și } w_{21} \neq w_{22}, \\ \text{sau} \\ w_{11} &\neq -w_{12} \text{ și } w_{21} \neq -w_{22}, \end{aligned}$$

răspunsul la problema noastră este: *Imposibil*.

Rămân de analizat următoarele cazuri, derivate din negarea condiției multiple de mai sus:

- *Cazul particular 1:* $w_{11} = w_{12}$ și $w_{11} = -w_{12}$ (de unde rezultă $w_{11} = w_{12} = 0$);
- *Cazul particular 2:* $w_{11} = w_{12}$ și $w_{21} = -w_{22}$;
- *Cazul particular 3:* $w_{21} = w_{22}$ și $w_{11} = -w_{12}$;
- *Cazul particular 4:* $w_{21} = w_{22}$ și $w_{21} = -w_{22}$ (de unde rezultă $w_{21} = w_{22} = 0$).

Cazul 4 este similar cazului 1, iar cazul 3 este similar cazului 2. Se poate arăta relativ ușor (deși nu este chiar imediat) că în fiecare dintre aceste cazuri particulare obținem același răspuns: *Imposibil*.

5.

(Expresivitatea rețelelor neuronale: funcții booleene)

■ • ○ CMU, 2010 fall, Aarti Singh, HW5, pr. 4.3
CMU, 2011 spring, Roni Rosenfeld, HW4, pr. 1.d

Să se arate că orice funcție booleană (cu n argumente / variabile booleene și câte o singură valoare booleană pentru fiecare combinație posibilă de valori pentru cele n argumente), poate fi reprezentată printr-o rețea neuronală cu doar două niveluri, folosind neuroni care au componenta de activare de tip funcție-prag.

Răspuns:

Stim că orice funcție booleană poate fi reprezentată în mod echivalent ca o expresie din logica propozițiilor. La rândul ei, aceasta se poate scrie ca o conjuncție de disjuncții de literalii (sau invers, ca o disjuncție de conjuncții de literalii), un literal fiind fie o variabilă fie negația ei.⁷⁴⁷ Pentru exemplificare, vedetă rezolvarea problemei 3. (Demonstrația dată aici poate fi văzută ca o generalizare naturală a exemplului particular de acolo.) Pentru a reprezenta o astfel de expresie cu ajutorul rețelelor de perceptriони-prag vom proceda astfel:

⁷⁴⁷Pentru fiecare combinație de valori l_1, \dots, l_n ale variabilelor X_1, \dots, X_n pentru care funcția booleană dată (f) ia valoarea de adevară T , se va considera conjuncția $L_1 \wedge \dots \wedge L_n$, unde L_i este X_i dacă $l_i = T$, și L_i este $\neg X_i$ dacă $l_i = F$. Expresia booleană care reprezintă funcția f este disjuncția de conjuncții astfel calculate.

- Fiecare disjuncție de literalii va fi reprezentată prin câte un perceptron-prag plasat pe nivelul ascuns. Ponderile pentru acest perceptron vor fi:

1 pentru literalii pozitivi,
 -1 pentru literalii negativi,
 $k - \frac{1}{2}$ pentru termenul liber ($x_0 = 1$), unde k este numărul de literali din disjuncție.

- Pentru conjuncția de disjuncții, vom pune un perceptron-prag pe nivelul exterior și-i vom asigna următoarele ponderi:

1 pentru fiecare conexiune hidden-to-output,
 $-(l - \frac{1}{2})$ pentru termenul liber, unde l este numărul de disjuncții, același cu numărul de perceptri-prag de pe nivelul ascuns.

Observații:

1. În mod absolut similar se poate proceda pornind de la rezultatul, cunoscut din logica propozițiilor, că orice expresie booleană se poate scrie ca o conjuncție de disjuncții.
2. Este posibil ca în rețea obținută, numărul de unități de pe nivelul ascuns să fie exponențial în raport cu numărul de variabile de intrare (n). Așadar, în astfel de cazuri, pentru valori mari ale lui n , rezultatul care a fost demonstrat la acest exercițiu, deși are o semnificație teoretică importantă, nu are aplicabilitate practică.

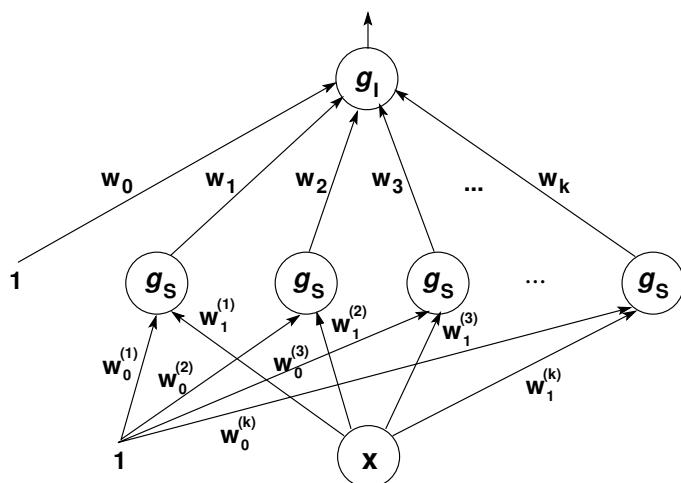
6. (Reprezentarea unor funcții de tip treaptă cu ajutorul unor rețele cu unități liniare sau de tip prag)

CMU, 2007 fall, Carlos Guestrin, HW2, pr. 4

În această problemă presupunem că dispunem (doar) de unități neuronale ale căror funcții de activare pot fi de două tipuri:

- funcția identitate: $g_I(x) = x$, și
- funcția treaptă: $g_S(x) = 1$ dacă $x \geq 0$ și 0 în rest.

De exemplu, rețeaua neuronală din figura alăturată are o intrare $x \in \mathbb{R}$, un singur nivel ascuns pe care sunt k unități având funcția de activare de tip treaptă, iar pe nivelul de ieșire un singur perceptron având funcția de activare de tip identitate.



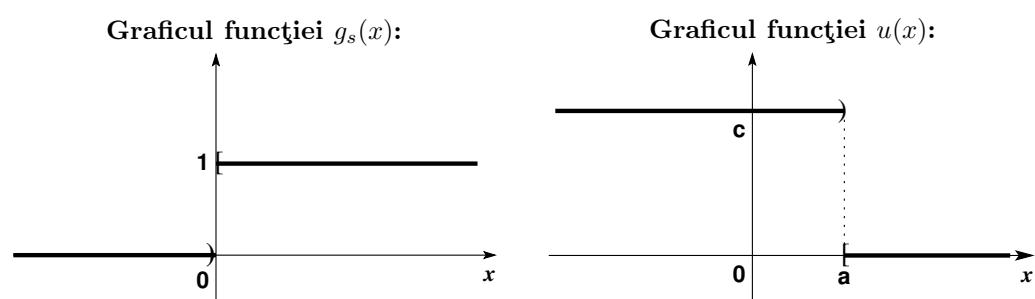
Ieșirea acestei rețele poate fi descrisă astfel:

$$out(x) = g_I \left(w_0 + \sum_{i=1}^k w_i g_S \left(w_0^{(i)} + w_1^{(i)} x \right) \right) = w_0 + \sum_{i=1}^k w_i g_S \left(w_0^{(i)} + w_1^{(i)} x \right)$$

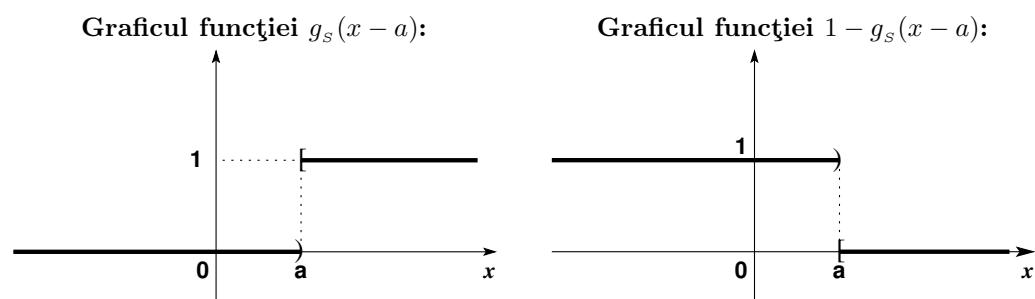
- a. Considerăm funcția de tip treaptă care este definită prin expresia $u(x) = c$ când $x < a$ și 0 în rest (unde a și c sunt constante reale fixate). Construiți o rețea neuronală având un singur nivel ascuns, intrarea x și outputul $u(x)$. Desenați structura rețelei neuronale, indicând funcția de activare pentru fiecare unitate (fie g_I fie g_S), și specificați valorile tuturor ponderilor (ca expresii în funcție de a și c).
- b. Se consideră constantele reale (fixate) a , b și c . Construiți o rețea neuronală având o intrare x și un nivel ascuns, al cărei output este c dacă $x \in [a, b]$ și 0 în rest. Desenați structura rețelei neuronale, indicând funcția de activare pentru fiecare unitate (fie g_I fie g_S), și specificați valorile tuturor ponderilor (ca expresii în funcție de a , b și c).

Răspuns:

- a. Pentru conveniență, vom reprezenta mai întâi graficele funcțiilor treaptă g_S și u din enunțul problemei.



Pornind de la aceste reprezentări grafice, putem să obținem în câteva pași funcția treaptă $u(x)$ ca funcție compusă, exprimată cu ajutorul lui g_S , după cum urmează:



Comparând ultimul grafic de mai sus cu graficul funcției u , rezultă că $u(x) = c(1 - g_S(x-a)) \Rightarrow u(x) = c - c \cdot g_S(x-a)$.

Pe de altă parte, putem observa că o rețea neuronală care are o singură unitate pe nivelul ascuns având funcția de activare de tip treaptă și o singură unitate

pe nivelul de ieșire având funcția de activare de tip identitate va avea ieșirea $out(x) = g_I(w_0 + w_1 \cdot g_S(w_0^{(1)} + w_1^{(1)}x))$, adică $out(x) = w_0 + w_1 \cdot g_S(w_0^{(1)} + w_1^{(1)}x)$.

Prin urmare, ponderile rețelei neuronale care reprezintă funcția u vor fi: $w_0 = c$, $w_1 = -c$, $w_0^{(1)} = -a$ și $w_1^{(1)} = 1$, iar rețeaua însăși va arăta ca în figura alăturată.

b. Funcția

$$v(x) = \begin{cases} c & \text{dacă } x \in [a, b] \\ 0 & \text{în rest} \end{cases}$$

este reprezentată grafic în figura alăturată. Se poate observa din acest grafic că funcția v se obține ușor din funcția u de la punctul anterior.

Într-adevăr, dacă notăm

$$u_a(x) = \begin{cases} c & \text{dacă } x < a \\ 0 & \text{în rest,} \end{cases}$$

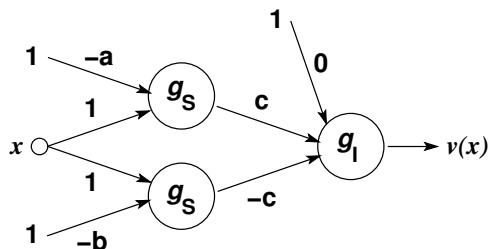
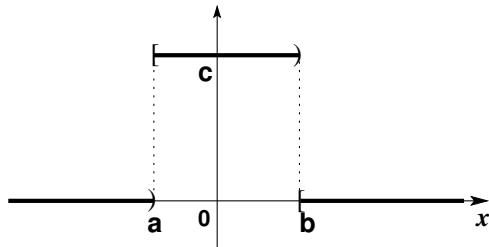
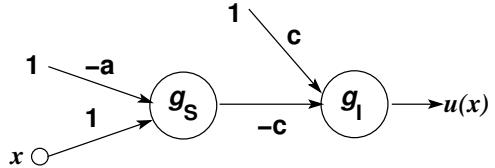
atunci

$$v(x) = u_b(x) - u_a(x) = c(1 - g_S(x - b)) - c(1 - g_S(x - a)).$$

Calculând, obținem $v(x) = c - c \cdot g_S(x - b) - c + c \cdot g_S(x - a)$

$$\Rightarrow v(x) = c \cdot g_S(x - a) - c \cdot g_S(x - b)$$

Așadar, putem construi o rețea neuronală cu două unități pe nivelul ascuns (fiecare având funcția de activare de tip treaptă). Ieșirile acestor două unități vor constitui intrările unei unități [cu funcție de activare] de tip liniar, aflată pe nivelul de ieșire. Setând corespunzător ponderile corespunzătoare acestui nivel, outputul rețelei va fi chiar $v(x)$. Reprezentarea acestei rețele este cea din figura alăturată.



7.

(Orice funcție definită pe un interval mărginit din \mathbb{R} , care este continuă în sens Lipschitz, poate fi aproximată oricât de bine cu ajutorul unei rețele neuronale care are un singur nivel ascuns)

■ • ○ CMU, 2011 fall, T. Mitchell, A. Singh, HW5, pr. 2.3

Considerăm o funcție oarecare $f(x)$, al cărei domeniu este de forma $[C, D] \subset \mathbb{R}$. Presupunem că f este continuă în sens Lipschitz,⁷⁴⁸ adică există o constantă

⁷⁴⁸Continuitatea în sens Lipschitz este o formă [mai] tare de uniform-continuitate: se poate demonstra că orice funcție continuă în sens Lipschitz este uniform continuă (și, deci, continuă).

$L \geq 0$ astfel încât

$$|f(x') - f(x)| \leq L|x' - x|, \forall x, x' \in [C, D].$$

Folosiți rezultatele obținute la problema 6 pentru a construi o rețea neuronală cu un singur nivel ascuns, care aproximează funcția f cu o eroare de cel mult $\varepsilon > 0$, adică

$$\forall x \in [C, D], |f(x) - \text{out}(x)| \leq \varepsilon,$$

unde $\text{out}(x)$ desemnează ieșirea acestei rețele neuronale pentru intrarea x .

Va trebui ca rețeaua să folosească doar funcțiile de activare de tip identitate și respectiv prag (notate cu g_I și g_S în problema 6). Indicați

- numărul K de unități ascunse,
- funcția de activare pentru fiecare unitate,
- o formulă pentru calcularea fiecărei ponderi de pe conexiunile input-to-hidden (notate cu $w_0^{(k)}$ și $w_1^{(k)}$, unde $k \in \{1, 2, \dots, K\}$), precum și pentru conexiunile hidden-to-output (și anume w_0, w_k , cu $k \in \{1, 2, \dots, K\}$).

Aceste ponderi pot fi specificate în funcție de C, D, L și ε , precum și în funcție de valorile $f(x)$ obținute pentru un număr finit de valori x , la libera alegere. (Va trebui să specificați în mod explicit care sunt valorile x pe care le veți folosi.)

Nu vi se cere să scrieți în mod explicit funcția $\text{out}(x)$.

Cum justificați faptul că rețeaua concepută de dumneavoastră atinge acuratețea indicată?

Răspuns:

Este imediat că din ipoteza $|f(x) - f(x')| \leq L|x - x'|$, în cazul în care $|x - x'| \leq \frac{\varepsilon}{L}$, rezultă

$$|f(x) - f(x')| \leq \varepsilon \quad (335)$$

Așadar, este de dorit să „acoperim“ intervalul $[C, D]$ cu intervale de lungime $\frac{2\varepsilon}{L}$ (sau mai mică):

$$[C, C + \frac{2\varepsilon}{L}), [C + \frac{2\varepsilon}{L}, C + 2\frac{2\varepsilon}{L}), \dots, [C + (K - 1)\frac{2\varepsilon}{L}, D))$$

și să luăm punctele x' de forma $\xi_i = (\alpha_i + \beta_i)/2$, pentru cu $i = 1, \dots, K$, unde

$$\begin{aligned} K &\stackrel{\text{not.}}{=} \left\lceil (D - C) \frac{L}{2\varepsilon} \right\rceil, \\ \alpha_1 &= C, \beta_1 = \alpha_2 = C + \frac{2\varepsilon}{L}, \dots, \beta_{K-1} = \alpha_K = C + (K - 1)\frac{2\varepsilon}{L}, \beta_K = D. \end{aligned}$$

Așadar, prin ξ_i am notat mijlocul intervalului i de mai sus, $[\alpha_i, \beta_i]$.

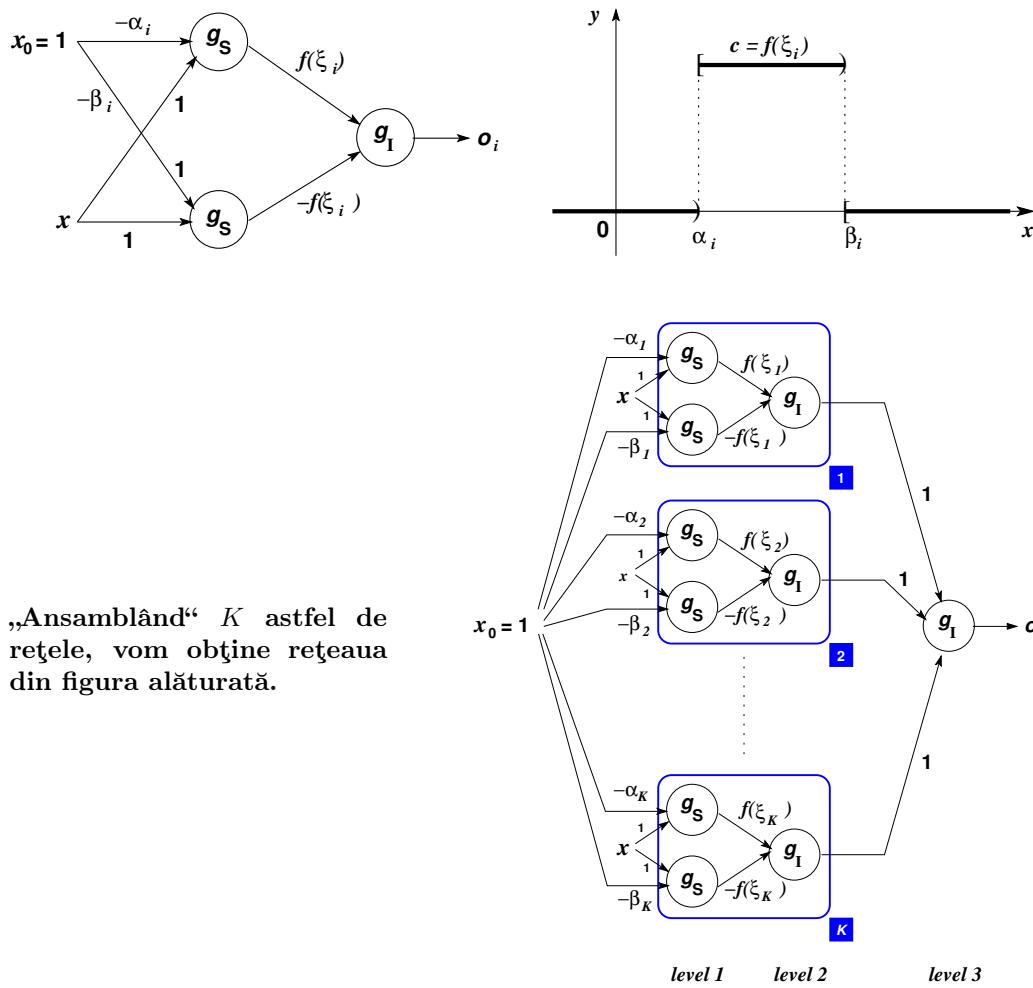
În continuare, vom arăta că putem construi o rețea neuronală cu un singur nivel ascuns și care folosește doar unități de tip liniar sau de tip prag, astfel încât

$$\text{pentru } \forall x \in [C, D], \text{ dacă } x \in [\alpha_i, \beta_i], \text{ atunci } \text{out}(x) \stackrel{\text{def.}}{=} f(\xi_i). \quad (336)$$

În consecință, dacă în relația (335) vom înlocui x' cu ξ_i , va rezulta imediat că $|f(x) - \text{out}(x)| \leq \varepsilon$.

Revenind acum la proprietatea (336), vom arăta că există o rețea având K unități pe nivelul ascuns (și o singură unitate de ieșire), astfel încât, dacă $x \in [\alpha_i, \beta_i]$, atunci unitatea i de pe nivelul ascuns se activează (producând valoarea $f(\xi_i)$, iar toate celelalte unități de pe nivelul ascuns rămân neactive (adică produc output 0). Prin urmare, rezultatul va fi exact cel dorit.

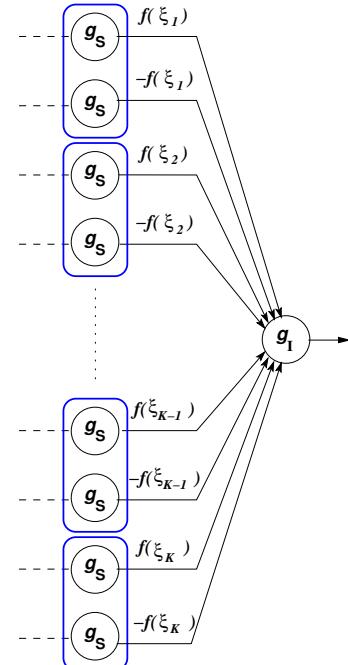
Conform problemei 6 punctul b, există o rețea neuronală care produce valoarea $c = f(\xi_i)$ pentru orice input $x \in [\alpha_i, \beta_i]$ și 0 în rest, și care folosește funcții de activare g_S pe nivelul ascuns și g_I pe nivelul de ieșire:



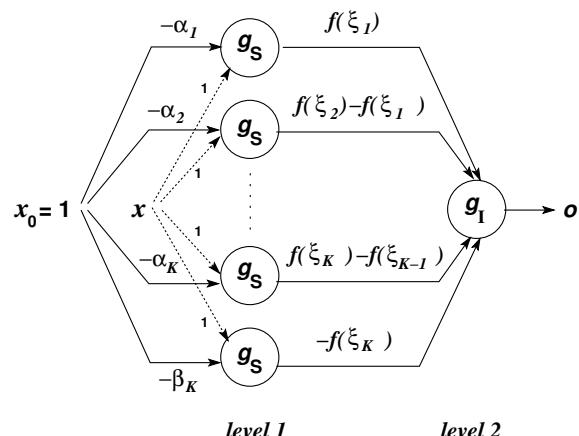
Outputul acestei rețele este exact cel dorit (deci satisfacă condiția din enunț):

$$\text{out}(x) = \begin{cases} f(\xi_1) & \text{pt. } x \in [\alpha_1 = C, \beta_1] \\ f(\xi_2) & \text{pt. } x \in [\alpha_2 = \beta_1, \beta_2] \\ \dots & \dots \\ f(\xi_K) & \text{pt. } x \in [\alpha_K = \beta_{K-1}, \beta_K = D] \end{cases} \Rightarrow |f(x) - \text{out}(x)| \leq \varepsilon, \forall x \in [C, D].$$

*Neajunsul este că rețeaua aceasta are două niveuri ascunse, în loc de unul singur, aşa cum se cere în enunț. Totuși, el poate fi „corectat“ imediat, comasând nivelurile 2 și 3 — formate doar din unități liniare — într-o singură unitate liniară (care va constitui nivelul de ieșire al noii rețele), ca în figura alăturată.*⁷⁴⁹



De asemenea, datorită faptului că $\beta_1 = \alpha_2, \beta_2 = \alpha_3, \dots, \beta_{K-1} = \alpha_K$, nivelul ascuns (al acestei noi rețele), format din cele $2K$ unități de tip prag, poate fi echivalat cu un altul, compus din doar $K+1$ unități de tip prag. La final, obținem rețeaua din figura alăturată.



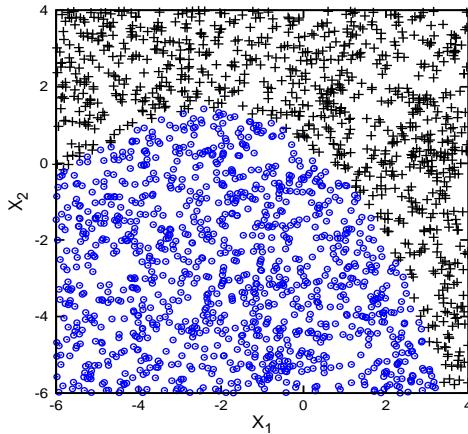
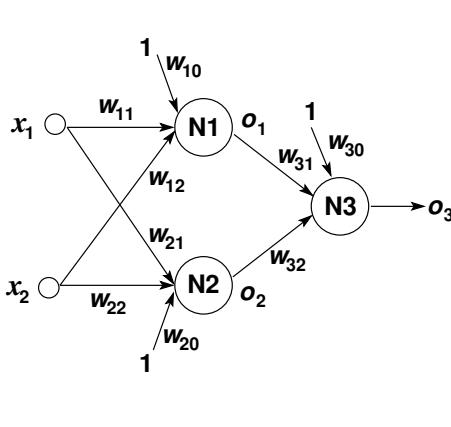
8.

(Rețele neuronale cu unități de tip sigmoidal: granițe / suprafete de decizie)

*CMU, 2008 fall, Eric Xing, HW2, pr. 2.1
CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh, HW3, pr. 2.1
CMU, 2011 fall, Eric Xing, HW1, pr. 3.2*

Una dintre chestiunile importante pe care trebuie să ni le punem atunci când facem cunoștință cu un nou clasificator este *ce tip de suprafețe de decizie* (sau: granițe de decizie; engl., decision boundaries) *poate să învețe respectivul clasificator*.

⁷⁴⁹Vedeți proprietatea enunțată la problema 32.b.



Fie rețeaua neuronală din figura de mai sus, partea stângă. Cele trei unități din componentă acestei rețele sunt de tip sigmoidal, adică au funcția de activare $\sigma(x) = \frac{1}{1 + e^{-x}}$. Imaginea din partea dreaptă a fost obținută calculând valorile produse de rețea pentru o serie de puncte generate în mod aleatoriu în dreptunghiul $(-6, 4) \times (-6, 4)$ din planul euclidian, după ce ponderile au fost instantiată cu următoarele valori: $w_{10} = -0.8, w_{11} = 0.8, w_{12} = 0.1, w_{20} = 0.3, w_{21} = 0.3, w_{22} = -0.4, w_{30} = 0.2, w_{31} = 1$ și $w_{32} = -1$. Clasificarea instanțelor a fost desemnată folosind semnele + și respectiv o.

- Exprimăți ieșirile o_1 și o_2 în funcție de intrările x_1 și x_2 și ponderile $w_{10}, w_{11}, w_{12}, w_{20}, w_{21}$ și w_{22} .
- Scrieți regula de decizie corespunzătoare acestei rețele neuronale.
- Presupunem că eliminăm funcția de activare sigmoidală din perceptronii de pe nivelul ascuns (așadar, N1 și N2 devin unități liniare), însă păstrăm aceleasi valori pentru ponderi. Scrieți regula de decizie pentru noua rețea.
- Presupunem că lăsăm valorile ponderilor din rețeaua de la punctul precedent să varieze liber. Poate această rețea să învețe conceptul reprezentat de rețeaua inițială?

Răspuns:

- $$o_1(x_1, x_2) = \sigma(w_{10} + w_{11}x_1 + w_{12}x_2) = \frac{1}{1 + \exp(-(w_{10} + w_{11}x_1 + w_{12}x_2))}$$

$$o_2(x_1, x_2) = \sigma(w_{20} + w_{21}x_1 + w_{22}x_2) = \frac{1}{1 + \exp(-(w_{20} + w_{21}x_1 + w_{22}x_2))},$$

unde simbolul \exp desemnează funcția exponențială cu baza e .

- $$o(x_1, x_2) = \sigma(w_{30} + w_{31}o_1(x_1, x_2) + w_{32}o_2(x_1, x_2))$$

$$= \frac{1}{1 + \exp(-(w_{30} + w_{31}o_1(x_1, x_2) + w_{32}o_2(x_1, x_2)))}.$$

Regula de decizie corespunzătoare rețelei neuronale din enunț este:

```
if  $o(x_1, x_2) \geq 1/2$  then assign the instance  $(x_1, x_2)$  the label +;
else assign it the label -;
```

sau, echivalent:

```
if  $w_{30} + w_{31}o_1(x_1, x_2) + w_{32}o_2(x_1, x_2) \geq 0$ 
then assign the instance  $(x_1, x_2)$  the label +;
else, assign it the label -.
```

Este de remarcat faptul că granița de decizie a rețelei neuronale, și anume curba de ecuație $w_{30} + w_{31}o_1(x_1, x_2) + w_{32}o_2(x_1, x_2) = 0$, este una *neliniară*, după cum se poate observa și din figura din enunț (partea dreaptă).

$$\begin{aligned} c. \quad o'_1(x_1, x_2) &= w_{10} + w_{11}x_1 + w_{12}x_2 \\ o'_2(x_1, x_2) &= w_{20} + w_{21}x_1 + w_{22}x_2 \\ o'(x_1, x_2) &= \sigma(w_{30} + w_{31}o'_1(x_1, x_2) + w_{32}o'_2(x_1, x_2)) \\ &= \sigma(w_{30} + w_{31}(w_{10} + w_{11}x_1 + w_{12}x_2) + w_{32}(w_{20} + w_{21}x_1 + w_{22}x_2)) \\ &= \sigma(w_{30} + w_{31}w_{10} + w_{32}w_{20} + (w_{31}w_{11} + w_{32}w_{21})x_1 + \\ &\quad (w_{31}w_{12} + w_{32}w_{22})x_2). \end{aligned}$$

Regula de decizie este:

```
if  $w_{30} + w_{31}w_{10} + w_{32}w_{20} + (w_{31}w_{11} + w_{32}w_{21})x_1 + (w_{31}w_{12} + w_{32}w_{22})x_2 \geq 0$ 
then assign the instance  $(x_1, x_2)$  the label +;
else, assign it the label - .
```

d. Se observă că la punctul c granița (suprafața) de decizie este o dreaptă. Așadar, noua rețea nu poate învăța în mod convenabil conceptul reprezentat de rețeaua inițială.

Observație: Este util de remarcat faptul că o rețea exact de același tip, având însă alte valori pentru ponderile w_{ij} , poate reprezenta funcția XOR.⁷⁵⁰

9.

(Unitatea sigmoidală: exemplificare)

■ • ○ CMU, 2003 fall, T. Mitchell, A. Moore, midterm, pr. 6.a

Considerăm o unitate neuronală sigmoidală care are intrările x_1, x_2 și x_3 . Așadar, ieșirea ei este de forma $y = \sigma(w_0 + w_1x_1 + w_2x_2 + w_3x_3)$ unde $\sigma(z) = \frac{1}{1 + e^{-z}}$. Valorile fiecărei dintre aceste trei intrări sunt 0 sau 1.

Asignați valori ponderilor w_0, w_1, w_2 și w_3 astfel încât outputul unității neuronale să fie strict mai mare decât 0.5 dacă și numai dacă valoarea logică a expresiei $(x_1 \wedge x_2) \vee x_3$ este *adevărat*.

Răspuns:

Vom scrie mai întâi tabela de valori a funcției / expresiei booleene date. Vom adăuga în această tabelă o coloană care va conține forma particulară a sumei ponderate $w_0 + w_1x_1 + w_2x_2 + w_3x_3$ — adică, *argumentul* pe care-l va lua funcția sigmoidală (σ) — pentru fiecare combinație de valori ale variabilelor x_1, x_2, x_3 . În plus, la fiecare linie vom preciza și *condiția* care trebuie să fie satisfăcută

⁷⁵⁰Vedeți CMU, 2008 fall, Eric Xing, midterm, pr. 4.2. Ar fi [un exercițiu] instructiv să vizualizați cum arată zonele de decizie în acel caz.

de către suma ponderată de pe linia respectivă, în aşa fel încât perceptronul să realizeze codificarea cerută în enunt.⁷⁵¹

x_1	x_2	x_3	$(x_1 \wedge x_2) \vee x_3$	$w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3$	$\vdots 0$	
0	0	0	0	w_0	< 0	
0	0	1	1	$w_0 + w_3$	> 0	
0	1	0	0	$w_0 + w_2$	< 0	
0	1	1	1	$w_0 + w_2 + w_3$	> 0	
1	0	0	0	$w_0 + w_1$	< 0	
1	0	1	1	$w_0 + w_1 + w_3$	> 0	
1	1	0	1	$w_0 + w_1 + w_2$	> 0	
1	1	1	1	$w_0 + w_1 + w_2 + w_3$	> 0	

Așadar, rezumând, pentru ca unitatea sigmoidală să codifice expresia booleană dată, trebuie ca ponderile w_i să satisfacă sistemul de inecuații scrise în ultima coloană a tabelului de mai sus.

În vederea găsirii unei soluții pentru acest sistem, putem observa că în expresia $(x_1 \wedge x_2) \vee x_3$ variabilele logice x_1 și x_2 joacă rol simetric în raport cu operatorul \wedge . Deci este natural să considerăm că ponderile alocate lui x_1 și x_2 [ca inputuri] în perceptronul sigmoidal pot fi egale. Introducând deci restricția $w_1 = w_2$, sistemul (337) va deveni:

$$\begin{cases} w_0 < 0 \\ w_0 + w_3 > 0 \\ w_0 + w_1 < 0 \\ w_0 + w_1 + w_3 > 0 \\ w_0 + 2w_1 > 0 \\ w_0 + 2w_1 + w_3 > 0 \end{cases} \quad (338)$$

Mai departe, analizând din nou expresia $(x_1 \wedge x_2) \vee x_3$, este natural să gândim că, în raport cu operatorul \vee , ponderea variabilei x_3 ar trebui să fie aceeași cu ponderile „cumulate“ ale variabilelor x_1 și x_2 . Prin urmare, „injectând“ în sistemul de inecuații (338) restricția $w_3 = w_1 + w_2 = 2w_1$, el va deveni:

$$\begin{cases} w_0 < 0 \\ w_0 + 2w_1 > 0 \\ w_0 + w_1 < 0 \\ w_0 + 3w_1 > 0 \\ w_0 + 4w_1 > 0 \end{cases} \quad (339)$$

Din forma acestui sistem, apare natural să explorăm ce anume se întâmplă dacă restricționăm spațiul de soluții impunând condiția $w_1 > 0$. În această ipoteză, sistemul (339) va avea aceeași soluție ca și inecuația dublă

$$w_0 + w_1 < 0 < w_0 + 2w_1,$$

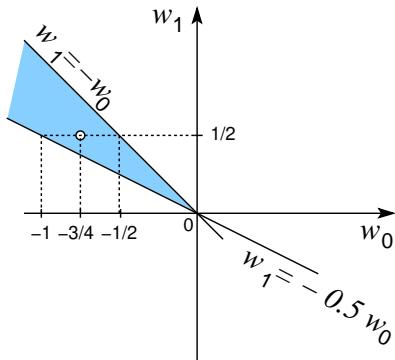
care este echivalentă cu

$$w_1 < -w_0 < 2w_1. \quad (340)$$

⁷⁵¹Stim că la clasificare cu unitatea sigmoidală se folosește echivalența $\sigma(z) > 1/2 \Leftrightarrow z > 0$ și, complementar, $\sigma(z) < 1/2 \Leftrightarrow z < 0$.

Pentru această ultimă inecuație dublă este foarte ușor să indicăm soluții. Iată una dintre ele: $w_1 = \frac{1}{2}$ și $w_0 = -\frac{3}{4}$. Prin urmare, $w_2 = \frac{1}{2}$ și $w_3 = 1$.

Observație (1): De fapt, dacă lucrăm într-un reper de coordonate w_0Ow_1 , orice punct (w_0, w_1) din cadranul al doilea, și care este situat între dreptele de ecuații $w_1 = -w_0$ și respectiv $w_1 = -\frac{1}{2}w_0$ este o soluție a inecuației duble (340), deci și a sistemului de inecuații (337), la care, aşa cum am văzut, au fost adăugate restricțiile $w_1 = w_2 = \frac{1}{2}w_3$. (Invers, adică a preciza care este întreg spațiul de soluții al sistemului (337) este mult în afara obiectivului acestui exercițiu.)



Observație (2): Rostul acestui [tip de] exercițiu este să arate că pentru a găsi valori corespunzătoare ponderilor unui perceptron în așa fel încât el să reprezinte / codifice o anumită funcție, putem apela la [rezolvarea de] sisteme de restricții / inecuații. Mai general, aşa cum precizează și Tom Mitchell în cartea *Machine Learning* la pag. 95, putem folosi metode de programare liniară sau neliniară. Pe de altă parte, acest exercițiu pune în evidență în mod indirect faptul că ar fi de dorit ca (alternativ) să dispunem de o procedură de calcul generală, cât mai simplă, prin care să atingem obiectivul menționat anterior. Modul în care, pentru cazul perceptronului liniar, se fundamentează din punct de vedere teoretic o astfel de procedură — bazată pe metoda gradientului descendente — va face obiectul problemei propuse 36, iar aplicarea ei va fi exemplificată la problemele 10 și 11. În raport cu programarea lininară sau cea neliniară, această procedură are avantajul că se scalează în mod elegant / convenabil la nivel de rețea neuronală.

6.1.2 Unități neuronale — algoritmi de antrenare

10.

(Unitatea liniară;

deducerea regulii de actualizare a ponderilor (în manieră “batch”),
în cazul particular al învățării unei anumite funcții booleane)

Liviu Ciortuz, 2011

Vom considera din nou funcția booleană $A \wedge \neg B$ (vedeți problema 2.a), dar nu vom mai folosi perceptronul-prag ci unitatea liniară (perceptronul liniar). Ne propunem aici să deducem forma specifică a *regulilor* de actualizare a ponderilor acestui perceptron, cu *obiectivul* de a „învăța“ ulterior valorile concrete ale acestor ponderi, astfel încât perceptronul rezultat să reprezinte funcția booleană $A \wedge \neg B$ (așa cum se va vedea la problema 11).

Considerând că antrenarea se va face în regim de lucru “batch” (folosind, desigur, ca instanțe de antrenament, liniile din tabela de adevăr a acestei

funcții booleene, cu codificarea 1 pentru *true*, și -1 pentru *false*), vă cerem să particularizați forma pe care o ia regula generală de actualizare a ponderilor ($\bar{w} \leftarrow \bar{w} + \Delta\bar{w}$) la învățarea acestei funcției booleene, procedând în două moduri diferite:

- a. calculând efectiv vectorul $\Delta\bar{w}$ (sau, pe componente, Δw_j) folosind formula dată la curs:

$$\Delta\bar{w} = \eta \sum_{i=1}^4 (t_i - o_i) \bar{x}_i$$

unde \bar{x}_i sunt instanțele din tabela de valori ale funcției $A \wedge \neg B$, $o_i = \bar{w} \cdot \bar{x}_i$ desemnează valoarea calculată de perceptron pentru $i = 1, \dots, 4$, iar t_i este valoarea target a funcției booleene date, pentru intrarea \bar{x}_i , cu $i = 1, \dots, 4$;⁷⁵²

- b. calculând mai întâi expresia funcției $E(\bar{w})$ care desemnează *semisuma pătratelor erorilor* corespunzătoare instanțelor \bar{x}_i ($i = 1, \dots, 4$):

$$E(\bar{w}) = \frac{1}{2} \sum_{i=1}^4 (t_i - o_i)^2 = \frac{1}{2} \sum_{i=1}^4 (t_i - \bar{w} \cdot \bar{x}_i)^2 = \dots,$$

precum și cele trei derive parțiale din expresia vectorului gradient

$$\nabla E(\bar{w}) = \left(\frac{\partial E(\bar{w})}{\partial w_0}, \frac{\partial E(\bar{w})}{\partial w_1}, \frac{\partial E(\bar{w})}{\partial w_2} \right),$$

folosind expresia funcției $E(\bar{w})$ care a fost calculată anterior, pentru ca la final, în spiritul *metodei gradientului descendente*, să obținem $\Delta\bar{w} = -\eta \nabla E(\bar{w})$.

Atenție! Dacă la final nu ați obținut același rezultat la punctele *a* și *b*, înseamnă că ați greșit la calcule.

Răspuns:

- a. Stim că $o_i = \bar{w} \cdot \bar{x}_i$ și $\Delta\bar{w} = \eta \sum_{i=1}^4 (t_i - o_i) \bar{x}_i$. În cazul nostru, $o_i = (w_0, w_1, w_2) \cdot (1, x_{i,1}, x_{i,2})$, deci $o_1 = w_0 - w_1 - w_2$, $o_2 = w_0 - w_1 + w_2$, $o_3 = w_0 + w_1 - w_2$, $o_4 = w_0 + w_1 + w_2$ și, ținând cont că $t_1 = t_2 = t_4 = -1$ și $t_3 = 1$, rezultă (în scriere matriceală):

$$\Delta\bar{w} = \eta \left[(-1 - o_1) \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix} + (-1 - o_2) \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} + (1 - o_3) \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} + (-1 - o_4) \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right].$$

Scriind această egalitate pe componente, vom avea mai întâi:

$$\begin{aligned} \Delta w_0 &= \eta(-1 - o_1 - 1 - o_2 + 1 - o_3 - 1 - o_4) = \eta(-2 - (o_1 + o_2 + o_3 + o_4)) \\ &= \eta(-2 - (w_0 - w_1 - w_2 + w_0 - w_1 + w_2 + w_0 + w_1 - w_2 + w_0 + w_1 + w_2)) \\ &= \eta(-2 - 4w_0) = -2\eta(2w_0 + 1) \end{aligned}$$

Așadar, regula de actualizare a ponderii w_0 va fi $w_0 \leftarrow w_0 - 2\eta(2w_0 + 1)$.

În mod similar, facând calculele, vom obține $w_1 \leftarrow w_1 - 2\eta(2w_1 - 1)$ și $w_2 \leftarrow w_2 - 2\eta(2w_2 + 1)$.

- b. Conform definiției, $E(w_0, w_1, w_2) = \frac{1}{2} \sum_i (t_i - o_i)^2$, cu $o_i = w_0 + w_1 x_{i,1} + w_2 x_{i,2}$. Prin urmare,

⁷⁵²Atenție! Aceste instanțe trebuie extinse cu componenta 1 corespunzătoare termenului liber din expresia separatorului liniar.

$$\begin{aligned}
E(w_0, w_1, w_2) &= \frac{1}{2} [(-1 - (w_0 - w_1 - w_2))^2 + (-1 - (w_0 - w_1 + w_2))^2 + \\
&\quad (1 - (w_0 + w_1 - w_2))^2 + (-1 - (w_0 + w_1 + w_2))^2] \\
&= \frac{1}{2} [(1 + w_0 - w_1 - w_2)^2 + (1 + w_0 - w_1 + w_2)^2 + \\
&\quad (1 - w_0 - w_1 + w_2)^2 + (1 + w_0 + w_1 + w_2)^2] \\
&= \frac{1}{2} [2(1 + w_0 - w_1)^2 + 2w_2^2 + 2(1 + w_2)^2 + 2(w_0 + w_1)^2] \\
&= (1 + w_0 - w_1)^2 + w_2^2 + (1 + w_2)^2 + (w_0 + w_1)^2 \\
&= 2(1 + w_0^2 + w_1^2 + w_2^2 + w_0 - w_1 + w_2)
\end{aligned}$$

Derivatele parțiale ale lui E în raport cu variabilele w_0 , w_1 și w_2 sunt:

$$\frac{\partial E}{\partial w_0} = 2(2w_0 + 1), \quad \frac{\partial E}{\partial w_1} = 2(2w_1 - 1), \quad \frac{\partial E}{\partial w_2} = 2(2w_2 + 1)$$

În concluzie, regulile de actualizare pentru ponderile unității liniare vor fi:

$$w_0 \leftarrow w_0 - 2\eta(2w_0 + 1), \quad w_1 \leftarrow w_1 - 2\eta(2w_1 - 1), \quad w_2 \leftarrow w_2 - 2\eta(2w_2 + 1)$$

Se observă că rezultatele obținute la punctele a și b coincid, după cum era de așteptat.

11.

(Perceptronul-prag;
aplicarea algoritmului de actualizare a ponderilor
în manieră incrementală / stochastică,
pentru a învăța o anumită funcție booleană)

Liviu Ciortuz, 2011

La problema 2.a am arătat că funcția booleană $A \wedge (\neg B)$ poate fi reprezentată cu ajutorul unui perceptron-prag — adică, un perceptron cu funcție de activare de tip prag — ale cărui ponderi pe intrări sunt date de ecuația dreptei $y = x - 1 \Leftrightarrow x - y - 1 = 0 \Leftrightarrow w_0 = -1, w_1 = 1, w_2 = -1$. (A se vedea linia punctată din graficul de mai jos.) Evident, există o infinitate de drepte care separă în mod corect cele patru instanțe de antrenament.

În exercițiul de față, spre deosebire de cum am procedat la problema 2.a, nu vom mai „ghici“, ci vom învăța un separator liniar care să reprezinte / codifice aceeași funcție booleană de mai sus. Învățarea se va face aplicând algoritmul de antrenare a perceptronului (adică de actualizare a ponderilor) din carteia *Machine Learning* a profesorului T. Mitchell, pag. 93.⁷⁵³

Convenție: Vom considera că valorile logice Adevărat / Fals sunt codificate de numerele $+1/-1$. De asemenea, variabila numerică x (rescrisă uneori, pentru

⁷⁵³ Atenție: La problema 10 s-a lucrat cu același concept — funcția booleană $A \wedge (\neg B)$ — însă acolo am folosit perceptronul liniar în varianta “batch”. Aici folosim perceptronul-prag în varianta incrementală / stochastică. Din punct de vedere analitic, forma regulii de actualizare este $\bar{w} \leftarrow \bar{w} + \Delta \bar{w}$, cu $\Delta w_i = -\eta \sum_n (t_n - o_n) \frac{\partial E}{\partial w_i} \cdot x_{n,i}$ pentru prima variantă și, respectiv, $\Delta w_i = -\eta (t_n - o_n) \frac{\partial E}{\partial w_i} \cdot x_{n,i}$ pentru a doua variantă. Rețineți că outputurile (o_n) produse de cei doi perceptroni pentru un același input x_n diferă în general, întrucât perceptronul-prag aplică funcția de activare *sign* la combinația liniară $\bar{w} \cdot \bar{x}$, în vreme ce perceptronul liniar nu o aplică.

conveniență, ca x_1) corespunde variabilei A , iar y (rescrisă ca x_2) corespunde variabilei B .

Vom considera că valorile inițiale ale ponderilor perceptronului sunt cele corespunzătoare dreptei $y = \frac{1}{3}x - \frac{1}{3}$ $\Leftrightarrow \frac{1}{3}x - y - \frac{1}{3} = 0 \Leftrightarrow w_0 = -\frac{1}{3}, w_1 = \frac{1}{3}, w_2 = -1$.

Se poate constata imediat că funcția $\frac{1}{3}x - y - \frac{1}{3}$ clasifică în mod eronat punctul $(-1, -1)$. Prin urmare, separatorul liniar reprezentat de ecuația $\frac{1}{3}x - y - \frac{1}{3} = 0$ — vedeti dreapta continuă din graficul de mai jos — nu este consistent cu conceptul logic $A \wedge \neg B$.

a. Folosind rata de învățare $\eta = 0.1$, să se aplice atâtea iterații câte sunt necesare pentru actualizarea ponderilor perceptronului-prag — în manieră incrementală / stochastică — astfel încât, la final, toate instanțele din tabelul funcției booleane $A \wedge \neg B$ (vedeti tabelul alăturat) să fie clasificate corect. Scrieți ecuația separatorului liniar rezultat.

A	B	$A \wedge \neg B$
-1	-1	-1
-1	+1	-1
+1	-1	+1
+1	+1	-1

Vă reamintim că regula de actualizare a ponderilor pentru perceptronul-prag, în varianta stochastică / incrementală este de forma:

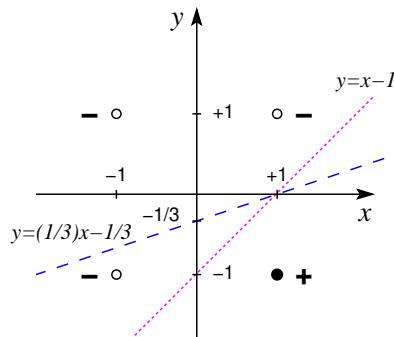
$$\bar{w} \leftarrow \bar{w} + \Delta \bar{w}, \text{ cu } \Delta \bar{w} = \eta(t - o)\bar{x}$$

unde $\bar{x} = (1, x_1, x_2)$ desemnează o instanță de antrenament oarecare, o este valoarea produsă de către perceptron pentru intrarea \bar{x} , iar t este valoarea țintă (engl., target) pentru aceeași intrare \bar{x} .

b. Ce s-ar fi întâmplat dacă am fi folosit o rată de învățare mult mai mică, de exemplu $\eta = 0.01$?

Răspuns:

Reprezentarea grafică a datelor de antrenament și a poziției inițiale a separatorului liniar ($\frac{1}{3}x - y - \frac{1}{3} = 0$) asociat perceptronului din enunț este cea din figura alăturată. Acest separator liniar greșește la clasificarea instanței logice $A = F, B = F$ (reprezentată numeric de către $x = -1$ și $y = -1$), fiindcă $\frac{1}{3}(-1) - (-1) - \frac{1}{3} = 1 - \frac{2}{3} = \frac{1}{3} > 0$. În schimb, toate celelalte instanțe de antrenament sunt clasificate corect (verificarea este imediată).



În vederea aplicării regulii de actualizare a ponderilor, vom calcula mai întâi $\Delta \bar{w} = \eta(t_1 - o_1) \cdot \bar{x}_1 = \eta(-1 - 1) \cdot \bar{x}_1$, unde $\bar{x}_1 = (1, -1, -1)$ este punctul incorect clasificat. Așadar,

$$\Delta \bar{w} = -2\eta(1, -1, -1) = -0.2 \cdot (1, -1, -1) = (-0.2, 0.2, 0.2)$$

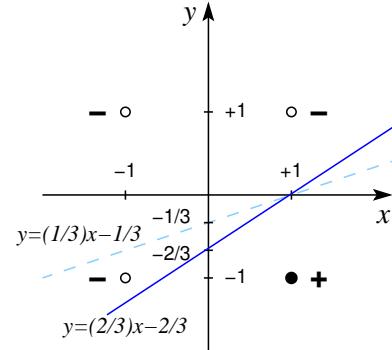
sau, rescris pe componente, $\Delta w_0 = -0.2$, $\Delta w_1 = 0.2$, $\Delta w_2 = 0.2$.⁷⁵⁴ Aplicând regula de actualizare a ponderilor, după efectuarea calculelor vom obține:

$$w_0 = -\frac{1}{3} - 0.2 = -\frac{8}{15}, \quad w_1 = \frac{1}{3} + 0.2 = \frac{8}{15}, \quad w_2 = -1 + 0.2 = -\frac{4}{5}$$

În consecință, perceptronul învăță o nouă poziție pentru separatorul liniar:

$$\frac{8}{15}x - \frac{4}{5}y - \frac{8}{15} = 0 \Leftrightarrow \frac{2}{3}x - y - \frac{2}{3} = 0 \Leftrightarrow y = \frac{2}{3}x - \frac{2}{3}$$

Valoarea funcției $\frac{2}{3}x - y - \frac{2}{3}$ pentru $x = -1$ și $y = -1$ este $-\frac{2}{3} + 1 - \frac{2}{3} = -\frac{1}{3} < 0$. Așadar, această instanță de antrenament este acum corect clasificată. Mai mult, se verifică imediat că toate instanțele de antrenament sunt clasificate corect de către noul separator liniar (vedeți graficul alăturat), deci algoritmul de actualizare a ponderilor perceptronului-prag se oprește după ce s-a efectuat o singură iterație.



Observație: Aplicarea variantei “batch” pentru algoritmul de învățare a ponderilor perceptronului-prag (bazată pe regulile deduse prin folosirea metodei gradientului descendente pentru perceptronul liniar) nu ar fi diferit de modul de lucru de mai sus (iar soluția obținută ar fi fost aceeași), fiindcă $t_n = o_n$ pentru orice $n \neq 1$. Doar scrierea / redactarea calculelor ar fi diferit ușor.

b. Dacă s-ar fi lucrat folosind o valoare mai mică pentru rata de învățare η , este foarte posibil să fi fost necesar să efectuăm mai multe iterări. Cele patru instanțe de antrenament fiind separabile liniar, convergența algoritmului (la valoarea minimă, adică 0, pentru eroarea totală) este asigurată pentru valori mici ale lui η . (Vedeți T. Mitchell, *Machine Learning*, pag. 89.)

⁷⁵⁴Dacă în locul perceptronului-prag am fi lucrat cu perceptronul liniar în varianta stochastică, aici am fi obținut

$$\begin{aligned} \Delta \bar{w} &= \eta(-1 - o(-1, -1))\bar{x}_1 = \eta(-1 - (-\frac{1}{3} + 1 - \frac{1}{3}))\bar{x}_1 = \eta(-1 - \frac{1}{3})\bar{x}_1 = -\eta \frac{4}{3}\bar{x}_1 = -\frac{1}{10} \frac{4}{3}\bar{x}_1 \\ &= -\frac{2}{15}(1, -1, -1) = \left(-\frac{2}{15}, \frac{2}{15}, \frac{2}{15}\right). \end{aligned}$$

După actualizare, aceasta ar fi dus la valorile

$$w_0 = -\frac{1}{3} - \frac{2}{15} = -\frac{7}{15}, \quad w_1 = \frac{1}{3} + \frac{2}{15} = \frac{7}{15}, \quad w_2 = -1 + \frac{2}{15} = -\frac{13}{15}.$$

Dreapta determinată de aceste ponderi are ecuația $\frac{17}{15}x_1 - \frac{13}{15}x_2 - \frac{7}{15} = 0 \Leftrightarrow x_2 = \frac{7}{13}x_1 - \frac{7}{13} = 0$. Pentru instanța \bar{x}_1 , expresia $\frac{17}{15}x_1 - \frac{13}{15}x_2 - \frac{7}{15}$ are valoarea $-\frac{17}{15} + \frac{13}{15} - \frac{17}{15} = -\frac{1}{15} < 0$.

12.

(Deducerea regulii de antrenare pentru un tip particular de unitate liniară)

• CMU, 1995 fall, Tom Mitchell, HW4, pr. 5

Considerăm un nou tip de unitate neuronală liniară care are două intrări x_1 și x_2 și o ieșire definită astfel:

$$\text{output} = w_0 + w_1 w_0 x_1 + w_2 w_0 x_2.$$

Derivați regula de actualizare a ponderilor cu metoda gradientului descedent, folosind următoarea definiție pentru *funcția de cost / pierdere* (engl., loss function):

$$E = \frac{1}{2} \sum_{d \in D} [\text{target}_d - \text{output}_d]^2,$$

unde D este setul de instanțe de antrenament.

Răspuns:

Conform metodei gradientului descedent, modificarea ponderilor se va face după formula $w_i \leftarrow w_i - \eta \frac{\partial E}{\partial w_i}$. Calculăm deci derivata funcției de eroare:

$$\begin{aligned} \frac{\partial E}{\partial w_i} &= \frac{\partial}{\partial w_i} \frac{1}{2} \sum_{d \in D} (t_d - o_d)^2 = \frac{1}{2} \sum_{d \in D} \frac{\partial}{\partial w_i} (t_d - o_d)^2 \\ &= \frac{1}{2} \sum_{d \in D} 2(t_d - o_d) \frac{\partial}{\partial w_i} (t_d - o_d) = \sum_{d \in D} -(t_d - o_d) \frac{\partial o_d}{\partial w_i} \\ &= - \sum_{d \in D} (t_d - o_d) \frac{\partial}{\partial w_i} (w_0 + w_1 w_0 x_{1,d} + w_2 w_0 x_{2,d}) \\ &= \begin{cases} - \sum_{d \in D} (t_d - o_d)(1 + w_1 x_{1,d} + w_2 x_{2,d}) & \text{dacă } i = 0 \\ - \sum_{d \in D} (t_d - o_d)w_0 x_{1,d} & \text{dacă } i = 1 \\ - \sum_{d \in D} (t_d - o_d)w_0 x_{2,d} & \text{dacă } i = 2. \end{cases} \end{aligned}$$

Așadar, actualizarea pondererilor o vom face folosind formulele:

$$\begin{aligned} w_0 &\leftarrow w_0 + \eta \sum_{d \in D} (t_d - o_d)(1 + w_1 x_{1,d} + w_2 x_{2,d}) \\ w_1 &\leftarrow w_1 + \eta \sum_{d \in D} (t_d - o_d)w_0 x_{1,d} \\ w_2 &\leftarrow w_2 + \eta \sum_{d \in D} (t_d - o_d)w_0 x_{2,d} \end{aligned}$$

unde η este o constantă pozitivă (rata de învățare).

13. (Aplicarea metodei gradientului descendente pentru modelarea unui dataset particular, cu „zgomot“ de tip gaussian)

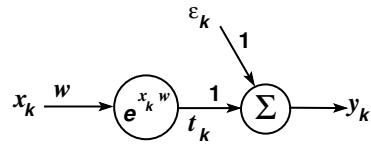
prelucrare de Liviu Ciortuz, după CMU, 2001 fall, Andrew Moore, final exam, pr. 10.2

Presupunem că avem la dispoziție un set de date, în care fiecare instanță are un atribut de intrare $x \in \mathbb{R}$ și un atribut de ieșire y . Bănuim că relația dintre valorile celor două attribute este de forma

$$y_k = e^{wx_k} + \varepsilon_k,$$

unde w este o constantă reală, iar $\varepsilon_k \in \mathbb{R}$ este un „zgomot“, care este independent de x_k .

Putem să ne imaginăm y_k ca fiind produs de o rețea neuronală [având pe nivelul ascuns un perceptron de un tip particular], care are o singură pondere nespecificată w , ca în figura alăturată. Se va considera că $t_k \stackrel{\text{not.}}{=} e^{wx_k}$, este targetul de învățat.



a. Considerând că instanțele x_1, \dots, x_n (fixate) au fost generate în mod independent și că „zgomotul“ ε urmează o distribuție de tip gaussian cu media 0 și varianță σ^2 (pentru orice $k \in \{1, \dots, n\}$),⁷⁵⁵ demonstrați că valoarea lui w pentru care se maximizează verosimilitatea condițională a datelor de antrenament, adică

$$\operatorname{argmax}_w P(y_1, \dots, y_n | x_1, \dots, x_n, w) \stackrel{\text{indep.}}{=} \operatorname{argmax}_w \prod_{k=1}^n P(y_k | x_k, w),$$

este exact aceeași cu valoarea (lui w) pentru care se minimizează suma pătratelor erorilor:

$$\operatorname{argmin}_w \sum_{k=1}^n (e^{wx_k} - y_k)^2.$$

Acest rezultat oferă o explicație suplimentară, probabilistă pentru utilizarea criteriului sumei pătratelor erorilor — a cărui justificare ar rămâne altminteri doar intuitivă — drept criteriu de optimizat la aplicarea metodei gradientului descendente pentru perceptri și rețele neuronale.

Indicație: Puteți face demonstrația nu doar pentru cazul de față, ci mai general, și anume pentru toate situațiile în care [ca target] vrem să învățăm o funcție oarecare f (în cazul nostru e^{wx}), datele de antrenament sunt de forma (x_k, y_k) , iar diferența dintre y_k și $f(x_k)$ (în cazul nostru ε_k) umează o distribuție gaussiană de medie 0.

b. Aplicați metoda gradientului descendente — adică, scrieți regula de actualizare a ponderii w — pentru a afla valoarea ponderii w care minimizează suma pătratelor erorilor.⁷⁵⁶

⁷⁵⁵ Pentru definiția distribuției gaussiene unidimensionale, vedeți problema 32 de la capitolul de *Fundamente*.

⁷⁵⁶ În enunțul original, în locul *sumei* este specificată *media* (engl., mean) *pătratelor erorilor*. Este vorba de medie în sensul de medie aritmetică. Cele două formulări sunt echivalente din punctul de vedere al optimizării (fiindcă diferă doar printr-un factor pozitiv), deci ambele sunt corecte. Am preferat să folosim termenul *sumă*, întrucât la punctul *a* se lucrează probabilist, deci în contextul respectiv termenul *medie* (ambiguu în românește!) ar avea [și] o altă semnificație.

Răspuns:

- a. Considerând că deviația standard a distribuției următoare de ε_k este σ și ținând cont că $P(\varepsilon_k) \stackrel{\text{def.}}{=} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\varepsilon_k - \mu)^2}{2\sigma^2}}$, cu $\mu = 0$, vom avea:

$$\begin{aligned}
 \operatorname{argmax}_w \prod_{k=1}^n P(y_k | x_k, w) &= \operatorname{argmax}_w \prod_{k=1}^n P(y_k - f(x_k) | x_k, w) \\
 &= \operatorname{argmax}_w \prod_{k=1}^n P(\varepsilon_k | x_k, w) = \operatorname{argmax}_w \prod_{k=1}^n P(\varepsilon_k | w) \\
 &= \operatorname{argmax}_w \prod_{k=1}^n P(y_k - t_k | w) = \operatorname{argmax}_w \prod_{k=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_k - t_k)^2}{2\sigma^2}} \\
 &= \operatorname{argmax}_w \ln \prod_{k=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_k - t_k)^2}{2\sigma^2}} = \operatorname{argmax}_w \sum_{k=1}^n \ln \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_k - t_k)^2}{2\sigma^2}} \\
 &= \operatorname{argmax}_w \sum_{k=1}^n \left(\ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{(y_k - t_k)^2}{2\sigma^2} \right) = \operatorname{argmax}_w \sum_{k=1}^n \left(-\frac{(y_k - t_k)^2}{2\sigma^2} \right) \\
 &= \operatorname{argmax}_w \left(-\sum_{k=1}^n \frac{(y_k - t_k)^2}{2\sigma^2} \right) = \operatorname{argmin}_w \sum_{k=1}^n \left(\frac{(y_k - t_k)^2}{2\sigma^2} \right) \\
 &= \operatorname{argmin}_w \sum_{k=1}^n (y_k - t_k)^2.
 \end{aligned}$$

- b. Ca și la exercițiile precedente, în locul sumei pătratelor erorilor vom considera drept criteriu de optimizat pentru metoda gradientului descendente semi-suma pătratelor erorilor, $E(w) = \frac{1}{2} \sum_k (y_k - t_k)^2$, unde $t_k = e^{wx_k}$, întrucât aceasta este ușor mai convenabilă la calcule și nu modifică rezultatul problemei. (A se vedea echivalența $(cf)'(x) = 0 \Leftrightarrow cf'(x) = 0$, unde c este constantă iar f este funcție derivabilă.)

Conform metodei gradientului descendente, regula de actualizare a ponderii w va fi (în varianta "batch") de forma

$$w \leftarrow w - \eta \frac{\partial E(w)}{\partial w}.$$

unde $\eta > 0$ este rata de învățare.

Calculăm derivata lui E în raport cu w :

$$\begin{aligned}
 \frac{\partial E(w)}{\partial w} &= \frac{\partial}{\partial w} \frac{1}{2} \sum_k (y_k - t_k)^2 = -\frac{1}{2} \sum_k 2(y_k - t_k) \frac{\partial t_k}{\partial w} \\
 &= -\sum_k (y_k - t_k) \frac{\partial}{\partial w} (e^{wx_k}) = -\sum_k (y_k - t_k) e^{wx_k} x_k = \sum_k (e^{wx_k} k - y_k) e^{wx_k} x_k.
 \end{aligned}$$

Așadar, regula de actualizare a ponderii w va fi:

$$w \leftarrow w + \eta \sum_k (y_k - o_k) e^{wx_k} x_k.$$

14.

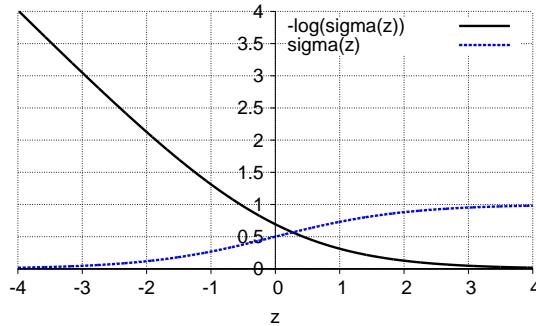
(O variantă a regulii de actualizare a ponderilor pentru perceptronul liniar, obținută prin aplicarea metodei gradientului descendente pe o altă funcție de pierdere / cost (engl., loss function) în locul sumei pătratelor erorilor: funcția log-sigmoidală)

■ • *Liviu Ciortuz, după University of Utah, 2008 fall, Hal Daumé III, HW4, pr. 2-4*

În acest exercițiu, vom construi o nouă variantă de unitate liniară (perceptron fără prag) care va optimiza în locul semisumei pătratelor erorilor, o *funcție de cost* de tip log-sigmoidal:⁷⁵⁷

$$l(y, h(x)) = -\ln \underbrace{\sigma(yh(x))}_z,$$

unde $\sigma(z) = 1/(1 + \exp(-z))$.



Convenție: Vom considera că o ipoteză oarecare h produce *valori reale* și că valorile pozitive desemnează clasa +1, iar cele negative clasa -1.

a. Verificați că funcția log-sigmoidală de mai sus este într-adevăr o *funcție de cost*, adică: atunci când clasa desemnată de ipoteza $h(x)$ este corectă ($y = h(x)$), valoarea $l(y, h(x))$ este mică și, invers, atunci când clasa desemnată de $h(x)$ este incorrectă ($y \neq h(x)$), valoarea $l(y, h(x))$ este mare.

b. Folosind o unitate liniară, dorim să învățăm o funcție liniară minimizând în locul semisumei pătratelor erorilor — cum am procedat la problemele 10, 11, 12 și 13 —, o funcție de cost de tip log-sigmoidal, și anume $f(\bar{w}) = \sum_n -\ln \sigma(y_n(\bar{w} \cdot \bar{x}_n + b))$, unde n este indicele folosit pentru indexarea instanțelor de antrenament $(\bar{x}_1, y_1), (\bar{x}_2, y_2), \dots$.⁷⁵⁸ Calculați gradientul acestei funcții în raport cu \bar{w} și în raport cu b .

c. Arătați că funcția log-sigmoidală de la punctul precedent este convexă atât în \bar{w} cât și în b . Așadar, se justifică următoarea cerință: formulați atât pentru w_i cât și pentru b *regulile de actualizare* (engl., update rules) care sunt necesare pentru antrenarea perceptronului liniar folosind metoda gradientului descendente în conjuncție cu funcția de cost log-sigmoidală.

Indicație: Pentru ușurință demonstrației, puteți folosi următoarele *propozitii*:⁷⁵⁹

i. Dacă f și g sunt funcții convexe, iar g este o funcție crescătoare (nu neapărat strict crescătoare) de o variabilă reală, atunci funcția h definită prin relația

⁷⁵⁷La capitolul *Metode de regresie* (vedeți problema 13) am numit această funcție *funcția de pierdere logistică*.

⁷⁵⁸Pentru clarificare, facem următoarea *precizare*: La problemele 8 și 9, funcția σ se aplica la calcularea outputului perceptronului sigmoidal: $o_n = \sigma(\bar{w} \cdot \bar{x}_n)$. Mai departe, o_n contribuia la calcularea funcției de eroare $E(\bar{w}) = 1/2 \sum_n (t_n - o_n)^2$ folosită la aplicarea metodei gradientului descendente. În schimb, la prezenta problemă funcția σ se aplică ieșirii produse de perceptronul liniar, sub forma $-\ln \sigma(t_n o_n)$, ceea ce corespunde termenului $(t_n - o_n)^2$ din $E(\bar{w})$.

⁷⁵⁹Vedeți https://en.wikipedia.org/wiki/Convex_function.

$h(x) = g(f(x))$ este convexă.⁷⁶⁰

ii. Dacă f este funcție concavă, iar g este funcție convexă și descrescătoare (nu neapărat strict descrescătoare) de o variabilă reală, atunci funcția h definită prin relația $h(x) = g(f(x))$ este convexă.⁷⁶¹

Răspuns:

a. Funcția l se poate scrie sub forma:

$$l(y, h(x)) = -\ln \sigma(y h(x)) = -\ln \frac{1}{1 + e^{-y h(x)}} = \ln(1 + e^{-y h(x)}).$$

Din proprietățile logaritmului stim că $\ln 1 = 0$, $\ln a < 0$ dacă $0 < a < 1$, și $\ln a > 0$ dacă $a > 1$.

Dacă ipoteza h este corectă pentru x , adică $h(x)$ are același semn cu y , atunci $e^{-y h(x)}$ este o valoare pozitivă mică (și cu atât mai mică cu cât $|h(x)|$ este mai mare), deci și valoarea funcției $l(y, h(x))$ este mică.

Dacă ipoteza h este incorrectă pentru x , adică $h(x)$ are semn contrar semnului lui y , atunci $e^{-y h(x)}$ este o valoare mare (și cu atât mai mare cu cât $|h(x)|$ este mai mare), deci și valoarea funcției $l(y, h(x))$ este mare.

De exemplu, dacă $y = 1$ și $h(x) = -1$ (sau $y = -1$ și $h(x) = 1$) atunci pierderea / costul este $-\ln \frac{1}{1 + e} = \ln(1 + e) \approx 1.31$, iar dacă $y = h(x) = 1$ (sau, ambele, -1) atunci costul este $-\ln \frac{1}{1 + e^{-1}} = \ln(1 + e^{-1}) \approx 0.31$.

b. Calculăm derivatele parțiale ale funcției $f(b, \bar{w}) = \sum_n -\ln \sigma(y_n(\bar{w} \cdot \bar{x}_n + b))$ în raport cu w_i și respectiv b . Vom folosi următoarele proprietăți:

- derivata sumei de funcții derivabile este suma derivatelor, adică $(f + g)' = f' + g'$;
- derivata unei compunerii de funcții derivabile este $(f \circ g)' = (f' \circ g)g'$;
- derivata logaritmului este $\ln'(x) = \frac{1}{x}$;
- derivata funcției sigmoidale este $\sigma'(x) = \sigma(x)(1 - \sigma(x))$.⁷⁶²

Vom obține:

$$\begin{aligned} \frac{\partial f}{\partial w_i} &= \frac{\partial}{\partial w_i} \sum_n -\ln \sigma(y_n(\bar{w} \cdot \bar{x}_n + b)) = -\sum_n \frac{\partial}{\partial w_i} \ln \sigma(y_n(\bar{w} \cdot \bar{x}_n + b)) \\ &= -\sum_n \frac{1}{\sigma(y_n(\bar{w} \cdot \bar{x}_n + b))} \frac{\partial}{\partial w_i} \sigma(y_n(\bar{w} \cdot \bar{x}_n + b)) \\ &= -\sum_n \frac{1}{\sigma(y_n(\bar{w} \cdot \bar{x}_n + b))} \overline{\sigma(y_n(\bar{w} \cdot \bar{x}_n + b))} (1 - \sigma(y_n(\bar{w} \cdot \bar{x}_n + b))) y_n x_{n,i} \\ &= -\sum_n (1 - \sigma(y_n(\bar{w} \cdot \bar{x}_n + b))) y_n x_{n,i}. \end{aligned}$$

⁷⁶⁰De exemplu, dacă f este funcție convexă, tot așa va fi și $e^{f(x)}$, întrucât e^x este funcție convexă și monoton crescătoare.

⁷⁶¹Ca exemplu puteți considera $f(x) = \ln x$ și $g(x) = e^{-x}$. Va rezulta funcția $h(x) = g(f(x)) = 1/x$ pentru $x > 0$, care este, într-adevăr, convexă.

⁷⁶²Vedeți ex. 32 de la capitolul *Metode de regresie*.

Similar, vom obține

$$\frac{\partial f}{\partial b} = - \sum_n (1 - \sigma(y_n(\bar{w} \cdot \bar{x}_n + b))) y_n.$$

Observație: La curs, b a fost asimilat cu w_0 , iar $x_{n,0} = 1$ prin definiție. Dacă s-ar fi procedat la fel și aici, atunci nu am fi avut de calculat decât o derivată la punctul a și una la punctul b .

c. Faptul că funcția $f(b, \bar{w})$ definită la punctul b este convexă în raport cu variabilele \bar{w} și b se demonstrează ușor, folosind cea de-a doua propoziție din Indicația din enunț. Într-adevăr, funcția univariată $l(z) \stackrel{\text{def.}}{=} -\ln \sigma(z)$ este convexă și strict descrescătoare (vedeți graficul din enunț),⁷⁶³ iar funcția $y(\bar{w} \cdot \bar{x} + b)$ este liniară în raport cu \bar{w} și cu b , deci este concavă (și convexă în același timp). Prin urmare, funcția $-\ln \sigma(y(\bar{w} \cdot \bar{x} + b))$ este convexă. În fine, se ține cont de faptul că o sumă de funcții convexe având același domeniu de definiție este de asemenea funcție convexă.

Observație: Pentru o demonstrație alternativă, și anume folosind matricea hessiană, vedeți problema 14 de la capitolul *Metode de regresie*, unde se lucrează cu regresia logistică.

În concluzie, valoarea optimă a funcției f este un minim, iar algoritmul de învățare automată bazat pe metoda gradientului descendente va găsi (la limită, eventual) acest minim. Antrenarea acestui perceptron se face similar cu antrenarea unității liniare, folosind reguli de actualizare de forma:

$$w_i \leftarrow w_i + \Delta w_i \text{ și } b \leftarrow b + \Delta b.$$

În cazul de fată, Δw_i și Δb sunt definite de expresiile

$$\Delta w_i = -\eta \frac{\partial f}{\partial w_i} = \eta \sum_n (1 - \sigma(y_n(\bar{w} \cdot \bar{x}_n + b))) y_n \bar{x}_{n,i}$$

și respectiv

$$\Delta b = -\eta \frac{\partial f}{\partial b} = \eta \sum_n (1 - \sigma(y_n(\bar{w} \cdot \bar{x}_n + b))) y_n.$$

15.

(Unități neuronale sigmoidale
[și rețele neuronale de unități sigmoidale]:
deducerea expresiei funcției de eroare de tip cross-entropie;
două variante: cu și, respectiv, fără „zgomot“)

■ • ○ *prelucrare de Liviu Ciortuz, după CMU, 2011 fall, Eric Xing, HW1, pr. 3.3*

Stim că pentru a antrena o unitate neuronală — și vom vedea mai târziu că este la fel și în cazul rețelelor neuronale — avem nevoie să definim o funcție de eroare / pierdere / cost, care poate fi minimizată cu ajutorul algoritmului de „învățare“ a ponderilor unității / rețelei. Până acum,⁷⁶⁴ am folosit [ca funcție de eroare] suma pătratelor erorilor. Se

⁷⁶³Faptul că funcția $l(z)$ este descrescătoare și convexă se poate demonstra în mod direct, cu mijloace clasice de analiză matematică.

⁷⁶⁴Adică: la toate problemele de până acum, cu excepția problemei 14.

poate arăta că obiectivul de a minimiza această funcție este implicit de principiul verosimilității maxime (engl., MLE), în contextul „procesării“ unui „semnal“ însotit de un „zgomot“ care urmează o distribuție de tip Gaussian.⁷⁶⁵ În acest exercițiu vom arăta că în cazul antrenării unităților sigmoidale se folosește un tip de funcție de cost care poate fi interpretată ca o cross-entropie (vedeți ex 64 de la capitolul de *Fundamente*).⁷⁶⁶

Fie datele de antrenament $D = (X, t) = \{(x_i, t_i)\}, i = 1, 2, \dots, N$. De exemplu, x_i (element dintr-un spațiu \mathbb{R}^d) poate fi imaginea unei fețe, în vreme ce t_i este o etichetă binară care are valoarea 0 dacă fața respectivă este a unui băiat/bărbat, respectiv 1 în cazul unei fete/femei.

Vom considera o unitate neuronală de tip sigmoidal.⁷⁶⁷ Notând cu y valoarea reală produsă de această unitate, rezultă că $y \in (0, 1)$. Este natural să interpretăm acest output ca fiind probabilitatea ca eticheta booleană t să fie 1. Formal, putem scrie $y \stackrel{\text{def}}{=} P(t = 1 | x; w)$. În consecință, este natural să căutăm valori convenabile pentru ponderile w folosind principiul verosimilității maxime (MLE), sub forma următoare:

$$w_{MLE} = \underset{w}{\operatorname{argmax}} \prod_{i=1}^N P(t_i | x_i; w).$$

a.⁷⁶⁸ Arătați că a maximiza expresia $\prod_{i=1}^N P(t_i | x_i; w)$ revine la a minimiza expresia următoare, despre care putem spune, după o observare atentă, că este o cross-entropie:⁷⁶⁹

$$E(w) = - \sum_{i=1}^N [t_i \ln y_i + (1 - t_i) \ln(1 - y_i)],$$

unde y_i este outputul produs de rețeaua neuronală pentru exemplul x_i .

b. Să presupunem că este posibil ca etichetele datelor de antrenament să fi fost puse eronat, și anume cu probabilitatea ε . Considerând că datele de antrenament sunt distribuite în mod identic și independent unele de altele, scrieți expresia funcției de eroare / cost $E^*(w, \varepsilon)$ care corespunde log-verosimilității cu semn schimbat (engl., the negative log likelihood).

c. Verificați că funcția de eroare $E(w)$ se obține din $E^*(w, \varepsilon)$ pentru cazul particular $\varepsilon = 0$.

d. Explicați de ce anume funcția de eroare $E^*(w, \varepsilon)$ va face ca modelul obținut să fie [mai] robust în raport cu date incorrect etichetate, spre deosebire de funcția uzuală $E(w)$.

Răspuns:

⁷⁶⁵Pentru a vedea cum se face deducerea expresiei funcției de eroare ca sumă a pătratelor erorilor într-un context particular (însă sugestiv), cititorul poate consulta problema 13.

⁷⁶⁶De fapt, tipul *sigmoidal* al funcției de activare nu este important în acest context. Singurul fapt determinant este acela că output-ul este un număr din intervalul (0, 1), deci poate fi interpretat ca o probabilitate.

⁷⁶⁷Mai general, se poate considera o rețea formată din unități sigmoidale. (De fapt, aşa a fost formulată problema originală de la CMU.) Într-un astfel de caz este însă suficient să presupunem că tipul unității / unităților de pe nivelul de ieșire din rețea este cel sigmoidal.

⁷⁶⁸Acest punct corespunde problemei 13.a de la capitolul *Metode de regresie*.

⁷⁶⁹A se vedea problema 64 de la capitolul de *Fundamente*.

a. Conform enunțului, $y_i \stackrel{\text{def.}}{=} P(t_i = 1|x_i; w)$. Așadar, putem scrie:

$$P(t_i|x_i; w) = \begin{cases} y_i & \text{dacă } t_i = 1, \\ 1 - y_i & \text{dacă } t_i = 0. \end{cases}$$

Se verifică imediat că, din punct de vedere matematic, relația de mai sus se poate exprima în mod echivalent sub forma

$$P(t_i|x_i; w) = (y_i)^{t_i} (1 - y_i)^{1-t_i}.$$

Această expresie, în ciuda faptului că este mai puțin intuitivă, este mult mai convenabilă pentru calculele pe care trebuie să le facem în vederea aplicării principiului verosimilității maxime:⁷⁷⁰

$$\begin{aligned} w_{MLE} &\stackrel{\text{def.}}{=} \operatorname{argmax}_w P(t_1, \dots, t_n | x_1, \dots, x_n; w) \stackrel{i.i.d.}{=} \operatorname{argmax}_w \prod_{i=1}^N P(t_i | x_i; w) \\ &= \operatorname{argmax}_w \sum_{i=1}^N \ln P(t_i | x_i; w) = \operatorname{argmax}_w \sum_{i=1}^N \ln(y_i)^{t_i} (1 - y_i)^{1-t_i} \\ &= \operatorname{argmax}_w \sum_{i=1}^N (t_i \ln y_i + (1 - t_i) \ln(1 - y_i)) = \operatorname{argmax}_w (-E(w)) \\ &= \operatorname{argmin}_w E(w). \end{aligned}$$

Pentru penultima egalitate de mai sus, veți formula de definiție a funcției E din enunț.

Observație: Pentru demonstrarea faptului că $E(w)$ este funcție convexă (echivalent: $-E(w)$ este funcție concavă), veți problema 14 de la capitolul *Metode de regresie*.

b. În cazul în care etichetarea datelor de antrenament s-a făcut în mod eronat, cu probabilitatea ε , rezultă imediat că

$$x_i \text{ a fost etichetat cu } \begin{cases} 1 & \text{cu probabilitatea } 1 - \varepsilon, \text{ dacă } t_i = 1, \\ 0 & \text{cu probabilitatea } \varepsilon, \text{ dacă } t_i = 1, \\ 0 & \text{cu probabilitatea } 1 - \varepsilon, \text{ dacă } t_i = 0, \\ 1 & \text{cu probabilitatea } \varepsilon, \text{ dacă } t_i = 0. \end{cases}$$

Prin urmare, putem scrie în manieră condensată

$$P(t_i|x_i; w) = \begin{cases} y_i(1 - \varepsilon) + (1 - y_i)\varepsilon & \text{dacă } t_i = 1, \\ (1 - y_i)(1 - \varepsilon) + y_i\varepsilon & \text{dacă } t_i = 0. \end{cases}$$

Ba chiar și mai mult, această expresie este echivalentă cu următoarea:

$$P(t_i|x_i; w) = [y_i(1 - \varepsilon) + (1 - y_i)\varepsilon]^{t_i} [(1 - y_i)(1 - \varepsilon) + y_i\varepsilon]^{1-t_i}.$$

Făcând un calcul similar cu cel de la punctul a, va rezulta:⁷⁷¹

$$E^*(w, \varepsilon) = - \sum_{i=1}^N \{t_i \ln[y_i(1 - \varepsilon) + (1 - y_i)\varepsilon] + (1 - t_i) \ln[(1 - y_i)(1 - \varepsilon) + y_i\varepsilon]\}.$$

⁷⁷⁰De fapt, în cazul de față vom lucra cu verosimilitate condițională.

⁷⁷¹Se verifică imediat că $[y_i(1 - \varepsilon) + (1 - y_i)\varepsilon] + [(1 - y_i)(1 - \varepsilon) + y_i\varepsilon] = 1$.

c. Înlocuind parametrul ε cu valoarea 0 în expresia pe care tocmai am obținut-o mai sus, vom avea:

$$\begin{aligned} E^*(w, 0) &= - \sum_{i=1}^N \{[t_i \ln y_i + (1 - y_i) \times 0] + (1 - t_i) \ln[(1 - y_i) + y_i \times 0]\} \\ &= - \sum_{i=1}^N [t_i \ln y_i + (1 - t_i) \ln(1 - y_i)] = E(w). \end{aligned}$$

d. Vom considera ca *exemplu* cazul extrem (însă instructiv) în care toate etichetele au fost puse incorrect. Dacă am folosi funcția de eroare $E(w)$, atunci modelul produs de algoritmul de retro-propagare ar fi complet eronat. Dacă în schimb vom folosi funcția de eroare $E^*(w, \varepsilon)$ cu $\varepsilon = 1$, atunci algoritmul de retro-propagare va încerca să învețe un model care clasifică datele exact invers față de etichetele originale, ceea ce corespunde obiectivului nostru.

Notă: Problema 50 cere să se adapteze algoritmul de retro-propagare folosit pentru antrenarea rețelelor neuronale de tip "feed-forward" (vedeți pr. 20) pentru cazul în care funcția de optimizat este de tip cross-entropie, aşa cum a fost definită în acest exercițiu.

16.

(Algoritmul de antrenare a perceptronului-prag,
versiunea Rosenblatt: aplicare)

• prelucrare de Liviu Ciortuz, după

CMU, 2013 fall, W. Cohen, E. Xing, Sample questions, pr. 2

CMU, 2013 fall, A. Smola, G. Gordon, midterm ex. practice, pr. 10

CMU, 2013 spring, A. Smola, B. Poczos, HW2, pr. 2

Se consideră o secvență de exemple $\bar{x}_i \in \mathbb{R}^d$, pentru $i = 1, \dots, n$, împreună cu etichetele corespunzătoare $y_i \in \{-1, 1\}$. Vom antrena perceptronul-prag folosind următorul algoritm, cunoscut în literatura de specialitate sub numele de *perceptronul Rosenblatt*:

```

initialize  $\bar{w} \leftarrow \bar{0}$ 
for  $i = 1, \dots, n$ 
    if  $y_i (\bar{w} \cdot \bar{x}_i) \leq 0$  then
         $\bar{w} \leftarrow \bar{w} + y_i \bar{x}_i$ 
    end if
end for

```

unde operatorul \cdot reprezintă produsul scalar, vectorii $\bar{x}_i \in \mathbb{R}^d$ desemnează instanțele de antrenament, iar etichetele asignate lor sunt notate (ca de obicei) cu y_i .

Observație: În pseudo-codul folosit mai sus se consideră că termenul liber w_0 fie este absent fie este inclus în \bar{w} și, în acest ultim caz, prima componentă a vectorului \bar{x}_i (notată cu $x_{0,i}$) este prin convenție 1.

a. Calculați actualizările vectorului de parametri \bar{w} folosind algoritmul de mai sus pe următorul set de exemple (considerând că nu se folosește termenul liber w_0):

$$\begin{aligned}\bar{x}_1 &= (0, 0, 0, 1, 0, 0, 1), \quad y_1 = 1 \\ \bar{x}_2 &= (1, 1, 0, 0, 0, 1, 0), \quad y_2 = -1 \\ \bar{x}_3 &= (0, 0, 1, 1, 0, 0, 0), \quad y_3 = 1 \\ \bar{x}_4 &= (1, 0, 0, 0, 1, 1, 0), \quad y_4 = -1 \\ \bar{x}_5 &= (1, 0, 0, 0, 0, 1, 0), \quad y_5 = -1\end{aligned}$$

b. Verificați dacă separatorul obținut la finalul execuției algoritmului va clasifica perfect toate datele de antrenament.

c. Care sunt diferențele ce caracterizează acest algoritm în raport cu algoritmul prezentat în cartea *Machine Learning* a lui Tom Mitchell (pag. 88-94), care este bazat pe aşa-numita *regulă delta*?⁷⁷² Vă puteți referi de exemplu la modul de lucru “batch” vs. modul incremental / stochastic, inițializarea ponderilor, rata de învățare, și criteriul de oprire.

Răspuns:

a. Conform algoritmului dat, avem:

$$\begin{aligned}i = 1 : \quad \bar{w} &= \bar{0} \stackrel{\text{not.}}{=} (0, 0, 0, 0, 0, 0) \Rightarrow y_1(\bar{w} \cdot \bar{x}_1) = 0 \leq 0 \\ &\Rightarrow \bar{w} \leftarrow y_1 \bar{x}_1 = \bar{x}_1 = (0, 0, 0, 1, 0, 0, 1) \\ i = 2 : \quad y_2(\bar{w} \cdot \bar{x}_2) &= -(0, 0, 0, 1, 0, 0, 1) \cdot (1, 1, 0, 0, 0, 1, 0) = 0 \leq 0 \\ &\Rightarrow \bar{w} \leftarrow \bar{w} + y_2 \bar{x}_2 = \bar{x}_1 - \bar{x}_2 = (-1, -1, 0, 1, 0, -1, 1) \\ i = 3 : \quad y_3(\bar{w} \cdot \bar{x}_3) &= (-1, -1, 0, 1, 0, -1, 1) \cdot (0, 0, 1, 1, 0, 0, 0) = 1 > 0 \\ i = 4 : \quad y_4(\bar{w} \cdot \bar{x}_4) &= -(-1, -1, 0, 1, 0, -1, 1) \cdot (1, 0, 0, 0, 1, 1, 0) = 2 > 0 \\ i = 5 : \quad y_5(\bar{w} \cdot \bar{x}_5) &= -(-1, -1, 0, 1, 0, -1, 1) \cdot (1, 0, 0, 0, 0, 1, 0) = 2 > 0.\end{aligned}$$

b. Știm deja de la punctul precedent că $y_i(\bar{w} \cdot \bar{x}_i) > 0$ pentru $i \in \{3, 4, 5\}$. Așadar, va trebui să mai verificăm doar dacă are loc această relație și pentru $i \in \{1, 2\}$:

$$\begin{aligned}i = 1 : \quad y_1(\bar{w} \cdot \bar{x}_1) &= (-1, -1, 0, 1, 0, -1, 1) \cdot (0, 0, 0, 1, 0, 0, 1) = 2 > 0 \\ i = 2 : \quad y_2(\bar{w} \cdot \bar{x}_2) &= -(-1, -1, 0, 1, 0, -1, 1) \cdot (1, 1, 0, 0, 0, 1, 0) = 3 > 0.\end{aligned}$$

În concluzie, separatorul $\bar{w} = (-1, -1, 0, 1, 0, -1, 1)$ clasifică perfect datele de antrenament.

c. Este imediat că perceptronul Rosenblatt corespunde modului de lucru incremental: el face actualizarea ponderilor după fiecare instanță de antrenament \bar{x}_i pentru care $\text{sign}(\bar{w} \cdot \bar{x}_i)$, i.e., clasificarea produsă de actualul \bar{w} , nu coincide cu eticheta y_i .

Inițializarea ponderilor se face în acest algoritm cu 0. În algoritmul clasic (cel din cartea lui Tom Mitchell), inițializarea ponderilor w_j , pentru $j = 1, \dots, d$, se face cu numere mici (în modul), alese în mod aleatoriu.

Forma regulii de actualizare a ponderilor perceptronului Rosenblatt nu coincide cu *regula delta*:

$$\bar{w} \leftarrow \bar{w} + \eta(y_i - o_i)\bar{x}_i \tag{341}$$

⁷⁷²Vedeți formula (4.10) din cartea citată, pag. 93.

unde $\eta > 0$ este rata de învățare. Legătura dintre cele două forme se poate face în felul următor:

Este evident că regula (341) are efect doar în cazul în care $y_i \neq o_i$, ceea ce echivalează cu $y_i(\bar{w} \cdot \bar{x}_i) \leq 0$ din algoritm Rosenblatt. Perceptronul-prag folosește funcția de activare *sign*, aşadar faptul că y_i este diferit de o_i implică fie $y_i = 1$ și $o_i = -1$, fie $y_i = -1$ și $o_i = 1$. În ambele situații rezultă $y_i - o_i = 2y_i$. Prin urmare, folosind pre-condiția $y_i(\bar{w} \cdot \bar{x}_i) \leq 0$, regula (341) devine

$$\bar{w} \leftarrow \bar{w} + 2\eta y_i x_i$$

Luând $\eta = 1/2$, obținem forma regulii din algoritm Rosenblatt.

Algoritmul Rosenblatt face o singură trecere prin setul de instanțe $\{x_i\}_i$. Evident, se poate extinde în mod natural acest algoritm pentru a proceda ca în cazul clasic, permitând trecerea de mai multe ori prin setul de instanțe, eventual până la convergență, adică până când în cursul unei aceleiași iterații nu se mai fac deloc ajustări ale ponderilor w (presupunând că că setul de instanțe de antrenament este liniar separabil).

17. (Perceptronul Rosenblatt, câteva proprietăți simple:
atât numărul de greșeli cât și poziția finală a separatorului
depind de ordinea de furnizare a exemplelor)

■ • CMU, 2015 spring, T. Mitchell, N. Balcan, HW6, pr. 3.ab
CMU, 2017 spring, Barnabas Poczos, HW3, pr. 1.bc

Considerăm că rulăm algoritmul *Perceptron* al lui Rosenblatt,⁷⁷³ folosind un anumit sir de exemple S . (Vă reamintim că un *exemplu* este o *instanță* — punct din \mathbb{R}^d — împreună cu *eticheta* sa). Fie S' același set de exemple ca și S , însă prezentate într-o altă ordine.

- a. Face oare algoritmul *Perceptron* același număr de greșeli pe sirul / succesiunea de exemple S ca și pe sirul S' ? Dacă da, de ce? Dacă nu, indicați două astfel de siruri (S și S') pe care algoritmul *Perceptron* face un număr diferit de greșeli.
- b. Din punct de vedere geometric, algoritmul *Perceptron* clasifică instanțele date conform poziției lor relativ la un hiperplan de separare. Vectorul de ponderi (\bar{w}) învățat de către *Perceptron* va fi normal [adică, perpendicular] pe acest hiperplan? Răspundeți prin *Adevărat* sau *Fals* și justificați în mod riguros.

Răspuns:

- a. Considerăm următorul set de exemple (S), în ordinea indicată:

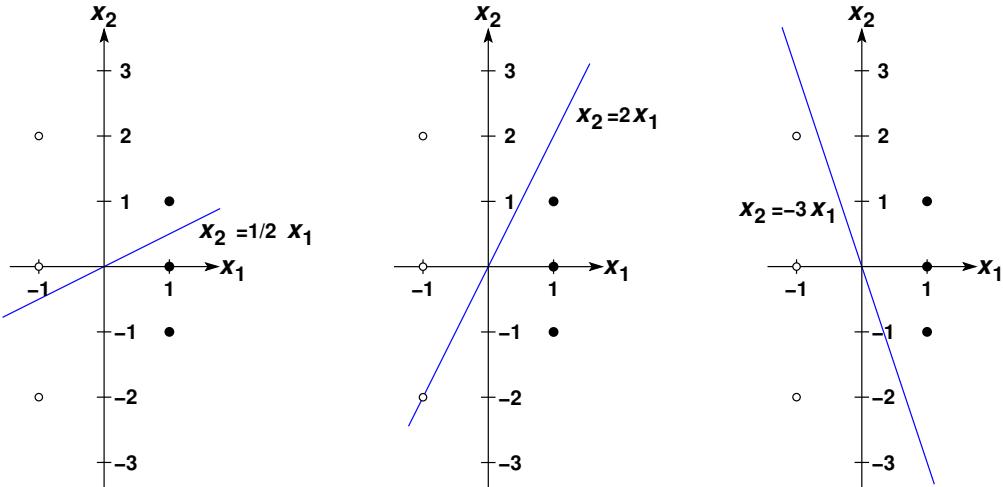
exemplul	1	2	3	4	5	6
instanța (x_1, x_2)	$(-1, 2)$	$(1, 0)$	$(1, 1)$	$(-1, 0)$	$(-1, -2)$	$(1, -1)$
eticheta y	-1	+1	+1	-1	-1	+1

⁷⁷³Pentru pseudo-cod, vedeți enunțul problemei 16.

Sintetizăm execuția algoritmului *Perceptron* pe aceste date, fără a folosi termenul “bias” (w_0), alcătuind tabelul următor:

iterația(i)	\bar{x}_i	y_i	$y_i \bar{w} \cdot \bar{x}_i$	\bar{w}	greșală	ec. separatorului
initializare	—	—	(0, 0)	—	—	
1	(-1, 2)	-1	0 ≤ 0	(1, -2)	da	$x_2 = 0.5x_1$
2	(1, 0)	+1	1 > 0	(1, -2)	nu	$x_2 = 0.5x_1$
3	(1, 1)	+1	-1 ≤ 0	(2, -1)	da	$x_2 = 2x_1$
4	(-1, 0)	-1	2 > 0	(2, -1)	nu	$x_2 = 2x_1$
5	(-1, -2)	-1	0 ≤ 0	(3, 1)	da	$x_2 = -3x_1$
6	(1, -1)	+1	2 > 0	(3, 1)	nu	$x_2 = -3x_1$

Așadar, algoritmul a comis trei greșeli. Cei trei separatori obținuți în cursul aplicării algoritmului *Perceptron* pe sirul de exemple S sunt prezentate în figura următoare.



Este imediat că modul în care se construiește separatorul — a cărui expresie analitică este o combinație liniară de instanțe greșit catalogate — este dependent de exemplele prezентate. Întrebarea care a fost pusă în enunț este dacă pentru două succesiuni / ordini diferite, rezultatele finale (i.e., expresiile separatorilor obținuți) coincid (întotdeauna) sau nu. În continuare vom arăta că în mod obișnuit rezultatele finale (i.e., pozițiile finale ale separatorilor obținuți) nu coincid.

Considerăm S' secvența de exemple obținută din S prin inversarea [ordinii] exemplelor \bar{x}_1 și \bar{x}_2 . Este imediat că la prima iterare perceptronul greșește și apoi calculează $\bar{w} = (1, 0)$, ceea ce conduce la ecuația separatorului $x_1 = 0$, adică axa Ox_2 . Evident, aceasta dreaptă este un separator perfect pentru întregul set de exemple, deci algoritmul nu va mai comite până la final nicio [altă] greșală. *Concluzia* este că, pe un același set de exemple de antrenament, algoritmul *Perceptron* poate comite un număr diferit de greșeli (și, de asemenea, un alt separator) dacă exemplele îi sunt prezentate în două secvențe diferite.

Observație importantă: Deși în ambele cazuri de mai sus (S și S') algoritmul *Perceptron* găsește un separator liniar pentru datele de intrare, acest fapt nu este garantat în general, chiar dacă datele sunt liniar separabile. Vedeti

de exemplu problema 38. (Motivele sunt: rata de învățare (implicită) prea mare și / sau neparcurgerea setului de date decât o singură dată.)

b. Cunoaștem din geometria analitică faptul că orice doi vectori \bar{w} și \bar{x} pentru care are loc relația $\bar{w} \cdot \bar{x} = 0$ sunt ortogonali (adică, perpendiculari unul pe celălalt).

În cazul perceptronilor (de tip liniar, prag sau sigmoidal), ecuația separatorului este tocmai de forma $\bar{w} \cdot \bar{x} = 0$, unde \bar{x} este un punct arbitrar de pe dreapta [sau, în general, hiperplanul] separator. Considerând \bar{x}_1 și \bar{x}_2 două astfel de puncte, diferența lor, $\bar{x}_1 - \bar{x}_2$, va fi un vector care are direcția dreptei-separator. Din relațiile $\bar{w} \cdot \bar{x}_1 = 0$ și $\bar{w} \cdot \bar{x}_2 = 0$ rezultă $\bar{w} \cdot (\bar{x}_1 - \bar{x}_2) = 0$. Prin urmare, vectorul de ponderi \bar{w} este perpendicular pe direcția dreptei-separator.

Proprietatea aceasta se poate verifica în mod particular pe datele de la punctele precedente. De exemplu, la punctul a , pe sirul de exemple S , la iterată 1 avem $\bar{w} = (1, -2)$ și se poate observa direct pe grafic că acest vector este perpendicular pe direcția dreptei $y = \frac{1}{2}x$.

Așadar, răspunsul la întrebarea din enunț este *Adevărat*.

18.

(Convergența algoritmului de antrenare a perceptronului-prag, varianta Rosenblatt)

*prelucrare de Liviu Ciortuz, 2014, după
 ■ □ • Tommy Jaakkola, MIT, ML course, 2009 fall, lecture notes 2,
 cf. Block and Novikoff's results, 1962
 (see also CMU, 2016 fall, E. Xing, Z. Bar-Joseph, HW3, pr. 3
 CMU, 2014 fall, E. Xing, B. Poczos, HW1, pr. 4.3)*

Presupunem că folosim perceptronul de tip prag în varianta Rosenblatt (vezi și problema 16) și vrem să învățăm un concept, folosind instanțele de antrenament $x_1, \dots, x_n, \dots \in \mathbb{R}^d$ împreună cu etichetele corespunzătoare $y_1, \dots, y_n, \dots \in \{-1, 1\}$.

Demonstrați că în cazul în care sunt îndeplinite condițiile *i-iv* de mai jos, algoritmul de actualizare a ponderilor perceptronului „converge“, adică termină într-un număr finit de pași. Formal, exprimăm acest fapt astfel: $\exists m \in \mathbb{N}$ astfel încât $y_t w^{(m)} \cdot x_t > 0$ pentru orice $t \in \{1, \dots, n, \dots\}$, unde $w^{(m)}$ este vectorul de ponderi obținut de perceptron la iterată m .

Iată acum *condițiile* menționate mai sus:

- i. Instanțele x_1, \dots, x_n, \dots sunt separabile liniar prin originea sistemului de coordinate, cu o margine finită $\gamma > 0$. Din punct de vedere formal, aceasta înseamnă că există $w^* \in \mathbb{R}^d$ astfel încât $y_t w^* \cdot x_t \geq \gamma$ pentru $t = 1, \dots, n, \dots$
- ii. Toate instanțele x_1, \dots, x_n, \dots sunt conținute într-o sferă din \mathbb{R}^d cu centrul în origine, adică $\exists R > 0$ astfel încât $\|x_t\| \stackrel{\text{def}}{=} \sqrt{x_t \cdot x_t} \leq R$ pentru orice t .
- iii. Învățarea se face în manieră *incrementală*, folosind următoarea regulă de actualizare a ponderilor:

$$w^{(k+1)} = w^{(k)} + y_{t_k} x_{t_k} \text{ pentru un } t_k \in \{1, \dots, n, \dots\} \text{ a.i. } y_{t_k} w^{(k)} \cdot x_{t_k} \leq 0, \quad (342)$$

ceea ce înseamnă că instanța x_{t_k} este clasificată eronat de către perceptron la iterată k .

iv. Startarea procesului de învățare se face cu $w^{(0)} = 0 \in \mathbb{R}^d$.

Indicație: Arătați că la fiecare iterare (k) a algoritmului, sunt satisfăcute următoarele proprietăți:

- a. $w^* \cdot w^{(k)} \geq k\gamma$;
- b. $\|w^{(k)}\|^2 \leq kR^2$;
- c. $k \leq \left(\frac{\|w^*\|}{\gamma} R\right)^2$.

Ultima inegalitate indică [o margine superioară pentru] numărul maxim de iterări executate de către perceptron.

Observația 1: Notând cu θ_t unghiul format de vectorii x_t și w^* în \mathbb{R}^d și ținând cont de faptul că

$$\cos(\theta_t) = \cos(x_t, w^*) = \frac{x_t \cdot w^*}{\|x_t\| \|w^*\|},$$

deci

$$y_t w^* \cdot x_t = y_t \|w^*\| \|x_t\| \cos(\theta_t),$$

ceea ce, coroborat cu condiția i , implică $y_t \|x_t\| \cos(\theta_t) \|w^*\| \geq \gamma$, adică $y_t \|x_t\| \cos(\theta_t) \geq \frac{\gamma}{\|w^*\|}$. Din punct de vedere geometric, aceasta înseamnă că în spațiul \mathbb{R}^d distanța de la orice instanță x_t la hiperplanul de separare care are ecuația $w^* \cdot x = 0$ (și care trece prin originea sistemului de coordonate) este mai mare sau egală cu $\frac{\gamma}{\|w^*\|}$.

Observația 2: Separatorul $w^{(k)}$ este o combinație liniară de instanțele x_i , întrucât regula de actualizare este $w^{(k+1)} = w^{(k)} + y_i x_i$. Observația aceasta este valabilă și la antrenarea perceptronului-prag clasic, bazat pe regula delta.

Răspuns:

Algoritmul de actualizare a ponderilor perceptronului determină schimbarea poziției separatorului la fiecare iterare (k) la care avem de a face cu un exemplu clasificat greșit. Intuitiv, ne așteptăm ca poziția separatorului la iterare $k+1$ (adică hiperplanul de ecuație $w^{(k+1)} \cdot x = 0$) să se apropie de poziția separatorului w^* care definește conceptul de învățat. În consecință, cosinusul unghiului dintre $w^{(k)}$ și w^* ar trebui să crească de la o iterare la alta. Pentru a dovedi / verifica în mod riguros aceasta, vom ține cont că, prin definiție,

$$\cos(w^{(k)}, w^*) = \frac{w^{(k)} \cdot w^*}{\|w^{(k)}\| \|w^*\|}.$$

Mai întâi vom compara valorile produsului scalar de la numărătorul fracției de mai sus, la două iterări successive. Folosind regula (342), putem scrie:

$$w^{(k+1)} \cdot w^* = (w^{(k)} + y_{t_k} x_{t_k}) \cdot w^* = w^{(k)} \cdot w^* + y_{t_k} x_{t_k} \cdot w^*.$$

Întrucât $y_{t_k} x_{t_k} \cdot w^* \geq \gamma$ (vedeți condiția i din enunț), rezultă că valoarea produsului scalar $w^{(k)} \cdot w^*$ crește la fiecare iterare cu o cantitate cel puțin egală cu γ . Cum $w^{(0)} = 0$ conform restricției *iv*, este imediat că $w^{(k)} \cdot w^* \geq k\gamma$ la iterare k . Cu aceasta, tocmai am demonstrat pe de o parte că relația *a* este adevărată,

iar pe de altă parte că numărătorul fracției care definește $\cos(w^{(k)}, w^*)$ crește cel puțin liniar în raport cu k .

A analiza evoluția numitorului fracției care definește $\cos(w^{(k)}, w^*)$ revine la a compara $\|w^{(k+1)}\|$ cu $\|w^{(k)}\|$, întrucât $\|w^*\|$ este constant în raport cu k .

$$\begin{aligned} \|w^{(k+1)}\|^2 &\stackrel{\text{def.}}{=} w^{(k+1)} \cdot w^{(k+1)} \stackrel{(342)}{=} (w^{(k)} + y_{t_k} x_{t_k}) \cdot (w^{(k)} + y_{t_k} x_{t_k}) \\ &= w^{(k)} \cdot w^{(k)} + 2y_{t_k} w^{(k)} \cdot x_{t_k} + y_{t_k}^2 x_{t_k} \cdot x_{t_k} \\ &= \|w^{(k)}\|^2 + 2y_{t_k} w^{(k)} \cdot x_{t_k} + y_{t_k}^2 \|x_{t_k}\|^2 \\ &= \|w^{(k)}\|^2 + 2y_{t_k} w^{(k)} \cdot x_{t_k} + \|x_{t_k}\|^2 \\ &\leq \|w^{(k)}\|^2 + R^2. \end{aligned}$$

Inegalitatea de mai sus derivă din faptul că $y_{t_k} w^{(k)} \cdot x_{t_k} \leq 0$ (i.e., exemplul t_k este clasificat greșit la iterația k , conform condiției *iii* din enunț) și din ipoteza că orice instanță de antrenament este conținută în sferă de rază R și având centrul în originea spațiului \mathbb{R}^d , conform condiției *ii*. Cum $w^{(0)} = 0$, este imediat că $\|w^{(k)}\|^2 \leq kR^2$, deci $\|w^{(k)}\| \leq \sqrt{k}R$ la iterația k . Cu aceasta, am demonstrat relația *b*.

Acum, combinând rezultatele *a* și *b*, obținem următoarea inegalitate:

$$\cos(w^{(k)}, w^*) = \frac{w^{(k)} \cdot w^*}{\|w^{(k)}\| \|w^*\|} \geq \frac{k\gamma}{\sqrt{k}R \|w^*\|} = \sqrt{k} \frac{\gamma}{R \|w^*\|}.$$

Stim că valoarea funcției cos este întotdeauna cel mult egală cu 1. În consecință,

$$1 \geq \cos(w^{(k)}, w^*) \geq \sqrt{k} \frac{\gamma}{R \|w^*\|} \Rightarrow k \leq \left(R \frac{\|w^*\|}{\gamma} \right)^2,$$

deci am demonstrat relația *c*. Aceasta înseamnă că în condițiile specificate în enunț, antrenarea perceptronului-prag se va face în cel mult $\left\lfloor \left(R \frac{\|w^*\|}{\gamma} \right)^2 \right\rfloor$ iterații, unde perechea de simboluri $\lfloor \rfloor$ desemnează funcția parte întreagă inferioară.

Observația 3: Marginea superioară calculată mai sus este remarcabilă, întrucât ea nu depinde nici de instanțele x_i și nici de dimensiunea (d) a spațiului din care sunt selectate aceste instanțe.

Observația 4: Restricția referitoare la *separabilitatea prin origine* a instanțelor de antrenament (vedeți condiția *i*, prima parte) — și anume, termenul „liber“ w_0 asociat separatorului w^* trebuie să fie 0 — nu este de fapt limitativ.⁷⁷⁴ Cazul general al separabilității liniare în \mathbb{R}^d , adică $y_t(w_0^* + w^* \cdot x_t) \geq 0$ pentru orice $t \in \{1, 2, \dots, n\}$, poate fi redus la cazul particular al separabilității prin origine în \mathbb{R}^{d+1} dacă pe de o parte se consideră $w' = (w^0, w^1, \dots, w^d)$, cu $w^* = (w^1, \dots, w^d)$, iar pe de altă parte se mapează toate instanțele de antrenament $x_t = (x_t^1, \dots, x_t^d)$ din \mathbb{R}^d în \mathbb{R}^{d+1} astfel: $x'_k = (x_t^0, x_t^1, \dots, x_t^d)$, cu $x_t^0 = 1$ pentru orice t . Cu această mapare, rezultă $y_t w' \cdot x'_t \geq 0$ pentru orice t (respectiv $y_t w' \cdot x'_t \geq \gamma$ când se lucrează cu margine finită, ca în enunțul acestei probleme). Nici restricția *iv* ($w^{(0)} = 0$) nu este cu adevărat limitativă; în general algoritmii de antrenare a unităților / rețelelor neuronale inițializează ponderile w la valori mici (în modul).

⁷⁷⁴În esență, ea ne ușurează calculele pe parcursul demonstrației, fiindcă ne plasează într-o situație particulară.

Similar, este imediat că *restricția* potrivit căreia sfera în care sunt conținute instanțele x_1, \dots, x_n, \dots trebuie să aibă centrul în originea lui \mathbb{R}^d (vedeți condiția *ii*, partea a doua) poate fi eliminată fără ca rezultatul de convergență să fie afectat.

Observația 5: Deducerea marginii superioare de la punctul c de mai sus în cazul în care se folosește o rată de învățare oarecare $\eta > 0$ se face extinzând în mod natural demonstrația de mai sus (vedeți pr. 40). În concluzie, mai rămân de examinat două condiții: separabilitatea liniară cu margine $\gamma > 0$ și conținerea instanțelor de antrenament într-o sferă de rază finită. Problema 39.B va cere să se demonstreze că ambele condiții sunt esențiale pentru convergența perceptronului Rosenblatt în regim de învățare online.

19.

(Perceptronul Rosenblatt [dual] kernel-izat)

*prelucrare de Liviu Ciortuz, după***■ CMU, 2013 spring, A. Smola, B. Poczos, HW2, pr. 2.d****Stanford, 2016 fall, A. Ng, J. Duchi, HW2, pr. 2****MIT, 2006 fall, Tommy Jaakkola, HW2, pr. 3**

Majoritatea clasificatorilor liniari pot fi kernel-izați, în vederea clasificării de date care sunt neseparabile liniar. În acest exercițiu vă cerem să elaborați varianta kernel-izată a algoritmului *Perceptron* [dual]. (Pentru pseudo-codul variantei simple, nekernel-izate a algoritmului *Perceptron*, vedeți pr. 16.)

Ca input, acest algoritm va lua secvența de instanțe etichetate $\{(x_i, y_i)\}_{i=1}^n$ cu $x_i \in \mathbb{R}^d$ și $y_i \in \{-1, 1\}$, precum și funcția-nucleu $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$, cu proprietatea că există $m \in \mathbb{N}^*$ (în general, $m > d$) și o funcție („mapare“) $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^m$ astfel încât $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$ pentru orice x_i, x_j .⁷⁷⁵ Perceptronul kernel-izat [dual] va lucra nu cu instanțele x_i , ci cu imaginile lor, $\phi(x_i)$, și va încerca să găsească pentru acestea un separator liniar în spațiul \mathbb{R}^m .

Comentariu: Analizând algoritmul *Perceptron*, veți constata că vectorul de ponderi $w \in \mathbb{R}^d$ este o combinație liniară de instanțe x_i , desemnată prin expresia $w = \sum_{i=1}^n \alpha_i x_i$. Din acest motiv este suficient să găsim valorile coeficienților α_i care asigură separabilitatea datelor. Această proprietate se dovedește deosebit de utilă în cazul funcțiilor de mapare (ϕ) cu vectori de „trăsături“ ($\phi(x)$) de dimensiuni foarte mari sau chiar infiniti, aşa cum este cazul funcțiilor-nucleu gaussiene, numite și *funcții cu bază radială* (engl., Radial Basis Functions, RBF). Așadar, în loc să actualizeze vectorul de ponderi w (cum face *Perceptronul simplu*), *Perceptronul kernel-izat [dual]* actualizează direct coeficienții $\alpha_1, \dots, \alpha_n$. Se va vedea că în timpul actualizărilor, operațiile care se fac asupra instanțelor de antrenament date (x_i) sunt *doar* produse scalare de forma $\phi(x_i) \cdot \phi(x_j)$, care, aşa cum am precizat mai sus, pot fi văzute ca valori ale funcției-nucleu, $K(x_i, x_j)$. Această [a doua] proprietate asigură eficiența calculelor.

Arătați cum anume se scrie regula de actualizare a coeficienților $\alpha_1, \dots, \alpha_n$ la procesarea unui nou exemplu (x_i, y_i) de către algoritmul *Perceptron kernel-izat [dual]*, și cum se face predicția ($y = 1$ sau -1) pentru o instanță nouă x .

Răspuns:

În spațiul \mathbb{R}^m (care se mai numește, în contextul kernel-izării, *spațiul de trăsături*), regula de actualizare a ponderilor w se scrie astfel:

⁷⁷⁵Din punct de vedere practic, este necesar ca funcția K să poată fi calculată în mod eficient.

$$w^{(i)} \leftarrow w^{(i-1)} + y_i \phi(x_i) \text{ dacă } y_i w^{(i-1)} \cdot \phi(x_i) \leq 0,$$

unde operatorul \cdot desemnează produsul scalar.

În consecință, întrucât facem inițializarea $w^{(0)} = \bar{0}$, vectorul $w \in \mathbb{R}^m$ va fi întotdeauna o combinație liniară de vectori de trăsături, $\phi(x_i)$. Aceasta înseamnă că există coeficienții $\alpha_l \in \mathbb{R}$ astfel încât $w^{(i)} = \sum_{l=1}^{i-1} \alpha_l \phi(x_l)$ după procesarea primelor i instanțe de antrenament. Așadar, vectorul $w^{(i)}$ poate fi reprezentat în mod compact prin coeficienții α_l (cu $l = 1, \dots, i$) din această combinație liniară. În particular, valoarea inițială $w^{(0)}$ corespunde cazului când suma nu conține niciun termen (adică avem o listă vidă de coeficienți α_l).

Arătăm acum că *putem calcula în mod eficient coeficienții α_i* .

La iteratăia i trebuie să calculăm produsul scalar $w^{(i-1)} \cdot \phi(x_i)$. Întrucât produsul scalar în spațiul de trăsături \mathbb{R}^m este o operație costisitoare atunci când m este mare, vom ține cont că

$$w^{(i-1)} \cdot \phi(x_i) = \left(\sum_{l=1}^{i-1} \alpha_l \phi(x_l) \right) \cdot \phi(x_i) = \sum_{l=1}^{i-1} \alpha_l (\phi(x_l) \cdot \phi(x_i)) = \sum_{l=1}^{i-1} \alpha_l K(x_l, x_i),$$

ceea ce înseamnă că acești coeficienți se pot calcula într-adevăr în mod eficient.

În mod similar, se poate face în mod eficient predicția clasei / etichetei pentru o instanță nouă (de test) $x \in \mathbb{R}^d$:

$$w^{(i)} \cdot \phi(x) = \sum_{l=1}^i \alpha_l \phi(x_l) \cdot \phi(x) = \sum_{l=1}^i \alpha_l K(x_l, x).$$

Sumarizând, algoritmul pentru antrenarea Perceptronului kernel-izat [dual] se poate scrie în pseudo-cod astfel:

```

initialize  $\alpha_i = 0$  for  $i = 1, \dots, n$ ;
for  $i = 1, \dots, n$  do
    if  $y_i \sum_{l=1}^{i-1} \alpha_l K(x_l, x_i) \leq 0$  then
         $\alpha_i \leftarrow y_i$ ;
    end if
end for

```

Observație importantă: Se poate arăta relativ ușor că raționamentele (și rezultatele) din acest exercițiu se pot extinde și la cazul perceptronului care folosește o altă codificare a ieșirilor y_i (de exemplu, 0 și 1 în loc de -1 și +1) și, corespunzător, o altă funcție de activare h (recte funcția prag 0 – 1 în locul funcției *sign*), o rată de învățare oarecare $\eta > 0$ — dar face inițializarea vectorului de ponderi $w \in \mathbb{R}^d$ cu $\bar{0}$, ceea ce este echivalent cu $\alpha_i = 0$, pentru $i = 1, \dots, n$ — și care eventual parcurge de mai multe ori setul de date de antrenament. În acest caz, regula de actualizare a perceptronului kernel-izat [dual] este de forma

$$\alpha_i \leftarrow \alpha_i + \eta(y_i - h(w^{(i-1)} \cdot \phi(x_i))) \text{ dacă } y_i \neq h(w^{(i-1)} \cdot \phi(x_i)),$$

unde, exact ca mai sus,

$$w^{(i-1)} \cdot \phi(x_i) = \left(\sum_{l=1}^{i-1} \alpha_l \phi(x_l) \right) \cdot \phi(x_i) = \sum_{l=1}^{i-1} \alpha_l K(x_l, x_i).$$

6.1.3 Rețele “feed-forward” — algoritmul de retropropagare

20.

(Rețele feed-forward, algoritmul de retro-propagare:
o generalizare simplă în raport cu
Machine Learning de Tom Mitchell, pag. 101-103)

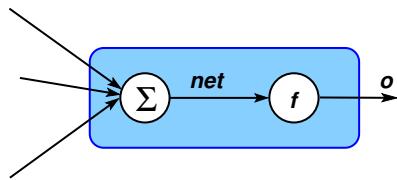
prelucrare de Liviu Ciortuz, după
■ • CMU, 2008(?) spring, HW2, pr. 2.2-4

La acest exercițiu veți deriva regula de actualizare pentru algoritmul de retropropagare (varianta stochastică) pentru o rețea neuronală artificială de tip “feed-forward”, cu două niveluri de unități sigmoidale. Se consideră d unități de intrare, H unități ascunse și K unități de ieșire.

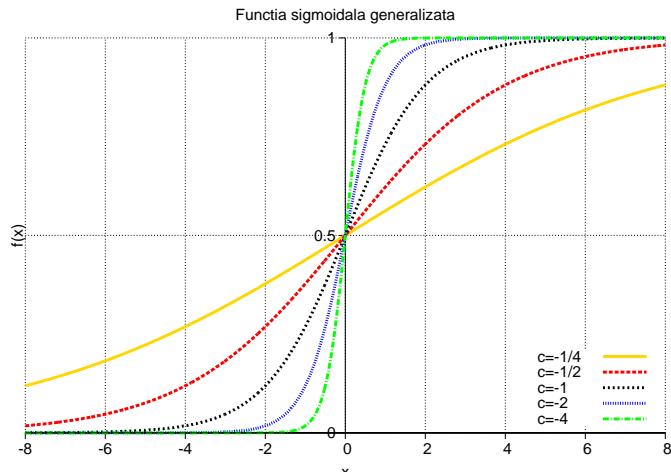
Funcția de eroare cu care se lucrează este

$$E(\bar{w}) = \frac{1}{2} \sum_k (t_k - o_k)^2,$$

unde $o_k = f(\text{net}_k) = \frac{1}{1 + e^c \text{net}_k}$, iar c este o constantă.



Observație: Este de remarcat faptul că se poate folosi constanta c pentru a „controla“ pantă cu care se face smoothing-ul (netezirea) funcției treaptă 0/1. Valori negative mici (adică, mari în modul) ale lui c implică o trecere foarte rapidă dinspre valori apropiate de 0 înspre valori apropiate de 1. Similar, valori negative mari (adică, apropiate de 0) pentru c implică o trecere lentă de la valori ≈ 0 la valori ≈ 1 .

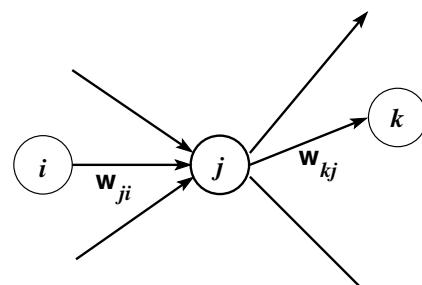


adaptare după: <http://en.wikipedia.org/wiki/File:Logistic-curve.svg>

a. Calculați derivata funcției sigmoidale generalizate $f(z) = \frac{1}{1 + e^{cz}}$ și exprimați rezultatul în funcție de însuși $f(z)$, adică fără a-l implica direct pe z .

b. Fie w_{kj} ponderea conexiunii de la unitatea ascunsă j către unitatea de ieșire k . Arătați că regula de actualizare pentru w_{kj} este de formă $w_{kj} \leftarrow w_{kj} + \eta \delta_k y_j$, unde y_j este ieșirea unității ascunse j , iar $\delta_k = f'(\text{net}_k)(t_k - o_k)$, cu $\text{net}_k = \sum_{j'} w_{kj'} y_{j'}$.

c. Arătați că regulile de actualizare pentru ponderile care corespund conexiunilor input-to-hidden sunt de formă $w_{ji} \leftarrow w_{ji} + \eta \tilde{\delta}_j x_i$, unde x_i este intrarea i , iar $\tilde{\delta}_j = f'(\text{net}_j)[\sum_{k=1}^K w_{kj} \delta_k]$, cu $\text{net}_j = \sum_{i'} w_{ji'} x_{i'}$.



Răspuns:

Demonstrația de mai jos urmează, în esență, aceeași linie de gândire ca și cea din cartea lui Tom Mitchell, *Machine Learning* (1997), pag. 101–103, unde s-a lucrat cu $c = -1$. Noi am preferat însă o redactare care să urmeze mai îndeaproape aplicarea formulelor de analiză matematică.

a. Mai întâi calculăm derivata lui f în raport cu argumentul z :

$$\frac{\partial f(z)}{\partial z} = \frac{\partial}{\partial z} \left(\frac{1}{1 + e^{cz}} \right) = -\frac{1}{(1 + e^{cz})^2} \cdot \frac{\partial}{\partial z} (1 + e^{cz}) = -\frac{1}{(1 + e^{cz})^2} \cdot c \cdot e^{cz}.$$

Apoi, în expresia pe care tocmai am obținut-o, forțăm apariția lui $f(z)$:

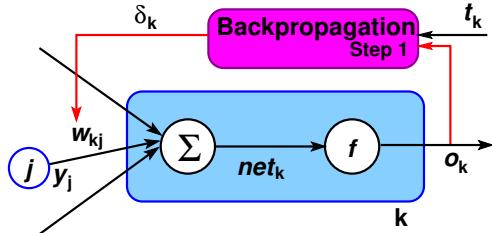
$$\begin{aligned} \frac{\partial f(z)}{\partial z} &= -\frac{1}{(1 + e^{cz})^2} \cdot c \cdot e^{cz} = -c \cdot \frac{1}{1 + e^{cz}} \cdot \frac{e^{cz}}{1 + e^{cz}} = -c \cdot \frac{1}{1 + e^{cz}} \cdot \frac{1 + e^{cz} - 1}{1 + e^{cz}} \\ &= -c \cdot \frac{1}{1 + e^{cz}} \cdot \left(1 - \frac{1}{1 + e^{cz}} \right) = -cf(z)(1 - f(z)). \end{aligned}$$

Observație importantă: Punctele b și c vor fi rezolvate de fapt în raport cu o funcție oarecare f derivabilă, lăsată nespecificată. Forma particulară dată în enunț pentru funcția f (folosită la punctul a) nu are nicio consecință particulară asupra raționamentului dezvoltat în continuare.

b. Conform metodei gradientului descendente, actualizarea ponderilor w_{kj} de pe conexiunile care intră într-o unitate de ieșire k se va face conform formulei

$$w_{kj} \leftarrow w_{kj} - \eta \frac{\partial E}{\partial w_{kj}},$$

unde η este o constantă pozitivă (rata de învățare).

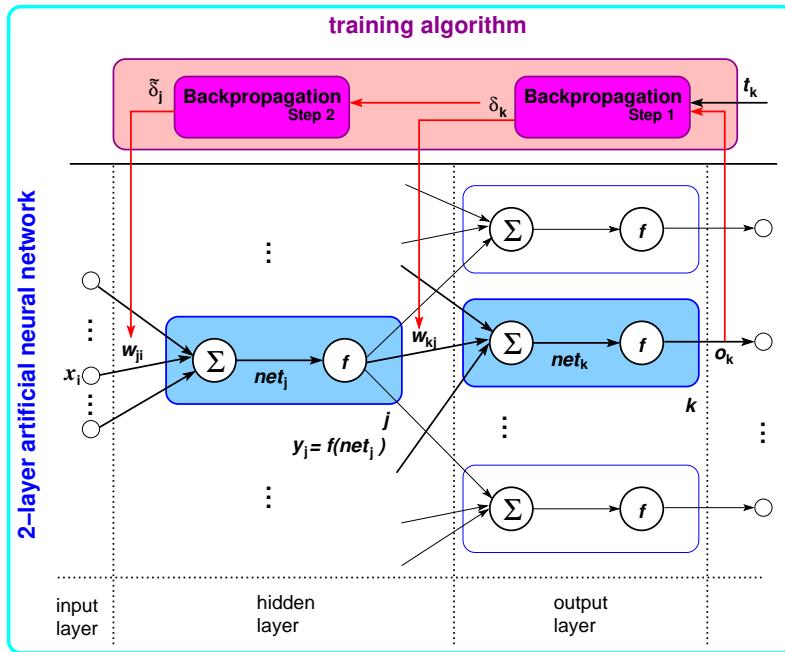


Intuitiv, ponderea w_{kj} influențează valoarea funcției de eroare E (doar) prin intermediul lui net_k , valoarea care este transmisă componentei de activare a unității k de pe nivelul de ieșire. (S-a notat cu net_k suma $\sum_{j'} w_{kj'} y_{j'}$.)

Așadar, este necesară aplicarea formulei pentru derivarea unei compunerii de funcții derivabile:

$$\begin{aligned} \frac{\partial E}{\partial w_{kj}} &= \frac{\partial E}{\partial net_k} \frac{\partial net_k}{\partial w_{kj}} = \left(\frac{\partial}{\partial net_k} \frac{1}{2} \sum_{k'=1}^K (t_{k'} - o_{k'})^2 \right) \left(\frac{\partial}{\partial w_{kj}} \sum_{j'=0}^{n_k} w_{kj'} y_{j'} \right) \\ &= \left(\frac{\partial}{\partial net_k} \frac{1}{2} (t_k - f(net_k))^2 \right) y_j \\ &= y_j \left(\frac{1}{2} 2(t_k - f(net_k)) \frac{\partial}{\partial net_k} (t_k - f(net_k)) \right) = -y_j(t_k - f(net_k))f'(net_k), \end{aligned}$$

unde n_k este numărul de intrări ale unității k . În particular, $n_k = H$ dacă se consideră toate conexiunile posibile (hidden-to-output).



Dacă notăm $\delta_k = (t_k - f(\text{net}_k))f'(\text{net}_k)$, atunci avem $\frac{\partial E}{\partial w_{kj}} = -\delta_k y_j$, iar regula de actualizare a ponderilor unității ascunse j devine:

$$w_{kj} \leftarrow w_{kj} - \eta \frac{\partial E}{\partial w_{kj}} = w_{kj} + \eta \delta_k y_j$$

c. Ca și la punctul precedent, actualizarea ponderilor w_{ji} de pe conexiunile care intră într-o unitate ascunsă j se va face după formula $w_{ji} \leftarrow w_{ji} - \eta \frac{\partial E}{\partial w_{ji}}$. Ponderea w_{ji} influențează valoarea funcției de eroare E (doar) prin intermediul lui net_j , valoarea care intră în componenta de activare a unității ascunse j .

Folosind din nou formula pentru derivarea unei compunerii de funcții derivabile și ținând cont că $\text{net}_j = \sum_{i'} w_{ji'} x_{i'}$, vom avea:

$$\begin{aligned} \frac{\partial E}{\partial w_{ji}} &= \frac{\partial E}{\partial \text{net}_j} \frac{\partial \text{net}_j}{\partial w_{ji}} = \left(\frac{\partial}{\partial \text{net}_j} \frac{1}{2} \sum_{k=1}^K (t_k - o_k)^2 \right) \left(\frac{\partial}{\partial w_{ji}} \sum_{i'=0}^d w_{ji'} x_{i'} \right) \\ &= \left(\frac{1}{2} \sum_{k=1}^K 2(t_k - o_k) \frac{\partial}{\partial \text{net}_j} (t_k - o_k) \right) x_i = -x_i \sum_{k=1}^K (t_k - o_k) \frac{\partial o_k}{\partial \text{net}_j} \end{aligned}$$

Pe de altă parte, valoarea net_j influențează valoarea o_k (doar) prin intermediul lui net_k . Așadar, urmează că

$$\begin{aligned} \frac{\partial o_k}{\partial \text{net}_j} &= \frac{\partial o_k}{\partial \text{net}_k} \frac{\partial \text{net}_k}{\partial \text{net}_j} = f'(\text{net}_k) \frac{\partial}{\partial \text{net}_j} \sum_{j'=0}^{n_k} w_{kj'} y_{j'} \\ &= f'(\text{net}_k) \frac{\partial}{\partial \text{net}_j} \sum_{j'=0}^{n_k} w_{kj'} f(\text{net}_{j'}) = f'(\text{net}_k) f'(\text{net}_j) w_{kj} \end{aligned}$$

Înlocuind acest rezultat în egalitatea precedentă, vom avea:

$$\begin{aligned}\frac{\partial E}{\partial w_{ji}} &= -x_i \sum_{k=1}^K (t_k - o_k) f'(net_k) f'(net_j) w_{kj} \\ &= -x_i f'(net_j) \sum_{k=1}^K (t_k - o_k) f'(net_k) w_{kj} = -x_i f'(net_j) \sum_{k=1}^K \delta_k w_{kj}\end{aligned}$$

Folosind notația $\tilde{\delta}_j = f'(net_j) \sum_{k=1}^K \delta_k w_{kj}$, egalitatea de mai sus devine $\frac{\partial E}{\partial w_{ji}} = -x_i \tilde{\delta}_j$, iar regula de actualizare a ponderii se transformă în:

$$w_{ji} \leftarrow w_{ji} + \eta \tilde{\delta}_j x_i$$

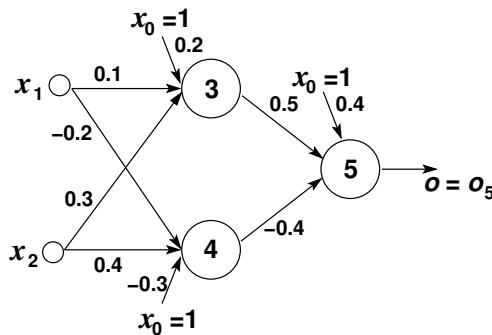
Observație: În mod similar cu demonstrația din carte Machine Learning de Tom Mitchell, pag. 101-103, demonstrația de aici poate fi extinsă în mod facil la rețele neuronale “feed-forward” cu un număr oarecare de niveluri ascunse. De asemenea, mai este posibilă încă o generalizare: funcția de activare f poate fi diferită de la o unitate la alta (sigmoidală, generalizat-sigmoidală (cu o anumită constantă c fixată), liniară etc). În demonstrație va fi suficient să atașăm simbolului f indicele unității neuronale respective.

21.

(Aplicarea algoritmului de retro-propagare a erorilor pe o rețea feed-forward formată din unități sigmoidale dispuse pe 2 niveluri)

prelucrare de Liviu Ciortuz,
după un exemplu din „Apprentissage artificiel“,
■ A. Cornuéjols, L. Miclet, 2010, pag. 332, 341

Rețeaua neuronală din figura următoare este formată din unități sigmoidale.



a. Care este rezultatul produs de această rețea pentru intrarea $x \stackrel{\text{not.}}{=} (x_1, x_2) = (1, 1)$?

Recomandare: Pentru ușurința urmăririi calculelor, vă cerem să completați un tabel de forma următoare:

i	$net_i = \sum_j w_{ij} x_j$	$o_i = \sigma(net_i)$
3		
4		
5		

b. La acest punct, veți executa manual prima iterare a algoritmului de retro-propagare pe rețea dată. Presupunem că în cazul intrării $x = (1, 1)$ ieșirea produsă de rețea ar trebui să fie $t = 0$. Luând rata de învățare $\eta = 1$, precizați care vor fi

- valorile ponderilor $w_{30}, w_{31}, w_{32}, w_{40}, w_{41}, w_{42}, w_{50}, w_{51}, w_{52}$, după aplicarea acestei prime iterării în antrenarea rețelei;⁷⁷⁶
- noul output produs de rețea (după actualizarea ponderilor) pe aceeași intrare $x = (1, 1)$.

c. Comparați rezultatele de la punctele a și b. Ce constatați?

Răspuns:

a. Completăm tabelul dat, ținând cont că ieșirile neuronilor de pe nivelul ascuns (unitățile sigmoidale 3 și 4) constituie intrările neuronului de pe nivelul exterior (unitatea sigmoidală 5).

i	$net_i = \sum_j w_{ij}x_j$	$o_i = \sigma(net_i)$
3	$0.2 + 0.1 + 0.3 = 0.6$	$1/(1 + e^{-0.6}) \simeq 0.646$
4	$-0.3 - 0.2 + 0.4 = -0.1$	$1/(1 + e^{0.1}) \simeq 0.475$
5	$0.4 + 0.5 \cdot 0.646 - 0.4 \cdot 0.475 = 0.533$	$1/(1 + e^{-0.533}) \simeq 0.630$

Așadar, outputul produs de rețea este 0.63.

b. Pentru a vedea detaliile algoritmului de retro-propagare a erorii, cititorul poate consulta cartea *Machine Learning* de Tom Mitchell, pag. 98 sau problema 20 din prezentul capitol, punctele b și c (unde parametrul c va fi considerat ca având valoarea 1).

Precizare: Ca să facilităm înțelegerea modului în care se aplică acest algoritm, în schema de mai jos am arătat cum anume vor fi calculate noile valori ale ponderilor w , precum și cantitățile δ implicate în execuția unei iterării a algoritmului. Ilustrarea se rezumă la parcurgerea (în ordinea output-to-input) unuia dintre drumurile din această rețea, și anume drumul 5, 4, 2. Extensia la celelalte căi este facilă. În această imagine (dar numai aici) am notat cu w' noile valori pentru ponderi (calculate în ultimul pas al iterării), pentru a nu fi confundate cu vechile valori, care intervin în calculul cantităților δ .

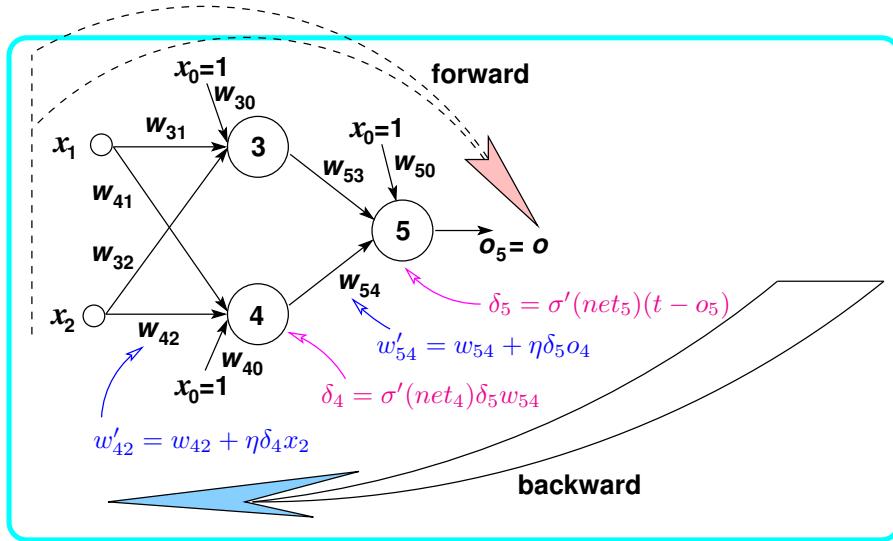
Conform algoritmului de retro-propagare,

$$\begin{aligned}\delta_5 &\stackrel{\text{def.}}{=} -\frac{\partial E}{\partial net_5} = \sigma'(net_5)(t - o) = \sigma(net_5)(1 - \sigma(net_5))(t - o) \\ &= o(1 - o)(t - o) = 0.63(1 - 0.63)(0 - 0.63) = -0.147\end{aligned}$$

Outputul neuronului 3 constituie unul din inputurile neuronului 5. În notația folosită de Tom Mitchell, vom scrie $Downstream(3) = \{5\}$. Așadar,

$$\delta_3 \stackrel{\text{def.}}{=} -\frac{\partial E}{\partial net_3} = o_3 \cdot (1 - o_3) \cdot \delta_5 \cdot w_{53} = 0.646 \cdot (1 - 0.646) \cdot (-0.147) \cdot 0.5 \simeq -0.017$$

⁷⁷⁶Semnificația pentru w_{ji} este cea din carte Machine Learning a lui Tom Mitchell, la pag. 98: w_{ji} este pondera de pe arcul / legătura de la unitatea (sau intrarea) i către unitatea j . A se vedea și reprezentarea grafică a rețelei din rezolvarea de mai jos.



În mod similar, $Downstream(4) = \{5\}$, și

$$\delta_4 \stackrel{\text{def.}}{=} -\frac{\partial E}{\partial net_4} = o_4 \cdot (1 - o_4) \cdot \delta_5 \cdot w_{54} = 0.475 \cdot (1 - 0.475) \cdot (-0.147) \cdot (-0.4) \simeq 0.015$$

Întrucât am terminat de calculat toate cantitățile de tip δ_j , vom trece acum la actualizarea ponderilor rețelei. Mai întâi vom calcula noile valori pentru ponderile de pe conexiunile hidden-to-output, deci vom avea:

$$w_{5j} \leftarrow w_{5j} + \Delta w_{5j} \text{ cu } \Delta w_{5j} = \eta \delta_5 o_j = \delta_5 o_j \text{ pentru } j \in \{0, 3, 4\}.$$

Așadar,

$$\begin{cases} \Delta w_{50} = -0.147 \cdot 1 = -0.147 \\ \Delta w_{53} = -0.147 \cdot 0.646 \simeq -0.095 \\ \Delta w_{54} = -0.147 \cdot 0.475 \simeq -0.070 \end{cases} \Rightarrow \begin{cases} w_{50} \leftarrow 0.4 - 0.147 \simeq 0.253 \\ w_{53} \leftarrow 0.5 - 0.095 = 0.405 \\ w_{54} \leftarrow -0.4 - 0.070 = -0.470 \end{cases}$$

Apoi vom actualiza ponderile care corespund conexiunilor de tip input-to-hidden. Știm că $\Delta w_{3i} = \eta \delta_3 x_i$ pentru $i \in \{0, 1, 2\}$, așadar inputul $x_0 = x_1 = x_2 = 1$ va implica $\Delta w_{30} = \Delta w_{31} = \Delta w_{32} = 1 \cdot (-0.017) \cdot 1 = -0.017$.

În consecință,

$$\begin{cases} w_{30} \leftarrow 0.2 - 0.017 = 0.183 \\ w_{31} \leftarrow 0.1 - 0.017 = 0.083 \\ w_{32} \leftarrow 0.3 - 0.017 = 0.283 \end{cases}$$

Similar, fiindcă $\Delta w_{4i} = \eta \delta_4 x_i$ pentru $i \in \{0, 1, 2\}$, rezultă că $\Delta w_{40} = \Delta w_{41} = \Delta w_{42} = 1 \cdot 0.015 \cdot 1 = 0.015$ și, prin urmare:

$$\begin{cases} w_{40} \leftarrow -0.3 + 0.015 = -0.285 \\ w_{41} \leftarrow -0.2 + 0.015 = -0.185 \\ w_{42} \leftarrow 0.4 + 0.015 = 0.415 \end{cases}$$

Acum putem calcula noua valoare produsă de rețeaua neuronală:

i	$net_i = \sum_j w_{ij}x_j$	$o_i = \sigma(net_i)$
3	$0.183 + 0.083 + 0.283 = 0.549$	$1/(1 + e^{-0.549}) \simeq 0.634$
4	$-0.285 - 0.185 + 0.415 = -0.055$	$1/(1 + e^{0.055}) \simeq 0.486$
5	$0.253 + 0.405 \cdot 0.634 - 0.47 \cdot 0.486 = 0.281$	$1/(1 + e^{-0.281}) \simeq 0.569$

Așadar, outputul produs de rețea după efectuarea primei iterări din cadrul algoritmului de retro-propagare a erorii este 0.569.

c. Valoarea produsă de rețea după ce a fost aplicată o iterare din algoritmul de retro-propagare (0.569) este mai aproape de valoarea-target (0) decât fusese outputul produs de rețea înainte de aplicarea acestei prime iterări (0.63). Este de așteptat ca rezultatul să se îmbunătățească la execuția următoarelor iterări ale algoritmului.

22.

(Regularizare: prevenirea overfitting-ului; cazul perceptronului liniar și al rețelelor feed-forward formate din astfel de perceptri)

■ • *Tom Mitchell, "Machine Learning", 1997, pr. 4.10*

În vederea reducerii riscului de overfitting, pentru o unitate liniară (engl., un-thresholded perceptron) se poate considera funcția de eroare

$$E(\bar{w}) = \frac{1}{2} \sum_{d \in D} (t_d - o_d)^2 + \gamma \sum_i w_i^2,$$

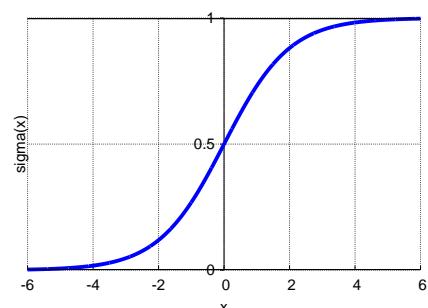
unde:

d indexează instanțele de antrenament din mulțimea $D \subset \mathbb{R}^n$,
 $\bar{w} = (w_0, w_1, \dots, w_n)$ sunt ponderile pentru intrările perceptronului,
 t_d este target-ul de învățat pentru instanță d ,
 o_d este outputul obținut de perceptron pentru aceeași instanță d ,
 γ este o constantă pozitivă.

Observație:

Suma $\sum_i w_i^2$ se poate scrie ca $\|w\|^2$, unde $\|w\|$ notează norma vectorului w . Intuitiv, minimizarea funcției obiectiv $E(\bar{w})$ va implica menținerea lui $\|w\|^2$ la o valoare destul de redusă.

Efectul practic, în cazul folosirii de unități sigmoidale este următorul: granița (suprafața) de decizie calculată de către rețea pentru ponderi w mici (în modul) este aproape liniară — a se vedea graficul funcției σ în jurul originii —, ceea ce o împiedică să se „muleze” în jurul neregularităților din datele de antrenament.



- a. Folosind metoda gradientului descendente, derivați regula de actualizare a ponderilor w_i pentru unitatea liniară care utilizează funcția de eroare dată mai sus.
- b. La acest punct vom considera că avem o rețea de tip feed-forward cu două niveluri de unități liniare și că forma funcției de eroare pentru întreaga rețea este similară cu cea indicată la punctul precedent. Pentru acest tip de rețea, vă cerem să deduceți regulile de actualizare a ponderilor. Pentru conveniență, veți considera varianta stochastică / incrementală, adică veți lucra cu o singură instanță de antrenament pe ciclu / „epocă“ de antrenare.

Observație: Deși orice rețea neuronală formată doar din unități liniare este echivalentă cu o unitate liniară, am propus punctul de mai sus pentru a servi drept model pentru rezolvarea problemei propuse 49.

- c. Arătați că minimizarea funcției $E(\bar{w})$ într-adevăr va conduce la găsirea unor valori mai apropiate de 0 pentru w_i decât în varianta clasică.

Răspuns:

Se poate arăta imediat că expresia

$$E(\bar{w}) = \frac{1}{2} \sum_{d \in D} (t_d - o_d)^2 + \gamma \sum_i w_i^2 = \frac{1}{2} \sum_{d \in D} (t_d - \sum_i w_i x_{i,d})^2 + \gamma \sum_i w_i^2$$

este o funcție convexă în raport cu toate argumentele w_i . Așadar, în vederea minimizării erorii, se justifică folosirea metodei gradientului descendente.

- a. Avem:

$$\begin{aligned} \frac{\partial E}{\partial w_i} &= \frac{\partial E}{\partial w_i} \left(\frac{1}{2} \sum_{d \in D} (t_d - o_d)^2 + \gamma \sum_{i'} w_{i'}^2 \right) = 2 \cdot \frac{1}{2} \sum_d (t_d - o_d) \frac{\partial}{\partial w_i} (t_d - o_d) + 2\gamma w_i \\ &= - \sum_d (t_d - o_d) \frac{\partial}{\partial w_i} \sum_{i'} w_{i'} x_{i',d} + 2\gamma w_i = - \sum_d (t_d - o_d) x_{i,d} + 2\gamma w_i \end{aligned}$$

Așadar,

$$w_i \leftarrow w_i - \eta \frac{\partial E}{\partial w_i} = w_i + \eta \left[\sum_d (t_d - o_d) x_{i,d} - 2\gamma w_i \right] = (1 - 2\eta\gamma) w_i + \eta \sum_d (t_d - o_d) x_{i,d}$$

- b. Înainte de a proceda la rezolvarea propriu-zisă a acestui punct al problemei, vom face următoarea *observație importantă*:

Cititorul atent va remarcă, desigur, analogia cu problema 20 (și problema 47). Această analogie este justificată, însă doar într-o anumită măsură. Va trebui să ținem cont că:

- i. Aici f corespunde funcției identitate, fiindcă lucrăm cu unitatea liniară. Așadar, $f'(net) = 1$ pentru orice valoare a expresiei net .

- ii. Nu vom mai putea lucra cu expresiile

$$\delta_k \stackrel{\text{def.}}{=} \frac{\partial E}{\partial net_k} \text{ și } \tilde{\delta}_j \stackrel{\text{def.}}{=} \frac{\partial E}{\partial net_j}$$

unde k este indicele unei unități neuronale de pe nivelul de ieșire iar j este indicele unei unități de pe nivelul ascuns. Explicația ține de faptul că expresia E nu se poate scrie ca o compunere de funcții $E(net_-(w_-))$, ci doar ca $E(net_-(w_-), w_-)$, deci pur și simplu $E(\bar{w})$.⁷⁷⁷ Așadar, nu vom putea scrie, ca în rezolvarea problemei 20 (și a problemei 47), că

⁷⁷⁷În această notație, caracterul ‘-’ ține loc de indice, lăsându-i valoarea nespecificată.

$\frac{\partial E}{\partial w_-} = \frac{\partial E}{\partial net_-} \frac{\partial net_-}{\partial w_-}$. În consecință, va trebui să calculăm direct derivatele partiale $\frac{\partial E}{\partial w_-}$. Similar cu demonstrația din cartea *Machine Learning* de Tom Mitchell de la pag. 101-103 și cu rezolvarea problemei 20 punctele b și c , vom calcula aceste derivate partiale mai întâi pentru variabilele care reprezintă ponderile de pe conexiunile hidden-to-output ale rețelei și apoi pentru variabilele care corespund ponderilor de pe conexiunile input-to-hidden.

Pentru conveniență, la acest punct al problemei vom urma convenția de scriere a variabilelor w folosind doi indici: w_{kj} va corespunde unei conexiuni hidden-to-output (unde j indică o unitate de pe nivelul ascuns, iar k o unitate de pe nivelul de ieșire), iar w_{ji} va corespunde unei conexiuni input-to-hidden (i – input, j – hidden). Notațiile t_k și o_k vor corespunde valorii-target și respectiv outputului unității liniare k de pe nivelul de ieșire al rețelei.

Avem:

$$E(\bar{w}) = \frac{1}{2} \sum_{k \in Outputs} (t_k - o_k)^2 + \gamma \|\bar{w}\|^2$$

și, în consecință:

$$\begin{aligned} \frac{\partial E}{\partial w_{kj}} &= \frac{\partial}{\partial w_{kj}} \left[\frac{1}{2}(t_k - o_k)^2 + \gamma w_{kj}^2 \right] + \sum_{k' \neq k} \frac{\partial}{\partial w_{kj}} \left[\frac{1}{2}(t_{k'} - o_{k'})^2 + \gamma w_{k'j}^2 \right] \\ &= -\frac{1}{2}2(t_k - o_k) \frac{\partial}{\partial w_{kj}} o_k + 2\gamma w_{kj} \\ &= -(t_k - o_k) \frac{\partial}{\partial w_{kj}} \sum_{j'} w_{kj'} y_{j'} + 2\gamma w_{kj} = -(t_k - o_k) y_j + 2\gamma w_{kj} \end{aligned}$$

Așadar,

$$w_{kj} \leftarrow w_{kj} - \eta \frac{\partial E}{\partial w_{kj}} = w_{kj} + \eta[(t_k - o_k)y_j - 2\gamma w_{kj}] = (1 - 2\eta\gamma)w_{kj} + \eta(t_k - o_k)y_j$$

Fie acum indicii j și i fixați ca mai sus, corespunzător unei conexiuni input-to-hidden. Notând că și Tom Mitchell $Downstream(j)$ mulțimea tuturor indicilor k cu proprietatea că există conexiune de la unitatea j către unitatea k , vom avea:

$$\begin{aligned} \frac{\partial E}{\partial w_{ji}} &= \frac{\partial}{\partial w_{ji}} \left\{ \left[\sum_{k \in Downstream(j)} \frac{1}{2}(t_k - o_k)^2 \right] + \gamma w_{ji}^2 \right\} \\ &= 2\gamma w_{ji} - \sum_{k \in Downstream(j)} (t_k - o_k) \frac{\partial}{\partial w_{ji}} o_k \\ &= 2\gamma w_{ji} - \sum_{k \in Downstream(j)} \left[(t_k - o_k) \frac{\partial}{\partial w_{ji}} \left(\sum_{j'} w_{kj'} y_{j'} \right) \right] \\ &= 2\gamma w_{ji} - \sum_{k \in Downstream(j)} \left\{ (t_k - o_k) \frac{\partial}{\partial w_{ji}} \left[\sum_{j'} w_{kj'} \left(\sum_{i'} w_{j'i'} x_{i'} \right) \right] \right\} \\ &= 2\gamma w_{ji} - \sum_{k \in Downstream(j)} \left[(t_k - o_k) \frac{\partial}{\partial w_{ji}} \left(\sum_{j'} \sum_{i'} w_{kj'} w_{j'i'} x_{i'} \right) \right] \end{aligned}$$

$$= 2\gamma w_{ji} - x_i \left(\sum_{k \in \text{Downstream}(j)} (t_k - o_k) w_{kj} \right)$$

Așadar,

$$\begin{aligned} w_{ji} \leftarrow w_{ji} - \eta \frac{\partial E}{\partial w_{ji}} &= w_{ji} + \eta \left[x_i \left(\sum_{k \in \text{Downstream}(j)} w_{kj} (t_k - o_k) \right) - 2\gamma w_{ji} \right] \\ &= (1 - 2\eta\gamma)w_{ji} + \eta x_i \left(\sum_{k \in \text{Downstream}(j)} w_{kj} (t_k - o_k) \right) \end{aligned}$$

c. Analizând regulile de actualizare a ponderilor deduse la punctul b, se observă că [în ambele cazuri, input-to-hidden și hidden-to-output] ele sunt de forma $w \leftarrow (1 - 2\eta\gamma)w + \Delta w$. Deosebirea față de cazul clasic ($w \leftarrow w + \Delta w$), este evidentă. Întrucât algoritmul de retro-propagare initializează ponderile w cu valori apropiate de 0 — iar în practică se folosesc valori mici pentru constantele η și γ —, lucrând cu varianta funcției de eroare propusă în această problemă vom obține valori ale ponderilor w mai aproape de 0 (decât în varianta clasică).

23. (De ce nu sunt convenabile funcțiile-prag pentru antrenarea perceptronilor și a rețelelor neuronale artificiale?)

CMU, 2011 spring, Tom Mitchell, HW5, pr. 3.4

Care sunt motivele pentru care în general funcțiile-prag nu convin pentru antrenare perceptronilor și a rețelelor neuronale artificiale [folosind metoda gradientului]?

Răspuns:

Răspunsul cel mai des citat, și anume că funcția-prag nu este derivabilă nu este în totalitate corect. Funcția modul ($|x|$) este folosită adeseori la aplicarea metodei gradientului, deși este derivabilă peste tot în afară de un singur punct (și anume, în 0), ca și funcția-prag. Un avantaj important al funcției modul este faptul că este continuă. Din contră, cel mai mare dezavantaj al funcției-treaptă este faptul că în orice punct în care este derivabilă, derivata sa este 0. Se știe că la aplicarea metodei gradientului descendant actualizarea parametrilor w se face după o regulă de forma $w \leftarrow w \pm \eta \nabla_w E(w)$. Prin urmare, în cazul în care derivatele partiale ale funcției $E(w)$ sunt constant nule, atunci regula gradientului va lăsa parametrul w neschimbăt. În consecință, funcția-prag nu poate servi la actualizarea valorilor parametrilor atunci când se folosește metoda gradientului.

24. (Rețele cu unități neuronale având funcții de activare liniare și / sau sigmoidale: corespondența cu suprafețe de decizie liniare / neliniare)

CMU, 2008 fall, Eric Xing, midterm, pr. 4.1

Fie o rețea neuronală având două niveluri, cu două unități pe nivelul ascuns și o singură unitate pe nivelul de ieșire (output). Vom lucra cu *funcția de*

activare liniară $y = C \cdot a \stackrel{\text{not.}}{=} C \cdot \sum_i w_i x_i$, unde C este o constantă, iar a este suma ponderată a intrărilor și, de asemenea, cu funcția sigmoidală (sau, logistică):

$$y = \sigma(a) \stackrel{\text{not.}}{=} \frac{1}{1 + e^{-a}}.$$

a. Presupunem că toate unitățile sunt liniare. Poate oare rețeaua aceasta să genereze / determine granițe de decizie (engl., decision boundaries) pe care un *model de regresie* standard de forma $y = b_0 + b_1 x_1 + b_2 x_2 + \varepsilon$ nu le poate genera?

b. Presupunem că unitățile ascunse ale rețelei noastre sunt de tip *sigmoidal*, iar unitatea de [pe nivelul de] ieșire este de tip *liniar*. Este oare posibil ca în acest caz rețeaua să genereze granițe de decizie neliniare?

c. Folosind funcția de activare *sigmoidală* pentru toate unitățile (atât cele de pe nivelul ascuns cât și cea de pe nivelul de ieșire), este oare posibil să aproximăm suprafețe de decizie oricăr de complicate prin combinarea multor [porțiuni de] granițe de decizie neliniare? Ce schimbări ar trebui să operați asupra rețelei de mai sus pentru a putea aproxima orice graniță de decizie?

Răspuns:

- a. Nu. Orice rețea formată doar din unități liniare poate fi redusă la un model liniar simplu.
- b. Da. Răspunsul se justifică imediat din punct de vedere matematic.
- c. Da; avem nevoie de unități ascunse adiționale. Atunci când folosim mai multe unități ascunse putem obține suprafețe de decizie mai complicate.

25.

(Adevărat sau Fals?)

- a. *CMU, 2003 fall, T. Mitchell, A. Moore, midterm, pr. 6.b.5*

Regula gradientului descendente aplicată în regim “batch” pentru o unitate neuronală care are intrările x_1 și x_2 și ieșirea $w_0 + w_1(x_1 + 1) + w_2(x_2^2)$ produce:

$$\begin{aligned}\Delta w_0 &= \eta \sum_d (t_i - o_i) \\ \Delta w_1 &= \eta \sum_i [(t_i - o_i)x_{i,1} + (t_i - o_i)] \\ \Delta w_2 &= \eta \sum_i [(t_i - o_i)2x_{i,2}]\end{aligned}$$

unde:

- t_i este valoarea dorită (engl., target) pentru outputul unității neuronale pentru exemplul i ;
- o_i este outputul unității neuronale pentru exemplul i ;
- $x_{i,1}$ este valoarea intrării / atributului x_1 din exemplul i ;
- $x_{i,2}$ este valoarea intrării / atributului x_2 din exemplul i .

b. *CMU, 2003 fall, T. Mitchell, A. Moore, midterm, pr. 6.b.3*

Varianta incrementală (sau, stochastică) a metodei gradientului descendente se comportă întotdeauna mai bine decât varianta “batch”.

c. *CMU, 2003 fall, T. Mitchell, A. Moore, midterm, pr. 6.b.2*

Suprafața de eroare urmată (conform metodei gradientului descendente) de către algoritmul de retro-propagare se modifică dacă modificăm datele de antrenament.

d. *CMU, 2003 fall, T. Mitchell, A. Moore, final exam, pr. 7.e*

Perceptronul poate fi (dar nu neapărat este) capabil să obțină o mai bună performanță de clasificare dacă (anterior antrenării) intrările sale sunt mapate într-un „spațiu de trăsături“ folosind o funcție-nucleu cu baza radială.

Răspuns:

a. Fals.

Valoarea funcției de ieșire pentru exemplul i este $o_i(w_0, w_1, w_2) = w_0 + w_1(x_{i,1} + 1) + w_2(x_{i,2}^2)$, iar funcția de eroare este $E(w_0, w_1, w_2) = \frac{1}{2} \sum_i (t_i - o_i)^2$. Stim că $w_j = w_j + \Delta w_j$, unde $\Delta w_j = -\eta \frac{\partial E}{\partial w_j}$. Este imediat că $\frac{\partial E}{\partial w_j} = \sum_i (t_i - o_i) \frac{\partial (t_i - o_i)}{\partial w_j}$. Calculând derivatele parțiale din această expresie, vom obține:

$$\frac{\partial (t_i - o_i)}{\partial w_0} = -1 \Rightarrow \frac{\partial E}{\partial w_0} = -\sum_i (t_i - o_i) \Rightarrow \Delta w_0 = \eta \sum_i (t_i - o_i);$$

$$\begin{aligned} \frac{\partial (t_i - o_i)}{\partial w_1} &= -(x_{i,1} + 1) \Rightarrow \frac{\partial E}{\partial w_1} = -\sum_i (t_i - o_i)(x_{i,1} + 1) \\ &\Rightarrow \Delta w_1 = \eta \sum_i (t_i - o_i)(x_{i,1} + 1); \end{aligned}$$

$$\frac{\partial (t_i - o_i)}{\partial w_2} = -x_{i,2}^2 \Rightarrow \frac{\partial E}{\partial w_2} = -\sum_i (t_i - o_i)x_{i,2}^2 \Rightarrow \Delta w_2 = \eta \sum_i (t_i - o_i)x_{i,2}^2.$$

Se observă că Δw_2 pe care tocmai l-am calculat mai sus diferă de cel dat în enunț, care este aşadar greșit.

b. Fals. Comparativ cu varianta “batch”, varianta incrementală a metodei gradientului descendente are *avantajul* de a evita unele optime (în spate, minime) locale. Însă varianta incrementală este doar o aproximare a metodei “batch” (care, pentru perceptronul liniar, obține în mod garantat optimul (global)). În plus, atunci când cele două variante folosesc rate ale învățării egale, varianta incrementală este mai lentă decât varianta “batch”.

c. Adevărat. Algoritmul de retro-propagare aplică metoda gradientului descendente funcției de eroare, care poate fi de exemplu $\frac{1}{2} \sum_d (t_i - o_i)^2$. În calculul funcției de eroare intervin atât target-ul (eticheta) t_i cât și atributele instanțelor de antrenament (acestea din urmă prin intermediul outputului o_i). În consecință, orice schimbare în datele de antrenament ca atare — cât și extinderea sau reducerea setului de date de antrenament — poate afecta funcția de eroare și, în consecință, rezultatul algoritmului de retro-propagare.

d. Adevărat. Chestiunea aceasta se poate pune pentru orice clasificator care învăță un separator liniar. Maparea mulțimii de instanțe antrenament $S = \{x_1, x_2, \dots, x_l\} \subset \mathbb{R}^n$ folosind o funcție-nucleu cu baza radială sau, mai general, o funcție oarecare $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}^m$ cu $m > n$ poate conduce la găsirea unui separator liniar pentru mulțimea $\{\Phi(x_1), \Phi(x_2), \dots, \Phi(x_l)\}$ în \mathbb{R}^m , chiar dacă mulțimea S nu este liniar separabilă. Totuși, găsirea unui astfel de separator (în urma aplicării funcției Φ) nu este garantată.

În general, chiar dacă nu se obține separabilitate liniară în urma mapării cu funcția Φ , este posibil (dar nu obligatoriu) ca în \mathbb{R}^m să se obțină un optim mai convenabil pentru funcția de optimizat. În particular, pentru perceptron, ca funcție de optimizat se ia în mod obișnuit semisuma pătratelor erorilor. „Un optim mai convenabil“ se va traduce în practică în faptul că funcția învățată în \mathbb{R}^m se comportă mai bine pe date de test (prin comparație, eventual, cu funcția învățată în \mathbb{R}^n).

Observație: Funcțiile cu baza radială sunt un caz particular de funcții-nucleu. (A se vedea capitolul *Mașini cu vectori-suport*.) Funcțiile-nucleu prezintă avantajul că produsele scalare $\Phi(x) \cdot \Phi(y)$ pot fi calculate în mod eficient, chiar dacă m este mult mai mare decât n . Această proprietate este foarte convenabilă pentru implementarea mașinilor cu vectori-suport.

26.

(Rețele neuronale profunde:

dispariția – ori, dimpotrivă, „explozia“ – vectorului gradient)

■ □ • ○ CMU, 2015 fall, E. Xing, Z. Bar-Joseph, HW3, pr. 1.3

În acest exercițiu veți studia anumite dificultăți care apar la aplicarea algoritmului de retro-propagare pentru antrenarea rețelelor neuronale profunde (engl., deep neural networks).⁷⁷⁸ Pentru conveniență, vom considera cea mai simplă rețea neuronală profundă, și anume una în care există câte un singur neuron pe fiecare nivel, iar outputul produs de neuronul de pe nivelul j este

$$z_j = f(\text{net}_j) \text{ cu } \text{net}_j \stackrel{\text{not.}}{=} \begin{cases} b_1 + w_{11}x_1 + \dots + w_{1d}x_d & \text{pentru } j = 1; \\ b_j + w_j z_{j-1} & \text{pentru } j = 2, \dots, m, \end{cases} \quad (343)$$

unde

- f este o anumită funcție de activare, a cărei derivată în punctul x este $f'(x)$,
- m este numărul de niveluri din această rețea neuronală,
- b_j este bias-ul (adică, termenul liber) al unității neuronale de pe nivelul j .

Vom nota cu L funcția de eroare / cost / pierdere (engl., loss function) care este folosită ca funcție obiectiv la antrenare.

- a. Calculați derivata parțială a funcției L în raport cu b_1 , bias-ul (adică, termenul liber) pentru neuronul de pe primul nivel.

Indicație: Pentru acest scop, nu este necesar să precizăm cine anume este funcția L și, la fel, cine este funcția f . Presupunând doar că L se exprimă în

⁷⁷⁸Se spune că o rețea neuronală este profundă dacă are cel puțin două niveluri ascunse.

funcție de f și este derivabilă, veți exprima $\frac{\partial L}{\partial b_1}$ în funcție de net_1, \dots, net_m și w_2, \dots, w_m . Veți folosi regula pentru derivarea compunerilor de funcții: $(f_1(f_2(x)))' = f'_1(f_2(x)) \cdot f'_2(x)$.

b. Presupunem că funcția de activare f este obișnuita funcție sigmoidală, $\sigma(x) = 1/(1 + \exp(-x))$ și că ponderile \bar{w} sunt inițializate astfel încât $|w_j| < 1$ pentru $j = 1, \dots, m$.

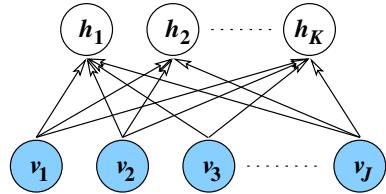
Explicați de ce derivata parțială de la punctul precedent, $\frac{\partial L}{\partial b_1}$, tinde la 0 atunci când m are valori mari. *Comentariu:* Din acest motiv, se spune că — în astfel de condiții — avem de a face cu „dispariția“ gradientului (engl., *gradient vanishing*).

Explicați de ce chiar și atunci când valoarea lui $|w|$ este mare, derivata parțială menționată mai sus tinde la 0 (deci, tinde să dispară), nu tinde la infinit (adică, să „explodeze“).

c. Una dintre modalitățile propuse pentru rezolvarea (parțială) a problemei „dispariției“ gradientului este ca în locul funcției sigmoidale să se folosească pentru activare *funcția liniar-rectificată* (engl., rectified linear function), notată cu ReL. Funcția de activare ReL este definită prin $\max\{0, x\}$. Explicați de ce funcția ReL ne poate ajuta să evităm fenomenul de „dispariție“ a gradientului.

d. O altă modalitate de rezolvare (parțială) a problemei „dispariției“ ori „exploziei“ gradientului este *pre-antrenarea nivel-cu-nivel* (engl., layer-wise pre-training).

Modelul *mașinii Boltzmann restricționate* (engl., Restricted Boltzmann machine), abreviat RBM, este unul dintre modelele cele mai des folosite pentru pre-antrenarea [unei rețele neuronale] nivel-cu-nivel. Figura alăturată prezintă un exemplu de RBM conținând K unități ascunse (h_1, \dots, h_K) și J intrări (v_1, \dots, v_J).



În vederea antrenării unei RBM, definim [ca funcție obiectiv] *distribuția probabilistă comună* — peste vectorii de valori posibile $\bar{v} \stackrel{\text{not.}}{=} (v_1, \dots, v_J)$ și $\bar{h} \stackrel{\text{not.}}{=} (h_1, \dots, h_K)$ — având forma [generală] următoare:

$$P(\bar{v}, \bar{h}) = \frac{1}{Z} \exp \left(\sum_i \theta_i \phi_i(\bar{v}, \bar{h}) \right),$$

unde

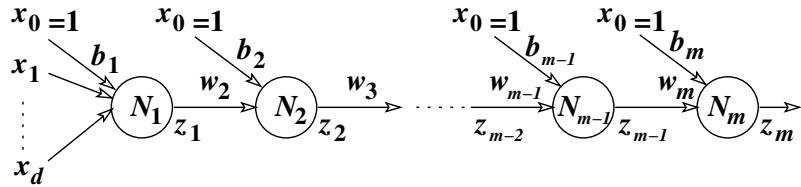
- $Z = \sum_{\bar{v}', \bar{h}'} \exp \left(\sum_i \theta_i \phi_i(\bar{v}', \bar{h}') \right)$ este constanta de normalizare;
- $\phi_i(\bar{v}, \bar{h})$ sunt anumite trăsături / attribute;
- θ_i sunt parametrii care corespund ponderilor din RBM.

Considerând că pentru antrenarea unei RBM se folosește metoda gradientului descent, arătați că expresia derivatei patiale a logaritmului funcției de probabilitate marginală $P(\bar{v})$ în raport cu parametrul / ponderea θ_i este:

$$\frac{\partial \ln P(\bar{v})}{\partial \theta_i} = \sum_{\bar{h}} \phi_i(\bar{v}, \bar{h}) P(\bar{h}|\bar{v}) - \sum_{\bar{v}', \bar{h}'} \phi_i(\bar{v}', \bar{h}') P(\bar{v}', \bar{h}').$$

Răspuns:

a. Ilustrăm mai jos rețeaua neuronală profundă definită în enunț.



Apoi vom scrie în mod desfășurat relațiile (343):

$$\begin{aligned} z_1 &= f(\text{net}_1) = f(b_1 + w_{11}x_1 + \dots + w_{1d}x_d) \\ z_2 &= f(\text{net}_2) = f(b_2 + w_2z_1) \\ &\dots \\ o = z_m &= f(\text{net}_m) = f(b_m + w_mz_{m-1}) \end{aligned}$$

Funcția de pierdere L se va exprima astfel:

$$\begin{aligned} L(w_{11}, \dots, w_{1d}, w_2, \dots, w_m, b_1, \dots, b_m) \\ \stackrel{\text{not.}}{=} L(f(\underbrace{\text{net}_m}_{b_m + w_m z_{m-1}})) \\ = L(f(b_m + w_m f(\underbrace{\text{net}_{m-1}}_{b_{m-1} + w_{m-1} z_{m-2}}))) \\ \dots \\ = L(f(b_m + w_m f(b_{m-1} + w_{m-1} f(b_{m-2} + \dots + w_3 f(\underbrace{\text{net}_2}_{b_2 + w_2 z_1} \dots)))) \\ = L(f(b_m + w_m f(b_{m-1} + w_{m-1} f(b_{m-2} + \dots + w_3 f(b_2 + w_2 z_1 \dots)))) \\ = L(f(b_m + w_m f(b_{m-1} + w_{m-1} f(b_{m-2} + \dots + w_3 f(b_2 + w_2 f(\underbrace{\text{net}_1}_{b_1 + w_{11} x_1 + \dots + w_{1d} x_d} \dots)))) \end{aligned}$$

Pentru a deriva funcția L în raport cu b_1 , ne vom aminti mai întâi regula de derivare a unei compuneri multiple de funcții $f_1(f_2(\dots f_n(x) \dots))$. Pe lângă forma clasică pe care o știm din liceu, $f'_1(f_2(\dots f_n(x) \dots)) \cdot f'_2(\dots f_n(x) \dots) \cdot f'_n(x)$, ea se poate scrie sub forma aşa-numitei *reguli de înlățuire*: $\frac{\partial f_1}{\partial f_2} \cdot \frac{\partial f_2}{\partial f_3} \cdots \frac{\partial f_2}{\partial f_1} \cdot \frac{\partial f_1}{\partial x}$.

Adoptând / aplicând această regulă pentru a calcula derivata parțială cerută în enunț, vom avea:

$$\frac{\partial L}{\partial b_1} = \frac{\partial L}{\partial \text{net}_m} \cdot \frac{\partial \text{net}_m}{\partial \text{net}_{m-1}} \cdots \frac{\partial \text{net}_2}{\partial \text{net}_1} \cdot \frac{\partial \text{net}_1}{\partial b_1}$$

$$\begin{aligned}
&= L'(f(\text{net}_m)) \cdot f'(\text{net}_m) \cdot \\
&\quad w_m \cdot f'(\text{net}_{m-1}) \cdot \\
&\quad w_{m-1} \cdot f'(\text{net}_{m-2}) \cdot \\
&\quad \dots \\
&\quad w_2 \cdot f'(\text{net}_1) \cdot \\
&\quad 1 \\
&= L'(f(\text{net}_m)) \cdot f'(\text{net}_1) \cdot \prod_{k=2}^m (f'(\text{net}_k) \cdot w_k)
\end{aligned}$$

b. Știm că $\sigma(x) \in (0, 1)$ și $\sigma'(x) = \sigma(x)(1 - \sigma(x))$ pentru orice $x \in \mathbb{R}$, iar maximul funcției $z(1 - z)$ este $1/4$ și se atinge în punctul $z = 1/2$. Prin urmare, $|\prod_{k=2}^m \sigma'(\text{net}_k)| \leq \left(\frac{1}{4}\right)^{m-1}$. Întrucât $|w_k| < 1$ pentru orice k (conform ipotezei din enunț), rezultă imediat că $\left|\frac{\partial L}{\partial b_1}\right| = |L'(\sigma(\text{net}_m)) \cdot \sigma'(\text{net}_1)| \cdot \prod_{k=2}^m |\sigma'(\text{net}_k) \cdot w_k|$ tinde la 0 pentru $k \rightarrow +\infty$ și pentru orice input \bar{x} fixat.

Pentru a contracara acest fenomen de „dispariție“ a gradientului, ar trebui să impunem condiții de forma $|w_k \sigma'(\text{net}_k)| > 1$. Trebuie însă să remarcăm faptul că $\sigma(\text{net}_k)$ depinde de w_k : $\sigma(\text{net}_k) = \sigma(b_k + w_k z_{k-1})$. În consecință, dacă îi vom da lui w_k posibilitatea să ia valori mari în modul, atunci $z_k \stackrel{\text{not.}}{=} b_k + w_k z_{k-1}$ va lua de asemenea valori mari în modul, iar $\sigma(z_k)$ va tinde fie la 0 fie la 1, deci $\sigma'(z_k) = \sigma(z_k)(1 - \sigma(z_k))$ va tinde la 0. În sfârșit, se poate verifica în mod riguros că $\lim_{z \rightarrow \pm\infty} z\sigma'(z) = 0$ (lăsăm această demonstrație ca exercițiu cititorului).

c. Dacă f este funcția de activare ReL, atunci $f'(x) = 0$ pentru $x < 0$ și $f'(x) = 1$ pentru $x > 0$. Prin urmare, folosind această funcție putem împiedica „dispariția“ gradientului.

d. Pornind de la expresia distribuției probabiliste $P(\bar{v}, \bar{h})$ care a fost dată în enunț, explicitând mai întâi constanta de normalizare Z , vom putea scrie apoi expresia distribuției marginale $P(\bar{v})$:

$$\begin{aligned}
P(\bar{v}, \bar{h}) &= \frac{1}{\sum_{\bar{v}', \bar{h}'} \exp(\sum_k \theta_k \phi_k(\bar{v}', \bar{h}'))} \exp\left(\sum_k \theta_k \phi_k(\bar{v}, \bar{h})\right) \\
\Rightarrow P(\bar{v}) &= \frac{1}{\sum_{\bar{v}', \bar{h}'} \exp(\sum_k \theta_k \phi_k(\bar{v}', \bar{h}'))} \sum_{\bar{h}} \exp\left(\sum_k \theta_k \phi_k(\bar{v}, \bar{h})\right)
\end{aligned}$$

Prin urmare,

$$\ln P(\bar{v}) = \ln \sum_{\bar{h}} \exp\left(\sum_k \theta_k \phi_k(\bar{v}, \bar{h})\right) - \ln \sum_{\bar{v}', \bar{h}'} \exp\left(\sum_k \theta_k \phi_k(\bar{v}', \bar{h}')\right).$$

În consecință, derivata parțială a funcției $\ln P(\bar{v})$ în raport cu parametrul θ_i se poate calcula astfel:

$$\begin{aligned}
\frac{\partial \ln P(\bar{v})}{\partial \theta_i} &= \\
&= \frac{\sum_{\bar{h}} \exp\left(\sum_k \theta_k \phi_k(\bar{v}, \bar{h})\right) \cdot \phi_i(\bar{v}, \bar{h})}{\sum_{\bar{h}} \exp\left(\sum_k \theta_k \phi_k(\bar{v}, \bar{h})\right)} - \underbrace{\frac{\sum_{\bar{v}', \bar{h}'} \exp\left(\sum_k \theta_k \phi_k(\bar{v}', \bar{h}')\right) \cdot \phi_i(\bar{v}, \bar{h})}{\sum_{\bar{v}', \bar{h}'} \exp\left(\sum_k \theta_k \phi_k(\bar{v}', \bar{h}')\right)}}_Z
\end{aligned}$$

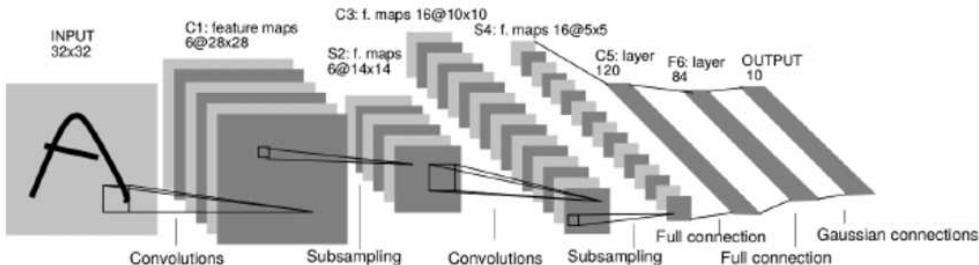
$$\begin{aligned}
&= \frac{\sum_{\bar{h}} \frac{1}{Z} \exp \left(\sum_k \theta_k \phi_k(\bar{v}, \bar{h}) \right) \cdot \phi_i(\bar{v}, \bar{h})}{\sum_{\bar{h}} \frac{1}{Z} \exp \left(\sum_k \theta_k \phi_k(\bar{v}, \bar{h}) \right)} - \sum_{\bar{v}', \bar{h}'} \frac{1}{Z} \exp \left(\sum_k \theta_k \phi_k(\bar{v}', \bar{h}') \right) \cdot \phi_i(\bar{v}, \bar{h}) \\
&= \frac{\sum_{\bar{h}} P(\bar{v}, \bar{h}) \cdot \phi_i(\bar{v}, \bar{h})}{P(\bar{v})} - \sum_{\bar{v}', \bar{h}'} P(\bar{v}', \bar{h}') \phi_i(\bar{v}', \bar{h}') \\
&= \sum_{\bar{h}} \frac{P(\bar{v}, \bar{h})}{P(\bar{v})} \cdot \phi_i(\bar{v}, \bar{h}) - \sum_{\bar{v}', \bar{h}'} P(\bar{v}', \bar{h}') \phi_i(\bar{v}', \bar{h}') \\
&= \sum_{\bar{h}} P(\bar{h}|\bar{v}) \cdot \phi_i(\bar{v}, \bar{h}) - \sum_{\bar{v}', \bar{h}'} P(\bar{v}', \bar{h}') \phi_i(\bar{v}', \bar{h}')
\end{aligned}$$

27.

(Determinarea numărului de parametri și de conexiuni din rețea neuronală convolutivă LeNet)

*prelucrare de Liviu Ciortuz, după**□ • ○ CMU, 2015 fall, E. Xing, Z. Bar-Joseph, HW3, pr. 1.2.1*

Ca să puteți rezolva acest exercițiu, trebuie să fiți deja familiari cu modelul de rețea neuronală convolutivă LeNet, a cărei reprezentare grafică o reproducem mai jos.⁷⁷⁹



Vă cerem să determinați numărul total de parametri și numărul total de conexiuni din această rețea. Câtă parametri și câte conexiuni sunt în fiecare dintre nivelurile convoluționale și de sub-selectie, C1, S2, C3 și S4? Câtă parametri sunt în fiecare dintre nivelurile total-conectate (engl., fully-connected layers), C5, F6 și OUTPUT?

Precizări:

- Mărimea *filtrului* pentru fiecare dintre nivelurile convoluționale și nivelurile de sub-selectie (engl., sub-sampling layers):
 - **C1:** 5×5 , adică, fiecare unitate neuronală din C1 are un *spațiu de recepție* (engl., receptive field) de 5×5 în nivelul precedent lui;
 - **S2:** 2×2 ;
 - **C3:** 5×5 ;
 - **S4:** 2×2 .

⁷⁷⁹Vă recomandăm să citiți secțiunile A și B din articolul *Gradient-based learning applied to document recognition*, de Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Proceedings of the IEEE, November 1998, 86 (11), p. 2278-2324. De asemenea, puteți vedea următorul exercițiu de tip implementare: CMU, 2016 spring, W. Cohen, N. Balcan, HW7 (sau CMU, 2016 fall, N. Balcan, M. Gormley, HW6).

ii. Spre deosebire de nivelul C1, care este total conectat cu nivelul precedent (INPUT), nivelul C3 nu este total conectat cu nivelul S2. Tabelul de mai jos arată care dintre hărțile de trăsături (sau, *planele*) de pe nivelul S2 sunt conectate cu (toate!) planele de pe nivelul C3.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	X				X	X	X			X	X	X	X		X	X
1	X	X				X	X	X			X	X	X	X		X
2	X	X	X				X	X	X			X		X	X	X
3		X	X	X			X	X	X	X			X		X	X
4			X	X	X			X	X	X	X		X	X		X
5				X	X	X			X	X	X	X		X	X	X

Răspuns:

Vom prezenta o sinteză a calculelor, precum și răspunsurile finale, în tabelul următor:

	nr. parametri	nr. conexiuni
ConvLayers:		
C1	$(5 * 5 + 1) * 6 = 156$	$28 * 28 * 26 * 6 = 122304$
S2	$2 * 6 = 12$	$14 * 14 * 6 * 5 = 5880$
C3	$(5 * 5 * 3 + 1) * 6 +$ $(5 * 5 * 4 + 1) * 9 +$ $(5 * 5 * 6 + 1) = 1516$	$1516 * 10 * 10 = 151600$ $5 * 5 * 16 * 16 = 2000$
S4	$2 * 16 = 32$	
FCLayers:		
C5	$(5 * 5 * 16 + 1) * 120 = 48120$	48120
F6	$(120 + 1) * 84 = 10164$	10164
OUTPUT	$(84 + 1) * 10 = 850$	850
Total	60850	340918

6.2 Rețele neuronale artificiale — Probleme propuse

6.2.1 Chestiuni introductive

28. (Perceptronul-prag: funcția calculată; zone de decizie)

- CMU, 1997 fall, T. Mitchell, A. Moore, midterm exam, pr. 3.1

Considerăm un perceptron-prag cu intrările x_1 și x_2 și având ponderile $w_0 = 0$, $w_1 = 1$ și $w_2 = -3$. Desenează granița de decizie a acestui perceptron și indicați zonele din planul bidimensional în care perceptronul produce rezultatele $+1$ și -1 .

29. (Perceptronul-prag: exemplificare pentru disjuncții / conjuncții generalizate)

- prelucrare de Liviu Ciortuz, după CMU, 2010 fall, Ziv Bar-Joseph, HW3, pr. 1.1

Imaginează-ți că ești responsabil cu design-ul unor preceptroni de tip prag și că trebuie să implementezi câteva funcții logice.

a. Concep un perceptron-prag care implementează operatorul logic AND. Se vor considera două intrări, x_1 și $x_2 \in \{-1, +1\}$, și ieșirea $y \in \{-1, +1\}$.

Scrie mai întâi *tabela de adevăr* a operatorului AND.

În planul bidimensional al intrărilor x_1 și x_2 , desenează *separatorul* (adică granița de decizie) a perceptronului. Determină *ecuația* $f(x_1, x_2) = 0$ corespunzătoare acestei separatori. Verifică semenele funcției f pe *zonele de decizie determinate* de separator.

Desenează perceptronul-prag AND specificând ponderile w_1 , w_2 și w_0 .

Extinde rezultatul precedent la

- conjuncții de n variabile (A_1 AND A_2 AND ... AND A_n);
- conjuncții de n literalii (l_1 AND l_2 AND ... AND l_n) unde l_i este fie A_i fie $\neg A_i$.

b. Procedează similar pentru funcția logică OR.

30. (Perceptroni-prag și rețele de astfel de perceptri: exemplificare)

- TU Dresden, 2006 summer, S. Hölldobler, A. Grossmann, HW3, pr. 2

Fie setul de instanțe de antrenament din \mathbb{R}^2 prezentate în tabelul de mai jos (partea dreaptă).

- a. Indicați fie un perceptron cu funcție de activare de tip prag fie o rețea formată din astfel de perceptroni — alegeți varianta cea mai simplă! — care să fie consistent(ă) cu aceste exemple. (Nu trebuie să apelați la niciun algoritm de învățare automată!)

<i>Exemplu</i>	<i>X</i>	<i>Y</i>	<i>Clasa</i>
1	1	6	—
2	1	2	+
3	2	4	+
4	5	2	—
5	6	5	—
6	7	8	—

- b. Îndepliniți aceleași cerințe ca mai sus după ce în prealabil ați adăugat la tabelul dat și exemplul din tabelul alăturat.

<i>Exemplu</i>	<i>X</i>	<i>Y</i>	<i>Clasa</i>
7	7	1	+

31.

(O proprietate:
extensia funcției XOR pe \mathbb{R}^2 (vedeți pr. 1.d)
nu poate fi calculată de rețele de perceptroni-prag
care au un singur nivel ascuns)

■ □ • ○ *prelucrare de Liviu Ciortuz, după CMU, 2010 fall, Aarti Singh, HW5, pr. 4.1.2*

În rezolvarea problemei 1.d s-a arătat că funcția $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ definită prin $f(x_1, x_2) = 1$ pentru $x_1, x_2 \geq 0$ sau $x_1, x_2 < 0$, și -1 în rest este calculabilă de către o rețea neuronală cu *două niveluri ascunse*, folosind doar unități de tip prag.

Demonstrați că nu există nicio rețea cu *un singur nivel ascuns*, constituită doar din astfel de unități, care să calculeze funcția f . Veți presupune că numărul de unități de pe nivelul ascuns poate fi oricât de mare, însă finit.

Indicație: Întrucât rețeaua are un singur nivel ascuns, fiecare unitate de pe acest nivel poate fi pusă în corespondență cu o dreaptă din planul bidimensional, iar outputul calculat de către respectiva unitate este $+1$ pentru punctele (x_1, x_2) situate de o parte a dreptei și -1 pentru punctele de cealaltă parte. Considerăm o vecinătate [de formă circulară] a originii reperului de coordinate, astfel încât dacă dreapta (adică, separatorul, sau granița de separare) asociată unei unități ascunse (oricare dintre acestea!) intersectează această vecinătate atunci ea trece (în mod necesar) prin origine. Este oare posibil ca o rețea neuronală [alcătuită doar din unități-prag și] având un singur nivel ascuns să reprezinte funcția f în această vecinătate?

32.

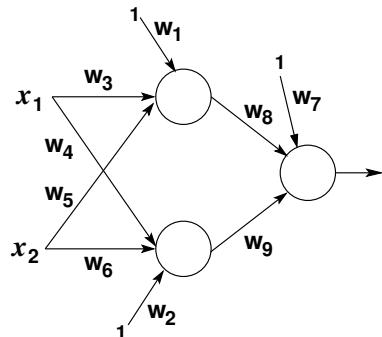
(Rețele neuronale; o proprietate de bază:
unitățile liniare de pe nivelurile ascunse pot fi „absorbite“
pe nivelul următor)

• ○ *CMU, ? spring, ML course 10-701, midterm, pr. 4*

Considerăm rețeaua neuronală pentru clasificare binară din figura de mai jos (partea dreaptă).

Pentru unitățile ascunse folosim o funcție liniară de activare $h(z) = cz$, iar pentru unitatea de pe nivelul de ieșire funcția de activare sigmoidală $\sigma(z) = \frac{1}{1 + e^{-z}}$.

a. Desenați o rețea neuronală care este echivalentă cu rețeaua dată, dar care nu are niciun nivel ascuns. Scrieți ponderile \tilde{w} ale acestei noi rețele în funcție de c și de w_i .



b. Este adevărat că orice rețea neuronală de tip feed-forward, cu mai multe niveluri și având toți neuronii situați pe nivelurile ascunse de tip unități liniare poate fi reprezentată ca o rețea neuronală fără niciun nivel ascuns? Explicați succint răspunsul dat.

33.

(Alegerea ponderilor pentru o rețea cu structură / compoziție specificată, în aşa fel încât să codifice funcția XOR)

• ○ * CMU, 2011 spring, Tom Mitchell, HW5, pr. 3.3

Fie o rețea neuronală, constând dintr-un nivel ascuns format din două unități sigmoidale, și un nivel de ieșire pe care se află o singură unitate, de tip prag. La niciuna dintre aceste trei unități nu se folosește termenul liber $x_0 = 1$. Găsiți valori pentru ponderile din rețea astfel încât ea să codifice conceptul logic XOR.

Indicație: Întrucât unitățile de pe nivelul ascuns sunt de tip sigmoidal, se va considera că intrările x_1 și x_2 iau valorile 0 și / sau 1.

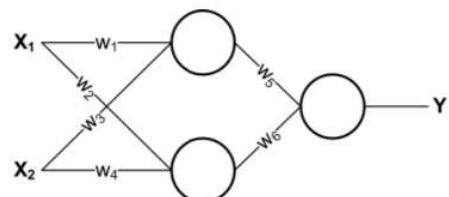
Observație: În raport cu problema 2.b, aici la toți neuronii s-a eliminat termenul liber $x_0 = 1$ (ca și la problema 4), iar tipul funcției de activare, și anume prag, a fost înlocuit cu cel sigmoidal. Din cele trei probleme se observă că astfel de modificări pot afecta capacitatea de reprezentare a unei rețele neuronale (chiar dacă topologia de ansamblu este aceeași în toate cele trei cazuri).

34.

(Exemplificarea legăturii dintre rețele neuronale, regresia liniară, regresia logistică și boosting)

□ • ○ CMU, 2015 fall, E. Xing, Z. Bar-Joseph, HW3, pr. 1.1

În figura alăturată este reprezentată o rețea neuronală cu două straturi, care învață o funcție $f : \mathbb{R}^2 \rightarrow \mathbb{R}$. Ponderile $w = (w_1, \dots, w_6)$ pot fi numere reale oarecare. Vom nota cu $X = (X_1, X_2) \in \mathbb{R}^2$ inputul rețelei, iar cu $Y \in \mathbb{R}$ outputul ei.



În ce privește funcția de activare implementată de fiecare dintre unitățile neuronale care compun această rețea, vom presupune că sunt posibile două variante:

- S : funcția compusă „semn-sigmoidală“ (engl., signed sigmoid),

$$S(a) \stackrel{\text{def.}}{=} \text{sign}(\sigma(a) - 0.5) = \text{sign}\left(\frac{1}{1 + \exp(-a)} - 0.5\right);$$

- L : funcția liniară $L(a) \stackrel{\text{def.}}{=} ca$, unde c este o constantă reală, fixată.

În ambele cazuri, $a \stackrel{\text{not.}}{=} \sum_i w_i X_i$.

a. Asignați fiecărei unități neuronale din figură câte o funcție de activare (S sau L) la în aşa fel încât această rețea să simuleze o regresie liniară: $Y = \beta_1 X_1 + \beta_2 X_2$.

b. Asignați fiecărei unități neuronale din figură câte o funcție de activare (S sau L) la în aşa fel încât această rețea să simuleze o regresie logistică binară: $Y = \arg \max_y P(Y = y|X)$, unde $P(Y = 1|X) = \frac{\exp(\beta_1 X_1 + \beta_2 X_2)}{1 + \exp(\beta_1 X_1 + \beta_2 X_2)}$, iar $P(Y = -1|X) = \frac{1}{1 + \exp(\beta_1 X_1 + \beta_2 X_2)}$. Exprimăți β_1 și β_2 în funcție de w_1, \dots, w_6 .

c. Asignați fiecărei unități neuronale din figură câte o funcție de activare (S sau L) la în aşa fel încât această rețea să simuleze un clasificator de tip boosting care combină doi clasificatori de tip regresie logistică $f_1 : X \rightarrow Y_1$ și $f_2 : X \rightarrow Y_2$, în aşa fel încât să se obțină outputul final $Y = \text{sign}(\alpha_1 Y_1 + \alpha_2 Y_2)$. Folosiți pentru f_1 și f_2 aceleași „distribuții“ ca la punctul b. Exprimăți α_1 și α_2 în funcție de w_1, \dots, w_6 .

35.

(Expresivitate: rețele neuronale vs. arbori de decizie)

• CMU, 2009 spring, Ziv Bar-Joseph, HW2, pr. 3

La acest exercițiu veți compara puterea de reprezentare a arborilor de decizie pe de o parte și a rețelelor neuronale pe de altă parte, privitor la capacitatea de a codifica două funcții, și anume *funcția majoritate* (engl., majority function) și *funcția paritate* (engl., parity function). Ambele funcții primesc ca input n -uple $x_1, \dots, x_n \in \{0, 1\}^n$, iar outputul lor este în multimea $\{0, 1\}$. Funcția *majoritate* returnează valoarea 1 dacă mai mult de jumătate dintre componente din intrare sunt 1. De exemplu, tuplul $(1, 1, 1, 1, 0)$ conține mai multe cifre de 1 decât de 0, deci va produce outputul 1. Funcția *paritate* returnează 1 dacă și numai dacă un număr par de componente din intrare sunt 1. De exemplu, tuplul $(1, 0, 1, 0, 1, 0)$ conține 3 cifre de 1, deci produce outputul 0.

a. Poate oare perceptronul[-prag] să fie folosit ca să codifice

- funcția majoritate?
- funcția paritate?

Pentru fiecare dintre aceste două funcții, în cazul în care ați răspuns afirmativ, prezentați perceptronul corespunzător, iar dacă ați răspuns negativ, justificați.

b. Poate oare un arbore de decizie să codifice

- funcția majoritate?
- funcția paritate?

Pentru fiecare dintre aceste două funcții, în cazul în care ați răspuns afirmativ, prezentați arborele de decizie corespunzător, iar dacă ați răspuns negativ, justificați.

c. Pentru fiecare dintre cele două funcții de la punctul a la care ați răspuns negativ, este oare posibil să obținem codificarea dorită folosind o rețea neuronală cu două niveluri (un nivel ascuns și un nivel de ieșire)? Dacă este așa, desenați rețeaua care rezolvă problema. Altminteri, precizați câte niveluri ascunse sunt necesare pentru a putea rezolva problema.

6.2.2 Unități neuronale — algoritmi de antrenare

36.

(Antrenarea pereceptronului liniar:
deducerea regulii de actualizare)

*Liviu Ciortuz, 2017, urmând
■ Tom Mitchell, Machine Learning, 1997, p. 89-93*

În acest exercițiu vi se cere mai întâi să elaborați / redați partea de fundamente teoretică pentru antrenarea pereceptronului liniar, date fiind instanțele $(\bar{x}_1, t_1), \dots, (\bar{x}_n, t_n)$, cu $\bar{x}_i \in \mathbb{R}^d$ și $t_i \in \mathbb{R}$ pentru $i = 1, \dots, n$.

Așadar, veți considera pereceptronul liniar având intrările $x_0 = 1, x_1, \dots, x_d \in \mathbb{R}$ și ponderile $w_0, w_1, \dots, w_d \in \mathbb{R}$.

a. În linie cu *metoda gradientului descendente*, deduceți care este forma *regulii de actualizare* a vectorului de ponderi $\bar{w} \in \mathbb{R}^{d+1}$ (sau, echivalent, forma regulilor de actualizare a ponderilor w_i , cu $i = 0, \dots, d$) pentru găsirea acelei valori care minimizează semisuma pătratelor erorilor comise de pereptron,⁷⁸⁰ adică

$$E(\bar{w}) \stackrel{\text{def.}}{=} \frac{1}{2} \sum_{i=1}^n (t_i - o_i)^2.$$

În această ultimă expresie, o_i este outputul pereceptronului liniar pentru intrarea $\bar{x}_i \stackrel{\text{renot.}}{=} (x_0 = 1, x_{i,1}, \dots, x_{i,d})$, adică $o_i = w_0 + \sum_{j=1}^d w_j x_{i,j}$.

Vă reamintim că la aplicarea metodei gradientului descendente pentru găsirea unui punct de minim al unei funcții derivabile $f : \mathbb{R}^{d+1} \rightarrow \mathbb{R}$, regula de actualizare a parametrilor \bar{w} ai funcției f are forma $\bar{w} \leftarrow \bar{w} + \Delta \bar{w}$, cu $\Delta \bar{w} \stackrel{\text{def.}}{=} -\eta \nabla_{\bar{w}} f(\bar{w})$, unde $\eta > 0$ este un număr real mic, numit *rata de învățare*, iar $\nabla_{\bar{w}} f(\bar{w})$ este vectorul *gradient*, adică $\left(\frac{\partial}{\partial w_0} f(\bar{w}), \frac{\partial}{\partial w_1} f(\bar{w}), \dots, \frac{\partial}{\partial w_d} f(\bar{w}) \right)$.

⁷⁸⁰Conform problemei 13.b, am putea scrie — în ipoteza că datele au fost generate probabilist, cu o componentă „zgomot” urmând o distribuție gaussiană unidimensională de medie 0, aşa cum s-a precizat acolo —,

$$\bar{w}_{\text{ML}} = \arg \min_{\bar{w}} E(\bar{w}) = \arg \min_{\bar{w}} \frac{1}{2} \sum_{i=1}^n (t_i - \bar{w} \cdot \bar{x}_i)^2.$$

b. Dacă în locul perceptronului liniar vom considera perceptronul-prag, respectiv perceptronul sigmoidal, prin ce va difera forma regulii de actualizare a ponderilor față de rezultatul de la punctul a? (Pentru perceptronul-prag, se va presupune că $t_i \in \{-1, +1\}$, în vreme ce pentru perceptronul sigmoidal vom considera $t_i \in \{0, 1\}$, pentru $i = 1, \dots, n$.)

c. Elaborați algoritmul de antrenare a perceptronului liniar (și apoi, dacă dorîți, și a celui sigmoidal, dar veți preciza doar diferențele!). La *initializare*, ponderile w_i vor primi ca valori numere reale mici. În ce privește *condiția de oprire* a algoritmului, se poate opta pentru una din următoarele variante (eventual combinate):

- toate instanțele de antrenament sunt corect clasificate (în ipoteza că datele \bar{x}_i , cu $i = 1, \dots, n$, sunt liniar separabile);
- efectuarea unui număr prestabilit de iterății;
- verificarea condiției $|E(\bar{w}^{(t+1)}) - E(\bar{w}^{(t)})| < \varepsilon$, unde pragul numeric $\varepsilon > 0$ este stabilit inițial, iar $E(\bar{w}^{(t)})$ este valoarea *funcției de eroare* E (i.e., ceea ce mai sus am numit semisuma pătratelor erorilor) pe setul de date de antrenament, la finalul iterăției t .

d. Ce cunoașteți (de la curs) despre convergența algoritmului de la punctul c?

37.

(Deducerea regulii de actualizare a ponderilor pentru un tip particular de perceptron)

• ★ CMU, 1997 fall, Tom Mitchell, midterm, pr. 3.d

Se consideră o variantă particulară a perceptronului, [fără funcție de activare,] la care ieșirea o depinde de intrările x_i astfel:

$$o = w_0 + w_1 x_1 + w_1 x_1^3 + w_2 x_2 + w_2 x_2^3 + \dots + w_n x_n + w_n x_n^3$$

Derivați un algoritm de antrenare bazat pe metoda descreșterii gradientului, care să minimizeze suma pătratelor erorilor pentru acest tip de perceptron. Dați răspunsul sub forma

$$w_i \leftarrow w_i + \dots \quad \text{pentru } 1 \leq i \leq n.$$

38.

(Algoritmul *Perceptron* (al lui Rosenblatt): aplicare pe date din \mathbb{R}^2 , folosind termen liber / bias)

□ • CMU, 2015 spring, Alex Smola, midterm, pr. 4

Considerăm următoarele exemple de antrenament $(x, y) \in \mathbb{R}^2 \times \{\pm 1\}$, în ordinea indicată:

exemplul	1	2	3	4	5	6	7	8
instanța (x_1, x_2)	(10, 10)	(0, 0)	(8, 4)	(3, 3)	(4, 8)	(0.5, 0.5)	(4, 3)	(2, 5)
eticheta y	+1	-1	+1	-1	+1	-1	+1	+1

Aplicați algoritmul *Perceptron* (al lui Rosenblatt) pe această succesiune de exemple. La inițializare, vectorul de ponderi va primi valoarea $w = (1, 1)$, iar bias-ul (adică, termenul liber) va fi $b = 0$.⁷⁸¹

39.

(Învățare online cu perceptronul Rosenblatt)

*prelucrare de Liviu Ciortuz, după***(•) (*) CMU, 2008 spring, T. Mitchell, W. Cohen, HW2, pr. 4**
CMU, 2008 spring, T. Mitchell, W. Cohen, midterm exam, pr. 3.2

Problema 18 poate fi ușor reformulată în termenii învățării automate online.⁷⁸² În această situație, instanțele de antrenament nu sunt furnizate în avans perceptronului-prag. În schimb — după ce vectorul de ponderi w a fost inițializat cu valori arbitrară, sau cu 0 ca în problema 18 —, la fiecare iterație (k) a algoritmului de antrenare, supervisorul prezintă sistemului de învățare automată căte o instanță de antrenament x_{t_k} , fără a divulga și eticheta y_{t_k} . După ce a primit vectorul x_{t_k} , perceptronul calculează ieșirea o_{t_k} (în funcție de $w^{(k)}$), iar supervisorul îi comunică eticheta y_{t_k} . Dacă $y_{t_k} \neq o_{t_k}$, perceptronul își actualizează / modifică vectorul de ponderi w .

Dacă presupunem că instanțele de antrenament respectă condițiile specificate în problema 18 și ne punem întrebarea *care este numărul maxim de greșeli pe care le poate face perceptronul în regim online înainte de a „converge“*, răspunsul va fi: $\left\lfloor \left(R \frac{\|w^*\|}{\gamma} \right)^2 \right\rfloor$. Demonstrația este practic aceeași cu cea de la problema 18.⁷⁸³

A. Considerăm un set de date de antrenament notat cu $D = \{(x_1, y_1), \dots, (x_T, y_T)\}$, unde

- i. fiecare variabilă x_i reprezintă un document (d_i) în limba engleză, în care nu există mai mult de 100 de cuvinte (pentru această problemă, vom considera că nu există mai mult de 100 000 de cuvinte englezesti);
- ii. fiecărui document x_i îi este asociat un vector „rar“ (engl., sparse) $b_1, \dots, b_{100\,000}$, unde b_w este 1 în cazul în care cuvântul w — mai exact, cuvântul cu indicele w din dicționarul / lista cuvintelor din limba engleză — este prezent în documentul d_i , și 0 în caz contrar;
- iii. $y_i \in \{+1, -1\}$ este eticheta care desemnează clasa documentului d_i ;
- iv. există un vector w^* astfel încât $y_i w^* \cdot x_i > 1$ pentru orice document d_i .

Indicați căte o *margine superioară* pentru

- a. R — distanța maximă dintre originea sistemului de coordonate din $\mathbb{R}^{100\,000}$ și orice exemplu x_i ;

⁷⁸¹[LC:] Inițializarea cu 0 a bias-ului nu înseamnă că el va păstra această valoare pe parcursul întregii execuții a algoritmului. Atenție: faptul că se folosește bias înseamnă că va trebui să considerați în mod implicit (sau, să adăugați explicit) componenta $x_0 = 1$ la fiecare instanță de antrenament!

⁷⁸²Toate metodele de învățare prezentate până acum sunt modele off-line, adică nu permit / prevăd nicio interacțiune a supervisorului cu algoritmul de antrenare.

⁷⁸³Acesta este motivul pentru care am specificat acolo un număr posibil infinit de instanțe de antrenament x_1, \dots, x_n, \dots

b. m – numărul de erori făcute de perceptronul Rosenblatt înainte de a converge la o ipoteză corectă atunci când se face antrenarea pe datele din mulțimea D în regim de învățare online.

Observație: Se va considera că $\|w^*\| = 1$.

B. Acum vă cerem să concepeți un algoritm care produce o secvență de exemple care forțează perceptronul să producă un sir de lungime arbitrară (m) de greșeli, dacă oricare dintre cele două presupuneri de mai jos este eliminată, dar datele de antrenament rămân în continuare separabile:⁷⁸⁴

- instanțele de antrenament sunt liniar-separabile (de către un vector w^*), cu marginea $\gamma > 0$;⁷⁸⁵
- toate instanțele de antrenament sunt situate într-o sferă cu raza R , cu centrul în originea sistemului de coordonate din \mathbb{R}^d .⁷⁸⁶

40. (De la convergența antrenării perceptronului Rosenblatt la convergența antrenării perceptronului-prag.

Alte două proprietăți legate de convergența antrenării perceptronului Rosenblatt)

- CMU, 2013 spring, A. Smola, B. Poczos, HW2, pr. 2.a-c
CMU, 2015 spring, A. Smola, HW6, pr. 2

Fie o secvență formată din n instanțe de antrenament $x_i \in \mathbb{R}^d$, împreună cu etichetele corespunzătoare $y_i \in \{-1, 1\}$. Vă readucem aminte⁷⁸⁷ că algoritmul lui Rosenblatt pentru antrenarea unui perceptron-prag — în cele ce urmează vom folosi pentru acest algoritm numele de *Perceptron* — procedează astfel:

```

initialize  $w \leftarrow 0$ 
for  $i = 1, \dots, n$  do
    if  $y_i w \cdot x_i \leq 0$  then
         $w \leftarrow w + y_i x_i$ 
    end if
end for

```

Pentru simplitate, am inclus termenul liber (sau, bias-ul, de la engl., “bias”) w_0 în vectorul w , adică am renosat cu w vectorul $[w_0, w]$ și, similar, cu x_i vectorul $[1, x_i]$. Așadar, în această problemă nu este necesar să tratăm în mod explicit termenul liber.

Presupunem că $\|x_i\| \leq R$ pentru orice i . La problema 18 am demonstrat că atunci când există un vector de ponderi w^* astfel încât $\|w^*\| = 1$ și pentru orice i are loc inegalitatea

$$y_i w^* \cdot x_i \geq \gamma,$$

numărul de actualizări ale lui w este mărginit superior de

$$R^2/\gamma^2. \quad (344)$$

⁷⁸⁴Adică, există w^* astfel încât $y_i w^* \cdot x_i > 0$.

⁷⁸⁵Negarea acestei condiții înseamnă că pentru orice $\gamma > 0$ există o valoare a lui i astfel încât $0 < y_i w^* \cdot x_i < \gamma$.

⁷⁸⁶Negarea acestei condiții înseamnă că pentru orice $R > 0$ există o valoare a lui i astfel încât $\|x_i\| > R$.

⁷⁸⁷Vedeți problema 16.

a. Știm că regula [mai] generală pentru actualizarea perceptronului-prag este $w \leftarrow w + 2\eta y_i x_i$, unde $\eta > 0$ este rata de învățare.⁷⁸⁸ Așadar, algoritmul *Perceptron* al lui Rosenblatt corespunde cazului special $\eta = 1/2$.

Găsiți o margine superioară (engl., upper bound) pentru numărul de actualizări ale perceptronului-prag, similar cu modul în care a fost obținută marginea (344). Cum variază această margine în funcție de rata η ?

b. Din relația (344) știm că pentru valori mici ale lui γ problema antrenării *Perceptronului* poate fi dificil de rezolvat. De fapt, complexitatea ei poate fi chiar exponențială în raport cu numărul de instanțe de antrenament. Ca să ilustrăm acest fapt, vom folosi următorul *exemplu*:

$$y_i = (-1)^{i+1} \text{ și } x_i = (\underbrace{(-1)^i, \dots, (-1)^i}_{i \text{ elemente}}, (-1)^{i+1}, 0, \dots, 0) \text{ pentru } i = 1, \dots, m.$$

Demonstrați că sunt necesare $O(2^m)$ actualizări pentru ca algoritmul *Perceptron* să găsească o soluție optimă w^* , satisfăcând inegalitatea $y_i x_i \cdot w^* > 0$ pentru orice i , indiferent de modul în care sunt selectate instanțele la fiecare iterație.

c. Acum vom renunța la presupunerea că instanțele (x_i, y_i) sunt liniar separabile.⁷⁸⁹ (Viața nu este aşa de simplă! Însă, din fericire, nu este nici din cale afară de complicată.) Fie u un vector oarecare din \mathbb{R}^d , cu $\|u\| = 1$, și $\gamma > 0$. Definim *deviația* lui x_i prin expresia $d_i = \max(0, \gamma - y_i u \cdot x_i)$, iar $\delta = (\sum_{i=1}^n d_i^2)^{-1/2}$. Arătați că numărul de actualizări făcute de către algoritmul *Perceptron* — într-un „spațiu de trăsături“ ales în mod convenabil — este mărginit superior de $(R + \delta)^2 / \gamma^2$.⁷⁹⁰

Sugestie: Ați putea încerca ca, pornind de la instanțele x_i să construiți în mod convenabil instanțe separabile x'_i și apoi să refolosiți / adaptați demonstrația de la punctul precedent. De exemplu, puteți defini $x'_i = [x_i, 0, \dots, 0, c, 0, \dots, 0]$, vector în spațiul \mathbb{R}^{d+n} , având pe poziția $d+i$ constanta $c \in \mathbb{R}$. Cum se poate oare construi în mod corespunzător un vector de ponderi u' astfel încât u' să separe instanțele x'_i ? Apoi — odată ce ati obținut o margine superioară similară cu (344) —, cum ar trebui să fie aceeași constantă c astfel încât să minimizăm această margine superioară? Este oare această margine valabilă / bună [și] pentru instanțele originale x_i ?⁷⁹¹

⁷⁸⁸În enunțul original se dă pentru regula de actualizare forma $w \leftarrow w + \eta y_i x_i$. Noi însă am folosit definitia clasică: $w \leftarrow w - \eta \Delta w$, cu $\Delta w = \frac{\partial E_i}{\partial w} = -(y_i - o_i)x_i$, unde o_i este outputul perceptronului pentru inputul x_i . De remarcat că $y_i \neq o_i$ implică $y_i = 1$ și $o_i = -1$ sau $y_i = -1$ și $o_i = 1$, deci $y_i - o_i = 2y_i$ în ambele situații. Prin urmare, $\Delta w = -2y_i x_i$, iar forma regulii de actualizare este, în această abordare, $w \leftarrow w + 2\eta y_i x_i$.

⁷⁸⁹Deci nu există w^* astfel încât $y_i w^* \cdot x_i > 0$ pentru $i = 1, \dots, n$.

⁷⁹⁰*Observație:* Ideea principală din spatele demonstrației de la acest punct este maparea / „proiectarea“ instanțelor x_i într-un spațiu de dimensiune mai mare, în care separarea liniară să fie posibilă. Aceasta coincide cu ideea folosirii *funcțiilor-nucleu* (engl., kernel functions), care mapează / transformă x_i în $\phi(x_i)$. Pentru varianta kernel-izată [duală] a algoritmului *Perceptron*, vedeți problema 19.

⁷⁹¹Pentru o variantă de mapare mai simplă decât cea indicată aici, vedeți problema 73 de la capitolul de *Fundamente*.

41. (Proprietăți ale unei variante a algoritmului *Perceptron* kernel-izat, în cazul când nucleul este de tip RBF)

*prelucrare de L. Ciortuz, după
□ • ○ MIT, 2009 fall, Tommi Jaakkola, midterm, pr. 2*

Algoritmul *Perceptron* — cu care am făcut cunoștință la problema 16 — constituie probabil una dintre cele mai simple modalități de a rezolva probleme de clasificare. La problema 19 am prezentat varianta kernel-izată a acestui algoritm și am menționat funcția-nucleu de tip Gaussian (care se mai numește și *funcție cu bază radială*, RBF).⁷⁹² În pseudo-codul următor redăm algoritmul *Perceptron* kernel-izat într-o variantă ușor extinsă în raport cu pseudo-codul din problema 19, în sensul că aici se permite parcurgerea de mai multe ori a setului de date de antrenament.

Date de intrare: $x_1, \dots, x_n \in \mathbb{R}^d$ și $y_1, \dots, y_n \in \{-1, +1\}$;

Inițializare: $\alpha_1 = \dots = \alpha_n = 0$;

Procedură:

Parcure ciclic $i = 1, \dots, n$ până când nu se mai produce nicio greșeală
dacă $y_i(\sum_{j=1}^n \alpha_j y_j K(x_j, x_i)) \leq 0$, atunci $\alpha_i \leftarrow \alpha_i + 1$;

În cele de mai jos, funcția K va fi considerată funcție-nucleu de tip RBF, deci $K(x, x') = \exp\left(-\frac{1}{2\sigma^2}\|x - x'\|^2\right)$ și $\sigma > 0$.

a. Converge oare întotdeauna acest algoritm? (În acest context, *a converge* înseamnă că la un moment dat algoritmul nu mai actualizează variabilele α_i .) Trebuie oare să adăugăm vreo condiție suplimentară pentru a ne asigura că algoritmul nostru se oprește?

b. Indicați — și apoi justificați în mod riguros — care dintre afirmațiile următoare sunt *adevărate*.

i. Dacă algoritmul converge, atunci el găsește o soluție⁷⁹³ pentru care putem calcula „marginea“ de separare.⁷⁹⁴ Această „margine“ depinde de ordinea în care sunt parcuse exemplele de antrenament.

ii. Presupunem că toate instanțele de antrenament x_i , cu $i = 1, \dots, n$, sunt distințe. În acest caz, dacă parametrul nucleului RBF (și anume, σ) are o valoare suficient de mică, algoritmul converge în mod cert după ce va fi făcut maximum n clasificări eronate (engl., mistakes).

iii. Numărul de clasificări eronate făcute de acest algoritm (în cazul în care converge) depinde de valoarea parametrului σ .

Indicație: Pentru a putea răspunde corect la întrebările de mai sus, vă recomandăm să faceți referire la rezultatele teoretice demonstate la problema 29 de la capitolul *Mașini cu vectori-support* din prezenta culegere,⁷⁹⁵ precum și la problema 18 din capitolul de față (*Rețele neuronale artificiale*).

⁷⁹²Vedeți problema 74 de la capitolul de *Fundamente*.

⁷⁹³Adică un *separator liniar* în spațiul de trăsături \mathbb{R}^m , conform „mapării“ $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^m$ specifice nucleului RBF.

⁷⁹⁴Adică distanța de la [hiperplanul] separator până le cele mai apropiate „imagini“ $\phi(x_i)$.

⁷⁹⁵Ar fi util să vedeați și problema 10 de la capitolul *Metode de regresie*.

42.

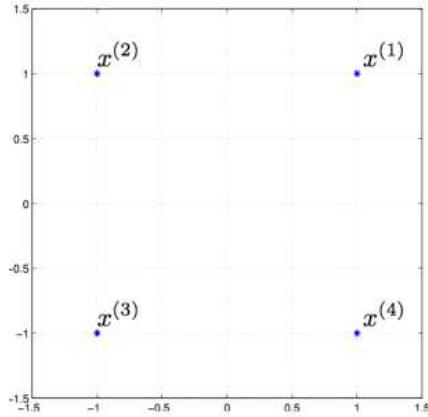
(Perceptronul Rosenblatt, cu termen liber (engl., offset): varianta simplă vs. varianta kernel-izată; aplicare pe un set simplu de date din \mathbb{R}^2)

prelucrare de Liviu Ciortuz, după

□ • · MIT, 2014 spring, T. Jaakkola, R. Barzilay, final exam, pr. 1

În acest exercițiu vom considera patru instanțe de antrenament, $x^{(1)}, \dots, x^{(4)} \in \mathbb{R}^2$, după cum se observă în figura alăturată. Aceste instanțe au respectiv etichetele $y^{(1)}, \dots, y^{(4)} \in \{-1, +1\}$. Veți explora diferite variante de a eticheta aceste instanțe și [mai precis] efectul pe care îl au aceste etichetări asupra algoritmului *perceptron cu termen liber* (engl., offset).

Acest algoritm este caracterizat de următorul pseudo-cod:



Initialize: $w = 0$ (vector), $w_0 = 0$

Cycle through $i = 1, \dots, n$

If $y^{(i)}(w \cdot x^{(i)} + w_0) \leq 0$ then

update $w \leftarrow w + y^{(i)}x^{(i)}$ and $w_0 \leftarrow w_0 + y^{(i)}$

Observație: Pseudo-codul acesta se obține în mod natural din cel al pseudo-codului perceptronului Rosenblatt (vedeți problema 16) dacă se consideră instanțele de forma $x \stackrel{\text{not.}}{=} (1, x_1, \dots, x_d)$, cu $d \in \mathbb{N}^*$.

Pentru o instanță de test x , algoritmul va prezice eticheta

$$\text{sign}(w \cdot x + w_0).$$

a. Vă cerem să elaborați varianta kernel-izată a acestui tip de *perceptron cu termen liber*. Pentru aceasta, ca de obicei, se va considera mai întâi că fiecare instanță x este „mapată” / transformată sub forma unui vector de trăsături $\phi(x)$, iar apoi algoritmul va fi reformulat astfel încât să folosească instanțele x doar prin intermediul funcției-nucleu $K(x, x') \stackrel{\text{def.}}{=} \phi(x) \cdot \phi(x')$.

Odată antrenat, algoritmul va prezice pentru instanța x eticheta y , care este exact valoarea expresiei

$$\text{sign}\left(\left(\sum_{i=1}^n \alpha_i y^{(i)} \phi(x^{(i)})\right) \cdot \phi(x) + w_0\right) = \text{sign}\left(\sum_{i=1}^n \alpha_i y^{(i)} K(x^{(i)}, x) + w_0\right),$$

unde $n = 4$ este numărul de exemple de antrenament, iar existența și rolul coeficienților $\alpha_i \in \mathbb{N}$ sunt justificate în mod similar cu cazul Perceptronului Rosenblatt kernel-izat (vedeți problema 19).⁷⁹⁶

⁷⁹⁶Veți observa că atât la problema 19 cât și la problema de față se parcurge o singură dată setul de date de antrenament. Însă la ambele probleme este posibilă extinderea algoritmului în așa fel încât să se poată face mai multe astfel de parcurgeri. De asemenea, în problema de față rolul coeficientului α_i va fi disociat de eticheta $y^{(i)}$, el indicând pur și simplu de câte ori instanța $x^{(i)}$ a cauzat erori / greșeli în timpul antrenamentului, provocând deci „actualizarea“ perceptronului.

Pentru o justificare cu caracter mai general a existenței și rolului coeficienților $\alpha_i \in \mathbb{R}$ în ce privește exprimarea [forme] soluțiilor unor probleme de clasificare sau de regresie care fac optimizarea unor funcții de cost / pierdere, vedeți *Teorema de reprezentare* care a fost demonstrată la problema 88 de la capitolul *Fundamente*.

Vă cerem să completați în pseudo-codul de mai jos detaliile algoritmului *perceptron cu termen liber kernel-izat*.

```

Initialize: ...
Cycle through  $i = 1, \dots, n$ 
If ... then
    update ...

```

b. Presupunem că rulăm algoritmul acesta kernel-izat folosind nucleul RBF

$$K(x, x') = \exp\left(-\frac{1}{2}\|x - x'\|^2\right).$$

Considerăm acum exemplele de antrenament din figura de mai sus. Are oare importanță cum sunt etichetate aceste patru puncte în ce privește convergența algoritmului? Justificați pe scurt răspunsul dumneavoastră.

Indicație: Vedeți proprietățile prezentate la problema 18 din prezentul capitol și problema 26.a de la capitolul *Mașini cu vectori-suport*.

c. Găsiți o etichetare pentru cele patru instanțe de antrenament

$$y^{(1)} = \dots, y^{(2)} = \dots, y^{(3)} = \dots, y^{(4)} = \dots$$

care să satisfacă următoarele două criterii:

- i. Varianta kernel-izată a algoritmului perceptron cu offset, folosind nucleul RBF menționat mai sus, converge după ce face o singură actualizare a coeficientilor (α, α_0).
- ii. Varianta simplă (deci ne-kernel-izată) a algoritmului perceptron cu offset converge, însă necesită mai multe actualizări ale ponderilor (w, w_0).

Faceți toate calculele necesare.

43. (Varianta perceptronului Rosenblatt pentru clasificare ternară: kernel-izare)

*prelucrare de Liviu Ciortuz, după
□ • · MIT, 2008 fall, Tommi Jaakkola, midterm, pr. 3.4*

În acest exercițiu vom considera varianta algoritmului Perceptron (al lui Rosenblatt) pentru rezolvarea unei probleme de clasificare ternară (adică, cu 3 clase). Setul de date de antrenament constă din n exemple de forma (x_i, y_i) , pentru $i = 1, \dots, n$, cu $x_i \in \mathbb{R}^d$ și $y_i \in \{1, 2, 3\}$. Pseudo-codul acestui algoritm este următorul:⁷⁹⁷

⁷⁹⁷Veți observa în acest pseudo-cod — analizați instrucțiunea if — că, spre deosebire de perceptronul kernelizat [dual] de la problema 19, unde se lucrează cu un singur vector de ponderi (w), aici se lucrează cu trei vectori de ponderi (w_1, w_2 și w_3), cîte un vector pentru fiecare clasă. (S-a procedat similar și în cazul regresiei liniare multivariante și în cel al regresiei logistice n -are (regresia *softmax*). Vedeți problemele 30 și respectiv 18 de la capitolul *Metode de regresie*.)

De asemenea, veți observa că spre deosebire de perceptronul kernelizat [dual], unde la fiecare greșeală de clasificare produsă de către perceptron se actualizează vectorul de ponderi w , aici la fiecare greșeală de clasificare se actualizează doi vectori: w_{y_i} — unde y_i este eticheta asignată instanței curente, x_i — și w_z — unde z este clasa indicată (în mod eronat) de către perceptronul ternar. Tot așa se procedează și în cazul perceptronului n -ar. Vedeți <http://proceedings.mlrf.press/v97/beygelzimer19a/beygelzimer19a-sup.pdf>. (LC: Mulțumesc studentului Răzvan Ciocoiu pentru aceste două precizări.)

Initializare:

```
for  $y \in \{1, 2, 3\}$ 
     $w_y = 0 \in \mathbb{R}^d$ 
```

Algoritm:

Repeat until convergence:

```
 $z = \arg \max_{y \in \{1, 2, 3\}} w_y \cdot x_i$ 
for  $i = 1, \dots, n$ 
    if  $z \neq y_i$ 
         $w_{y_i} = w_{y_i} + x_i$ 
         $w_z = w_z - x_i$ 
```

Clasificarea unei instanțe de test x :
 $f(x) = \arg \max_{y \in \{1, 2, 3\}} w_y \cdot x$

Vă cerem să elaborați o versiune kernel-izată a acestui algoritm de tip perceptron. Veți considera că nucleul pe care îl veți folosi este $K(x, z)$. Completăți pseudo-codul următor în aşa fel încât să rezulte algoritmul kernel-izat.

Initializare:

```
for  $y \in \{1, 2, 3\}$ , for  $i = 1, \dots, n$ 
     $\alpha_{i,y} = 0$ 
```

Algoritm:

Repeat until convergence:

Clasificarea unei instanțe de test x :
 $\arg \max_{y \in \{1, 2, 3\}} \sum_{i=1}^n \alpha_{i,y} K(x, x_i)$

44.

(Antrenarea perceptronului:
elaborare cod C / pseudo-cod)

Liviu Ciortuz, 2012

Elaborați fie în pseudo-cod fie în limbajul de programare C funcții pentru

- construirea (adică reprezentarea ca structură de date) a unui perceptron. Veți lucra cu n intrări (numere reale), iar funcția de activare va fi de *tip* prag, liniar sau sigmoidal. Într-o versiune ulterioară veți putea adăuga tipurile: generalizat-sigmoidal (a se vedea problema 20) sau tangentă hiperbolică (\tanh , ca în problema 47).
- calculul outputului unui perceptron (construit la punctul precedent), pentru un input dat;
- algoritmul de antrenare (adică algoritmul de actualizare a ponderilor corespunzătoare intrărilor perceptronului), conform tipului de funcție de activare indicat mai sus. Veți elabora atât varianta “batch” cât și varianta stochastică / incrementală (a se vedea cartea lui Tom Mitchell, *Machine Learning*, pag. 92-94).

Indicație: Acest algoritm se va opri atunci când una din următoarele condiții sunt satisfăcute:

- toate instanțele de antrenament sunt corect clasificate, sau

– s-au executat deja un număr *maxim* de iterații dat ca parametru. Alternativ, se poate testa dacă funcția de optimizat (de exemplu, semisuma pătratelor erorilor, sau suma costurilor / pierderilor de tip log-sigmoidal (vedeți problema 14)) a scăzut sub un anumit prag ε (dat ca parametru) la ultima iterație executată, în comparație cu iterația precedentă.

Vă sugerăm să testați implementarea dumneavoastră pe datele de la problemele 2.a, 29, 30.a și 9 (pentru punctele *a* și *b*), și respectiv problema 11 (pentru punctul *c*). Ulterior, veți putea implementa varianta în care funcția de activare este de tip log-sigmoidal (ca la problema 14), sau de tip tangentă hiperbolică (ca la problema 47).

6.2.3 Rețele “feed-forward” — algoritmul de retropropagare

45.

(Rețele “feed-forward”: algoritmul de retro-propagare, aplicare)

• * CMU, 2014 spring, Seyoung Kim, HW2, pr. 1.3

Considerăm că toate unitățile rețelei neuronale din figura de mai jos sunt de tip sigmoidal.

Vă readucem aminte că acest fapt înseamnă că pentru inputul x_1, \dots, x_d , o astfel de unitate neuronală calculează outputul

$$\frac{1}{1 + e^{-(w_0 + \sum_i w_i x_i)}},$$

unde w_i este ponderea corespunzătoare intrării x_i , pentru fiecare $i = 1, \dots, d$.

În cele ce urmează vom considera că $w_0 = 0$ pentru fiecare unitate din această rețea.

a. Presupunem că inputul rețelei este $i_1 = 1$.

Cât va fi valoarea outputului calculat de către nodul o_1 ?

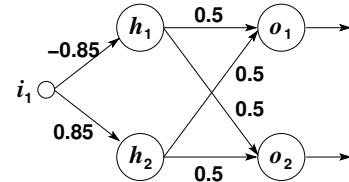
Dar valoarea outputului calculat de către nodul o_2 ?

b. De data aceasta presupunem că outputurile corecte pentru inputul $i_1 = 1$ sunt $t_1 = t_2 = 1$. Folosind formulele date pentru cantitățile δ_i în algoritmul de retro-propagare — a se vedea pr. 20 și / sau cartea *Machine Learning* de Tom Mitchell, pag. 98 —, calculați:

- valoarea lui δ_{o_1} ;
- valoarea lui δ_{o_2} .

c. În final, folosind aceste valori pentru δ și considerând că rata de învățare este $\eta = 0.1$, calculați:

- valoarea lui δ_{h_1} ;
- noua valoare a ponderii de pe conexiunea care leagă unitatea h_1 de unitatea o_1 ;
- noua valoare a ponderii de pe conexiunea care leagă unitatea h_1 de unitatea o_2 ;
- noua valoare a ponderii de pe conexiunea care leagă intrarea i_1 de unitatea h_1 .



46.

(Algoritmul de retro-propagare: Adevărat sau Fals?)

• * CMU, 1994 fall, A. Blum, T. Mitchell, HW4, pr. 1

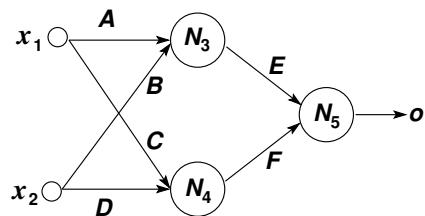
Rețeaua neuronală din figura alăturată este formată din unități de același tip, de exemplu (pentru fixarea ideilor) sigmoidală.

Presupunem că executăm algoritmul de retro-propagare pe această rețea, folosind un set oarecare de date de antrenament și dând inițial valori egale tuturor ponderilor din rețea.

a. Precizați care dintre afirmațiile de mai jos vor fi întotdeauna adevărate:

- i. Ponderile A și B nu vor difera niciodată.
- ii. Ponderile A și C nu vor difera niciodată.
- iii. Ponderile E și F nu vor difera niciodată.

b. Justificați de ce NU este bine să se initializeze toate ponderile dintr-o rețea neuronală cu o aceeași valoare atunci când se rulează algoritmul de retro-propagare.



47.

(Rețele neuronale — algoritmul de retro-propagare pentru rețele de unități neuronale având funcția de activare \tanh)

* prelucrare de Liviu Ciortuz, după T. Mitchell, "Machine Learning", 1997, pr. 4.8 și CMU, 2011 spring, Roni Rosenfeld, HW4, pr. 2.a

Rescrieți algoritmul de retro-propagare pentru rețele neuronale aciclice în care unitățile neuronale folosesc funcția de activare tangentă hiperbolică (\tanh) în locul funcției sigmoidale.

Recomandare: Pentru a reduce din generalitatea problemei, vă veți limita la a elabora varianta stochastică / incrementală a acestui algoritm pentru rețele de tip feed-forward cu două niveluri.

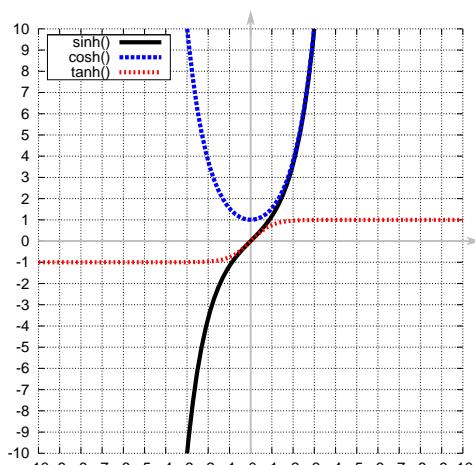
Vă reamintim că

$$\tanh(y) = \frac{e^y - e^{-y}}{e^y + e^{-y}} = \frac{e^{2y} - 1}{e^{2y} + 1} = \frac{\sinh(y)}{\cosh(y)},$$

iar

$$\sinh(y) = \frac{e^y - e^{-y}}{2} \text{ și } \cosh(y) = \frac{e^y + e^{-y}}{2}$$

Din definiția de mai sus rezultă imediat că funcția \tanh este derivabilă și deci continuă pe întreg domeniul de definiție (axa reală), iar din imaginea alăturată se observă că funcția \tanh este o formă „netezită“ (engl., smoothed) a funcției treaptă $sign$.



Adaptare după

http://en.wikipedia.org/wiki/File:Sinh_cosh_tanh.svg

Așadar, vom considera că outputul unităților neuronale din rețea este de forma $o(\bar{x}) = \tanh(\bar{w} \cdot \bar{x})$.

a. Demonstrați că $\tanh'(y) = 1 - \tanh^2(y)$ pentru orice $y \in \mathbb{R}$.

b. Arătați că pentru regula de actualizare a ponderilor unităților neuronale de pe nivelul de ieșire, $w_{kj} \leftarrow w_{kj} + \Delta w_{kj}$, avem

$$\Delta w_{kj} \stackrel{\text{def.}}{=} -\eta \frac{\partial E}{\partial w_{kj}} = \eta(t_k - o_k)(1 - \tanh^2(\text{net}_k))y_j$$

unde notațiile sunt similare cu cele din cartea *Machine Learning* de Tom Mitchell.⁷⁹⁸

c. Similar, pentru ponderile unităților neuronale de pe nivelul ascuns:

$$\Delta w_{ji} \stackrel{\text{def.}}{=} -\eta \frac{\partial E}{\partial w_{ji}} = \eta(1 - \tanh^2(\text{net}_j))x_i \left(\sum_{k \in \text{Downstream}(j)} \delta_k w_{kj} \right),$$

unde

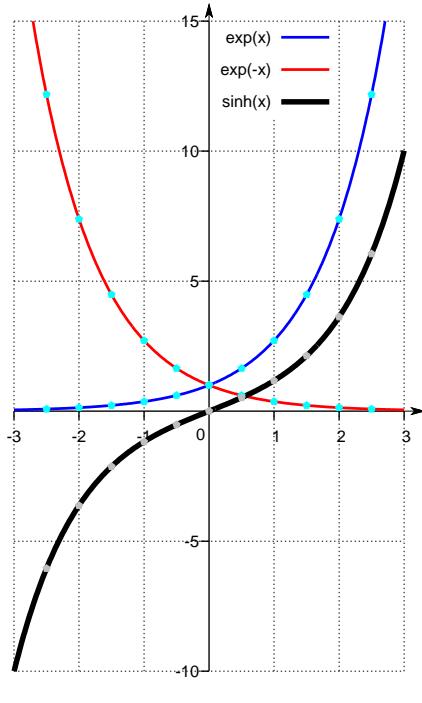
$$\delta_k = -\frac{\partial E_d(\bar{w})}{\partial \text{net}_k} \text{ iar } E_d(\bar{w}) = \sum_{k \in \text{outputs}} \frac{1}{2}(t_{k,d} - o_{k,d})^2,$$

d fiind o instanță de antrenament oarecare sau (echivalent) indicele asociat ei.

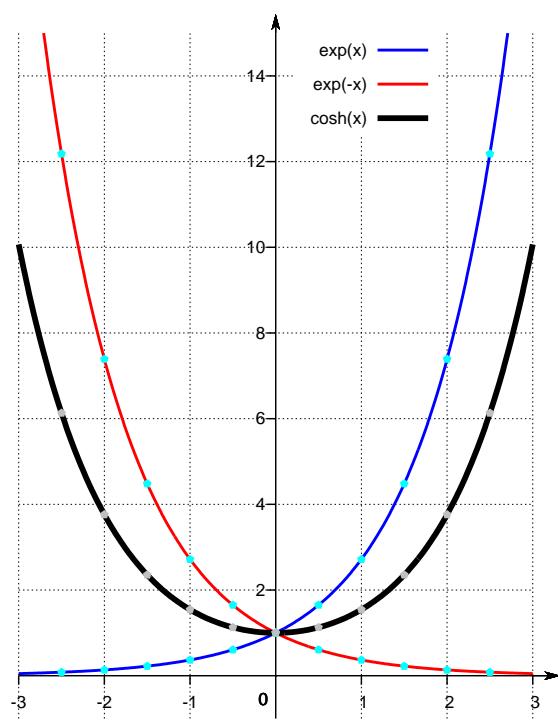
Adăugăm câteva *observații* cu caracter instructiv / informativ în legătură cu funcția \tanh :

(1) Pentru a vedea relația dintre funcția \sinh (și respectiv \cosh) pe de o parte și funcțiile exponentiale (e^x) și invers-exponentiale (e^{-x}) pe de altă parte, cititorul poate observa graficele de mai jos.

⁷⁹⁸Aceleași notații au fost folosite și în problema 20.



Adaptat după:
http://en.wikipedia.org/wiki/File:Hyperbolic_and_exponential_sinh.svg
 Observație: Funcția \sinh este semidiferența funcțiilor e^x și e^{-x} .



Adaptat după: http://en.wikipedia.org/wiki/File:Hyperbolic_and_exponential_cosh.svg
 Observație: Funcția \cosh este media funcțiilor e^x și e^{-x} .

(2) Legătura dintre funcția sigmoidală $\sigma(x) = \frac{1}{1 + e^{-x}}$ (care mai este numită și *funcția logistică*) și funcția *tanh* este dată de relația:

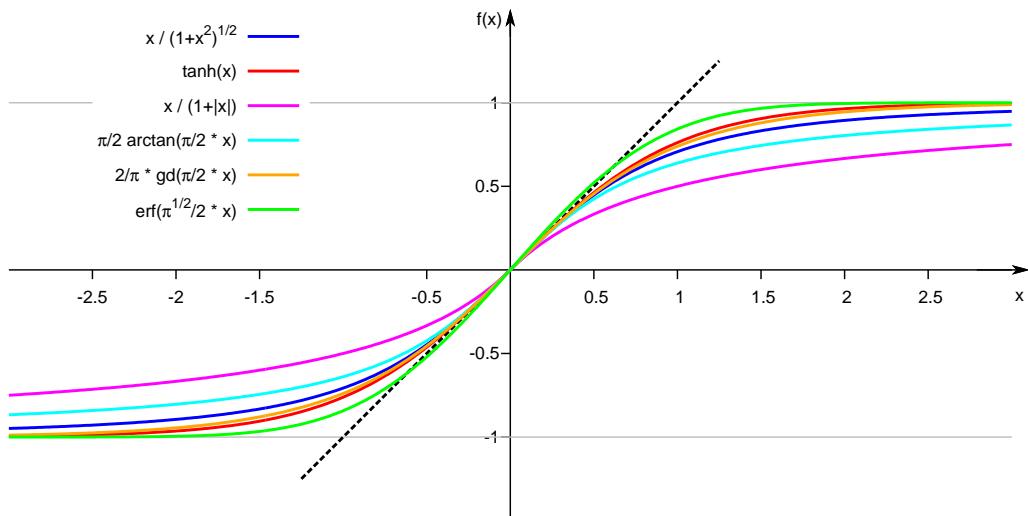
$$\sigma(x) = \frac{1}{2} \left(1 + \tanh \frac{x}{2} \right)$$

(3) Spre deosebire de funcția sigmoidală (care este o variantă „netezită“ a funcției treaptă 0/1), funcția *tanh* este simetrică față de originea sistemului de coordonate. Din această cauză, în literatura de specialitate se apreciază că funcția *tanh* favorizează o convergență mai rapidă a algoritmului de antrenare a rețelei neuronale.

(4) Există și alte funcții care „netezesc“ funcțiile-treaptă, de exemplu: $\frac{x}{1+|x|}$, $\frac{x}{\sqrt{1+x^2}}$, funcția arctg (inversa funcției tangentă trigonometrică), funcția de eroare gaussiană (notată *erf*, prin abreviere de la engl. error function) și funcția Gudermannian (*gd*), unde

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt, \text{ iar } gd(x) = \int_0^x \frac{dt}{\cosh(t)} = 2 \arctg(e^x) - \frac{\pi}{2}$$

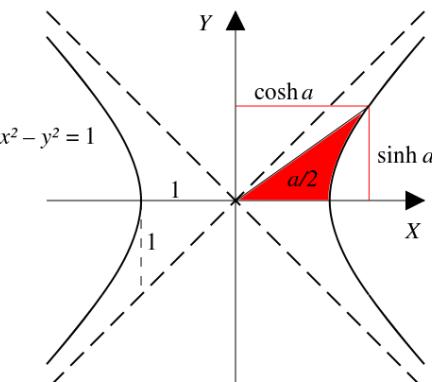
Pentru aceste funcții se pot defini versiuni „normalizate“, în așa fel încât panta tangentei lor pentru $x = 0$ să fie 1. A se vedea graficul următor.



Adaptat după: [http://en.wikipedia.org/wiki/File:Gjl-t\(x\).svg](http://en.wikipedia.org/wiki/File:Gjl-t(x).svg)

(5) Legătura dintre funcțiile trigonometrice hiperbolice și funcțiile trigonometrice clasice („circulare“) poate fi ilustrată astfel:

Pentru funcțiile trigonometrice circulare avem proprietatea fundamentală: $\sin^2 \alpha + \cos^2 \alpha = 1$. În consecință, dacă notăm $x = \cos \alpha$ și $y = \sin \alpha$, rezultă că punctul (x, y) este situat pe cercul de ecuație $x^2 + y^2 = 1$ (având centrul O și raza 1). Pentru funcțiile trigonometrice hiperbolice avem o altă proprietate, care se verifică imediat: $\cosh^2 a - \sinh^2 a = 1$. Așadar, dacă notăm $x' = \cosh a$ și $y' = \sinh a$, rezultă că punctul (x', y') este situat pe hiperbola de ecuație $(x')^2 - (y')^2 = 1$, ale cărei asymptote la $+\infty$ și $-\infty$ sunt prima și cea de-a doua bisectoare a sistemului de coordinate în plan. (De remarcat proprietatea: pentru $\cosh a$ și $\sinh a$, aria zonei hașurate din imagine este $a/2$.)



Sursa: http://en.wikipedia.org/wiki/File:Hyperbolic_functions-2.svg

48.

(Aplicarea algoritmului de retro-propagare pe o rețea feed-forward formată din unități sigmoidale dispuse pe 2 niveluri, utilizând și componenta „moment“)

adaptare făcută de către Liviu Ciortuz, după
• * T. Mitchell, "Machine Learning", 1997, pr. 4.7

Se consideră o rețea neuronală de tip feed-forward cu două niveluri, având două intrări a și b , o unitate ascunsă c și o unitate de ieșire d . Ambele unități sunt de tip sigmoidal. Această rețea are 5 ponderi: $w_{ca}, w_{cb}, w_{c0}, w_{dc}, w_{d0}$, unde w_{c0} reprezintă ponderea asociată termenului liber ($x_0 = 1$) pentru unitatea c și similar pentru unitatea d .

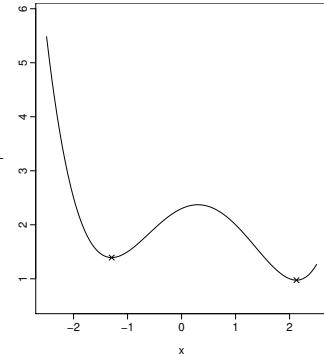
Inițializați ponderile cu valorile $(0.1, 0.1, 0.1, 0.1, 0.1)$, apoi calculați valorile lor după fiecare dintre primele două iterații ale algoritmului de retro-propagare în varianta ‘‘batch’’. Considerăm rata de învățare $\eta = 0.3$, momentul $\alpha = 0.9$ și datele de antrenament din tabelul alăturat.⁷⁹⁹

a	b	t
1	0	1
0	1	0

Comentariu:

Vă readucem aminte următoarele:

1. La aplicarea metodei gradientului descendente, pentru evitarea ‘‘căderii’’ în unele minime locale, regula de actualizare a parametrilor w_{ji} , și anume $w_{ji}^{(n)} \leftarrow w_{ji}^{(n-1)} + \Delta w_{ji}^{(n)}$, se poate extinde cu încă un termen, $\alpha \Delta w_{ji}^{(n-1)}$, numit *termenul moment*, datorită analogiei cu mișcarea unui punct material din fizică. (Constanta $\alpha > 0$ desemnează cât anume dorim să păstrăm din ‘‘inertia’’ de mișcare a particulei considerate.) Veți lua $\Delta w_{(ji)}^{(0)} = 0$ pentru orice j și orice i .



2. Folosind notațiile de la pag. 101 din cartea *Machine Learning* de Tom Mitchell, avem $\Delta w_{ji}^{(n)} = -\eta \frac{\partial E}{\partial w_{ji}}(w^{(n-1)})$. Făcând abstracție de n , această cantitate Δw_{ji} este $\eta (\sum_l \delta_{j,l} x_{j,i,l})$, cu $\delta_{j,l} = o_{j,l}(1 - o_{j,l})(t_{j,l} - o_{j,l})$ atunci când j este indicele unei unități de pe nivelul de ieșire și, respectiv, $\delta_{j,l} = o_{j,l}(1 - o_{j,l}) \left(\sum_{k \in \text{Downstream}(j)} \delta_{k,l} w_{kj} \right)$ atunci când j este indicele unei unități de pe un nivel ascuns. Indicele l parcurge setul de instanțe de antrenament.

49.

(Regularizare:
prevenirea overfitting-ului pentru o rețea feed-forward
cu unități de tip (generalizat-)sigmoidale;
extensie pentru problemele 20 și 47)

*prelucrare de Liviu Ciortuz, după
Tom Mitchell, ‘‘Machine Learning’’, 1997, pr. 4.10
CMU, 2011 spring, Roni Rosenfeld, HW4, pr. 2.b*

Considerăm o rețea neuronală de tip feed-forward care folosește o funcție de eroare adecvată pentru prevenirea overfitting-ului (a se vedea problema 22):

$$E_D(w) = \sum_{d \in D} \sum_{k \in \text{outputs}} \frac{1}{2} (t_{kd} - o_{kd})^2 + \gamma \sum_{i,j} w_{ji}^2,$$

unde D este mulțimea instanțelor de antrenament.

Deduceți regulile de actualizare a ponderilor de pe conexiunile din rețea, corespunzător acestei definiții a erorii presupunând că rețeaua folosește

⁷⁹⁹Dacă limităm exercițiul acesta la aplicarea manuală a primei iterații, atunci nu avem de-a face cu componenta *moment*, însă el rămâne un exercițiu instructiv prin faptul că se solicită aplicarea algoritmului de retro-propagare în regim ‘‘batch’’, adică non-incremental (ceea ce nu mai apare, deocamdată, în niciun alt exercițiu din acest capitol). Aceasta implică folosirea formulelor de actualizare a ponderilor în forma din *Comentariul* al doilea de mai jos.

- a. doar unități (generalizat-)sigmoidale (ca la problema 20);
 b. doar unități cu funcția de activare de tip $tanh$ (ca la problema 47).

50. (Algoritmul de retro-propagare pentru rețele feed-forward cu funcția obiectiv de tip cross-entropie)

• CMU, 2008 fall, E. Xing, HW2, pr. 2.2
 CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh, HW3, pr. 2.2

Fie o rețea neuronală de tip feed-forward având d intrări, H unități pe nivelul ascuns și K unități pe nivelul de ieșire. Toate unitățile sunt de tip sigmoidal. Vom considera N exemple de antrenament, iar funcția de eroare cu care vom lucra este de tip cross-entropie:

$$E(w) = - \sum_{n=1}^N \sum_{k=1}^K [t_{nk} \ln y_{nk} + (1 - t_{nk}) \ln(1 - y_{nk})]$$

unde

t_{nk} este target-ul pentru al n -lea exemplu și pentru ieșirea k a rețelei,

y_{nk} este răspunsul rețelei pentru ieșirea k atunci când cel de-al n -lea exemplu este furnizat rețelei.

În această problemă veți calcula regula de actualizare pentru algoritmul de tip retro-propagare (folosind metoda gradientului descendente) în varianta stochastică. După cum știți, regula de actualizare a ponderii de pe conexiunea care leagă intrarea i de unitatea j este:

$$w_{ji}^{(t)} = w_{ji}^{(t-1)} - \eta \frac{\partial E_n}{\partial w_{ji}^{(t-1)}},$$

unde E_n este funcția de eroare pentru exemplul n .

Vom nota cu net_j intrarea în componenta de activare a unității j :

$$net_j = \sum_i w_{ji} z_i,$$

unde z_i este outputul unității ascunse i (dacă unitatea j este pe nivelul de ieșire) sau intrarea i (dacă unitatea j este pe nivelul ascuns).

Observații: (i) Întrucât veți calcula regula de actualizare pentru un singur exemplu, puteți elimina indicele n din notațiile de mai sus. (ii) În mod implicit vom considera că nu se folosesc termeni liberi (engl., bias) pentru neuronii rețelei.

- a. Câți parametri (i.e., câte ponderi) are rețeaua aceasta? (Veți presupune, la acest punct, că se folosesc termeni liberi.)
 b. Calculați derivata funcției de activare, $\sigma'(x)$, în funcție de însuși $\sigma(x)$.

- c. Definim $\delta_j = \frac{\partial E_n}{\partial net_j}$. Arătați că $\frac{\partial E_n}{\partial w_{ji}} = \delta_j z_i$.⁸⁰⁰
- d. Arătați că pentru orice unitate de ieșire k , avem $\delta_k = y_k - t_k$.
- e. Considerăm intrarea (x_1, x_2, \dots, x_d) . Deducreți regula de actualizare pentru $w_{kj}^{(t)}$, valoarea ponderii de pe conexiunea dintre unitatea ascunsă j și unitatea de ieșire k la iterată t , în funcție de x_i și de valoarea ponderilor din rețea la iterată $t - 1$.
- f. Pentru o unitate ascunsă j oarecare, avem $\delta_j = \frac{\partial E_n}{\partial net_j} = \sum_k \frac{\partial E_n}{\partial net_k} \frac{\partial net_k}{\partial net_j}$. Arătați că $\delta_j = \sigma(net_j)(1 - \sigma(net_j)) \sum_k \delta_k w_{kj}$.
- g. Considerăm intrarea (x_1, x_2, \dots, x_d) . Deducreți regula de actualizare pentru $w_{ji}^{(t)}$, valoarea ponderii de pe conexiunea dintre intrarea i și unitatea ascunsă j la iterată t , în funcție de x_i și de valoarea ponderilor din rețea la iterată $t - 1$.

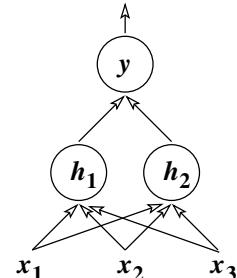
51.

(Aplicarea metodei gradientului descendente pentru o rețea formată din unități cu funcția de activare ReL)

 • ○ CMU, 2015 spring, Alex Smola, final, pr. 2.2

În figura de mai jos vi se dă structura unei rețele neuronale simple, având un singur nivel ascuns.

Inputul este 3-dimensional: $x = (x_1, x_2, x_3)$. Nivelul ascuns este constituit din două unități, deci vom putea scrie outputul acestui nivel sub forma $h = (h_1, h_2)$. Nivelul de ieșire este constituit dintr-o singură unitate, y . Pentru conveniență, nu vom folosi termeni liber (adică, bias-uri). Pentru toate unitățile neuronale din această rețea vom folosi funcția de activare liniar-rectificată (engl., linear rectified activation function, ReL), definită prin expresia $f(z) = \max(0, z)$ pentru orice $z \in \mathbb{R}$.



Considerăm funcția de eroare / cost / pierdere (engl., loss function) $l(y, t) = \frac{1}{2}(y - t)^2$, unde t este valoarea target pentru inputul x , iar y este valoarea produsă de rețea pentru același input. De asemenea, vom nota cu W și V matricele de ponderi care corespund conexiunilor dintre nivelul de intrare și nivelul ascuns, și respectiv dintre nivelul ascuns și nivelul de ieșire. Aceste matrice sunt inițializate astfel:

$$W = \begin{bmatrix} 1 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix} \text{ și } V = \begin{bmatrix} 0 & 1 \end{bmatrix}.$$

⁸⁰⁰Egalitatea $\frac{\partial E_n}{\partial w_{ji}} = \delta_j z_i$ are loc nu doar pentru funcția E din enunț (o cross-entropie), ci și, de exemplu, pentru semisuma pătratelor erorilor (vedeți pr. 20). Mai general, ea are loc pentru acele funcții de cost E în care ponderile w sunt folosite doar în interiorul termenilor de forma $net_j \stackrel{\text{def.}}{=} \sum_i w_{ii}x_i$. În rest — vedeți, de exemplu, cazul folosirii termenilor pentru regularizare, pr. 22 —, această manieră de lucru (centrată pe calculul cantităților δ) nu mai este adekvată / utilizabilă.

- a. Folosind simbolii f , W and V , scrieți [în manieră matriceală] expresia funcției $x \rightarrow y$ calculate de către rețea neuronală dată. (Nu este nevoie să folosiți aici valorile care au fost specificate mai sus pentru matricele W și V .)
- b. Presupunem că inputul rețelei este $x = (1, 2, 1)$, iar valoarea target este $t = 1$. Calculați valoarea *numerică* pentru outputul y , elaborând în mod clar toți pașii intermediari. Puteți refolosi rezultatele de la punctul precedent. Folosirea formei matriceale este recomandată, dar nu este obligatorie.
- c. Calculați vectorul gradient pentru funcția de eroare l , în raport cu ponderile rețelei. În mod specific, calculați mai întâi [în manieră *symbolică*] expresiile gradientului

- relativ la matricea de ponderi V , adică $\frac{\partial l}{\partial V}$;
- relativ la matricea de ponderi W , adică $\frac{\partial l}{\partial W}$.

Calculați apoi valoarea *numerică* a gradientului funcției l , pentru valorile specificate mai sus pentru W , V , x și y .

52.

(Adevărat / Fals?)

- CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh, midterm, pr. 1.5
CMU, 2012 fall, T. Mitchell, Z. Bar-Joseph, midterm, pr. 1.a

- a. Granița de decizie învățată de către o rețea neuronală este întotdeauna neliniară.
- b. Este posibil ca la antrenarea perceptronului liniar — aşadar, urmărind să minimizăm suma pătratelor erorilor, folosind metoda gradientului descendente — să obținem puncte [multiple] de optim local.
- c. Indiferent de mărimea rețelei neuronale, algoritmul de retro-propagare poate întotdeauna să găsească valorile optime globale pentru ponderile rețelei.

53.

(Antrenarea rețelelor feed-forward
în diverse variante: elaborare cod C / pseudo-cod)*Liviu Ciortuz, 2012*

Elaborați fie în pseudo-cod fie în limbajul de programare C funcții pentru

- a. configurarea (adică reprezentarea ca structură de date) a unei rețele neuronale artificiale de tip feed-forward;

Indicație: Într-o primă versiune, veți lucra cu n intrări (numere reale), un singur nivel ascuns și o singură unitate pe nivelul de ieșire, iar toate unitățile neuronale vor fi de același tip (de exemplu, sigmoidal). Ulterior, veți elabora o variantă extinsă a acestei funcții, în așa fel încât să se permită lucrul cu K niveluri ascunse, iar pentru fiecare unitate din rețea să se poată alege unul din tipurile de perceptri indicate la problema 44.

- b. calculul outputului unei rețele de tipul indicat la punctul precedent, pentru un input dat.

Indicații:

1. La punctele *a* și *b* veți putea folosi funcțiile corespunzătoare de la implementarea perceptronului (vedeți problema 44).
2. Vă sugerăm să testați implementarea dumneavoastră (la acest stadiu) pe datele de la problemele 1.ab, 2.b, 3, 6, 8, [7,] 30.b și 33.

c. algoritmul de antrenare (adică algoritmul de retro-propagare) pentru rețele neuronale, de asemenea de tipul indicat mai sus. Veți elabora atât varianta "batch" cât și varianta stochastică / incrementală (a se vedea cartea lui Tom Mitchell, *Machine Learning*, pag. 92-94).

Indicație: Vă sugerăm să elaborați mai întâi o versiune de bază a acestei funcții, pe care să o testați pe datele de la problema rezolvată 21 (verificând la final rezultatele) și de la problemele propuse 45 și 46. Ulterior puteți extinde această versiune (fie în manieră parametrizată, fie orientată pe obiecte) în așa fel încât să poată opera cu diverse variante ale *funcției obiectiv* de optimizat: fie semisuma pătratelor erorilor, fie varianta cu *regularizare*, pentru prevenirea overfitting-ului (ca în problemele 22 și 49), fie o funcție de tip cross-entropie (ca în problema 50). În sfârșit, încă o variantă posibilă este cea în care regulile de actualizare a ponderilor conțin și o componentă de tip „moment“, pentru evitarea unor optime locale ale funcției de eroare (a se vedea problema 48).

54.

(Rețele neuronale convolutive:
determinarea mărimii hărții de trăsături
de pe un anumit nivel)

CMU, 2015 fall, E. Xing, Z. Bar-Joseph, midterm, pr. 4.1

Considerăm un nivel convolutiv *C* urmat de un nivel de tip selecție a maximului (engl., max pooling layer) *P*. Inputul nivelului *C* are 50 de *canale* / planuri, fiecare dintre ele având dimensiunea 12×12 . Nivelul *C* are 20 de *filtre*, fiecare dintre acestea fiind de mărime 4×4 . Adausul / bordarea convolutivă (engl., the convolution padding) este de mărime 1, iar lungimea *pasului*⁸⁰¹ (engl., the stride) este 2.

Nivelul *P* efectuează o selecție a maximului din fiecare *hartă de trăsături* (engl., feature maps) din outputul nivelului *C*, cu *ferestre* (sau, câmpuri locale receptive; engl., local receptive fields) de dimensiune 3×3 și lungimea *pasului* este 1.

Cât este mărimea fiecărei *hărți de trăsături* (engl., feature map) din outputul nivelului *P*?

⁸⁰¹Pasul se referă la deplasarea *ferestrei* care corespunde filtrului de pe nivelul respectiv.



© M. Romanică

7 Clusterizare

Sumar

7.0 Noțiuni de bază

- instanță neetichetată vs. instanță etichetată (exemplu de antrenament);
- învățare nesupervizată (de exemplu: clusterizare, estimarea parametrilor distribuțiilor probabiliste și reducerea dimensionalității) vs. învățare supervizată (de exemplu: clasificare și regresie);
- [funcție / măsură de] distanță definită pe $\mathbb{R}^d \times \mathbb{R}^d$: ex. 2 de la capitolul *Învățare bazată pe memorare*;
- cluster / grup / grupare / bin (engl.) vs. clasă;
- tipuri de clusterizare: ierarhică vs. neierarhică;
- tipuri de ierarhii: ierarhii (arbori de clusterizare, dendrograme) obișnuite vs. ierarhii plate (engl., flat hierarchies);
exemple: ex. 1.a și respectiv ex. 1.b, ex. 6.a;
- tipuri de apartenență a unei instanțe la un cluster: “hard” vs. “soft” (ultima numai pt. clusterizare neierarhică).

7.1. Clusterizare ierarhică

7.1.1. Noțiuni specifice

- [funcție de] similaritate între clustere, definită pe baza [extinderii] noțiunii de distanță la $\mathcal{P}(X) \times \mathcal{P}(X)$, unde $X \subset \mathbb{R}^d$ este mulțimea de instanțe, iar $\mathcal{P}(X)$ este mulțimea părților lui X ;
tipuri de [funcții de] similaritate:

“single-linkage”:⁸⁰² $d(A, B) = \min\{d(x, y) | x \in A, y \in B\}$

“complete-linkage”:⁸⁰³ $d(A, B) = \max\{d(x, y) | x \in A, y \in B\}$

“average-linkage”: $d(A, B) = \frac{1}{|A| |B|} \sum_{x \in A, y \in B} d(x, y)$

metrica lui Ward (bazată pe centroizi): ex. 33, 34, 35.

În general, putem considera $sim(A, B) = 1/(1 + d(A, B))$ sau chiar $sim(A, B) = 1/d(A, B)$ dacă lucrăm doar cu clustere non-singleton;

proprietate / restricție: $sim(A \cup B, C) \leq \min\{sim(A, C), sim(B, C)\}$ pentru orice clustere A, B selectate de algoritmul de clusterizare ierarhică la execuția unei iterări oarecare [din bucla principală a algoritmului] și orice alt cluster C ;

- [funcție de] coeziune internă a unui cluster (sau: între elementele / instanțele dintr-un cluster);

exemplu (pentru clustere non-singleton): inversul mediei aritmetice a distanțelor dintre perechi de puncte din clusterul respectiv, adică

$$coh(A) = \left(\frac{1}{C_{|A|}^2} \sum_{x, y \in A} d(x, y) \right)^{-1} = \frac{C_{|A|}^2}{\sum_{x, y \in A} d(x, y)}.$$

⁸⁰²Sau: nearest-neighbour.

⁸⁰³Sau: furthest-neighbour.

7.1.2. Algoritmi de clusterizare ierarhică

- tipuri de algoritmi de clusterizare ierarhică:
bottom-up (*clusterizare aglomerativă*) vs. top-down (*clusterizare divizivă*);
- pseudo-coduri (cf. Manning & Schütze, *Foundations of Statistical Natural Language Processing*, 2002, pag. 502):

bottom-up:

```

Given: a set  $X = \{x_1, \dots, x_n\}$  of objects
       a function  $\text{sim}: \mathcal{P}(X) \times \mathcal{P}(X) \rightarrow R$ 
for  $i = 1, n$  do
     $c_i = \{x_i\}$  end
 $C = \{c_1, \dots, c_n\}$ 
 $j = n + 1$ 
while  $|C| > 1$ 
     $(c_{n_1}, c_{n_2}) = \text{argmax}_{(c_u, c_v) \in C \times C} \text{sim}(c_u, c_v)$ 
     $c_j = c_{n_1} \cup c_{n_2}$ 
     $C = C \setminus \{c_{n_1}, c_{n_2}\} \cup \{c_j\}$ 
     $j = j + 1$ 

```

top-down:

```

Given: a set  $X = \{x_1, \dots, x_n\}$  of objects
       a function  $\text{coh}: \mathcal{P}(X) \rightarrow R$ 
       a function  $\text{split}: \mathcal{P}(X) \rightarrow \mathcal{P}(X) \times \mathcal{P}(X)$ 
 $C = \{X\} (= \{c_1\})$ 
 $j = 1$ 
while  $\exists c_i \in C$  such that  $|c_i| > 1$ 
     $c_u = \arg \min_{c_v \in C} \text{coh}(c_v)$ 
     $c_{j+1} \cup c_{j+2} = \text{split}(c_u)$ 
     $C = C \setminus \{c_u\} \cup \{c_{j+1}, c_{j+2}\}$ 
     $j = j + 2$ 

```

- analiza (ca algoritmi *per se*): ambii algoritmi sunt iterativi și “greedy”;
- complexitate: $\mathcal{O}(n^2 \log n)$ (cf. CMU, 2021 fall, Aarti Singh, Lecture Notes 21);
- exemple de aplicare: ex. 1-5, ex. 27-33 pentru bottom-up, respectiv ex. 6 pentru top-down;
- o legătură cu alți algoritmi de învățare automată: ex. 37;
- implementări: ex. 35, 38, 39.

7.1.3 Proprietăți

- (P0) rezultatul aplicării algoritmilor de clusterizare ierarhică depinde de ce măsură de distanță și ce funcție de similaritate sunt folosite;
- (P1) rezultatul aplicării celor doi algoritmi de clusterizare ierarhică pe un set de date oarecare nu este în mod neapărat unic determinat: ex. 3.b;
- (P2) clusterizarea folosind similaritate de tip “single-linkage” are tendința să creeze clustere alungite; invers, folosind similaritate “complete-linkage” sau “average-linkage”, se formează clustere de formă mai degrabă sferică: ex. 2, 5 și 31;

- (P3) dacă atunci când folosim “single-linkage” și “complete-linkage” se obțin dendrograme identice, nu rezultă în mod neapărat că folosind “average-linkage” vom obține aceeași dendrogramă: ex. 3.b;
- (P4) numărul maxim de niveluri dintr-o dendrogramă (văzută ca arbore în sensul teoriei grafurilor) este $n - 1$, unde n este numărul de instanțe de clusterizat: ex. 4.a; numărul minim de niveluri: $\lceil \log_2 n \rceil$; ex. 4.b;
- (P5) există o anumită corespondență între clusterizare ierarhică cu similaritate de tip
 - “single-linkage” și aflarea *arborelui [de acoperire] de cost minim* dintr-un graf: ex. 6;
 - “complete-linkage” și aflarea unei *clici* (subgraf maximal complet) dintr-un graf (vedeți Manning & Schütze, *op. cit.*, pag. 506-507);
- (P6) algoritmul de clusterizare aglomerativă la al cărui pseudo-cod am făcut referire mai sus are complexitate $\mathcal{O}(n^3)$: ex. 27; atunci când se folosește single-linkage sau complete-linkage, există însă versiuni / algoritmi de complexitate $\mathcal{O}(n^2)$: SLINK (1973) și respectiv CLINK (1976);
- la clusterizare ierarhică aglomerativă cu similaritate “average-linkage”:
 - (P7) dacă se folosește ca măsură de similaritate între 2 instanțe cosinusul unghiului dintre vectorii care reprezintă instanțele și se „normalizează“ acești vectori (i.e., se lucrează cu 2 vectori coliniari cu ei, dar de normă egală cu 1), atunci calculul coeziunii [interne a] unui cluster nou format, precum și calculul „distanței“ dintre două clustere se pot face în timp constant: ex. 36.

7.2. Clusterizare partitională

7.2.1 Notiuni specifice

- centroid (centru de greutate) al unui cluster: ex. 33,
- K*-partiție, *K*-configurație [initială] a centroizilor: ex. 11;
- o funcție de evaluare a „calității“ clusterelor (sau: funcție de „coezione“ / „distorsiune“ / „eroare“ totală):

„suma celor mai mici pătrate“: $J_K(C, \mu) = \sum \|x_i - \mu_{C(x_i)}\|^2$, unde C este *K*-partiție, μ este *K*-configurație de centroizi, iar $\mu_{C(x_i)}$ este centroidul cel mai apropiat de x_i : ex. 12.

7.2.2 Algoritmul *K*-means

- pseudo-cod (cf. Manning & Schütze, *op. cit.*, pag. 516):

Given: a set $X = \{x_1, \dots, x_n\} \subseteq \mathbb{R}^m$,
 a distance measure d on \mathbb{R}^m ,
 a function for computing the mean $\mu : \mathcal{P}(\mathbb{R}^m) \rightarrow \mathbb{R}^m$

Select (arbitrarily) k initial centers f_1, \dots, f_k in \mathbb{R}^m ;
 while the *stopping criterion* is not satisfied
 for all clusters c_j do $c_j = \{x_i \mid \forall f_l \quad d(x_i, f_l) \leq d(x_i, f_j)\}$ end
 for all means f_j do $f_j \leftarrow \mu(c_j)$ end

alternativ, vedeți enunțul ex. 12 (sau, echivalent, folosind variabile-indicator: ex. 45);

- complexitate: $\mathcal{O}(lKn)$, unde l este numărul de iterații executate (cf. CMU, 2021 fall, Aarti Singh, Lecture Notes 21);
- exemple de aplicare: ex. 7-11, ex. 17.a, ex. 21.a, ex. 22.a, ex. 40, ex. 41.
- exemple de *euristici pentru inițializarea centroizilor*: inițializare arbitrară / random în \mathbb{R}^d sau în $X = \{x_1, x_2, \dots, x_n\} \subseteq \mathbb{R}^d$ (setul de date de clusterizat); aplicare (în prealabil) a unui algoritm de clusterizare ierarhică; folosind o anumită distribuție probabilistă definită pe X : *K-means++* (David Arthur, Sergei Vassilvitskii, 2007): ex. 50.
- exemple de *criterii de oprire*: după efectuarea unui număr maxim de iterații (fixat inițial); când componența clusterelor nu se mai modifică de la o iterație la alta; când pozițiile centroizilor nu se mai modifică de la o iterație la alta; când descreșterea valorii criteriului J_K de la o iterație la alta nu mai este strictă sau nu mai este peste un anumit prag ε fixat în prealabil.
- ca algoritm *per se*:

K-means este un algoritm de *căutare*:

spațiul de căutare este mulțimea tuturor K -partițiilor care se pot forma pe dataset-ul de intrare;

(P0) întrucât acest spațiu de căutare (deși este finit) este exponențial (K^n), *K-means* explorează doar parțial spațiul de căutare, procedând *iterativ*: el pleacă de la o „soluție“ (K -partiție) aleasă eventual în mod arbitrar / aleatoriu și o „îmbunătățește“ la fiecare iterație;

(P1) soluția găsită este dependentă de inițializarea centroizilor: ex. 10;

(P1') mai mult, chiar la o aceeași inițializare, rezultatele pot diferi(!) dacă avem instanțe multiple / redundante, situate la egală distanță de 2 centroizi la o iterație oarecare: ex. 11.b (cazul 2), ex. 12.b;

(P1'') rezultatele lui *K-means* sunt dependente [și] de măsura de distanță folosită: ex. 49;

K-means poate fi văzut și ca *algoritm de optimizare* — vedeti criteriul J_K de mai sus;

(P2) strategia de căutare / optimizare folosită de *K-means* este de tipul *descreștere pe coordonate* (engl., coordinate descent), i.e. descreștere iterativă, mergând alternativ pe fiecare din cele două coordonate ale criteriului $J_K(C^t, \mu^t)$: ex. 12.a;

(P2') algoritmul *K-means* nu garantează atingerea optimului global (i.e., minimul) pentru criteriul J_K : ex. 12.b, ex. 47.b.
- ca algoritm de *învățare automată*:

[urmat de] „generalizare“: o instanță nouă x se asociază clusterului având centroidul cel mai apropiat de x ;

(P3) „granițele“ de separare dintre [perechile de] clustere produse de *K-means* sunt liniare atunci când se folosește distanța euclidiană: ex. 11.b (cazul 1);

(P3') este însă posibil să se obțină separatori neliniari dacă se folosește o versiune „kernelizată“ a algoritmului *K-means*: ex. 51;

(P4) rezultatele lui *K-means* pot fi influențate de prezența outlier-elor: ex. 10; în astfel de cazuri este de preferat să se folosească distanța Manhattan în locul distanței euclidiene: ex. 49.B.

- o euristică pentru alegerea unei valori convenabile / „naturale” pentru K (pentru un dataset dat) — criteriul “elbow”: ex. 44 (și CMU, 2012f, E. Xing, A. Singh, HW3, ex. 1.de);
- adaptarea algoritmului K -means pentru cazul în care în locul distanței euclidiene se folosește distanța Manhattan: ex. 49.B;
- recapitulare: ex. 54;
- implementare: ex. 55.

7.2.3 Alte proprietăți ale algoritmului K -means

- în legătură cu criteriul definit mai sus, $J_K : \mathcal{P}_K \times (\mathbb{R}^d)^K \rightarrow [0, +\infty)$, unde \mathcal{P}_K este mulțimea tuturor K -partițiilor peste mulțimea de instanțe, $X = \{x_1, x_2, \dots, x_n\} \subseteq \mathbb{R}^d$:
 - (P5) $J_K(C^{t-1}, \mu^{t-1}) \geq J_K(C^t, \mu^t)$ la orice iterație ($t > 0$) a algoritmului K -means: ex. 12.a (exemplificare: ex. 7.a, ex. 42);
 - (P5') în consecință, dacă se impune restricția ca la fiecare iterare inegalitatea de mai sus să fie satisfăcută în varianta strictă ($J_K(C^{t-1}, \mu^{t-1}) > J_K(C^t, \mu^t)$), atunci algoritmul K -means termină într-un număr finit de pași: ex. 12.b.
- (P6) pentru $K > 0$ fixat, $|\mathcal{P}_K| = K^n$, deci este finit, și există $\underline{J}_K \stackrel{\text{not.}}{=} \min_C J_K(C, \mu_C)$; acest minim (\underline{J}_K) se poate obține prin explorarea exhaustivă a spațiului \mathcal{P}_K , însă consumul de timp este prohibitiv în practică: ex. 12.b;
- (P7) valoarea 0 pentru \underline{J} este atinsă, și anume atunci când $K = n$, C este K -partiția de clustere singleton $C_i = \{x_i\}$, iar $\mu_i = x_i$, pentru $i = 1, \dots, n$ (ex. 44);
- (P8) $\underline{J}_1 \geq \underline{J}_2 \geq \dots \geq \underline{J}_{n-1} \geq \underline{J}_n = 0$: ex. 13.
- (P9) dacă $d = 1$, deci $x_1, x_2, \dots, x_n \in \mathbb{R}$,
 - orice K -partiție (C_1, \dots, C_K) pentru care se atinge \underline{J}_K este de forma unei colecții de „intervale”: $C_1 = \{x_1, \dots, x_{i_1}\}$, $C_2 = \{x_{i_1+1}, \dots, x_{i_2}\}$, \dots , $C_K = \{x_{i_{K-1}+1}, \dots, x_n\}$, cu $i_1 < i_2 < \dots < i_{K-1} < i_K = n$;
 - există un algoritm [de programare dinamică] de complexitate $\mathcal{O}(Kn^2)$ care calculează \underline{J}_K : ex. 47.
- (P10) în vreme ce maximizează *coezionea intra-clustere*, i.e. minimizează criteriul J , algoritmul K -means maximizează (în mod aproximativ!) o sumă ponderată a pătratelor distanțelor dintre centroizii clusterelor și centrul de greutate al întregului set de instanțe:

$$\sum_{k=1}^K \left(\frac{\sum_{i=1}^n \gamma_{ik}}{n} \right) \|\mu_k - \bar{x}\|^2,$$

unde \bar{x} este media instanțelor x_1, x_2, \dots, x_n (ex. 45, ex. 46).

7.3. Clusterizare prin modelare probabilistă

7.3.1 Noțiuni preliminare

- variabile aleatoare (discrete, resp. continue);
media, varianța și covarianța variabilelor aleatoare;

- vector de variabile aleatoare; matrice de covarianță pentru un astfel de vector; proprietăți: matricea de covarianță trebuie să fie în mod necesar simetrică și pozitiv definită: ex. 20 de la capitolul de *Fundamente*;
- distribuție (funcție de densitate) de probabilitate (p.d.f.); parametri ai unei distribuții de probabilitate; distribuția gaussiană, cazurile uni- și multidimensional: vedeti secțiunea corespunzătoare din *Sumarul* capitolului de *Fundamente*;
- mixtură de distribuții probabiliste: văzută ca o formă particulară de *combinăție liniară* de distribuții de probabilitate $\pi_1\Psi_1 + \pi_2\Psi_2 + \dots + \pi_k\Psi_k$ (cu $\pi_i \geq 0$ și $\sum_{i=1}^k \pi_i = 1$), definită [și mai specific] scriind distribuția $P(X)$ ca o sumă ponderată de probabilități condiționate: $\sum_z P(X|Z)P(Z)$, unde X sunt variabilele „observabile“, iar variabila Z (eventual multiplă) poate fi „neobservabilă“ / „latentă“ / „ascunsă“; exemple:
 - o mixtură de distribuții categoriale, respectiv o mixtură de distribuții Bernoulli: ex. 29 și ex. 114 și ex. 113 de la capitolul de *Fundamente*;
 - o mixtură de distribuții gaussiene multidimensionale: ex. 118 de la capitolul de *Fundamente*;
 - o mixtură de distribuții oarecare: ex. 119 de la capitolul de *Fundamente*;
- funcție de *verosimilitate* a unui set de date (D), în raport cu o distribuție probabilistă dată: $L(\theta) = P(D|\theta)$, unde prin θ se notează parametrii respectivei distribuții. Exemplificare: ex. 43.abd, ex. 42 de la capitolul de *Fundamente*;
- MLE (Maximum Likelihood Estimation): estimarea [valorilor] parametrilor unei distribuții probabiliste în sensul maximizării verosimilității datelor disponibile. Exemplificare: capitolul de *Fundamente*, ex. 43-54, ex. 124-135. Aplicare în cazul distribuției gaussiene unidimensionale: ex. 15.ab de la capitolul *Clasificare bayesiană*;
- regula de decizie pentru algoritm Bayes [Naiv] Gaussian:⁸⁰⁴ pentru cazul unidimensional: ex. 15, ex. 44, ex. 16, și ex. 45 de la capitolul *Clasificare bayesiană*;
- Observație: Algoritmul EM este [sau, mai degrabă, poate fi folosit ca] o metodă de estimare a parametrilor unei mixturi de distribuții probabiliste. *Alternativ*, pentru același obiectiv pot fi folosite alte metode, de exemplu *metoda gradientului ascendent*: ex. 69.

7.3.2 Algoritmul EM pentru clusterizare prin estimarea parametrilor unui model de mixturi de distribuții gaussiene (EM/GMM)

- pseudo-cod:⁸⁰⁵ cazul unidimensional, varianta când doar parametrul μ este lăsat liber: ex. 15 (cf. *Machine Learning*, Tom Mitchell, 1997, pag. 193); aplicare: ex. 16, ex. 17.b, ex. 56, ex. 57;

⁸⁰⁴În cazul separării linare, în literatura de specialitate se folosește termenul de *analiză discriminativă gaussiană*.

⁸⁰⁵Nu doar pentru pseudo-cod, ci și (sau, mai ales) pentru o privire de ansamblu unitară, atât pentru cazul unidimensional cât și pentru cazul multidimensional (ambele urmând a fi sistematizate mai jos), puteți consulta documentul *An Introduction to Expectation-Maximization*, de Dahua Lin, MIT, ML course 6768, 2012 fall.

- cazul unidimensional, varianta când toți parametrii (π , μ și σ) sunt lăsați liberi: ex. 18; aplicare: ex. 17.c, ex. 58, ex. 59;
- cazul unidimensional, alte variante: [ex. 19] ex. 60, ex. 61, ex. 62;
- cazul multidimensional, varianta când toți parametrii (π , μ și Σ) sunt lăsați liberi: ex. 24;
- cazul multidimensional, alte variante: ex. 20, ex. 63;
- aplicarea algoritmului EM/GMM, cazul bidimensional: ex. 21.b, ex. 22.b, ex. 23, ex. 25, ex. 65, ex. 66, ex. 67;
- schema algoritmica EM: vedeti Tom Mitchell, *Machine Learning book*, 1997, pag. 194-195; ex. 19;
 - pentru *proprietăți generale* ale schemei algoritmice EM, vedeti secțiunea corespunzătoare din *Sumarul capitolului Schema algoritmica EM*;
 - câteva *proprietăți* ale algoritmului EM/GMM:
 - (P0) rezultatele algoritmului EM/GMM depind (ca și la K-means) de valorile atribuite parametrilor la inițializare (ex. 17.c);
 - (P1) anumite valori atribuite inițial parametrilor algoritmului EM/GMM pot provoca rularea la infinit a algoritmului, fără ca [la pasul M] valorile parametrilor să se modifice de la o iterare la alta: ex. 19.c;
 - (P2) spre deosebire de cazul algoritmului K-means, suprafețele / granițele de separare create de algoritmul EM/GMM nu sunt în mod neapărat liniare. (Vedeți de exemplu situațiile întâlnite la rezolvarea ex. 17.c, pag. 877, sau la ex. 67.c și ex. 68.c);
 - comparativ cu algoritmul K-means,
 - (P3) algoritmul EM/GMM este în general mai lent — mișcarea centroizilor poate explora într-o manieră mai fină spațiul (vedeți de exemplu ex. 21) —, dar din acest motiv el poate să obțină uneori rezultate mai bune / convenabile. (Vedeți spre exemplu ex. 22);
 - (P4) apare un fenomen de “atracție” reciprocă a mediilor gaussianelor (aceste medii fiind echivalentul centroizilor din algoritm K-means), datorită faptului că fiecare instanță aparține (cu o anumită probabilitate) la fiecare cluster. Atracția mediilor este cu atât mai puternică cu cât varianțele sunt mai mari. (Vedeți spre exemplu ex. 17.b);
 - (P5) EM/GMM este mai robust la influența outlier-elor.
 - (P6) atunci când $\Sigma = \sigma^2 I$, iar $\sigma^2 \rightarrow 0$ (și sunt satisfăcute încă două restricții), algoritmul EM/GMM倾de să se comporte asemenea algoritmului K-means:⁸⁰⁶ ex. 64;
 - o legătură interesantă între algoritmul EM/GMM și metoda gradientului ascendent, în cazul în care matricele de covarianță sunt de forma $\sigma_k^2 I$: ex. 69;
 - algoritmul EM/GMM *semi-supervizat*: ex. 70;
 - o variantă a algoritmului EM/GMM semi-supervizată pentru cazul când matricele de covarianță sunt de forma $\sigma_k^2 I$ și este satisfăcută presupozitia de independentă condițională de tip Bayes Naiv: ex. 71 (are loc o legătură interesantă cu clasificatorul Bayes Naiv gaussian).

⁸⁰⁶Formularea se referă la cazul multidimensional, dar proprietatea este valabilă și în cazul unidimensional.

7.1 Clusterizare — Probleme rezolvate

7.1.1 Clusterizare ierarhică

1. (Clusterizare ierarhică aglomerativă: exemplificare pe date din \mathbb{R} , folosind similaritate de tip “single-linkage”)

CMU, 2006 spring, Carlos Guestrin, HW5, pr. 1

- a. Desenați *dendrograma* (adică arborele de clusterizare ierarhică) pentru următoarea mulțime de 10 puncte pe axa reală:

$$S = \{-2.2, -2.0, -0.3, 0.1, 0.2, 0.4, 1.6, 1.7, 1.9, 2.0\}$$

Folosiți *similaritate* de tip “single-linkage”, adică $d(C_i, C_j) = \min_{x \in C_i, x' \in C_j} |x - x'|$.

Observație: În dendrogramă, înălțimea (h) corespunzătoare nodului rădăcină al unui cluster va fi considerată direct proporțională cu media aritmetică a distanțelor dintre punctele din clusterul respectiv,⁸⁰⁷ iar *coezionea* (c) clusterului va fi definită aici ca inversul acestei medii.⁸⁰⁸

- b. Bazat pe arborele obținut, justificați — în manieră informală — faptul că 3 este numărul natural de clustere care se formează în acest set de date.

Răspuns:

- a. Numerotăm cele 10 puncte date în ordinea crescătoare a valorilor lor: $x_1 = -2.2, x_2 = -2, x_3 = -0.3, x_4 = 0.1, x_5 = 0.2, x_6 = 0.4, x_7 = 1.6, x_8 = 1.7, x_9 = 1.9, x_{10} = 2.0$.

Dendrograma mulțimii de instanțe date se construiește în manieră *bottom-up* astfel:

- $d(x_4, x_5) = d(x_7, x_8) = d(x_9, x_{10}) = 0.1 = \min_{1 \leq i < j \leq 10} d(x_j, x_i)$, prin urmare primele 3 clustere care se vor forma sunt $C_1 = \{x_4, x_5\}, C_2 = \{x_7, x_8\}, C_3 = \{x_9, x_{10}\}$. În mod evident, $h(C_1) = h(C_2) = h(C_3) = 0.1$ iar $c(C_1) = c(C_2) = c(C_3) = 1/0.1 = 10$.

- Avem:

$$\begin{aligned} d(x_1, x_2) &= 0.2 \\ d(C_1, x_6) &= d(x_5, x_6) = 0.2 \\ d(C_2, C_3) &= d(x_8, x_9) = 0.2 \end{aligned}$$

Celelalte distanțe sunt toate mai mari decât 0.2, aşadar următorul nivel al arborelui va fi alcătuit din clusterele $C_4 = \{x_1, x_2\}, C_5 = \{x_4, x_5, x_6\}$ și $C_6 = \{x_7, x_8, x_9, x_{10}\}$. Făcând calculele, obținem: $h(C_4) = h(C_5) = 0.2$ și $h(C_6) = \frac{1.4}{6} = 0.2(3)$, iar $c(C_4) = c(C_5) = 1/0.2 = 5$ și $c(C_6) = \frac{6}{1.4} \approx 4.285$.

⁸⁰⁷În rezolvarea acestei probleme, factorul de proporționalitate va fi considerat 1.

⁸⁰⁸Evident, această definiție pentru coeziune are sens doar în cazul clusterelor non-singleton.

• Avem:

$$\begin{aligned} d(C_4, x_3) &= d(x_2, x_3) = 1.7 & d(C_4, C_5) &= d(x_2, x_4) = 2.1 \\ d(C_5, x_3) &= d(x_4, x_3) = 0.4 & d(C_4, C_6) &= d(x_2, x_7) = 3.6 \\ d(C_6, x_3) &= d(x_7, x_3) = 1.9 & d(C_5, C_6) &= d(x_6, x_7) = 1.2 \end{aligned}$$

Așadar, pe următorul nivel al ierarhiei se va afla clusterul $C_7 = C_5 \cup \{x_3\} = \bigcup_{3 \leq i \leq 6} \{x_i\}$. Înălțimea și respectiv coeziunea acestui cluster sunt: $h(C_7) = \frac{2.2}{6} = 0.3(6)$ și $c(C_7) = \frac{6}{2.2} = 2.(72)$.

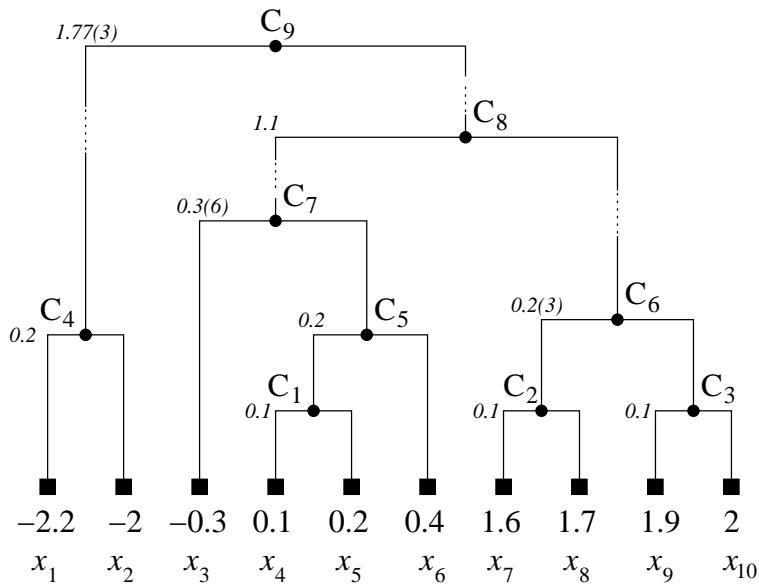
• Avem:

$$\begin{aligned} d(C_4, C_6) &= d(x_2, x_7) = 3.6 \\ d(C_4, C_7) &= d(x_2, x_3) = 1.7 \\ d(C_6, C_7) &= d(x_6, x_7) = 1.2 \end{aligned}$$

În consecință, pe următorul nivel al ierarhiei se va afla clusterul $C_8 = C_6 \cup C_7 = \bigcup_{3 \leq i \leq 10} \{x_i\}$. Făcând calculele, obținem $h(C_8) = \frac{30.8}{28} = 1.1$ și $c(C_8) = 1/1.1 = 0.9(09)$.

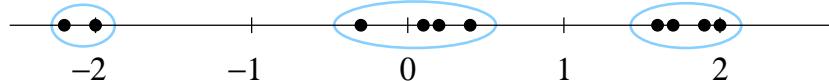
• Pentru clusterul $C_9 = C_4 \cup C_8$, avem $h(C_9) = \frac{79.8}{45} = 1.77(3)$ și $c(C_9) \approx 0.564$.

Dendrograma multșimii date va arăta deci astfel:



b. După alcătuirea dendrogramei, împărțirea mulșimii de instanțe date în clustere cu coeziune comparabilă se face „tăind“ dendrograma cu o linie paralelă cu baza. Împărțirea în clustere fine corespunde liniilor paralele apropiate de bază, iar împărțirea în clustere ample corespunde liniilor paralele apropiate de vîrful / rădăcina dendrogramei.

În cazul nostru, se observă că pentru o mare parte a înălțimii dendrogramei (și anume între $0.3(6)$ și 1.1), rezultatul operației descrise mai sus rămâne ne-schimbăt (mulșimea dată descompunându-se corespunzător în clusterele C_4 , C_7 și C_6), ceea ce sugerează că „numărul natural“ de clustere este 3. Acest lucru se poate vedea și dacă reprezentăm cele 10 puncte pe axa reală:



2. (Clusterizare ierarhică aglomerativă: exemplificare pe date din \mathbb{R}^2 , folosind tipurile de similaritate “single-linkage” și “complete-linkage”)

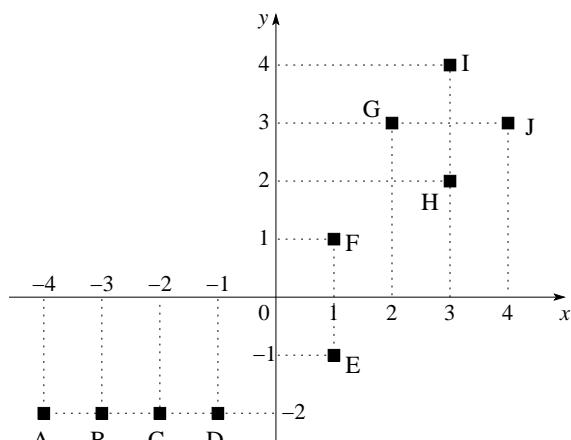
Edinburgh, 2009 fall, C. Williams, V. Lavrenko, HW4, pr. 1

Considerăm următorul set de date, în care fiecare instanță (x, y) este un punct în plan. Pentru conveniență, în cele ce urmează, vom identifica aceste instanțe prin literele asociate punctelor respective.

$$\begin{aligned} A &: (-4, -2), B : (-3, -2), C : (-2, -2), D : (-1, -2), E : (+1, -1) \\ F &: (+1, +1), G : (+2, +3), H : (+3, +2), I : (+3, +4), J : (+4, +3) \end{aligned}$$

- a. Reprezentați grafic datele din enunț.
 - b. Realizați clusterizarea ierarhică a datelor în maniera bottom-up, folosind similaritate de tip “single-linkage” și distanța euclidiană între puncte.
- Observație:* Dacă la o iterație a algoritmului de clusterizare distanțele (adică similaritățile) dintre două perechi de clustere au aceeași valoare, prioritatea la alcătuirea noului cluster este dictată de ordinea alfabetică.
- c. Realizați clusterizarea datelor, folosind de această dată similaritate de tip “complete-linkage”.
 - d. Discutați diferența dintre clusterizările obținute în urma folosirii celor două tipuri de [măsuri de] similaritate.

Răspuns:

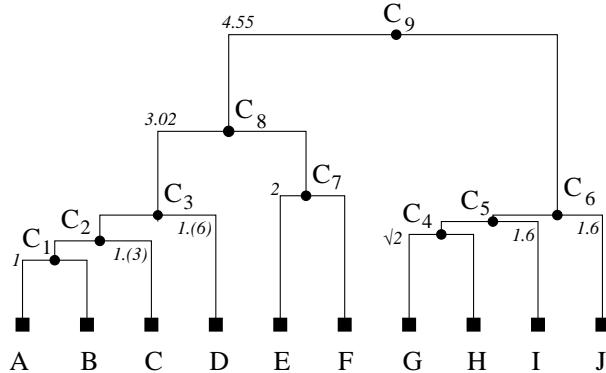


- a. Reprezentarea în plan a celor 8 puncte date în enunț este cea din figura alăturată.

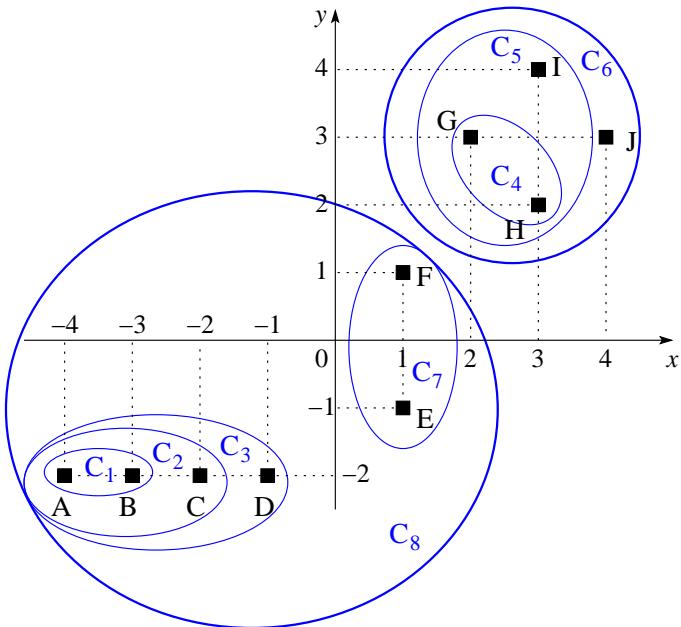
- b. Clusterizarea folosind similaritate de tip “single-linkage” (pentru care distanța dintre două clustere este definită ca fiind distanța dintre cele mai apropiate două puncte, câte unul din fiecare cluster) se realizează astfel:

- Minimul distanțelor mutuale dintre punctele date este 1. Acest minim este obținut pentru mai multe perechi de puncte, deci se va recurge la ordinea alfabetică pentru a stabili prioritatea. Prin urmare, primul cluster format va fi $C_1 = \{A, B\}$. Evident, $h(C_1) = c(C_1) = 1$, unde h și c sunt înălțimea asociată clusterului C_1 , respectiv coeziunea lui, calculate după aceeași metodă ca la problema 1.
- Acum, distanța minimă este obținută (înținând cont și de regula alfabetică) între clusterul C_1 și punctul C , deci $C_2 = C_1 \cup \{C\} = \{A, B, C\}$. Prin urmare, $h(C_2) = 4/3 = 1.(3)$, iar $c(C_2) = 3/4 = 0.75$.
- În mod similar, tot la distanța 1, vom avea $C_3 = C_2 \cup \{D\} = \{A, B, C, D\}$, iar $h(C_3) = 10/6 = 1.(6)$ și $c(C_3) = 6/10 = 0.6$.
- La acest nou pas se consideră clusterul C_3 și punctele E, F, \dots, J . Distanța minimă este $\sqrt{2}$, și anume — având din nou în vedere ordinea alfabetică — între punctele G și H , deci $C_4 = \{G, H\}$, cu $h(C_4) = \sqrt{2} \approx 1.414$ și $c(C_4) = 1/\sqrt{2} \approx 0.707$.
- Apoi $C_5 = C_4 \cup \{I\} = \{G, H, I\}$ și $C_6 = C_5 \cup \{J\} = \{G, H, I, J\}$, clustere care au înălțimile (egale!) $h(C_5) = h(C_6) = \frac{2}{3}(\sqrt{2} + 1) \approx 1.609$ și, deci, și coeziunile egale: $c(C_5) = c(C_6) = \frac{3}{2}\frac{1}{\sqrt{2} + 1} \approx 0.621$.
- Acum se consideră clusterele C_3 și C_6 și punctele E și F . Distanța minimă este 2, și anume cea dintre punctele E și F , deci $C_7 = \{E, F\}$, cu $h(C_7) = 2$ și $c(C_7) = 0.5$.
- Între clusterele C_3, C_6 și C_7 , distanța minimă este $\sqrt{5}$, deci $C_8 = C_3 \cup C_7 = \{A, B, C, D, E, F\}$, cu $h(C_8) \approx 3.02$ și $c(C_8) \approx 0.331$.
- Rămâne în final $C_9 = C_8 \cup C_6$, cu $h(C_9) \approx 4.552$ și $c(C_9) \approx 0.219$.

Dendrograma obținută este redată în figura alăturată.



Varianta „aplatizată“ a acestei ierarhii (engl., flat hierarchy) este prezentată în desenul alăturat.



c. Clusterizarea ierarhică de tip “complete-linkage” (pentru care similaritatea dintre două clustere este dată de distanța dintre cele mai depărtate două puncte, câte unul din fiecare cluster) se realizează astfel:

- Minimul distanțelor dintre oricare două puncte este 1, iar acest minim este obținut pentru mai multe perechi de puncte. Luând în considerare ordinea alfabetică, primul cluster format va fi $C_1 = \{A, B\}$, cu $h(C_1) = c(C_1) = 1$.
- La acest nou pas, distanța minimă este $d(C, D) = 1$. Deci, spre deosebire de cazul “single-linkage”, vom avea $C_2 = \{C, D\}$, cu $h(C_2) = c(C_2) = 1$.
- În continuare, cea mai mică distanță este $\sqrt{2}$, între G și H , deci $C_3 = \{G, H\}$, cu $h(C_3) = \sqrt{2} \approx 1.414$ și $c(C_3) = 1/\sqrt{2} \approx 0.707$.
- Tot $\sqrt{2}$ este și distanța dintre I și J , deci $C_4 = \{I, J\}$, cu $h(C_4) = \sqrt{2}$ și $c(C_4) = 1/\sqrt{2}$.
- Apoi, $C_5 = \{E, F\}$, între aceste două puncte distanța fiind 2, deci $h(C_5) = 2$ și $c(C_5) = 0.5$.
- În acest moment se consideră clusterele C_1, C_2, C_3, C_4 și C_5 . Distanțele dintre ele sunt:

$$\begin{array}{ll}
 d(C_1, C_2) = d(A, D) = 3 & d(C_2, C_4) = d(C, I) = d(C, J) = 7.81 \\
 d(C_1, C_3) = d(A, H) = 8.06 & d(C_2, C_5) = d(C, F) = 4.24 \\
 d(C_1, C_4) = d(A, J) = 9.43 & d(C_3, C_4) = d(G, J) = d(H, I) = 2 \\
 d(C_1, C_5) = d(A, F) = 5.83 & d(C_3, C_5) = d(E, G) = 4.12 \\
 d(C_2, C_3) = d(C, G) = d(C, H) = 6.4 & d(C_4, C_5) = d(E, I) = 5.39
 \end{array}$$

Luând minimul acestor distanțe, și anume $d(C_3, C_4) = 2$, se obține $C_6 = C_3 \cup C_4 = \{G, H, I, J\}$, cu $h(C_6) = \frac{2}{3}(\sqrt{2} + 1) \approx 1.609$ și $c(C_6) \approx 0.621$.

- Urmează $C_7 = C_1 \cup C_2 = \{A, B, C, D\}$, cu $h(C_7) = 10/6 = 1.(6)$ și $c(C_7) = 0.6$.

- Între C_5, C_6 și C_7 , distanțele sunt:

$$d(C_5, C_6) = d(E, I) = 5.39$$

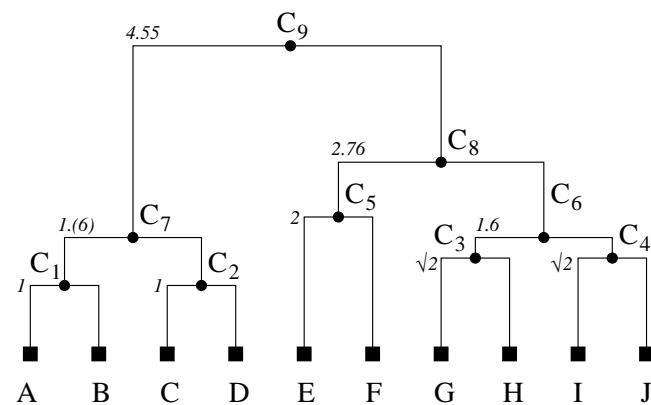
$$d(C_5, C_7) = d(A, F) = 5.83$$

$$d(C_6, C_7) = d(A, J) = 9.43$$

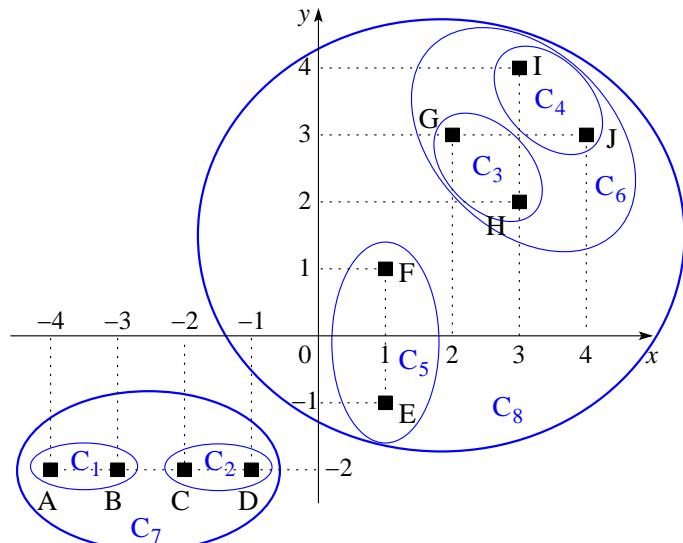
Întrucât cea mai mică dintre aceste distanțe este 5.39, rezultă $C_8 = C_5 \cup C_6 = \{E, F, G, H, I, J\}$, cu $h(C_8) \approx 2.76$ și $c(C_8) \approx 0.36$.

- Rămâne în final $C_9 = C_7 \cup C_8$, distanța dintre cele două clustere constitutive fiind de 9.43. Înălțimea și coeziunea clusterului C_9 sunt respectiv 4.55 și 0.21.

Dendrogramea rezultată este cea din figura alăturată.



Varianta „aplatizată“ a acestei ierarhii (engl., flat hierarchy) este prezentată în desenul alăturat.



d. Clusterizarea ierarhică folosind similaritate de tip “single-linkage” are tendința de a înlăntui / alungi clusterele, spre deosebire de cea “complete-linkage” care grupează clusterele mai degrabă sub formă sferică.

3. (Clusterizare ierarhică aglomerativă: exemplificare în \mathbb{R} , folosind tipurile de similaritate “single-linkage” și “complete-linkage”, comparativ cu “average-linkage”)

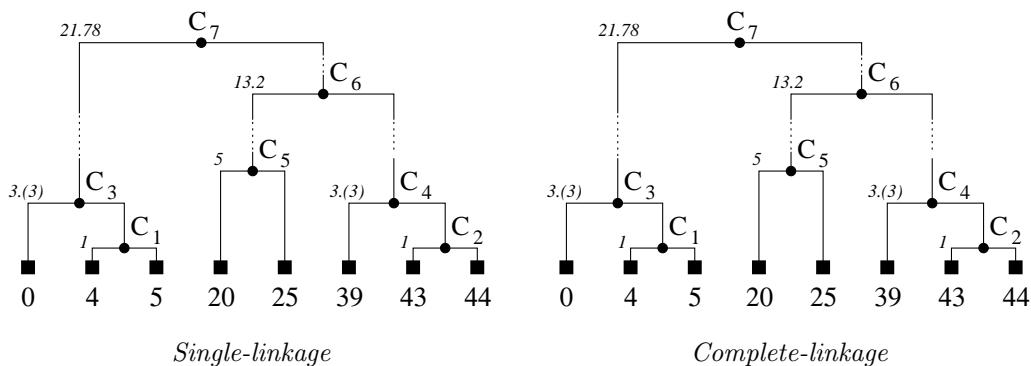
prelucrare făcută de Liviu Ciortuz, după CMU, 2009 spring, Ziv Bar-Joseph, final exam, pr. 9.1

Considerăm următorul set de date: $\{0, 4, 5, 20, 25, 39, 43, 44\}$. Presupunem că vrem să obținem cele mai importante 2 clustere din dendrogramă (cele situate imediat sub nodul rădăcină), numite în cele ce urmează *top-clustere*.

- Care sunt cele două top-clustere atunci când se folosește similaritate *single-linkage* și respectiv *complete-linkage*?
- Dacă *single-linkage* și *complete-linkage* produc rezultate identice (relativ la componența [top-]clusterelor), rezultă că funcția de similaritate *average-linkage* va conduce și ea la același rezultat? Pentru a elabora răspunsul dumneavoastră, veДЕti ce se întâmplă lucrând pe datele de mai sus.

Răspuns:

- Dendrogramele corespunzătoare similarităților *single-linkage* și *complete-linkage* sunt următoarele:⁸⁰⁹



Se observă că atât în cazul *single-linkage* cât și în cazul *complete-linkage*, cele două top-clustere sunt $\{0, 4, 5\}$ și $\{20, 25, 39, 43, 44\}$. Mai mult, cele două dendrograme sunt identice.

- Clusterizarea ierarhică cu similaritate *average-linkage* folosește ca „distanță“ între două clustere oarecare media aritmetică a distanțelor calculate pentru toate perechile de puncte ce se pot forma luând un punct dintr-un cluster și un punct din celălalt cluster. Aplicarea acestui algoritm se realizează astfel:

- $C_1 = \{4, 5\}$ și $C_2 = \{43, 44\}$, ambele constituise din câte o pereche de clustere “singleton” aflate la distanță 1. Înălțimile clusterelor rezultate sunt $h(C_1) = h(C_2) = 1$.⁸¹⁰
- $C_3 = \{0\} \cup C_1 = \{0, 4, 5\}$, distanța dintre clusterul “singleton” $\{0\}$ și clusterul C_1 fiind $\frac{4+5}{2} = 4.5$. Rezultă $h(C_3) = \frac{4+5+1}{3} = 3.(3)$.

⁸⁰⁹Calculele sunt absolut similare celor de la rezolvarea problemei 2.

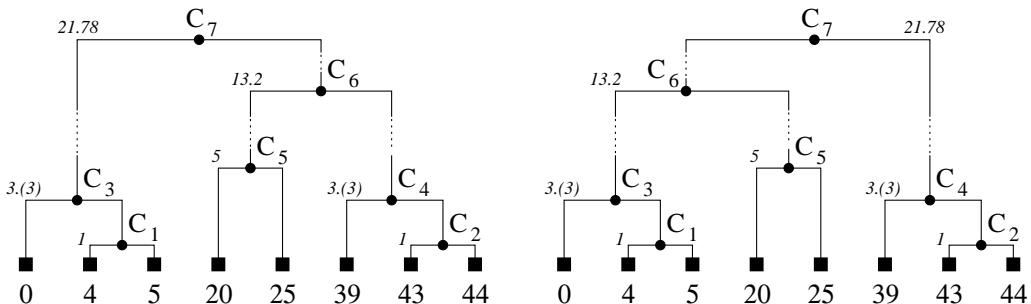
⁸¹⁰Pentru calculul înălțimilor am procedat ca la problemele 1 și 2.

- $C_4 = \{39\} \cup C_2 = \{39, 43, 44\}$. Distanța dintre clusterul $\{39\}$ și clusterul C_2 este $\frac{(43 - 39) + (44 - 39)}{2} = \frac{4 + 5}{2} = 4.5$. Înălțimea noului cluster este $h(C_4) = 3.(3)$.
- $C_5 = \{20, 25\}$, cu distanța dintre clusterele “singleton” constitutive 5 și înălțimea rezultantă tot 5.
- În acest moment există 3 clustere: $C_3 = \{0, 4, 5\}$, $C_4 = \{39, 43, 44\}$ și $C_5 = \{20, 25\}$. Sunt relevante două dintre cele trei distanțe mutuale dintre aceste clustere, și anume:

$$d(C_3, C_5) = \frac{20 + 25 + 16 + 21 + 15 + 20}{6} = \frac{117}{6} = 19.5$$

$$d(C_4, C_5) = \frac{19 + 14 + 23 + 18 + 24 + 19}{6} = \frac{117}{6} = 19.5$$

Se observă că aceste două distanțe sunt egale. Deoarece în enunț nu s-a specificat — pentru un astfel de caz — vreo regulă de prioritate de care să se țină cont atunci când se combină două clustere,⁸¹¹ cele două top-clustere care se pot obține în cazul nostru sunt fie $\{0, 4, 5\}$ și $\{20, 25, 39, 43, 44\}$ — adică exact ca la punctul a —, fie $\{0, 4, 5, 20, 25\}$ și $\{39, 43, 44\}$. Acest fapt este ilustrat în cele două dendrograme de mai jos:



Concluzie: Chiar dacă, lucrând pe un același set de date, la clusterizare ierarhică folosind similaritate *single-linkage* și respectiv *complete-linkage* se obțin rezultate identice, este posibil ca atunci când se folosește *average-linkage* să obținem un alt rezultat.

4.

(Clusterizare ierarhică aglomerativă, folosind similaritate “single-linkage”: exemplificare pe date din \mathbb{R} ; corespondența cu numărul maxim / minim de niveluri din arborele clasic corespunzător dendrogramiei)

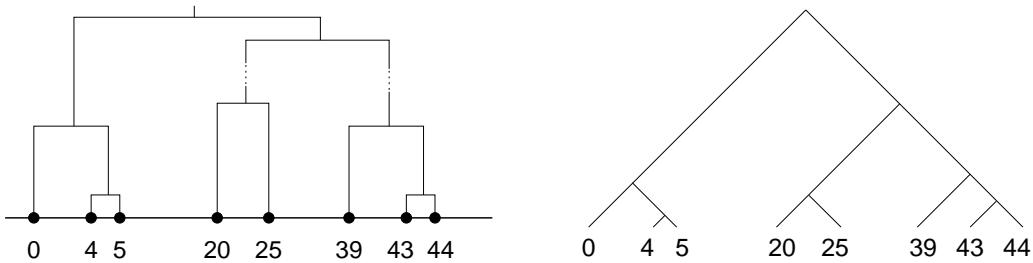
CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW3, pr. 5.2

În acest exercițiu veți folosi metoda clusterizării ierarhice aglomerative cu similaritate de tip “single-linkage” la compararea a două clustere oarecare.⁸¹²

⁸¹¹Vedeți regulile folosite la problemele 1 și 2.

⁸¹²În acest caz, distanța dintre două clustere este definită ca minimul distanțelor $d(x, y)$, unde x aparține primului cluster, iar y aparține celui de-al doilea cluster.

Vom defini *înălțimea* unei ierarhii ca fiind $l - 1$, unde l este numărul de niveluri din arborele (în sens clasic) corespunzător acestei ierarhii. De exemplu, pentru numerele 0, 4, 5, 20, 25, 39, 43, 44, ierarhia “single-linkage” (arătată în desenul de mai jos, partea stângă) are înălțimea 4. Pentru conveniență (și pentru a elimina posibilele confuzii datorate înălțimilor diferitelor noduri din dendrogramă), am desenat și arborele clasic care corespunde acestei ierarhii (vedeți desenul din partea dreaptă); evident, el are 5 niveluri.



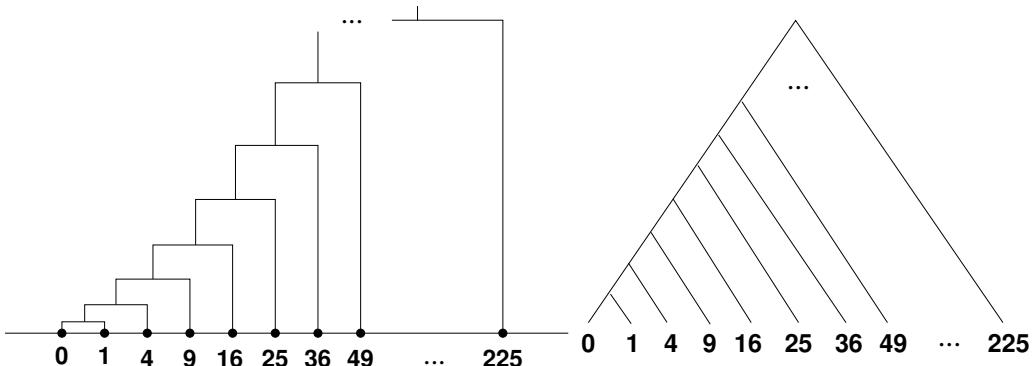
- Care este înălțimea maximă a unei ierarhii care se poate construi cu N puncte? Dar înălțimea minimă?
- Dați un exemplu de 16 puncte din mulțimea numerelor întregi care la acest tip de clusterizare produce (i.) o ierarie de înălțime maximă, (ii.) o ierarie de înălțime minimă.

Răspuns:

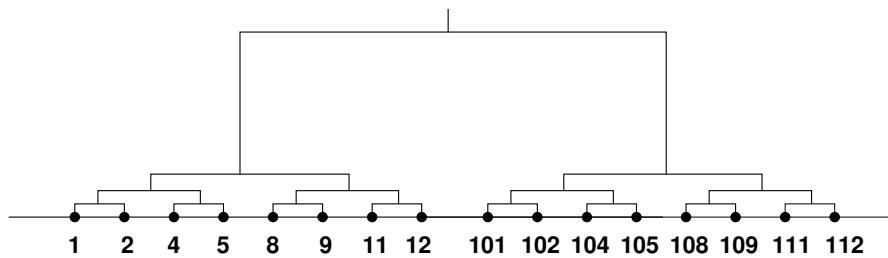
- Pentru a avea o ierarie de înălțime maximă, putem considera spre exemplu cazul în care distanțele dintre punctele care se clusterizează succesiv sunt în ordine strict crescătoare, obținând un arbore cu N niveluri. Deci înălțimea maximă a unei ierarhii este $N - 1$.

Pentru a avea înălțime minimă, trebuie ca arborele corespunzător să fie pe cât posibil un arbore binar echilibrat. Deci înălțimea minimă este partea întreagă superioară din $\log_2(N)$.

- O mulțime de 16 puncte care conduce la o ierarie de tip “single-linkage” de înălțime maximă este: 0, 1, 4, 9, 16, 25, 36, 49, 64, 81, 100, 121, 144, 169, 196, 225. Ierarhia este următoarea (am folosit ambele variante de reprezentare grafică):



Pentru o ierarie de tip “single-linkage” de înălțime minimă putem considera următoarele 16 puncte: 1, 2, 4, 5, 8, 9, 11, 12, 101, 102, 104, 105, 108, 109, 111, 112. Ierarhia este următoarea:



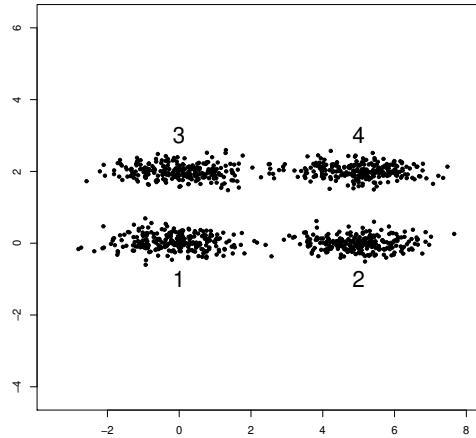
5.

(Clusterizare ierarhică: aplicare în manieră intuitivă pe date din \mathbb{R}^2)

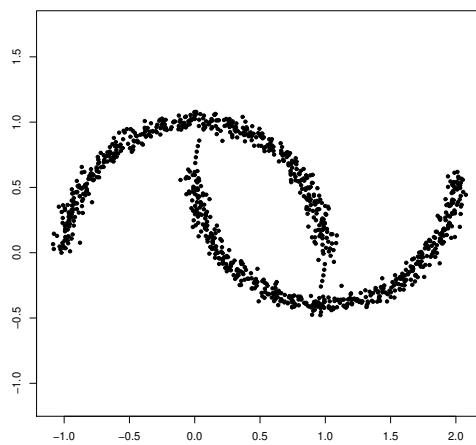
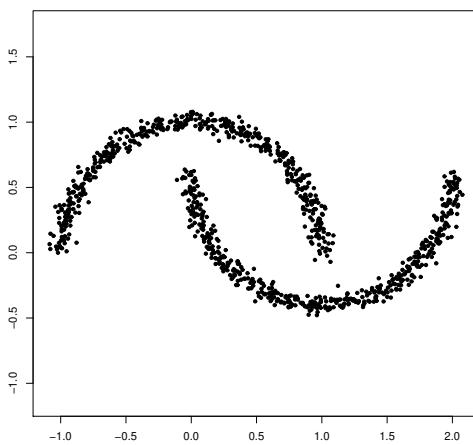
• ○ CMU, 2012 fall, E. Xing, A. Singh, HW3, pr. 1.2.bc

a. Considerăm datele din figura alăturată.

Ce rezultat vom obține dacă extragem cele două clustere de “top” din arborele obținut prin clusterizare ierarhică pe acest set de date folosind măsura de similaritate “single-linkage”? (Formulați răspunsul raportându-vă la etichetele 1–4 care identifică cele patru grupuri din date.) Procedați similar pentru “complete-linkage” și “average-linkage”.



b. Poate vreuna dintre cele trei măsuri de similaritate să separe cu succes cele două „semiluni“ din figura de mai jos, partea stângă? Dar în ce privește figura din partea dreaptă? Justificați pe scurt.



Răspuns:

- a. “Single-linkage” va asigna aglomerările de puncte notate cu 1 și 2 la un cluster, iar pe cele notate cu 3 și 4 la celălalt cluster. “Complete-linkage” și “average-linkage” vor asigna aglomerările de puncte notate cu 1 și 3 la un cluster, iar pe cele notate cu 2 și 4 la alt cluster.
- b. “Single-linkage” va reuși să separe cu succes cele două semiluni din figura din partea stângă (din enunț), în vreme ce “complete-linkage” și “average-linkage” nu vor reuși. Niciuna dintre cele trei metode nu va avea succes pe datele din figura din partea dreaptă.

6. (Un algoritm de clusterizare ierarhică divizivă: aplicare; comparație cu variantele algoritmului de clusterizare aglomerativă; echivalență cu algoritmii de aflare a MST din teoria grafurilor)

*prelucrare făcută de Liviu Ciortuz, după
■ • CMU, 2009 spring, Ziv Bar-Joseph, final exam, pr. 9.3*

Clusterizarea ierarhică poate fi de două tipuri: *divizivă* (top-down) sau *aglomerativă* (bottom-up). Întrebarea la care am dori să răspundem prin această problemă este dacă un algoritm din categoria top-down poate să fie echivalent (relativ la rezultatul obținut ca atare, sau chiar la modul efectiv de elaborare a clusterizării) cu un algoritm din categoria bottom-up.

Să considerăm următorul *algoritm de clusterizare de tip top-down*:

• **Intrare:** o mulțime de instanțe $S = \{x_1, x_2, \dots, x_n\}$ și o măsură de distanță definită pe $S \times S$;

• **Ieșire:** o dendrogramă, adică un arbore de clusterizare binară în ale cărui noduri frunză sunt elementele din S , satisfăcând proprietățile dezirabile pentru clusterizare ierarhică (i.e., coeziune cât mai mare în interiorul fiecărui cluster și distanțe maxime între clustere, pe fiecare nivel din dendrogramă);

• **Procedură:**

Pasul 1: Se construiește graful neorientat, complet⁸¹³ și ponderat, având ca mulțime de noduri chiar mulțimea S (așadar, fiecare nod din graf este reprezentat de o instanță $x_i \in S$), iar ponderea / costul fiecărei muchii din graf este exact distanța dintre punctele reprezentate de nodurile adiacente acestei muchii.

Pasul 2: Se calculează un arbore de acoperire de cost minim (engl., Minimum Spanning Tree) corespunzător grafului obținut la **Pasul 1**. Această operație revine la alegerea unei submulțimi de muchii din graful complet care *i.* constituie un arbore T ce folosește / conexează toate nodurile din graf, iar *ii.* suma costurilor / lungimilor muchiilor sale este minimă (în raport cu toți arborii care satisfac condiția precedentă). Acest arbore se poate obține folosind *algoritmul lui Kruskal*⁸¹⁴ sau *algoritmul lui Prim*⁸¹⁵.

⁸¹³Un graf neorientat este complet dacă între oricare două noduri ale sale există o [singură] muchie.

⁸¹⁴Publicat în 1956 de americanul Joseph Kruskal (1928-2010) în *Proceedings of American Mathematical Society*, pag. 48-50.

⁸¹⁵Algoritm descoperit de Vojtech Jarnik în 1930, redescoperit de Robert Prim în 1957 și din nou de către Edsger Dijkstra în 1959. Din această cauză, el este numit uneori în literatura de specialitate *algoritmul DJP*. Atât acest algoritm cât și algoritmul lui Kruskal au complexitatea $O(n \log n)$.

Pasul 3: Se elimină din arborele de cost minim obținut la *Pasul 2* muchia cu costul maxim, obținându-se astfel doi arbori. Aceștia vor corespunde celor două clustere de pe nivelul cel mai de sus în dendrograma pe care o construim în manieră top-down.

Pasul 4: Se repetă recursiv *Pasul 3* atât timp cât este posibil, obținând astfel o clusterizare de tip top-down pe mulțimea de instanțe date.

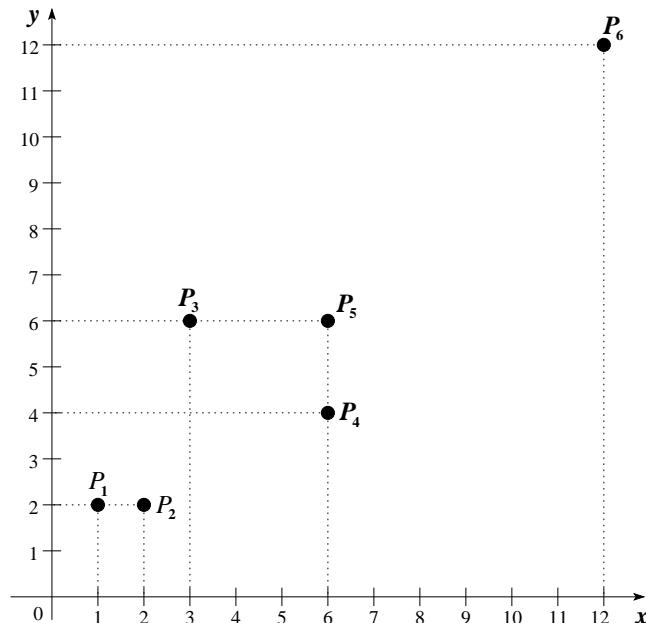
- a. Aplicați acest algoritm pe setul de date din \mathbb{R}^2 din tabelul alăturat, folosind distanța euclidiană. Desenați rezultatul sub forma unei ierarhii „aplatizate“ sau ca dendrogramă (fără a face calculul exact al înălțimilor).

Punctul	x	y
P_1	1	2
P_2	2	2
P_3	3	6
P_4	6	4
P_5	6	6
P_6	12	12

- b. Este oare acest algoritm diviziv echivalent⁸¹⁶ cu unul dintre algoritmii de clusterizare aglomerativă studiați? Justificați atât pe datele acestea cât și în cazul general. Atenție la situațiile în care se poate alege între mai mulți candidați.

Răspuns:

- a. Reprezentarea celor șase puncte în plan este dată în figura alăturată.



Vom aplica pașii algoritmului descris în enunț:

Pasul 1: La calcularea distanțelor dintre perechile de puncte din setul de instanțe date, se poate completa un tabel, aşa cum apare mai jos. Deoarece aceste distanțe se pot calcula folosind teorema lui Pitagora, iar ulterior ele vor fi folosite doar pentru a face diverse comparații între ele, este suficient să le exprimăm cu ajutorul radicalului. Notăm faptul că acest tabel al distanțelor

⁸¹⁶Adică: produc exact același rezultat și eventual exact în aceeași ordine (vedeți pasul 2 și respectiv pașii 3-4).

poate fi considerat *matricea de adiacență* a grafului complet pe care vom lucra în continuare.

	(1, 2) P_1	(2, 2) P_2	(3, 6) P_3	(6, 4) P_4	(6, 6) P_5	(12, 12) P_6
(1, 2) P_1	0	$\sqrt{1}$	$\sqrt{20}$	$\sqrt{29}$	$\sqrt{41}$	$\sqrt{221}$
(2, 2) P_2	$\sqrt{1}$	0	$\sqrt{17}$	$\sqrt{20}$	$\sqrt{32}$	$\sqrt{200}$
(3, 6) P_3	$\sqrt{20}$	$\sqrt{17}$	0	$\sqrt{13}$	$\sqrt{9}$	$\sqrt{117}$
(6, 4) P_4	$\sqrt{29}$	$\sqrt{20}$	$\sqrt{13}$	0	$\sqrt{4}$	$\sqrt{100}$
(6, 6) P_5	$\sqrt{41}$	$\sqrt{32}$	$\sqrt{9}$	$\sqrt{4}$	0	$\sqrt{72}$
(12, 12) P_6	$\sqrt{221}$	$\sqrt{200}$	$\sqrt{117}$	$\sqrt{100}$	$\sqrt{72}$	0

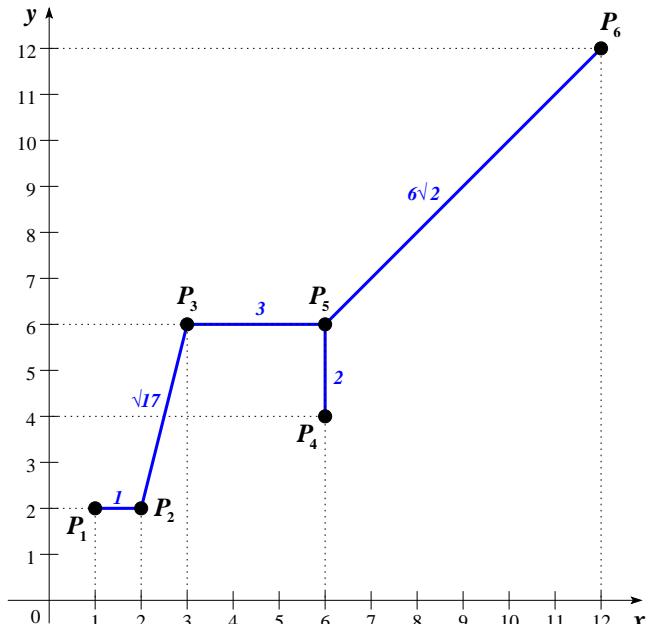
Pasul 2: Vom determina un arbore de cost minim folosind *algoritmul lui Kruskal*. Pentru un graf conex,⁸¹⁷ acest algoritm poate fi descris pe scurt astfel:

Se pornește de la graful parțial de cost minim [care are toate nodurile grafului conex dat, dar] care nu are nicio muchie. Apoi, în mod iterativ se alege câte o muchie de cost minim din graful inițial, care n-a fost încă folosită și care unește două componente conexe ale grafului parțial de la iterația anterioară. Algoritmul se încheie după $n - 1$ iterații, unde n este numărul de noduri din graf, adică atunci când graful parțial construit este conex. Se poate demonstra că acesta este chiar un arbore de cost minim.

Aplicat pe setul de date din enunț, algoritmul va alege în ordine următoarele muchii:

- (P_1, P_2) cu costul 1,
- (P_4, P_5) cu costul 2,
- (P_3, P_5) cu costul 3,
- (P_2, P_3) cu costul $\sqrt{17}$,
- (P_5, P_6) cu costul $6\sqrt{2}$.

Arborele de cost minim rezultat este reprezentat în figura alăturată.



Pasul 3: Eliminând muchia de cost maxim, adică muchia (P_5, P_6) , care are costul $\sqrt{72} = 6\sqrt{2}$, rezultă clusterele: $C_1 = \{P_1, P_2, P_3, P_4, P_5\}$ și $C_2 = \{P_6\}$.

La iterația următoare se elimină muchia (P_2, P_3) cu costul $\sqrt{17}$, rezultând clusterele: $C_3 = \{P_1, P_2\}$ și $C_4 = \{P_3, P_4, P_5\}$.

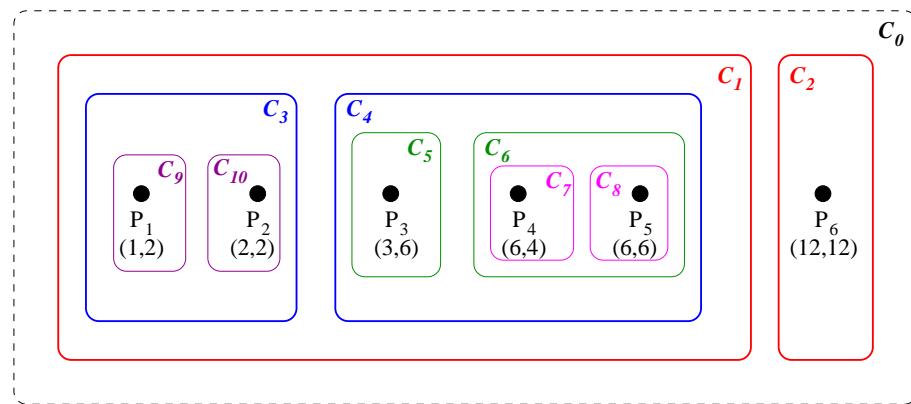
Apoi se elimină muchia (P_3, P_5) cu costul 3, rezultând clusterele: $C_5 = \{P_3\}$ și $C_6 = \{P_4, P_5\}$.

⁸¹⁷Un graf neorientat este conex dacă între oricare două noduri ale sale există cel puțin un drum.

Au mai rămas două muchii, care se vor elimina pe rând, și anume: mai întâi muchia (P_4, P_5) , producând clusterele $C_7 = \{P_4\}$ și $C_8 = \{P_5\}$, și în final muchia (P_1, P_2) , rezultând clusterele $C_9 = \{P_1\}$ și $C_{10} = \{P_2\}$.

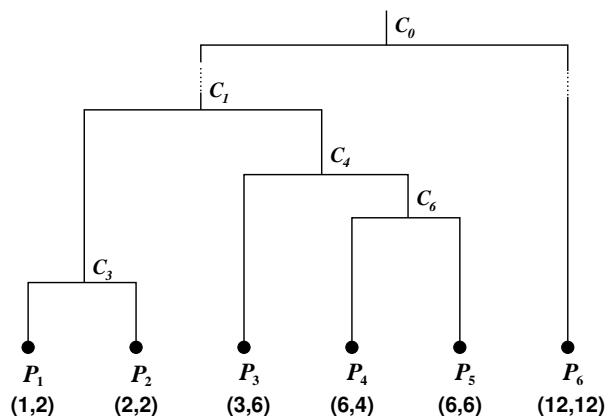
Observație: Eliminarea muchiilor din arbore se realizează exact în ordine inversă față de ordinea în care au fost alese inițial de către algoritmul lui Kruskal.

Putem reprezenta acum într-o formă aplatizată efectul aplicării algoritmului de clusterizare ierarhică top-down pe setul de instanțe date:



b. Dacă se folosește algoritmul lui Kruskal pentru determinarea arborelui de cost minim, atunci algoritmul de clusterizare top-down din enunț este echivalent (în privința rezultatelor obținute) cu algoritmul de clusterizare aglomerativă care folosește similaritate de tip “single-linkage”. Chiar mai mult, alegerea muchiilor arborelui de cost minim corespunde (în ordine inversă, pas cu pas) calculului distanțelor de tip “single-linkage” dintre clustere.

Dendrograma obținută — fără a include calculul înălțimilor — de către algoritmul de clusterizare aglomerativă folosind similaritate de tip “single-linkage” pe datele din enunț este cea din figura alăturată.



Observație 1: Este posibil ca atunci când se folosește *algoritmul lui Prim* pentru determinarea arborelui de cost minim să se obțină rezultate diferite față de cazul când se folosește algoritmul lui Kruskal. Pentru un graf conex, algoritmul lui Prim poate fi descris pe scurt astfel:

Inițial, arborele este alcătuit dintr-un singur nod, ales în mod aleatoriu dintre nodurile grafului. La fiecare iterație se adaugă la acest arbore muchia de cost minim (și nodul corespunzător) care unește unul dintre nodurile nefolosite cu arborele deja existent. Algoritmul se încheie după $n - 1$ iterării, unde n este numărul de noduri din graf, adică

atunci când graful parțial construit conectează toate nodurile grafului inițial. Acest graf parțial este chiar un arbore de cost minim.

Observația 2: Dacă graful considerat admite un singur arbore de cost minim — se poate verifica imediat că aşa se întâmplă și în cazul de față —, atunci rezultatul aplicării algoritmilor lui Kruskal și respectiv al lui Prim este, evident, același. În caz contrar, arborii obținuți de fiecare dintre cei doi algoritmi depind de alegerile care pot fi făcute atunci când la un pas oarecare există mai multe muchii de cost minim. În plus, arborele obținut de către algoritmul lui Prim depinde de alegerea primului nod. Chiar dacă prin aplicarea celor doi algoritmi se obține același rezultat (adică același arbore de cost minim), în general ordinea de alegere a muchiilor diferă. Algoritmul lui Prim extinde pas cu pas un arbore care în final devine arbore(le) de cost minim, în vreme ce algoritmul lui Kruskal construiește o *pădure de arbori* pe care-i unește până la final într-un arbore de cost minim.

Așadar, în cazul în care la *Pasul 2* al algoritmului de clusterizare top-down prezentat în enunț se folosește algoritmul lui Prim, în general nu se poate stabili o echivalentă directă între modul de lucru al algoritmului de clusterizare top-down pe de o parte (*Pasul 2*) și modul de lucru al algoritmului de clusterizare aglomerativă cu similaritate de tip “single-linkage” pe de altă parte,⁸¹⁸ chiar dacă rezultatul final (adică, dendrograme) este exact același ca în cazul folosirii algoritmului lui Kruskal.

Indiferent dacă se folosește algoritmul lui Kruskal sau algoritmul lui Prim, la pașii 3-4 din algoritmul de clusterizare top-down se procedează exact în ordine inversă față de construirea dendrogramei prin clusterizare bottom-up folosind similaritate single-linkage (făcând alegerile corespunzătoare, în cazul situațiilor care comportă posibilități de alegere multiple).

7.1.2 Algoritmul *K-means*

7. (Algoritmul *K-means*:⁸¹⁹ un exemplu simplu de aplicare pe date din \mathbb{R}^2)
prelucrare de Liviu Ciortuz, după
 • Univ. of Utah, 2008 spring, HW3A, pr. 2

Considerăm în planul euclidian bidimensional un set de date format din următoarele puncte: $A(-1, 0)$, $B(1, 0)$, $C(0, 1)$, $D(3, 0)$ și $E(3, 1)$.

a. Rulați manual două iterații ale algoritmului 2-means pe acest set de date,⁸²⁰ pornind de la următoarele poziții inițiale ale centroizilor: $(-1, 0)$ și $(3, 1)$. Veți folosi distanța euclidiană. La fiecare iterație indicați cum sunt asignate punctele la clustere și care sunt pozițiile actualizate ale centroizilor. A convers algoritmul după ce au fost efectuate cele două iterații?

⁸¹⁸Si cu atât mai puțin pentru similaritate “complete-linkage” sau “average-linkage”.

⁸¹⁹La problema aceasta, precum și la fiecare dintre problemele următoare, dacă nu se specifică altfel, se va folosi forma algoritmului *K-means* din cartea *Foundations of Statistical Natural Language Processing* (capitolul 14) de C. Manning and H. Schütze, editura MIT Press, 2002.

⁸²⁰Pentru pseudo-codul algoritmului *K-means*, puteți vedea enunțul problemei 12.

b. Definim

$$J(\mu^{(t)}) = \sum_{x^{(i)} \in \{A, B, C, D, E\}} \|x^{(i)} - \mu^{(t)}(x^{(i)})\|^2$$

unde $\mu^{(t)}$ desemnează ansamblul centroizilor la momentul / iterația t , iar $\mu^{(t)}(x^{(i)})$ este centroidul clusterului la care este asignată instanța $x^{(i)}$ la iterația t . Simbolul $\|\cdot\|$ reprezintă norma euclidiană. Pentru un vector oarecare $x \stackrel{\text{not.}}{=} (x_1, \dots, x_d) \in \mathbb{R}^d$, avem $\|x\|^2 \stackrel{\text{def.}}{=} x \cdot x \stackrel{\text{def.}}{=} x_1^2 + \dots + x_d^2 \in \mathbb{R}^+$. Arătați pur și simplu prin calcul numeric că pentru $t = 1$ avem

$$J(\mu^{(t)}) \leq J(\mu^{(t-1)}).$$

(Am desemnat cu $t = 0$ iterația inițială.)

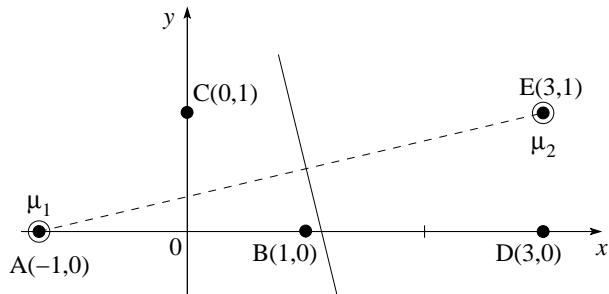
Răspuns:

a. Conform enunțului, centroizii inițiali sunt chiar două dintre punctele considerate: $\mu_1 \equiv A(-1, 0)$ și $\mu_2 \equiv E(3, 1)$. Atribuirea celorlalte puncte la clustere se poate face în manieră *analitică*, calculând distanțele dintre puncte și centroizi:

$$\begin{aligned} d(B, \mu_1) &= 2 \text{ și } d(B, \mu_2) = \sqrt{5} \Rightarrow B \text{ se atribuează clusterului cu centroidul } \mu_1 \\ d(C, \mu_1) &= \sqrt{2} \text{ și } d(C, \mu_2) = 3 \Rightarrow C \text{ se atribuează clusterului cu centroidul } \mu_1 \\ d(D, \mu_1) &= 4 \text{ și } d(D, \mu_2) = 1 \Rightarrow D \text{ se atribuează clusterului cu centroidul } \mu_2 \end{aligned}$$

Alternativ, adică *geometric*, se reprezintă punctele date și centroizii inițiali în planul euclidian, după care se trasează mediatoarea segmentului $\mu_1\mu_2$. În figura de mai jos am reprezentat instanțele cu un cerculeț umplut (\bullet), iar centroizii cu un cerculeț simplu. Cele două clustere sunt separate de către mediatoarea segmentului care unește cei doi centroizi.

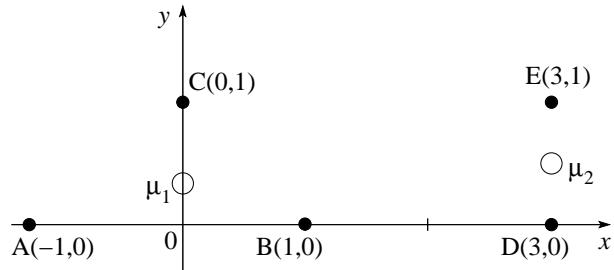
Se observă că punctele B și C sunt situate de aceeași parte a mediatoarei ca și centroidul μ_1 , deci rezultă că B și C vor apartine clusterului cu centroidul μ_1 . Similar, punctul D va apartine clusterului cu centroidul μ_2 .



Având aşadar componența inițială a clusterelor $\{A(-1, 0), B(1, 0), C(0, 1)\}$ și, respectiv, $\{D(3, 0), E(3, 1)\}$, la următoarea iterație algoritmului se calculează noile poziții ale centroizilor (vedeți calculul alăturat).

$$\left. \begin{aligned} x_{\mu_1} &= \frac{x_A + x_B + x_C}{3} = 0 \\ y_{\mu_1} &= \frac{y_A + y_B + y_C}{3} = \frac{1}{3} \end{aligned} \right\} \Rightarrow \mu_1(0, \frac{1}{3})$$

$$\left. \begin{aligned} x_{\mu_2} &= \frac{x_D + x_E}{2} = 3 \\ y_{\mu_2} &= \frac{y_D + y_E}{2} = \frac{1}{2} \end{aligned} \right\} \Rightarrow \mu_2(3, \frac{1}{2})$$



Acum, reprezentarea grafică este cea din figura alăturată.

Apoi se re-asignează punctele la centroizi, operație în urma căreia clusterele se pot modifica:

$$d(A, \mu_1) = \frac{\sqrt{10}}{3} \text{ și } d(A, \mu_2) = \frac{\sqrt{65}}{2} \Rightarrow A \text{ se asignează la centroidul } \mu_1$$

$$d(B, \mu_1) = \frac{\sqrt{10}}{3} \text{ și } d(B, \mu_2) = \frac{\sqrt{17}}{2} \Rightarrow B \text{ se asignează la centroidul } \mu_1$$

$$d(C, \mu_1) = \frac{2}{3} \text{ și } d(C, \mu_2) = \frac{\sqrt{37}}{2} \Rightarrow C \text{ se asignează la centroidul } \mu_1$$

$$d(D, \mu_1) = \frac{\sqrt{82}}{3} \text{ și } d(D, \mu_2) = \frac{1}{2} \Rightarrow D \text{ se asignează la centroidul } \mu_2$$

$$d(E, \mu_1) = \frac{\sqrt{85}}{3} \text{ și } d(E, \mu_2) = \frac{1}{2} \Rightarrow E \text{ se asignează la centroidul } \mu_2$$

Se observă deci analitic (dar și grafic, vedeti figura alăturată) că după această (a doua) iterație avem aceeași componentă a clusterelor (ca și la finalul precedentei iterării), și anume: $\{A(-1,0), B(1,0), C(0,1)\}$ și $\{D(3,0), E(3,1)\}$. Prin urmare, centroizii rămân $\mu_1(0, 1/3)$ și $\mu_2(3, 1/2)$.

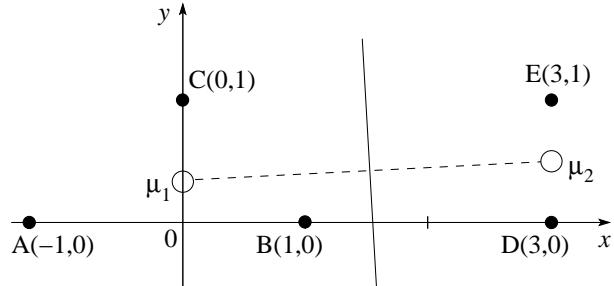
Putem concluziona că după aceste două iterării algoritmul K-means a converz.

Observație importantă:

Atât în această problemă cât și în cele următoare, ne-am străduit să elaborăm soluția algoritmului K-means atât bazat pe calculul distanțelor (am numit-o deci soluția *analitică*), cât și bazat pe reprezentarea datelor în planul euclidian (am numit-o soluția *geometrică*). Aceasta din urmă are avantajul că oferă un suport vizual care ajută studentul să înțeleagă lucrurile într-o manieră mai intuitivă.

La implementare, cele două tipuri de soluții apelează la procedee de calcul care diferă parțial.⁸²¹

În condiții de seminar sau de examen, cele două procedee pot fi combinate — iar acolo unde detaliile sunt evidente, calculele pot fi lăsate de o parte dacă nu se specifică altfel —, datorită restricțiilor de timp impuse.



⁸²¹ Este mai simplu de realizat implementarea bazată pe metoda analitică. Însă pentru vizualizarea rezultatelor intermedii sau finale — aşa cum se procedează la rezolvarea majorității problemelor din această secțiune a prezentului capitol — se utilizează și componente ale variantei geometrice.

b. Vom reprezenta punctele / instanțele de clusterizat sub formă de vectori-coloană, cu două componente (una pentru abscisa x și cealaltă pentru ordonata y).

$$\begin{aligned} J(\mu^{(0)}) &= \left\{ \left[\begin{pmatrix} -1 \\ 0 \end{pmatrix} - \begin{pmatrix} -1 \\ 0 \end{pmatrix} \right]^2 + \left[\begin{pmatrix} 1 \\ 0 \end{pmatrix} - \begin{pmatrix} -1 \\ 0 \end{pmatrix} \right]^2 + \left[\begin{pmatrix} 0 \\ 1 \end{pmatrix} - \begin{pmatrix} -1 \\ 0 \end{pmatrix} \right]^2 \right\} \\ &\quad + \left\{ \left[\begin{pmatrix} 3 \\ 0 \end{pmatrix} - \begin{pmatrix} 3 \\ 1 \end{pmatrix} \right]^2 + \left[\begin{pmatrix} 3 \\ 1 \end{pmatrix} - \begin{pmatrix} 3 \\ 1 \end{pmatrix} \right]^2 \right\} \\ &= \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}^2 + \begin{pmatrix} 2 \\ 0 \end{pmatrix}^2 + \begin{pmatrix} 1 \\ 1 \end{pmatrix}^2 \right\} + \left\{ \begin{pmatrix} 0 \\ -1 \end{pmatrix}^2 + \begin{pmatrix} 0 \\ 0 \end{pmatrix}^2 \right\} \\ &= 0 + 4 + 2 + 1 + 0 = 7. \end{aligned}$$

$$\begin{aligned} J(\mu^{(1)}) &= \left\{ \left[\begin{pmatrix} -1 \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ 1/3 \end{pmatrix} \right]^2 + \left[\begin{pmatrix} 1 \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ 1/3 \end{pmatrix} \right]^2 + \left[\begin{pmatrix} 0 \\ 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 1/3 \end{pmatrix} \right]^2 \right\} \\ &\quad + \left\{ \left[\begin{pmatrix} 3 \\ 0 \end{pmatrix} - \begin{pmatrix} 3 \\ 1/2 \end{pmatrix} \right]^2 + \left[\begin{pmatrix} 3 \\ 1 \end{pmatrix} - \begin{pmatrix} 3 \\ 1/2 \end{pmatrix} \right]^2 \right\} \\ &= \left\{ \begin{pmatrix} -1 \\ -1/3 \end{pmatrix}^2 + \begin{pmatrix} 1 \\ -1/3 \end{pmatrix}^2 + \begin{pmatrix} 0 \\ 2/3 \end{pmatrix}^2 \right\} + \left\{ \begin{pmatrix} 0 \\ -1/2 \end{pmatrix}^2 + \begin{pmatrix} 0 \\ 1/2 \end{pmatrix}^2 \right\} \\ &= \frac{10}{9} + \frac{10}{9} + \frac{4}{9} + \frac{1}{4} + \frac{1}{4} = \frac{24}{9} + \frac{1}{2} = \frac{8}{3} + \frac{1}{2} = \frac{19}{6}. \end{aligned}$$

Prin urmare, se verifică inegalitatea $J(\mu^{(1)}) \leq J(\mu^{(0)})$.

Observație:

Proprietatea de descreștere [nu neapărat strictă] a valorilor funcției J în cursul execuției algoritmului K -means (la orice iterare și pe orice set de date de clusterizat) va face obiectul problemei 12.

8. (Algoritmul K -means: exemplificare pentru alegerea centroizilor inițiali astfel încât clusterele obținute să satisfacă anumite restricții)

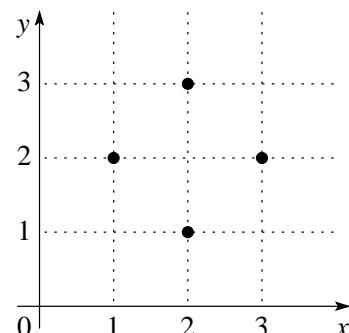
CMU, 2009 spring, Ziv Bar-Joseph, HW5, pr. 2.2

Considerăm 4 puncte dispuse în spațiu aşa cum se arată în figura alăturată.

a. Aplicând algoritmul K -means pentru $K = 2$, alegeți centroizii inițiali în aşa fel încât să obțineți o clusterizare în care fiecare cluster să conțină câte 2 elemente.

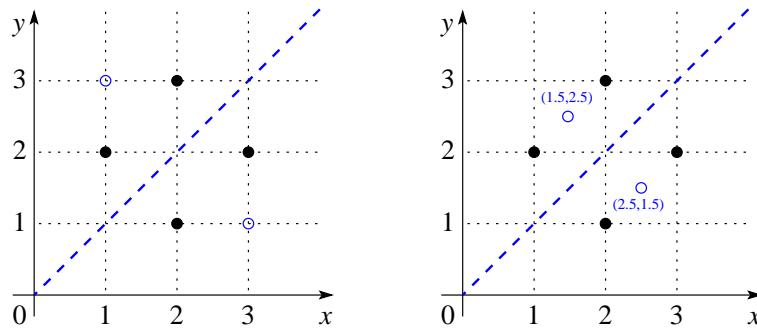
b. Aceeași cerință, de data aceasta pentru a obține 2 clustere dintre care unul să conțină un element, iar celălalt 3 elemente.

Observație. Se va folosi distanța euclidiană.

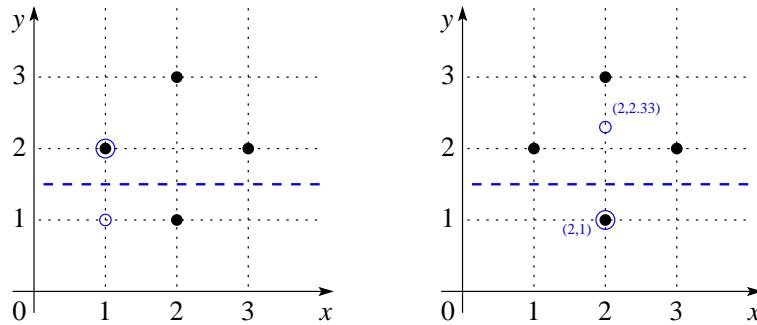


Răspuns:

- a. Pentru a se obține două clustere de căte două elemente, se pot alege centroizii inițiali $\mu_1 = (1, 3)$ și $\mu_2 = (3, 1)$. După prima iterare (care se dovedește a fi și ultima), se vor obține centroizii $\mu'_1 = (1.5, 2.5)$ și $\mu'_2 = (2.5, 1.5)$, ca în figura de mai jos. Instantele au fost reprezentate prin cerculețe umplute (\bullet), iar simbolurile \circ marchează pozițiile centroizilor. Liniile punctate constituie separatorii celor două clustere.



- b. Pentru a se obține două clustere cu 1 și respectiv 3 elemente, se pot alege centroizii inițiali $\mu_1 = (1, 1)$ și $\mu_2 = (1, 2)$. După o primă iterare, se vor obține centroizii $\mu'_1 = (2, 1)$ și $\mu'_2 = (2, 2, (3))$, ca în figura de mai jos, după care algoritmul converge.



9.

(Algoritmul K -means: aplicare pe date din \mathbb{R}^2)
CMU, 2009 spring, Ziv Bar-Joseph, final exam, pr. 1.8

Se dă un set de puncte din planul euclidian, conform tabelului alăturat.

Vă cerem să rulați manual algoritmul K -means pentru a identifica două clustere în acest set de date. Centroizii inițiali sunt chiar punctele P_1 și P_{10} . Se folosesc distanțe euclidiene.

Care este compoziția clusterelor după prima iterare? Dar la convergența algoritmului?

Punctul	x	y
P_1	1.90	0.97
P_2	1.76	0.84
P_3	2.32	1.63
P_4	2.31	2.09
P_5	1.14	2.11
P_6	5.02	3.02
P_7	5.74	3.84
P_8	2.25	3.47
P_9	4.71	3.60
P_{10}	3.17	4.96

Răspuns:

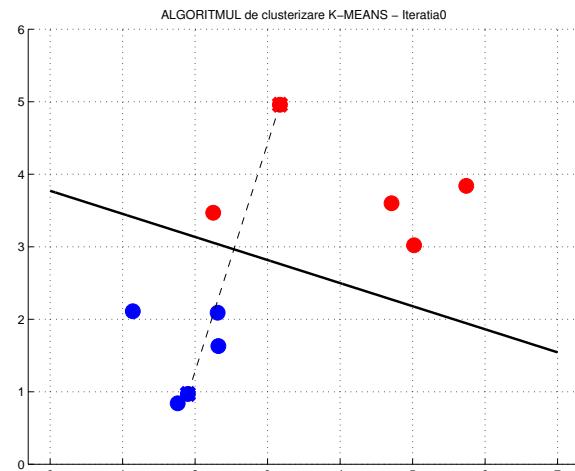
La prima iterație a algoritmului K -means, fiecare punct P_i va fi asignat unui dintre cele două clustere (C_1 și C_2), în funcție de distanțele dintre punctul respectiv și centroizii inițiali.

Construim un tabel cu aceste distanțe și asignările deduse din relația de ordine dintre ele. Vom nota cu $d(P_i, P_j)$ distanța de la punctul P_i la punctul P_j .

P_i	$d(P_i, P_1)$	$d(P_i, P_{10})$	$d(P_i, P_1) \leq d(P_i, P_{10})$	Clusterul
P_1	0	4.18	<	C_1
P_2	0.19	4.35	<	C_1
P_3	0.78	3.43	<	C_1
P_4	1.19	2.99	<	C_1
P_5	1.37	3.49	<	C_1
P_6	3.73	2.68	>	C_2
P_7	4.97	2.80	>	C_2
P_8	2.52	1.75	>	C_2
P_9	3.84	2.05	>	C_2
P_{10}	4.18	0	>	C_2

Așadar, la finalul primei iterări obținem clusterele $C_1 = \{P_1, P_2, P_3, P_4, P_5\}$ și $C_2 = \{P_6, P_7, P_8, P_9, P_{10}\}$.

Același rezultat se poate obține și procedând *geometric*. Pentru aceasta, în planul euclidian bidimensional trăsăm mediatoarea segmentului P_1P_{10} și apoi analizăm poziția fiecărui punct dat în raport cu această mediatoare. Rezultatul este arătat în figura alăturată.



Pozиїile centroizilor (marcate pe grafic prin caracterul \times) se recalculează astfel:

$$\begin{aligned} x_{\mu_1} &= \frac{x_{P_1} + x_{P_2} + x_{P_3} + x_{P_4} + x_{P_5}}{5} = 1.886 \\ y_{\mu_1} &= \frac{y_{P_1} + y_{P_2} + y_{P_3} + y_{P_4} + y_{P_5}}{5} = 1.528 \end{aligned}$$

Prin urmare, $\mu_1 = (1.886, 1.528)$. Analog, obținem $\mu_2 = (4.178, 3.778)$.

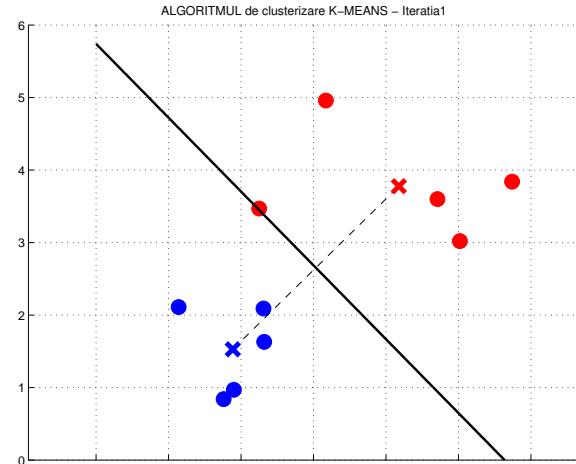
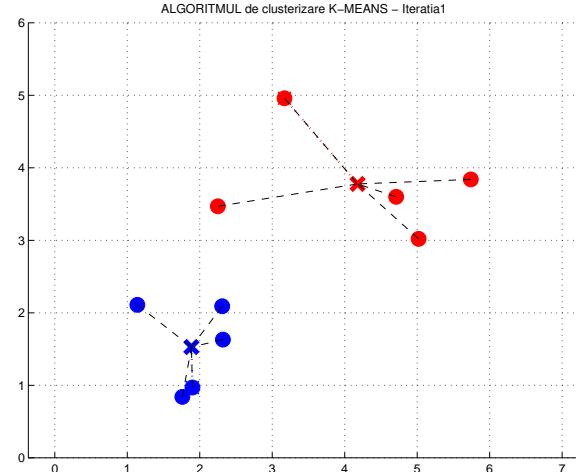
Dacă am decis să procedăm *analitic*, în continuare vom calcula distanțele de la fiecare punct P_i la centroizii μ_1 și respectiv μ_2 .

În figura alăturată am marcat noile poziții ale centroizilor (μ_1 și μ_2), precum și asignarea fiecărui dintre punctele P_i la cel mai apropiat centroid, în funcție de distanțe.

Așadar, punctele P_1, \dots, P_5 formează — și de data aceasta! — primul cluster (cel cu centroidul μ_1), iar restul punctelor, P_6, \dots, P_{10} , constituie cel de-al doilea cluster (cel cu centroidul μ_2).

Alternativ, adică în ipoteza că am optat pentru o *rezolvare geometrică*, am observat că punctul $P_8(2.25, 3.47)$ are o poziție „la limită“ față de mediatoarea segmentului care unește centroizii μ_1 și μ_2 , așa cum se observă din figura de mai jos.

Pentru a decide cărui cluster îi va apartine acest punct, fie calculăm $d^2(P_8, \mu_1)$ și $d^2(P_8, \mu_2)$ — și vom obține valorile 3.904 și respectiv 3.812 —, fie comparăm semnul expresiei $f(2.52, 1.75)$ cu cel al lui $f(x_{\mu_1}, y_{\mu_1})$ (sau $f(x_{\mu_2}, y_{\mu_2})$), unde f este funcția din membrul stâng al ecuației $f(x, y) = 0$ care definește mediatoarea segmentului determinat de punctele μ_1 și μ_2 .



În ambele variante se ajunge la aceeași concluzie: punctul P_8 va apartine clusterului cu centroidul μ_2 . (Valoarea expresiei $f(2.52, 1.75)$ are același semn ca $f(x_{\mu_2}, y_{\mu_2})$.)

Clusterele finale sunt deci: $\{P_1, P_2, P_3, P_4, P_5\}$ și $\{P_6, P_7, P_8, P_9, P_{10}\}$.

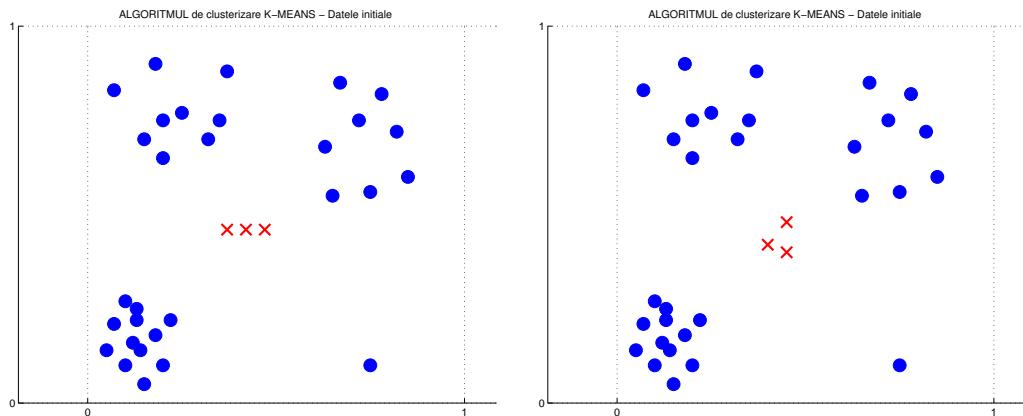
10.

(Algoritmul *K-means*: aplicare în \mathbb{R}^2 ; compararea rezultatelor obținute în cazul a două inițializări diferite ale centroizilor)

■ • CMU, 2006 spring, Carlos Guestrin, HW5, pr. 1

Se consideră setul de date reprezentat în cele două figuri de mai jos. Sunt date două inițializări posibile ale centroizilor, câte una în fiecare din cele două figuri. Instanțele sunt reprezentate de cerculete umplute (●), iar simbolurile

- corespund pozițiilor inițiale ale centroizilor.⁸²²



a-b. Execuați / aplicați algoritmul *K*-means separat pentru fiecare dintre cele două inițializări ale centroizilor, arătând la fiecare iterație cum evoluează centroizii și separatorii, până la convergență.

Observație: La execuția algoritmului se va considera că atunci când un centroid nu are puncte asignate lui, atunci el va rămâne pe loc la iterarea respectivă.

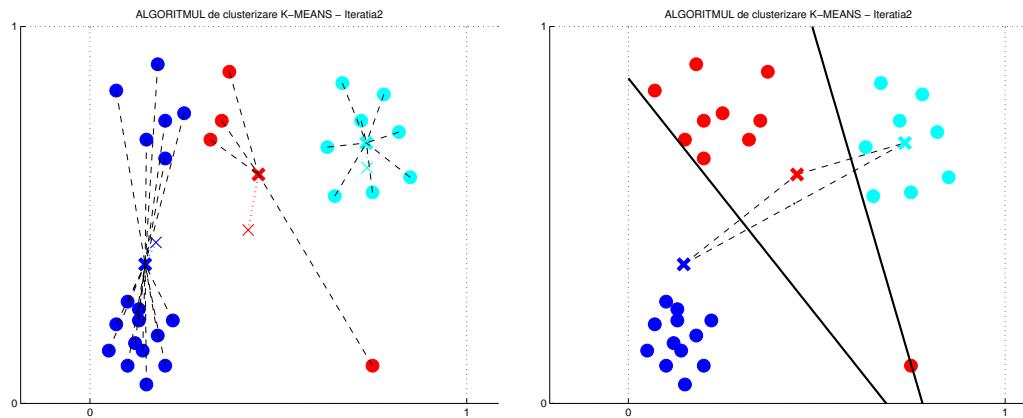
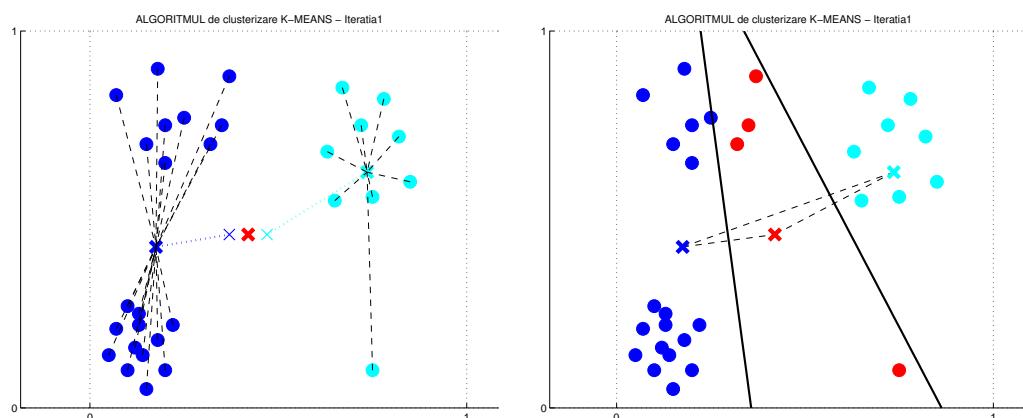
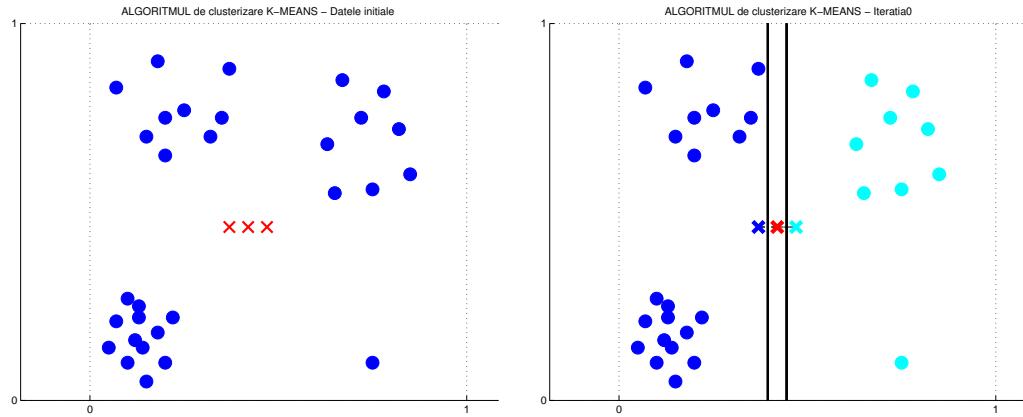
c. Având în vedere rezultatele obținute în cele două cazuri, ce puteți spune despre comportamentul algoritmului *K*-means?

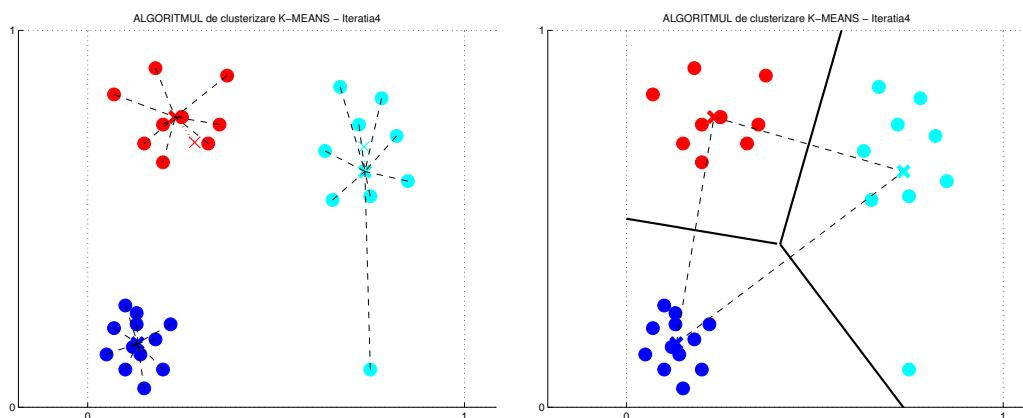
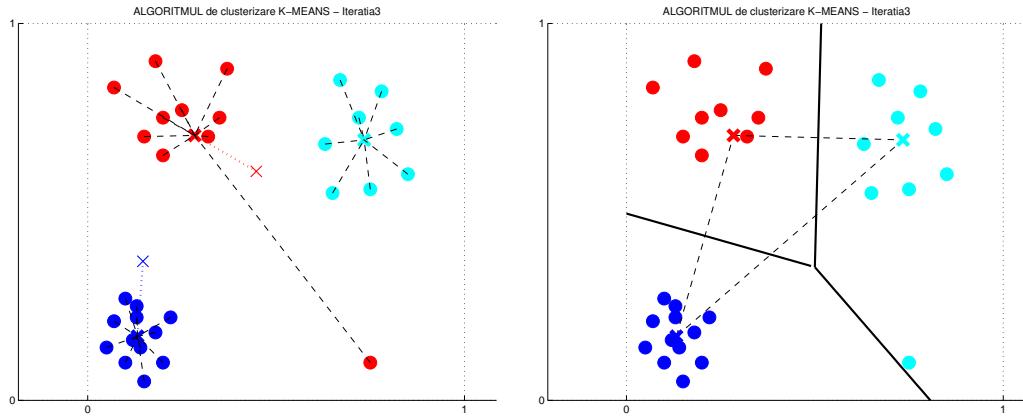
Răspuns:

a. Pentru prima inițializare a centroizilor, algoritmul *K*-means parcurge următoarele iterări:⁸²³

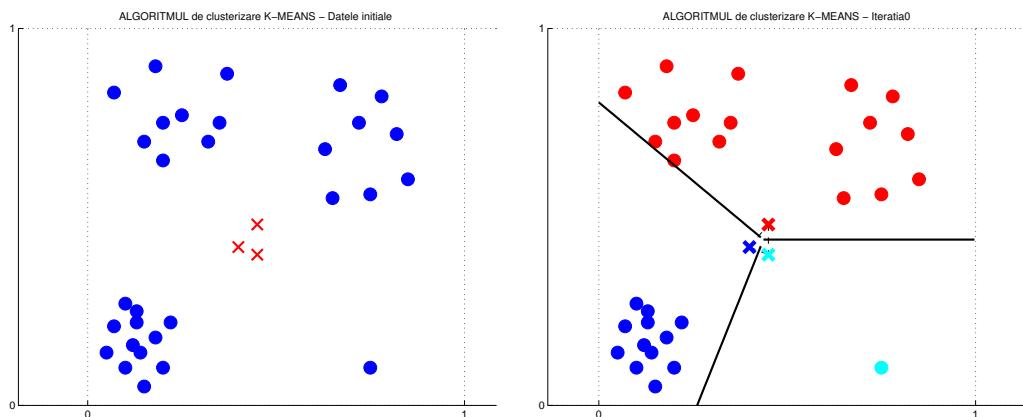
⁸²²Coordonatele exacte ale instanțelor, precum și cele ale centroizilor vă sunt puse la dispoziție în următoarele fișiere, depuse pe site-ul acestei cărți:
<http://profinfo.uaic.ro/~ciortuz/ML.ex-book/res/CMU.2006s.HW5.pr1.cl.dat>,
<http://profinfo.uaic.ro/~ciortuz/ML.ex-book/res/CMU.2006s.HW5.pr1.a.init.dat>,
<http://profinfo.uaic.ro/~ciortuz/ML.ex-book/res/CMU.2006s.HW5.pr1.b.init.dat>.

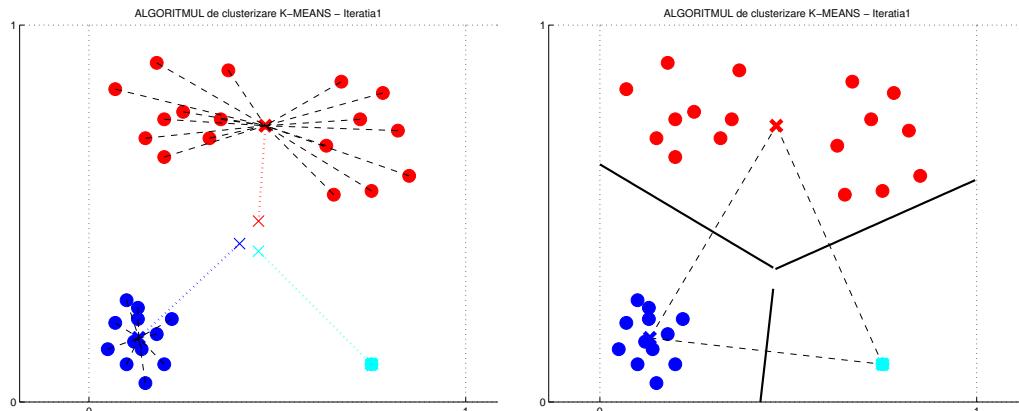
⁸²³La fiecare iterare, în afară de cea inițială (care a fost numerotată cu 0), am desenat în graficul din partea stângă noile poziții ale centroizilor — vedeti simbolul × mai îngroșat, spre deosebire de vechea lui poziție, care a fost redată neîngroșat pentru a sugera mișcarea —, iar în graficul din partea dreaptă am indicat noua componentă a clusterelor, în funcție de pozițiile actuale ale separatorilor (mediatoarele segmentelor care unesc centroizii).





b. Pentru a două inițializare a centroizilor, algoritmul *K-means* parcurge următoarele iterării:





c. Este evident din acest exercițiu că rezultatul algoritmului *K*-means depinde de poziționarea inițială a centroizilor. În cazul inițializării de la punctul *a* au fost necesare 4 iterații până a se ajunge la convergență, pe când la punctul *b* algoritmul a convergat după doar o iterație.

Mai este încă ceva important *de remarcat*: faptul că la prima variantă de inițializare, punctul din dreapta jos, care este un *outlier* (rom., excepție, caz particular, aberație) este până la urmă asociat clusterului format de punctele din partea dreaptă (sus), în vreme ce la cea de-a doua variantă de inițializare el constituie un cluster aparte / “singleton”, obligând în mod indirect grupările de puncte din stânga-sus și dreapta-sus să formeze împreună un singur cluster. Detectia outlier-elor este un capitol important din învățarea automată. Ignorarea lor poate conduce la rezultate eronate sau chiar aberante ale modelărilor.

11.

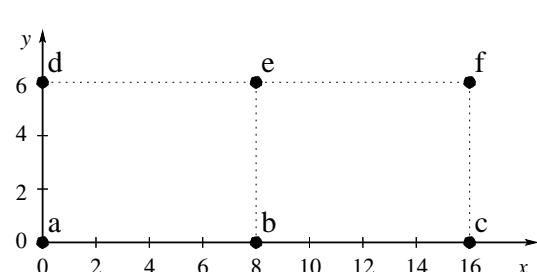
(Algoritmul *K*-means: aplicare în \mathbb{R}^2 ; realizarea corespondenței dintre *K*-configurația [obținută la o iterație] curentă a algoritmului și una sau mai multe *K*-configurații de start)

• ○ *CMU, 2003 fall, T. Mitchell, A. Moore, final exam, pr. 9*

Considerăm mulțimea S formată de următoarele 6 puncte din plan: $a = (0, 0)$, $b = (8, 0)$, $c = (16, 0)$, $d = (0, 6)$, $e = (8, 6)$ și $f = (16, 6)$. Reprezentarea datelor în planul euclidian este dată în figura alăturată. Se rulează algoritmul *K*-means, unde $K = 3$, și se folosește distanța euclidiană.

Introducem următoarele două noțiuni:

- ***K*-partiție:** desemnează o partiționare a mulțimii S în K submulțimi nevide. Spre exemplu, $\{a, b, e\}, \{c, d\}, \{f\}$ este o 3-partiție;



- **K -configurație de start:** desemnează o submulțime de K puncte din mulțimea S , care vor juca rolul de centroizi inițiali. Iată un exemplu de 3-configurație de start, care generează una dintre 3-partițiile considerate mai jos (și anume, $\{a, b\}, \{d, e\}, \{c, f\}$): $\{a, d, c\}$.

În mod evident, orice K -partiție induce o mulțime de K centroizi. O K -partiție este numită *stabilă* dacă după execuția unei noi iterații a algoritmului K -means, partiția respectivă rămâne neschimbată.

a. Câte 3-configurații de start există pentru mulțimea S dată mai sus?

b. Completă tabelul următor:

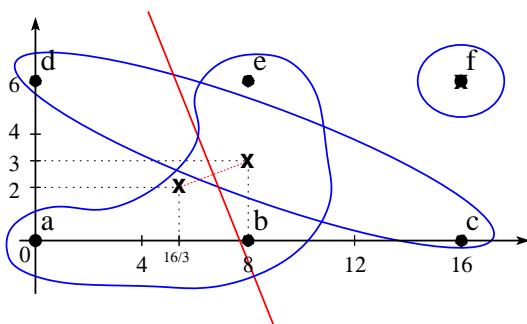
3-partiție	Stabilitate?	Un exemplu de 3-configurație de start care poate determina 3-partiția din prima coloană, după 0 sau mai multe iterații ale algoritmului K -means. (Dacă nu există o asemenea configurație, se va trece -.)	Numărul de 3-configurații de start care pot genera această 3-partiție
$\{a, b, e\}, \{c, d\}, \{f\}$	Nu	-	0
$\{a, b\}, \{d, e\}, \{c, f\}$	Da	$\{b, c, e\}$	4
$\{a, d\}, \{b, e\}, \{c, f\}$			
$\{a\}, \{d\}, \{b, c, e, f\}$			
$\{a, b\}, \{d\}, \{c, e, f\}$			
$\{a, b, d\}, \{c\}, \{e, f\}$			

Răspuns:

a. Pentru a forma o 3-configurație de start, se aleg oricare 3 puncte distincte din mulțimea S (care are 6 puncte), deci numărul acestor 3-configurații este $C_6^3 = \frac{6!}{3! \cdot 3!} = \frac{6 \cdot 5 \cdot 4}{1 \cdot 2 \cdot 3} = 20$.

b. Vom analiza pe rând fiecare dintre cele șase cazuri din tabel, inclusiv primele două cazuri pentru care avem deja anumite informații date în tabel, dar care merită a fi discutate. În figurile următoare, centroizii inițiali sunt reprezentați cu cerculete, iar centroizii actuali cu x .

Se poate verifica ușor faptul că în *cazul 1*, 3-partiția dată nu este stabilă. Ea „evolvează” în partiția $\{a, d\}, \{b, e\}, \{c, f\}$. Însă, ținând cont de faptul că pozițiile inițiale ale centroizilor sunt alese doar din mulțimea S , vom putea arăta mai jos că *nu* există nicio 3-configurație de start care să rezulte în 3-partiția dată.

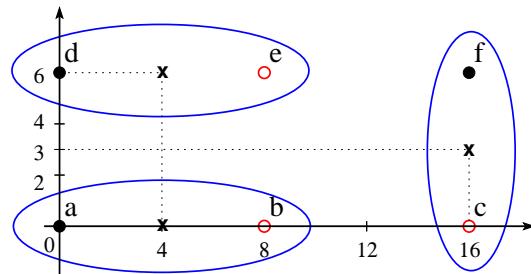


Pentru algoritmul K -means, atunci când se folosește distanța euclidiană, are loc următoarea proprietate: la finalul oricărei iterații a algoritmului (deci și la terminarea algoritmului), orice două clustere sunt separabile liniar (și anume, cu ajutorul mediatoarei segmentului care unește centroizii acestor două clustere, dacă instanțele sunt din \mathbb{R}^2). Demonstrarea acestei proprietăți

este imediată, ținând cont de modul în care sunt asignate instanțele la centroizi. (Vedeți pasul al doilea al corpului iterativ al algoritmului K -means, de exemplu în formularea din enunțul problemei 12.)

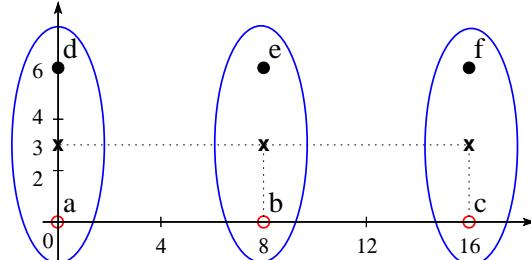
În condițiile de față (*cazul 1*), se observă imediat că mediatoarea segmentului care unește centroizii clusterelor $\{a, b, e\}$ și $\{c, d\}$ le separă pe fiecare în câte două submulțimi nevide, în loc să situeze un cluster de o parte a ei (adică, a mediatoarei) și celălalt cluster de cealaltă parte.

În *cazul 2*, se observă imediat că în afară de 3-configurația de start dată ($\{b, c, e\}$, deja stabilă), mai există 3 alte asemenea configurații de start care generează 3-partiția considerată ($(\{a, b\}, \{d, e\}, \{c, f\})$): $\{b, e, f\}$, $\{a, d, c\}$ și $\{a, d, f\}$.

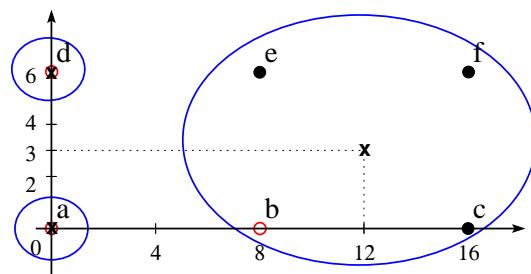


Observație: În acest caz,⁸²⁴ am presupus că este permisă libertatea de asociere a instanțelor care sunt egal distanțate față de doi centroizi. Aceeași presupozitie o vom folosi și în continuare, acolo unde va fi nevoie (*cazul 5* și *cazul 6*).

În *cazul 3*, se observă că 3-partiția dată ($\{a, d\}, \{b, e\}, \{c, f\}$) este stabilă și sunt ușor de enumerat cele 8 configurații de start care „termină” în această 3-partiție a mulțimii S : $\{a, b, c\}$, $\{a, b, f\}$, $\{a, e, c\}$, $\{a, e, f\}$, $\{d, b, c\}$, $\{d, b, f\}$, $\{d, e, c\}$, $\{d, e, f\}$.



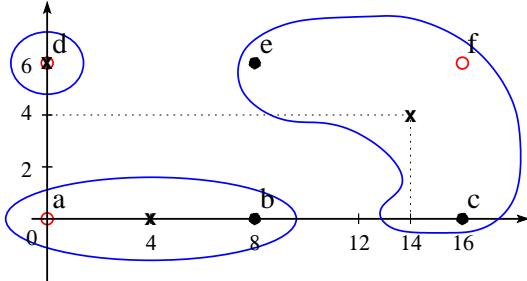
În *cazul 4*, unde 3-partiția de start ($\{a\}, \{d\}, \{b, c, e, f\}$) este de asemenea stabilă, este imediat că instanțele a și d trebuie să facă parte în mod obligatoriu din configurația (eventual, configurațiile) de start. Într-adevăr, dacă măcar una dintre instanțele a și d ar lipsi din configurația de start, se observă că algoritmul K -means nu ar putea să ajungă în 3-partiția dată.



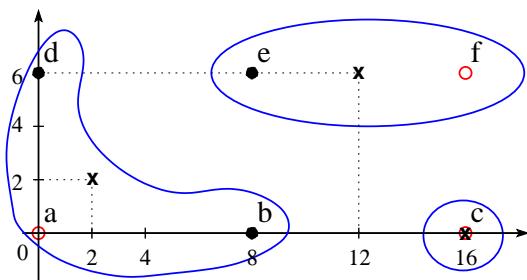
Pe de altă parte, c și f nu pot să apară în configurația de start (din exact același motiv). Rămân de analizat pozițiile b și e ; se verifică ușor că acestea pot completa (pe rând) 3-configurațiile de start, care sunt prin urmare $\{a, b, d\}$ și $\{a, d, e\}$.

⁸²⁴Vedeți configurațiile de start $\{b, c, e\}$ și $\{b, e, f\}$ menționate mai sus.

În cazul 5, 3-partiția dată ($\{a, b\}, \{d\}, \{c, e, f\}$) este stabilă. Se observă că pe de o parte instanța d trebuie să facă parte din orice configurație de start care ar putea să genereze 3-partiția dată, iar pe de altă parte instanțele b , c și e nu pot face parte din nicio astfel de configurație de start. Celelalte instanțe din S formează singura 3-configurație de start validă: $\{a, d, f\}$.



Cazul 6 este similar cazului 5 (datorită simetriei celor două 3-partiții date față de punctul $(8; 3)$), așadar singura 3-configurație de start posibilă este $\{a, c, f\}$.



În final, centralizăm în tabel rezultatele obținute:

Caz	3-partiție	Stabilitate?	Un exemplu de 3-configurație de start care poate determina 3-partiția dată, după 0 sau mai multe iterări ale algoritmului K-means.	Numărul de 3-configurații de start care generează această 3-partiție
1	$\{a, b, e\}, \{c, d\}, \{f\}$	Nu	—	0
2	$\{a, b\}, \{d, e\}, \{c, f\}$	Da	$\{b, c, e\}$	4
3	$\{a, d\}, \{b, e\}, \{c, f\}$	Da	$\{a, b, c\}$	8
4	$\{a\}, \{d\}, \{b, c, e, f\}$	Da	$\{a, b, d\}$	2
5	$\{a, b\}, \{d\}, \{c, e, f\}$	Da	$\{a, d, f\}$	1
6	$\{a, b, d\}, \{c\}, \{e, f\}$	Da	$\{a, c, f\}$	1

Observație: Restul de 5 configurații de start — în total, în varianta completă (vedeți coloana a cincea) a tabelului de mai sus sunt 15 configurații de start, dintre care una, $\{a, d, f\}$, apare de două ori — conduc la alte trei 3-partiții decât cele menționate în coloana a doua a tabelului.

Observație importantă: Este util de precizat acum — odată ce au fost „fixate“ noțiunile de K -partiție și K -configurație — că algoritmul K -means poate fi văzut ca un *algoritm de căutare*. K -means pornește de la o K -configurație (aleasă eventual în mod arbitrar) și o îmbunătățește în mod succesiv, prin intermediul K -partițiilor corespunzătoare, sau invers. *Spațiul de căutare* corespunzător algoritmului K -means este multimea tuturor K -partițiilor care se pot forma peste X , setul de instanțe de clusterizat.

12.

(*K-means, ca algoritm de optimizare a criteriului coeziunii intra-clustere* („suma celor mai mici pătrate“))

*prelucrare de Liviu Ciortuz, după
■ CMU, 2009 spring, Ziv Bar-Joseph, HW5, pr. 2.1*

Vă reamintim algoritmul *K-means*, datorat lui Lloyd,⁸²⁵ predat la curs:

Input: $x_1, \dots, x_n \in \mathbb{R}^d$ și $K \in \mathbb{N}^*$.

Output: o anumită *K*-partiție pentru $\{x_1, \dots, x_n\}$, adică o descompunere a acestei multimi într-o colecție de *K* multimi disjuncte (care nu sunt în mod neapărat nevide).

Procedură:

[Inițializare / Iterația 0:] $t \leftarrow 0$;

se fixează în mod arbitrar μ_1^0, \dots, μ_K^0 , centroizii inițiali ai clusterelor, și se asignează fiecare instantă x_i la centroidul cel mai apropiat, formând astfel clusterele C_1^0, \dots, C_K^0 .

[Corpul iterativ:] Se execută iterația $++t$:

Pasul 1: se calculează noile poziții ale centroizilor:⁸²⁶

$$\mu_j^t = \frac{1}{|C_j^{t-1}|} \sum_{x_i \in C_j^{t-1}} x_i \text{ pentru } j = \overline{1, K};$$

Pasul 2:

se reasignează fiecare x_i la [clusterul cu] centroidul cel mai apropiat, adică se stabilește noua componentă a clusterelor la iterarea t : C_1^t, \dots, C_K^t ;

[Terminare:] până când o anumită condiție este îndeplinită

(de exemplu: până când pozițiile centroizilor — sau: componentă clusterelor — nu se mai modifică de la o iterare la alta).

a. Demonstrați că, de la o iterare la alta, algoritmul *K-means* mărește *coezionea de ansamblu* a clusterelor. Veți proceda astfel: considerând funcția

$$J(C^t, \mu^t) \stackrel{\text{def.}}{=} \sum_{i=1}^n \|x_i - \mu_{C^t(x_i)}^t\|^2 \stackrel{\text{def.}}{=} \sum_{i=1}^n (x_i - \mu_{C^t(x_i)}^t) \cdot (x_i - \mu_{C^t(x_i)}^t),$$

unde

$C^t = (C_1^t, C_2^t, \dots, C_K^t)$ este colecția de clustere (i.e., *K*-partiția) la momentul t , $\mu^t = (\mu_1^t, \mu_2^t, \dots, \mu_K^t)$ este colecția de centroizi ai clusterelor (*K*-configurația) la momentul t ,

$C^t(x_i)$ desemnează clusterul la care este asignat elementul x_i la iterarea t , operatorul \cdot desemnează produsul scalar al vectorilor din \mathbb{R}^d ,

arătați că $J(C^t, \mu^t) \geq J(C^{t+1}, \mu^{t+1})$ pentru orice t .

Indicație: Inegalitatea de mai sus rezultă din două inegalități (care corespund pașilor 1 și 2 de la iterarea t):

$$J(C^t, \mu^t) \stackrel{(1)}{\geq} J(C^t, \mu^{t+1}) \stackrel{(2)}{\geq} J(C^{t+1}, \mu^{t+1}).$$

⁸²⁵Lloyd, S. P. (1957). "Least square quantization in PCM". Bell Telephone Laboratories Paper.

⁸²⁶Formula aceasta corespunde folosirii distanței euclidiene. Pentru alte măsuri de distanță, este posibil să fie necesare alte formule (vedeți problema 49.B).

La prima inegalitate (cea corespunzătoare pasului 1) se poate considera că parametrul C^t este fixat iar μ este variabil, în vreme ce la a doua inegalitate (cea corespunzătoare pasului 2) se consideră μ^t fixat și C variabil. Prima inegalitate se poate obține însumând o serie de inegalități, și anume câte una pentru fiecare cluster C_j^t : se demonstrează (de exemplu, cu ajutorul proprietăților derivatei) că $J(C_j^t, \mu) \geq J(C_j^t, \mu^{t+1})$ pentru $\forall \mu$, deci în particular și pentru μ^t . A doua inegalitate se demonstrează imediat.

b. Ce puteți spune despre oprirea algoritmului K -means? Termină oare acest algoritm într-un număr finit de pași, ori dimpotrivă — dat fiind faptul că există doar un număr finit de K -partiții ale mulțimii de instanțe $\{x_1, \dots, x_n\}$, și anume K^n — este posibil ca el să reviziteze de o infinitate de ori o K -configurație anterioară, $\mu = (\mu_1, \dots, \mu_K)$?

Răspuns:

a. Pentru conveniență, ne vom limita la cazul $d = 1$.⁸²⁷ Extinderea demonstrației la cazul $d > 1$ nu comportă dificultăți.⁸²⁸

Vom demonstra mai întâi inegalitatea (1): $J(C^t, \mu^t) \geq J(C^t, \mu^{t+1})$.

După definiție, $J(C^t, \mu^t) = \sum_{i=1}^n (x_i - \mu_{C^t(x_i)}^t)^2$.

Conform notației din enunț, C_j^t reprezintă clusterul j de la iterarea t , iar μ_j^t este centroidul corespunzător acestui cluster. Dacă notăm cu $x_{i_1}, x_{i_2}, \dots, x_{i_l}$ instanțele din compoziția clusterului C_j^t , unde $l \stackrel{\text{not.}}{=} |C_j^t|$, atunci (prințr-un ușor abuz de notație) vom putea scrie:

$$J(C_j^t, \mu_j^t) = \sum_{p=1}^l (x_{i_p} - \mu_j^t)^2,$$

iar $J(C^t, \mu^t)$ se va rescrie sub formă $J(C^t, \mu^t) = \sum_{j=1}^K J(C_j^t, \mu_j^t)$.

Dacă se consideră C_j^t fixat, iar μ_j^t variabil (vedeți pasul 1 al iterăției t), atunci putem minimiza funcția de gradul al doilea care „măsoară“ coeziunea din interiorul clusterului C_j^t :

$$f(\mu) \stackrel{\text{def.}}{=} J(C_j^t, \mu) = l\mu^2 - 2 \left(\sum_{p=1}^l x_{i_p} \right) \cdot \mu + \sum_{p=1}^l x_{i_p}^2$$

fie în mod direct (adică făcând apel la proprietățile funcției de gradul al doilea):

$$\arg \min_{\mu} J(C_j^t, \mu) = \frac{1}{l} \sum_{p=1}^l x_{i_p},$$

fie cu ajutorul derivatei vectoriale:⁸²⁹

$$\frac{\partial}{\partial \mu} f(\mu) = \sum_{p=1}^l (-2) (x_{i_p} - \mu) = -2 \left(\sum_{p=1}^l x_{i_p} - l\mu \right),$$

⁸²⁷Pentru o discuție interesantă asupra aplicării algoritmului K -means în cazul $d = 1$, vedeți problema 47.

⁸²⁸Se dezvoltă în mod corespunzător expresia lui J sau se folosesc derivatele sale parțiale de ordinul întâi.

⁸²⁹Vedeți formula (5g) din documentul *Matrix Identities* de Sam Roweis.

știind că punctul de minim al funcției $f(\mu) = J(C_j^t, \mu)$ se obține pentru rădăcina ecuației $\frac{\partial}{\partial \mu} f(\mu) = 0$, adică

$$\mu = \frac{1}{l} \sum_{p=1}^l x_{i_p} = \frac{1}{|C_j^t|} \sum_{p=1}^l x_{i_p} \stackrel{\text{def.}}{=} \mu_j^{t+1}.$$

Aceasta înseamnă că $J(C_j^t, \mu) \geq J(C_j^t, \mu_j^{t+1})$, pentru $\forall \mu$. În particular, pentru $\mu = \mu_j^t$ vom avea: $J(C_j^t, \mu_j^t) \geq J(C_j^t, \mu_j^{t+1})$. Inegalitatea aceasta este valabilă pentru toate clusterele $j = 1, \dots, K$. Dacă însumăm toate aceste inegalități, rezultă: $J(C^t, \mu^t) \geq J(C^t, \mu^{t+1})$.

Mai rămâne de demonstrat inegalitatea (2): $J(C^t, \mu^{t+1}) \geq J(C^{t+1}, \mu^{t+1})$, corespunzătoare *pasului 2* din algoritm.

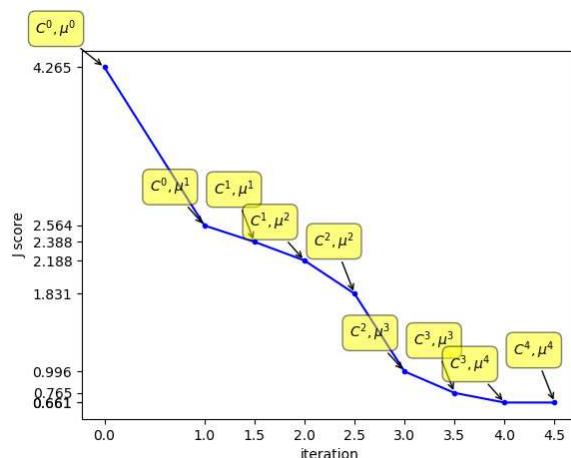
La acest pas, o instanță oarecare x_i , unde $i \in \{1, \dots, n\}$, este reasignată de la clusterul cu centroidul μ_j^{t+1} , la un alt centroid μ_q^{t+1} , dacă $\|x_i - \mu_j^{t+1}\|^2 \geq \|x_i - \mu_q^{t+1}\|^2$ pentru orice $j' = 1, \dots, K$. Această condiție este echivalentă cu următoarea: $(x_i - \mu_{j'}^{t+1})^2 \geq (x_i - \mu_q^{t+1})^2$ pentru orice j' . În contextul iterației t , acest lucru implică

$$(x_i - \mu_{C^t(x_i)}^{t+1})^2 \geq (x_i - \mu_{C^{t+1}(x_i)}^{t+1})^2.$$

Sumând membru cu membru inegalitățile de acest tip obținute pentru $i = \overline{1, n}$, rezultă: $J(C^t, \mu^{t+1}) \geq J(C^{t+1}, \mu^{t+1})$, ceea ce era de demonstrat.

Exemplu:

Pentru a ilustra grafic monotonia criteriului J , alăturat vă punem la dispoziție rezultatul obținut pe datele de la problema 10.a.⁸³⁰



b. Evident, numărul K -partițiilor care se pot forma cu cele n instanțe date este finit (K^n). A căuta minimul „criteriului” J se poate face parcurgând în mod exhaustiv multimea acestor K -partiții,⁸³¹ dar acest proces ar fi inefficient (sau, practic imposibil de realizat) pentru valori mari ale lui n . Algoritmul

⁸³⁰Graficul a fost realizat de către studentul Gheorghe Balan de la Universitatea „Al. I. Cuza“ din Iași în semestrul I al anului universitar 2017-2018.

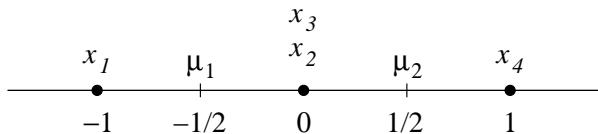
⁸³¹Este suficient să explorăm doar spațiul format de valorile primului argument al criteriului J deoarece, conform inegalității (1), pentru orice poziționare a centroizilor μ , valoarea $J(C, \mu)$ este minorată de $J(C, \mu')$ unde μ' reprezintă pozițiile recalulate ale centroizilor [adică, centrele de greutate ale] clusterelor din K -partiția C , după cum se procedează la pasul 1 al algoritmului K -means.

K-means explorează — pornind de la o anumită inițializare a celor K centroizi —, doar un subset de K -partiții, asigurându-ne însă că are loc proprietatea $J(C^0, \mu^1) \geq J(C^1, \mu^2) \geq \dots \geq J(C^{t-1}, \mu^t) \geq J(C^t, \mu^{t+1})$, conform punctului precedent al acestei probleme. (În inegalitatea multiplă de mai sus, μ^{i+1} constituie vectorul de centroizi, i.e., mediile instanțelor din clusterele care compun K -partiția C^i .)

Dacă algoritmul revizitează o K -partiție, atunci rezultă că pentru un anumit t avem $J(C^{t-1}, \mu^t) = J(C^t, \mu^{t+1})$. Este posibil ca acest fapt să se întâpte, și anume atunci când:

- există instanțe multiple (i.e., $x_i = x_j$, deși $i \neq j$),
- criteriul de oprire al algoritmului K -means este de forma „până când componența clusterelor nu se mai modifică“,
- se presupune că, în cazul în care o instanță x_i este situată la egală distanță față de doi sau mai mulți centroizi, ea poate fi asignată în mod aleatoriu la oricare dintre ei.

Așa se întâmplă în *exemplul* din figura următoare



dacă se consideră că la o iterare t avem $x_2 = 0 \in C_1^t$ și $x_3 = 0 \in C_2^t$, iar la iterare următoare alegem ca $x_3 = 0 \in C_1^{t+1}$ și $x_2 = 0 \in C_2^{t+1}$ și, din nou, invers la iterare $t + 2$.

Evident, dacă se impune restricția ca $x_2 = 0$ și $x_3 = 0$ să fie asignați la un același cluster,⁸³² atunci nu vom mai avea ciclare (iar $J(C, \mu)$ va avea în final valoarea minimă: $0 + \left(\frac{2}{3}\right)^2 + 2\left(\frac{1}{3}\right)^2 = \frac{2}{3}$.

Observația 1: Dacă se păstrează criteriul dat ca exemplu în enunțul problemei – adică se iterează până când centroizii „staționează“ – algoritmul se poate opri fără ca la ultima iterare $J(C, \mu)$ să fi atins minimul posibil. În cazul exemplului de mai sus, vom avea $\frac{1}{4} + 2 \cdot \frac{1}{4} + \frac{1}{4} = 1 > \frac{2}{3}$.

Observația 2: Dacă nu există instanțe multiple care să fie situate la distanțe egale față de doi sau mai mulți centroizi la o iterare oarecare a algoritmului K -means (precum sunt x_2 și x_3 în *exemplul* de mai sus), sau dacă se impune restricția ca în astfel de situații instanțele identice să fie asignate la un singur cluster, este evident că algoritmul K -means se oprește într-un număr finit de pași.

Observația 3: Atingerea minimului global al funcției $J(C, \mu)$ — unde C este o variabilă care parcurge multimea tuturor K -partiților care se pot forma cu instanțele $\{x_1, \dots, x_n\}$ — nu este garantată pentru algoritmul K -means. Valoarea funcției J care se obține la oprirea algoritmului K -means este dependentă de plasarea inițială a centroizilor μ , precum și de modul concret în care sunt

⁸³²O condiție mai generală poate fi următoarea: asignarea instanțelor x_i la centroizi să se facă astfel încât la pasul 1 inegalitatea (1) să fie satisfăcută în sens strict: $J(C^t, \mu^t) > J(C^t, \mu^{t+1})$.

alcătuite clusterelor în cazul în care o instanță oarecare se află la distanță egală de doi sau mai mulți centroizi, după cum am arătat în exemplul de mai sus.

Observația 4: Datorită rezultatului obținut la punctul *a* al acestui exercițiu, algoritmul de clusterizare *K-means* poate fi văzut [și reformulat] ca un *algoritm de optimizare*, folosind metoda *descreșterii pe coordonate*.⁸³³ Ca *input*, vom considera (ca și în pseudo-codul anterior) punctele $x_1, \dots, x_n \in \mathbb{R}^d$, împreună cu numărul natural $K \geq 1$. *Obiectivul* este acela de a minimiza o funcție obiectiv care măsoară (indirect) coeziunea intra-clustere. Atunci când se folosește distanța euclidiană, această funcție este

$$J(L, \mu) = \sum_{i=1}^n \|x_i - \mu_{l_i}\|^2,$$

unde $\mu = (\mu_1, \dots, \mu_K)$ sunt centroizii clusterelor ($\mu_j \in \mathbb{R}^d$), iar lista $L = (l_1, \dots, l_n)$ memorează asignările instanțelor x_i la clustere ($l_i \in \{1, \dots, K\}$).⁸³⁴

Algoritmul *K-means* face *initializarea* centroizilor clusterelor μ cu anumite valori, după care procedează astfel:

Pasul 1: Păstrând μ fixat, găsește acea asignare L a instanțelor la clustere care minimizează funcția $J(L, \mu)$;⁸³⁵

Pasul 2: Păstrând asignarea L fixată, găsește acea valoare pentru μ pentru care se minimizează $J(L, \mu)$.⁸³⁶

Criteriul de oprire: Dacă [aceasta nu este prima iteratie și] niciuna dintre asignările din lista L nu s-a modificat în raport cu precedenta iteratie, se trece la pasul următor (Terminare); altfel se repetă de la Pasul 1.

Terminare: Returnează L și μ .

13.

(Algoritmul *K-means*: o proprietate de [anti]monotonie a valorilor minime pentru criteriul „sumei celor mai mici pătrate“ (J) în funcție de K)

- ○ CMU, 2012 fall, E. Xing, A. Singh, HW3, pr. 1.1.d
CMU, 2010 fall, Aarti Singh, HW3, pr. 5.4

Cum evoluează valoarea minimă a funcției obiectiv J , așa-numita sumă a celor mai mici pătrate (vedeți problema 12), la execuția algoritmului *K-means* pe un set de date arbitrar ales (dar fixat) atunci când valoarea lui K crește de la 1 la n : crește, scade, rămâne constantă, variază arbitrar ori conform unei anumite legi / proprietăți? Justificați în mod riguros.

⁸³³Ceea ce urmează în continuare este preluat — cu ușoare adaptări — din CMU, 2012 fall, E. Xing, A. Singh, HW3, pr. 1.

⁸³⁴Găsirea minimului global al acestei funcții este un task *dificil* din punct de vedere computațional.

⁸³⁵Aceasta revine la a găsi pentru fiecare instanță x_i care este cel mai apropiat centroid. În cazul în care instanța se află la distanță minimă față de mai mulți centroizi, se pot folosi diverse euristici pentru a decide cum anume se face asignarea.

⁸³⁶Acesta este un pas simplu, care revine — în cazul în care se folosește distanța euclidiană — la a calcula pentru fiecare cluster media instanțelor (văzute ca vectori) care alcătuiesc clusterul respectiv.

Indicație: În răspunsul pe care-l veți da, veți nota cu $X = \{x_1, \dots, x_n\}$ setul de instanțe de clusterizat și cu $\underline{J}_K(X)$ valoarea minimă a criteriului J pe mulțimea tuturor K -partițiilor C ale lui X . Formal, putem scrie astfel:

$$\begin{aligned}\underline{J}_K(X) &= \min\{J_K(X, C) \mid C \text{ este } K\text{-partiție a lui } X\} \\ \text{cu } J_K(X, C) &= \sum_{i=1}^n \|x_i - \mu_{C(x_i)}\|^2,\end{aligned}$$

unde $\mu_{C(x_i)}$ este centroidul clusterului $C(x_i)$, la care este asignat x_i în K -partiția (adică setul de clustere) C . Conform problemei 12.a, atunci când se folosește distanța euclidiană se consideră $\mu_{C(x_i)} = \frac{1}{|C(x_i)|} \sum_{x_j \in C(x_i)} x_j$.

Răspuns:

Este imediat că orice K -partiție C a lui X poate fi pusă în mod natural în corespondență cu o $(K+1)$ -partiție C' a lui X (să o notăm pe aceasta din urmă cu $(C_1, \dots, C_K, C_{K+1})$), în care un cluster (de exemplu C_{K+1}) este vid, iar celelalte clustere sunt în corespondență bijectivă de tipul funcției *identitate* cu clusterele K -partiției C . Evident, $J_K(X, C) = J_K(X, C')$.

În acest fel, mulțimea tuturor K -partițiilor lui X (notată cu \mathcal{P}) este pusă în corespondență bijectivă cu o *submulțime* din mulțimea tuturor $(K+1)$ -partițiilor lui X (ultima, notată cu \mathcal{P}').⁸³⁷ Așadar, putem spune că într-un anumit sens, mulțimea \mathcal{P} este mai amplă decât mulțimea \mathcal{P}' . Rezultă că

$$\underline{J}_{K+1}(X) \leq \underline{J}_K(X), \text{ pentru orice } K = 1, \dots, n-1.$$

Aceasta este o proprietate de *anti-monotonie* a valorilor minime pentru criteriul „sumei celor mai mici pătrate“ (J) în funcție de K .

Altă variantă de demonstrație:

Se poate vedea că orice K -partiție poate fi pusă în corespondență cu o $(K+1)$ -partiție al cărei scor (ca „sumă a pătratelor erorilor“) este mai mic sau egal cu cel al K -partiției.

În cazul $K < n$ (unde n este numărul de instanțe de clusterizat), este clar că există o instanță care poate fi extrasă din clusterul ei ca să devină [ea însăși] un cluster singleton. Clusterul din care a provenit instanța respectivă va avea acum valoarea „sumei pătratelor erorilor“ mai mică decât [sau egală cu] vechea valoare (ceea ce se justifică imediat, fiindcă centroidul clusterului se va muta în poziția optimă, adică poziția care corespunde minimului „sumei pătratelor erorilor“ pentru clusterul care acum are cu o instanță mai puțin decât avea înainte), iar clusterul singleton va avea valoarea „sumei pătratelor erorilor“ 0. În cazul $K = n$, toate $(K+1)$ -partițiile vor avea un cluster vid, în consecință, valoarea minimă a criteriului J_{K+1} este aceeași cu minimul lui J_K .

⁸³⁷ Matematicienii spun că mulțimea \mathcal{P} este „scufundată“ în mulțimea \mathcal{P}' .

14. (Clusterizarea ierarhică și algoritmul K -means: comparații)

CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, midterm, pr. 4.b

Enunțați câteva avantaje ale clusterizării ierarhice în raport cu clusterizarea bazată pe algoritmul K -means. Similar, enunțați câteva avantaje ale clusterizării bazate pe algoritmul K -means în raport cu clusterizarea ierarhică.

Răspuns:

Iată câteva avantaje ale clusterizării ierarhice:

- nu necesită fixarea dinainte a unui anumit număr (K) de clustere pe care dorim să le obținem;
- arborele ierarhic obținut poate fi tăiat la orice nivel, pentru a obține câte clustere dorim;
- pentru multe aplicații, ierarhia obținută este ușor de interpretat;
- se poate lucra cu date dispuse spațial într-o formă elongată (engl., long stringy data).

Și câteva avantaje ale clusterizării bazate pe algoritmul K -means:

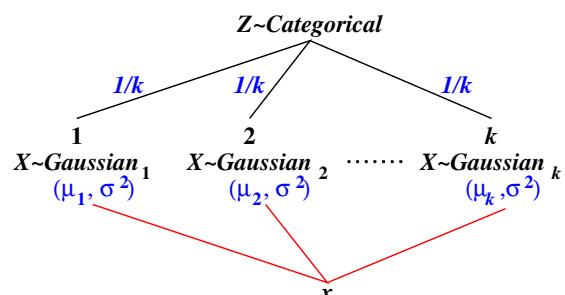
- pe anumite seturi de date, rezultatul poate fi obținut mult mai rapid decât clusterizarea ierarhică;
- beneficiază de un cadru teoretic elegant;
- se pot incorpora ușor date noi și, de asemenea, este posibilă reactualizarea clusterelor în mod facil;
- poate fi kernel-izat (vedeți ex. 51).

7.1.3 Algoritmul EM pentru modele de mixturi gaussiene

15. (Algoritmul EM pentru GMM, cazul unidimensional, cu $\sigma_1 = \dots = \sigma_k$ și $\pi_1 = \dots = \pi_k$: demonstrația formulelor de „actualizare“)

Machine Learning, Tom Mitchell, 1997, pag. 193, 195–196.

Ne propunem să facem estimarea parametrilor unui model de mixtură de k distribuții gaussiene (engl., Gaussian Mixture Model, GMM), despre care se presupune că au [toate] aceeași varianță σ^2 , iar probabilitățile a priori de selecție sunt egale ($1/k$).



Prin urmare, se vor estima doar mediile acestor k distribuții gaussiene. Pentru aceasta, vom considera următorul *algoritm EM*:

Inițializare: Date fiind instanțele $x_1, \dots, x_n \in \mathbb{R}$, se aleg în mod arbitrar valori pentru mediile $\mu \stackrel{\text{not.}}{=} (\mu_1, \dots, \mu_k)$. Vom nota aceste valori inițiale cu $\mu_1^{(0)}, \dots, \mu_k^{(0)}$. Pentru $t = 0, 1, \dots$, atâtă timp cât nu este îndeplinit criteriul de oprire,⁸³⁸ execută:

Pasul E:

$$\begin{aligned} E[Z_{ij}] &\stackrel{\text{not.}}{=} E[Z_{ij}|X = x_i; \mu^{(t)}, \sigma^2] = \dots = P(Z_{ij} = 1|X = x_i; \mu^{(t)}, \sigma^2) \\ &\stackrel{F. Bayes}{=} \dots \\ &\stackrel{p.d.f.}{=} \dots \quad \text{pentru } i = 1, \dots, n \text{ și } j = 1, \dots, k. \end{aligned}$$

Pasul M:

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^n E[Z_{ij}]x_i}{\sum_{i=1}^n E[Z_{ij}]}, \quad \text{pentru } j = 1, \dots, k, \quad (345)$$

unde Z_{ij} , cu $i \in \{1, \dots, n\}$ și $j \in \{1, \dots, k\}$, sunt variabile-indicator neobservabile (sau „ascunse“ / „latente“), luând valoarea 1 dacă x_i a fost generat de gaussiana j și 0 în caz contrar.

Elaborați formula de la pasul E și demonstrați formula (345) pentru pasul M.

Sugestie: Pentru pasul M,

- veți scrie mai întâi $p(y_i|\mu)$ verosimilitatea unei instanțe „complete“, $y_i \stackrel{\text{not.}}{=} (x_i, z_{i1}, \dots, z_{ik})$;⁸³⁹
- apoi veți scrie $\ln P(Y|\mu)$, funcția de log-verosimilitate a datelor complete $Y \stackrel{\text{not.}}{=} \{y_1, \dots, y_n\}$;
- veți calcula funcția „auxiliară“ $Q(\mu|\mu^{(t)}) \stackrel{\text{def.}}{=} E[\ln P(Y|\mu)]$, care este media funcției de log-verosimilitate a datelor complete în raport cu distribuția probabilistă a posteriori a variabilelor neobservabile Z_{ij} în raport cu datele observabile x_i și cu valorile parametrilor μ de la iterată t , adică $\mu^{(t)}$;
- în final, veți calcula valorile mediilor μ_j pentru care se obține valoarea optimă a funcției „auxiliare“ $Q(\mu|\mu^{(t)})$.

Răspuns:

Pasul E:

Așa cum a fost sugerat în enunț, mai întâi arătăm că media variabilei-indicator Z_{ij} (văzută ca variabilă aleatoare) este $P(Z_{ij} = 1 | x_i, \mu^{(t)})$.⁸⁴⁰ Apoi calculăm această probabilitate condiționată folosind formula lui Bayes combinată cu formula probabilității totale, iar la final ținem cont de expresia funcției de densitate pentru distribuția gaussiană.⁸⁴¹

$$E[Z_{ij}] \stackrel{\text{def.}}{=} 0 \cdot P(Z_{ij} = 0 | x_i, \mu^{(t)}) + 1 \cdot P(Z_{ij} = 1 | x_i, \mu^{(t)}) = P(Z_{ij} = 1 | x_i, \mu^{(t)})$$

⁸³⁸De exemplu, se execută un anumit număr de iterării, fixat în prealabil.

⁸³⁹Pentru un exemplu simplu de definire și calculare a funcției de verosimilitate, vedeți ex. 42 de la capitolul *Fundamente*.

⁸⁴⁰Pentru o formulare generală a acestei proprietăți, vedeți problema 98 de la capitolul de *Fundamente*.

⁸⁴¹Vedeți problema 32 de la capitolul de *Fundamente*.

$$\begin{aligned}
T.Bayes & \stackrel{P(x_i | Z_{ij} = 1; \mu^{(t)}) \cdot \overbrace{P(Z_{ij} = 1 | \mu^{(t)})}^{1/k}}{\sum_{l=1}^k P(x_i | Z_{il} = 1; \mu^{(t)}) \cdot \underbrace{P(Z_{il} = 1 | \mu^{(t)})}_{1/k}} = \frac{\mathcal{N}(X = x_i | \mu = \mu_j^{(t)})}{\sum_{l=1}^k \mathcal{N}(X = x_i | \mu = \mu_l^{(t)})} \\
& \stackrel{p.d.f.}{=} \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x_i - \mu_j^{(t)})^2}}{\sum_{l=1}^k \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x_i - \mu_l^{(t)})^2}} = \frac{e^{-\frac{1}{2\sigma^2}(x_i - \mu_j^{(t)})^2}}{\sum_{l=1}^k e^{-\frac{1}{2\sigma^2}(x_i - \mu_l^{(t)})^2}}.
\end{aligned}$$

Vă readucem aminte că probabilitățile a priori $P(Z_{il} = 1 | \mu^{(t)})$ au fost considerate de la început identice, indiferent de valoarea lui l .

Pasul M:

Pentru a calcula probabilitatea unei date complete $y_i \stackrel{not.}{=} (x_i, z_{i1}, \dots, z_{ik})$, vom folosi regula de înmulțire a probabilităților și, din nou, expresia funcției de densitate pentru distribuția gaussiană.

$$\begin{aligned}
p(y_i | \mu) & \stackrel{not.}{=} p(x_i, z_{i1}, \dots, z_{ik} | \mu) = p(x_i | z_{i1}, \dots, z_{ik}; \mu) \underbrace{p(z_{i1}, \dots, z_{ik} | \mu)}_{1/k} \\
& = \frac{1}{k} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2} \sum_{j=1}^k z_{ij}(x_i - \mu_j)^2}
\end{aligned}$$

Funcția de log-verosimilitate a datelor complete $Y = \{y_1, \dots, y_n\}$ se calculează ușor, ținând cont de independența creării datelor x_i , precum și de rezultatul obținut anterior.

$$\begin{aligned}
\ln P(Y | \mu) & \stackrel{i.i.d.}{=} \ln \prod_{i=1}^n p(y_i | \mu) = \sum_{i=1}^n \ln p(y_i | \mu) \\
& = \sum_{i=1}^n (-\ln k + \ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{j=1}^k z_{ij}(x_i - \mu_j)^2).
\end{aligned}$$

Observați introducerea în ultima expresie a variabilelor-indicator Z_{ij} . Se ține cont de faptul că pentru fiecare $i = 1, \dots, n$, avem $\sum_{j=1}^k Z_{ij} = 1$, deci doar una dintre aceste variabile-indicator are valoarea 1 (și anume, cea care corespunde gaussienei care a generat instanța x_i), în vreme ce toate celelalte variabile-indicator au valoarea 0.

Media funcției de log-verosimilitate a datelor complete (aşa-numita *funcție auxiliară*) se obține aplicând proprietatea de *liniaritate a mediei*.⁸⁴²

$$Q(\mu | \mu^{(t)}) = E[\ln P(Y | \mu)] \stackrel{lin.}{=} \sum_{i=1}^n \left(-\ln k + \ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{j=1}^k E[Z_{ij}](x_i - \mu_j)^2 \right). \quad (346)$$

Se poate observa că aceasta este o funcție de gradul al doilea în raport cu variabilele μ_j .⁸⁴³ Se constată apoi că optimul ei se poate obține prin maximizare separată (sau, pe rând) în raport cu fiecare variabilă μ_j .

⁸⁴²Vedeți problema 9.a de la capitolul de *Fundamente*.

⁸⁴³Așadar, pentru $k = 2$ graficul unei astfel de funcții este un *paraboloid*.

$$\begin{aligned}
Q(\mu|\mu^{(t)}) &= n(-\ln k + \ln \frac{1}{\sqrt{2\pi}\sigma}) - \frac{1}{2\sigma^2} \sum_{i=1}^n E[Z_{ij}] \sum_{j=1}^k E[Z_{ij}](x_i - \mu_j)^2 \\
&= n(-\ln k + \ln \frac{1}{\sqrt{2\pi}\sigma}) - \frac{1}{2\sigma^2} \sum_{j=1}^k \sum_{i=1}^n E[Z_{ij}](x_i - \mu_j)^2 \\
&= \text{const} - \frac{1}{2\sigma^2} \sum_{j=1}^k \sum_{i=1}^n E[Z_{ij}](x_i^2 - 2\mu_j x_i + \mu_j^2) \\
&= \text{const}' - \frac{1}{2\sigma^2} \sum_{j=1}^k \left[\left(\sum_{i=1}^n E[Z_{ij}] \right) \mu_j^2 - 2 \left(\sum_{i=1}^n E[Z_{ij}] x_i \right) \mu_j \right].
\end{aligned}$$

Pentru fiecare $j = 1, \dots, k$, coeficientul dominant al lui μ_j este strict pozitiv (fiindcă $E[Z_{ij}] = P(Z_{ij} = 1 | x_i, \mu^{(t)}) > 0$, conform rezultatului de la pasul E). Așadar, soluția optimă este dată de relația

$$\mu_j^{(t+1)} \stackrel{\text{def.}}{=} \underset{\mu_j}{\operatorname{argmax}} Q(\mu|\mu^{(t)}) = -\frac{-2\frac{1}{2\sigma^2} \sum_{i=1}^n E[Z_{ij}] x_i}{\frac{1}{2\sigma^2} \sum_{i=1}^n E[Z_{ij}]} = \frac{\sum_{i=1}^n E[Z_{ij}] x_i}{\sum_{i=1}^n E[Z_{ij}]}.$$

Observație: Rezultatul acesta (în enunț, formula (345)) reprezintă media ponderată a instanțelor x_i , $i = 1, \dots, n$. Așadar, această formulă reprezintă o generalizare (sau, o variantă ponderată) a formulei pentru estimarea mediei (în sens MLE) pentru o distribuție gaussiană unidimensională (vedeți rezolvarea problemei 50.a de la capitolul *Fundamente*).

16. (Algoritmul EM/GMM, cazul $\sigma_1^2 = \sigma_2^2 = 1$, $\pi_1 = \pi_2 = 1/2$: executarea unei iterări)

*prelucrare de Liviu Ciortuz, după
■ • ○ CMU, 2012 spring, Ziv Bar-Joseph, final exam, pr. 3.1*

Fie un model de mixtură gaussiană (engl., Gaussian mixture model, GMM) cu două componente având varianțe cunoscute și probabilități *a priori* egale pentru selecția celor două distribuții:

$$\frac{1}{2} \mathcal{N}(x|\mu_1, 1) + \frac{1}{2} \mathcal{N}(x|\mu_2, 1), \quad x \in \mathbb{R}.$$

În continuare se va considera că $n = 2$, $x_1 = 0.5$ și $x_2 = 2$, iar valorile inițiale pentru μ_1 și μ_2 sunt 1 și respectiv 2.

Execuți în mod manual o iterare a algoritmului EM din enunțul problemei 15 (preluat din carte Machine Learning a lui Tom Mitchell, pag. 193) pe aceste date, astfel:

- a. (Pasul E) Calculați mai întâi $P(Z_{ij} = 1 | x_i, \mu)$, probabilitățile a posteriori de apartenență a datelor observate (x_1 și x_2) la cele două componente ale mixturii. Am folosit notația $\mu = (\mu_1, \mu_2)$, iar Z_{ij} este variabilă-indicator având valoarea 1 dacă instanța x_i a fost generată de gaussiană j și 0 în cazul contrar.

Indicație: În vederea efectuării calculelor, pentru conveniență puteți considera valorile distribuției normale / gaussiene standard $\mathcal{N}(x|\mu = 0, \sigma^2 = 1)$ în punctele 0, 0.5, 1, 1.5 și 2 ca fiind respectiv 0.4, 0.35, 0.24, 0.13 și 0.05.

b. (Pasul M) Re-calculați valorile parametrilor μ_1 și μ_2 în funcție de probabilitățile calculate la punctul precedent. Care credeți că va fi tendința de mișcare a mediilor la următoarele iterații?

c. Funcția de log-verosimilitate a datelor observabile este definită aşa cum este de așteptat, ca logaritmul unei funcții de probabilitate (marginală în raport cu distribuția comună $P(x, z|\mu)$):

$$\ell(\mu_1, \mu_2) \stackrel{\text{def.}}{=} \ln P(x_1, x_2|\mu_1, \mu_2) \stackrel{\text{indep.}}{=} \sum_{i=1}^2 \ln P(x_i|\mu_1, \mu_2) = \sum_{i=1}^2 \ln \left(\sum_{z_{ij}} P(x_i, z_{ij}|\mu_1, \mu_2) \right),$$

unde simbolul $\sum_{z_{ij}}$ desemnează parcurgerea tuturor asignărilor posibile pentru variabilele neobservabile Z_{ij} (adică, mai întâi $z_{11} = 1$ și $z_{21} = 1$, apoi $z_{11} = 1$ și $z_{21} = 0$, după care $z_{11} = 0$ și $z_{21} = 1$ și, în final, $z_{11} = 0$ și $z_{21} = 0$). Explicați expresia acestei funcții (păstrând μ_1 și μ_2 nespecificați). După aceea, calculați valoarea ei la începutul și respectiv la sfârșitul primei iterații a algoritmului EM pe datele x_1 și x_2 specificate în prima parte a acestui exercițiu.

Răspuns:

a. Folosind teorema lui Bayes, vom exprima probabilitățile de apartenență ale instanțelor x_1 și respectiv x_2 la clusterul reprezentat de prima gaussiană:

$$\begin{aligned} P(Z_{i1} = 1|x_i, \mu) &\stackrel{T.B.}{=} \frac{P(x_i|Z_{i1} = 1, \mu_1)P(Z_{i1} = 1)}{P(x_i|Z_{i1} = 1, \mu_1)P(Z_{i1} = 1) + P(x_i|Z_{i2} = 1, \mu_2)P(Z_{i2} = 1)} \\ &= \frac{P(x_i|Z_{i1} = 1, \mu_1) \cdot \frac{1}{2}}{P(x_i|Z_{i1} = 1, \mu_1) \cdot \frac{1}{2} + P(x_i|Z_{i2} = 1, \mu_2) \cdot \frac{1}{2}} \\ &= \frac{P(x_i|Z_{i1} = 1, \mu_1)}{P(x_i|Z_{i1} = 1, \mu_1) + P(x_i|Z_{i2} = 1, \mu_2)} \text{ pentru } i \in \{1, 2\}. \end{aligned}$$

Tinând cont de faptul că valorile oricărei *distribuții gaussiene unidimensionale* pot fi puse în corespondență cu valorile *distribuției gaussiene standard*,⁸⁴⁴ utilizând valorile furnizate în *indicația* din enunț și folosind proprietatea de simetrie a valorilor distribuției gaussiene standard în raport cu originea, vom avea:

$$\begin{aligned} P(Z_{11} = 1|x_1, \mu) &= \frac{\mathcal{N}(0.5|1, 1)}{\mathcal{N}(0.5|1, 1) + \mathcal{N}(0.5|2, 1)} = \frac{\mathcal{N}(0.5|0, 1)}{\mathcal{N}(0.5|0, 1) + \mathcal{N}(1.5|0, 1)} \\ &= \frac{0.35}{0.35 + 0.13} = \frac{0.35}{0.48} = \frac{35}{48} \\ P(Z_{21} = 1|x_2, \mu) &= \frac{\mathcal{N}(2|1, 1)}{\mathcal{N}(2|1, 1) + \mathcal{N}(2|2, 1)} = \frac{\mathcal{N}(1|0, 1)}{\mathcal{N}(1|0, 1) + \mathcal{N}(0|0, 1)} \\ &= \frac{0.24}{0.24 + 0.4} = \frac{0.24}{0.64} = \frac{3}{8} \end{aligned}$$

⁸⁴⁴Vom folosi formula pentru „standardizare“ $\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sigma} \mathcal{N}\left(\frac{x-\mu}{\sigma}|0, 1\right)$, care este ușor de demonstrat, precum și proprietatea $\mathcal{N}(x|0, 1) = \mathcal{N}(-x|0, 1)$.

În sfârșit, datorită faptului că $Z_{i1} + Z_{i2} = 1$ pentru $\forall i \in \{1, 2\}$ (cu $Z_{ij} \in \{0, 1\}$) și, de asemenea, datorită complementarității evenimentelor aleatoare, vom obține imediat și celelalte două probabilități cerute:

$$\begin{aligned} P(Z_{12} = 1|x_1, \mu) &= P(Z_{11} = 0|x_1, \mu) = 1 - P(Z_{11} = 1|x_1, \mu_1) = \frac{13}{48} \\ P(Z_{22} = 1|x_2, \mu) &= P(Z_{21} = 0|x_2, \mu) = 1 - P(Z_{21} = 1|x_2, \mu_1) = \frac{5}{8} \end{aligned}$$

b. Formulele (345) pentru actualizarea mediilor μ_j pot fi scrise folosind notațiile de aici (adică, explicitând direct mediile variabilelor-indicator Z_{ij}), astfel:⁸⁴⁵

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^2 P(Z_{ij} = 1|x_i, \mu^{(t)}) x_i}{\sum_{i=1}^2 P(Z_{ij} = 1|x_i, \mu^{(t)})}$$

Prin urmare,

$$\mu_1^{(1)} = \frac{\frac{35}{48} \cdot 0.5 + \frac{3}{8} \cdot 2}{\frac{35}{48} + \frac{3}{8}} = \frac{107}{106} \approx 1.009 \text{ și } \mu_2^{(1)} = \frac{\frac{13}{48} \cdot 0.5 + \frac{5}{8} \cdot 2}{\frac{13}{48} + \frac{5}{8}} = \frac{133}{86} \approx 1.547$$

Se observă că, în raport cu pozițiile inițiale, mediile celor două gaussiene au „migrat“ astfel: μ_1 foarte puțin spre dreapta, iar μ_2 considerabil mai la stânga. În noile poziții, mediile $\mu_1^{(1)}$ și $\mu_2^{(1)}$ sunt aproape simetrice în raport cu mijlocul segmentului determinat de punctele $x_1 = 0.5$ și $x_2 = 2$. Aceasta implică faptul că la pasul următor vom avea $P(Z_{11} = 1|x_1, \mu^{(1)}) \approx P(Z_{22} = 1|x_2, \mu^{(1)})$ și $P(Z_{12} = 1|x_1, \mu^{(1)}) \approx P(Z_{21} = 1|x_2, \mu^{(1)})$. Executând încă o iterație (nu dăm aici detaliile) obținem $\mu_1^{(2)} \approx 1.110$ și $\mu_2^{(2)} \approx 1.390$. Ne așteptăm ca pe parcursul următoarelor iterări mediile $\mu_1^{(t)}$ și $\mu_2^{(t)}$ să migreze către mijlocul intervalului $[0.5, 2]$.

c. Vom face mai întâi explicitarea expresiei care a fost dată în enunț pentru funcția de verosimilitate ℓ , specificând apoi termenii ei de bază într-un tabel:

$$\begin{aligned} \ell(\mu_1, \mu_2) &= \sum_{i=1}^2 \ln \left(\sum_{z_{ij}} P(x_i, Z_{ij} = z_{ij} | \mu_1, \mu_2) \right) \\ &= \sum_{i=1}^2 \ln \left(\sum_{z_{ij}} P(x_i | z_{ij}, \mu_1, \mu_2) \cdot \underbrace{P(z_{ij} | \mu_1, \mu_2)}_{1/2} \right) \\ &= \sum_{i=1}^2 \ln \left(\frac{1}{2} \sum_{z_{ij}} P(x_i | z_{ij}, \mu_1, \mu_2) \right) \\ &= \sum_{i=1}^2 \left[-\ln 2 + \ln \left(\sum_{j=1}^2 P(x_i | z_{ij} = 1, \mu_j) \right) \right] \end{aligned}$$

⁸⁴⁵Expresia funcției auxiliare $Q(\mu, \mu^{(t)})$, al cărei punct de optim are abscisa $\mu^{(t+1)} \stackrel{\text{not.}}{=} (\mu_1^{(t+1)}, \mu_2^{(t+1)})$, se obține — pentru acest caz particular — din formula (346) de la problema 15.

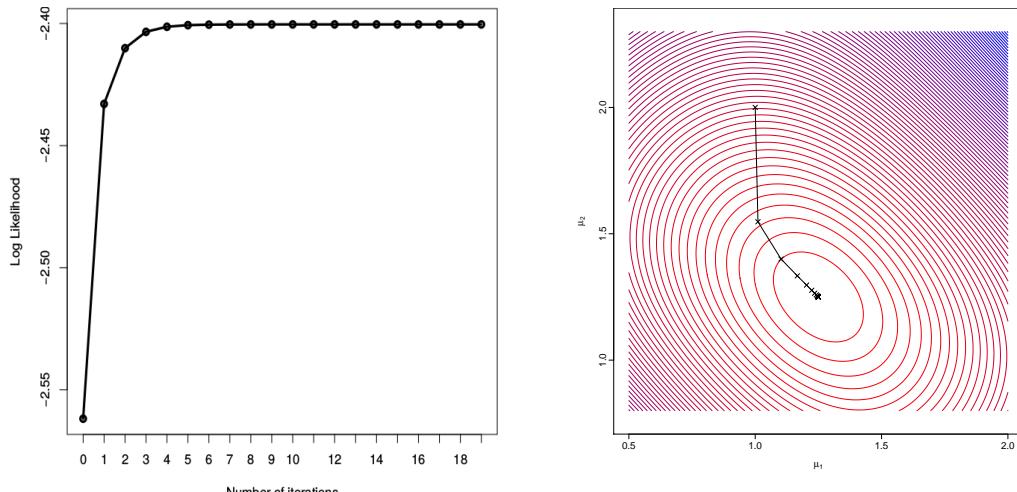
z_{ij}	$P(x_i z_{ij}, \mu_1, \mu_2)$
$z_{11} = 1, z_{12} = 0$	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_1 - \mu_1)^2\right)$
$z_{11} = 0, z_{12} = 1$	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_1 - \mu_2)^2\right)$
$z_{21} = 1, z_{22} = 0$	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_2 - \mu_1)^2\right)$
$z_{21} = 0, z_{22} = 1$	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_2 - \mu_2)^2\right)$

Prin urmare,

$$\begin{aligned}\ell(\mu_1, \mu_2) &= -2 \ln 2 + \ln \left(\frac{1}{\sqrt{2\pi}} \left(\exp\left(-\frac{1}{2}(x_1 - \mu_1)^2\right) + \exp\left(-\frac{1}{2}(x_1 - \mu_2)^2\right) \right) \right) \\ &\quad + \ln \left(\frac{1}{\sqrt{2\pi}} \left(\exp\left(-\frac{1}{2}(x_2 - \mu_1)^2\right) + \exp\left(-\frac{1}{2}(x_2 - \mu_2)^2\right) \right) \right) \\ &= -2 \ln 2 - \ln(2\pi) + \ln \left(\exp\left(-\frac{1}{2}(x_1 - \mu_1)^2\right) + \exp\left(-\frac{1}{2}(x_1 - \mu_2)^2\right) \right) \\ &\quad + \ln \left(\exp\left(-\frac{1}{2}(x_2 - \mu_1)^2\right) + \exp\left(-\frac{1}{2}(x_2 - \mu_2)^2\right) \right)\end{aligned}$$

Valoarea acestei funcții ca urmare a inițializării mediilor μ_1 și μ_2 cu valorile 1 și respectiv 2 este -2.561833 , iar la finalul primei iterații a algoritmului EM ea devine -2.462877 . Se observă că valoarea funcției de log-verosimilitate crește, așa cum era de așteptat (vedeți problema 2 de la capitolul *Schema algoritmică EM*).

Pentru [confirmarea și] extinderea acestui rezultat, prezentăm mai jos rezultatele obținute cu ajutorul unei implementări a algoritmului EM.⁸⁴⁶

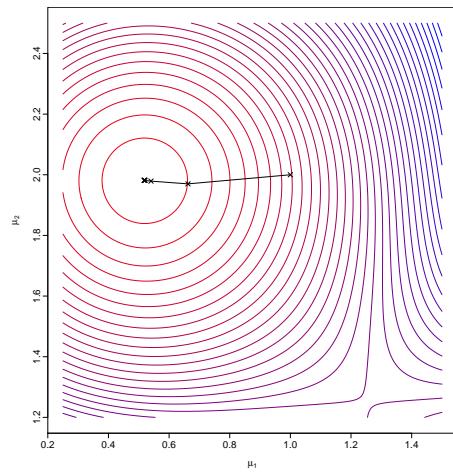


- în graficul din partea stângă sunt reprezentate valorile funcției de log-verosimilitate obținute de EM la inițializare și apoi la finalul fiecăreia din primele 19 iterații,
- în graficul din partea dreaptă avem reprezentarea sub formă de curbe de izocontur a valorilor log-verosimilității în funcție de cei doi parametri, μ_1 și

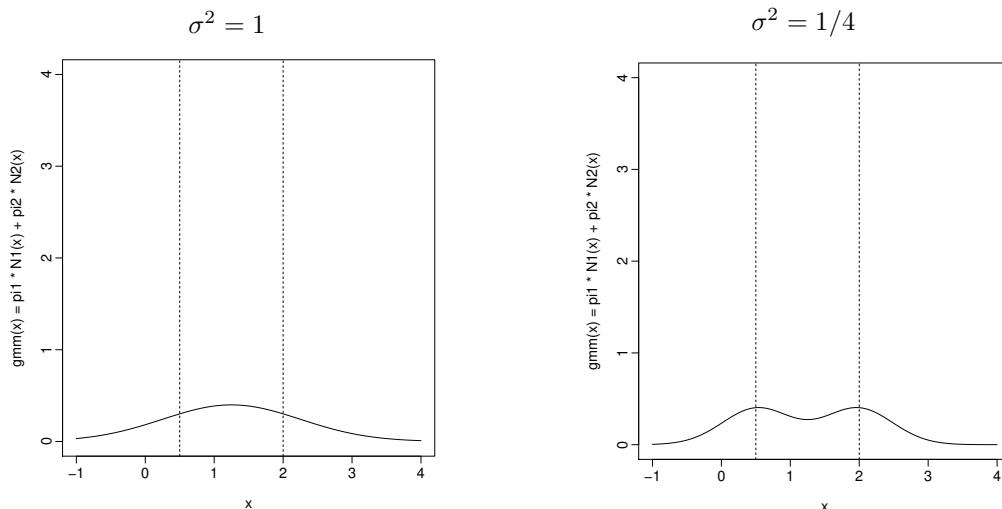
⁸⁴⁶Această implementare a fost realizată în 2018 de către Sebastian Ciobanu.

μ_2 . Pe acest al doilea grafic a fost adăugat un „drum“ care pune în evidență succesiunea de valori pentru perechile (μ_1, μ_2) de-a lungul iterațiilor executate de algoritmul EM.

Observație (1): Se poate vedea în figura de mai sus, partea dreaptă, că mediile μ_1 și μ_2 vor tinde la convergență la aceeași valoare, 1.25, valoare care reprezintă jumătatea intervalului $[0.5, 2]$. Este interesant de constatat că dacă am fi lucrat cu valori mici ale varianței (de exemplu, $\sigma^2 = 1/4$, caz ilustrat în figura alăturată), atunci la convergență μ_1 ar fi tins la valoarea $x_1 = 0.5$, în vreme ce μ_2 ar fi tins la valoarea $x_2 = 2$. La problema 64 se va vedea că în general, atunci când $\sigma^2 \rightarrow 0$, algoritmul EM/GMM tinde să obțină (la convergență) aceleasi rezultate ca și algoritmul K-means!



Observație (2): Modelele probabiliste obținute „învățate“ în urma aplicării algoritmului EM pentru cele două cazuri ($\sigma^2 = 1$ și $\sigma^2 = 1/4$) sunt prezentate în graficele următoare.⁸⁴⁷ În primul caz se poate observa că mixtura obținută coincide practic cu o distribuție gaussiană a cărei medie este situată la jumătatea intervalului determinat de cele două instanțe, x_1 și x_2 . În cel de-al doilea caz se poate constata că mediile celor două gaussiene care compun mixtura coincid cu (de fapt, sunt situate foarte aproape de) cele două instanțe, x_1 și x_2 .



⁸⁴⁷ Aceste grafice, ca și cel precedent, au fost realizate de Sebastian Ciobanu.

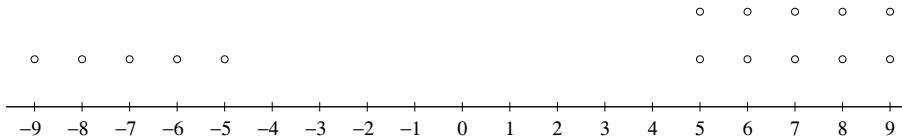
17. (Algoritmii K -means și EM pentru GMM, cazul unidimensional: aplicare)

prelucrare făcută de L. Ciortuz și A. Munteanu, după Edinburgh, 2009 fall, C. Williams, V. Lavrenko, HW4, pr. 3

Se consideră următorul set de instanțe – puncte în spațiul unidimensional:

$$-9, -8, -7, -6, -5, 5, 5, 6, 6, 7, 7, 8, 8, 9, 9$$

Aceste date au reprezentarea următoare:



Obiectivul este să se aplice pe de o parte algoritmul K -means și pe de altă parte algoritmul EM pentru un model de mixtură de distribuții gaussiene (GMM) ca să împărțim acest set de date în două clustere. Drept centroizi inițiali se consideră punctele -20 și -10 . Este evident că această alegere este una nefavorabilă / nestandard, însă dorim să comparăm evoluțiile celor doi algoritmi în acest caz.

a. Să se elaboreze calculele numerice corespunzătoare aplicării algoritmului K -means pe acest set de date. Care sunt clusterele finale și centroizii corespunzători?

b. Se consideră algoritmul EM/GMM descris în enunțul problemei 15. El a fost conceput pentru a face estimarea parametrilor unui model de mixtură de două distribuții gaussiene, pentru care se presupune că ambele distribuții au aceeași varianță σ^2 , iar probabilitățile a priori de selecție sunt egale ($1/2$). Prin urmare, se vor estima doar mediile celor două distribuții gaussiene (i.e., centroizii clusterelor corespunzătoare), aplicându-se formulele următoare:⁸⁴⁸

Pasul E:

$$E[Z_{ij}] = P(Z_{ij} = 1 | X = x_i; \mu, \sigma^2) \stackrel{F.Bayes}{=} \frac{\exp\left(-\frac{1}{2\sigma^2}(x_i - \mu_j)^2\right)}{\sum_{p=1}^k \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu_p)^2\right)}$$

Pasul M:

$$\mu_j = \frac{\sum_{i=1}^n E[Z_{ij}] x_i}{\sum_{i=1}^n E[Z_{ij}]}$$

unde simbolul \exp denotă funcția exponentială (e^x), $n = 15$ este numărul de puncte, $k = 2$ este numărul de clustere, $\mu \stackrel{not.}{=} (\mu_1, \mu_2)$, iar Z_{ij} sunt variabilele-indicator „ascunse“ / „latente“, cu $i = \overline{1, n}$ și $j = \overline{1, k}$.

Să se calculeze valorile numerice corespunzătoare execuției primei iterării a acestui algoritm, pornind ca și la punctul a cu valorile inițiale $\mu_1 = -20$, $\mu_2 = -10$ și considerând varianță (fixată) $\sigma^2 = 1$.

⁸⁴⁸Pentru deducerea acestor formule, vedeti problema 15 sau cartea *Machine Learning*, de Tom Mitchell, 1997, pag. 193, 195–196.

c. Se consideră o altă variantă a algoritmului EM (de asemenea pentru GMM, cazul unidimensional), în care însă se estimează toți parametrii (μ , σ , π). Ecuațiile de actualizare pentru parametrii primei distribuții gausiene (desemnată prin a) sunt date mai jos, iar cele pentru cea de-a doua gaussiană (b) sunt obținute pur și simplu înlocuind a cu b :⁸⁴⁹

Pasul E:

$$a_i \stackrel{\text{not.}}{=} P(a | x_i) \stackrel{F.Bayes}{=} \frac{p(x_i | a) \cdot \pi_a}{p(x_i | a) \cdot \pi_a + p(x_i | b) \cdot \pi_b}, \text{ unde}$$

$$p(x_i | a) \stackrel{\text{def.}}{=} \frac{1}{\sqrt{2\pi}\sigma_a} \cdot \exp\left(-\frac{(x_i - \mu_a)^2}{2\sigma_a^2}\right)$$

Pasul M:

$$\mu_a = \frac{a_1 x_1 + a_2 x_2 + \dots + a_n x_n}{a_1 + a_2 + \dots + a_n}$$

$$\sigma_a^2 = \frac{a_1(x_1 - \mu_a)^2 + a_2(x_2 - \mu_a)^2 + \dots + a_n(x_n - \mu_a)^2}{a_1 + a_2 + \dots + a_n}$$

$$\pi_a = (a_1 + a_2 + \dots + a_n)/n$$

Presupunem că valorile inițiale pentru parametri sunt: $\mu_a = -20$, $\mu_b = -10$, $\sigma_a^2 = \sigma_b^2 = 1$ și $\pi_a = \pi_b = 0.5$. Efectuați o singură iterație a acestui algoritm EM, pentru a determina noile valori ale parametrilor μ_a , μ_b , σ_a^2 , σ_b^2 , π_a și π_b . Continuând rularea acestui algoritm EM, valorile parametrilor vor rămâne neschimbate? Dacă da, de ce? Dacă nu, de ce?

Răspuns:

a. Algoritmul *K-means* va executa următoarele iterări:

Iterația 1:

$$\begin{aligned} \mu_1 &= -20 & C_1 &= \emptyset \\ \mu_2 &= -10 & C_2 &= \{-9, -8, -7, -6, -5, 5, 5, 6, 6, 7, 7, 8, 8, 9, 9\} \end{aligned}$$

Iterația 2:

$$\begin{aligned} \mu_1 &= -20 & C_1 &= \{-9\} \\ \mu_2 &= \frac{-9 + \dots + 9}{15} = \frac{7}{3} = 2.(3) & C_2 &= \{-8, -7, -6, -5, 5, 5, 6, 6, 7, 7, 8, 8, 9, 9\} \end{aligned}$$

Iterația 3:

$$\begin{aligned} \mu_1 &= -9 & C_1 &= \{-9, -8, -7, -6, -5\} \\ \mu_2 &= \frac{-8 + \dots + 9}{14} = \frac{44}{14} = \frac{22}{7} \approx 3.143 & C_2 &= \{5, 5, 6, 6, 7, 7, 8, 8, 9, 9\} \end{aligned}$$

Iterația 4:

$$\begin{aligned} \mu_1 &= \frac{-9 - 8 - 7 - 6 - 5}{5} = -7 & C_1 &= \{-9, -8, -7, -6, -5\} \\ \mu_2 &= \frac{2(5 + 6 + 7 + 8 + 9)}{10} = 7 & C_2 &= \{5, 5, 6, 6, 7, 7, 8, 8, 9, 9\} \end{aligned}$$

⁸⁴⁹Pentru deducerea acestor formule, vedeti rezolvarea problemei 18.

După această iterație algoritmul K -means se oprește, întrucât centroizii nu se mai modifică. Așadar, clusterele finale sunt:

$$C_1 = \{-9, -8, -7, -6, -5\} \text{ și } C_2 = \{5, 5, 6, 6, 7, 7, 8, 8, 9, 9\}, \text{ cu } \mu_1 = -7 \text{ și } \mu_2 = 7.$$

b. O iterare a algoritmului EM/GMM în această variantă, cu valorile inițiale $\mu_1 = -20$, $\mu_2 = -10$ și $\sigma^2 = 1$ constă în calcularea mediilor variabilelor „ascunse” Z_{ij} (pasul E) și apoi recalcularea mediilor, adică a parametrilor μ_1 și μ_2 (pasul M).

Așadar, la pasul E, pentru punctul $x_1 = -9$, vom avea:

$$E[Z_{11}] = \frac{\exp(-\frac{1}{2}(-9+20)^2)}{\exp(-\frac{1}{2}(-9+10)^2) + \exp(-\frac{1}{2}(-9+20)^2)} = \frac{\exp(-\frac{121}{2})}{\exp(-\frac{1}{2}) + \exp(-\frac{121}{2})} = \frac{e^{-60}}{1 + e^{-60}}$$

$$E[Z_{12}] = \frac{\exp(-\frac{1}{2}(-9+10)^2)}{\exp(-\frac{1}{2}(-9+10)^2) + \exp(-\frac{1}{2}(-9+20)^2)} = \frac{\exp(-\frac{1}{2})}{\exp(-\frac{1}{2}) + \exp(-\frac{121}{2})} = \frac{1}{1 + e^{-60}}$$

Este evident că $P(Z_{11} = 1|x_1) = E[Z_{11}] \approx 0$, iar $P(Z_{12} = 1|x_1) = E[Z_{12}] \approx 1$. Așadar, dacă decidem să asociem fiecare instanță x_i la clusterul / centroidul desemnat de $\text{argmax}_j E[Z_{ij}]$, punctul $x_1 = -9$ va fi asignat celui de-al doilea cluster.

Pentru punctul $x_2 = -8$:

$$E[Z_{21}] = \frac{\exp(-\frac{1}{2} \cdot 144)}{\exp(-\frac{1}{2} \cdot 4) + \exp(-\frac{1}{2} \cdot 144)} = \frac{e^{-70}}{1 + e^{-70}}$$

$$E[Z_{22}] = \frac{\exp(-\frac{1}{2} \cdot 4)}{\exp(-\frac{1}{2} \cdot 4) + \exp(-\frac{1}{2} \cdot 144)} = \frac{1}{1 + e^{-70}}$$

Ca și mai sus, este evident că $E[Z_{21}] \approx 0$, iar $E[Z_{22}] \approx 1$. Deci punctul $x_2 = -8$ va fi asociat tot celui de-al doilea cluster.

Pentru punctul $x_3 = -7$:

$$E[Z_{31}] = \frac{\exp(-\frac{1}{2} \cdot 169)}{\exp(-\frac{1}{2} \cdot 9) + \exp(-\frac{1}{2} \cdot 169)} = \frac{e^{-80}}{1 + e^{-80}}$$

$$E[Z_{32}] = \frac{\exp(-\frac{1}{2} \cdot 9)}{\exp(-\frac{1}{2} \cdot 9) + \exp(-\frac{1}{2} \cdot 169)} = \frac{1}{1 + e^{-80}}$$

Este evident că $E[Z_{31}] \approx 0$, iar $E[Z_{32}] \approx 1$. Deci punctul $x_3 = -7$ va fi asociat de asemenea celui de-al doilea cluster.

Se observă că pentru celelalte puncte se va păstra acest comportament, deci $C_1 = \emptyset$ și $C_2 = \{-9, -8, -7, -6, -5, 5, 5, 6, 6, 7, 7, 8, 8, 9, 9\}$, iar $E[Z_{i1}] \approx 0, E[Z_{i2}] \approx 1$ pentru $i = \overline{1, 15}$.

Apoi, la pasul de **maximizare**, trebuie să calculăm noile valori ale mediilor:

$$\mu_1 = \frac{\frac{e^{-60}}{1+e^{-60}} \cdot (-9) + \frac{e^{-70}}{1+e^{-70}} \cdot (-8) + \frac{e^{-80}}{1+e^{-80}} \cdot (-7) + \dots}{\frac{e^{-60}}{1+e^{-60}} + \frac{e^{-70}}{1+e^{-70}} + \frac{e^{-80}}{1+e^{-80}} + \dots} \approx -9$$

$$\mu_2 \approx \frac{1 \cdot (-9) + \dots + 1 \cdot 9}{15} = \frac{35}{15} = \frac{7}{3} = 2.(3)$$

Observație (1): Folosind o implementare în Matlab a acestui algoritm EM/GMM, valorile numerice pe care le-am obținut pentru această iterație sunt:

x_i	-9	-8	-7	-6	-5
$E[Z_{i1}]$	$8.75 \cdot 10^{-27}$	$3.97 \cdot 10^{-31}$	$1.8 \cdot 10^{-35}$	$8.19 \cdot 10^{-40}$	$3.72 \cdot 10^{-44}$
$E[Z_{i2}]$	≈ 1	≈ 1	≈ 1	≈ 1	≈ 1

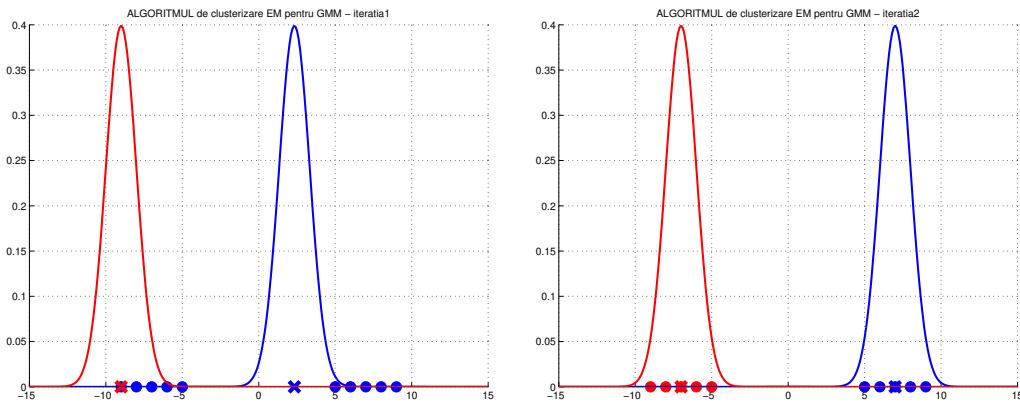
x_i	5	6	7	8	9
$E[Z_{i1}]$	$1.38 \cdot 10^{-87}$	$6.28 \cdot 10^{-92}$	$2.85 \cdot 10^{-96}$	$1.29 \cdot 10^{-100}$	$5.87 \cdot 10^{-105}$
$E[Z_{i2}]$	≈ 1	≈ 1	≈ 1	≈ 1	≈ 1

$$\mu_1 = -8.999955 \text{ și } \mu_2 = 2.333333$$

La următoarea iterație a acestui algoritm EM/GMM, componenta clusterelor devine:

$$C_1 = \{-9, -8, -7, -6, -5\} \text{ și } C_2 = \{5, 5, 6, 6, 7, 7, 8, 8, 9, 9\}, \text{ cu } \mu_1 \approx -7, \mu_2 \approx 7,$$

după cum se poate observa și din graficul de mai jos, partea dreaptă:⁸⁵⁰



La următoarea iterație, algoritmul converge.⁸⁵¹

Observație (2): Așadar, pentru cazul $\sigma^2 = 1$, soluția obținută de acest algoritm EM/GMM coincide cu cea a algoritmului K-means (deși numărul de iterării diferă). În cele ce urmează vom arăta că pentru alte valori ale lui σ^2 se pot obține rezultate semnificativ diferite pentru *valorile finale ale mediilor*:

⁸⁵⁰Graficul din partea stângă corespunde finalului primei iterării.

⁸⁵¹Criteriul de oprire pe care l-am folosit în implementarea noastră pentru algoritm EM/GMM a fost următorul: la două iterării succesive, valorile mediilor nu se modifică cu mai mult de 10^{-5} .

Pentru $\sigma^2 = 20$, algoritmul EM/GMM converge după 10 iterații, obținându-se aceleași clustere C_1 și C_2 , dar cu

$$\mu_1 = -6.657120 \text{ și } \mu_2 = 6.940543.$$

Se observă că aceste medii sunt foarte aproape de valorile 7 și -7 obținute de către algoritmul K -means, doar că au fost puțin „atras“ spre punctele din clusterul opus (și anume, mai mult μ_2 decât μ_1 , pentru că în clusterul din dreapta sunt de două ori mai multe puncte decât în clusterul din stânga).

Pentru $\sigma^2 = 30$, algoritmul EM/GMM converge după 29 iterații, obținându-se aceleași clustere ca mai sus, dar cu

$$\mu_1 = -4.564930 \text{ și } \mu_2 = 6.698046.$$

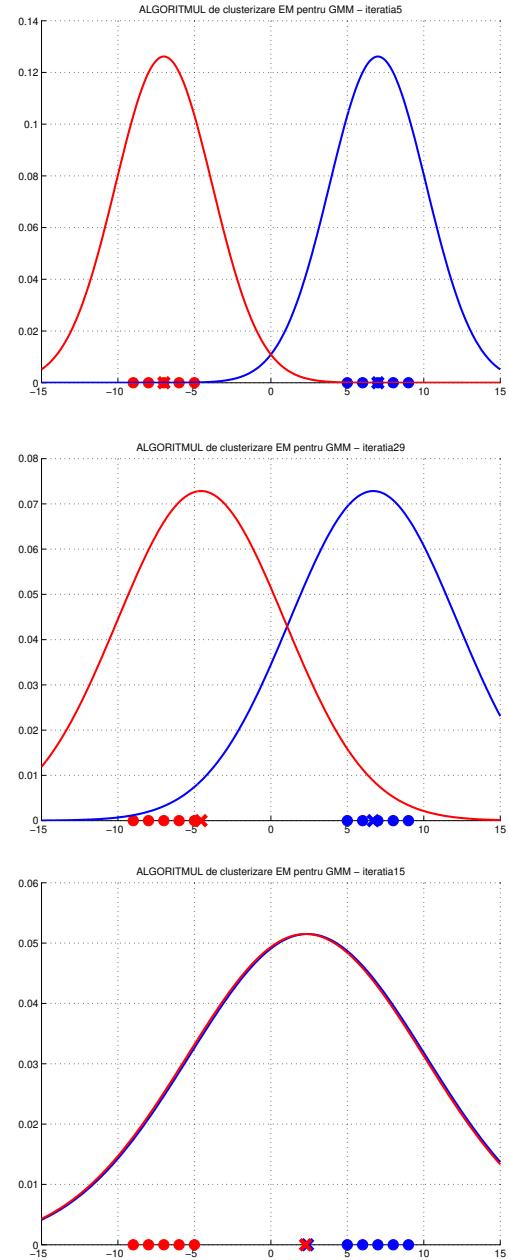
Media primei distribuții gaussiene a migrat în mod considerabil spre cele 10 puncte din cel de-al doilea cluster, pe când media celei de-a doua distribuții gaussiene este mult mai puțin „atras“ de cele 5 puncte din clusterul opus.

Pentru $\sigma^2 = 60$, algoritmul EM/GMM converge după 45 iterații, obținându-se medii aproape identice:

$$\mu_1 = 2.333316 \text{ și } \mu_2 = 2.333351.$$

Cu toate acestea, probabilitățile de apartenență ale punctelor la cele două clustere sunt sensibil diferite. De exemplu, pentru punctul -9 aceste probabilități sunt $E[Z_{11}] = 0.5000022$ și $E[Z_{12}] = 0.4999978$. Acest fapt determină împărțirea punctelor în aceleași două clustere ca mai sus.

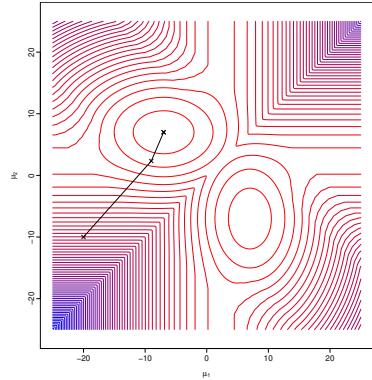
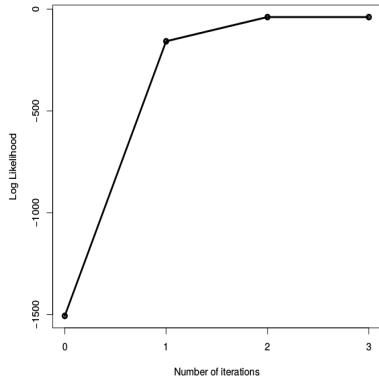
În imaginile următoare prezentăm,⁸⁵² pentru diferențele cazuri de mai sus (i. $\sigma^2 = 1$, ii. $\sigma^2 = 20$, iii. $\sigma^2 = 30$, iv. $\sigma^2 = 60$), valorile funcțiilor de log-verosimilitate [a datelor „observabile“] obținute de algoritmul EM la iterații succesive (vedeți partea stângă) și respectiv graficele funcțiilor de log-verosimilitate reprezentate prin curbe de izocontur, pe care au fost puse în evidență valorile parametrilor μ_1 și μ_2 obținute de EM la iterații succesive (partea dreaptă).⁸⁵³



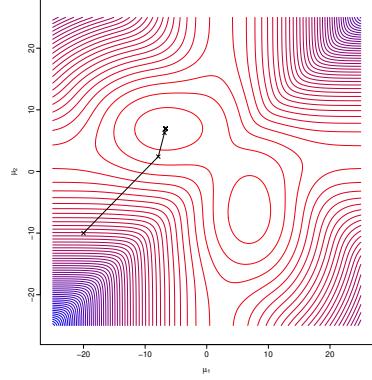
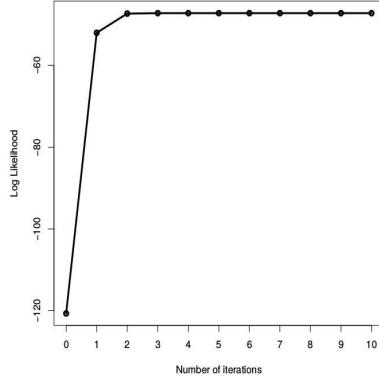
⁸⁵²Graficele au fost realizate de către studentul Sebastian Ciobanu în 2018.

⁸⁵³Pentru un exemplu relativ simplu de calculare a funcției de log-verosimilitate a datelor observabile în cazul unei mixturi de două distribuții gaussiene unidimensionale, vedeți ex. 16.c.

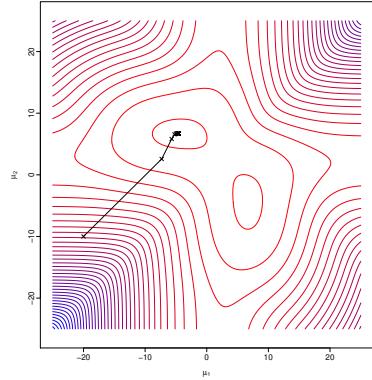
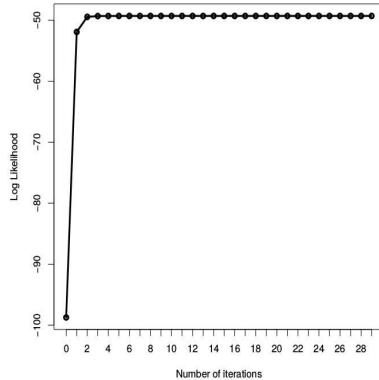
Cazul i.
 $\sigma^2 = 1.$



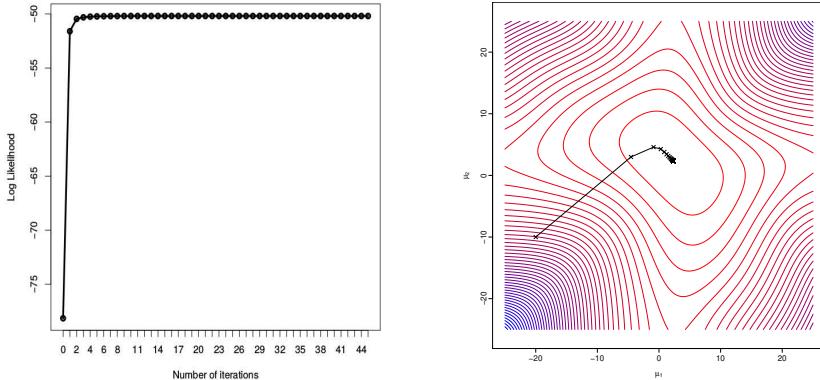
Cazul ii.
 $\sigma^2 = 20.$



Cazul iii.
 $\sigma^2 = 30.$



Cazul iv.
 $\sigma^2 = 60$.



Se observă, ca și la problema 16, că valorile funcției de log-verosimilitate [a datelor „observabile“] obținute de algoritmul EM cresc, aşa cum era de așteptat (vedeți problema 2 de la capitolul *Schema algoritmică EM*).

c. Pentru a realiza o iterație a acestei variante algoritmului EM/GMM trebuie să calculăm „densitățile“ (adică, valorile funcțiilor densitate de probabilitate ale celor două gaussiene) $p(x_i | a)$ și $p(x_i | b)$ și probabilitățile a posteriori $a_i \stackrel{not.}{=} P(a | x_i)$ și $b_i \stackrel{not.}{=} P(b | x_i)$ corespunzătoare fiecărui punct x_i din setul de date considerat.

Spre exemplu, pentru punctul -9 , vom obține:

$$\begin{aligned} p(-9 | a) &= \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{(-9+20)^2}{2}\right) = \frac{1}{e^{60}\sqrt{2\pi e}} \approx 2 \cdot 10^{-27} \\ p(-9 | b) &= \frac{1}{\sqrt{2\pi}} \cdot \exp\left(-\frac{(-9+10)^2}{2}\right) = \frac{1}{\sqrt{2\pi e}} \approx 0.24 \\ a_{-9} &\stackrel{T. Bayes}{=} \frac{2 \cdot 10^{-27} \cdot 0.5}{0.24 \cdot 0.5 + 2 \cdot 10^{-27} \cdot 0.5} = 8 \cdot 10^{-27} \\ b_{-9} &\stackrel{T. Bayes}{=} \frac{0.24 \cdot 0.5}{0.24 \cdot 0.5 + 2 \cdot 10^{-27} \cdot 0.5} \approx 1 \end{aligned}$$

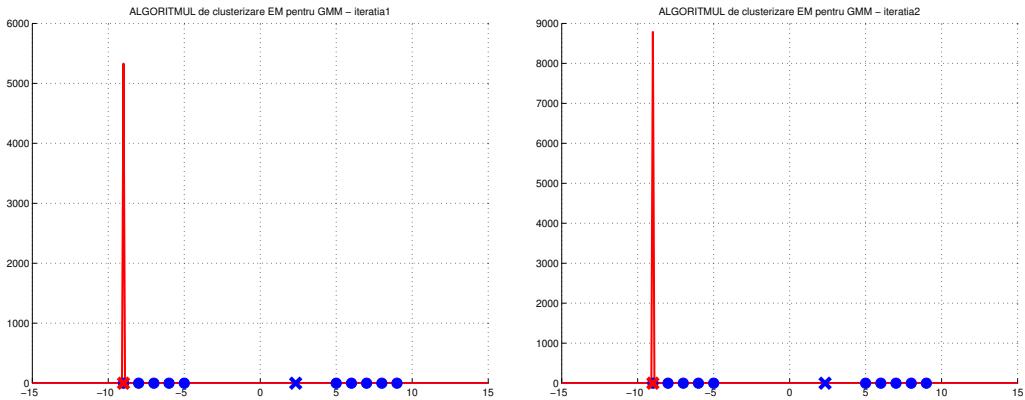
În mod analog se fac calculele pentru toate punctele. Folosind o [altă] implementare în Matlab, am obținut:

x_i	-9	-8	-7	-6	-5
$p(x_i a)$	$2 \cdot 10^{-27}$	$2 \cdot 10^{-32}$	$8 \cdot 10^{-38}$	10^{-43}	10^{-50}
$p(x_i b)$	0.24	0.05	0.004	0.0001	$1.4 \cdot 10^{-6}$
a_i	$8 \cdot 10^{-27}$	10^{-31}	10^{-35}	10^{-40}	10^{-44}
b_i	≈ 1	≈ 1	≈ 1	≈ 1	≈ 1

x_i	5	6	7	8	9
$p(x_i a)$	10^{-137}	10^{-148}	10^{-159}	10^{-171}	10^{-184}
$p(x_i b)$	10^{-50}	10^{-56}	10^{-64}	10^{-71}	10^{-79}
a_i	10^{-87}	10^{-92}	10^{-96}	10^{-100}	10^{-105}
b_i	≈ 1				

Apoi se calculează $\pi_a = \frac{1}{n} \sum_{i=1}^{15} a_i \simeq 5 \cdot 10^{-28}$ și $\pi_b = \frac{1}{n} \sum_{i=1}^{15} b_i \simeq 1$, deci toate punctele vor fi asignate clusterului b , iar noile medii și varianțe vor fi:⁸⁵⁴

$$\begin{aligned}\mu_a &\simeq -9, & \mu_b &= 2.3 \\ \sigma_a &= 0.000045, & \sigma_b &= 45.55\end{aligned}$$



La cea de-a două iterație, varianța componentei a se va apropiă foarte mult de 0, ceea ce va cauza un fapt suprinzător: cu toate că $\mu_a \simeq -9$, punctul $x_1 = -9$ va fi asociat celuilalt cluster, datorită varianței foarte mici a lui a , precum și datorită probabilității foarte mici de selecție a acestei gaussiene ($\pi_a \simeq 0$). Deci clusterele finale vor fi:

$$C_a = \emptyset \text{ și } C_b = \{-9, -8, -7, -6, -5, 5, 5, 6, 6, 7, 7, 8, 8, 9, 9\}$$

Observație (3): Soluția pe care tocmai am obținut-o cu această variantă a algoritmului EM/GMM nu este satisfăcătoare. Totuși, ca și la punctul b , vom arăta că rezultatul se poate schimba, în funcție de valorile inițiale considerate. Concret, dacă se variază initializarea varianțelor, se pot obține alte clustere, ca în exemplele următoare:

Pentru $\sigma_a^2 = \sigma_b^2 = 9$, algoritmul converge după 8 iterării, obținându-se:

$$\begin{aligned}\mu_a &\simeq -8, & \mu_b &= 3.071348 \\ \sigma_a &= 0.000308, & \sigma_b &= 40.638356 \\ \pi_a &= 0.0666659, & \pi_b &= 0.933340\end{aligned}$$

Punctul $x_2 = -8$ va fi asociat gaussienei a , încrucișat probabilățile de apartenență la cele două clustere sunt: $a_2 = E[Z_{21}] = 0.999897$, $b_2 = E[Z_{22}] = 0.000102$, însă restul punctelor vor fi asignate celuilalt cluster.

Deci, la finalul execuției algoritmului vom avea:

$$C_a = \{-8\} \text{ și } C_b = \{-9, -7, -6, -5, 5, 5, 6, 6, 7, 7, 8, 8, 9, 9\}.$$

⁸⁵⁴La următoarele trei grafice nu am pus în evidență p.d.f.-urile corespunzătoare gaussienei b din cauza varianțelor prea mari. (În astfel de situații, curbele gaussiene nu se disting în mod semnificativ de axa Ox .)

Este un rezultat interesant, care se datorează nerestricționării varianțelor celor două distribuții. Se pune astfel în evidență faptul că spre deosebire de cazul algoritmului K -means, *suprafețele de separare determinate de algoritmul EM nu sunt în mod neapărat liniare!*

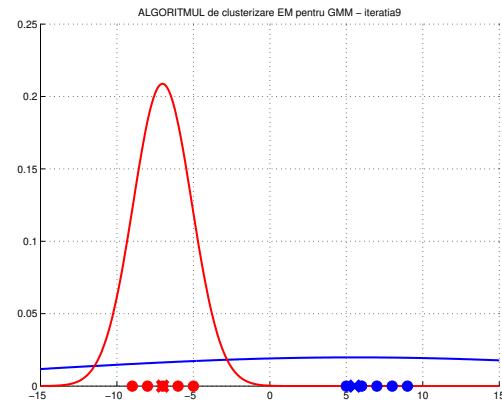
Observație (4): La ultimele două rezultate se observă că valoarea funcției densitate de probabilitate (p.d.f.) poate fi foarte mare dacă parametrul σ ia valori foarte mici. Într-o astfel de situație, valoarea funcției de log-verosimilitate a datelor complete poate deveni foarte mare.⁸⁵⁵ Astfel de situații (caracterizate de „puncte de singularitate“ pentru p.d.f.) nu sunt însă în general dezirabile pentru clusterizare.

Pentru $\sigma_a^2 = \sigma_b^2 = 20$, algoritmul EM/GMM converge după 9 iterării, obținându-se:

$$\begin{aligned}\mu_a &= -7.018194, & \mu_b &= 5.550311 \\ \sigma_a &= 1.910617, & \sigma_b &= 20.137057 \\ \pi_a &= 0.255955, & \pi_b &= 0.744044\end{aligned}$$

Clusterurile obținute în acest caz sunt cele așteptate, și anume:

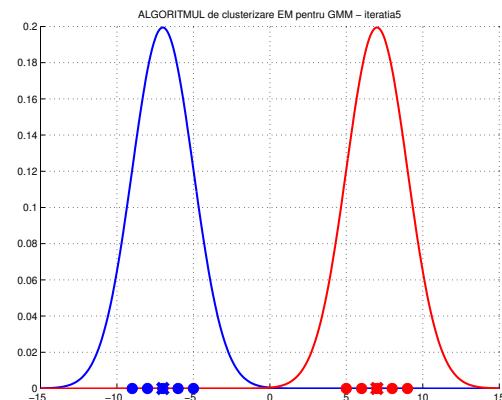
$$\begin{aligned}C_a &= \{-9, -8, -7, -6, -5\} \\ C_b &= \{5, 5, 6, 6, 7, 7, 8, 8, 9, 9\}\end{aligned}$$



Pentru $\sigma_a^2 = 20$ și $\sigma_b^2 = 4$, acest algoritm EM/GMM converge după doar 5 iterării, obținându-se un rezultat similar cu cel al algoritmului de la punctul b (unde s-au folosit valori identice pentru σ^2):

$$\begin{aligned}\mu_a &= 7, & \mu_b &= -7 \\ \sigma_a &= 2, & \sigma_b &= 2 \\ \pi_a &= 0.(6), & \pi_b &= 0.(3)\end{aligned}$$

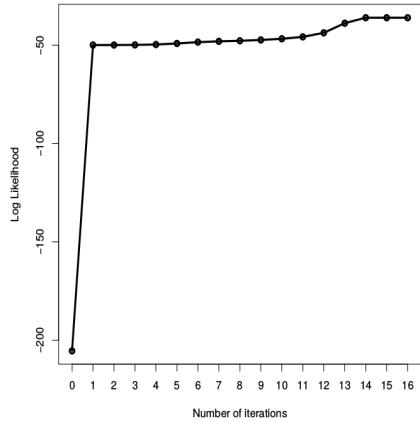
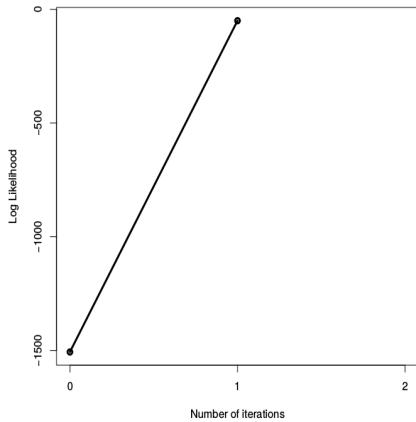
Mai mult, clusterele și mediile / centroizii sunt aceiași ca și în cazul algoritmului K -means, însă — interesant! — ele devin inversate: $C_a = \{5, 5, 6, 6, 7, 7, 8, 8, 9, 9\}$, $C_b = \{-9, -8, -7, -6, -5\}$



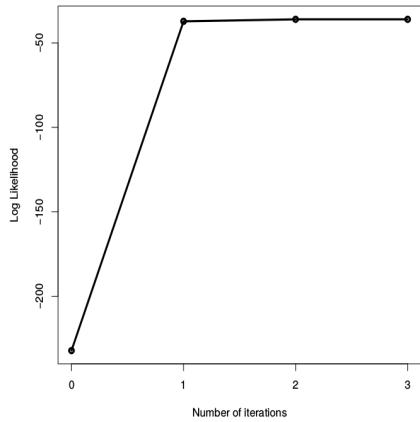
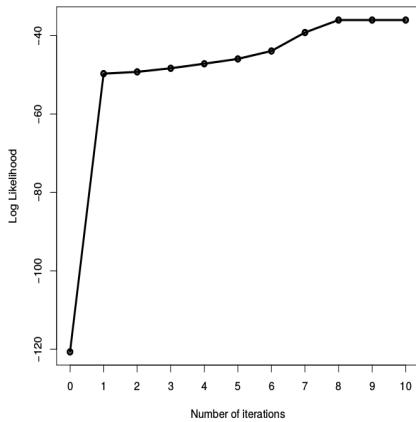
În imaginile următoare prezentăm,⁸⁵⁶ pentru diferitele cazuri de mai sus (*i.e.* $\sigma_a^2 = \sigma_b^2 = 1$, *ii.* $\sigma_a^2 = \sigma_b^2 = 9$, *iii.* $\sigma_a^2 = \sigma_b^2 = 20$, *iv.* $\sigma_a^2 = 20, \sigma_b^2 = 4$), valorile funcțiilor de log-verosimilitate obținute de algoritmul EM la iterării succesive.

⁸⁵⁵La distribuții discrete, funcția de log-verosimilitate a datelor complete (a cărei medie este maximizată de algoritmul EM) nu poate avea decât valori negative. (Justificarea este imediată, fiindcă funcția masă de probabilitate nu poate lua valori decât în intervalul $[0, 1]$.) La distribuții continue, funcția de log-verosimilitate a datelor complete poate lua și valori pozitive.

⁸⁵⁶Graficele au fost realizate de către studentul Sebastian Ciobanu în 2018.



Cazurile *i.* $\sigma_a^2 = \sigma_b^2 = 1$ (partea stângă) și, *ii.* $\sigma_a^2 = \sigma_b^2 = 9$ (partea dreaptă).



Cazurile *iii.* $\sigma_a^2 = \sigma_b^2 = 20$ (partea stângă) și, *iv.* $\sigma_a^2 = 20, \sigma_b^2 = 4$ (partea dreaptă).

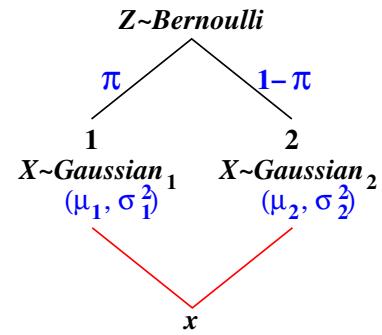
Observație (5): T. Hastie, R. Tibshirani și J. Friedman recomandă în cartea *The Elements of Statistical Learning* (Springer, ed. a II-a, 2009) ca la pasul de inițializare în algoritmul EM/GMM să se folosească pentru varianțele celor două distribuții gaussiene individuale valoarea varianței empirice a întregului set de date.⁸⁵⁷

⁸⁵⁷Pentru un exemplu de calcul al varianței empirice, vedeți problema 123.b de la capitolul de *Fundamente*.

18.

(Algoritmul EM: estimarea tuturor parametrilor unei mixturi de două distribuții gaussiene unidimensionale)

*Liviu Ciortuz, 2012, după
■ • ○ MIT, ML course 6768, 2012 fall, Dahua Lin,
An Introduction to Expectation-Maximization*



Folosind algoritmul EM, rezolvați problema estimării parametrilor unei mixturi de două distribuții gaussiene unidimensionale în cazul cel mai general, adică lăsând liberi toți parametrii (μ – mediile, σ^2 – varianțele și π – probabilitățile a priori de selecție a celor două gaussiene).

Răspuns:

Vom urma etapele indicate de schema algoritmică EM, prezentată în secțiunea 6.12.2 din carte Machine Learning de Tom Mitchell (sau, echivalent problema 1, pag. 959 de la capitolul Schema algoritmică EM din această culegere).

Pasul E:

Vom considera instanțele $x_1, \dots, x_n \in \mathbb{R}$, vom nota parametrii $\mu = (\mu_1, \mu_2)$, $\sigma = (\sigma_1, \sigma_2)$ și $\pi = (\pi_1, \pi_2)$ și vom folosi variabilele aleatoare $Z_{i1}, Z_{i2} \in \{0, 1\}$ cu restricția $Z_{i1} + Z_{i2} = 1$ pentru orice $i \in \{1, \dots, n\}$. În aceste condiții, probabilitățile / mediile

$$p_{ij} \stackrel{\text{not.}}{=} P(Z_{ij} = 1 \mid X_i, \mu, \sigma, \pi) \stackrel{\text{calcul}}{=} E[Z_{ij} \mid X_i, \mu, \sigma, \pi], \text{ pentru } i = 1, \dots, n \text{ și } j = 1, 2,$$

se vor obține imediat folosind teorema lui Bayes:

$$\begin{aligned} p_{ij} &= \frac{P(X_i = x_i \mid Z_{ij} = 1, \mu, \sigma, \pi) \cdot P(Z_{ij} = 1 \mid \mu, \sigma, \pi)}{\sum_{j'} P(X_i = x_i \mid Z_{ij'} = 1, \mu, \sigma, \pi) \cdot P(Z_{ij'} = 1 \mid \mu, \sigma, \pi)} \\ &= \frac{\frac{1}{\sqrt{2\pi}\sigma_j} \cdot \exp\left(-\frac{(x_i - \mu_j)^2}{2\sigma_j^2}\right) \cdot \pi_j}{\frac{1}{\sqrt{2\pi}\sigma_1} \cdot \exp\left(-\frac{(x_i - \mu_1)^2}{2\sigma_1^2}\right) \cdot \pi_1 + \frac{1}{\sqrt{2\pi}\sigma_2} \cdot \exp\left(-\frac{(x_i - \mu_2)^2}{2\sigma_2^2}\right) \cdot \pi_2} \\ &= \frac{\frac{\pi_j}{\sigma_j} \cdot \exp\left(-\frac{(x_i - \mu_j)^2}{2(\sigma_j)^2}\right)}{\frac{\pi_1}{\sigma_1} \cdot \exp\left(-\frac{(x_i - \mu_1)^2}{2(\sigma_1)^2}\right) + \frac{\pi_2}{\sigma_2} \cdot \exp\left(-\frac{(x_i - \mu_2)^2}{2(\sigma_2)^2}\right)}. \end{aligned}$$

Verosimilitatea oricărei instanțe „complete“ $y_i \stackrel{\text{not.}}{=} (x_i, z_{i1}, z_{i2})$ se va exprima sub forma:

$$\begin{aligned} P(X_i = x_i, Z_{i1} = z_{i1}, Z_{i2} = z_{i2} \mid \mu, \sigma, \pi) &= P(X_i = x_i \mid Z_{i1} = z_{i1}, Z_{i2} = z_{i2}, \mu_i, \sigma_i, \pi_i) \cdot P(Z_{i1} = z_{i1}, Z_{i2} = z_{i2} \mid \mu_i, \sigma_i, \pi_i) \\ &= \frac{1}{\sqrt{2\pi}\sigma_{j'}} \cdot \exp\left(-\frac{(x_i - \mu_{j'})^2}{2\sigma_{j'}^2}\right) \cdot \pi_{j'}, \text{ unde } z_{ij'} = 1 \text{ iar } z_{ij''} = 0 \text{ pentru } j'' \neq j'. \end{aligned}$$

Tinând cont de egalitatea $z_{i1} + z_{i2} = 1$ și de faptul că $z_{ij} \in \{0, 1\}$, putem continua dezvoltarea acestei expresii astfel:

$$\begin{aligned} P(X_i = x_i, Z_{i1} = z_{i1}, Z_{i2} = z_{i2} | \mu, \sigma, \pi) \\ = \frac{1}{\sqrt{2\pi} \sigma_1^{z_{i1}} \sigma_2^{z_{i2}}} \cdot \exp \left(-\frac{1}{2} \sum_{j \in \{1,2\}} z_{ij} \frac{(x_i - \mu_j)^2}{\sigma_j^2} \right) \cdot \pi_1^{z_{i1}} \pi_2^{z_{i2}}. \end{aligned}$$

Observație (1): Acest mod de a scrie $P(X_i, Z_{i1}, Z_{i2} | \mu, \sigma, \pi)$, deși pare cumva artificial, este de un ajutor crucial pentru că include atât Z_{i1} cât și Z_{i2} . Ne pregătim astfel pentru aplicarea [ulterioră a] proprietății de liniaritate a mediei, la construirea funcției „auxiliare“ Q .

Logaritmând ultima expresie obținută mai sus, vom obține *log-verosimilitatea* instanței „complete“ $y_i = (x_i, z_{i1}, z_{i2})$:

$$\begin{aligned} \ln P(X_i = x_i, Z_{i1} = z_{i1}, Z_{i2} = z_{i2} | \mu, \sigma, \pi) \\ = -\frac{1}{2} \ln(2\pi) - \sum_{j=1}^2 z_{ij} \ln \sigma_j - \frac{1}{2} \sum_{j=1}^2 z_{ij} \frac{(x_i - \mu_j)^2}{\sigma_j^2} + \sum_{j=1}^2 z_{ij} \ln \pi_j. \end{aligned}$$

Notând $X = (X_1, \dots, X_n)$, $Z_1 = (Z_{11}, Z_{21}, \dots, Z_{n1})$ și $Z_2 = (Z_{12}, Z_{22}, \dots, Z_{n2})$, și tinând cont de independentă statistică a datelor (x_i, z_{i1}, z_{i2}) cu $i = 1, \dots, n$, *log-verosimilitatea* datelor complete va fi dată de expresia:

$$\begin{aligned} \ln P(X, Z_1, Z_2 | \mu, \sigma, \pi) &= \ln \prod_{i=1}^n P(X_i = x_i, Z_{i1}, Z_{i2} | \mu, \sigma, \pi) = \\ &\sum_{i=1}^n \ln P(X_i = x_i, Z_{i1}, Z_{i2} | \mu, \sigma, \pi) = \\ &-\frac{n}{2} \ln(2\pi) - \sum_{i=1}^n \sum_{j=1}^2 Z_{ij} \ln \sigma_j - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^2 Z_{ij} \frac{(x_i - \mu_j)^2}{\sigma_j^2} + \sum_{i=1}^n \sum_{j=1}^2 Z_{ij} \ln \pi_j. \end{aligned}$$

Mai departe, pentru a calcula *media log-verosimilității* datelor complete, vom ține cont de proprietatea de liniaritate a mediilor. Așadar, vom avea:

$$\begin{aligned} E[\ln P(X, Z_1, Z_2 | \mu, \sigma, \pi)] &= \\ &-\frac{n}{2} \ln(2\pi) - \sum_{i=1}^n \sum_{j=1}^2 E[Z_{ij}] \ln \sigma_j - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^2 E[Z_{ij}] \frac{(x_i - \mu_j)^2}{\sigma_j^2} + \sum_{i=1}^n \sum_{j=1}^2 E[Z_{ij}] \ln \pi_j. \end{aligned}$$

Acum vom pune în evidență faptul că media aceasta se calculează în funcție de probabilitatea condiționată $P_{Z|X, \mu^{(t)}, \sigma^{(t)}, \pi^{(t)}}$, unde $\mu^{(t)}, \sigma^{(t)}, \pi^{(t)}$ notează valorile parametrilor la inițializare ($t = 0$) sau, mai general, la iteratărea precedentă:

$$\begin{aligned} Q(\mu, \sigma, \pi | \mu^{(t)}, \sigma^{(t)}, \pi^{(t)}) &\stackrel{\text{not.}}{=} E_{Z|X, \mu^{(t)}, \sigma^{(t)}, \pi^{(t)}} [\ln P(X, Z_1, Z_2 | \mu, \sigma, \pi)] = \\ &-\frac{n}{2} \ln(2\pi) - \sum_{i=1}^n \sum_{j=1}^2 E[Z_{ij} | X_i, \mu^{(t)}, \sigma^{(t)}, \pi^{(t)}] \ln \sigma_j \\ &- \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^2 E[Z_{ij} | X_i, \mu^{(t)}, \sigma^{(t)}, \pi^{(t)}] \frac{(x_i - \mu_j)^2}{\sigma_j^2} \\ &+ \sum_{i=1}^n \sum_{j=1}^2 E[Z_{ij} | X_i, \mu^{(t)}, \sigma^{(t)}, \pi^{(t)}] \ln \pi_j. \end{aligned}$$

Extinzând în mod natural notăția $p_{ij} = E[Z_{ij} | X_i, \mu, \sigma, \pi]$, care a fost introdusă mai sus la $p_{ij}^{(t)} = E[Z_{ij} | X_i, \mu^{(t-1)}, \sigma^{(t-1)}, \pi^{(t-1)}]$, vom rescrie în mod simplificat expresia aşa-numitei *funcții auxiliare* Q :

$$\begin{aligned} Q(\mu, \sigma, \pi | \mu^{(t)}, \sigma^{(t)}, \pi^{(t)}) &= \\ -\frac{n}{2} \ln(2\pi) - \sum_{i=1}^n \sum_{j=1}^2 p_{ij}^{(t)} \ln \sigma_j - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^2 p_{ij}^{(t)} \frac{(x_i - \mu_j)^2}{\sigma_j^2} + \sum_{i=1}^n \sum_{j=1}^2 p_{ij}^{(t)} \ln \pi_j. \end{aligned} \quad (347)$$

Pasul M:

Vom căuta optimul funcției $Q(\mu, \sigma, \pi | \mu^{(t)}, \sigma^{(t)}, \pi^{(t)})$ cu ajutorul derivatelor parțiale de ordinul întâi.

Pentru calculul derivatelor parțiale în raport cu π_j , vom ține cont de restricția $\pi_1 + \pi_2 = 1$:

$$\begin{aligned} Q(\mu, \sigma, \pi | \mu^{(t)}, \sigma^{(t)}, \pi^{(t)}) &= -\frac{n}{2} \ln 2\pi - \sum_{i=1}^n \sum_{j=1}^2 p_{ij}^{(t)} \ln \sigma_j \\ &\quad - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^2 p_{ij}^{(t)} \frac{(x_i - \mu_j)^2}{\sigma_j^2} + \sum_{i=1}^n (p_{i1}^{(t)} \ln \pi_1 + p_{i2}^{(t)} \ln(1 - \pi_1)). \end{aligned}$$

Așadar,

$$\begin{aligned} \frac{\partial}{\partial \pi_1} Q(\mu, \sigma, \pi | \mu^{(t)}, \sigma^{(t)}, \pi^{(t)}) = 0 &\Leftrightarrow \frac{1}{\pi_1} \sum_{i=1}^n p_{i1}^{(t)} = \frac{1}{1 - \pi_1} \sum_{i=1}^n p_{i2}^{(t)} \Leftrightarrow \\ \sum_{i=1}^n p_{i1}^{(t)} &= \pi_1 \left(\sum_{i=1}^n p_{i1}^{(t)} + \sum_{i=1}^n p_{i2}^{(t)} \right) \Leftrightarrow \sum_{i=1}^n p_{i1}^{(t)} = \pi_1 \underbrace{\sum_{i=1}^n (p_{i1}^{(t)} + p_{i2}^{(t)})}_{1} \Leftrightarrow \sum_{i=1}^n p_{i1}^{(t)} = n\pi_1, \end{aligned}$$

de unde rezultă $\pi_1^{(t+1)} \leftarrow \frac{1}{n} \sum_{i=1}^n p_{i1}^{(t)}$. Apoi, ținând cont de relațiile $p_{i1}^{(t)} + p_{i2}^{(t)} = 1$ pentru $i = 1, \dots, n$ și $\pi_1^{(t+1)} + \pi_2^{(t+1)} = 1$, vom obține imediat $\pi_2^{(t+1)} \leftarrow \frac{1}{n} \sum_{i=1}^n p_{i2}^{(t)}$.

Mai departe, vom deriva funcția Q în raport cu μ_1 :

$$\frac{\partial}{\partial \mu_1} Q(\mu, \sigma, \pi | \mu^{(t)}, \sigma^{(t)}, \pi^{(t)}) = 0 \Leftrightarrow \frac{1}{\sigma_1^2} \sum_{i=1}^n p_{i1}^{(t)} (x_i - \mu_1) = 0 \Leftrightarrow \sum_{i=1}^n p_{i1}^{(t)} (x_i - \mu_1) = 0,$$

de unde avem $\mu_1^{(t+1)} \leftarrow \frac{\sum_{i=1}^n p_{i1}^{(t)} x_i}{\sum_{i=1}^n p_{i1}^{(t)}}$. Similar, vom obține $\mu_2^{(t+1)} \leftarrow \frac{\sum_{i=1}^n p_{i2}^{(t)} x_i}{\sum_{i=1}^n p_{i2}^{(t)}}$.

(Se poate demonstra ușor că $\sum_{i=1}^n p_{i1}^{(t)} > 0$ și $\sum_{i=1}^n p_{i2}^{(t)} > 0$.)

În final, vom deriva funcția Q în raport cu σ_1 :

$$\frac{\partial}{\partial \sigma_1} Q(\mu, \sigma, \pi | \mu^{(t)}, \sigma^{(t)}, \pi^{(t)}) = 0 \Leftrightarrow -\frac{1}{\sigma_1} \sum_{i=1}^n p_{i1}^{(t)} + \frac{1}{\sigma_1^3} \sum_{i=1}^n p_{i1}^{(t)} (x_i - \mu_1)^2 = 0,$$

de unde rezultă că $\left(\sigma_1^{(t+1)}\right)^2 \leftarrow \frac{\sum_{i=1}^n p_{i1}^{(t)} (x_i - \mu_1^{(t+1)})^2}{\sum_{i=1}^n p_{i1}^{(t)}}$.

Similar, vom obține $\left(\sigma_2^{(t+1)}\right)^2 \leftarrow \frac{\sum_{i=1}^n p_{i2}^{(t)} (x_i - \mu_2^{(t+1)})^2}{\sum_{i=1}^n p_{i2}^{(t)}}$.

Observație (2):

Cititorul atent va sesiza că în final, la calculul expresiei lui $\sigma_1^{(t+1)}$, s-a înlocuit μ_1 cu $\mu_1^{(t+1)}$. Explicația, în manieră *intuitivă*, este următoarea: per ansamblu, aici calculăm maximul funcției Q nu doar în raport cu σ_1 ci și cu ceilalți parametri, deci în particular și cu μ_1 , a cărui valoare optimă la iterată $t + 1$ este $\mu_1^{(t+1)}$. La calculul celorlalte valori optime ale parametrilor ($\pi^{(t+1)}$ și $\mu^{(t+1)}$) nu a fost necesară o astfel de coroborare a soluțiilor. Cazul lui $\sigma^{(t+1)}$ este însă ușor diferit față de cazul celorlalți parametri. În mod *riguros*, se poate demonstra că rădăcinile derivatelor partiale de ordinul întâi ale funcției Q constituie într-adevăr valorile [parametrilor] pentru care această funcție își atinge *maximul*.⁸⁵⁸

Observație (3):

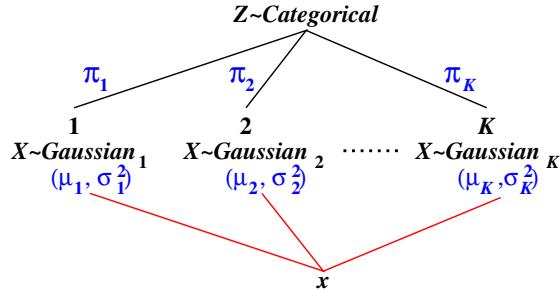
Este foarte util să comparăm *i.* forma relațiilor de actualizare pentru parametrii μ_j și σ_j^2 de la pasul M al algoritmului EM pentru modelarea unei mixturi de distribuții gaussiene unidimensionale cu *ii.* formulele obținute la estimarea în sens MLE a parametrilor μ și σ^2 ai unei distribuții gaussiene tot unidimensionale (vedeți problemele 50.a și 51.a de la capitolul de *Fundamente*). Se poate observa ușor că aceste două seturi de formule sunt foarte asemănătoare! Așa cum era așteptat, deosebirea dintre ele este dată de prezența probabilităților p_{ij} (calculate la pasul E) din algoritm EM. Așadar, formulele *i.* sunt versiuni *ponderate* ale formulelor *ii.* (deci reprezintă *medii ponderate*, respectiv *varianțe ponderate*).

Observație (4):

O comparație similară se poate face și între formula de actualizare a mediilor μ_j de la pasul M al algoritmului EM și formula pentru recalcularea centroizilor μ_j la pasul al doilea al algoritmului K -means (a se vedea *în special* varianta lui K -means de la problema 45, unde se folosesc variabilele-indicator γ_{ij}).

În final, putem face următoarea *generalizare*, pentru mixturi de $K > 2$ gaussiene unidimensionale:

$$\sum_{j=1}^K \pi_j N(x; \mu_j, \sigma_j^2).$$



Este de observat faptul că, în acest caz, în mixtură, distribuția Bernoulli este înlocuită cu o distribuție categorială, ale cărei valori $(1, \dots, K)$ sunt luate cu probabilitățile π_1, \dots, π_K . Pentru deducerea algoritmului EM în acest caz, singura schimbare de fond necesară în raport cu demonstrația de mai sus se referă la actualizarea parametrilor π_j :

Întrucât $\pi_1 + \dots + \pi_K = 1$, va trebui să rezolvăm următoarea *problemă de optimizare cu restricții*:

⁸⁵⁸Ideea demonstrației este similară cu cea de la problema 53.a de la capitolul de *Fundamente*. Și anume, după fiecare etapă de calcul al unei derivate partiale de la pasul M se pot scrie relații de forma

$$\begin{aligned} Q(\pi_1, \pi_2, \mu, \sigma | \pi^{(t)}, \mu^{(t)}, \sigma^{(t)}) &\leq Q(\pi_1^{(t+1)}, \pi_2, \mu, \sigma | \pi^{(t)}, \mu^{(t)}, \sigma^{(t)}), \forall \pi_1, \pi_2, \mu, \sigma \\ Q(\pi_1^{(t+1)}, \pi_2, \mu, \sigma | \pi^{(t)}, \mu^{(t)}, \sigma^{(t)}) &\leq Q(\pi_1^{(t+1)}, \pi_2^{(t+1)}, \mu, \sigma | \pi^{(t)}, \mu^{(t)}, \sigma^{(t)}), \forall \pi_2, \mu, \sigma \end{aligned}$$

s.a.m.d.

$$\max_{\pi, \mu, \sigma} Q(\pi, \mu, \sigma | \pi^{(t)}, \mu^{(t)}, \sigma^{(t)})$$

$$\text{cu restricția } \sum_{j=1}^K \pi_j = 1,$$

în care expresia lui Q provine din relația (347).⁸⁵⁹ Folosind *metoda multiplicatorilor Lagrange*, vom introduce variabila $\lambda \in \mathbb{R}$, iar problema de mai sus va deveni:

$$\max_{\pi, \mu, \sigma} \left(Q(\pi, \mu, \sigma | \pi^{(t)}, \mu^{(t)}, \sigma^{(t)}) + \lambda \left(1 - \sum_{j=1}^K \pi_j \right) \right).$$

Derivând funcția obiectiv a acestei probleme în raport cu variabila π_j , unde $j \in \{1, \dots, K\}$, obținem:

$$\frac{\partial}{\partial \pi_j} Q(\mu, \sigma, \pi | \mu^{(t)}, \sigma^{(t)}, \pi^{(t)}) = 0 \Leftrightarrow \sum_{i=1}^n p_{ij}^{(t)} \frac{1}{\pi_j} = \lambda \Leftrightarrow \pi_j^{(t+1)} = \frac{1}{\lambda} \sum_{i=1}^n p_{ij}^{(t)}.$$

Întrucât $\sum_{j=1}^K \pi_j^{(t+1)} = 1$, urmează că

$$\lambda = \sum_{j=1}^K \sum_{i=1}^n p_{ij}^{(t)} = \underbrace{\sum_{i=1}^n \sum_{j=1}^K p_{ij}^{(t)}}_1 = \sum_{i=1}^n 1 = n.$$

Așadar,

$$\pi_j^{(t+1)} \leftarrow \frac{1}{n} \sum_{i=1}^n p_{ij}^{(t)}.$$

Se observă că $\pi_j^{(t+1)} \geq 0$ pentru orice j , întrucât termenii $p_{ij}^{(t)}$ reprezintă niște probabilități (vedeți pasul E).

În concluzie, corpul iterativ al algoritmului EM pentru estimarea [tuturor] parametrilor unei mixturi de K distribuții gaussiene este:

Pasul E: Pentru $i = 1, \dots, n$ și $j = 1, \dots, K$, calculează

$$p_{ij}^{(t)} \stackrel{\text{not.}}{=} P(z_{ij} = 1 | x_i; \mu^{(t)}, (\sigma^2)^{(t)}, \pi^{(t)}) = \frac{\mathcal{N}(x_i | \mu_j^{(t)}, (\sigma_j^2)^{(t)}) \cdot \pi_j^{(t)}}{\sum_{l=1}^K \mathcal{N}(x_i | \mu_l^{(t)}, (\sigma_l^2)^{(t)}) \cdot \pi_l^{(t)}}$$

$$\text{unde } \mathcal{N}(x_i | \mu_j, \sigma_j^2) \stackrel{\text{def.}}{=} \frac{1}{\sqrt{2\pi}\sigma_j} \cdot \exp\left(-\frac{(x_i - \mu_j)^2}{2\sigma_j^2}\right).$$

Pasul M: Pentru $j = 1, \dots, K$, calculează

$$\begin{aligned} \pi_j^{(t+1)} &\leftarrow \frac{1}{n} \sum_{i=1}^n p_{ij}^{(t)} \\ \mu_j^{(t+1)} &\leftarrow \frac{\sum_{i=1}^n p_{ij}^{(t)} x_i}{\sum_{i=1}^n p_{ij}^{(t)}} \\ (\sigma_j^{(t+1)})^2 &\leftarrow \frac{\sum_{i=1}^n p_{ij}^{(t)} (x_i - \mu_j^{(t+1)})^2}{\sum_{i=1}^n p_{ij}^{(t)}} \end{aligned}$$

⁸⁵⁹Observație: Nu mai adăugăm la această problemă de optimizare și restricțiile $\pi_j \geq 0$ pentru $j = 1, \dots, K$, întrucât vom arăta la final că soluțiile obținute satisfac în mod implicit aceste restricții.

Evident, aceste formule le generalizează pe cele care au fost deduse anterior, când s-a considerat $K = 2$.

Observație (5): Se va putea vedea la problema 24 că formulele care au fost obținute aici pot fi derivate și pe altă cale, și anume particularizând numărul de dimensiuni (d) la valoarea 1 în formulele care alcătuiesc corpul iterativ al algoritmului EM pentru mixturi de distribuții gaussiene multidimensionale (vedeți pag. 900).

19.

(Algoritmul EM:

chestiuni calitative / metodologice privind aplicarea
în cazul unei mixturi de două distribuții gaussiene unidimensionale
presupunând probabilitățile de selecție egale și fixate)

■ • ○ CMU, 2007 spring, Eric Xing, final exam, pr. 1.8

A fost odată, cu mult timp în urmă, un sat care era situat într-o regiune cu sute de lacuri. În acele lacuri trăiau două specii de pești, însă în fiecare lac, nu erau decât pești dintr-o singură specie. (Peștii din lacuri diferite puteau fi de specii diferite.) Peștii din aceste două specii arătau identic, miroseau identic atunci când erau gătiți și aveau exact același gust delicios, însă una dintre specii era otrăvitoare și oricine mâncă pești din specia respectivă murea. Singura diferență observabilă la aceste specii de pești consta în efectul asupra nivelului pH-ului (aciditatea) apei din lacul în care trăiau. Lacurile cu pești neotrăvitori aveau pH-ul distribuit conform unei distribuții gaussiene de medie (μ_{sigur}) și varianță (σ_{sigur}^2) necunoscute, iar pH-ul din lacurile cu pești otrăvitori era distribuit conform unei alte distribuții gaussiene de medie (μ_{mortal}) și varianță (σ_{mortal}^2) de asemenea necunoscute. (Peștii otrăvitori cauzau o aciditate care era în general mai mare / ridicată decât în cazul celorlalți pești.)

Așa cum era firesc, pentru a ieși din încurcătură — adică pentru a determina caracterul otrăvitor ori neotrăvitor al peștilor din fiecare lac (sau dintr-un lac oarecare, din care încă nu s-au consumat pești) în funcție de nivelul pH-ului apei —, sătenii au apelat la învățarea automată. Cu toate acestea, au avut loc dezbateri aprinse cu privire la modalitatea corectă de aplicare a algoritmului Expectation-Maximization la problema lor.

Pentru fiecare dintre modalitățile prezentate mai jos, indicați dacă ea constituie o aplicare corectă a algoritmului EM și dacă va conduce la o estimare rezonabilă pentru parametrii μ și σ^2 ai fiecărei clase.

a. Se ghicesc valorile inițiale pentru parametrii μ și σ^2 corespunzători fiecărei specii. (1) Pentru fiecare lac, se determină — folosind teorema lui Bayes și distribuțiile gaussiene de parametri μ și σ^2 — care este cea mai probabilă specie de pești asociată lacului respectiv. (2) Se recalculează valorile pentru μ și σ^2 utilizând metoda estimării de verosimilitate maximă. Se repetă iterativ pașii (1) și (2) până se ajunge la convergență.

b. Pentru fiecare lac, se ghicește probabilitatea (inițială) că lacul respectiv ar fi populat cu pești neotrăvitori. (1) Folosind aceste probabilități, se estimează în sensul verosimilității maxime valorile μ și σ corespunzătoare fiecărei clase. (2) Utilizând aceste estimări pentru μ și σ , se recalculează probabilitățile de

‘siguranță’ pentru fiecare lac. Se repetă iterativ pașii (1) și (2) până se ajunge la convergență.

c. Se calculează media și varianța nivelului de pH pentru toată multimea lacurilor. Se folosesc aceste valori pentru a inițializa parametrii μ și σ^2 corespunzători fiecărei specii de pești. (1) Folosind aceste valori pentru μ și σ^2 , se calculează pentru fiecare lac probabilitatea să conțină pești otrăvitori. (2) Se găsesc valorile de verosimilitate maximă pentru μ și σ^2 . Se repetă iterativ pașii (1) și (2) până se ajunge la convergență.

Răspuns:

Din enunțul problemei, în mod implicit putem presupune că $P(\text{sigur}) = P(\text{mortal}) = 1/2$. Varianta algoritmui EM/GMM utilizată în acest caz diferă doar ușor de varianta (mai generală) formulată la problema 18: probabilitățile π_{sigur} și π_{mortal} se consideră totdeauna 1/2 (deci nu se recalculează la pasul M).

a. Această modalitate de aplicare a algoritmului este cea clasică, deci corectă. Pasul (1) face în mod implicit calculul mediilor $E[z_{i,\text{sigur}}] = p(z_{i,\text{sigur}} = 1|x_i; \mu, \sigma^2)$ și $E[z_{i,\text{mortal}}] = p(z_{i,\text{mortal}} = 1|x_i; \mu, \sigma^2)$, unde notațiile folosite sunt cele clasice, iar pasul (2) reprezintă maximizarea funcției auxiliare, în spătă recalcularea parametrilor μ și σ^2 .

b. Această metodă de aplicare a algoritmului EM nu este la fel de evidentă ca metoda la punctul precedent, dar vom arăta că este totuși corectă.

O primă diferență față de metoda anterioară constă în faptul că în loc să se ghicească valorile inițiale pentru parametrii μ și σ^2 corespunzători fiecărei clase / specii, se ghicesc probabilitățile inițiale pentru ‘siguranță’ fiecărui lac, adică $p(z_{i,\text{sigur}} = 1|x_i; \mu, \sigma^2)$ (chiar dacă numărul acestor probabilități este în mod normal mult mai mare decât al parametrilor μ și σ^2). Stim că $z_{i,\text{sigur}} + z_{i,\text{mortal}} = 1$, aşadar ghicirea probabilității $p(z_{i,\text{sigur}} = 1|x_i; \mu, \sigma^2)$ antrenează în mod implicit setarea probabilității $p(z_{i,\text{mortal}} = 1|x_i; \mu, \sigma^2)$. Prin urmare, la acest pas de inițializare se încearcă să se „ghicească“ valorile care ar trebui calculate la prima execuție a pasului E din varianta clasică a algoritmului EM, dacă s-ar dispune de valori inițiale pentru parametrii μ și σ^2 .

A doua diferență față de metoda clasică constă în inversarea pașilor din corpul iterativ al algoritmului EM: pasul (1) reprezintă maximizarea, iar pasul (2) estimarea.

Se observă că aceste două „diferențe“ se compensează reciproc: valorile pentru parametrii μ și σ^2 care sunt calculate / estimate la prima execuție a pasului de maximizare (1) — chiar dacă probabilitățile „ghicite“ inițial, pe care se bazează aceste calcule, sunt imperfecte sau chiar grav eronate — vor juca rolul valorilor inițiale din varianta (clasică) de la punctul a. După acest prim pas, prezenta variantă a algoritmului EM se comportă aidoma cu cea de la punctul a.

c. Această metodă nu este corectă / convenabilă deoarece utilizează valori inițiale egale pentru media (și respectiv pentru varianța) nivelului de pH pentru ambele clase de lacuri. Vom arăta că din acest motiv algoritmul rămâne „blocat“ la valorile calculate pentru μ și σ^2 la prima iterație. Iată explicația:

La pasul (1) – estimare, se calculează:

$$P(\text{mortal} | x_i; \mu, \sigma^2) = \frac{P(x_i | \text{mortal}, \mu, \sigma^2) \cdot P(\text{mortal})}{P(x_i | \text{mortal}, \mu, \sigma^2) \cdot P(\text{mortal}) + P(x_i | \text{sigur}, \mu, \sigma^2) \cdot P(\text{sigur})}$$

Am văzut că din enunțul problemei putem presupune că $P(\text{sigur}) = P(\text{mortal}) = 1/2$. Întrucât s-a considerat $\mu_{\text{sigur}} = \mu_{\text{mortal}}$ și $\sigma_{\text{sigur}} = \sigma_{\text{mortal}}$, rezultă că $P(x | \text{sigur}, \mu, \sigma^2) = P(x | \text{mortal}, \mu, \sigma^2)$. Deci, conform formulei lui Bayes, vom obține $P(\text{sigur} | x; \mu, \sigma^2) = 1/2$. Prin urmare, și $P(\text{mortal} | x; \mu, \sigma^2) = 1/2$.

La pasul (2) – maximizare, valorile obținute pentru parametrii μ și σ^2 vor fi:⁸⁶⁰ $\mu_{\text{sigur}} = \mu_{\text{mortal}} = \frac{1}{n} \sum_{i=1}^n x_i$, iar $\sigma_{\text{sigur}}^2 = \sigma_{\text{mortal}}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_{\text{mortal}})^2$, unde x_i sunt nivelerile de pH ale lacurilor (și anume, cele pentru care s-a măsurat acest nivel).

Se observă ușor că valorile obținute pentru $P(\text{sigur} | x, \mu, \sigma^2)$, $P(\text{mortal} | x, \mu, \sigma^2)$, μ_{sigur} , μ_{mortal} , σ_{sigur}^2 și σ_{mortal}^2 vor rămâne neschimbate după execuția primei iterării, deci practic algoritmul „staționează“ / „buclează“.

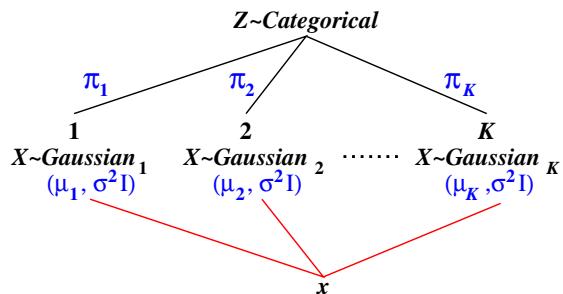
20. (Algoritmul EM/GMM, cazul multidimensional,
cu $\Sigma_j = \sigma^2 I$ pt. $j = 1, \dots, K$:
deducerea formulelor de actualizare)

■ ○ U. Utah, 2009 fall, ML (CS5350/6350), Piyush Rai

A. Fie mixtura de distribuții gaussiene

$$gmm(x) = \sum_{j=1}^K \pi_j \mathcal{N}(x | \mu_j, \sigma^2 I),$$

unde $x \in \mathbb{R}^d$, probabilitățile a priori de selecție $\pi_j \in \mathbb{R}$ satisfac (ca de obicei) restricțiile $\pi_j \geq 0$ pentru $j = 1, \dots, K$ și $\sum_{j=1}^K \pi_j = 1$, mediile gaussienelor $j = 1, \dots, K$ sunt vectorii $\mu_j \in \mathbb{R}^d$, iar matricele de covarianță ale acestor gaussiene sunt identice, ba chiar au forma particulară $\sigma^2 I$, cu $\sigma \in \mathbb{R}$ și $\sigma > 0$, matricea I fiind matricea identitate.



Se consideră instanțele $x_1, \dots, x_n \in \mathbb{R}^d$ generate cu distribuția probabilistă de mai sus (gmm). Vom asocia fiecărei instanțe x_i un vector-indicator (mai precis, un vector de variabile aleatoare $z_i \in \{0, 1\}^K$, cu $z_i \stackrel{\text{not.}}{=} (z_{i1}, \dots, z_{iK})$ și $z_{ij} = 1$ dacă și numai dacă x_i a fost generat de gaussiana $\mathcal{N}(x | \mu_j, \sigma^2 I)$.

Deducreți regulile de actualizare din cadrul [pasului E și al pasului M al] algoritmului EM. După cum știm, acest algoritm face estimarea parametrilor $\pi \stackrel{\text{not.}}{=} (\pi_1, \dots, \pi_K)$, $\mu \stackrel{\text{not.}}{=} (\mu_1, \dots, \mu_K)$ și σ , asigurând la convergență atingerea

⁸⁶⁰A se vedea regulile de actualizare de la problema 18.

unui maxim local pentru funcția de log-verosimilitate a datelor „observabile“ x_1, \dots, x_n .

Sugestie: Pentru rezolvarea problemei, vă recomandăm să parcurgeți următoarele etape:

- a. Se știe că expresia funcției de densitate a distribuției gaussiene multidimensionale (mai precis, d -dimensionale) de medie $\mu \in \mathbb{R}^d$ și matrice de covarianță $\Sigma \in \mathbb{R}^{d \times d}$ este

$$\frac{1}{(\sqrt{2\pi})^d \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right),$$

unde x și μ sunt considerați vectori-colonă din \mathbb{R}^d , iar operatorul \top desemnează operația de transpunere a vectorilor / matricelor.

Aduceți expresia de mai sus la forma cea mai simplă pentru cazul $\Sigma = \sigma^2 I$.

Vă recomandăm să folosiți faptul că $\|x - \mu\|^2 = (x - \mu)^\top (x - \mu) = (x - \mu) \cdot (x - \mu)$, unde operatorul \cdot desemnează produsul scalar al vectorilor.

- b. Pentru *pasul E* al algoritmului EM, veți demonstra mai întâi că media $E[z_{ij}] \stackrel{\text{not.}}{=} E[z_{ij}|x_i; \pi, \mu, \sigma]$, unde $x_i = (x_{i,1}, \dots, x_{i,d}) \in \mathbb{R}^d$, $\mu \stackrel{\text{not.}}{=} (\mu_1, \dots, \mu_K) \in (\mathbb{R}^d)^K$ și $\sigma \in \mathbb{R}^+$, are valoarea $P(z_{ij} = 1|x_i; \pi, \mu, \sigma)$, iar apoi veți elabora formula de calcul a acestei probabilități, folosind teorema lui Bayes.

- c. Arătați că expresia funcției de log-verosimilitate a datelor „complete“ în raport cu parametrii π, μ și σ este

$$\ln p(x, z|\pi, \mu, \sigma) = \sum_{i=1}^n \sum_{j=1}^K z_{ij} (\ln \pi_j + \ln \mathcal{N}(x_i|\mu_j, \sigma^2 I)),$$

unde $x \stackrel{\text{not.}}{=} (x_1, \dots, x_n)$ și $z \stackrel{\text{not.}}{=} (z_1, \dots, z_n)$.

Deducreți apoi expresia „funcției auxiliare“ $Q(\pi, \mu, \sigma|\pi^{(t)}, \mu^{(t)}, \sigma^{(t)}) \stackrel{\text{def.}}{=} E[\ln p(x, z|\pi, \mu, \sigma)]$, cu precizarea că media aceasta este calculată în raport cu distribuția / distribuțiile $P(z_{ij}|x_i, \pi^{(t)}, \mu^{(t)}, \sigma^{(t)})$.

Pasul M: Problema de optimizare EM în acest context este

$$(\pi^{(t+1)}, \mu^{(t+1)}, \sigma^{(t+1)}) = \underset{\pi, \mu, \sigma}{\operatorname{argmax}} Q(\pi, \mu, \sigma|\pi^{(t)}, \mu^{(t)}, \sigma^{(t)}), \quad (348)$$

cu restricțiile $\pi_j^{(t+1)} \geq 0$ pentru $j = 1, \dots, K$ și $\sum_{j=1}^K \pi_j^{(t+1)} = 1$.

Această problemă se rezolvă optimizând funcția ei obiectiv în mod separat în raport cu variabilele π, μ și σ .

- d. Aplicați metoda multiplicatorilor lui Lagrange pentru a rezolva problema de optimizare cu restricții (348) în raport (doar) cu variabilele π .

- e. Optimizați funcția Q (i.e., rezolvați problema de optimizare (348)) în raport cu variabilele μ .

Indicație: Următoarea formulă de *derivare vectorială* (preluată din documentul *Matrix Identities*, de Sam Roweis, 1999) vă poate fi de folos:

$$(5g) \quad \frac{\partial}{\partial X} (Xa + b)^\top C(Xa + b) = (C + C^\top)(Xa + b)a^\top$$

f. Optimizați funcția Q (i.e., rezolvați problema de optimizare (348)) în raport cu variabila σ .

g. Sumarizați rezultatele obținute la punctele de mai sus (b și $d-f$), redactând în pseudocod algoritmul EM pentru rezolvarea mixturii de gaussiene din enunț.

B. Analizați ce se întâmplă atunci când modelul de mixtură de mai sus este generalizat la gaussiene cu matrice de covariantă diagonale oarecare (adică acestea nu mai sunt de forma $\sigma^2 I$). Concret, modelul devine:

$$z_i \sim \text{Categorical}(p_1, \dots, p_K)$$

$$x_i | z_i = j \sim \mathcal{N} \left(\begin{bmatrix} \mu_{j,1} \\ \vdots \\ \mu_{j,d} \end{bmatrix}, \begin{bmatrix} (\sigma_{j,1})^2 & 0 & \dots & 0 \\ 0 & (\sigma_{j,2})^2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & (\sigma_{j,d})^2 \end{bmatrix} \right)$$

Răspuns:

A. a. Întrucât $\Sigma = \sigma^2 I$, rezultă că $\Sigma^{-1} = \frac{1}{\sigma^2} I$. Avem, de asemenea, $|\Sigma| = (\sigma^2)^d$, ceea ce implică $\sqrt{|\Sigma|} = \sigma^d$, întrucât stim că $\sigma > 0$, prin definiție. Așadar,

$$\begin{aligned} \mathcal{N}(x|\mu, \Sigma = \sigma^2 I) &\stackrel{\text{def.}}{=} \frac{1}{(\sqrt{2\pi})^d \sqrt{|\Sigma|}} \exp \left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1} (x - \mu) \right) \\ &= \frac{1}{(\sqrt{2\pi})^d \sigma^d} \exp \left(-\frac{1}{2}(x - \mu)^\top \frac{1}{\sigma^2} (x - \mu) \right) \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^d} \exp \left(-\frac{1}{2\sigma^2} \|x - \mu\|^2 \right). \end{aligned}$$

b. Pentru comoditate, în cele ce urmează vom renota $E[z_{ij}|x_i; \pi, \mu, \sigma]$ prin $E[z_{ij}]$. Așadar,

$$\begin{aligned} E[z_{ij}] &\stackrel{\text{not.}}{=} E[z_{ij}|x_i; \pi, \mu, \sigma] \stackrel{\text{def.}}{=} 0 \cdot P(z_{ij} = 0|x_i; \pi, \mu, \sigma) + 1 \cdot P(z_{ij} = 1|x_i; \pi, \mu, \sigma) \\ &= P(z_{ij} = 1|x_i; \pi, \mu, \sigma) \\ &\stackrel{F. Bayes}{=} \frac{P(x_i|z_{ij} = 1; \pi, \mu, \sigma) \cdot P(z_{ij} = 1; \pi, \mu, \sigma)}{P(x_i; \pi, \mu, \sigma)} \\ &\stackrel{F.P.T.}{=} \frac{P(x_i|z_{ij} = 1; \pi, \mu, \sigma) \cdot P(z_{ij} = 1; \pi, \mu, \sigma)}{\sum_{j'=1}^K P(x_i|z_{ij'} = 1; \pi, \mu, \sigma) \cdot P(z_{ij'} = 1; \pi, \mu, \sigma)} \\ &\stackrel{a.}{=} \frac{\frac{1}{(\sqrt{2\pi}\sigma)^d} \exp \left(-\frac{1}{2\sigma^2} \|x_i - \mu_j\|^2 \right) \pi_j}{\sum_{j'=1}^K \frac{1}{(\sqrt{2\pi}\sigma)^d} \exp \left(-\frac{1}{2\sigma^2} \|x_i - \mu_{j'}\|^2 \right) \pi_{j'}} \\ &= \frac{\pi_j \exp \left(-\frac{1}{2\sigma^2} \|x_i - \mu_j\|^2 \right)}{\sum_{j'=1}^K \pi_{j'} \exp \left(-\frac{1}{2\sigma^2} \|x_i - \mu_{j'}\|^2 \right)}. \end{aligned} \tag{349}$$

c. Log-verosimilitatea datelor complete este

$$\begin{aligned}
 \ln p(x, z | \pi, \mu, \sigma) &\stackrel{\text{def.}}{=} \ln p((x_1, z_1), \dots, (x_n, z_n) | \pi, \mu, \sigma) \stackrel{i.i.d.}{=} \ln \prod_{i=1}^n p(x_i, z_i | \pi, \mu, \sigma) \\
 &= \sum_{i=1}^n \ln p(x_i, z_i | \pi, \mu, \sigma) \stackrel{\text{reg. mult.}}{=} \sum_{i=1}^n \ln \left(p(x_i | z_i; \pi, \mu, \sigma) \cdot \underbrace{p(z_i | \pi, \mu, \sigma)}_{\pi_j} \right) \\
 &= \sum_{i=1}^n \sum_{j=1}^K z_{ij} [\ln \mathcal{N}(x_i | \mu_j, \sigma) + \ln \pi_j],
 \end{aligned}$$

întrucât $z_i = (z_{i,1}, \dots, z_{i,j}, \dots, z_{i,K})$, cu $z_{i,j} = 1$ și $z_{i,j'} = 0$ pentru orice $j' \neq j$.

Acum, folosind rezultatul pe care l-am obținut la punctul a, putem scrie:

$$\ln p(x, z | \pi, \mu, \sigma) = \sum_{i=1}^n \sum_{j=1}^K z_{ij} \left[-\frac{d}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|x_i - \mu_j\|^2 + \ln \pi_j \right].$$

În fine, folosind proprietatea de liniaritate a mediei, obținem:

$$\begin{aligned}
 Q(\pi, \mu, \sigma | \pi^{(t)}, \mu^{(t)}, \sigma^{(t)}) &\stackrel{\text{def.}}{=} E[\ln p(x, z | \pi, \mu, \sigma)] \\
 &= \sum_{i=1}^n \sum_{j=1}^K E[z_{ij}] \left[-\frac{d}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|x_i - \mu_j\|^2 + \ln \pi_j \right].
 \end{aligned}$$

d. Deocamdată nu vom lua în calcul restricțiile $\pi_j \geq 0$. În consecință, *funcția lagrangeană* asociată problemei noastre de optimizare este:

$$Q(\pi, \mu, \sigma | \pi^{(t)}, \mu^{(t)}, \sigma^{(t)}) + \lambda \left(1 - \sum_{j=1}^K \pi_j \right), \quad (350)$$

unde $\lambda \in \mathbb{R}$ este *multiplicator Lagrange*. (După ce vom obține soluția acestei probleme de optimizare se va vedea că $\pi_j \geq 0$ pentru orice $j = 1, \dots, K$.)

Calculând derivata parțială a funcției Lagrange (350) în raport cu π_j (cu j fixat în mulțimea $\{1, \dots, K\}$) și apoi egalând-o cu zero,

$$\begin{aligned}
 \frac{\partial}{\partial \pi_j} (Q(\pi, \mu, \sigma | \pi^{(t)}, \mu^{(t)}, \sigma^{(t)}) + \lambda \left(1 - \sum_{j=1}^K \pi_j \right)) &= 0 \Leftrightarrow \\
 \sum_{i=1}^n E[z_{ij}] \frac{1}{\pi_j} &= \lambda,
 \end{aligned}$$

vom obține soluția

$$\pi_j^{(t+1)} = \frac{1}{\lambda} \sum_{i=1}^n E[z_{ij}]. \quad (351)$$

Apoi, ținând cont că aceste valori trebuie să satisfacă restricția $\sum_{j=1}^K \pi_j^{(t+1)} = 1$, rezultă că

$$\sum_{j=1}^K \frac{1}{\lambda} \sum_{i=1}^n E[z_{ij}] = 1 \Leftrightarrow \frac{1}{\lambda} \sum_{j=1}^K \sum_{i=1}^n E[z_{ij}] = 1 \stackrel{b.}{\Leftrightarrow} \frac{1}{\lambda} \sum_{i=1}^n \sum_{j=1}^K P(z_{ij} = 1 | x_i, \pi, \mu, \sigma) = 1 \Leftrightarrow$$

$$\frac{1}{\lambda} \sum_{i=1}^n \underbrace{\sum_{j=1}^K P(z_{ij} = 1 | x_i, \pi, \mu, \sigma)}_1 = 1 \Leftrightarrow \frac{1}{\lambda} \sum_{i=1}^n 1 = 1 \Leftrightarrow \frac{n}{\lambda} = 1 \Leftrightarrow \lambda = n.$$

Înlocuind această valoare a lui λ în relația (351), obținem

$$\pi_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^n E[z_{ij}]. \quad (352)$$

Este evident acum că $\pi_j^{(t+1)} \geq 0$.

e. Vă readucem aminte notația $\mu = (\mu_1, \dots, \mu_K) \in (\mathbb{R}^d)^K$, cu $\mu_j \in \mathbb{R}^d$ pentru $j = 1, \dots, K$. În loc să calculăm derivată parțială a funcției „auxiliare“ Q în raport cu fiecare componentă a vectorului μ_j , vom calcula direct gradientul lui Q în raport cu μ_j :

$$\begin{aligned} & \nabla_{\mu_j} Q(\pi, \mu, \sigma | \pi^{(t)}, \mu^{(t)}, \sigma^{(t)}) \\ &= \nabla_{\mu_j} \sum_{i=1}^n \sum_{j'=1}^K E[z_{ij'}] \left[-\frac{d}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|x_i - \mu'_j\|^2 + \ln \pi_{j'} \right] \\ &= \sum_{i=1}^n E[z_{ij}] \left[-\frac{1}{2\sigma^2} \nabla_{\mu_j} (x_i - \mu_j)^2 \right] \stackrel{(5g)}{=} -\frac{1}{2\sigma^2} \sum_{i=1}^n E[z_{ij}] 2(x_i - \mu_j)(-1) \\ &= \frac{1}{\sigma^2} \sum_{i=1}^n E[z_{ij}] (x_i - \mu_j) = \frac{1}{\sigma^2} \left[\left(\sum_{i=1}^n E[z_{ij}] x_i \right) - \left(\sum_{i=1}^n E[z_{ij}] \right) \mu_j \right]. \end{aligned}$$

Egalând această expresie cu zero (de fapt, vectorul-colonă $(0, \dots, 0)^\top \in \mathbb{R}^d$), vom obține soluția:

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^n E[z_{ij}] x_i}{\sum_{i=1}^n E[z_{ij}]} \quad (353)$$

f. Derivata parțială a funcției „auxiliare“ $Q(\pi, \mu, \sigma | \pi^{(t)}, \mu^{(t)}, \sigma^{(t)})$ în raport cu σ este:

$$\begin{aligned} & \frac{\partial}{\partial \sigma} Q(\pi, \mu, \sigma | \pi^{(t)}, \mu^{(t)}, \sigma^{(t)}) \\ &= \frac{\partial}{\partial \sigma} \sum_{i=1}^n \sum_{j=1}^K E[z_{ij}] \left[-\frac{d}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|x_i - \mu_j\|^2 + \ln \pi_j \right] \\ &= \sum_{i=1}^n \sum_{j=1}^K E[z_{ij}] \frac{\partial}{\partial \sigma} \left[-\frac{d}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|x_i - \mu_j\|^2 \right] \\ &= \sum_{i=1}^n \sum_{j=1}^K E[z_{ij}] \left[-\frac{d}{2} \cdot \frac{2\sigma}{\sigma^2} - \frac{-2}{2\sigma^3} \|x_i - \mu_j\|^2 \right] = \sum_{i=1}^n \sum_{j=1}^K E[z_{ij}] \left[-\frac{d}{\sigma} + \frac{1}{\sigma^3} \|x_i - \mu_j\|^2 \right] \\ &= \frac{1}{\sigma^3} (-d\sigma^2 \underbrace{\sum_{i=1}^n \sum_{j=1}^K E[z_{ij}]}_1 + \sum_{i=1}^n \sum_{j=1}^K E[z_{ij}] \|x_i - \mu_j\|^2) \\ &= \frac{1}{\sigma^3} (-nd\sigma^2 + \sum_{i=1}^n \sum_{j=1}^K E[z_{ij}] \|x_i - \mu_j\|^2) \end{aligned}$$

Egalând această expresie cu 0, obținem soluția:

$$(\sigma^{(t+1)})^2 = \frac{1}{nd} \sum_{i=1}^n \sum_{j=1}^K E[z_{ij}] \|x_i - \mu_j^{(t+1)}\|^2 \geq 0. \quad (354)$$

Nu este dificil de demonstrat faptul că această soluție corespunde maximizării funcției Q în raport cu σ^2 . (Se știe că $\sigma > 0$, iar expresia $-nd\sigma^2 + \sum_{i=1}^n \sum_{j=1}^K E[z_{ij}] \|x_i - \mu_j\|^2$, văzută ca funcție de σ^2 , este liniară și descrescătoare.)

g. Putem acum să scriem pseudo-codul algoritmului EM pentru rezolvarea acestei mixturi de distribuții gaussiene multidimensionale, cu ajutorul relațiilor (349), (352), (353) și (354):

- Inițializează [cu valori arbitrale] probabilitățile a priori π , mediile μ și varianța σ^2 .
- Iterează până când o anumită *condiție de oprire* este satisfăcută:

Pasul E: Calculează mediile (adică, probabilitățile a posterorii) pentru variabilele z :

$$p_{ij}^{(t)} \stackrel{\text{not.}}{=} E[z_{ij}|x_i; \pi^{(t)}, \mu^{(t)}, \sigma^{(t)}] = \frac{\pi_j^{(t)} \mathcal{N}(x_i | \mu_j^{(t)}, (\sigma^{(t)})^2 I)}{\sum_{j'=1}^K \pi_{j'}^{(t)} \mathcal{N}(x_i | \mu_{j'}^{(t)}, (\sigma^{(t)})^2 I)};$$

Pasul M: Calculează noile valori pentru π, μ și σ :

$$\begin{aligned} \pi_j^{(t+1)} &= \frac{1}{n} \sum_{i=1}^n p_{ij}^{(t)}; \\ \mu_j^{(t+1)} &= \frac{\sum_{i=1}^n p_{ij}^{(t)} x_i}{\sum_{i=1}^n p_{ij}^{(t)}}; \\ (\sigma^{(t+1)})^2 &= \frac{1}{dn} \sum_{i=1}^n \sum_{j=1}^K p_{ij}^{(t)} \|x_i - \mu_j^{(t)}\|^2. \end{aligned}$$

B. Se poate demonstra ușor (urmând linia demonstrației de la secțiunea A) că singura *regulă de actualizare* care se schimbă la reformularea algoritmului EM este (354). Ea devine:

$$(\sigma_{jk}^{(t+1)})^2 = \frac{\sum_{i=1}^n E[z_{ij}](x_{i,k} - \mu_{j,k}^{(t+1)})^2}{\sum_{i=1}^n E[z_{ij}]} \geq 0 \text{ pentru } j = 1, \dots, K \text{ și } k = 1, \dots, d. \quad (355)$$

Acesta este exact lucrul la care ne aşteptăm, date fiind rezultatele pe care le-am demonstrat / obținut la problema 20 de la capitolul de *Fundamente* și la problema 18 din prezentul capitol (i.e., *Clusterizare*).

Observație importantă: La rezolvarea problemelor 21, 22 și 23 am folosit varianta de algoritm EM/GMM (cazul multidimensional) descris la problema 20, unde se consideră că matricele de covarianță sunt diagonale, ba chiar $\Sigma_j = \sigma^2 I_d$, unde I_d este matricea identitate de tip pătratic cu d linii și d coloane.

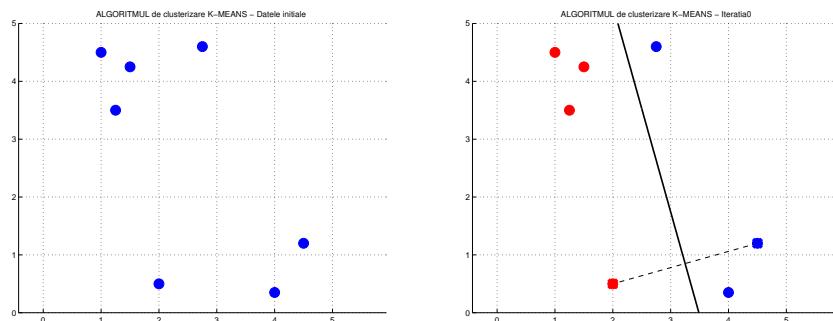
21. (Algoritmii K -means și EM/GMM: aplicare pe date din \mathbb{R}^2 ; compararea pozițiilor finale ale centroizilor / mediilor)

CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, midterm, pr. 4.a

Se consideră un set de date / instanțe din planul bidimensional, conform tabelului de mai jos (partea dreaptă).

- a. Folosind algoritmul K -means, se vor forma două clustere. Centroizii inițiali ai celor două clustere sunt instanțele 5 și respectiv 7. Componența inițială a clusterelor este definită ca de obicei, cu ajutorul unui separator liniar. (Vedeți imaginea de mai jos, partea dreaptă; instanțele sunt reprezentate prin cerculețe.)

Instanță	x	y
P_1	1.50	4.25
P_2	1.25	3.50
P_3	1.00	4.50
P_4	2.75	4.60
P_5	4.50	1.20
P_6	4.00	0.35
P_7	2.00	0.50



Pentru fiecare iterație executată de algoritmul K -means pe aceste date, veți desena

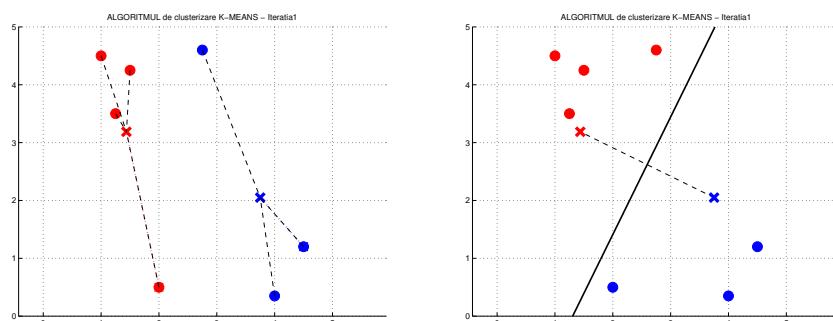
- centroizii clusterelor (reprezentați cu semnul \times);
- separatorii liniari pentru cele două clustere.

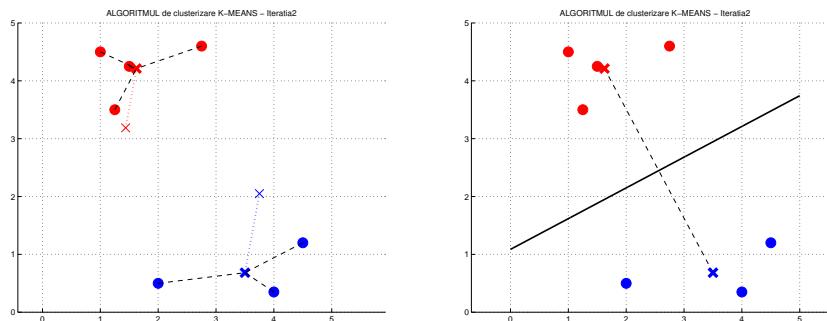
Folosiți oricără diagraame sunt necesare, până se ajunge la convergență.

- b. Cum va dифeri aplicarea pe aceste date a algoritmului EM pentru o mixtură de două distribuții gaussiene bidimensionale față de aplicarea algoritmului K -means?

Răspuns:

- a. Aplicând algoritmul K -means, obținem următoarea evoluție a clusterelor:

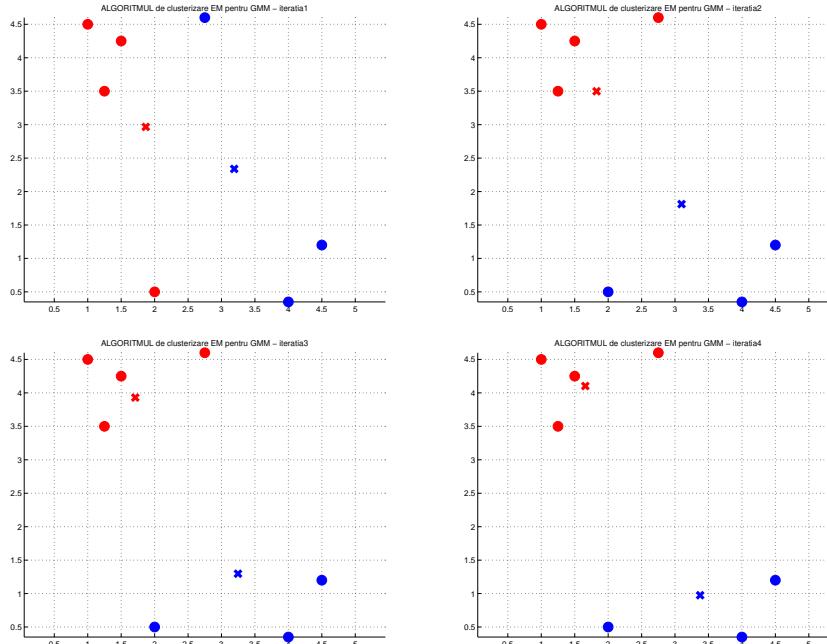




b. La aplicarea algoritmului EM, mediile celor două distribuții (corespondențul centroizilor clusterelor în K -means) pot evoluă mai încet de la o iterare la alta, deoarece sunt influențate de toate instanțele de antrenament (nu doar de cele din clusterul corespunzător, ca în cazul algoritmului K -means). Așadar, am putea avea:

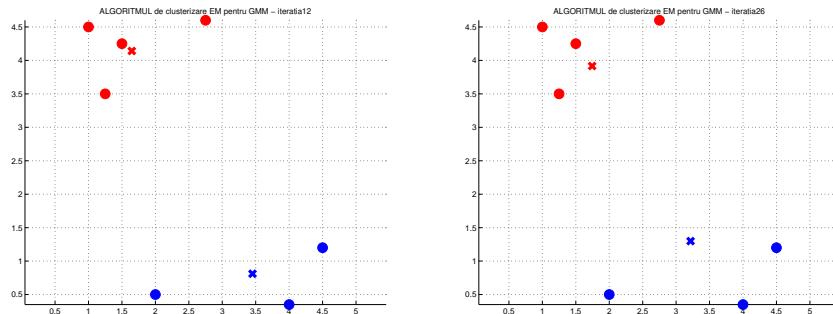
- mai multe interații până la convergență;
- o altă componentă a clusterelor la final;
- o altă poziționare finală a mediilor / centroizilor, chiar dacă se pleacă de la aceleasi poziții inițiale ca pentru algoritmul K -means.

Pentru exemplificare, dacă se aplică algoritmul EM pe aceste date, considerând varianța $\sigma^2 = 2$ (valoare fixată și identică pentru ambele distribuții gaussiene și pentru fiecare dintre cele două axe de coordonate), iar probabilitățile a priori de selecție ale celor gaussiene $1/2$, atunci primele 4 interații vor fi:



Se observă că „deplasarea“ mediilor diferă față de cea a centroizilor de la algoritmul K -means este, într-adevăr, mai „lentă“.

Pe datele considerate, algoritmul se oprește după 12 îterări, rezultatul final fiind cel din figura de mai jos, în partea stângă. Se observă că pozițiile finale ale mediilor sunt foarte apropiate de cele ale centroizilor de la terminarea algoritmului K -means.

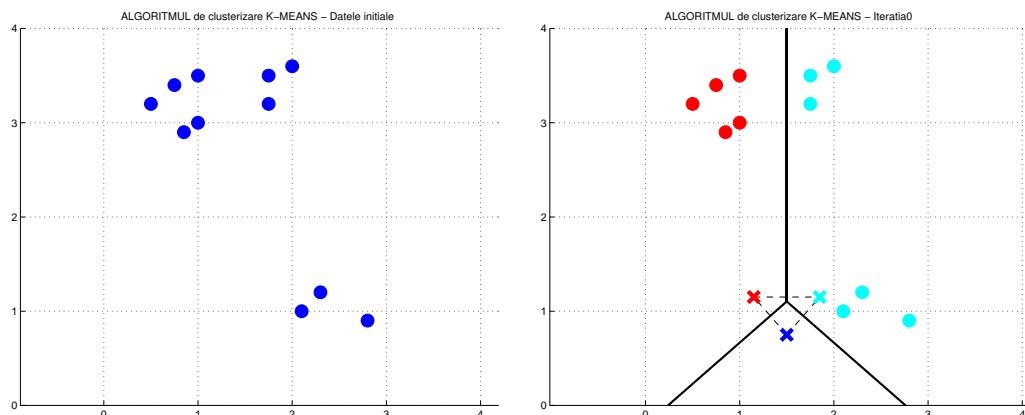


Dacă se consideră $\sigma^2 = 3$, atunci vor fi necesare mai multe îterări (și anume, 26) până la convergență, iar rezultatul este cel din figura de mai sus, în partea dreaptă. Se observă că, deși se obțin aceleași clustere, pozițiile finale ale mediilor sunt diferite față de cazul anterior ($\sigma^2 = 2$).

22. (Algoritmii K -means și EM/GMM: aplicare pe date din \mathbb{R}^2 ; compararea clusterelor finale)

CMU, 2003 fall, T. Mitchell, A. Moore, HW7, pr. 1.c

Execuția algoritmului K -means [eventual în mod manual] pe următorul set de date. Cerculețele reprezintă instanțele de clusterizat, iar cruciulițele (\times) sunt centroizii inițiali ai clusterelor. Separatorii definesc componenta inițială a clusterelor.



În eventualitatea că veți utiliza o implementare, coordonatele datelor și a centroizilor inițiali sunt date în tabelul din partea dreaptă.

a. Pentru fiecare iterație a algoritmului, desenați centroizii și separatorii care definesc fiecare cluster. Folosiți oricâte imagini sunt necesare până veți ajunge la convergență.

Observație: La execuția algoritmului, se consideră că în cazul în care un centroid nu are puncte asignate lui, atunci el rămâne pe loc în iterată respectivă.

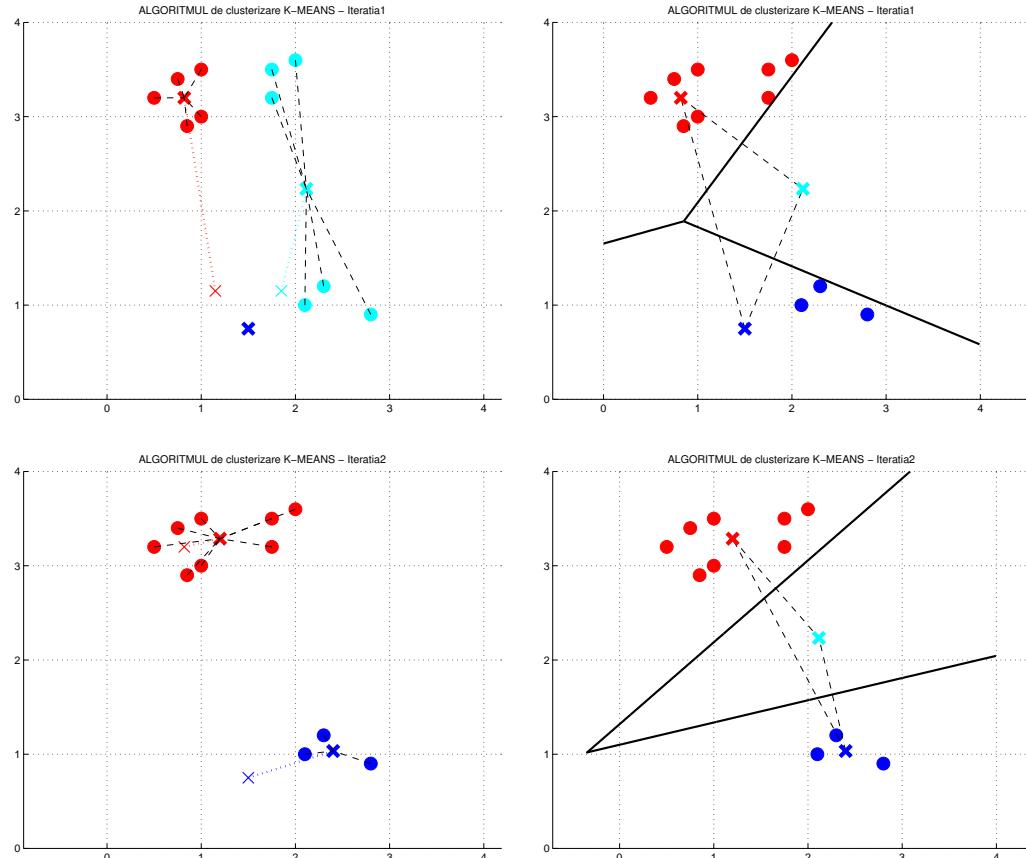
b. Dacă am rula algoritmul EM pentru o mixtură de 3 distribuții gaussiene bidimensionale pe același set de date, menținând aceeași poziționare a centroizilor inițiali ca mai sus, am ajunge la același rezultat? Justificați răspunsul. Precizați la ce clustere se va ajunge în final.

Punctele	x	y
1	0.50	3.20
2	0.75	3.40
3	1.00	3.50
4	1.00	3.00
5	0.85	2.90
6	1.75	3.20
7	1.75	3.50
8	2.00	3.60
9	2.10	1.00
10	2.30	1.20
11	2.80	0.90

Centroizii	x	y
μ_1	1.50	0.75
μ_2	1.15	1.15
μ_3	1.85	1.15

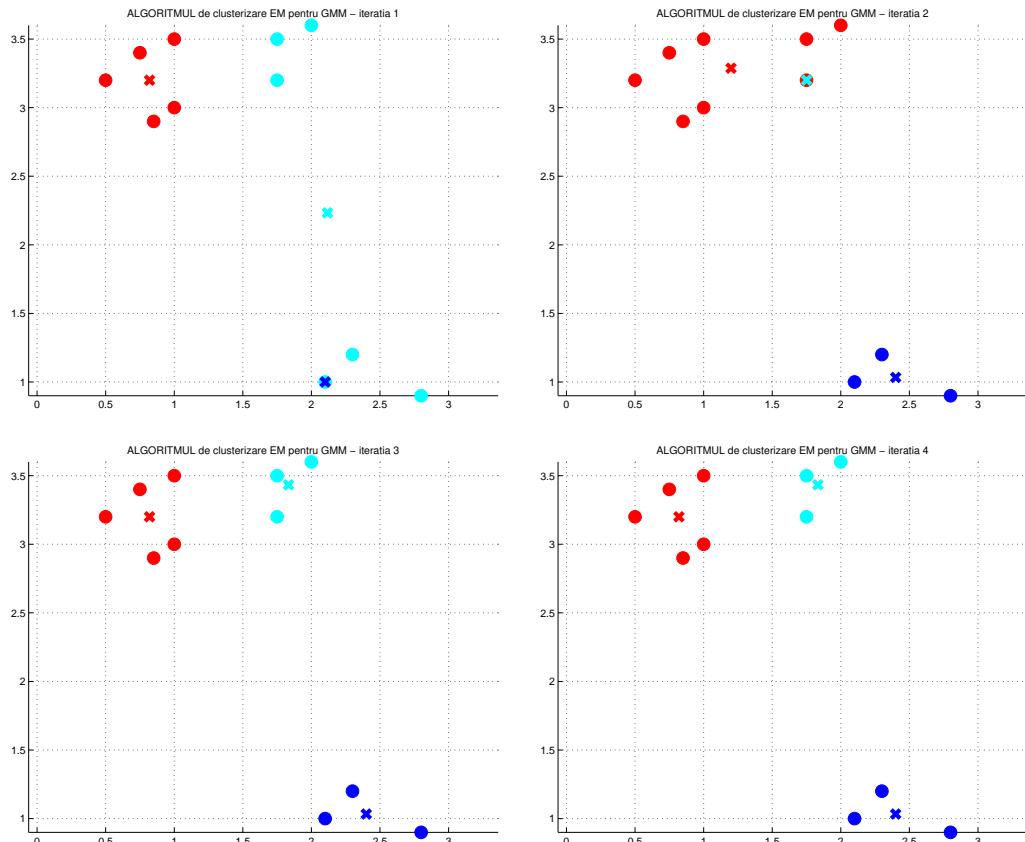
Răspuns:

a. Aplicând algoritmul K-means obținem:



Se observă că algoritmul va împărți punctele în doar două clustere, al treilea cluster obținut fiind vid.

b. Figurile de mai jos reprezintă cele 4 iterații ale algoritmului EM pe aceleasi date inițiale, considerându-se varianța fixă $\sigma^2 = 0.01$ (valoare identică pentru ambele distribuții gaussiene și pentru fiecare din cele două axe de coordonate).



Observăm că se ajunge la o altă configurație finală (mai bună) decât cea obținută de algoritmul K-means: clusterele noi au o mai mare coeziune și nu mai avem un cluster vid.

23.

(Algoritmul EM pentru GMM: aplicare pe date din \mathbb{R}^2 ; variante de „mișcare“ a mediilor;

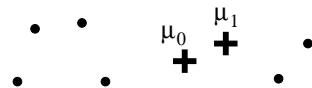
evoluția valorilor funcției de log-verosimilitate a datelor observabile)

CMU, 2010 spring, E. Xing, T. Mitchel, A. Singh, midterm, pr. 5.2

Considerăm aplicarea algoritmului EM/GMM pentru a clusteriza datele de mai jos în două clustere. Semnele + indică valorile inițiale ale mediilor (μ_0 și respectiv μ_1).⁸⁶¹

⁸⁶¹Coordonatele asociate acestor instanțe și respectiv ale mediilor vă sunt puse la dispoziție în următoarele fișiere, depuse pe site-ul acestei cărți:

<http://profes.info.uaic.ro/~ciortuz/ML.ex-book/res/CMU.2010s.EX+TM+AS.midterm.pr5.2.em.dat>,
<http://profes.info.uaic.ro/~ciortuz/ML.ex-book/res/CMU.2010s.EX+TM+AS.midterm.pr5.2.init.dat>.



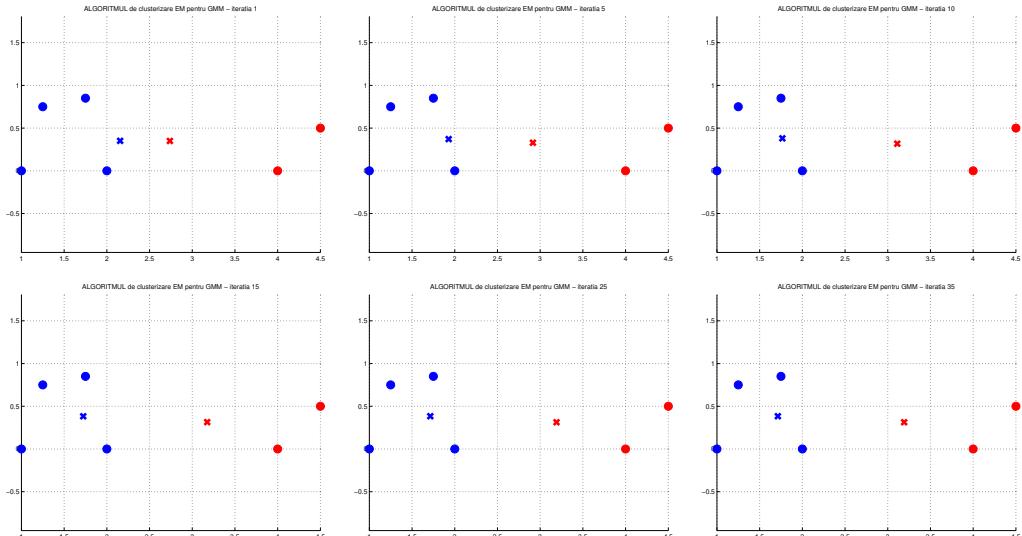
- a. În ce direcție se vor deplasa pozițiile mediilor μ_0 și μ_1 la execuția primelor iterații ale algoritmului EM?
- b. Considerând θ ansamblul parametrilor, precizați cum se comportă $\prod_j P(x_j | \theta)$, probabilitatea marginală a datelor „observabile“, după prima iterație a algoritmului EM: crește ori scade? Justificați pe scurt.

Răspuns:

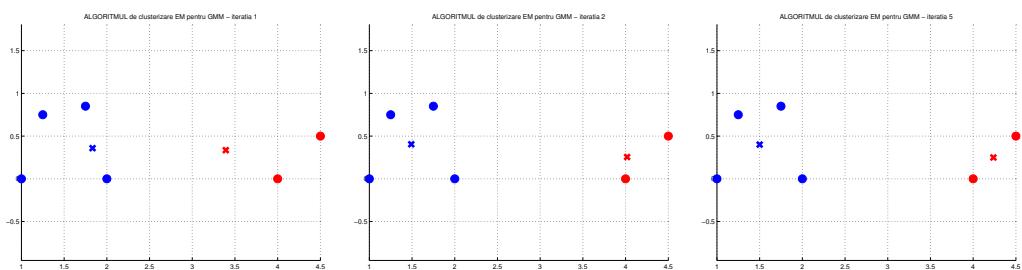
a. Folosind o implementare a variantei algoritmului EM/GMM prezentată la problema 20, vom arăta că mișcarea mediilor / centroizilor depinde de valorile varianțelor celor două distribuțiile gaussiene. Se disting următoarele cazuri:

Cazul 1: Pentru valori relativ mici ale lui σ^2 , se poate observa că la execuția algoritmului EM μ_0 se va muta din ce în ce mai mult către stânga, în vreme ce μ_1 se va muta inițial spre stânga, iar după aceea către dreapta.

De exemplu, pentru $\sigma^2 = 1.5$ se obțin următoarele iterații mai relevante până la convergență (la care se ajunge în 35 de iterații):

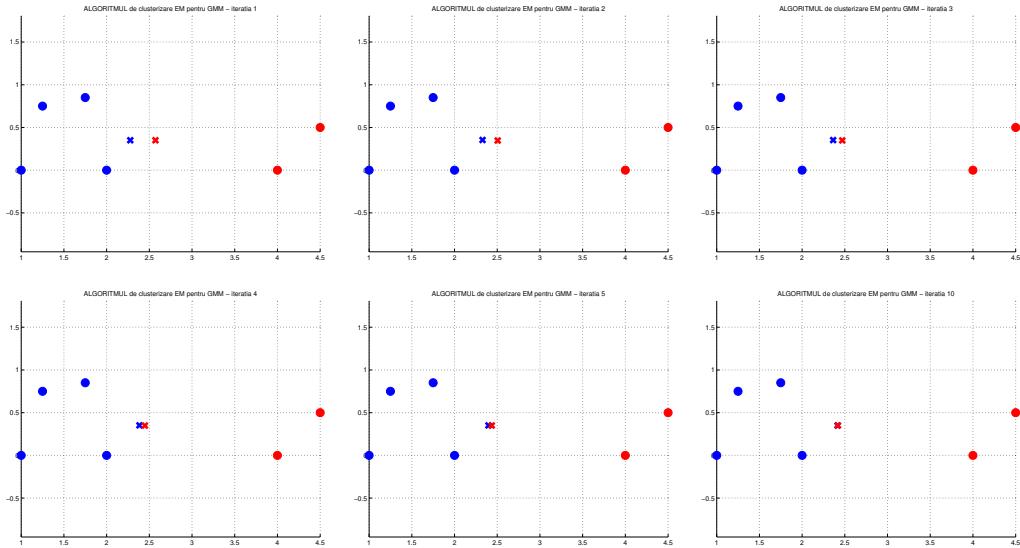


În schimb, pentru $\sigma^2 = 0.5$ (mai mic decât valoarea precedentă), se obțin următoarele iterații mai relevante până la convergență (la care se ajunge în 5 iterații):



Cazul 2: Pentru valori mai mari pentru σ^2 , se poate observa că μ_0 și μ_1 se vor deplasa unul spre celălalt, fiind posibil chiar să ajungă practic să coincidă.

De exemplu, pentru $\sigma^2 = 3$ se obțin următoarele iterații mai relevante până la convergență (la care se ajunge în 20 de iterații):



b. Probabilitatea $\prod_j P(x^j | \theta)$ reprezintă verosimilitatea datelor observabile (presupunând că aceste date sunt generate independent unele de altele). Conform rezultatului de convergență / corectitudine demonstrat la problema 2 de la capitolul *Schema algoritmică EM*, această probabilitate nu descrește — adică fie crește fie rămâne la fel — de la o iterație la alta a algoritmului EM. Creșterea se oprește doar la atingerea unui punct de optim local.

24.

(Algoritmul EM pentru mixturi de distribuții gaussiene multidimensionale; cazul general)

prelucrare de Liviu Ciortuz, după

■ ● ○ Andrew Ng, Stanford University, ML course, 2009 fall,
lecture notes, parts VIII and IX,
and CMU, 2010 fall, Aarti Singh, HW4, pr. 1.1

Presupunem, ca de obicei în învățarea automată, că avem un set de date $\{x_1, \dots, x_n\} \subset \mathbb{R}^d$. Ne vom situa în paradigma învățării nesupervizate, deci vom considera că aceste puncte nu au nicio etichetă atașată.

Dorim să modelăm aceste date prin specificarea [funcției de densitate a] unei distribuții probabiliste comune

$$p(x_i, z_i) = p(x_i|z_i)p(z_i), \text{ pentru } i = 1, \dots, n,$$

cu $z_i \sim \text{Categorial}(\pi)$ și $\pi \stackrel{\text{not.}}{=} (\pi_1, \dots, \pi_K)$.

Deci $p(x_i) = \sum_{z_i} p(x_i, z_i) = \sum_{z_i} p(x_i|z_i)p(z_i)$, formulă cu care suntem obișnuiți. K este numărul de valori pe care le pot lua variabilele z_i . Pentru $j = 1, \dots, K$,

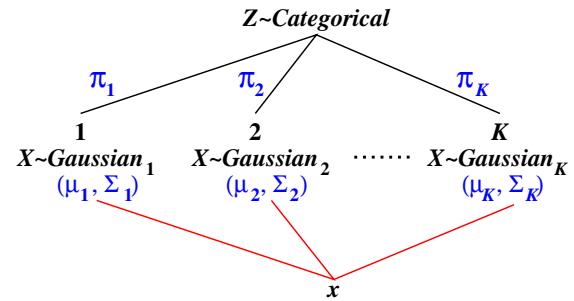
probabilitatea a priori $p(z_i = j)$ va fi desemnată prin parametrul π_j . Vom presupune că $(x_i|z_i = j) \sim \mathcal{N}(\mu_j, \Sigma_j)$. Așadar,

$$p(x_i|z_i = j; \mu_j, \Sigma_j) \stackrel{\text{def.}}{=} \frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_j)^\top \Sigma_j^{-1}(x - \mu_j)\right), \quad (356)$$

unde $\mu \in \mathbb{R}^d$, iar Σ_j este matrice de dimensiune $d \times d$, simetrică și pozitiv definită (ultima condiție este suficientă pentru a asigura inversabilitatea).⁸⁶² ⁸⁶³ Vectorii din \mathbb{R}^d sunt considerați vectori-coloană în contextul acestei probleme.

Prin urmare, modelul nostru postulează că

- fiecare instanță x_i a fost generată aleatoriu valorile variabilelor z_i din mulțimea $\{1, \dots, K\}$,
- iar apoi x_i a fost generat de una dintre cele K distribuții gaussiene, mai precis cea desemnată de z_i .



Acest model se numește *modelul mixturii de [distribuții] gaussiene*. Vom considera că z_i sunt variabile aleatoare *latente* (se mai spune că sunt variabile *ascunse*, sau *neobservabile*). Datorită acestui fapt, problema estimării în sensul MLE a parametrilor π , μ și Σ devine dificilă. (Am notat $\mu = (\mu_1, \dots, \mu_K)$ și $\Sigma = (\Sigma_1, \dots, \Sigma_K)$.)

Observație (1): Evident, dacă am cunoaște valorile variabilelor z_i , problema estimării parametrilor modelului nostru (adică π , μ și Σ) ar fi ușor de rezolvat.⁸⁶⁴ Concret, am putea scrie *verosimilitatea* datelor noastre ca

$$l(\pi, \mu, \Sigma) = \sum_{i=1}^n [\ln p(x_i|z_i; \mu, \Sigma) + \ln p(z_i; \pi)].$$

Făcând calculele, vom obține următoarele estimări MLE ale parametrilor π , μ și Σ :⁸⁶⁵

$$\pi_j = \frac{1}{n} \sum_{i=1}^n 1_{\{z_i=j\}}, \quad (357)$$

$$\mu_j = \frac{\sum_{i=1}^n 1_{\{z_i=j\}} x_i}{\sum_{i=1}^n 1_{\{z_i=j\}}}, \quad (358)$$

$$\Sigma_j = \frac{\sum_{i=1}^n 1_{\{z_i=j\}} (x_i - \mu_j)(x_i - \mu_j)^\top}{\sum_{i=1}^n 1_{\{z_i=j\}}}, \quad (359)$$

unde $1_{\{z_i=j\}}$ sunt „funcții indicator“, care desemnează gaussiana (j) care a generat o instanță oarecare (x_i) din setul de date pe care îl avem la dispoziție. Însă, atunci când valorile variabilelor z_i sunt necunoscute, nu este posibil să obținem estimările în sens MLE

⁸⁶²Vedeți proprietatea iv de la problema 36.c de la capitolul de *Fundamente*.

⁸⁶³În penultima linie a egalității de mai sus, prin simbolul \exp am desemnat ca de obicei funcția exponentială cu baza e . (Așadar, $\exp(y) \stackrel{\text{not.}}{=} e^y$ pentru orice $y \in \mathbb{R}$.) Simbolul \top din expresia din dreapta egalității (356) reprezintă operația de transpunere de matrice / vectori.

⁸⁶⁴Este o situație similară cu — și doar ușor mai elaborată decât — cea din problema 53 (punctul a) de la capitolul de *Fundamente*.

⁸⁶⁵Vedeți similaritatea dintre formulele (358) și (56), respectiv (359) și (60).

ale parametrilor π , μ și Σ prin calcularea analitică a rădăcinilor derivatelor parțiale de ordinul întâi ale funcției de log-verosimilitate $l(\pi, \mu, \Sigma)$.

În astfel de cazuri, *algoritmul EM* constituie o metodă convenabilă pentru estimarea parametrilor în sensul verosimilității maxime. Strategia sa constă în a construi în mod iterativ o *limită inferioară* pentru funcția de *log-verosimilitate* a datelor observabile x_1, \dots, x_n (pasul E), pentru ca apoi să calculeze *maximul* acelei limite inferioare (pasul M).

Folosind schema algoritmică EM,⁸⁶⁶ demonstrați că *regulile de actualizare* pentru parametrii π, μ și Σ în cazul unei mixturi de distribuții gaussiene multidimensionale sunt următoarele:⁸⁶⁷

Pasul E:

$$w_{ij} \stackrel{\text{not.}}{=} p(z_i = j | x_i; \pi', \mu', \Sigma') = \frac{p(x_i | z_i = j; \mu', \Sigma') p(z_i = j; \pi')}{\sum_{l=1}^K p(x_i | z_i = l; \mu', \Sigma') p(z_i = l; \pi')}$$

Pasul M:

$$\pi_j = \frac{1}{n} \sum_{i=1}^n w_{ij}, \quad (360)$$

$$\mu_j = \frac{\sum_{i=1}^n w_{ij} x_i}{\sum_{i=1}^n w_{ij}}, \quad (361)$$

$$\Sigma_j = \frac{\sum_{i=1}^n w_{ij} (x_i - \mu_j)(x_i - \mu_j)^\top}{\sum_{i=1}^n w_{ij}}. \quad (362)$$

În aceste formule, π' , μ' și Σ' reprezintă valorile parametrilor modelului nostru la iterația precedentă a algoritmului EM (respectiv, valorile atribuite inițial parametrilor, pentru prima iterație).⁸⁶⁸

Sugestie: Întrucât se lucrează cu instanțe din \mathbb{R}^d , vă recomandăm să folosiți deriveate [parțiale] vectoriale. Pentru acest tip de deriveate, variabila în raport cu care se derivează este din \mathbb{R}^d , nu din \mathbb{R} cum suntem obișnuiți. Unele din următoarele formule (preluate din documentul *Matrix Identities*, de Sam Roweis, 1999) vă pot fi de folos:

$$(1e) \quad (A^{-1})^\top = (A^\top)^{-1}$$

$$(2b) \quad |A^{-1}| = \frac{1}{|A|}$$

⁸⁶⁶Vedeți cartea *Machine Learning*, a lui Tom Mitchell, 1997, pag. 194-195 sau problema 1 (pag. 959) de la capitolul *Schema algoritmică EM* din această culegere.

⁸⁶⁷Este util de observat (și de analizat) corespondența dintre formulele (357) și (360), apoi (358) și (361) și, în sfârșit, (359) și (372).

⁸⁶⁸Pentru ultimele două reguli de actualizare de la pasul M trebuie să fie satisfăcute condițiile $\sum_{i=1}^n w_{ij} \neq 0$, pentru $j \in \{1, \dots, K\}$. Înținând cont de definiția dată mai sus pentru ponderile w_{ij} , cu $i \in \{1, \dots, n\}$ și $j \in \{1, \dots, K\}$, condiția $\sum_{i=1}^n w_{ij} \neq 0$ (pentru j fixat) revine la următoarea proprietate: $\exists i \in \{1, \dots, n\}$ astfel încât $p(z_i = j | x_i; \pi, \mu, \Sigma) \neq 0$, adică

$$\underbrace{p(x_i | z_i = j; \mu, \Sigma)}_{>0} \underbrace{p(z_i = j | \pi)}_{\pi_j} \neq 0.$$

Restricția aceasta este satisfăcută (în contextul mixturilor de distribuții gaussiene) pentru acei j pentru care $\pi_j > 0$.

Este natural să presupunem $\pi_j > 0$ pentru toți $j \in \{1, \dots, K\}$. Cazurile $\pi_j = 0$ nu sunt interesante (sunt „degenerate”), deci pot fi eliminate din start. Așadar, la pasul de inițializare a algoritmului EM, vom lua toți $\pi'_j > 0$. Vom arăta la finalul demonstrației că această premiză va implica faptul că restricția $\pi_j > 0$ este satisfăcută la fiecare iterație.

$$(4a) \frac{\partial}{\partial X} |AXB| = |AXB|(X^{-1})^\top = |AXB|(X^\top)^{-1}$$

$$(4b) \frac{\partial}{\partial X} \ln |X| = (X^{-1})^\top = (X^\top)^{-1}$$

$$(5a) \frac{\partial}{\partial X} a^\top X = \frac{\partial}{\partial X} X^\top a = a$$

$$(5b) \frac{\partial}{\partial X} X^\top AX = (A + A^\top)X$$

$$(5c) \frac{\partial}{\partial X} a^\top X b = ab^\top$$

$$(5e) \frac{\partial}{\partial X} a^\top X a = \frac{\partial}{\partial X} a^\top X^\top a = aa^\top$$

$$(5g) \frac{\partial}{\partial X} (Xa + b)^\top C(Xa + b) = (C + C^\top)(Xa + b)a^\top$$

Răspuns:

Pasul E este ușor. Vom calcula

$$w_{ij} \stackrel{not.}{=} p(z_i = j | x_i; \pi', \mu', \Sigma').$$

Folosind regula lui Bayes (combinată cu formula probabilității totale), obținem:

$$p(z_i = j | x_i; \pi', \mu', \Sigma') = \frac{p(x_i | z_i = j; \mu', \Sigma') p(z_i = j; \pi')}{\sum_{l=1}^K p(x_i | z_i = l; \mu', \Sigma') p(z_i = l; \pi')}$$

Aici, probabilitatea $p(x_i | z_i = j; \mu', \Sigma')$ este obținută prin evaluarea densității gaussienei având media μ'_j și matricea de covarianță Σ'_j pentru argumentul x_i , în vreme ce probabilitatea $p(z_i = j; \pi')$ este dată de π'_j , s.a.m.d.

Apoi, la *pasul M*, trebuie să maximizăm, în raport cu parametrii π , μ și Σ , funcția

$$\sum_{i=1}^n \sum_j w_{ij} \ln \frac{p(x_i, z_i; \pi, \mu, \Sigma)}{w_{ij}}. \quad (363)$$

Conform schemei algoritmice EM,⁸⁶⁹ funcția (363), ale cărei argumente sunt π , μ și Σ , constituie o margine inferioară pentru funcția de log-verosimilitate a datelor observabile, $\sum_{i=1}^n \ln p(x_i | \pi, \mu, \Sigma)$.⁸⁷⁰ Vom rescrie în mod convenabil expresia acestei funcții, ținând cont mai întâi de definiția probabilității conditionate, apoi de definiția distribuției gaussiene multidimensionale și, în sfârșit, de proprietățile logaritmului:

⁸⁶⁹Vedeți problema 1, pag. 959 de la capitolul *Schema algoritmică EM* din această culegere.

⁸⁷⁰De fapt, folosind notațiile de la problema 1 de la capitolul *Schema algoritmică EM*, pentru orice distribuție de probabilitate q peste variabilele neobservabile z , urmează că funcția $F(q(z), \theta')$ — unde, F este definit prin relația (382), iar $\theta' \stackrel{not.}{=} (\pi', \mu', \Sigma')$ — este o margine inferioară pentru $\ln P(x | \theta')$, log-verosimilitatea variabilelor observabile x .

Însă dintre toate aceste distribuții, cea mai bună — adică, cea pentru care se atinge $\max_{q(z)} F(q(z), \theta')$, este distribuția $q_{\pi', \mu', \Sigma'}$ definită astfel: $q(z_i = j) \stackrel{not.}{=} q_{\pi', \mu', \Sigma'}(z_i = j) \stackrel{not.}{=} p(z_i = j | x_i; \pi', \mu', \Sigma') \stackrel{not.}{=} w_{ij}$. (Vedeți demonstrația de la problema 1.c de la capitolul *Schema algoritmică EM*.)

$$\begin{aligned}
\sum_{i=1}^n \sum_{j=1}^K w_{ij} \ln \frac{p(x_i, z_i; \pi, \mu, \Sigma)}{w_{ij}} &= \sum_{i=1}^n \sum_{j=1}^K w_{ij} \ln \frac{p(x_i | z_i = j; \mu, \Sigma) p(z_i = j; \pi)}{w_{ij}} = \\
&\sum_{i=1}^n \sum_{j=1}^K w_{ij} \ln \frac{\frac{1}{(2\pi)^{d/2} |\Sigma_j|^{1/2}} \cdot \exp\left(-\frac{1}{2}(x_i - \mu_j)^\top \Sigma_j^{-1} (x_i - \mu_j)\right) \cdot \pi_j}{w_{ij}} = \\
&\sum_{i=1}^n \sum_{j=1}^K w_{ij} \left[-\ln((2\pi)^{d/2} |\Sigma_j|^{1/2} w_{ij}) - \frac{1}{2}(x_i - \mu_j)^\top \Sigma_j^{-1} (x_i - \mu_j) + \ln \pi_j \right] \quad (364)
\end{aligned}$$

Ca o *consecință* a calculelor de mai sus, obiectivul nostru a devenit maximizarea expresiei (364). Pentru aceasta, vom proceda în maniera clasică: vom căuta soluțiile derivatelor parțiale de ordinul întâi ale acestei expresii în raport cu parametrii μ , Σ și π .

Mai întâi vom calcula derivata parțială în raport unde μ_l , cu $l \in \{1, \dots, K\}$.⁸⁷¹

$$\begin{aligned}
&\frac{\partial}{\partial \mu_l} \sum_{i=1}^n \sum_{j=1}^K w_{ij} \left[-\ln((2\pi)^{d/2} |\Sigma_j|^{1/2} w_{ij}) - \frac{1}{2}(x_i - \mu_j)^\top \Sigma_j^{-1} (x_i - \mu_j) + \ln \pi_j \right] \\
&= -\frac{\partial}{\partial \mu_l} \sum_{i=1}^n \sum_{j=1}^K w_{ij} \frac{1}{2}(x_i - \mu_j)^\top \Sigma_j^{-1} (x_i - \mu_j) \\
&= -\frac{1}{2} \sum_{i=1}^n w_{il} \frac{\partial}{\partial \mu_l} (x_i - \mu_l)^\top \Sigma_l^{-1} (x_i - \mu_l) \quad (365)
\end{aligned}$$

$$= \frac{1}{2} \sum_{i=1}^n w_{il} \frac{\partial}{\partial \mu_l} (2\mu_l^\top \Sigma_l^{-1} x_i - \mu_l^\top \Sigma_l^{-1} \mu_l) \quad (366)$$

$$= \sum_{i=1}^n w_{il} (\Sigma_l^{-1} x_i - \Sigma_l^{-1} \mu_l). \quad (367)$$

Detalii de calcul:

O primă posibilitate este ca, pornind de la expresia (365) să se treacă la expresia (366) — folosind pe de o parte distributivitatea operației de înmulțire a matricelor față de operația de adunare sau de scădere și, pe de altă parte, formula (1e) —, iar apoi de la expresia (366) să se treacă la expresia (367), ținând cont de formulele (5c) — sau, mai simplu, (5a) — și (5b) din documentul *Matrix Identities* de Sam Rowens.

A doua posibilitate este ca, pornind de la expresia (365) să obținem direct expresia (367), folosind formulele (5g) și (1e) din documentul menționat. Formula (1e) se aplică matricei Σ_l^{-1} . Se ține cont că Σ_l este matrice simetrică și inversabilă.⁸⁷²

O a treia posibilitate este să se lucreze în mod clasic (nu derivând în raport cu un vector), și anume scriind derivele parțiale ale expresiei (364) în raport cu fiecare dintre componentele vectorului μ_l (și scriind de asemenea pe componente vectorul x_i și respectiv matricea Σ_l^{-1}).⁸⁷³

⁸⁷¹Atenție: μ_l este vector (coloană) și, în consecință, derivata parțială în raport cu μ_l va fi de asemenea un vector (coloană). Similar, derivata parțială în raport cu Σ_l va fi o matrice.

⁸⁷²Vedeți problema 20 de la capitolul de *Fundamente*.

⁸⁷³În același mod se pot demonstra și formulele (5g), (1e), (5c) și (5b) menționate mai sus (precum și cele care vor fi utilizate mai jos).

Egalând expresia (367) cu vectorul-colonă 0 (format din K elemente, toate având valoarea 0 $\in \mathbb{R}$), vom obține regula de actualizare pentru μ_l :

$$\mu_l = \frac{\sum_{i=1}^n w_{il} x_i}{\sum_{i=1}^n w_{il}}. \quad (368)$$

Detalii de calcul:

$$\sum_{i=1}^n w_{il} (\Sigma_l^{-1} x_i - \Sigma_l^{-1} \mu_l) = 0 \Leftrightarrow \sum_{i=1}^n w_{il} \Sigma_l^{-1} x_i = \sum_{i=1}^n w_{il} \Sigma_l^{-1} \mu_l$$

Înmulțind ultima egalitate la stânga cu matricea Σ_l , obținem $\sum_{i=1}^n w_{il} x_i = \sum_{i=1}^n w_{il} \mu_l$. Este imediat că membrul drept al acestei ultime egalități se poate scrie ca $(\sum_{i=1}^n w_{il}) \mu_l$. (Observați că $\sum_{i=1}^n w_{il}$ este un număr real.) Prin urmare, $\mu_l = \frac{\sum_{i=1}^n w_{il} x_i}{\sum_{i=1}^n w_{il}}$.

Să deducem acum relațiile de actualizare de la pasul pasul M pentru matricele Σ_j , unde $j = 1, \dots, K$. Reținând în expresia (364) doar termenii care depind de Σ_j , rezultă că trebuie să maximizăm (în raport cu Σ_j) expresia

$$\begin{aligned} & \sum_{i=1}^n \sum_{j=1}^K w_{ij} \left[\ln \frac{1}{|\Sigma_j|^{1/2}} - \frac{1}{2} (x_i - \mu_j)^\top \Sigma_j^{-1} (x_i - \mu_j) \right] \\ &= \sum_{i=1}^n \sum_{j=1}^K w_{ij} \left[\frac{1}{2} \ln |\Sigma_j^{-1}| - \frac{1}{2} (x_i - \mu_j)^\top \Sigma_j^{-1} (x_i - \mu_j) \right] \\ &= \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^K w_{ij} [\ln |\Sigma_j^{-1}| - (x_i - \mu_j)^\top \Sigma_j^{-1} (x_i - \mu_j)]. \end{aligned} \quad (369)$$

Detalii de calcul:

Pentru a obține prima egalitate de mai sus, se folosește formula (2b). Concret, $\frac{1}{|\Sigma_j|^{1/2}} = |\Sigma_j|^{-1/2} = (|\Sigma|^{-1})^{1/2} \stackrel{(2b)}{=} |\Sigma^{-1}|^{1/2}$.

În astfel de situații, pentru convenientă, se folosește următorul „artificiu de calcul“: în loc să se lucreze cu matricea de covarianță Σ_j (presupusă a fi inversabilă), se preferă să se lucreze cu matricea de precizie $\Lambda_j \stackrel{\text{not.}}{=} \Sigma_j^{-1}$.⁸⁷⁴ Așadar, vom maximiza dublul expresiei (369) în raport cu Λ_j . Derivând (în raport cu Λ_j), vom obține:

$$\frac{\partial}{\partial \Lambda_j} \sum_{i=1}^n w_{ij} [\ln |\Lambda_j| - (x_i - \mu_j)^\top \Lambda_j (x_i - \mu_j)] \quad (370)$$

$$= \sum_{i=1}^n w_{ij} \frac{1}{|\Lambda_j|} \frac{\partial}{\partial \Lambda_j} |\Lambda_j| - \sum_{i=1}^n w_{ij} \frac{\partial}{\partial \Lambda_j} [(x_i - \mu_j)^\top \Lambda_j (x_i - \mu_j)] \quad (371)$$

$$= \sum_{i=1}^n w_{ij} \Lambda_j^{-1} - \sum_{i=1}^n w_{ij} (x_i - \mu_j) (x_i - \mu_j)^\top \quad (372)$$

$$= \left(\sum_{i=1}^n w_{ij} \right) \Lambda_j^{-1} - \sum_{i=1}^n w_{ij} (x_i - \mu_j) (x_i - \mu_j)^\top. \quad (373)$$

⁸⁷⁴Aşa s-a procedat și la problema 53 de la capitolul de *Fundamente*, unde am arătat cum anume se face estimarea parametrilor distribuției gaussiene multidimensionale.

Detalii de calcul:

Expresia (371) se poate obține din expresia (370) folosind pentru primul termen formula de derivare a funcțiilor compuse, iar expresia (372) se poate obține din expresia (371) folosind pentru primul termen formula (4a) (luând $A = B = I$, matricea identitate), iar pentru termenul al doilea formula (5e). Alternativ, expresia (372) se poate obține direct din expresia (370),⁸⁷⁵ folosind pentru derivarea primului termen formula (4b) și pentru derivarea celui de-al doilea termen formula (5e). În plus, la fiecare dintre aceste două variante de rezolvare, se ține cont și de formula (1e).

Egalând expresia (373) cu 0 (matrice pătratică de dimensiune $K \times K$, ale cărei elemente au, toate, valoarea $0 \in \mathbb{R}$), vom obține:

$$\Sigma_j = \Lambda_j^{-1} = \frac{\sum_{i=1}^n w_{ij} (x_i - \mu_j)(x_i - \mu_j)^\top}{\sum_{i=1}^n w_{ij}}. \quad (374)$$

În sfârșit, vom deriva regula de actualizare de la pasul M pentru parametrii π_j , unde $j = 1, \dots, K$. Reținând în expresia (364) doar acei termeni care depind de π_j , rezultă că va trebui să maximizăm

$$\sum_{i=1}^n \sum_{j=1}^K w_{ij} \ln \pi_j. \quad (375)$$

Să observăm însă — exact ca la rezolvarea problemei 18, cazul $K > 2$ — că soluțiile optime π_j trebuie să satisfacă o restricție suplimentară, și anume $\sum_{j=1}^K \pi_j = 1$, din moment ce $\pi_j \stackrel{\text{not.}}{=} p(z_i = j | \pi)$, pentru orice $j \in \{1, \dots, K\}$ și orice $i \in \{1, \dots, n\}$. Vom proceda la fel și aici, folosind *metoda multimplicatorilor lui Lagrange* pentru a identifica acele valori ale variabilelor π_j care maximizează expresia (375) satisfăcând în același timp restricția $\sum_{j=1}^K \pi_j = 1$. Așadar, vom căuta punctul de optim al polinomului lagrangean

$$\mathcal{L}(\pi, \lambda) = \sum_{i=1}^n \sum_{j=1}^K w_{ij} \ln \pi_j + \lambda \left(\sum_{j=1}^K \pi_j - 1 \right),$$

unde $\lambda \in \mathbb{R}$ este aşa-numita variabilă („multiplicator“) Lagrange.⁸⁷⁶ Derivatele parțiale ale polinomului $\mathcal{L}(\pi, \lambda)$ în raport cu π_j , pentru $j = 1, \dots, K$ sunt:

$$\frac{\partial}{\partial \pi_j} \mathcal{L}(\pi, \lambda) = \sum_{i=1}^n \frac{w_{ij}}{\pi_j} + \lambda = \frac{1}{\pi_j} \left(\sum_{i=1}^n w_{ij} \right) + \lambda.$$

Egalând cu 0, rezultă:

$$\pi_j = -\frac{\sum_{i=1}^n w_{ij}}{\lambda}.$$

Impunând restricția $\sum_j \pi_j = 1$, rezultă $-\lambda = \sum_{j=1}^K \sum_{i=1}^n w_{ij} = \sum_{i=1}^n \sum_{j=1}^K w_{ij} = \sum_{i=1}^n 1 = n$. (Aici am folosit faptul că $w_{ij} \stackrel{\text{not.}}{=} p(z_i = j | x_i; \mu', \Sigma')$ și, din moment ce

⁸⁷⁵Remarcați că $\Lambda_j^\top = \Lambda_j$. Într-adevăr, întrucât $\Sigma_j = \Sigma_j^\top$, urmează că $\Lambda_j = \Sigma_j^{-1} = (\Sigma_j^\top)^{-1} \stackrel{(1e)}{=} (\Sigma_j^{-1})^\top = \Lambda_j^\top$.

⁸⁷⁶Ca și la problema 18, vom observa că nu este nevoie să ne mai ocupăm și de restricțiile $\pi_j \geq 0$ — care ar trebui adăugate în mod normal la formularea problemei de optimizare —, pentru că, aşa cum vom vedea în curând, soluția pe care o vom găsi pur și simplu derivând polinomul $\mathcal{L}(\pi)$ va satisface (în mod implicit) aceste restricții.

pentru fiecare $i = 1, \dots, n$ (fixat) aceasta reprezintă o funcție de probabilitate, rezultă că $\sum_j w_{ij} = 1$.) Așadar, relația de actualizare de la pasul M pentru parametrul π_j va fi:

$$\pi_j = \frac{1}{n} \sum_{i=1}^n w_{ij}. \quad (376)$$

Desigur, $\pi_j \geq 0$, întrucât $w_{ij} \geq 0$ pentru orice $j \in \{1, \dots, K\}$ și orice $i \in \{1, \dots, n\}$.⁸⁷⁷

Observație (2): Acum cititorul ar trebui să compare relațiile de actualizare de la pasul M⁸⁷⁸ cu formulele de estimare directă (MLE) atunci când variabilele z_i sunt cunoscute / observabile⁸⁷⁹: ele sunt identice, cu excepția faptului că în locul funcțiilor-indicator $1_{\{z_i=j\}}$ acum apar ponderile / mediile w_{ij} . Așadar, formulele de la pasul M al acestui algoritm EM (deci, în cazul învățării nesupervizate) sunt versiuni *ponderate* ale formulelor MLE (cazul supervizat). De asemenea, forma acestui algoritm EM ne amintește și de algoritmul de clusterizare K -means: totul este similar, cu excepția faptului că în loc de asignare “hard” a instanțelor x_i la clustere $c(i)$, avem acum asignări “soft” (w_{ij}). Ca și K -means, algoritmul EM este sensibil la problema optimului local, deci rularea repetată, cu diferite valori inițiale pentru unii parametri (sau, pentru toți parametrii) poate fi folositoare. Este limpede că putem asocia algoritmului EM o *interpretare* foarte naturală, și anume aceea a încercării repetate de a ghici necunoscutele z_i .

25.

(Algoritmii K -means și EM pentru GMM în \mathbb{R}^2 ; o întrebare de ordin calitativ)

CMU, 2010 spring, E. Xing, T. Mitchel, A. Singh, midterm, pr. 5.1

Considerăm setul de date de antrenament de mai jos și doi algoritmi de clusterizare: K -means și EM pentru modelare de mixturi de gaussiene (EM/GMM). Credeți că acești algoritmi vor produce în final aceeași centroizi pentru clustere dacă pornesc de la aceleași valori inițiale μ_1 și μ_2 ? Explicați pe scurt.



Răspuns:

Nu. Diferența esențială dintre cei doi algoritmi constă în faptul că algoritmul K -means folosește asignări de tip “hard”, adică fiecare punct aparține unui singur cluster, în timp ce algoritmul EM/GMM utilizează asignări de tip “soft”, ceea ce înseamnă că fiecare punct este asignat cu o anumită probabilitate fiecărui cluster.

Așadar, în cazul K -means, la fiecare nouă iterație poziția fiecărui centroid este determinată de media punctelor asignate clusterului corespunzător, pe când în cazul algoritmului EM/GMM poziția fiecărui centroid este o anumită medie ponderată obținută din coordonatele tuturor punctelor. Ponderile reprezintă

⁸⁷⁷Conform notei de subsol 868, dacă valorile atribuite inițial parametrilor π_j satisfac restricția $\pi_j > 0$ pentru orice $j \in \{1, \dots, K\}$, rezultă că $w_{ij} > 0$ pentru orice $i \in \{1, \dots, n\}$, deci și noile valori obținute pentru π_j la pasul M (pentru $j = 1, \dots, K$) vor fi tot strict pozitive.

⁸⁷⁸Și anume (376), (368) și (374). Acestea coincid cu formulele (360), (361) și (362) din ultima parte a enunțului.

⁸⁷⁹Vedeți *Observația* (1) din enunț și / sau problema 53.a de la capitolul de *Fundamente*.

probabilitatea de generare a punctului de către gaussiana asociată respectivului centroid. Prin urmare, centroizii finali obținuți de EM/GMM vor fi (cel puțin sensibil) deplasăți față de centroizii determinați de K -means.

26.

(Adevărat sau Fals?)

a.

Liviu Ciortuz

Aplicând algoritmul de clusterizare EM pentru o mixtură de K distribuții gaussiene obținem același rezultat ca și în cazul folosirii algoritmului K -means.

b.

CMU, 2006 (10-701/15-781), final exam, pr. 1.e

În afară de algoritmul EM, pentru a învăța parametrii unei mixturi de distribuții gaussiene se mai poate folosi și metoda gradientului.

Răspuns:

a. Fals. Mai precis: posibil da, posibil nu; depinde de valorile inițiale atribuite varianțelor distribuțiilor gaussiene (chiar dacă valorile inițiale ale centroizilor / mediilor se corespund). Vedeți problema 22.

b. Adevărat. Algoritmul EM este însă preferat, fiindcă uneori derivele parțiale ale funcției de verosimilitate a datelor observabile în raport cu parametrii de estimat sunt dificil sau chiar imposibil de calculat.

7.2 Clusterizare — Probleme propuse

7.2.1 Clusterizare ierarhică

27. (Clusterizare ierarhică aglomerativă, folosind similaritate de tip “single-”, “complete-” și “average-linkage”)
*prelucrare de Liviu Ciortuz, după
 ■ • ○ CMU, 2012 fall, T. Mitchell, Z. Bar-Joseph, HW4, pr. 2.a*

Tabelul de mai jos reprezintă matricea de distanțe pentru [o mulțime formată din] șase obiecte.

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>
<i>A</i>	0					
<i>B</i>	0.12	0				
<i>C</i>	0.51	0.25	0			
<i>D</i>	0.84	0.16	0.14	0		
<i>E</i>	0.28	0.77	0.70	0.45	0	
<i>F</i>	0.34	0.61	0.93	0.20	0.67	0

- a. Aplicați algoritmul de clusterizare ierarhică aglomerativă pe aceste date, folosind mai întâi similaritate *single-linkage* și apoi similaritate *complete-linkage*. La fiecare pas al algoritmului, rescrieți în mod corespunzător matricea de distanțe. (De la o iterație la alta, se micșorează cu 1 numărul liniilor precum și al coloanelor folosite.) La final, desenați dendrogramele rezultate.

Indicație: Înălțimea corespunzătoare fiecărui cluster non-singleton (adică, a fiecărui nod intern) din dendrogramă va fi considerată ca fiind egală cu distanța (i.e., conform măsurii de similaritate) dintre cele două sub-clustere constitutive.

- b. Dacă ați lucrat corect, atunci cele două dendrograme obținute la punctul *a* nu coincid [nici măcar] ca structură. Modificați două valori din matricea de distanțe dată mai sus, în aşa fel încât de data aceasta cele două dendrograme care se obțin să fie identice ca structură.

- c. Procedați similar cu cerințele de la punctul *a*, dar de această dată pentru *average-linkage*. La actualizarea matricei de distanțe (sau, de „proximitate“) veți ține cont de formula:

$$\begin{aligned} \Delta(X \cup Y, Z) &\stackrel{\text{def.}}{=} \frac{1}{(|X| + |Y|)|Z|} \sum_{x \in X \cup Y} \sum_{z \in Z} d(x, z) \\ &\stackrel{\text{calcul}}{=} \frac{1}{|X| + |Y|} (|X| \Delta(X, Z) + |Y| \Delta(Y, Z)). \end{aligned} \quad (377)$$

unde X, Y și Z sunt clustere disjuncte două câte două, iar notația $|X|$ desemnează cardinalul lui X (adică, numărul de elemente din X).

- d. Demonstrați formula enunțată la punctul *c*.

28.

(Clusterizare ierarhică aglomerativă, folosind similaritate de tip “single-linkage”, “complete-linkage” și “average-linkage”; aplicare pe date din \mathbb{R})

* CMU, 2009 spring, Ziv Bar-Joseph, final exam., pr. 9.2

Dorim să clusterizăm numerele de la 1 la 1024, folosind algoritmul de clusterizare ierarhică aglomerativă. Dacă la o iterație oarecare distanțele dintre două perechi de clustere au aceeași valoare, ne vom folosi de ordonarea numerică.⁸⁸⁰

Să se compare rezultatele obținute cu trei variante ale funcției de similaritate discutate la curs — și anume, “single-linkage”, “complete-linkage” și “average-linkage” —, specificându-se pentru fiecare dintre aceste variante care este numărul de elemente asignate fiecăruiă dintre cele două clustere [care compun clusterul corespunzător nodului rădăcină] de la vârful dendrogramei rezultate.

29.

(Clusterizare ierarhică aglomerativă; aplicare pe date din \mathbb{R})

CMU, 2012 spring, Ziv Bar-Joseph, midterm, pr. 9.1

Ne propunem să clusterizăm în manieră ierarhică aglomerativă instanțele $2^0, 2^1, 2^2, \dots, 2^n$ (deci, în total, $n+1$ puncte), folosind distanța euclidiană. Trasăți dendrogramale (adică arborii de clusterizare ierarhică) care se obțin pentru fiecare dintre următoarele trei tipuri de funcții de similaritate: single-, complete- și average-linkage. Vom considera că fiecare nod neterminal din dendrogramă — adică rădăcina fiecărui cluster non-singleton —, are înălțimea egală cu valoarea măsurii de similaritate (adică, distanță) dintre cele două clustere din care a fost obținut.

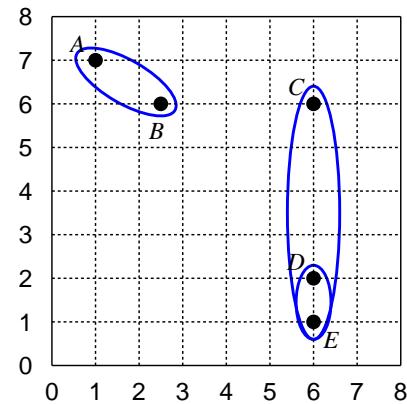
⁸⁸⁰Vedeți de exemplu cum s-a procedat la problema 1.

30. (Clusterizare ierarhică aglomerativă: un exemplu simplu de aplicare, cu single-, complete- și average-linkage, pe date din \mathbb{R}^2)

*prelucrare de Liviu Ciortuz, 2022, după
• CMU, 2021 fall, Aarti Singh, Recitation 9, pr. 3*

Considerăm punctele $A(1, 7)$, $B(2.5, 6)$, $C(6, 6)$, $D(6, 2)$ și $E(6, 1)$ din planul euclidian. Pe acest dataset veți aplica algoritmul de clusterizare ierarhică aglomerativă (i.e., bottom-up), folosind pe rând (separat) funcțiile de similaritate single-linkage, complete-linkage și average-linkage.

Care dintre aceste funcții de similaritate va conduce după executarea a trei iterări consecutive la ierarhia aplatizată prezentată în figura alăturată? (*Observație:* Pentru simplitate, nu am mai desenat și elipsa corespunzătoare clusterului care include toate punctele de clusterizat.)



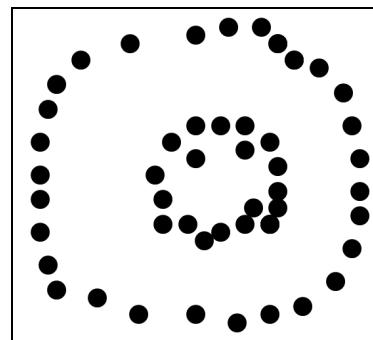
31. (Clusterizare ierarhică aglomerativă: aplicare în manieră intuitivă pe date din \mathbb{R}^2)

** CMU, 2010 fall, Ziv Bar-Joseph, midterm, pr. 8.b*

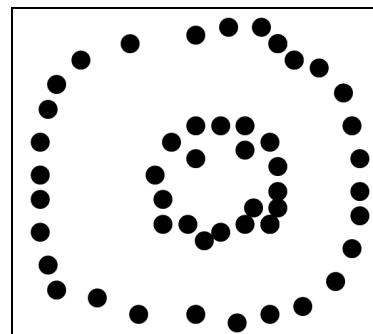
- a. Folosind distanța euclidiană, aplicați algoritmul de clusterizare ierarhică aglomerativă cu similaritate de tip “single-linkage” pe datele din figura alăturată.

Indicați pe desen care sunt instanțele care formează cele două clustere de la vârful dendrogramei.

Observație: Nu este necesar să construiți efectiv dendrograma.



- b. Care este rezultatul dacă se folosește similaritate de tip “average-linkage”?

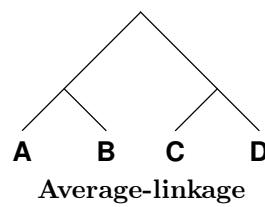
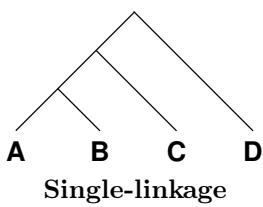


32.

(Clusterizare ierarhică aglomerativă:
raționamente calitative)

• o CMU, 2014 fall, W. Cohen, Z. Bar-Joseph, midterm, pr. 9.AB

A. În figurile de mai jos sunt reprezentate rezultatele clusterizării a patru puncte (A , B , C și D) folosind o aceeași matrice de distanțe T și două tipuri diferite de măsuri de similaritate, după cum este indicat sub fiecare dintre cei doi arbori de clusterizare considerați.



În cele ce urmează, vom nota spre exemplu cu $T(A, B)$ distanța dintre A și B dată în matricea de intrare, și cu

$$T'(\{A, B\}, \{C, D\}) = \frac{T(A, C) + T(A, D) + T(B, C) + T(B, D)}{4}$$

distanța corespunzătoare măsurii de similaritate “average-linkage” (adică distanță medie) dintre clusterele $\{A, B\}$ și respectiv $\{C, D\}$. Ca de obicei, notăm cu $\min(X, Y)$ cea mai mică dintre valorile X și Y .

Facem *presupunerea* că toate distanțele din matricea de intrare sunt diferite și că în niciuna dintre inegalitățile de mai jos termenii care se compară nu sunt egali (așadar, răspunsul este fie $>$ fie $<$).

La fiecare din cele trei puncte de mai jos, precizați dacă inegalitatea indicată are loc, nu are loc, sau este imposibil de precizat. Justificați în mod riguros alegerea făcută.

a. $\min(T(A, C), T(B, C)) > T(C, D)$.

adevărat fals imposibil de precizat

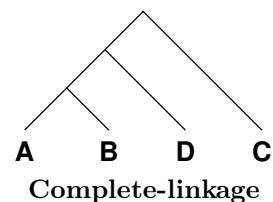
b. $T'(\{A, B\}, \{C\}) > T(C, D)$.

adevărat fals imposibil de precizat

c. $T(A, D) > T(A, C)$.

adevărat fals imposibil de precizat

B. Considerăm arborele de clusterizare din figura alăturată. Este oare posibil ca acest arbore să fi fost obținut pornind de la matricea de distanțe T de la punctul A de mai sus folosind similaritate de tip “complete-linkage” (așa cum se specifică în figură)?



33.

(Clusterizare ierarhică [aglomerativă]:
o altă funcție / măsură de similaritate între clustere:
metrică / distanță lui Ward)

*prelucrare făcută de Liviu Ciortuz, după
■ • ○ CMU, 2010 fall, Aarti Singh, HW3, pr. 4.1*

În acest exercițiu veți analiza o modalitate alternativă pentru a defini distanța dintre două clustere disjuncte, care a fost propusă de către Joe H. Ward în 1963.⁸⁸¹ O vom numi *metrică lui Ward*.

Această metrică definește distanța dintre două clustere disjuncte X și Y (ambele finite și incluse în \mathbb{R}^d cu $d \in \mathbb{N}^*$) ca fiind dată de creșterea sumei pătratelor distanțelor de la fiecare instanță x_i la *centroidul* clusterului la care ea este asignată, atunci când reunim cele două clustere ca să formăm un cluster nou. Din punct de vedere formal, vom scrie:

$$\Delta(X, Y) = \sum_{x_i \in X \cup Y} \|x_i - \mu_{X \cup Y}\|^2 - \sum_{x_i \in X} \|x_i - \mu_X\|^2 - \sum_{x_i \in Y} \|x_i - \mu_Y\|^2 \quad (378)$$

unde, spre exemplu, μ_X este centroidul („centrul de greutate“) clusterului X , iar x_i este o instanță generică dintr-un cluster [oarecare, fixat]. Prin definiție, aici vom considera $\mu_X = \frac{1}{n_X} \sum_{x_i \in X} x_i$, unde n_X este numărul de elemente din X . Mărimea $\Delta(X, Y)$ din formula (378) poate fi interpretată ca reprezentând *costul pentru combinarea* celor două clustere X și Y într-un singur cluster.⁸⁸²

- a. Aduceți expresia din partea dreaptă a formulei (378) la o formă mai simplă. (Redactați toți pașii intermediari.) *Sugestie:* Formula obținută trebuie să fie doar în funcție de numărul de elemente din cele două clustere (n_X și respectiv n_Y) și de distanță $\|\mu_X - \mu_Y\|^2$ dintre centroizii acestor clustere, μ_X și respectiv μ_Y .
- b. Comentați metrica lui Ward. Ce credeți că urmărește ea să realizeze? *Sugestie:* Varianta simplificată a formulei de mai sus vă va ajuta să răspundeti la întrebarea care a fost pusă aici.
- c. Presupuneți că vi se dau două *perechi* de clustere, P_1 și P_2 , iar centroizii celor două clustere din P_1 sunt situați la o distanță mai mare decât distanța dintre centroizii clusterelor din perechea P_2 . Oare, folosind metrică lui Ward, algoritmul de clusterizare aglomerativă va alege *întotdeauna* să „combine“ mai întâi cele două clustere din P_2 (fiindcă sunt situate la o ‘distanță’ mai mică)? De ce (nu)? Justificați răspunsul dumneavoastră cu ajutorul unui exemplu simplu.
- d. La clusterizare, de obicei nu este ușor să decizi care este numărul potrivit de clustere în care trebuie grupate datele. Folosind metrică lui Ward în conjuncție cu clusterizarea aglomerativă, puteți concepe o euristică simplă care să vă ajute să alegeți [bine] numărul de clustere, K ?

⁸⁸¹Ward, J. H., Jr., *Hierarchical Grouping to Optimize an Objective Function*, Journal of the American Statistical Association, 58 (1963), 236–244.

⁸⁸²Remarcați faptul că instanțele x_i fiind dintr-un spațiu \mathbb{R}^d , urmează că $\mu_X \in \mathbb{R}^d$ (ca și ceilalți centroizi), iar $\|x\|^2 = x \cdot x$ pentru orice $x \in \mathbb{R}^d$, operatorul · desemnând produsul scalar al vectorilor în \mathbb{R}^d .

34.

(Clusterizare ierarhică aglomerativă: aplicare pe date din \mathbb{R}^2 , folosind metrica lui Ward)

*Liviu Ciortuz, 2018, folosind datele de la
■ ○ * Edinburgh, 2009 fall, C. Williams, V. Lavrenko, HW4, pr. 3*

La problema 33 am prezentat o funcție de similaritate numită *metrica lui Ward*. Potrivit acestei metriki, distanța dintre două clustere disjuncte X și Y se definește astfel:

$$\Delta(X, Y) = \sum_{x_i \in X \cup Y} \|x_i - \mu_{X \cup Y}\|^2 - \sum_{x_i \in X} \|x_i - \mu_X\|^2 - \sum_{x_i \in Y} \|x_i - \mu_Y\|^2 \quad (379)$$

unde, spre exemplu, μ_X este centroidul [sau „centrul de greutate“ al] clusterului X , iar x_i este o instanță generică dintr-un cluster [oarecare, fixat].

Prin definiție, aici vom considera $\mu_X = \frac{1}{n_X} \sum_{x_i \in X} x_i$, unde n_X este numărul de elemente din X . (Similar sunt definiți centroizii μ_Y și $\mu_{X \cup Y}$.)

Se poate arăta (vedeți tot problema 33) că

$$\Delta(X, Y) = \frac{n_X n_Y}{n_X + n_Y} \|\mu_X - \mu_Y\|^2. \quad (380)$$

Observații:

1. Pentru perechi de clustere (X, Y) și (X', Y') astfel încât $n_X = n_{X'}$ și $n_Y = n_{Y'}$,⁸⁸³ formula (380) arată că la clusterizare ierarhică folosind merica lui Ward [la o iterare oarecare] este „favorizată“ acea pereche pentru care centroizii (μ_X și μ_Y , respectiv $\mu_{X'}$ și $\mu_{Y'}$) sunt mai apropiati.
2. Invers, dacă $\|\mu_X - \mu_Y\| = \|\mu_{X'} - \mu_{Y'}\|$, atunci este favorizată perechea pentru care ponderea⁸⁸⁴ (adică $\frac{n_X n_Y}{n_X + n_Y}$, respectiv $\frac{n_{X'} n_{Y'}}{n_{X'} + n_{Y'}}$) este mai mică.

Aceste două *observații* vă vor ajuta să simplificați / reduceți foarte mult calculele pe care ar trebui să le faceți pentru a rezolva următoarea cerință!

Aplicați algoritmul de clusterizare ierarhică aglomerativă pe setul de date de la problema 2,

$$\begin{aligned} A &: (-4, -2), B : (-3, -2), C : (-2, -2), D : (-1, -2), E : (+1, -1) \\ F &: (+1, +1), G : (+2, +3), H : (+3, +2), I : (+3, +4), J : (+4, +3) \end{aligned}$$

utilizând [de această dată] metrica lui Ward.⁸⁸⁵ Ca rezultat al clusterizării, veți reprezenta *dendrograma* sub formă *aplatizată*, folosind elipse (și indici) pentru a indica clusterele formate.

Coincide rezultatul obținut aici cu vreunul din rezultatele de la problema 2?

Precizare: Dacă la o iterare a algoritmului de clusterizare distanțele (adică similaritățile) dintre două perechi de clustere au aceeași valoare, veți considera că prioritatea la alcătuirea noului cluster este dictată de ordinea alfabetică.

⁸⁸³Sau, mai general, atunci când $\frac{n_X n_Y}{n_X + n_Y} = \frac{n_{X'} n_{Y'}}{n_{X'} + n_{Y'}}$.

⁸⁸⁴Această pondere este jumătate din *media armonică* a cardinalilor n_X și n_Y .

⁸⁸⁵Acest tip de clusterizare ierarhică este cunoscută în literatura de specialitate și sub numele de *clusterizare ierarhică bazată pe centroizi* (engl., centroid-based).

35.

(Clusterizare ierarhică aglomerativă: calcularea eficientă a matricei de „proximitate“)

 CMU, 2010 fall, Aarti Singh, HW3, pr. 4.2

La curs, atunci când am prezentat algoritmul de clusterizare ierarhică aglomerativă, am arătat că este necesar să actualizăm matricea de distanțe (numită și *matricea de proximitate*; engl., the closeness matrix) la fiecare iterare a algoritmului.⁸⁸⁶ Calcularea distanțelor dintre clusterul nou-obținut la iterată curentă (prin „contopirea“ clusterelor X și Y) și toate celelalte clustere se poate face, în mod *naiv*, folosind instanțele [individuale] din clustere.

a. Presupunem că dorim să obținem / realizăm o procedură mai eficientă de actualizare a matricei de proximitate, folosind *doar* elementele din forma curentă a matricei de proximitate, adică fără să folosim instanțele [individuale] din clustere. Pentru care dintre următoarele tipuri de funcții de similaritate putem să atingem acest obiectiv?

- i. single-linkage,
- ii. complete-linkage,
- iii. average-linkage,
- iv. metrika lui Ward.⁸⁸⁷

Prezentați raționamentul dumneavoastră, până la nivel de detaliu.

b. Dacă la vreunul dintre cazurile *i–iv* nu este posibil să atingem obiectivul pe care ni l-am propus la punctul *a* — repetăm, folosind *doar* conținutul curent al matricei de proximitate —, am putea totuși să reușim dacă (*în plus*) vom memora doar câteva informații adiționale pentru fiecare cluster? Pentru fiecare din cazurile identificate [la acest punct] veți specifica în mod clar ce informații adiționale trebuie memorate.

36.

(Clusterizare ierarhică: particularizare pentru cazul când similaritatea se exprimă cu ajutorul funcției cosinus între instanțe „normalizate“)

Liviu Ciortuz, 2016, pornind de la
 Foundations of Statistical Natural Language Processing,
 C. Manning, H. Schütze, MIT Press, 1999, pag. 507-509

Stim că produsul scalar al doi vectori $x = (x_1, \dots, x_d)$ și $y = (y_1, \dots, y_d)$ din \mathbb{R}^d se definește astfel: $x \cdot y = \sum_{i=1}^d x_i y_i$. Dacă atât x cât și y sunt nenuli, atunci definim $\cos(x, y) = \frac{x \cdot y}{\|x\| \|y\|}$.⁸⁸⁸ Dacă x și y sunt vectori unitari, adică $\|x\| = \|y\| = 1$, atunci definiția de mai sus devine: $\cos(x, y) = x \cdot y$. Atunci când x și y nu sunt unitari, îi putem „normaliza“, înmulțindu-i cu scalarii $1/\|x\|$ și respectiv $1/\|y\|$.

În continuare vom presupune că folosim doar vectori unitari / normalizați.

Funcția cos definită mai sus este folosită uneori în clusterizarea ierarhică, ca măsură de similaritate între instanțe, de exemplu la clusterizarea textelor. De

⁸⁸⁶Vedeți, ca exemplu, rezolvarea problemei 27.

⁸⁸⁷Vedeți problema 33.

⁸⁸⁸Se poate arăta că această definiție extinde în mod natural definiția funcției trigonometrice cosinus, aşa cum o stim de la geometria plană din școală. Cosinusul este [măsurat în] funcție de unghiul dintre cei doi vectori.

obicei, vectorii considerați într-un astfel de context sunt nenegativi (i.e., au toate componente pozitive sau 0), așa că atunci valoarea maximă a lui cos va fi 1 (și anume, pentru cazul $x = y$), iar 0 va fi valoarea sa minimă (și anume, pentru cazul când x și y sunt vectori ortogonali).

Observație: Trebuie să știm că, deși uneori această funcție este numită „distanță cosinus“, ea nu este de fapt o măsură de distanță, fiindcă nu satisfac condițiile de nenegativitate, identitatea indiscernabililor și inegalitatea triunghiului.⁸⁸⁹

a. Folosind funcția cos, definim *coezunea* (internă) a unui cluster oarecare X astfel:⁸⁹⁰

$$Coh(X) = \frac{1}{C_{|X|}^2} \cdot \frac{1}{2} \cdot \sum_{x \neq y \in X} \cos(x, y),$$

adică exact *media similarității* instanțelor din clusterul X . În formula aceasta, am notat cu $|X|$ cardinalul clusterului X , adică numărul de instanțe din X .

Arătați că $Coh(X)$ poate fi exprimată folosind doar $|X|$ și centrul de greutate (sau, *centroidul*) clusterului X . Acesta din urmă este notat cu μ_X și este definit ca $\frac{1}{|X|} \sum_{x \in X} x$.

b. Date fiind două clustere oarecare disjuncte A și B , particularizați expresia funcției de coezune (Coh) care a fost definită la punctul precedent pentru clusterul $A \cup B$. Ulterior veți exprima $Coh(A \cup B)$ folosind doar $|A|$, $|B|$, μ_A și μ_B .

c. Arătați că la clusterizare ierarhică cu similaritate de tip “average-linkage” bazată pe măsura cos, date fiind două clustere oarecare disjuncte A și B , „distanță“ dintre clusterul $A \cup B$ și un cluster oarecare X disjunct de A și de B se poate exprima folosind doar cardinalii $|A|$, $|B|$ și $|X|$ și centroizii μ_A , μ_B și μ_X .

Sugestie: Ca o consecință ce decurge din definiția similarității “average-linkage”, această „distanță“ se poate calcula cu ajutorul coezuinilor $Coh((A \cup B) \cup X)$, $Coh(A \cup B)$ și $Coh(X)$.

Observație: Ca și la problema 35.cd, va rezulta că la fiecare iterație a algoritmului de clusterizare aglomerativă, matricea de „proximitate“ poate fi actualizată folosind doar conținutul acestei matrice la iterarea precedentă, cardinalii clusterelor ($n_X \stackrel{\text{not.}}{=} |X|$) și centroizii lor (μ_x).

37. (O legătură între clusterizarea ierarhică și arborii de decizie)

• ○ CMU, 2012 spring, Ziv Bar-Joseph, midterm, pr. 9.3

Considerăm un set de instanțe etichetate. Ne propunem să folosim un *arbore de clusterizare ierarhică* (adică, o dendrogramă) pe post de *arbore de decizie*. Pentru aceasta, selectăm o *măsură de similaritate* de tip *linkage* și construim arboarele de clusterizare ierarhică bazat pe setul de date de antrenament furnizat, fără a ține cont de etichetele instanțelor. Considerăm o instanță oarecare

⁸⁸⁹Pentru formalizarea acestor proprietăți, vedeți problema 2 de la capitolul *Învățare bazată pe memorare*.

⁸⁹⁰Prin C_n^2 notăm ca de obicei numărul de combinări de n obiecte luate câte 2.

de test (engl., *test sample* sau *query sample*) x_q . Vom parcurge cu ea arborele pe care tocmai l-am construit, în felul următor.

La fiecare nod din arbore calculăm distanțele dintre [clusterul singleton format doar din instanța] x_q și fiecare dintre sub-clusterele care sunt descendenții imediați ai nodului respectiv, folosind măsura de similaritate pe care am ales-o mai înainte, după care selectăm ramura corespunzătoare minimului dintre aceste distanțe. În acest fel, indiferent care ar fi instanța de test considerată, putem să o propagăm (de sus în jos) în arborele de clusterizare ierarhică. La final, adică atunci când se ajunge la un nod frunză (care reprezintă un cluster singleton), asignăm lui x_q eticheta nodului respectiv.

Vă cerem să identificați o *măsură de similaritate* de tip *linkage* care are proprietatea că atunci când este folosită în maniera descrisă mai sus funcționează exact ca un arbore de decizie care (la rândul lui) este echivalent cu un anumit algoritm de clasificare automată pe care l-am prezentat la curs. Care este acest clasificator? (Justificați într-o manieră suficient de detaliată.)

38.

(Clusterizare ierarhică bottom-up: implementare)

Liviu Ciortuz, 2015

- a. Elaborați — de preferință în limbajul de programare C/C++ — implementarea algoritmului de clusterizare ierarhică bottom-up.

Veți considera că instanțele de clusterizat sunt puncte într-un spațiu euclidian \mathbb{R}^d (d fiind un număr natural, $d \geq 1$). Veți presupune de asemenea că instanțele sunt înregistrate într-un fișier text, câte o instanță $x_i = (x_{i_1}, x_{i_2}, \dots, x_{i_d})$ pe fiecare linie, și că între fiecare două atribute / coordonate succesive ale unei astfel de instanțe se află cel puțin un [caracter de tip] spațiu. Numele acestui fișier text, eventual numărul de dimensiuni al spațiului în care se lucrează (d), precum și tipul funcției de similaritate (*single-linkage*, *complete-linkage*, *average-linkage* sau *metrica lui Ward*⁸⁹¹) vor fi indicate ca argumente în linia de comandă a programului.

Ca punct de plecare puteți considera pseudo-codul din *Foundations of Statistical Natural Processing*, C. Manning, H. Schütze, MIT Press, 2002, pag. 496, pe care însă îl veți extinde cu calculul înălțimilor asociate nodurilor din dendrograma rezultantă.

Distanțele dintre instanțele date vor fi calculate și apoi memorate într-o matrice.⁸⁹² Pentru a face în mod eficient actualizarea acestei matrice la fiecare iterare a algoritmului de clusterizare ierarhică bottom-up, veți folosi rezultatele de la problema 35.

Înălțimea unui de nod oarecare din dendrogramă va fi dată de media tuturor distanțelor $d(x, y)$, unde x și y sunt puncte în clusterul respectiv (vedeți, ca exemplu, problema 1).⁸⁹³

Ca să „reprazentați” în mod convenabil ieșirea programului puteți folosi o notație bazată pe expresii cu paranteze imbricate.⁸⁹⁴ Ulterior, veți putea extinde eventual programul, cuplându-l cu o interfață grafică, atât pentru colectarea intrărilor (vedeți de exemplu problema 31) cât și pentru ieșiri (trasarea efectivă a dendrogramelor).

Vă sugerăm să testați programul pe datele din problemele 1, 2, 27 și 28.

- b. Extindeți implementarea în aşa fel încât, dacă se solicită (printr-o anumită opțiune la linia de comandă), să se livreze toate soluțiile posibile. Testați această variantă a programului dumneavoastră folosind datele de la problema 3.

⁸⁹¹Vedeți problema 33.

⁸⁹²Inițial veți lucra cu o matrice pătratică. Ulterior, pentru motive de economie de memorie, veți putea înlocui această matrice cu una triunghiulară (intuitiv, aceasta din urmă reprezintă zona de desupra diagonalei principale a matricei pătratice, memorată sub o formă convenabilă).

⁸⁹³Coeziunea unui cluster se definește ca fiind inversul acestei medii.

⁸⁹⁴De exemplu, pentru dendrograma obținută la problema 1 se poate folosi notația simplă (“flat hierarchy”):

$$((x_1, x_2), ((x_3, ((x_4, x_5), x_6)), ((x_7, x_8), (x_9, x_{10}))))$$

sau, o variantă extinsă cu informații despre ordinea de formare a clusterelor și înălțimile nodurilor (interne) corespunzătoare clusterelor în dendrogramă.

$$((x_1, x_2)_{C1}^{0.2}, ((x_3, ((x_4, x_5)_{C1}^{0.1}, x_6)_{C5}^{0.2})_{C7}^{0.3(6)}, ((x_7, x_8)_{C2}^{0.1}, (x_9, x_{10})_{C3}^{0.1}{}_{C6}^{0.2(3)})_{C8}^{1.1})_{C9}^{1.77(3)}.$$

39.

(Clusterizare ierarhică top-down: implementare)

□ Liviu Ciortuz, 2016

Implementați — de preferință în C/C++ — algoritmul de clusterizare ierarhică top-down prezentat în problema rezolvată 6. Pentru conveniență, puteți porni de la o implementare oarecare disponibilă pe web pentru algoritmul lui Kruskal sau algoritmul lui Prim pentru aflarea arborelui de cost minim (MST) dintr-un graf. Testați implementarea realizată de dumneavoastră pe datele de la problema pe care am menționat-o anterior.

7.2.2 Algoritmul K -means

40.

(Algoritmul K -means: aplicare în \mathbb{R}^2)

■ ○ T.U. Dresden, 2006 summer, S. Hölldobler, A. Grossmann, HW3

Folosiți algoritmul K -means și distanța euclidiană pentru a grupa următoarele 8 instanțe din \mathbb{R}^2 în 3 clustere:

$$A(2, 10), B(2, 5), C(8, 4), D(5, 8), E(7, 5), F(6, 4), G(1, 2), H(4, 9).$$

Se vor lua drept centroizi inițiali punctele A , D și G .

a. Rulați prima iterație a algoritmului K -means. Pe un grid de valori 10×10 veți marca instanțele date, pozițiile centroizilor la începutul primei iterări și compoñența fiecărui cluster la finalul acestei iterări. (Trasați mediatoarele segmentelor determinate de centroizi, ca *separatori* ai clusterelor.)

b. Câte iterări sunt necesare pentru ca algoritmul K -means să conveargă? Desenați pe câte un grid rezultatul rulării fiecărei iterări.

41.

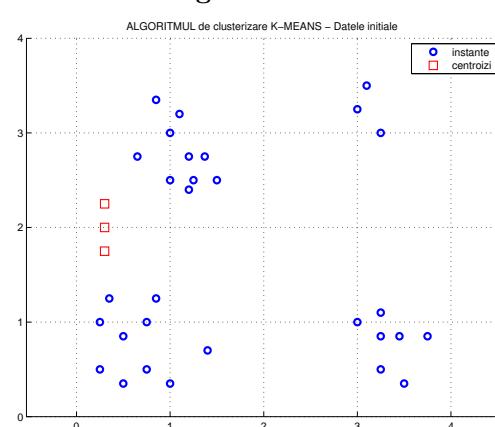
(Algoritmul K -means: aplicare pe date din \mathbb{R}^2)

* CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW3, pr. 5

Aplicați algoritmul K -means pe setul de date din imaginea următoare.

Cerculețele reprezintă instanțele de clusterizat, iar pătrățelele sunt centroizii inițiali ai clusterelor. Pentru fiecare iterăție a algoritmului desenați centroizii și separatorii care definesc fiecare cluster. Folosiți oricătre imagini aveți nevoie până ajungeți la convergență.

Coordonatele acestor instanțe, precum și cele ale centroizilor vă sunt puse la dispozitie în fișiere depuse pe site-ul acestei cărți.⁸⁹⁵



⁸⁹⁵<http://profes.info.uaic.ro/~ciortuz/ML.ex-book/res/CMU.2004f.TM+AM.HW3.pr5.cl.dat>,
<http://profes.info.uaic.ro/~ciortuz/ML.ex-book/res/CMU.2004f.TM+AM.HW3.pr5.init.dat>.

Observație: La execuția algoritmului se consideră că în cazul în care un centroid nu are puncte asignate lui, atunci el rămâne pe loc în iterată respectivă.

42.

(Algoritmul K -means: aplicare în \mathbb{R} și \mathbb{R}^2 ; verificarea monotoniei „criteriului“ coeziunii intra-clustere (J))
 ■ (pt. punctul a) □ * *Liviu Ciortuz, 2020*

În acest exercițiu veți folosi algoritmul K -means în varianta dată în enunțul exercițiului 12.⁸⁹⁶

Vă readucem aminte definiția aşa-numitului criteriu J , care este o măsură a coeziunii intra-clustere:

$$J(C^{(t)}, \mu^{(t)}) = \sum_{i=1}^n (x_i - \mu_{C^{(t)}(x_i)}^{(t)})^2,$$

unde $C^{(t)}$ este ansamblul clusterelor la momentul / iterată t , apoi $\mu^{(t)}$ desemnează ansamblul centroizilor la iterată t și, în sfârșit, $\mu_{C^{(t)}(x_i)}^{(t)}$ este centroidul clusterului la care este asignată instanța x_i la iterată t .

a. Fie următorul set de date din \mathbb{R} : -9, -8, -7, -6, -5, 5, 5, 6, 6, 7, 7, 8, 8, 9, 9. Considerăm $\mu_1^{(0)} = -20$ și $\mu_2^{(0)} = -10$.⁸⁹⁷ Demonstrați în manieră analitică (NU numeric!) că pentru $t = 1$ avem

$$J(C^{(t)}, \mu^{(t)}) \leq J(C^{(t-1)}, \mu^{(t-1)}).$$

b. Cerințele de la acest punct sunt similare cu cele de la punctul precedent, însă de data aceasta veți lucra pe următorul set de date din \mathbb{R}^2 : $A(-1, 0)$, $B(1, 0)$, $C(0, 1)$, $D(3, 0)$, $E(3, 1)$; drept centroizi inițiali veți considera $\mu_1^{(0)} = A$ și $\mu_2^{(0)} = E$.⁸⁹⁸

43.

(Algoritmul K -means: convergență)
 * *CMU, 2009 spring, Ziv Bar-Joseph, final exam, pr. 1.7*

Fie un set de date neetichetate și o K -partiție a acestui set de date, generată la sfârșitul unei iterări oarecare a algoritmului K -means. Este posibil ca algoritmul K -means să revizeze această K -partiție?

Observație: Notiunea de K -partiție a fost introdusă la problema 11. Aveți grija că notiunea de K -partiție revizitată de mai sus nu este aceeași cu cea de K -partiție stabilă de la problema 11.

⁸⁹⁶Atenție! În eventualitatea că la o iterată oarecare a algoritmului K -means un cluster este vid, centroidul său rămâne pe loc la iterată respectivă.

⁸⁹⁷Acestea sunt datele de la exercițiul 17.

⁸⁹⁸Acestea sunt datele de la exercițiul 7.

44. (Algoritmul K -means: discuție asupra alegerii valorii lui K)

• o CMU, 2010 fall, Aarti Singh, HW3, pr. 5.4
CMU, 2015 spring, T. Mitchell, N. Balcan, HW7, pr. 1.1

Unul dintre dezavantajele algoritmului K -means este acela că trebuie specificată valoarea parametrului K . Ce părere aveți despre următoarea *strategie*:

Se poate alege K în mod automat, încercând toate valorile lui posibile ($K = 1, \dots, n$, unde n este numărul de instanțe de clusterizat), și reținând apoi acea valoare a lui K pentru care s-a obținut cea mai mică valoare a criteriului „sumei celor mai mici pătrate“, J_K .⁸⁹⁹

Justificați de ce această strategie este bună (sau rea).

45. (Algoritmul K -means ca algoritm de optimizare:
maximizarea aproximativă a distanțelor dintre centroizii clusterelor)

■ o CMU, 2010 fall, Aarti Singh, HW3, pr. 5.2

În această problemă vom lucra cu o versiune a algoritmului K -means ușor modificată față de cea dată în enunțul problemei 12.

Fie $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ o mulțime de instanțe, iar K numărul de clustere cu care vom lucra. De data aceasta vom specifica asignările instanțelor la clustere folosind o *matrice-indicator* $\gamma \in \{0, 1\}^{n \times K}$, cu $\gamma_{ij} = 1$ dacă și numai dacă \mathbf{x}_i aparține clusterului j . Vom impune ca fiecare instanță să aparțină câte unui singur cluster, deci $\sum_{j=1}^K \gamma_{ij} = 1$.

După cum s-a arătat la problema 12, algoritmul K -means „estimează“ matricea γ făcând minimizarea criteriului (sau, a „măsurii de distorsiune“) J , pe care, folosind matricea γ , îl rescriem sub forma

$$J(\gamma, \mu_1, \mu_2, \dots, \mu_K) = \sum_{i=1}^n \sum_{j=1}^K \gamma_{ij} \|\mathbf{x}_i - \mu_j\|^2,$$

unde $\|\cdot\|$ desemnează norma vectorială L_2 . Concret, algoritmul K -means alternează „estimarea“ matricei γ cu re-calcularea centroizilor μ_j .

Acestea fiind spuse, putem da acum noua versiune a algoritmului K -means:⁹⁰⁰

- Se initializează în mod arbitrar centroizii $\mu_1, \mu_2, \dots, \mu_K$ și se ia $C = \{1, \dots, K\}$.
- Atâtă timp cât valoarea lui J descrește în mod strict,⁹⁰¹ repetă:

Pasul 1: Calculează γ astfel:

$$\gamma_{ij} \leftarrow \begin{cases} 1, & \text{dacă } \|\mathbf{x}_i - \mu_j\|^2 \leq \|\mathbf{x}_i - \mu_{j'}\|^2, \forall j' \in C, \\ 0, & \text{în caz contrar.} \end{cases}$$

În caz de egalitate, alege în mod arbitrar căruia cluster (dintre cele eligibile) să-i aparțină \mathbf{x}_i .

⁸⁹⁹ Pentru definiția riguroasă a acestui criteriu vedeți problema 12.

⁹⁰⁰ Această versiune a algoritmului K -means poate servi pentru a introduce [o variantă simplă pentru] algoritmul EM/GMM — vedeți pseudo-codul de la exercițiul 15 — ca versiune probabilistă a algoritmului K -means. De asemenea, această versiune a algoritmului K -means poate servi pentru a demonstra o proprietate interesantă, și anume, *convergența* [în anumite condiții a] rezultatelor acestor doi algoritmi; vedeți exercițiul 64.

⁹⁰¹ Această condiție de oprire este ușor diferită de cea formulată de Lloyd: oprește execuția algoritmului atunci când matricea γ nu se mai schimbă.

Pasul 2: Recalculează μ_j folosind matricea γ actualizată:

Pentru fiecare $j \in C$, dacă $\sum_{i=1}^n \gamma_{ij} > 0$, asignează

$$\mu_j \leftarrow \frac{\sum_{i=1}^n \gamma_{ij} \mathbf{x}_i}{\sum_{i=1}^n \gamma_{ij}}.$$

Altfel, menține neschimbăt centroidul μ_j .

Vom nota cu $\bar{\mathbf{x}}$ media instanțelor date și vom considera următoarele trei cantități:⁹⁰²

$$\text{Variația totală: } T(X) = \frac{\sum_{i=1}^n \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2}{n}$$

$$\text{Variația intra-clustere: } W_j(X) = \frac{\sum_{i=1}^n \gamma_{ij} \|\mathbf{x}_i - \mu_j\|^2}{\sum_{i=1}^n \gamma_{ij}} \text{ pentru } j = 1, \dots, K$$

$$\text{Variația inter-clustere: } B(X) = \sum_{j=1}^K \left(\frac{\sum_{i=1}^n \gamma_{ij}}{n} \right) \|\mu_j - \bar{\mathbf{x}}\|^2.$$

a. Care este relația dintre aceste trei cantități?

Observație: Veți ține cont că această relație poate să conțină un termen suplimentar care nu este menționat mai sus.

b. Folosind relația stabilită la punctul a, arătați că putem interpreta algoritmul K -means ca tînzând să minimizeze (și anume descrescând, dar nu neapărat strict monoton) o medie ponderată a variației intra-clustere în timp ce el tinde să maximizeze (crescând) însă doar în mod *aproximativ* variația inter-clustere.

46.

(Algoritmul K -means: aplicare pe date din \mathbb{R}^3
calcularea variației / coeziunii intra- și inter-clustere)

prelucrare de Liviu Ciortuz, 2021, după

□ • Andreas Wickert, Luis Sa-Couto,

Machine Learning – A Journey to Deep Learning, 2021, pag. 370-385

Fie următoarele instanțe de antrenament, neetichetate:

$$x_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, x_2 = \begin{pmatrix} 8 \\ 8 \\ 4 \end{pmatrix}, x_3 = \begin{pmatrix} 3 \\ 3 \\ 0 \end{pmatrix}, x_4 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, x_5 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, x_6 = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}.$$

a. Aplicați algoritmul K -means cu $K = 2$ pe acest set de date, până ajungeți la convergență. Pentru inițializarea centroizilor, veți lua primele K instanțe din setul de date furnizat.

⁹⁰²Veți observa că $T(X)$, prima dintre aceste cantități este o „măsură” a împărăștierii instanțelor \mathbf{x}_i față de $\bar{\mathbf{x}}$, centrul de greutate al întregii mulțimi; ea nu depinde de μ sau de γ . Mai exact, $T(X)$ este media pătratelor distanțelor de la instanțele din mulțimea X la $\bar{\mathbf{x}}$. Similar, $W_j(X)$ este media pătratelor distanțelor de la instanțele din clusterul j la centroidul μ_j , care este totușa cu raportul dintre coeziunea clusterului j și cardinalul acestui cluster. În fine, $B(X)$ este suma ponderată a pătratelor distanțelor de la centroizii μ_j la $\bar{\mathbf{x}}$, centrul de greutate al mulțimii de instanțe X .

- b. Aplicați algoritmul K -means cu $K = 3$ pe același set de date, până ajungeți la convergență. Pentru inițializarea centroizilor, veți lua primele K instanțe din setul de date furnizat.
- c. Ce valoare a lui $K \in \{2, 3\}$ produce o clusterizare mai bună dacă se folosește drept criteriu de evaluare *variația / coeziunea intra-clustere* (adică, suma pătratelor distanțelor de la fiecare instanță la centroidul cel mai apropiat)?
- d. Ce valoare a lui $K \in \{2, 3\}$ produce o clusterizare mai bună dacă se folosește drept criteriu de evaluare *variația / coeziunea inter-clustere*? (Vedeți definiția de la problema 45.)

47.

(Clusterizare partițională în \mathbb{R} :un algoritm de programare dinamică, cu complexitate $\mathcal{O}(Kn^2)$ pentru calculul minimului criteriului „celor mai mici pătrate“ (J))*prelucrare de Liviu Ciortuz, după**□ • ○ CMU, 2015 spring, T. Mitchell, N. Balcan, HW7, pr. 1.5-8*

Stim că, date fiind instanțele $x_1, \dots, x_n \in \mathbb{R}^d$ cu $d \in \mathbb{N}^*$, algoritmul de clusterizare K -means urmărește să găsească centroizii a K clustere, și anume $\mu_j \in \mathbb{R}^d, j \in \{1, \dots, K\}$, astfel încât suma pătratelor distanțelor de la fiecare instanță la cel mai apropiat centroid să fie minimă (vedeți problema 12). Altfel spus, K -means încearcă să găsească acei μ_1, \dots, μ_K care minimizează valoarea așa-numitului criteriu al „celor mai mici pătrate“:

$$J \stackrel{\text{def.}}{=} \sum_{i=1}^n \min_{j \in \{1, \dots, K\}} \|x_i - \mu_j\|^2,$$

Pentru a realiza aceasta, algoritmul K -means procedează iterativ, alternând asignarea fiecărei instanțe x_i ($i = 1, \dots, n$) la cel mai apropiat centroid cu actualizarea centroizilor clusterelor (concret, μ_j va deveni media instanțelor asignate la clusterul j).

În general, găsirea minimului criteriului J pentru o valoare fixată a lui K este o problemă NP-dificilă (engl., NP-hard). Totuși, se poate arăta că ea poate fi rezolvată în timp polinomial (în n și K) dacă instanțele de clusterizat sunt într-un spațiu unidimensional ($d = 1$). La acest exercițiu ne vom concentra atenția asupra acestui caz.

- a. Să considerăm situația în care $K = 3$ și ne sunt date 4 instanțe, $x_1 = 1, x_2 = 2, x_3 = 5$ și $x_4 = 7$. Care este clusterizarea optimă pentru acest set de date? Cât este valoarea corespunzătoare pentru funcția obiectiv J ?
- b. Am putea fi tentați să credem că în cazul $d = 1$ algoritmul K -means converge în mod cert la valoarea minimă (globală) a criteriului J . Considerând din nou instanțele date la punctul a, arătați că există o asignare suboptimală a lor la clustere, pe care algoritmul K -means nu poate îmbunătăți. (Indicați asignarea, arătați de ce este suboptimală și explicați de ce anume ea nu va putea fi îmbunătățită.)
- c. Presupunem că sortăm instanțele noastre astfel încât $x_1 \leq x_2 \leq \dots \leq x_n$. Demonstrați că orice *asignare optimă* a acestor instanțe la clustere — în sensul că orice instanță este asociată la cel mai apropiat centroid — are proprietatea

că fiecare cluster nevid corespunde unui anumit „interval“ de instanțe. Adică, pentru fiecare cluster nevid j există $i_1, i_2 \in \{1, \dots, n\}$, cu $i_1 \leq i_2$, astfel încât clusterul acesta constă din instanțele $\{x_{i_1}, x_{i_1+1}, \dots, x_{i_2}\}$.⁹⁰³

d. Concepți un algoritm de clusterizare de tip programare dinamică având complexitatea $O(Kn^2)$, ca înlocuitor pentru algoritmul K -means în cazul unidimensional.

Indicație: Dat fiind rezultatul de la punctul c, ceea ce trebuie să „optimizăm“ / setăm sunt cele $K - 1$ granițe / margini (engl., boundaries) ale clusterelor, marginea cu numărul de ordine i fiind cea mai mare instanță din clusterul i .

48. (K-means pentru comprimarea imaginilor)
 • o CMU, 2017 fall, Nina Balcan, HW5, pr. 3, question 12

La curs am arătat că algoritmul K -means poate fi folosit pentru comprimarea imaginilor.

Presupunem că sunt necesari 24 de biți pentru a memora valorile celor trei componente {R, G, B} în care se descompune culoarea fiecărui pixel. Cât biți sunt necesari pentru a memora o imagine care are N pixeli?

Vom presupune acum că folosim algoritmul K -means (cu K clustere) pentru a comprima imagini și că identificăm [culoarea pentru] fiecare pixel folosind id-ul clusterului său (un număr din multimea $\{0, \dots, K - 1\}$). Cât biți sunt necesari acum pentru a memora imaginea [care are N pixeli]?

Care este raportul de compresie?

49. (Algoritmul K -means: aplicare, folosind distanța euclidiană, respectiv distanța Manhattan / norma L_1 robustețea algoritmului K -means la prezența outlierelor)
 prelucrare de Liviu Ciortuz, după
 • o CMU, 2010 fall, Aarti Singh, HW3, pr. 5.3
 CMU, 2014 spring, B. Poczos, A. Singh, HW3, pr. 1.4
 CMU, 2014 spring, Seyoung Kim, HW3, pr. 1.1

Fie setul date alăturat (X), fiecare rând / linie reprezentând o instanță.

A. K -means folosind distanța euclidiană

Aplicați algoritmul K -means pe acest set de date, folosind $K = 3$ și distanța euclidiană. La pasul de inițializare nu veți seta pozițiile centroizilor ci, în schimb, veți inițializa clusterele, după cum urmează:

$C1 : \{A, B, F\}$, $C2 : \{C, H, I\}$, $C3 : \{D, E, G\}$.

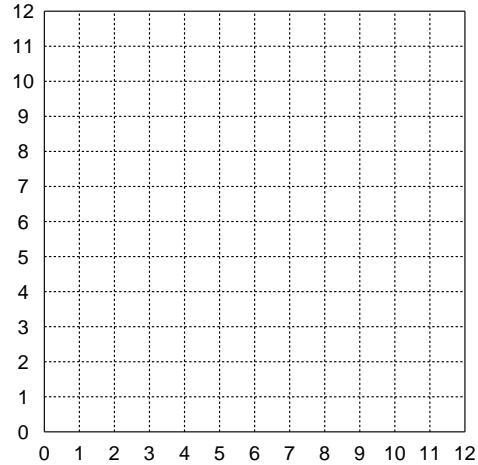
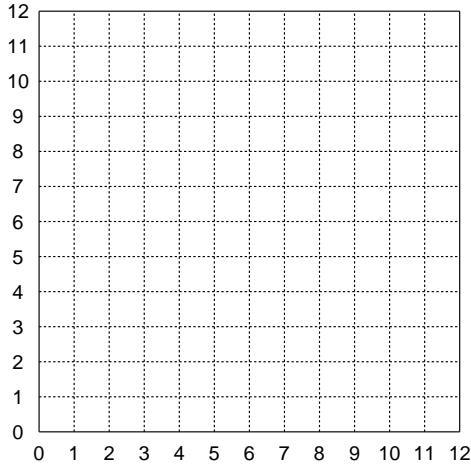
În consecință, veți inversa pașii 1 și 2 din ciclul iterativ al algoritmului K -means.

A	(1, 1)
B	(3, 3)
C	(6, 6)
D	(6, 12)
E	(9, 9)
F	(11, 11)
G	(0, 3)
H	(3, 0)
I	(9, 3)

⁹⁰³Se poate observa imediat că de fapt această proprietate este satisfăcută mereu pe parcursul aplicării algoritmului K -means pe seturi de date din \mathbb{R} .

Folosiți gridurile de mai jos; puteți adăuga și altele decă veți considera că este necesar.

Puteți proceda fie în *maniera analitică* (calculând distanțele de la instanțe la centroizi), fie trasând mediatoarele segmentelor care unesc centroizii.



B. K-means folosind distanța Manhattan

În această parte a exercițiului vom folosi distanța Manhattan, desemnată prin $\|\cdot\|_1$, ceea ce înseamnă că vom defini [ca nouă „măsură de distorsiune“] funcția

$$J_1(\gamma, \mu_1, \mu_2, \dots, \mu_K) = \sum_{i=1}^n \sum_{k=1}^K \gamma_{ik} \|\mathbf{x}_i - \mu_k\|_1.$$

Vă reamintim faptul că *distanța Manhattan* (d_M) dintre un vector $x = (x_1, \dots, x_p)$ și un alt vector $y = (y_1, \dots, y_p)$, ambele fiind din \mathbb{R}^p , este definită astfel:

$$d_M = \sum_{i=1}^p |x_i - y_i|.$$

Vom minimiza funcția J_1 folosind următoarea variantă a algoritmului K-means

- Inițializează centroizii μ_1, \dots, μ_K .
- Atât timp cât valoarea lui J_1 încă scade, repetă următorii pași:
 1. Calculează [variabilele-indicator] γ astfel:

$$\gamma_{ik} \leftarrow \begin{cases} 1, & \|\mathbf{x}_i - \mu_k\|_1 \leq \|\mathbf{x}_i - \mu_{k'}\|_1, \forall k' \in C, \\ 0, & \text{în caz contrar.} \end{cases}$$

În cazul în care minimul se atinge pentru mai mulți astfel de k , veți asigna x_i la centroidul cu indicele cel mai mic pentru care se atinge distanța minimală.

2. Recalculează μ_k folosind variabilele γ care tocmai au fost actualizate: pentru orice $k = 1, \dots, K$, dacă $\sum_{i=1}^n \gamma_{ik} > 0$, redefinește centroidul $\mu_k \stackrel{\text{not}}{=} ((\mu_k)_1, \dots, (\mu_k)_p)$ astfel:

$$(\mu_k)_d \leftarrow \text{median}(\{(\mathbf{x}_i)_d \mid \gamma_{ik} = 1\}), \text{ pentru } d = 1, \dots, p, \quad (381)$$

unde funcția median () este specificată în *Definiția* de mai jos; în caz contrar, lasă centroidul μ_k neschimbat.

Definiție: Fie o mulțime formată din n numere reale, $\{x_1, x_2, \dots, x_n\}$. Dacă $x_1 \leq x_2 \leq \dots \leq x_n$, atunci

- i. pentru cazul în care n este impar, *valoarea mediană* a secvenței x_1, x_2, \dots, x_n este $x_{\lceil n/2 \rceil}$, unde simbolul $\lceil z \rceil$ desemnează partea întreagă superioară a numărului z ;
- ii. pentru cazul în care n este par, *valoarea mediană* a secvenței x_1, x_2, \dots, x_n este orice număr din intervalul $[x_{n/2}, x_{(n/2)+1}]$.

Proprietate: Soluția problemei de minimizare

$$\min_{\mu} \sum_{i=1}^n |x_i - \mu|$$

este *valoarea mediană* pentru mulțimea $\{x_1, x_2, \dots, x_n\}$.⁹⁰⁴

a. Demonstrația următoare fundamentează formula (381):

$$\begin{aligned} & \min_{\mu_1, \dots, \mu_K} \sum_{i=1}^n \sum_{k=1}^K \gamma_{ik} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|_1 \\ &= \min_{\mu_1, \dots, \mu_K} \sum_{k=1}^K \sum_{i=1}^n \gamma_{ik} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|_1 \\ &= \sum_{k=1}^K \min_{\boldsymbol{\mu}_k} \sum_{i=1}^n \gamma_{ik} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|_1 \\ &= \sum_{k=1}^K \min_{\boldsymbol{\mu}_k} \sum_{i=1}^n \sum_{d=1}^p \gamma_{ik} |(\mathbf{x}_i)_d - (\boldsymbol{\mu}_k)_d| \\ &= \sum_{k=1}^K \min_{\boldsymbol{\mu}_k} \sum_{d=1}^p \sum_{i=1}^n \gamma_{ik} |(\mathbf{x}_i)_d - (\boldsymbol{\mu}_k)_d| \\ &= \sum_{k=1}^K \sum_{d=1}^p \min_{(\boldsymbol{\mu}_k)_d} \sum_{i=1}^n \gamma_{ik} |(\mathbf{x}_i)_d - (\boldsymbol{\mu}_k)_d| \end{aligned}$$

Justificați fiecare pas al acestei demonstrații.

b. Dacă setul de date de clusterizat conține *outliere*, ce versiune a algoritmului *K-means* ar fi indicat să o folosiți — aceasta (cu distanță Manhattan / L_1) ori cea originală (cu distanță euclidiană / L_2)? Justificați.

c. Aplicați algoritmul *K-means* pe același set de date ca la punctul A — tot cu $K = 3$ și aceleși inițializări ale clusterelor ca acolo —, folosind însă distanța Manhattan în scrierea criteriului / funcției de „distorsiune“, aşa cum s-a arătat în versiunea lui *K-means* de mai sus.

Atenție: Inițializarea de aici diferă de cea din pseudo-codul dat în enunț; veți proceda în consecință, adică veți inversa pașii 1 și 2 din ciclul iterativ al

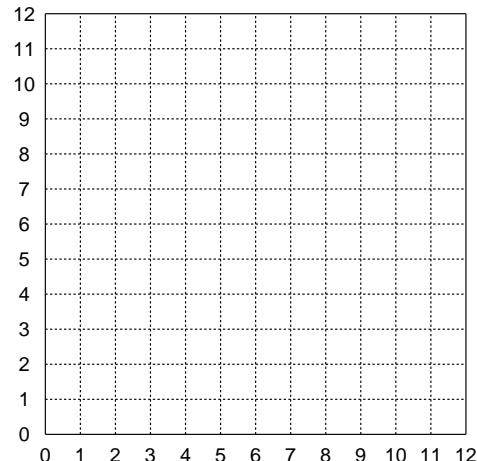
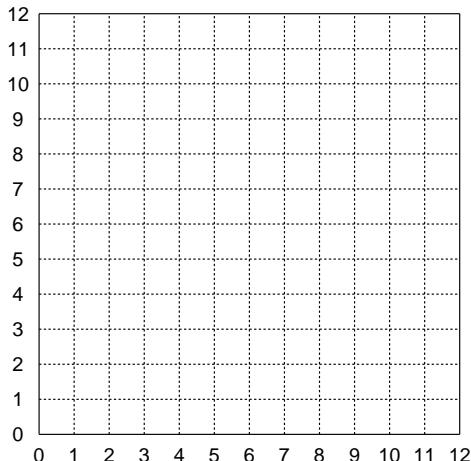
⁹⁰⁴În consecință, răspunsul la problema noastră nu este în mod neapărat unic!

algoritmului K -means. Si (din nou!) nu veți inițializa pozițiile centroizilor în mod arbitrar!

Folosiți gridurile următoare; puteți adăuga și altele decă veți considera că este necesar.

Observație (1): Nu este nevoie să calculați valorile variabilelor-indicator γ_{ij} ; în schimb, pur și simplu asignați fiecare instanță la cel mai apropiat centroid. Pentru aceasta va trebui să faceți efectiv calculul distanțelor Manhattan; nu veți trasa mediatore!

Observație (2): În situația în care va trebui să aplicați cazul ii al definiției pe care am dat-o mai sus pentru *valoarea mediană*, veți considera că funcția $\text{median}()$ returnează mijlocul intervalului $[x_{n/2}, x_{(n/2)+1}]$.



d. Demonstrați *proprietatea* care a fost dată în enunț [particularizată] pentru următoarele secvențe de numere:

- i. 1, 3, 11
- ii. 0, 1, 3, 3

C. Comparați rezultatele de la punctele A și B.c. Dacă ele sunt diferite, care este justificarea?

50. (Algoritmul K -means++: comparație cu algoritmul K -means)

Liviu Ciortuz, 2018, după

- *CMU, 2017 fall, Nina Balcan, midterm, pr. 3.2*
- *CMU, 2012 fall, E. Xing, A. Singh, HW3, pr. 1*

Comentariu:

În exercițiul de față vom considera că la faza de inițializare algoritmul K -means alege centroizii în mod arbitrar dintre instanțele de clusterizat.

Dacă datele de clusterizat sunt [a priori] bine separate în K clustere, atunci este foarte posibil ca la finalul inițializării să existe [măcar] un cluster din care algoritmul nu a selectat niciun punct. În astfel de situații, algoritmul K -means nu va produce clusterele dorite de noi.

În schimb, varianta *K-means++* a algoritmului *K-means*, propusă de David Arthur și Sergei Vassilvitskii în 2007, încearcă să selecteze pentru pozițiile inițiale ale celor *K* centroizi instanțe care sunt [pe cât se poate] mai distanțate unele de altele. În acest fel, pot fi selectate, cu o probabilitate [destul de] mare, instanțe din toate clusterelor.

Formalizare:

K-means++ face inițializarea centroizilor în maniera următoare:

- i. alege primul centroid, μ_1 , în manieră uniform aleatorie dintre instanțele de clusterizat, x_1, \dots, x_n . Cu alte cuvinte, alegem mai întâi un indice i în mod uniform aleatoriu din mulțimea $\{1, \dots, n\}$ și fixăm $\mu_1 = x_i$.
- ii. pentru $j = 2, \dots, K$:

- Pentru fiecare instanță x_i , calculează distanța D_i până la cel mai apropiat centroid ales / fixat la o iterație anterioară:

$$D_i = \min_{j'=1, \dots, j-1} \|x_i - \mu_{j'}\|.$$

- Alege centroidul μ_j în mod aleatoriu dintre instanțele x_1, \dots, x_n , cu probabilitate direct proporțională cu D_1^2, \dots, D_n^2 . Altfel spus, alegem un indice i în mod aleatoriu din mulțimea $\{1, \dots, n\}$ cu probabilități egale cu $\frac{D_1^2}{\sum_{i'=1}^n D_{i'}^2}, \dots, \frac{D_n^2}{\sum_{i'=1}^n D_{i'}^2}$, și fixăm $\mu_j = x_i$.
- iii. Returnează $\mu \stackrel{\text{not.}}{=} (\mu_1, \dots, \mu_K)$, setul de asignări ale pozițiilor inițiale ale centroizilor clusterelor pentru algoritmul lui Lloyd (*K-means*).

Vom ilustra acum diferența dintre *K-means++* și *K-means* [la inițializare], folosind un set de date simplu, format din cinci puncte în planul euclidian bidimensional: $A(-1, 0)$, $B(1, 0)$, $C(0, 1)$, $D(3, 0)$ și $E(3, 1)$.⁹⁰⁵

- a. Presupunem că aplicăm algoritmul *K-means* cu $K = 2$, făcând inițializarea centroizilor cu instanțe din mulțimea pe care tocmai am precizat-o. Presupunem că a fost deja ales centroidul $\mu_1 = A$. Dacă selecția următorului centroid (μ_2) se face în manieră uniform aleatorie — aşa cum procedează îndeobște algoritmul *K-means* — din mulțimea $\{B, C, D, E\}$, care este probabilitatea ca μ_2 să fie din submulțimea $\{B, C\}$? Justificați.
- b. Aplicăm acum pe același set de date algoritmul *K-means++*, tot cu $K = 2$. Presupunem, ca și la punctul a, că a fost selectat centroidul $\mu_1 = A$. Care este acum probabilitatea ca μ_2 să fie selectat din submulțimea $\{B, C\}$? Rezultă oare într-adevăr o îmbunătățire semnificativă față de [inițializarea făcută de] algoritmul *K-means*? Justificați riguros.

⁹⁰⁵ Aceste date sunt preluate de la problema 7.

51.

(Algoritmul K -means: varianta „kernelizată“)*prelucrare de Liviu Ciortuz, după***■ □ • ○ CMU, 2015 spring, T. Mitchell, N. Balcan, HW7, pr. 1.2-4**

Atunci când aplicăm algoritmul K -means folosind distanța euclidiană, lucrăm în mod implicit cu *presupozitia* că orice două clustere sunt liniar-separabile. Însă, evident, datele nu satisfac întotdeauna această presupozitie. Un *exemplu* clasic este acela în care avem două clustere constituite din puncte situate în două zone / benzi circulare concentrice din planul euclidian (\mathbb{R}^2).

În general, pentru algoritmii de învățare automată care determină *separatori liniari* este posibil să folosim funcții-nucleu (engl., *kernel functions*, sau, pe scurt, *kernels*) ca să obținem versiuni neliniare. Algoritmul K -means nu face excepție.

Vă reamintim că există două *aspects principale* ale problemelor rezolvate cu ajutorul algoritmilor „kernelizați“:

- i. Soluția unei astfel de probleme se exprimă ca o combinație liniară de „reprezentări“ / „imaginii“ ale instanțelor într-un așa-numit „spațiu de trăsături“;⁹⁰⁶
- ii. Algoritmul folosește doar produsele scalare dintre [vectorii care reprezintă] imagini de instanțe, nu imaginile propriu-zise ($\Phi(x_i)$).⁹⁰⁷

În cele ce urmează, vom arăta — în manieră ușor simplificată⁹⁰⁸ la punctele a , b și c — că aceste două aspecte pot fi satisfăcute în cazul algoritmului K -means.

- a. Fie γ_{ij} o *variabilă-indicator*, care este egală cu 1 dacă instanța x_i este [în prezent] asignată la clusterul j și 0 în caz contrar. Arătați că pentru orice j , centroidul clusterului j (notat cu μ_j , unde $j \in \{1, \dots, K\}$) care se actualizează în corpul iterativ al algoritmului K -means poate fi calculat folosind o formulă de genul $\mu_j = \sum_{i=1}^n \alpha_{ij} x_i$, unde coeficienții α_{ij} — pe care îi veți determina — pot fi calculați în funcție de [toate] variabilele-indicator γ .
- b. Arătați că pentru două puncte oarecare x și x' din \mathbb{R}^d , expresia $\|x - x'\|^2$ poate fi calculată folosind doar (combinații liniare de) produse scalare de elemente (în speță x și x') din \mathbb{R}^d .
- c. Tinând cont de rezultatele de la punctele a și b , arătați că expresiile $\|x_i - \mu_j\|^2$, care sunt utilizate în corpul iterativ al algoritmului K -means, se pot calcula folosind doar (combinații liniare de) produse scalare dintre instanțele x_1, \dots, x_n .
- d. Considerând dată o funcție de „mapare“ $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}^m$ și funcția-nucleu corespunzătoare $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$, cu $K(x, y) = \Phi(x) \cdot \Phi(y)$, scrieți pseudo-codul algoritmului K -means kernel-izat.⁹⁰⁹ Concret, veți porni cu anumite puncte inițiale ca centroizi. Apoi, în mod iterativ veți folosi răspunsul de la punctul

⁹⁰⁶Formal, date fiind instanțele $x_i \in \mathbb{R}^d$, cu $i = 1, \dots, n$ și $d \in \mathbb{N}^*$, se va lucra cu o funcție de „mapare“ $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}^m$, cu $m \gg d$, operatorul relațional ' \gg ' având semnificația „mult mai mare decât“. Imaginea lui x_i prin funcția Φ este $\Phi(x_i)$. Spațiul de trăsături este \mathbb{R}^m .

⁹⁰⁷Această *restricție* se justifică prin faptul că se consideră doar funcții de mapare Φ pentru care dacă definim $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ prin $K(x, y) = \Phi(x) \cdot \Phi(y)$, atunci K este calculabilă în mod eficient în \mathbb{R}^d . K se numește funcție-nucleu. *Observație*: nu există nicio legătură între această funcție K și numărul K care este apără în algoritmul de clusterizare K -means.

⁹⁰⁸Și anume, lucrând în \mathbb{R}^d în loc de în \mathbb{R}^m .

⁹⁰⁹Această variantă a algoritmului K -means caută separatori liniari între clusterele ce se vor forma cu $\Phi(x_1), \dots, \Phi(x_n) \in \mathbb{R}^m$, însă folosește efectiv doar instanțele $x_i \in \mathbb{R}^d$ (prin intermediul funcției-nucleu K) nu și imaginile lor prin funcția de mapare (Φ) în \mathbb{R}^m .

c ca să găsiți cel mai apropiat centroid pentru fiecare instanță și veți seta valorile variabilelor-indicator γ_{ij} , după care veți face uz de răspunsul de la punctul *a* pentru a recalcula pozițiile centroizilor.

52. (Comparație între clusterizarea ierarhică și algoritmul *K-means*)
 • ○ * CMU, 2012 fall, E. Xing, A. Singh, HW3, pr. 1.2.a

Pornind de la o dendrogramă (arbore de clusterizare ierarhică) oarecare, putem defini o partitioare a datelor în *K* clustere „tăind“ ramurile arborelui în dreptul unor anumite nivele, sub rădăcina arborelui.

De exemplu, pentru $K = 2$ putem defini două clustere pornind de la cei doi subarbوري ai rădăcinii dendrogramei. Pentru $K = 4$, putem folosi subarbوري nodurilor descendente din nodul-rădăcină, și aşa mai departe. (Observați că dacă valoarea lui *K* nu este putere a lui 2, atunci ar trebui să definim un anumit criteriu pentru a stabili care este prioritatea în alegerea subarbutorilor.)

Folosind această procedură pentru a forma o partitie (a setului de date de intrare) din rezultatul unui clustering de tip ierarhizat, care dintre cele trei măsuri de similaritate prezentate la curs — “single-linkage”, “complete-linkage” și “average-linkage” — va conduce cel mai probabil la formarea unor clustere foarte asemănătoare cu cele obținute de către algoritmul *K-means*? (Presupunem că valoarea lui *K* este o putere a lui 2).

Vă readucem aminte că măsurile de similaritate “single-linkage”, “complete-linkage” și “average-linkage” dintre două mulțimi *X* și *Y* dintr-un spațiu dotat cu o măsură de distanță sunt definite în raport cu minimul, maximul și respectiv media distanțelor dintre perechile de puncte (x, y) cu $x \in X$ și $y \in Y$.

53. (Algoritmul *K-means*: un caz special; comparație cu algoritmi de clasificare)
 ○ * CMU, 2003 fall, T. Mitchell, A. Moore, HW7, pr. 1.b

Pentru această problemă folosim algoritmul *K-means*, unde *K* este egal cu numărul de instanțe / puncte de clusterizat, iar fiecare cluster este format dintr-un singur punct. Considerând că o anumită măsură de distanță, vom putea clasifica o instanță nouă (de test) asociind-o cu clusterul al cărui centroid se află cel mai aproape de instanța respectivă. Atenție: centroizii clusterelor nu se modifică în faza de test.

Cu ce metodă de clasificare automată este echivalent modelul descris aici?

54. (Algoritmul *K-means*: Adevărat sau Fals?)
 • ○ CMU, 2021 fall, M. Gormley, H. Chai,
Exam 3 practice problems, pr. 6.1

- a. Dat fiind un anumit set de date și o valoare fixată pentru *K*, algoritmul *K-means* va produce întotdeauna același rezultat, dacă
 – păstrăm aceleași inițializări pentru centroizi;

- pentru acest set de date și această valoare fixată pentru K nu apar niciodată situații de „indecizie“ (engl., tie) la asignarea instanțelor la centroizi. (Adică, la pasul de reasignare a instanțelor la centroizi, nicio instanță x_i nu se află la distanță minimală (și egală) față de doi centroizi diferiți, c_j și $c_{j'}$.)
- b. Algoritmul K -means converge pe orice set de date la optimul global al criteriului J (adică, echivalent, la optimul global al coeziunii intra-clustere). Veți considera valoarea lui K ca fiind oarecare, dar fixată.
- c. Este posibil ca la două iterații consecutive ale algoritmului K -means, asignările instanțelor la clustere să nu se schimbe deloc.
- d. Algoritmul K -means nu este deloc sensibil la prezența outlier-elor în setul de date de clusterizat.
- e. K de la algoritmul K -Means și k de la algoritmul k -NN au aceeași semnificație.
- f. Care credeți că este deosebirea esențială (sau, cea mai importantă) dintre algoritmii K -Means și k -NN?

55.

(Algoritmul K -means: implementare)

□ Liviu Ciortuz, 2016

Implementați — de preferință în C/C++ — algoritmul K -means. Ca și la problema 38, veți considera că instanțele de clusterizat sunt puncte într-un spațiu euclidian \mathbb{R}^d și că ele sunt înregistrate într-un fișier în format CSV (Comma Separated Value). Numele acestui fișier și eventual d vor fi indicate ca argumente în linia de comandă a programului.

Testați implementarea pe datele de la problema 40. Ulterior, puteți extinde implementarea cu funcții de reprezentare grafică, ca la rezolvarea problemeelor 7, 8, 9, 10 și 41. O altă extensie utilă este calculul valorii criteriului J la fiecare iterație, aşa cum s-a arătat la problemele 12 și 45.

Încă o direcție posibilă pentru extinderea acestei implementări se referă la varianta kernel-izată a algoritmului K -means, al cărei obiectiv l-a constituit problema 51. Implementați și apoi testați această variantă pe setul de date corespunzător conceptului XOR, folosind funcția-nucleu polinomială de ordinul al doilea

$$K(x, x') = (x \cdot x' + 1)^2 \quad \forall x, x' \in \mathbb{R}^2.$$

7.2.3 Algoritmul EM pentru modele de mixturi gaussiene

56.

(EM/GMM, cazul unidimensional:
aplicarea manuală a unei iterări,
pentru o mixtură de tipul următor: μ_1, μ_2 liberi,
 $\pi_1 = \pi_2, \sigma_1 = \sigma_2 = 1/\sqrt{2}$)

*prelucrare de Liviu Ciortuz, 2021, după
□ • CMU, 2015 spring, Roni Rosenfeld,
“Numerical example of one EM iteration over a Mixture of Gaussians”*

Acesta este un mini-exemplu numeric pentru executarea unei singure iterări a algoritmului EM aplicat la problema estimării mediilor a două distribuții gaussiene care formează o mixtură.

Deviațiile standard ale celor două gaussiene sunt $\sigma_1 = \sigma_2 = 1/\sqrt{2}$.

Probabilitățile a priori corespunzătoare acestor gaussiene sunt $\pi_1 = \pi_2 = 0.5$.

Vom nota cu j indexul ale cărui valori identifică aceste gaussiene, cu i indexul pentru instanțe, și cu t indexul pentru iterările algoritmului EM.

Vom inițializa mediile celor două gaussiene cu niște valori convenabile (pentru ușurința calculelor): $\mu_1^{(0)} = 3, \mu_2^{(0)} = 6$.

Vom considera că variabila z_{ij} ia valoarea 1 dacă instanța x_i a fost generată de către gaussiana j , și 0 în cazul contrar. Variabilele z_{ij} sunt latente / „ascunse“ / „neobservabile“.

*Remember:*⁹¹⁰

Verosimilitatea unei instanțe „observabile“ se exprimă astfel:

$$L(x_i|\mu^{(t)}) = \sum_j \pi_j \mathcal{N}(x_i|\mu_j^{(t)}, \sigma_j^2).$$

Pasul E este:

$$E[z_{ij}|\mu^{(t)}] = \frac{\pi_j \mathcal{N}(x_i|\mu_j^{(t)}, \sigma_j^2)}{\sum_{j'} \pi_{j'} \mathcal{N}(x_i|\mu_{j'}^{(t)}, \sigma_{j'}^2)}.$$

Pasul M este:

$$\mu_j^{(t+1)} = \frac{\sum_i E[z_{ij}|\mu_j^{(t)}] \cdot x_i}{\sum_i E[z_{ij}|\mu^{(t)}]}.$$

a. Completăți următorul tabel și apoi calculați valorile celor două medii ($\mu = (\mu_1, \mu_2)$) care maximizează verosimilitatea datelor furnizate, adică:

$$\mu_{MLE} = \underset{\mu_1, \mu_2}{\operatorname{argmax}} \prod_i L(x_i|\mu^{(t)}).$$

⁹¹⁰Modul în care se deduc regulile de actualizare corespunzătoare celor doi pași din ciclul repetitiv al algoritmului EM specific acestui caz este prezentat în multe cărți de specialitate (de exemplu, în cartea *Machine Learning* a profesorului Tom Mitchell, la secțiunea 6.12.1). În prezenta culegere de exerciții, demonstrația aceasta constituie obiectul problemei 15.

i	1	2	3
x_i	2	4	7
$\mathcal{N}(x_i \mu_1^{(0)}, \sigma_1^2)$			
$\mathcal{N}(x_i \mu_2^{(0)}, \sigma_2^2)$			
$L(x_i \mu_1^{(0)}, \mu_2^{(0)})$			
$E[z_{i1} x_i, \mu_1^{(0)}, \mu_2^{(0)})$			
$E[z_{i2} x_i, \mu_1^{(0)}, \mu_2^{(0)})$			

$$\mu_1^{(1)} = \dots \approx \dots$$

$$\mu_2^{(1)} = \dots \approx \dots$$

b. Care este valoarea verosimilității $L(x_1, x_2, x_3|\mu_1^{(1)}, \mu_2^{(1)})$? Comparați-o cu verosimilitatea $L(x_1, x_2, x_3|\mu_1^{(0)}, \mu_2^{(0)})$. Rezultatul pe care l-ați obținut este în acord cu proprietatea teoretică (referitoare la convergența algoritmului EM) care a fost prezentată curs?

57. (EM/GMM, cazul unidimensional, varianta generală: executarea manuală a pasului M)

□ • N. de Freitas, preluat de Kevin Murphy, în ex. 11.7 din "Machine Learning, A Probabilistic Perspective", MIT Press, 2012

În acest exercițiu urmărim să clusterizăm setul de date $x = \{1, 10, 20\}$ din \mathbb{R} , folosind algoritmul EM pentru o mixtură de două distribuții gaussiene (EM/GMM). Presupunem că rezultatul execuției pasului E este cel dat în următoarea matrice:

$$\begin{bmatrix} 1 & 0 \\ 0.4 & 0.6 \\ 0 & 1 \end{bmatrix},$$

în care elementul generic $p_{i,j}$ reprezintă probabilitatea ca „observația“ x_i să fi fost generată de către gaussiana j .

Vi se cere să executați pasul M al algoritmului EM. Veți scrie mai întâi formulele pentru actualizarea parametrilor mixturii formate din cele două gaussiene

(vedeți rezolvarea problemei 18), fără să trebuiască să faceți demonstrația lor; trebuie doar să aplicați aceste formule pe setul de date indicat mai sus. Elaborați toate detaliile necesare.

- Scrieți expresia așa-numitei *funcții auxiliare* (Q) care se optimizează la acest pas. Vă readucem aminte că funcția auxiliară este media funcției de log-verosimilitate a *datelor complete* (adică, cele *observabile* (x_i) și cele *neobservabile* (z_{ij})).
- Care vor fi noile valori ale probabilităților de selecție π_1 și π_2 după execuția pasului M? Dar ale mediilor μ_1 și μ_2 ? Dar ale varianțelor σ_1^2 și σ_2^2 ?

58. (Algoritmul EM/GMM, cazul unidimensional, varianta generală: aplicare pe un set de date din \mathbb{R})

• MIT, 2016 spring, T. Jaakkola, R. Barzilay, HW5, pr. 5

Fie mixtura de două distribuții gaussiene $P(x|\theta) = p_1\mathcal{N}(x|\mu_1, \sigma_1^2) + p_2\mathcal{N}(x|\mu_2, \sigma_2^2)$, care are parametrii $\theta = (p_1, p_2, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$, adică mediile, varianțele și ponderile (probabilitățile de selecție) pentru fiecare dintre cele două gaussiene. Valorile inițiale $\theta^{(0)}$ asignate parametrilor mixturii sunt: $p_1 = p_2 = 0.5$, $\mu_1 = 6$, $\mu_2 = 7$, $\sigma_1 = 1$ și $\sigma_2 = 2$.

Avem următoarele instanțe ale lui x : $x_1 = 0$, $x_2 = 1$, $x_3 = 5$, $x_4 = 6$ și $x_5 = 7$.

- Scrieți expresia funcției de log-verosimilitate, $\ell(x|\theta)$, pe care urmărim să o maximizăm. De ce este convenabil să folosim algoritmul EM [comparativ, de exemplu, cu metoda gradientului]? Furnizează oare *întotdeauna* acest algoritm soluția optimă?
- Referindu-ne la prima execuție a pasului E, care sunt instanțele x_i care vor fi cel mai probabil (dar nu în totalitate) asignate celei de-a două gaussiene? Altfel spus, care sunt punctele x_i pentru care $P(z_i = 2|x_i, \theta^{(0)}) > P(z_i = 1|x_i, \theta^{(0)})$?
- Referindu-ne acum la prima execuție a pasului M, în ce direcție se vor deplasa mediile celor două gaussiene?
- Tot în legătură cu prima execuție a pasului M, varianțele σ_1^2 și σ_2^2 vor crește sau vor descrește?
- Folosind o implementare a algoritmului EM/GMM specifică pentru cazul de față, faceți o rulare pe datele acestei probleme. Cât sunt valorile mediilor celor două gaussiene atunci când algoritmul ajunge la convergență? Care dintre varianțele celor două gaussiene obținute la final este mai mare?

59. (Algoritmul EM/GMM, cazul unidimensional, varianta generală: întrebări calitative)

• MIT, 2006 fall, Tommi Jaakkola, final exam, pr. 3

Estimăm parametrii unei mixturi de două distribuții gaussiene unidimensionale, aplicând versiunea generală pentru algoritmul EM/GMM (vedeți pr. 18) pe setul de date $S = \{0, 1, 5, 6, 7\}$. Funcția de densitate de probabilitate (engl.,

probability density function, p.d.f.) pentru această mixtură de distribuții este definită (pentru $\forall x \in \mathbb{R}$) astfel:

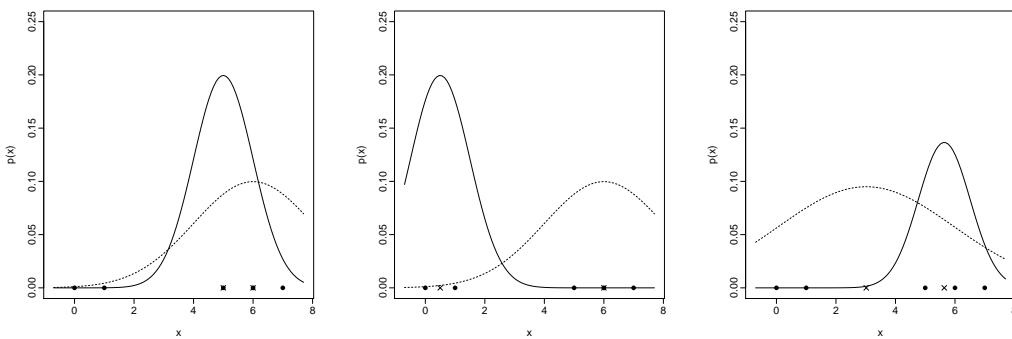
$$P(x|z, \theta) = P(z=1) \cdot \mathcal{N}(x|\mu_1, \sigma_1^2) + P(z=2) \cdot \mathcal{N}(x|\mu_2, \sigma_2^2) \text{ cu } \theta \stackrel{\text{not.}}{=} (\mu_1, \mu_2, \sigma_1, \sigma_2).$$

Orice student de la acest curs ar trebui să știe să rezolve o astfel de problemă de estimare de parametri. Ei bine, un student (dintre cei mai neatenți / zăpăciți), a încurcat ordinea figurilor care ilustrează grafic „actualizările“ (engl., updates) făcute de către algoritmul EM/GMM pe dataset-ul S . Mai mult, este posibil ca el să fi strecut din greșală o figură care nu aparține acestui „experiment“.

Comentariu:

În cele ce urmează, vă vom cere să indicați care dintre figurile de mai jos reprezintă iterații / „actualizări“ successive executate de algoritmul EM/GMM pe dataset-ul S . Veți explica de ce anume ordinea pe care ați indicat-o dumneavoastră are sens din punctul de vedere al modului cum lucrează algoritmul EM/GMM.

În toate aceste figuri se reprezintă grafic cu o linie continuă produsul $P(z=1) \cdot \mathcal{N}(x|\mu_1, \sigma_1^2)$ văzut ca funcție de x , iar cu o linie punctată produsul $P(z=2) \cdot \mathcal{N}(x|\mu_2, \sigma_2^2)$.



a.

i. Clasificați cu 'Adevărat' sau 'Fals' afirmația următoare:

În acest model de mixturuă, putem identifica asignarea a posteriori cea mai probabilă (engl., *the most likely posterior assignment*), adică acel j pentru care se maximizează probabilitatea $P(j|x)$, comparând valorile produselor $P(z=1) \cdot \mathcal{N}(x|\mu_1, \sigma_1^2)$ și $P(z=2) \cdot \mathcal{N}(x|\mu_2, \sigma_2^2)$.

ii. Justificați apoi în mod riguros răspunsul dumneavoastră.

b. Asignați două dintre figurile prezentate mai sus la iterații succesive (și corecte) ale algoritmului EM/GMM. Justificați succint alegerea pe care ați făcut-o.

c. Explicați [în mod succint] cum anume modelul de mixturuă pe care l-ați ales pentru 'iterația 2' decurge din modelul de mixturuă pe care l-ați ales pentru 'iterația 1'.

60.

(EM/GMM, cazul unidimensional:
efectuarea manuală a unei iterări,
pentru o mixtură de tipul $\mu_1 = \mu_2$ (liber),
 $\pi_1 = \pi_2, \sigma_1 = 1, \sigma_2 = 2$)

• U. Toronto, Radford Neal,
“Statistical Methods for Machine Learning and Data Mining” course,
2014 spring, Practice problems, set 2, pr. 4

Considerăm o mixtură de două distribuții gaussiene unidimensionale, pentru care funcția de densitate de probabilitate (p.d.f.) are pentru o „observație“ oarecare x expresia

$$\frac{1}{2}\mathcal{N}(x|\mu, 1) + \frac{1}{2}\mathcal{N}(x|\mu, 2^2).$$

În această formulă, $\mathcal{N}(x|\mu, \sigma^2)$ desemnează, pentru x , valoarea densității distribuției normale unidimensionale de medie μ și varianță σ^2 . Remarcați faptul că în acest model de mixtură probabilitățile de mixare / selecție sunt egale, apoi că mediile celor două componente sunt egale și, de asemenea, că deviațiile standard ale celor două componente au valorile fixate 1 și respectiv 2. Modelul acesta de mixtură are un singur parametru, μ .

Presupunem că dorim să estimăm valoarea parametrului μ prin metoda maximizării verosimilității, folosind algoritmul EM. Răspundeți la următoarele întrebări privitoare la modul în care operează pașii E și M ai acestui algoritm, atunci când considerăm cele trei instanțe / „observații“ de mai jos:

4.0, 4.6, 2.0.

Vă punem la dispoziție un tabel cu anumite valori ale funcției de densitate pentru distribuția normală standard, de care veți avea probabil nevoie în rezolvarea acestui exercițiu:

x	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$\mathcal{N}(x 0, 1)$	0.40	0.40	0.39	0.38	0.37	0.35	0.33	0.31	0.29	0.27	0.24
x		1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
$\mathcal{N}(x 0, 1)$		0.22	0.19	0.17	0.15	0.13	0.11	0.09	0.08	0.07	0.05

a. Găsiți valoarea probabilităților condiționate calculate la pasul E, presupunând că la execuția precedentului pas M estimarea obținută pentru parametrul modelului a fost $\mu = 4$ (iar $\sigma_1 = 1$ și $\sigma_2 = 2$, totdeauna). Remarcați faptul că, întrucât probabilitățile condiționate pentru cele două componente ale mixturii trebuie să se sumeze la valoarea 1, este suficient să se calculeze $p_{i1} \stackrel{\text{not.}}{=} P(\text{componenta}_1|x_i)$ pentru $i = 1, 2, 3$.

Atenție! Nu este suficient să faceți schimbarea de variabilă necesară pentru „standardizare“, și anume $x \leftarrow \frac{x - \mu}{\sigma}$. Trebuie să folosiți formula de legătură între o distribuție gaussiană unidimensională oarecare și cea standard:

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sigma} \mathcal{N}\left(\frac{x - \mu}{\sigma} | 0, 1\right).$$

În prealabil veți demonstra această formulă.

b. Folosind probabilitățile pe care le-ați calculat la punctul a , găsiți estimarea pentru parametrul μ care va fi obținută la următoarea execuție a pasului M. Vă reamintim că pasul M maximizează media (engl., expected value) log-probabilității comune pentru instanțele x_1, x_2 și x_3 și variabilele-indicator [pentru componentele mixturii], care sunt latente. Media aceasta se calculează în raport cu distribuția de probabilitate condiționată a variabilelor-indicator [pentru componentele mixturii], care a fost determinată / calculată la pasul E al aceleiași iterății.

Sugestie: Nu este necesar ca în prealabil să faceți elaborarea formulelor algoritmului EM pentru acest caz specific de mixtură de distribuții gaussiene unidimensionale. Este suficient să lucrați cu funcția „auxiliară“ Q corespunzătoare iterăției respective.

c. Calculați log-verosimilitatea datelor „observabile“ atunci când $\mu = 4$. Cum este valoarea acestei expresii (mai mică, egală, sau mai mare?) față de valoarea log-verosimilității acelorași date „observabile“ atunci când μ ia valoarea calculată la finalul pasului M descris la punctul b?

61.

(EM/GMM, cazul unidimensional, cu parametrii μ_1 și π_1 liberi, $\mu_2 = 3/2\mu_1$, $\pi_2 = 1 - \pi_1$, $\sigma_1 = \sigma_2 = 1$; elaborare + implementare + rulare)

• ○ U. Toronto, Radford Neal,
"Statistical Computation" course,
2003 spring, HW3, pr. 2

Să presupunem că un ornitolog⁹¹¹ face un studiu referitor la când anume păsările femele se întorc la cuiburile lor după ce fac „expediții“ / zboruri de strângere de hrana pentru pușorii lor. Pentru toate păsările femele studiate, ornitologul cunoaște momentele de timp când puii lor ies din ouă, iar toate măsurările de timp făcute de către ornitolog sunt relative la aceste momente.

Se crede că, după ce au scos puii din ouă, păsările se întorc din „expedițiile“ / zborurile de strângere a hranei la intervale [aproximativ] regulate de timp, θ , 2θ , 3θ , etc., însă cu o anumită variație aleatoare, despre care vom presupune că urmărează o distribuție normală / gaussiană de medie 0 și varianță 1. Se presupune că păsările acționează în mod independent unele de altele.

Pentru a evita să deranjeze păsările la momente critice, ornitologul nu a „observat“ / notat timpul de întoarcere din prima „expediție“, pentru nicio pasăre. În schimb, pentru fiecare pasăre, ornitologul a „observat“ / notat timpul de întoarcere fie de la a doua fie de la a treia „expediție“. (Păsările sunt foarte dificil de zărit, aşa că anumite întoarceri nu au fost observate.) Momentele de timp când au avut loc aceste întoarceri constituie datele noastre, X_1, \dots, X_n . Din nefericire, ornitologul nu cunoaște, în legătură cu fiecare instanță X_i , dacă ea reprezintă timpul de întoarcere corespunzător celei de-a doua „expediții“ ori a celei de-a treia. Probabilitatea de a observa o a doua întoarcere versus o a treia nu este cunoscută.

Datele pot fi modelate cu ajutorul unei mixturi, ale cărei componente sunt distribuțiile $\mathcal{N}(2\theta, 1)$ și $\mathcal{N}(3\theta, 1)$, cu probabilitățile de mixare / selecție p și

⁹¹¹Ornitologia este o ramură a zoologiei, care studiază păsările.

respectiv $1-p$. Cu alte cuvinte, expresia funcției de densitate de probabilitate pentru o „observație“ x este următoarea:

$$f(x) = p \cdot \frac{1}{\sqrt{2\pi}} \cdot \exp(-(x - 2\theta)^2/2) + (1-p) \cdot \frac{1}{\sqrt{2\pi}} \cdot \exp(-(x - 3\theta)^2/2).$$

Considerând datele X_1, \dots, X_n , va trebui să găsiți estimările de verosimilitate maximă pentru θ și p , folosind un algoritm EM pe care îl veți deduce dumneavoastră. Variabilele neobservabile ar trebui să fie variabilele-indicator relativ la care întoarcere — a doua ori a treia — se referă fiecare „observație“ / instanță X_i .

a. Deducreți formulele de actualizare care se folosesc la pașii E și M ai algoritmului EM corespunzător acestei probleme.

b. Elaborați o implementare a algoritmului EM pe care l-ați dezvoltat și testați-l pe datele pe care le furnizăm pe site-ul web asociat acestei cărți,⁹¹² precum și pe alte date pe care le considerați potrivite. Programul dumneavoastră ar trebui să execute un număr de iterații pe care-l veți specifica (nu trebuie să detectați în mod automat când anume se realizează convergența). De asemenea, ar trebui ca programul să aibă o opțiune de depanare (“debug”) la selectarea căreia să se imprime, la fiecare iterație, valorile parametrilor și ale log-verosimilității. (Aduceți-vă aminte că log-verosimilitatea datelor observabile nu trebuie să descrească de la o iterație la alta.) Programul trebuie să primească la intrare datele „observabile“, precum și valorile inițiale ale parametrilor.

Vă recomandăm să evaluați cât de bine s-a comportat algoritmul dumneavoastră. În particular, precizați dacă [au apărut indicii că] există maxime locale multiple pe suprafața funcției de verosimilitate.

62. (EM/GMM, cazul unidimensional: estimarea parametrilor μ și π pentru o mixtură de distribuții gaussiene unidimensionale, presupunând $\sigma_1^2 = \sigma_2^2$)

*prelucrare de Liviu Ciortuz, după
* CMU, 2010 fall, Aarti Singh, HW4, pr. 2.1-2*

În această problemă, vom rezolva problema mixturii a două distribuții gaussiene pentru o variantă mai generală decât cea care este prezentată în cartea *Machine Learning* a lui Tom Mitchell, pag. 193 (și preluată de noi în problema 15), unde se consideră că $\pi_1 = \pi_2 = 1/2$ și $\sigma_1^2 = \sigma_2^2$. Aici parametrii π_1 și π_2 sunt liberi, deci vom estima atât mediile μ cât și probabilitățile de selecție π , pe când acolo se estimau doar mediile μ . Vom vedea și de data aceasta — ca și în problema 15 / cartea lui T. Mitchell — că necunoașterea valorilor varianțelor $\sigma_1^2 = \sigma_2^2$ nu va impiedica calculelor.

Vă reamintim faptul că un model probabilist de tip mixtură este de fapt o funcție densitate de probabilitate care a fost obținută prin extragerea fiecărei instanțe X conform uneia din două distribuții posibile, $P(X|Z=0)$ sau $P(X|Z=1)$. Z este o variabilă aleatoare neobservabilă / „ascunsă“ care este definită peste două clase. Pur și simplu,

⁹¹²<https://profs.info.uaic.ro/~ciortuz/ML.ex-book/implementation-exercises/UToronto.2003s.HW3.pr2.EM-for-a-particular-uni-varGMM.data+R-code+sol/data2>.

Z indică pentru fiecare instanță X care anume dintre cele două distribuții considerate a generat-o.

Vom presupune că $P(Z)$ este o distribuție de tip Bernoulli și că fiecare funcție (densitate) de probabilitate condiționată $P(X|Z)$ este o distribuție gaussiană 1-dimensională cu varianță σ^2 . Așadar, funcția (densitate) de probabilitate marginală este:

$$P(X = x) = \sum_{z \in \{1,2\}} P(X = x|Z = z) \cdot P(Z = z),$$

iar ca funcție de parametrii μ și π ea se rescrie astfel:

$$P(X = x | \mu, \pi) = \sum_{z \in \{1,2\}} \pi_z \cdot \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x - \mu_z)^2}{2\sigma^2}}.$$

Parametrii acestui model sunt $\mu \stackrel{\text{not.}}{=} (\mu_1, \mu_2)$ și $\pi \stackrel{\text{not.}}{=} (\pi_1, \pi_2)$, unde μ_z este media gaussianei care corespunde clasei z , iar $\pi_z = P(Z = z)$ este probabilitatea de a extrage / genera o instanță din clasa z . (Rețineți faptul că $\pi_1 + \pi_2 = 1$.) Dorim să folosim algoritmul EM pentru a estima acești parametri din setul de date independent generate $\{x_i\}_{i=1}^m$, unde $x_i \in \mathbb{R}$.

a. Vom nota cu p_{iz} probabilitatea ca instanța i să fie extrasă din clasa z (adică, $p_{iz} = P(Z = z|X = x_i, \mu, \pi)$). În cadrul iterației t a algoritmului EM, la pasul E se calculează probabilitățile p_{iz} pentru toate perechile de valori posibile ale indicilor i, z , folosind valorile parametrilor care au fost calculate la iterarea precedentă, și anume $\mu^{(t-1)} \stackrel{\text{not.}}{=} (\mu_1^{(t-1)}, \mu_2^{(t-1)})$ și $\pi^{(t-1)} \stackrel{\text{not.}}{=} (\pi_1^{(t-1)}, \pi_2^{(t-1)})$. Scrieți / deduceți expresia de calcul pentru probabilitățile p_{iz} în funcție de acești parametri.

b. La pasul M de la iterarea t a algoritmului EM, se tratează cantitățile p_{iz} ca fiind niște count-uri fractionale pentru valorile neobservabile z și se actualizează / estimează valorile parametrilor μ și π ca și când punctul (x_i, z) ar fi fost generat / observat de p_{iz} ori. Deduceți regula de actualizare a parametrilor $\mu^{(t)}$ și $\pi^{(t)}$ în funcție de probabilitățile p_{iz} .

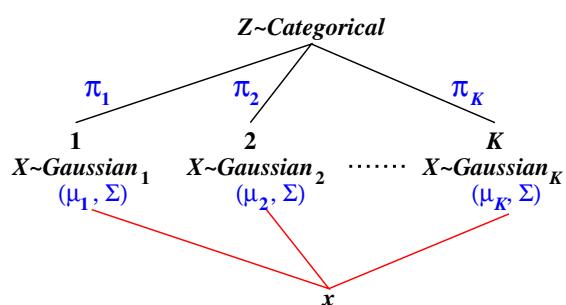
63.

(EM/GMM, cazul multidimensional:
deducerea regulii de actualizare pentru Σ ,
atunci când $\Sigma_j = \Sigma$ pt. $j = 1, \dots, K$)

• o CMU, 2010 fall, Aarti Singh, HW4, pr. 1.1-2

Un model de mixturi de distribuții gaussiane (engl., Gaussian mixture model, GMM) este o familie de distribuții ale căror densități de probabilitate (engl., probability density functions, p.d.f.) au forma următoare:

$$gmm(\mathbf{x}) = \sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \Sigma_j),$$



unde prin $\mathcal{N}(\cdot|\boldsymbol{\mu}, \Sigma)$ am notat funcția de densitate a distribuției gaussiane de medie $\boldsymbol{\mu}$ și

matrice de covarianță Σ ,⁹¹³ iar π_1, \dots, π_K sunt ponderile mixturii, care satisfac restricțiile $\sum_{j=1}^K \pi_j = 1$ și $\pi_j \geq 0$ pentru $j \in \{1, \dots, K\}$.

Pseudo-codul algoritmului EM (Expectation-Maximization) pentru învățarea [valorilor parametrilor] unui GMM pornind de la un set de instanțe $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ poate fi schițat astfel:⁹¹⁴

- Inițializează $\boldsymbol{\mu}_j$, Σ_j și π_j , pentru $j \in \{1, \dots, K\}$.
- Repetă următorii doi pași până la „convergență“:

Pasul E:

$$E[z_{ij}] = P(\mathbf{x}_i \in \text{cluster } j \mid \mathbf{x}_i, \{(\pi_{j'}, \boldsymbol{\mu}_{j'}, \Sigma_{j'})\}_{j'=1}^K),$$

Pasul M:⁹¹⁵

$$\{(\pi_j, \boldsymbol{\mu}_j, \Sigma_j)\}_{j=1}^K \leftarrow \arg \max \sum_{i=1}^n \sum_{j=1}^K E[z_{ij}] (\ln \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_j, \Sigma_j) + \ln \pi_j)$$

Considerăm o variantă simplificată de GMM, în care toate componentele mixturii au o aceeași matrice de covarianță, adică $\Sigma_j = \Sigma$ pentru $j = 1, \dots, K$. Dețineți regulile de actualizare pentru matricea Σ_j , de la pasul M. (Răspunsul dumneavoastră poate folosi valoarea lui μ_j deja actualizată la prezentul pas M.)

Indicație: Vă recomandăm să faceți apel la următoarele formule de derivare a matricelor în raport cu vectori / matrice de variabile:⁹¹⁶

$$|A|^{-1} = |A^{-1}| \quad (1e)$$

$$\frac{\partial \ln |\mathbf{X}|}{\partial \mathbf{X}} = (\mathbf{X}^{-1})^\top = (\mathbf{X}^\top)^{-1} \quad (4b)$$

$$\frac{\partial a^\top \mathbf{X} b}{\partial \mathbf{X}} = ab^\top \quad (5c)$$

unde operatorul \top desemnează transpunerea matricelor / vectorilor.⁹¹⁷

⁹¹³ $\mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^d} \sqrt{|\Sigma|}} \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$, unde d este dimensiunea spațiului (\mathbb{R}^d) în care se lucrează.

⁹¹⁴ Această procedură este ușor simplificată în raport cu formularea algoritmului EM/GMM folosită (în mod implicit) la problema 24.

⁹¹⁵ Pentru justificarea formulei date aici la Pasul M, a se vedea problema 24 de la acest capitol: termenul $-\sum_i \sum_j E[z_{ij}] \ln w_{ij} = -\sum_i \sum_j w_{ij} \ln w_{ij}$ din formula (364) de la pag. 902 a fost lăsat deoparte aici, întrucât el reprezintă o entropie (vedeți problema 1.c de la capitolul *Schema algoritmică EM* din prezenta culegere) și nu depinde de π_j , $\boldsymbol{\mu}_j$ și Σ_j .

⁹¹⁶ A se vedea *Matrix identities*, Sam Roweis, New York University, June 1999.

⁹¹⁷ Ca și la alte probleme de acest gen, vom considera că vectorii \mathbf{x} și $\boldsymbol{\mu}_j$ din \mathbb{R}^d sunt vectori-coloană.

64. (EM/GMM, cazul multidimensional; o proprietate importantă: atunci când $\Sigma_j = \sigma^2 I$, $\forall j$ și $\sigma^2 \rightarrow 0$, algoritmul EM/GMM tinde să se comporte asemenea algoritmului K-means)

■ □ • ○ CMU, 2008 fall, Eric Xing, HW4, pr. 2.2

Vă readucem aminte că, date fiind instanțele $\mathbf{x}_1, \dots, \mathbf{x}_n$, algoritmul K-means le va grupa în K clustere minimizând criteriul de „distorsiune” (sau: coeziune intra-clustere) $J_K = \sum_{i=1}^n \sum_{j=1}^K \gamma_{ij} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2$, unde $\boldsymbol{\mu}_j$ este centroidul clusterului j , iar $\gamma_{ij} = 1$ dacă \mathbf{x}_i aparține clusterului j și $\gamma_{ij} = 0$ în caz contrar. În acest exercițiu vom folosi pentru algoritmul K-means următoarea procedură:⁹¹⁸

- Inițializează cu valori arbitrarе $(\boldsymbol{\mu}_j)$ centroizii clusterelor ($j = 1, \dots, K$);
 - Itereză până se ajunge la „convergență”:
 - pentru fiecare punct \mathbf{x}_i ($i = 1, \dots, n$), revizuește asignarea sa la clustere: $\gamma_{ij} = 1$ dacă $j = \arg \min_{j'} \|\mathbf{x}_i - \boldsymbol{\mu}_{j'}\|^2$ și $\gamma_{ij} = 0$ în caz contrar;
 - actualizează pozițiile centroizilor clusterelor:
- $$\boldsymbol{\mu}_j = \frac{\sum_{i=1}^n \gamma_{ij} \mathbf{x}_i}{\sum_{i=1}^n \gamma_{ij}} \text{ pentru } j = 1, \dots, K.$$

Considerăm acum o variantă de GMM în care toate componentele mixturii au matricea de covarianță $\sigma^2 I$, unde $\sigma^2 > 0$, iar I este matricea identitate. Presupunem că ni se dă un set de instanțe $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ și că aplicăm algoritmul EM pentru a estima ponderile (π_j) și mediile ($\boldsymbol{\mu}_j$) acestei mixturi și pentru a obține pentru fiecare instanță \mathbf{x}_i probabilitățile de apartenență la fiecare cluster j .⁹¹⁹

Vom presupune că următoarele proprietăți sunt adevărate / satisfăcute pe parcursul întregii execuții a algoritmului EM:

- Există $\varepsilon > 0$ astfel încât $\pi_j \geq \varepsilon, \forall j \in \{1, \dots, K\}$ — adică ponderile mixturi sunt menținute la (o anumită) distanță față de 0 — pe parcursul tuturor iterațiilor.
- Pe parcursul tuturor iterațiilor,

$$\|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2 \neq \|\mathbf{x}_i - \boldsymbol{\mu}_{j'}\|^2 \quad \forall i \in \{1, \dots, n\}, j \neq j'.$$

Arătați că atunci când $\sigma^2 \rightarrow 0$, expresiile calculate la pasul E al algoritmului EM tind la valorile calculate de regula de actualizare (engl., update rule) pentru variabilele-indicator γ_{ij} din cadrul algoritmului K-means (adică $p(z_{ij} = 1 | x_i) \rightarrow \gamma_{ij}$), deci asignarea “soft” devine “hard”.

⁹¹⁸ Această procedură este ușor simplificată în raport cu formularea de la problema 45.

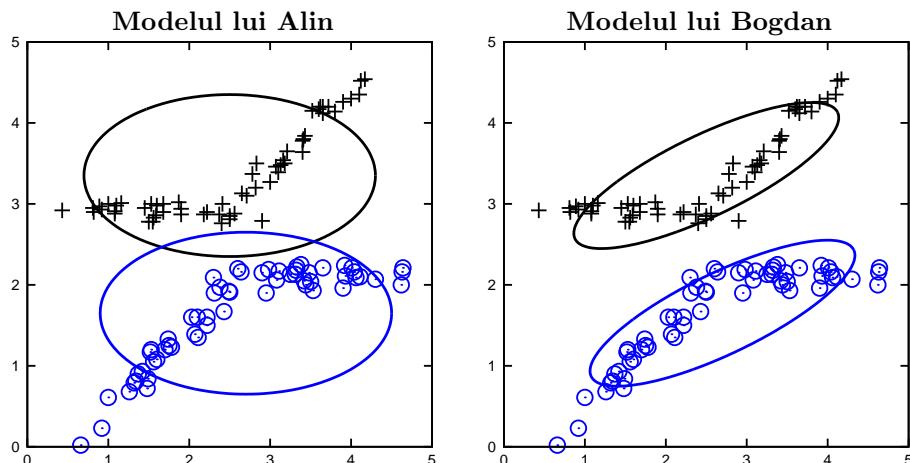
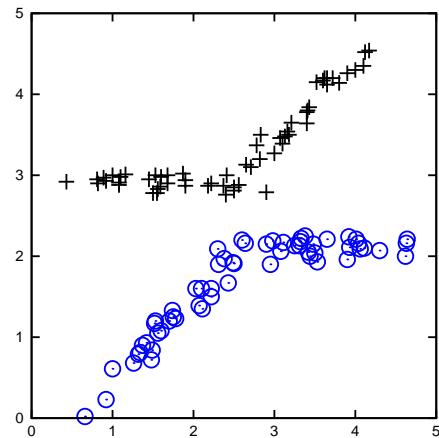
⁹¹⁹ Așadar, σ^2 este considerat fixat, iar π_j și $\boldsymbol{\mu}_j$ sunt liberi. Pentru derivarea completă a algoritmului EM/GMM corespunzător situației în care $\Sigma_j = \sigma^2 I$, $\forall j$, dar σ^2 , $\boldsymbol{\mu}_j$ și π_j sunt toți parametri variabili / liberi, a se vedea problema 20.

65.

(Algoritmul EM/GMM, cazul bidimensional; chestiuni de ordin calitativ, discutate pe date din \mathbb{R}^2)** CMU, 2006 fall, E. Xing, T. Mitchell, final exam, pr. 4*

Profesorul de la cursul de învățare automată a rugat pe trei dintre studenții săi (Alin, Bogdan și Cezar) să folosească GMM (Gaussian Mixture Models) pe setul de date ilustrat în figura alăturată. Exemplele etichetate cu $+$ sunt pozitive, iar cele etichetate cu \circ sunt negative.⁹²⁰

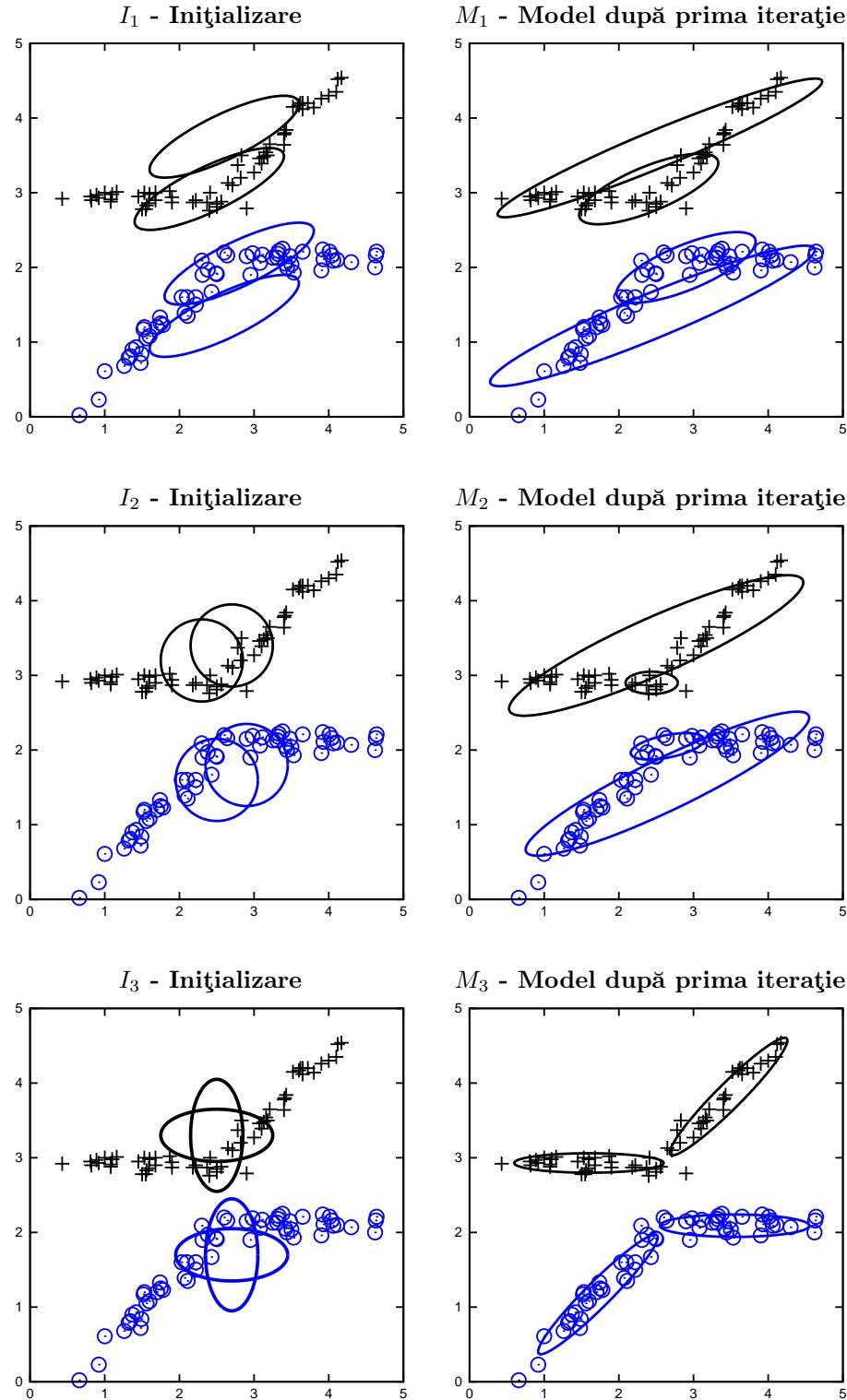
a. Alin și Bogdan au decis să folosească o distribuție gaussiană pentru exemplele pozitive și una pentru exemplele negative, obținând modelele următoare:



Dintre acestea, ce model preferi pentru acest set de date? Ce anume determină diferența dintre ele?

b. Cezar a decis să utilizeze [LC: o mixtură de] două distribuții gaussiene pentru exemplele pozitive și [LC: o altă mixtură de] două distribuții gaussiene pentru cele negative. El a utilizat algoritmul EM pentru estimarea parametrilor, și de asemenea, a încercat diferite inițializări. În coloana din stânga sunt reprezentate 3 inițializări diferite, iar în coloana din dreapta sunt reprezentate cele 3 modele obținute după prima iterație. Precizați corespondența dintre modelele din cele două coloane.

⁹²⁰Coordonatele acestor instanțe vă sunt puse la dispoziție într-un fișier depus pe site-ul acestei cărți:
<http://profs.info.uaic.ro/~ciortuz/ML.ex-book/res/CMU.2006f.EX+TM.final.pr4.em.m>.



66. (Comparări privind (in)adecvarea algoritmilor de clusterizare K -means și EM/GMM pe diverse tipuri de seturi de date din \mathbb{R}^2)

• ○ CMU, 2010 fall, Ziv Bar-Joseph, HW4, pr. 2.2-4

- a. Schițați în planul euclidian un set de date pentru care o mixtură de distribuții gaussiene sferice (i.e., pentru care matricea de covarianță este matricea identitate înmulțită cu un număr pozitiv) poate să modeleze bine datele, însă algoritmul K -means nu poate.
- b. Schițați (tot în planul euclidian) un set de date pentru care o mixtură de distribuții gaussiene diagonale (i.e., pentru care matricea de covarianță poate avea valori diferite de zero doar pe prima diagonală) poate să modeleze bine datele, dar algoritmul K -means și o mixtură de distribuții gaussiene sferice nu pot.
- c. Schițați (tot în planul euclidian) un set de date pentru care o mixtură de distribuții gaussiene având matricele de covarianță nerestricționate pot să modeleze bine datele, dar algoritmul K -means și o mixtură de distribuții gaussiene diagonale nu pot.

Cerință suplimentară: În fiecare dintre cazurile a, b și c, veți multiplica desenele de câte ori este nevoie în aşa fel încât să puteți pune în evidență granițele de separare / decizie induse de către fiecare dintre algoritmii menționați în enunț.

67. (Comparări privind (in)adecvarea algoritmilor de clusterizare ierarhică folosind diverse măsuri de similaritate, K -means și EM/GMM, pe diverse tipuri de seturi de date din \mathbb{R}^2)

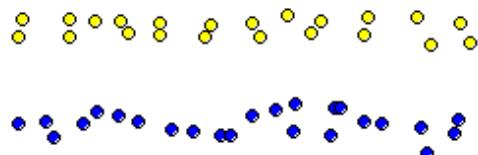
• ○ CMU, 2012 fall, T. Mitchell, Z. Bar-Joseph, HW4, pr. 2.b

Pentru fiecare dintre figurile următoare, precizați care anume dintre metodele de clusterizare listate mai jos va / vor produce rezultatele indicate, considerând $K = 2$. Alegeti de fiecare dată cea mai probabilă metodă (sau, cele mai probabile metode) și explicați pe scurt de ce ea va (respectiv, ele vor) lucra mai bine decât celelalte metode pe datele respective.

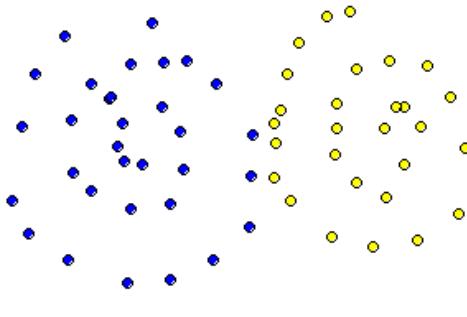
Iată lista de metode de clusterizare pe care le veți considera:

- clusterizare ierarhică cu similaritate *single-linkage*
- clusterizare ierarhică cu similaritate *complete-linkage*
- clusterizare ierarhică cu similaritate *average-linkage*
- K -means
- EM/GMM (fără a face vreo presupunere particulară în legătură cu matricea de covarianță).

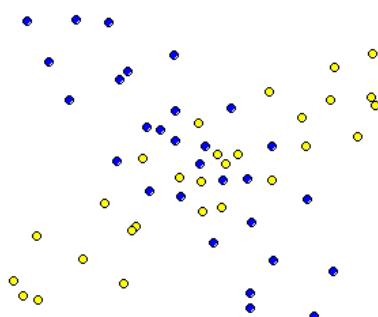
a.



b.



c.

68. (K -means vs. EM/GMM, cazul multidimensional (π, μ, Σ))

• ○ CMU, 2014 fall, W. Cohen, Z. Bar-Joseph, midterm, pr. 8

Redăm mai jos pseudo-codul algoritmului EM — de fapt, doar partea sa iterativă, conținând cei doi pași, E și M — pentru clusterizare prin modelare de mixturi de gaussiene (GMM), cazul multidimensional.⁹²¹ Îți cerem să faci schimbările minimale(!) necesare pentru a-l transforma într-un pseudo-cod corespunzător buclei principale a algoritmului de clusterizare K -means. Scrie pașii care trebuie modificați, folosind spațiul lăsat disponibil sub fiecare pas. Dacă un pas nu necesită schimbări, scrie „*Nicio modificare*“ în spațiul disponibil sub pasul respectiv. Dacă un anumit pas nu este necesar, scrie „*Elimină acest pas*“ în spațiul disponibil sub pasul respectiv.

Pasul E:

Calculează probabilitatea $w_{ij}^{(t)}$ de asignare a instanței x_i la clusterul [corespunzător gaussienei] j , ținând cont de valorile actuale ale parametrilor $\pi^{(t)}, \mu^{(t)}, \Sigma^{(t)}$, unde t identifică iterația curentă a algoritmului EM.

$$w_{ij}^{(t)} = p(z_i = j | x_i, \pi^{(t)}, \mu^{(t)}, \Sigma^{(t)}) \text{ pentru } i \in \{1, \dots, n\} \text{ și } j \in \{1, \dots, K\}$$

Pasul M:

A. Re-calculează probabilitățile a priori pentru fiecare cluster / componentă a mixturii:

$$\pi_j^{(t+1)} = \sum_{i=1}^n \frac{w_{ij}^{(t)}}{n} \text{ pentru } j \in \{1, \dots, K\}$$

B. Re-calculează [centroizii clusterelor, reprezentați de] mediile distribuțiilor gaussiene care „modelează“ clusterele:

⁹²¹Pentru versiunea completă a acestui algoritm, vedeți problema 24.

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^n w_{ij}^{(t)} x_i}{\sum_{i=1}^n w_{ij}^{(t)}} \text{ pentru } j \in \{1, \dots, K\}$$

C. Re-calculează varianțele distribuțiilor gaussiene care „modelează“ clusterele:

$$\Sigma_j^{(t+1)} = \frac{\sum_{i=1}^n \left\{ w_{ij}^{(t)} (x_i - \mu_j^{(t+1)}) (x_i - \mu_j^{(t+1)})^\top \right\}}{\sum_{i=1}^n w_{ij}^{(t)}} \text{ pentru } j \in \{1, \dots, K\}$$

69. (EM/GMM, cazul multidimensional, cazul când $\Sigma_k = \sigma_k^2 I$ și probabilitățile de selecție π_k sunt cunoscute / fixate: legătura dintre *metoda gradientului ascendent* pentru maximizarea funcției de log-verosimilitate a datelor observabile și o versiune particulară a algoritmului EM)
- • ○ CMU, 2010 fall, Aarti Singh, HW4, pr. 1.3

În acest exercițiu vom explora conexiunile dintre algoritmul EM și gradientul ascendent [ca metode de identificare a maximului funcției de verosimilitate a datelor].

Considerăm un model de mixtură de gaussiene (GMM) în care $\Sigma_k = \sigma_k^2 I$ pentru $k = 1, \dots, K$, deci „clopoțele“ corespunzătoare acestor K distribuții sunt cu „deschidere“ / secțiune circulară, însă mărimele acestor „deschideri“ sunt în general diferite. Pe lângă aceasta, vom presupune că π_k , ponderile (engl., weights) distribuțiilor din mixtură, sunt cunoscute. Expresia funcției de log-verosimilitate este următoarea:

$$l(\{\boldsymbol{\mu}_k, \sigma_k^2\}_{k=1}^K) = \ln \prod_{i=1}^n \left(\sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \sigma_k^2 I) \right) = \sum_{i=1}^n \ln \left(\sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \sigma_k^2 I) \right).$$

Formulăm mai jos un algoritm de identificare a maximului funcției de log-verosimilitate, bazat pe metoda gradientului ascendent:

- Inițializează parametrii $\boldsymbol{\mu}_k$ și σ_k^2 , pentru $k \in \{1, \dots, K\}$ cu valori alese în mod aleatoriu ($\boldsymbol{\mu}_k \in \mathbb{R}^d$ și $\sigma_k^2 \in \mathbb{R}^+$). Setează t , contorul de iteratie, la valoarea 1.
- Repetă pașii următori până când se ajunge la convergență:
 - Pentru $k = 1, \dots, K$,

$$\boldsymbol{\mu}_k^{(t+1)} \leftarrow \boldsymbol{\mu}_k^{(t)} + \eta_k^{(t)} \frac{\partial}{\partial \boldsymbol{\mu}_k} l \left(\{\boldsymbol{\mu}_k^{(t)}, (\sigma_k^2)^{(t)}\}_{k=1}^K \right),$$

unde $\eta_k^{(t)} > 0$ desemnează o rată de învățare;

- Pentru $k = 1, \dots, K$,

$$(\sigma_k^2)^{(t+1)} \leftarrow (\sigma_k^2)^{(t)} + s_k^{(t)} \frac{\partial}{\partial \sigma_k^2} l \left(\{(\mu_k^{(t+1)}, (\sigma_k^2)^{(t)})\}_{k=1}^K \right),$$

unde $s_k^{(t)} > 0$ desemnează de asemenea o rată de învățare;

- Incrementează contorul de iteratie: $t \leftarrow t + 1$.

Demonstrați că dacă alegem în mod convenabil mărimele ratelor de învățare (engl., step sizes) $\eta_k^{(t)}$ și $s_k^{(t)}$, acest algoritm bazat pe metoda gradientului ascendent este echivalent cu următorul algoritm de tip EM (care este însă ușor modificat în raport cu algoritmii de tip EM pe care i-am întâlnit până acum):

- Inițializează în mod aleatoriu parametrii $\mu_k \in \mathbb{R}$ și $\sigma_k^2 \in \mathbb{R}^+$, pentru $k \in \{1, \dots, K\}$. Setează t la valoarea 1.
- Repetă pașii următori până când se ajunge la convergență:

Pas E:

$$\tilde{z}_{ik}^{(t+0.5)} \leftarrow \text{Prob} \left(\mathbf{x}_i \in \text{cluster } k \mid \{(\mu_j^{(t)}, (\sigma_j^2)^{(t)})\}_{j=1}^K, \mathbf{x}_i \right);$$

Pas M:

$$\{\mu_k^{(t+1)}\}_{k=1}^K \leftarrow \arg \max_{\{\mu_k\}_{k=1}^K} \sum_{i=1}^n \sum_{k=1}^K \tilde{z}_{ik}^{(t+0.5)} \left(\ln \mathcal{N}(\mathbf{x}_i | \mu_k, (\sigma_k^2)^{(t)} I) + \ln \pi_k \right);$$

Pas E:

$$\tilde{z}_{ik}^{(t+1)} \leftarrow \text{Prob} \left(\mathbf{x}_i \in \text{cluster } k \mid \{(\mu_j^{(t+1)}, (\sigma_j^2)^{(t)})\}_{j=1}^K, \mathbf{x}_i \right);$$

Pas M:

$$\{(\sigma_k^2)^{(t+1)}\}_{k=1}^K \leftarrow \arg \max_{\{\sigma_k^2\}_{k=1}^K} \sum_{i=1}^n \sum_{k=1}^K \tilde{z}_{ik}^{(t+1)} \left(\ln \mathcal{N}(\mathbf{x}_i | \mu_k^{(t+1)}, \sigma_k^2 I) + \ln \pi_k \right);$$

Incrementează contorul de iteratie: $t \leftarrow t + 1$.

Ați observat, desigur, că principala modificare (în raport cu ceilalți algoritmi de tip EM) este faptul că a fost inserat un al doilea pas E între pasul M (i.e., aplicarea regulii de actualizare) pentru mediile μ_k și pasul M pentru σ_k^2 .

Sugestie: Mărimele alese pentru ratele de învățare $\eta_k^{(t)}$ și $s_k^{(t)}$ trebuie să fie puse în corespondență cu cei doi pași E.

70.

(Algoritmul EM semi-supervizat:
cazul mixturilor gaussiene multidimensionale;
concepție, implementare și aplicare)

□ • Stanford, 2020 summer, Andrew Ng, HW3, pr. 4.b-f

În această problemă ne vom referi la modelul de mixturi de distribuții gaussiene (engl., Gaussian Mixture Model, GMM), la care vom adapta algoritmul

EM semi-supervizat (vedeți problema 22 de la capitolul *Schema algoritmică EM*).

Vom considera un „scenariu“ în care datele sunt generate de către $k \in \mathbb{N}^*$ distribuții gaussiene, având atât mediile $\mu_j \in \mathbb{R}^d$ cât și matricele de covarianță $\Sigma_j \in \mathbb{S}_+^d$ necunoscute, cu $j \in \{1, \dots, k\}$. Avem n instanțe $x^{(i)} \in \mathbb{R}^d$, cu $i \in \{1, \dots, n\}$, iar fiecarei instanțe îi corespunde o variabilă latentă (sau: ascunsă, necunoscută) $z^{(i)} \in \{1, \dots, k\}$, care indică ce distribuție a generat instanța $x^{(i)}$. În mod concret, $z^{(i)} \sim \text{Categorial}(\pi)$, astfel încât $\sum_{j=1}^k \pi_j = 1$ și $\pi_j \geq 0$ pentru orice j , iar $(x^{(i)}|z^{(i)}) \sim \mathcal{N}(\mu_{z^{(i)}}, \Sigma_{z^{(i)}})$ sunt independente și identic distribuite (i.i.d.). Prin urmare, μ , Σ și π sunt parametrii modelului.

De asemenea, mai avem \tilde{n} instanțe $\tilde{x}^{(i)} \in \mathbb{R}^d$, cu $i \in \{1, \dots, \tilde{n}\}$, cărora le sunt asociate variabilele *observabile* $\tilde{z}^{(i)} \in \{1, \dots, k\}$, fiecare $\tilde{z}^{(i)}$ indicând distribuția care a generat instanța $\tilde{x}^{(i)}$. Remarcați faptul că $\tilde{z}^{(i)}$ sunt constante ale căror valori sunt cunoscute (în contrast cu $z^{(i)}$, care sunt variabile *aleatoare* având valori necunoscute). Ca și mai înainte, presupunem că $(\tilde{x}^{(i)}|\tilde{z}^{(i)}) \sim \mathcal{N}(\mu_{\tilde{z}^{(i)}}, \Sigma_{\tilde{z}^{(i)}})$ sunt i.i.d.

Ca să rezumăm, aici considerăm $n + \tilde{n}$ instanțe de antrenament, dintre care n sunt instanțe neetichetate $x^{(i)}$, cărora le sunt asociate variabilele neobservabile $z^{(i)}$, iar \tilde{n} sunt instanțe etichetate $\tilde{x}^{(i)}$, cărora le sunt asociate variabilele observabile $\tilde{z}^{(i)}$. Algoritmul EM tradițional a fost conceput să ia ca input doar n instanțe neetichetate, iar apoi să învețe parametrii modelului, μ , Σ și π . *Obiectivul* acestui exercițiu este să adaptați algoritmul EM semi-supervizat la modelul de mixturi de distribuții gaussiene (GMM), luând în considerare — pe lângă cele n instanțe neetichetate — și cele \tilde{n} instanțe etichetate, și elaborând regulile de actualizare specifice pașilor E și M din acest algoritm de învățare semi-supervizată pentru GMM.

a. [Pasul E semi-supervizat]

Indicați în mod clar care sunt variabilele latente care trebuie să fie re-estimate la pasul E. Deducreți regula de actualizare de la pasul E pentru re-estimarea tuturor variabilelor latente pe care le-ați indicat. În scrierea acestei reguli de actualizare veți putea să folosiți doar x , z , μ , Σ , π și constante universale.

b. [Pasul M semi-supervizat]

Indicați în mod clar care sunt parametrii care trebuie să fie re-estimați la pasul M. Deducreți regulile de actualizare de la pasul M pentru re-estimarea tuturor parametrilor pe care le-ați indicat. Mai precis, deducreți formulele analitice (engl., closed form expressions) pentru regulile de actualizare ale parametrilor $\mu^{(t+1)}$, $\Sigma^{(t+1)}$ și $\pi^{(t+1)}$, folosind funcția obiectiv semi-supervizată.

c. [Implementarea algoritmului EM/GMM clasic (nesupervizat)]

La acest punct veți lucra doar cu cele n instanțe neetichetate. Urmăriți instrucțiunile din fișierul `src/semi_supervised_em/gmm.py` [furnizat în directorul corespondător acestei probleme pe site-ul prezentei culegeri⁹²²] pentru a implementa algoritmul EM tradițional, iar apoi rulați-l pe setul de instanțe neetichetate, până se ajunge la convergență.

Rulați algoritmul de trei ori și folosiți funcția de reprezentare grafică furnizată pentru a construi o diagramă (engl., scatter plot) în care se pun în

⁹²²<https://profs.info.uaic.ro/ciortuz/ML.ex-book/implementation-exercises/>
Stanford.2020summer.AndrewNg.HW3.pr.4.semi-supervised-EM4GMM.DATA+CODE/

evidență asignările instanțelor la clustere. (Veți face câte un grafic pentru fiecare rulare.) Graficele trebuie să indice asignările instanțelor la clustere cu ajutorul culorilor asociate clusterelor. (Clusterul la care este asociată o instanță oarecare este acela pentru care se obține cea mai mare probabilitate [de apartenență] la pasul E al ultimei iterații.)

d. [Implementarea algoritmului EM/GMM semi-supervizat]

Acum veți folosi atât instanțele neetichetate cât și instanțele etichetate (în total, $n + \tilde{n}$ instanțe), cu câte 5 instanțe etichetate la fiecare cluster. Vă punem la dispoziție [tot pe site-ul acestei culegeri] schița un program pentru împărțirea setului de antrenament în matricele x și x_{tilde} corespunzătoare instanțelor neetichetate și respectiv instanțelor etichetate. Adăugați codul dumneavoastră la `src/semi_supervised_em/gmm.py` pentru a implementa algoritmul EM modificat, iar apoi rulați-l pe setul de instanțe de antrenament, până se ajunge la convergență. (Veți face câte un grafic pentru fiecare rulare, aşa cum ați procedat și la punctul precedent.)

e. [Comparație între EM/GMM nesupervizat și EM/GMM semi-supervizat] Descrieți succint diferențele pe care l-ați observat între algoritmul EM nesupervizat și algoritmul EM semi-supervizat, referindu-vă la următoarele chestiuni:

- i. Numărul de iterații necesare pentru a se ajunge la convergență.
- ii. Stabilitate (adică, cât de mult se schimbă asignările instanțelor în funcție de diferențele inițializări aleatoare?).
- iii. Calitatea per ansamblu a asignărilor instanțelor la clustere.

Observație: Setul de date de antrenament a fost eșantionat dintr-o mixtură de trei distribuții gaussiene cu varianță mică, precum și o a patra distribuție având varianță mare. Această informație ar trebui să vă ajute la evaluarea calității per ansamblu a asignărilor pe care le produc cei doi algoritmi.

71.

(O legătură între clasificatorul Bayes Naiv gaussian și algoritmul EM/GMM

[pentru rezolvarea unei mixturi de distribuții gaussiene multidimensionale satisfăcând presupoziția de independentă condițională de tip Bayes Naiv]; o variantă semi-supervizată a algoritmului EM/GMM)

□ • ○ CMU, 2010 spring, T. Mitchell, E. Xing, A. Singh, midterm, pr. 5.3

În acest exercițiu vom analiza relația dintre un tip particular de mixturi de distribuții gaussiene (engl., Gaussian Mixture Models, GMMs) multidimensionale și clasificatorul Bayes Naiv gaussian (engl., Gaussian Naive Bayes, GNB). Vom arăta mai întâi că ambii algoritmi folosesc în esență același model probabilist.

Pentru simplitate, în cele ce urmează vom presupune că *mixturile* cu care vom lucra sunt constituite din doar două [componente] gaussiene.⁹²³ Vom nota cu μ_0, μ_1, σ_0^2 și σ_1^2 mediile și varianțele celor două gaussiene și vom considera că

⁹²³ Veți putea constata singuri că rezultatele care vor fi obținute aici se pot extinde ușor la un număr oarecare (supra-unitar) de gaussiene.

proporțiile de mixare corespunzătoare celor două componente ale mixturii sunt π_0 și respectiv $1 - \pi_0$. De asemenea, vom folosi simbolul θ pentru a ne referi la întregul set de parametri $(\mu_0, \mu_1, \sigma_0^2, \sigma_1^2, \pi_0)$ care definesc modelul nostru de mixtură:

$$p(x) = \pi_0 \mathcal{N}(\mu_0, \sigma_0^2 I) + (1 - \pi_0) \mathcal{N}(\mu_1, \sigma_1^2 I).$$

Pe de altă parte, *clasificatorul GNB*, știm, face presupunerea că probabilitatea condiționată $p(Y|X)$ satisface relația

$$p(Y|X) = \frac{p(Y) \prod_{k=1}^d p(X^k|Y)}{p(X)},$$

iar fiecare atribut X^k , la condiționare cu [o valoare oarecare a lui] Y , este guvernăt de o distribuție gaussiană. În contextul exercițiului nostru, este natural să considerăm că variabila aleatoare Y este de tip Bernoulli, cu $P(Y = 0) = \pi_0$ și (din nou, pentru simplitate) că toate atributele au aceeași varianță, deci

$$P(X^k|Y = j) \sim \mathcal{N}(\mu_{jk}, \sigma^2).$$

Remarcați faptul că ambeii algoritmi specificați mai sus (GNB și EM/GMM) folosesc acest model probabilist „generativ“. Cu alte cuvinte, ambele modele presupun că generăm instanțe alegând mai întâi o valoare pentru Y (în funcție de parametrul π_0), iar apoi extragem / generăm un X conform distribuției gaussiane condiționate / determinate de valoarea lui Y . Singura diferență este că antrenăm clasificatorul GNB folosind instanțe etichetate pentru care valoarea lui Y este cunoscută, pe când în cazul [execuției] algoritmului EM/GMM presupunem că valorile lui Y sunt necunoscute.

- a. [Se știe că] la antrenarea clasificatorului GNB se alege setul de parametri θ care maximizează verosimilitatea datelor:

$$\underset{\theta}{\operatorname{argmax}} \prod_{i=1}^n p(x_i, y_i | \theta),$$

unde indicele i a fost folosit pentru a desemna exemplul de antrenament [LC: cu numărul de ordine] i . Vă cerem să scrieți expresia funcției obiectiv pe care urmărește să o maximizeze celălalt algoritm, EM/GMM, atunci când învață parametrii aceluiași model, însă fără a cunoaște valorile [pentru] y_i .

- b. Scrieți regulile de actualizare pentru pașii E și M din algoritmul EM standard pentru rezolvarea mixturii de gaussiane.

c. Clasificatorul GNB este antrenat folosind exemple / instanțe etichetate, în vreme ce algoritmul EM/GMM folosește instanțe neetichetate. Presupunem că avem un set de date de antrenament în care avem instanțe de ambele tipuri. Cunoaștem valorile etichetei y pentru instanțele x_1, x_2, \dots, x_m , în vreme ce instanțele x_{m+1}, \dots, x_{m+n} nu au valori cunoscute pentru y . Vă cerem să adaptați algoritmul EM/GMM pentru acest tip de date. Pentru aceasta, mai întâi veți scrie funcția obiectiv pe care o maximizează algoritmul EM/GMM modificat (pe care tocmai l-ați obținut). În expresia pe care o veți scrie, va trebui să distingeți între exemplele de antrenament pentru care y este cunoscut și cele pentru care y este necunoscut.

- d. Formulați pașii E și M corespunzători acestui nou algoritm EM/GMM.⁹²⁴

⁹²⁴Vedeți problema 22 de la capitolul *Schema algoritmică EM*.

72.

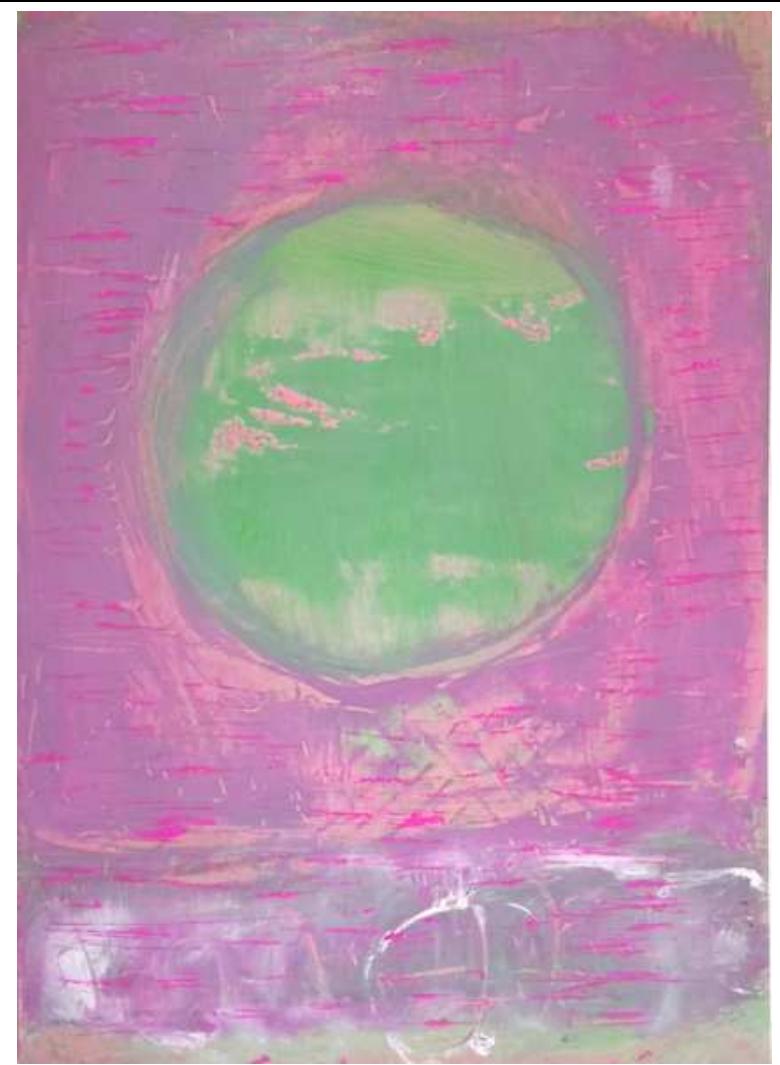
(Algoritmul EM/GMM: implementare)

 Liviu Ciortuz, 2016

Se consideră numărul natural $d \geq 1$, precum și o mulțime formată din n instanțe ($X = \{x_1, \dots, x_n\}$) din spațiul \mathbb{R}^d . Obiectivul acestei probleme este să implementați în manieră orientată pe obiecte (pentru aceasta, vă sugerăm să folosiți limbajul C++) algoritmul EM pentru a clusteriza instanțele din mulțimea X în K clustere, folosind o mixtură de distribuții gaussiene de forma $\sum_{k=1}^K \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$.

Cerințe:

1. Veți implementa câte o funcție / „metodă“ pentru fiecare componentă a algoritmului EM: pasul de inițializare, pasul E, pasul M și condiția de oprire. Adițional, veți implementa și o „metodă“ pentru calculul funcției de verosimilitate a datelor „observabile“ X în raport cu parametrii modelului probabilist asociat. În ce privește condiția de oprire, veți crea posibilitatea ca utilizatorul să poată să aleagă (printr-o opțiune la linia de comandă) între două variante: fie să se efectueze un număr (prestabilit) de iterații, fie să se testeze diferența dintre verosimilitatea datelor X la iterația t și verosimilitatea datelor X la iterația $t - 1$ în raport cu un prag (prestabilit) ε .
2. Veți trata separat — folosind supra-scrierea „metodelor“ precizate la punctul anterior — cazul distribuțiilor unidimensionale ($d = 1$) față de cazul distribuțiilor multidimensionale ($d > 1$). Pentru primul caz veți implementa ambele variante ale algoritmului EM/GMM care apar în enunțul problemei 17.bc (și veți face testarea pe datele de la problema aceea, precum și pe cele de la problema 16). Pentru a doilea caz veți implementa algoritmul care a fost dedus la problema 24 (pentru testare, folosiți datele de la problemele 21, 22, 23, 25 și 65).
3. Veți implementa — tot prin supra-scrierea „metodelor“ — încă două variante ale algoritmului EM/GMM cazul multidimensional, conform problemelor 20 și 63. (Pentru testare, puteți folosi datele de la problemele indicate la punctul precedent.)



© M. Romanică

8 Schema algoritmică EM

Sumar

Noțiuni preliminare

- distribuții probabiliste: vedeti secțiunea *Distribuții probabiliste uzuale* de la capitolul de *Fundamente*;
- estimarea parametrilor distribuțiilor probabiliste în sensul verosimilității maxime (MLE) și respectiv în sensul probabilității maxime a posteriori (MAP): vedeti secțiunea *Estimarea parametrilor unor distribuții probabiliste* de la capitolul de *Fundamente*;
- mixturi de distribuții probabiliste: vedeti ex. 29, ex. 113 și ex. 114 de la capitolul de *Fundamente*, ex. 5 și ex. 7.AB de la prezentul capitol;⁹²⁵
- metoda maximizării alternante pe coordonate (engl., *coordinate ascent*) pentru rezolvarea unor probleme de optimizare: ex. 1;⁹²⁶
- *metoda multiplicatorilor lui Lagrange* pentru rezolvarea unor probleme de optimizare cu restricții: vedeti secțiunea *Metode de optimizare în învățarea automată* de la capitolul de *Fundamente*, precum și ex. 10, ex. 12, ex. 14, ex. 27, ex. 11 și ex. 34 de la prezentul capitol.

Schema algoritmică EM

- pseudo-cod: *Machine Learning*, Tom Mitchell, 1997, pag. 194-195;
- fundamentare teoretică:
 - (P0) ex. 1: pentru funcția de log-verosimilitate a datelor observabile, $\ln P(X|\theta)$, există o margine inferioară, $F(q, \theta)$; algoritmul EM face maximizarea acestei margini inferioare aplicând *metoda [iterativă a] creșterii alternative pe coordonate* (engl., coordinate ascent);
 - (P1) ex. 2: monotonia valorilor funcției de log-verosimilitate a datelor observabile, care sunt calculate la iterării succesive ale lui EM, $\ln P(X|\theta^{(t)})$;
 - (P1') nu se garantează găsirea optimului global al funcției de log-verosimilitate a datelor observabile, $\log P(X|\theta)$, ci a unui optim local (dacă $\ln P(X|\theta^{(t)})$ este funcție continuă);
 - ex. 3: MAP EM – algoritmul EM pentru *estimarea* nu în sens MLE (cum este cazul adeseori), ci *în sens MAP*; pentru exemplificare, vedeti ex. 31.B;⁹²⁷
 - ex. 22: algoritmul EM semi-supervizat (particularizare pentru cazul mixturilor de distribuții Bernoulli: ex. 37);⁹²⁸
 - ex. 23: “hard” EM – algoritmul EM cu asignare “hard” a instanțelor la clustere;
 - ex. 24: *algoritmul EM generalizat* (engl., Generalized EM, GEM).

⁹²⁵Pentru mixturi de distribuții gaussiene vedeti secțiunea *Algoritmul EM pentru modele de mixturi gaussiene* de la capitolul de *Clusterizare*.

⁹²⁶Vedeti, de asemenea, utilizarea același metode de optimizare (eventual pentru minimizare în loc de maximizare) în cazul altor algoritmi de învățare automată: pentru algoritmul AdaBoost, ex. 22, ex. 28 și ex. 29 de la capitolul de *Arborei de decizie*; pentru algoritmul K-means, ex. 12 de la capitolul de *Clusterizare*; în sfârșit, pentru algoritmul SMO, ex. 22 de la capitolul de *Mașini cu vectori-suport*.

⁹²⁷O altă exemplificare: MAP EM pentru mixturi de distribuții [de același tip] din familia exponentială: CMU, 2015 spring, Alex Smola, HW9, pr. 1.

⁹²⁸Pentru cazul mixturilor de distribuții gaussiene, vedeti ex. 70 de la capitolul *Clusterizare*.

EM pentru modelarea de mixturi de distribuții probabiliste

- mixturi de distribuții *Bernoulli*: ex. 5, ex. 7 și ex. 6;
- mixturi de *vectori* de distribuții *Bernoulli*, cu presupunerea de independentă condițională a atributelor de intrare în raport cu atributul de ieșire (eticheta): ex. 8, ex. 9;
 - o versiune particulară — clusterizare în interiorul claselor —, cu *aplicare* la recunoașterea cifrelor scrise de mână: ex. 28;
- mixturi de distribuții *categoriale*: ex. 10 și ex. 27;
 - aplicații*: [EM pentru] *dezambiguizarea semantică* a cuvintelor dintr-un document (engl., word sense disambiguation): ex. 12 și respectiv pentru clusterizare de documente (engl., topic model): ex. 29;
- mixturi de *vectori* de distribuții *categoriale*, cu presupunerea de independentă condițională a atributelor de intrare în raport cu atributul de ieșire (eticheta): ex. 11 (algoritmul *Bayes Naiv nesupervizat*, cu *asignare “soft”* a instanțelor la clustere);
- mixturi de [vectori de] distribuții *Poisson*: ex. 33, ex. 34;
- mixturi de distribuții *Gamma*: ex. 35;
- EM pentru *estimarea probabilității de selecție* a unei componente din cadrul unei mixturi (i.e., combinație liniară) de două distribuții probabiliste oarecare: ex. 18.

EM pentru estimarea parametrilor unor distribuții probabiliste

- EM pentru estimarea parametrilor unor distribuții *binomiale*: ex. 30,⁹²⁹
- EM pentru estimarea parametrilor unor distribuții *multinomiale* (care se definesc cu ajutorul uneia sau mai multor distribuții *categoriale*): ex. 13, ex. 14, ex. 31, ex. 32.

EM pentru estimarea parametrilor unei distribuții, atunci când o parte din date lipesc

- cazul distribuției *Poisson*: ex. 17.

EM pentru estimarea parametrilor unei *sume* de două distribuții⁹³⁰

- cazul distribuțiilor *exponențiale*: ex. 15;
- cazul distribuțiilor *gaussiene*: ex. 16.

⁹²⁹În acest exercițiu, apar trei distribuții binomiale, dintre care una se definește cu ajutorul unei mixturi de două distribuții Bernoulli.

⁹³⁰Exlicație: ne referim aici la aplicarea algoritmului EM pentru estimarea parametrilor a două distribuții probabiliste atunci când se dau instanțe care sunt generate prin însumarea unor perechi de valori generate de cele două distribuții;

Alte instanțe ale schemei algoritmice EM

- algoritmul EM pentru învățare supervizată; cazul mixturilor de regresori liniari: ex. 36.

Alte probleme

- chestiuni metodologice (relativ la inițializarea parametrilor): ex. 38;
- probleme recapitulative (A/F): ex. 4, ex. 19, ex. 20, ex. 21 și ex. 39.

Analiza generală a algoritmilor EM

- ca algoritm de *învățare statistică*:
algoritmul EM poate fi văzut ca o metodă de *estimare a parametrilor* (engl., *parameter fitting*) în sensul *verosimilității maxime* (engl., *maximum likelihood estimation*, MLE);⁹³¹
- ca algoritm *per se*:
algoritm iterativ: pleacă de la o soluție (instanțiere pentru parametri) aleasă eventual în mod arbitrar / aleatoriu și o „îmbunătățește“ la fiecare iterație;
(P2) rezultatele algoritmului EM depind (ca și la K-means) de valorile atribuite parametrilor la inițializare;
(P3) anumite valori atribuite inițial parametrilor algoritmului EM pot provoca rularea la infinit a algoritmului, fără ca [la pasul M] valorile parametrilor să se modifice de la o iterare la alta: ex. 6.f;
- ca algoritm de *optimizare*:
în *esență* / *rezumat*, metoda de maximizare a funcției de *log-verosimilitate a datelor observabile* $\log P(X|\theta)$ este maximizarea la fiecare iterare t a unei funcții auxiliare Q_t , care constituie o margine inferioară a lui $\log P(X|\theta)$, și anume media funcției de *log-verosimilitate a datelor complete* în raport cu distribuția de probabilitate a *variabilelor neobservabile* la iterarea t ;
mai precis, la fiecare iterare t se calculează funcția „auxiliară“ $Q_t(\theta|\theta^{(t)})$, care reprezintă media funcției de log-verosimilitate a datelor „complete“ (cele „observabile“ plus cele „neobservabile“), unde $\theta^{(0)}$, constând din valorile inițiale ale parametrilor mixturi (θ), se alege în mod arbitrar, iar apoi $\theta^{(t+1)} = \text{argmax}_{\theta} Q_t(\theta|\theta^{(t)})$;
media reprezentată de funcția Q_t se calculează în funcție de distribuțiile condiționale ale variabilelor „neobservabile“ Z în raport cu datele observabile X și cu $\theta^{(t)}$;
- ca algoritm de *învățare automată*:
algoritmul EM este o metodă de identificare / învățare de ipoteze ML (Maximum Likelihood); vedeti capitolul / secțiunea 6.4 din cartea *Machine Learning*;
învățare în prezența unor variabile aleatoare neobservabile(!);
[urmată eventual de] „generalizare“: o instanță nouă x se asociază clusterului (i.e., distribuției) j pentru care se atinge $\max_{j'} P(X = x|h_{j'})P(h_{j'})$.

⁹³¹Acesta este cazul general. În cazuri particulare, acest principiu poate fi înlocuit cu *principiul probabilității maxime a posteriori* (engl., maximum a posteriori probability, MAP). Vedeti problema 3.

8.1 Schema algoritmică EM — Probleme rezolvate

Observație importantă: Pe tot parcursul acestui capitol, atunci când baza funcției log nu este specificată, o vom considera în mod implicit ca fiind supraunitară.⁹³²

8.1.1 Fundamente teoretice

1.

(Algoritmul EM, fundamentare teoretică:
pasul E [și pasul M])

*prelucrare de Liviu Ciortuz, după
■ CMU, 2008 fall, Eric Xing, HW4, pr. 1.1-3*

Algoritmul EM (Expectation-Maximization) este unul dintre cele mai importante procedee din învățarea automată. El permite crearea unor modele probabiliste care pe de o parte depind de un set de parametri θ iar pe de altă parte includ pe lângă variabilele obișnuite („observabile“ sau „vizibile“) x și variabile necunoscute („neobservabile“, „ascunse“ sau „latente“) z .⁹³³ În general, în astfel de situații / modele, nu se poate face în manieră analitică (adică, prin optimizare directă) o estimare a parametrilor modelului (θ), în aşa fel încât să se garanteze obținerea maximului verosimilității datelor observabile x .⁹³⁴ În schimb, algoritmul EM procedează în manieră iterativă, constituind astfel o modalitate foarte convenabilă de estimare a parametrilor θ .

Definim log-verosimilitatea datelor *complete* (observabile, x , și neobservabile, z) ca fiind $\log P(x, z | \theta)$, iar log-verosimilitatea datelor observabile ca fiind $\log P(x | \theta)$.

a. Log-verosimilitatea datelor *observabile* (x) se poate exprima în funcție de datele neobservabile (z), astfel:⁹³⁵

$$\ell(\theta) \stackrel{\text{not.}}{=} \log P(x | \theta) = \log \left(\sum_z P(x, z | \theta) \right)$$

⁹³²În secțiunea *Estimarea parametrilor unor distribuții probabiliste* de la capitolul de *Fundamente*, precum și în secțiunea *Algoritmul EM pentru modele de mixturi gaussiene* de la capitolul de *Clusterizare*, am folosit ca bază a logaritmului numărul e . Întrucât și acolo și aici ne interesează în general doar(!) să obținem maximul unor funcții de log-verosimilitate, alegerea bazei nu contează (atât timp cât este un număr supraunitar). Totuși, va trebui să observați că la punctul b și c al problemei 1, precum și la problema 2, intervin în calcule atât verosimilități cât și noțiuni legate de teoria informației (entropii, cross-entropii și etropii relative). La secțiunea *Elemente de teoria informației* de la capitolul de *Fundamente*, am folosit ca bază a logaritmului numărul 2. Așadar, în astfel de situații va trebui să avem grijă cum „armonizăm“ între ele enunțurile și relațiile.

⁹³³Un astfel de model sunt *mixturi de distribuții probabiliste*. Pentru exemple de mixturi de distribuții Bernoulli și respectiv distribuții categoriale, vedeți problemele 114, 113 și 29 de la capitolul de *Fundamente*. Rezolvarea mixturilor de distribuții gaussiene folosind algoritmul EM a constituit subiectul unei întregi secțiuni din capitolul de *Clusterizare*.

⁹³⁴Atunci când funcția de log-verosimilitate este derivabilă și toate variabilele sunt cunoscute / observabile, se poate calcula maximul acestei funcții fie în mod direct, calculând rădăcinile derivatelor parțiale, fie aplicând o metodă de aproximare, aşa cum este metoda gradientului. În cazul în care o parte din variabile sunt necunoscute / neobservabile, aflarea soluțiilor derivatelor parțiale ale funcției de log-verosimilitate necesită adeseori aplicarea unor metode de calcul numeric sofisticante. Algoritmul EM, care este foarte ușor de implementat, ne oferă o cale alternativă, foarte utilă în astfel de cazuri.

⁹³⁵Cititorul trebuie să rețină că x , vectorul de date observabile, este fixat (dat), în vreme ce z , vectorul de date neobservabile, este liber (variabil).

În continuare vom nota cu q o funcție / distribuție de probabilitate definită peste variabilele ascunse / neobservabile z .

Folosiți *inegalitatea lui Jensen*⁹³⁶ pentru a demonstra că are loc următoarea inegalitate:

$$\log P(x | \theta) \geq \sum_z q(z) \log \left(\frac{P(x, z | \theta)}{q(z)} \right)$$

pentru orice set de date observabile x (fixat), pentru orice valoare a parametrului θ și pentru orice distribuție probabilistă q definită peste variabilele neobservabile z .

Observație (1): Semnificația acestei inegalități este următoarea:
Funcția

$$F(q, \theta) \stackrel{\text{def.}}{=} \sum_z q(z) \log \left(\frac{P(x, z | \theta)}{q(z)} \right) \quad (382)$$

constituie o margine inferioară pentru funcția de log-verosimilitate a datelor observabile / incomplete, $\ell(\theta) \stackrel{\text{not.}}{=} \log P(x | \theta)$.⁹³⁷ Remarcați faptul că F este o funcție de două variabile, iar prima variabilă nu este de tip numeric (cum este θ), ci este de tip funcțional.⁹³⁸ Mai mult, se observă că expresia funcției F este de fapt (similară cu) o medie, $E_{q(z)} \left[\log \frac{P(x, z | \theta)}{q(z)} \right]$, atunci când x , q și parametrul θ se consideră fixați, iar z este lăsat să varieze.

b. Conform definiției divergenței Kullback-Leibler (numită și entropia relativă), care a fost dată la problema 63 de la capitolul de *Fundamente*, putem scrie:⁹³⁹

$$KL(q(z) || P(z | x, \theta)) = - \sum_z q(z) \log \left(\frac{P(z | x, \theta)}{q(z)} \right) \quad (383)$$

Arătați că — dacă folosim numărul 2 ca bază a logaritmului — avem:

$$\log P(x | \theta) = F(q(z), \theta) + KL(q(z) || P(z | x, \theta)).$$

Observație (2): Semnificația egalității care trebuie demonstrată la acest punct este foarte interesantă: diferența dintre funcția obiectiv $\ell(\theta) \stackrel{\text{not.}}{=} \log P(x | \theta)$ și marginea sa inferioară $F(q(z), \theta)$ — a se vedea punctul a — este $KL(q(z) || P(z | x, \theta))$. Aceasta este divergența Kullback-Leibler dintre distribuția (arbitră) considerată $q(z)$ și distribuția condițională a variabilei ascunse z în raport cu variabila observabilă x . Tocmai pe această chestiune se va „construi” punctul final, și cel mai important, al problemei noastre. Însă înainte de aceasta, este foarte util să formulăm încă o observație.

⁹³⁶Vedeți problema 79 de la capitolul de *Fundamente*, și în special punctul c al acestei probleme.

⁹³⁷În inegalitatea $\ell(\theta) \geq F(q, \theta)$, deși poate părea surprinzător, funcția ℓ are un singur argument, în vreme ce funcția F are două argumente. Diversele aspecte (foarte interesante!) ale relației dintre cele două funcții vor fi „aduse la lumină” în cele ce urmează.

⁹³⁸În continuare, pentru a aduce mereu aminte cititorului că distribuția q se referă la datele neobservabile z , vom folosi notația $q(z)$ în loc de q . În consecință, în cele ce urmează, în funcție de context, $q(z)$ va desemna fie la distribuția q , fie la valoarea acestei distribuții pentru o valoare oarecare [a variabilei neobservabile] z . (Suntem conștienți de faptul că această lejeră ambiguitate poate induce în eroare cititorul neexperimentat.)

⁹³⁹Veți observa că în relația (382) apare $P(x, z | \theta)$, iar în relația (383) apare $P(z | x, \theta)$.

Observație (3): Ideile de bază ale algoritmului EM sunt două:

1. În loc să calculeze maximul funcției de log-verosimilitate $\log P(x | \theta)$ în raport cu θ , algoritmul EM va maximiza marginea sa inferioară, $F(q(z), \theta)$, în raport cu ambele argumente, $q(z)$ și θ .
2. Pentru a căuta maximul (de fapt, un maxim local al) marginii inferioare $F(q(z), \theta)$, algoritmul EM aplică metoda *creșterii pe coordonate* (engl., coordinate ascent): după ce inițial se fixează $\theta^{(0)}$ eventual aleatoriu, se maximizează *iterativ* funcția $F(q(z), \theta)$, în mod *alternativ*: mai întâi în raport cu distribuția $q(z)$ și apoi în raport cu parametrul θ .

$$\text{Pasul E: } q^{(t)}(z) = \underset{q(z)}{\operatorname{argmax}} F(q(z), \theta^{(t)})$$

$$\text{Pasul M: } \theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}} F(q^{(t)}(z), \theta)$$

c. Fie $\theta^{(t)}$ valoarea obținută pentru parametrul / parametrii θ la iterată t a algoritmului EM. Considerând această valoare fixată, arătați că maximul lui F în raport cu argumentul / distribuția $q(z)$ este atins pentru distribuția $P(z | x, \theta^{(t)})$, iar — dacă baza logaritmului este 2 — valoarea maximului se poate exprima astfel:

$$\max_{q(z)} F(q(z), \theta^{(t)}) = E_{P(z|x, \theta^{(t)})} [\log P(x, z | \theta^{(t)})] + H(P(z | x, \theta^{(t)}))$$

Așadar, atunci când se consideră fixat cel de-al doilea argument al funcției F , valoarea maximă a acestei funcții (F) în raport cu primul său argument, $q(z)$, este suma a doi termeni: media log-verosimilității datelor complete (observabile și neobservabile) în raport cu distribuția posterioară a variabilelor neobservabile z (primul termen) și entropia acestei distribuții posterioare (al doilea termen), care nu depinde de $q(z)$.

Răspuns:

a. În contextul teoriei probabilităților, inegalitatea lui Jensen este exprimată astfel: dacă X este o variabilă aleatoare iar φ este o funcție convexă, atunci $\varphi(E[X]) \leq E[\varphi(X)]$. Dacă φ este funcție concavă, inegalitatea lui Jensen devine $\varphi(E[X]) \geq E[\varphi(X)]$.⁹⁴⁰

În cazul nostru, folosim funcția \log cu bază supraunitară, care este o funcție concavă, deci aplicând inegalitatea lui Jensen obținem: $\log(E[X]) \geq E[\log(X)]$.

Log-verosimilitatea datelor observabile este:

$$\begin{aligned} \log P(x | \theta) &= \log \left(\sum_z P(x, z | \theta) \right) = \log \left(\sum_z q(z) \frac{P(x, z | \theta)}{q(z)} \right) \\ &\stackrel{\text{def.}}{=} \log \left(E_{q(z)} \left[\frac{P(x, z | \theta)}{q(z)} \right] \right) \end{aligned}$$

De aici, conform inegalității lui Jensen — concret, înlocuind X din inegalitatea menționată cu $\frac{P(x, z | \theta)}{q(z)}$ —, rezultă:

$$\log P(x | \theta) \geq E_{q(z)} \left[\log \frac{P(x, z | \theta)}{q(z)} \right] \stackrel{\text{def.}}{=} \sum_z q(z) \log \frac{P(x, z | \theta)}{q(z)},$$

⁹⁴⁰Vedeți problema 79.c de la capitolul de *Fundamente*.

adică tocmai ceea ce trebuie să demonstrăm.

b. Pentru a demonstra egalitatea cerută, vom pleca de la definiția dată în enunț pentru funcția $F(q(z), \theta)$. Apoi, în expresia din definiție vom înlocui $P(x, z | \theta)$ cu $P(z | x, \theta) \cdot P(x | \theta)$ — conform regulii de înmulțire a probabilităților — și vom obține:

$$\begin{aligned} F(q(z), \theta) &\stackrel{\text{def.}}{=} \sum_z q(z) \log \left(\frac{P(x, z | \theta)}{q(z)} \right) \\ &= \sum_z q(z) \log \left(\frac{P(z | x, \theta) \cdot P(x | \theta)}{q(z)} \right) \\ &= \sum_z q(z) \left[\log \frac{P(z | x, \theta)}{q(z)} + \log P(x | \theta) \right] \\ &= \sum_z q(z) \log \left(\frac{P(z | x, \theta)}{q(z)} \right) + \sum_z q(z) \log P(x | \theta) \\ &= -KL(q(z) || P(z | x, \theta)) + \underbrace{\sum_z q(z)}_{=1} \log P(x | \theta) \end{aligned}$$

Rezultă că $\log P(x | \theta) = F(q(z), \theta) + KL(q(z) || P(z | x, \theta))$.

Observație (4): Conform proprietății $KL(p || q) \geq 0$ pentru $\forall p, q$, care a fost demonstrată și de noi la exercițiul 63 de la capitolul de *Fundamente*, rezultă că $KL(q(z) || P(z | x, \theta)) \geq 0$. Așadar, din egalitatea care tocmai a fost demonstrată la punctul b obținem (din nou!, după rezultatul de la punctul a) că $F(q(z), \theta)$ este o margine inferioară pentru log-verosimilitatea datelor observabile, $\ell(\theta) \stackrel{\text{not.}}{=} \log P(x | \theta)$.

c. Trebuie să maximizăm $F(q(z), \theta^{(t)})$ — marginea inferioară a log-verosimilității datelor observabile x — în raport cu distribuția $q(z)$.

Pe de o parte, rezultatul de la punctul a ne spune că $F(q(z), \theta) \leq \log P(x | \theta)$, pentru orice valoare a lui θ ; în particular, pentru $\theta^{(t)}$ avem

$$\log P(x | \theta^{(t)}) \geq F(q(z), \theta^{(t)})$$

Pe de altă parte, dacă în egalitatea demonstrată la punctul b se înlocuiește θ cu $\theta^{(t)}$, rezultă:

$$\log P(x | \theta^{(t)}) = F(q(z), \theta^{(t)}) + KL(q(z) || P(z | x, \theta^{(t)}))$$

În fine, dacă alegem $q(z) = P(z | x, \theta^{(t)})$, atunci termenul $KL(q(z) || P(z | x, \theta^{(t)}))$ din dreapta egalității de mai sus devine zero (a se vedea punctul a de la același exercițiu 63 de la capitolul de *Fundamente*). Așadar, valoarea $\max_{q(z)} F(q(z), \theta^{(t)})$ se obține pentru distribuția $q(z) = P(z | x, \theta^{(t)})$.

Acum vom calcula această valoare maximă:

$$\begin{aligned} \log P(x | \theta^{(t)}) &= \max_{q(z)} F(q(z), \theta^{(t)}) = F(P(z | x, \theta^{(t)}), \theta^{(t)}) \\ &\stackrel{\text{def. } F}{=} \sum_z P(z | x, \theta^{(t)}) \log \left(\frac{P(x, z | \theta^{(t)})}{P(z | x, \theta^{(t)})} \right) \\ &= E_{P(z | x, \theta^{(t)})} \left[\log \frac{P(x, z | \theta^{(t)})}{P(z | x, \theta^{(t)})} \right] \end{aligned}$$

$$\begin{aligned}
&= E_{P(z|x,\theta^{(t)})} [\log P(x, z|\theta^{(t)}) - \log P(z|x, \theta^{(t)})] \\
&= E_{P(z|x,\theta^{(t)})} [\log P(x, z|\theta^{(t)})] - E_{P(z|x,\theta^{(t)})} [\log P(z|x, \theta^{(t)})] \\
&= E_{P(z|x,\theta^{(t)})} [\log P(x, z|\theta^{(t)})] + H[P(z|x, \theta^{(t)})] \\
&= Q(\theta^{(t)} | \theta^{(t)}) + H[P(z | x, \theta^{(t)})],
\end{aligned}$$

unde $Q(\theta | \theta^{(t)}) \stackrel{\text{not.}}{=} E_{P(z|x,\theta^{(t)})}[\log P(x, z | \theta)]$. Așadar, am obținut rezultatul care a fost cerut în enunț.

Observație (5): [Fundamentarea pasului M]

Notând $G_t(\theta) \stackrel{\text{def.}}{=} F(P(z | x, \theta^{(t)}), \theta)$, din calculul de mai sus rezultă că $\log P(x | \theta^{(t)}) = G_t(\theta^{(t)}) = Q(\theta^{(t)} | \theta^{(t)}) + H[P(z | x, \theta^{(t)})]$. Se poate demonstra ușor — procedând similar cu calculul de mai sus — egalitatea

$$G_t(\theta) = Q(\theta | \theta^{(t)}) + H[P(z | x, \theta^{(t)})]$$

Observând că termenul $H[P(z | x, \theta^{(t)})]$ din această ultimă egalitate nu depinde de θ , rezultă imediat că

$$\operatorname{argmax}_{\theta} G_t(\theta) = \operatorname{argmax}_{\theta} Q(\theta | \theta^{(t)})$$

În consecință,

$$\theta^{(t+1)} \stackrel{\text{def.}}{=} \operatorname{argmax}_{\theta} \underbrace{F(P(z | x, \theta^{(t)}), \theta)}_{G_t(\theta)} = \operatorname{argmax}_{\theta} Q(\theta | \theta^{(t)})$$

Ultima egalitate de mai sus este responsabilă pentru următoarea reformulare (cea uzuală!) a corpului iterativ al algoritmului EM:

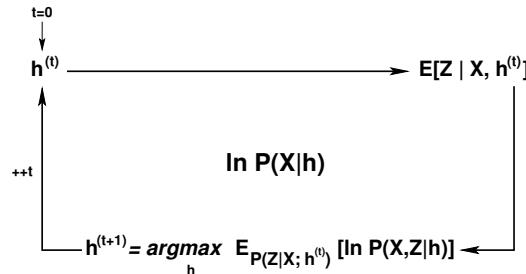
Pasul E': calculează $Q(\theta | \theta^{(t)}) = E_{P(z|x,\theta^{(t)})}[\log P(x, z | \theta)]$

Pasul M': calculează $\theta^{(t+1)} = \operatorname{argmax}_{\theta} Q(\theta | \theta^{(t)})$

Așadar, pasul E (“expectation”) al algoritmului EM va consta în calcularea funcției auxiliare $Q(\theta | \theta^{(t)})$. Pentru calculul expresiei acestei funcții (care este o medie, vedeți definiția ei de mai sus) se folosesc mediile variabilelor neobservabile, $E[z_i | x, \theta^{(t)}]$. Adeseori, în practică, la pasul E al algoritmului EM se face doar(!) calcularea mediilor acestor variabile neobservabile.⁹⁴¹ În continuare, la pasul M al aceluiași algoritm se vor calcula parametrii $\theta^{(t+1)} = \operatorname{argmax}_{\theta} Q(\theta | \theta^{(t)})$, de obicei ca rădăcini ale derivatelor parțiale ale funcției $Q(\theta | \theta^{(t)})$, atunci când aceste derive există.

Observație (6): Reprezentarea grafică următoare sintetizează modul în care funcționează algoritmul EM (sau, de fapt, mai general, schema algoritmică EM), conform rezultatelor demonstrează la acest exercițiu. În această imagine, $h^{(t)}$ corespunde lui $\theta^{(t)}$ din exercițiul nostru.

⁹⁴¹În cazul în care $E[z_i | x, \theta^{(t)}]$ reprezintă totodată un parametru al distribuției probabiliste care guvernează datele observabile — de exemplu, factorul de combinare a unor distribuții în cadrul unei mixturi, care este adeseori parametrul unei distribuții de tip Bernoulli —, calculul lui $E[z_i | x, \theta^{(t)}]$ este în sine o procedură de *estimare* a acestui parametru. Alți parametri (de exemplu μ și σ^2 pentru distribuții gaussiene) pot fi determinați în funcție de mediile acestor variabile neobservabile. Așa se explică de ce termenul “Expectation” este folosit de către unii autori în locul termenului “Expectation” în denumirea algoritmului EM.



2.

(Algoritmul EM: monotonie / convergență)

*prelucrare de Liviu Ciortuz, 2012, după
■ en.wikipedia.org/wiki/Expectation-maximization*

La problema precedentă (1) am văzut că în loc să urmărească în mod direct maximizarea funcției de log-verosimilitate a datelor observabile, adică $\ell(\theta) \stackrel{\text{def.}}{=} \log P(x | \theta)$, unde baza logaritmului (nespecificată) este considerată supraunitară, algoritmul EM procedează în mod iterativ, optimizând la pasul M al fiecărei iterații (t) o funcție „auxiliară” $Q(\theta | \theta^{(t)}) \stackrel{\text{def.}}{=} E_{P(z|x, \theta^{(t)})}[\log P(x, z | \theta)]$, reprezentând media log-verosimilității datelor complete (observabile și neobservabile) în raport cu distribuția a posteriori $P(z | x, \theta^{(t)})$.

Vom considera iterațiile $t = 0, 1, \dots$ și $\theta^{(t+1)} = \operatorname{argmax}_{\theta} Q(\theta | \theta^{(t)})$, cu $\theta^{(0)}$ ales în mod arbitrar.

În acest exercițiu veți demonstra că pentru orice t fixat (arbitrар) și pentru orice θ astfel încât $Q(\theta | \theta^{(t)}) \geq Q(\theta^{(t)} | \theta^{(t)})$ are loc inegalitatea:

$$\log P(x | \theta) - \log P(x | \theta^{(t)}) \geq Q(\theta | \theta^{(t)}) - Q(\theta^{(t)} | \theta^{(t)}) \quad (384)$$

Observație (1): Această inegalitate are o semnificație pe cât de simplă pe atât de importantă: orice îmbunătățire a valorii funcției $Q(\theta | \theta^{(t)})$ conduce la o îmbunătățire cel puțin la fel de mare a valorii funcției obiectiv, $\ell(\theta) \stackrel{\text{def.}}{=} \log P(x | \theta)$.

Observație (2):

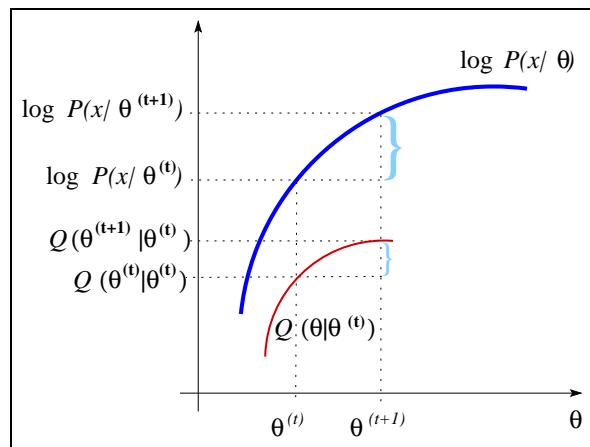
Dacă în inegalitatea (384) se înlocuiește θ cu

$$\theta^{(t+1)} = \operatorname{argmax}_{\theta} Q(\theta | \theta^{(t)}),$$

va rezulta

$$\log P(x | \theta^{(t+1)}) \geq \log P(x | \theta^{(t)}).$$

Altfel spus, la fiecare iterație a algoritmului EM, odată cu trecerea de la $\theta^{(t)}$ la $\theta^{(t+1)}$, valoarea funcției de log-verosimilitate a datelor observabile, $\ell(\theta) \stackrel{\text{def.}}{=} \log P(x | \theta)$ crește sau, în cel mai rău caz, rămâne pe loc.



În final, vom avea $\ell(\theta^{(0)}) \leq \ell(\theta^{(t)}) \leq \ell(\theta^{(t+1)}) \leq \dots$. Sirul acesta (monoton) este mărginit superior de 0 (vedeți definiția lui ℓ), deci converge la o anumită valoare ℓ^* . În anumite cazuri / condiții, această valoare este un maxim (în general, local) al funcției de log-verosimilitate.⁹⁴²

Observație (3): Conform aceleiași inegalități (384), la pasul M de la iterația t a algoritmului EM, este suficient ca în loc să se ia $\theta^{(t+1)} = \operatorname{argmax}_{\theta} Q(\theta | \theta^{(t)})$, să se aleagă $\theta^{(t+1)}$ astfel încât $Q(\theta^{(t+1)} | \theta^{(t)}) > Q(\theta^{(t)} | \theta^{(t)})$. Aceasta constituie *versiunea „generalizată“* a algoritmului EM.

Răspuns:

$P(x, z | \theta) = P(z | x, \theta) \cdot P(x | \theta)$, conform regulii de înmulțire a probabilităților. Prin logaritmarea acestei egalități, rezultă:

$$\log P(x | \theta) = \log P(x, z | \theta) - \log P(z | x, \theta)$$

Ca să obținem expresia funcției „auxiliare“ Q pe care o optimizează algoritmul EM la pasul M al iterației t , vom înmulții ambii membri ai egalității precedente cu $P(z | x, \theta^{(t)})$ și apoi vom suma după toate valorile posibile ale lui z :

$$\begin{aligned} \sum_z P(z | x, \theta^{(t)}) \cdot \log P(x | \theta) = \\ \sum_z P(z | x, \theta^{(t)}) \cdot \log P(x, z | \theta) - \sum_z P(z | x, \theta^{(t)}) \cdot \log P(z | x, \theta) \end{aligned} \quad (385)$$

Membrul stâng al acestei egalități poate fi rescris astfel:

$$\sum_z P(z | x, \theta^{(t)}) \cdot \log P(x | \theta) = \log P(x | \theta) \cdot \underbrace{\sum_z P(z | x, \theta^{(t)})}_{1}$$

În ceea ce privește membrul drept al aceleiași egalități (385), întrucât termenul al doilea, și anume $-\sum_z P(z | x, \theta^{(t)}) \cdot \log P(z | x, \theta)$ reprezintă o cross-entropie (vedeți problema 64 de la capitolul de *Fundamente*), o vom nota cu $CH(\theta | \theta^{(t)})$. Așadar, egalitatea (385) devine:

$$\log P(x | \theta) = Q(\theta | \theta^{(t)}) + CH(\theta | \theta^{(t)}) \quad (386)$$

Această egalitate este valabilă pentru toate valorile posibile ale parametrului θ . În particular pentru $\theta = \theta^{(t)}$, vom avea:

$$\log P(x | \theta^{(t)}) = Q(\theta^{(t)} | \theta^{(t)}) + CH(\theta^{(t)} | \theta^{(t)}) \quad (387)$$

Scăzând membru cu membru ultimele două egalități, obținem:

$$\log P(x | \theta) - \log P(x | \theta^{(t)}) = Q(\theta | \theta^{(t)}) - Q(\theta^{(t)} | \theta^{(t)}) + CH(\theta | \theta^{(t)}) - CH(\theta^{(t)} | \theta^{(t)})$$

Conform inegalității lui Gibbs — a se vedea problema 65 de la capitolul de *Fundamente* —,⁹⁴³ avem $CH(\theta | \theta^{(t)}) \geq CH(\theta^{(t)} | \theta^{(t)})$, deci în final rezultă:

$$\log P(x | \theta) - \log P(x | \theta^{(t)}) \geq Q(\theta | \theta^{(t)}) - Q(\theta^{(t)} | \theta^{(t)})$$

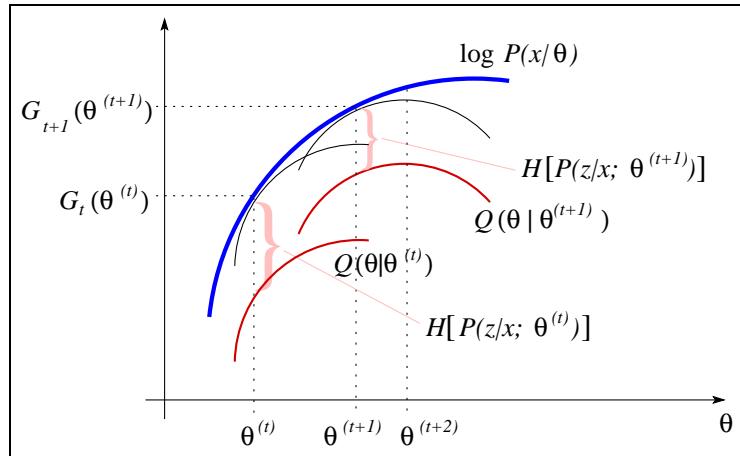
⁹⁴²Vedeți capitolul *The EM Algorithm*, de Shu Kay Ng et al., în *Handbook of Computational Statistics*, Springer, 2004, pag. 139.

⁹⁴³Folosind noțiunea de divergență Kullback-Leibler, inegalitatea aceasta se scrie sub forma $KL(p || q) \geq 0$. În notațiile / termenii din rezolvarea problemei noastre, această inegalitate se scrie $CH(p||q) \geq CH(p||p) = H(p)$.

Aceasta este chiar inegalitatea pe care trebuia s-o demonstrăm.

Observație (4):

Cu ajutorul figurii alăturate, putem să ilustrăm și totodată să sumarizăm în mod grafic rezultatele pe care le-am demonstrat la aceste ultime două probleme (1 și 2).



Observație (5): [Proprietăți ale algoritmului EM⁹⁴⁴]

Algoritmul EM este util datorită următoarelor *avantaje*: simplitate conceptuală, posibilitatea estimării datelor neobservabile, ușurința implementării, precum și faptul că la fiecare iterație se îmbunătățesc valorile parametrului / parametrilor θ . Rata de convergență la primele câteva iterări ale algoritmului EM este în general foarte bună, însă ea poate deveni foarte mică pe măsură ce ne apropiem de punctul de optim local. În general, algoritmul EM dă rezultate bune atunci când proporția datelor neobservabile este mică, iar dimensiunea datelor (adică, numărul de atribute ale instanțelor) nu este prea mare. EM poate necesita executarea multor iterări, iar în cazul dimensiunilor mari pasul E poate deveni foarte lent.

3. **(Algoritmul EM pentru estimarea parametrilor în sens MAP: fundamentare teoretică)**

□ • ○ *Stanford, 2016 fall. A. Ng, J. Duchi, HW4, pr. 1*

Algoritmul EM, aşa cum l-am folosit până la acest exercițiu, a fost conceput pentru a rezolva [unele] probleme de *estimare de parametri în sensul verosimilității maxime* (engl., maximum likelihood estimation, MLE). În astfel de probleme se urmărește să se maximizeze o expresie de forma

$$\prod_{i=1}^n p(x_i|\theta) = \prod_{i=1}^n \left(\sum_{z_i} p(x_i, z_i|\theta) \right), \quad (388)$$

unde

- θ este setul de parametri pentru distribuția probabilistă p ,
- x_i (cu $i = 1, \dots, n$) sunt variabile aleatoare observabile,
- z_i (cu $i = 1, \dots, n$) sunt variabile aleatoare neobservabile / latente.

⁹⁴⁴Cf. documentului *The EM Algorithm*, Ajit Singh, CMU, November 20, 2005. Vedeți de asemenea secțiunea 5.2.5 din capitolul *The EM Algorithm*, de Shu Kay Ng et al., din volumul *Handbook of Computational Statistics*, Springer, 2004.

În acest exercițiu presupunem că lucrăm într-un *cadrul bayesian*, adică dorim să găsim *estimarea în sensul probabilității maxime a posteriori* (engl., maximum a posteriori probability, MAP) pentru parametrii θ . Aceasta revine la a maximiza o expresie de forma următoare:⁹⁴⁵

$$\left(\prod_{i=1}^n p(x_i|\theta) \right) \cdot p(\theta) = \left(\prod_{i=1}^n \left(\sum_{z_i} p(x_i, z_i|\theta) \right) \right) \cdot p(\theta),$$

unde $p(\theta)$ desemnează o distribuție de probabilitate a priori definită peste valorile parameterilor θ .

Generalizați algoritmul EM astfel încât să realizeze estimări în sens MAP. Veți presupune că ambele distribuții — $p(x, z|\theta)$ și $p(\theta)$ — sunt concave în raport cu parametrul θ . Aceasta va implica faptul că pasul M al algoritmului EM este *realizabil* (engl., tractable) atunci când se cere doar să se maximizeze o combinație liniară de aceste cantități ($p(x, z|\theta)$ și $p(\theta)$).⁹⁴⁶

Asigurați-vă că pasul M din formularea pe care o veți adopta pentru noul algoritm EM este într-adevăr *realizabil*. De asemenea, demonstrați că valorile expresiei $(\prod_{i=1}^n p(x_i|\theta)) \cdot p(\theta)$, văzută ca funcție de θ , nu descresc — adică, fie cresc, fie rămân pe loc — de la o iterație la alta a noului algorithm.

Răspuns:

Vom deduce regulile de actualizare de la cei doi pași ai algoritmului EM într-un mod similar celui în care am procedat la problema 1 pentru estimare în sensul verosimilității maxime (MLE). La fiecare iterație (t) , creșterea este monotonă, din aceleași motive ca și la problema 1: la pasul E calculăm [o] cea mai bună *margine inferioară* (engl., lower bound) a log-probabilității a posteriori pentru datele observabile, în raport cu $\theta^{(t)}$, care reprezintă valorile curente ale parametrilor θ , iar la pasul M alegem pentru θ o nouă valoare, $\theta^{(t+1)}$, care este optimă în raport cu această margine inferioară.

Pornind de la expresia (388) din enunț, vom identifica marginea inferioară despre care tocmai am vorbit:

$$\begin{aligned} \log \left(\left(\prod_{i=1}^n p(x_i|\theta) \right) \cdot p(\theta) \right) &= \log p(\theta) + \sum_{i=1}^n \log p(x_i|\theta) \\ &= \log p(\theta) + \sum_{i=1}^n \log \sum_{z_i} p(x_i, z_i|\theta) = \log p(\theta) + \sum_{i=1}^n \log \sum_{z_i} q(z_i) \cdot \frac{p(x_i, z_i|\theta)}{q(z_i)} \end{aligned}$$

⁹⁴⁵Justificarea este aceasta: log-probabilitatea a posteriori pentru datele observabile este

$$\begin{aligned} \log p(\theta|x_1, \dots, x_n) &\stackrel{F. Bayes}{=} \log \frac{p(x_1, \dots, x_n|\theta) \cdot p(\theta)}{p(x_1, \dots, x_n)} = \log(p(x_1, \dots, x_n|\theta) \cdot p(\theta)) - \log p(x_1, \dots, x_n) \\ &\stackrel{i.i.d.}{=} \log \left(\left(\prod_{i=1}^n p(x_i|\theta) \right) \cdot p(\theta) \right) - \log p(x_1, \dots, x_n). \end{aligned}$$

Întrucât cel de-al doilea termen nu depinde de θ , rezultă

$$\operatorname{argmax}_{\theta} \log p(\theta|x_1, \dots, x_n) = \operatorname{argmax}_{\theta} \log \left(\left(\prod_{i=1}^n p(x_i|\theta) \right) \cdot p(\theta) \right).$$

⁹⁴⁶Grosso-modo, aceasta revine la a presupune că estimarea în sens MAP este *realizabilă* atunci când variabilele x, z sunt în totalitate observabile, exact cum procedam în cadrul „frecvenționist“ (adică, non-bayesian).

$$\geq \log p(\theta) + \sum_{i=1}^n \sum_{z_i} q(z_i) \cdot \log \frac{p(x_i, z_i | \theta)}{q(z_i)}. \quad (389)$$

La ultimul pas de mai sus am aplicat inegalitatea lui Jensen.⁹⁴⁷ Ca și la problema 1, inegalitatea (389) are loc cu egalitate atunci când la *pasul E* (de la iteratăția t) luăm

$$q(z_i) = p(z_i | x_i; \theta^{(t)}).$$

Pentru *pasul M*, trebuie să maximizăm marginea inferioară (vedeți relația (389)), adică vom calcula

$$\theta^{(t+1)} = \operatorname{argmax}_{\theta} \left(\log p(\theta) + \sum_{i=1}^n \sum_{z_i} q(z_i) \log \frac{p(x_i, z_i | \theta)}{q(z_i)} \right).$$

Pasul M este *realizabil*, întrucât ceea ce se cere aici este doar să maximizăm o combinație liniară de termenii $\log p(\theta)$ și $\log p(x, z | \theta)$, care, conform ipotezei din enunț, sunt concavi.

În sfârșit, nu este deloc dificil să demonstrăm faptul că produsul $(\prod_{i=1}^n p(x_i | \theta)) \cdot p(\theta)$, văzut ca o funcție de argumentul θ , crește monoton de la o iteratăție la alta a algoritmului EM. Practic, tot ceea ce trebuie făcut este să extindem ușor demonstrația de la problema 2. Într-adevăr, notând

$$Q'(\theta | \theta^{(t)}) = \log p(\theta) + \underbrace{\sum_{i=1}^n \sum_{z_i} q(z_i) \log \frac{p(x_i, z_i | \theta)}{q(z_i)}}_{Q(\theta | \theta^{(t)})},$$

relațiile (386) și (387) devin

$$\begin{aligned} \log P(x | \theta) + \log P(\theta) &= Q(\theta | \theta^{(t)}) + CH(\theta | \theta^{(t)}) + \log P(\theta) \\ \log P(x | \theta^{(t)}) + \log P(\theta^{(t)}) &= Q(\theta^{(t)} | \theta^{(t)}) + CH(\theta^{(t)} | \theta^{(t)}) + \log P(\theta^{(t)}). \end{aligned}$$

Scăzându-le membru cu membru, obținem:

$$\begin{aligned} \log P(x | \theta) + \log P(\theta) - (\log P(x | \theta^{(t)}) + \log P(\theta^{(t)})) \\ = Q(\theta | \theta^{(t)}) + \log P(\theta) - (Q(\theta^{(t)} | \theta^{(t)}) + \log P(\theta^{(t)})) + CH(\theta | \theta^{(t)}) - CH(\theta^{(t)} | \theta^{(t)}) \\ = Q'(\theta | \theta^{(t)}) - Q'(\theta^{(t)} | \theta^{(t)}) + \underbrace{CH(\theta | \theta^{(t)}) - CH(\theta^{(t)} | \theta^{(t)})}_{\geq 0}. \end{aligned}$$

Concluzia este că în mod garantat valoarea funcției de probabilitate a posteriori a datelor observabile nu descrește, ci fie crește fie rămâne pe loc, la fiecare iteratăție a acestui algoritm de tip EM pentru estimare de parametri în sens MAP.

4.

(Adevărat sau Fals?)

CMU, 2011 fall, Eric Xing, HW2, pr. 3.1.ac

- a. Algoritmul EM maximizează log-verosimilitatea datelor complete (observabile și neobservabile).

⁹⁴⁷Vedeți problema 79.c de la capitolul de *Fundamente*.

b. Funcția optimizată de către algoritmul EM este o margine inferioară (engl., lower bound) pentru log-verosimilitatea datelor complete.

Răspuns:

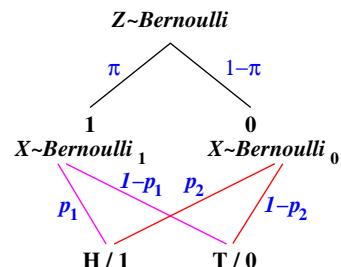
- a. Fals. Algoritmul EM aplică o metodă de optimizare (și anume *coordinate-ascent*) pentru a găsi un maxim (în general, un maxim local) al funcției $F(q, \theta)$,⁹⁴⁸ care este media log-verosimilității datelor complete (observabile și neobservabile) și, în același timp, o margine inferioară pentru log-verosimilitatea datelor incomplete / observabile x . (A se vedea problema 1 de la acest capitol.)
- b. Fals. Funcția F este o margine inferioară pentru log-verosimilitatea datelor incomplete / observabile.

8.1.2 Mixturi de distribuții Bernoulli / categoriale

5. (O mixtură de distribuții Bernoulli; estimarea unui parametru cu ajutorul algoritmului EM, folosind distribuția binomială⁹⁴⁹)
*formulare de Liviu Ciortuz, pornind de la
 ■ □ • CMU, 2014 fall, W. Cohen, Z. Bar-Joseph, HW1, pr. 2.b*

Presupunem că avem două monede: una este perfectă, având deci probabilitatea de apariție a feței cu stema $p_1 = 1/2$, iar cealaltă monedă este măsluită / imperfectă, având probabilitatea de apariție a stemei $p_2 = 1/3$.

Facem 100 de aruncări după cum urmează. De fiecare dată alegem una dintre cele două monede. Cu o probabilitate necunoscută π , alegem moneda perfectă, iar cu probabilitatea $1-\pi$ alegem moneda măsluită. Aruncăm moneda aleasă și notăm / reținem rezultatul aruncării, dar nu notăm și informația referitoare la ce monedă am folosit. Constatăm la final că din totalul celor 100 de aruncări am obținut de 40 de ori stema și de 60 de ori banul.



Observație: Pe lângă cele trei metode prezentate la pr. 45 de la capitolul de *Fundamente*, pentru estimarea în sensul verosimilității maxime (MLE) a parametrului π al mixturii din enunț se poate folosi și algoritmul EM.

- a. Elaborați varianta / „instanța“ algoritmului EM pentru problema de față, stabilind mai întâi cine sunt variabilele observabile și respectiv variabilele neobservabile. Scrieți expresia funcției care trebuie optimizată — funcția de log-verosimilitate a datelor „observabile“ — și expresia funcției auxiliare $Q(\pi|\pi^{(t)})$, care este media log-verosimilității datelor „complete“ (adică, „observabile“ și

⁹⁴⁸ Această funcție se mai numește „funcția de energie liberă“ (engl., free-energy functional).

⁹⁴⁹ Pentru estimarea aceluiasi parametru folosind metoda analitică, metoda gradientului și metoda lui Newton, vedeți ex. 45 de la capitolul *Fundamente*.

„neobservabile“), iar în final deduceti formulele de actualizare pentru cei doi pași (E și M) din corpul iterativ al algoritmului.

b. Implementati algoritmul EM pe care l-ați obtinut la punctul a. Comparați numărul de iterații [precum și timpul necesar] pentru a se ajunge la convergență pentru cei trei algoritmi: algoritmul EM, metoda lui Newton și metoda gradientului ascendent. Veți folosi în toate cele trei cazuri aceeași valoare inițială $\pi^{(0)} = 0.1$, precum și aceeași condiție de oprire, $|\pi^{(t+1)} - \pi^{(t)}| < 10^{-4}$.

Răspuns:

a. Asociind fiecărei „observații“ din cele 100 despre care este vorba în enunț o variabilă aleatoare $X_i \sim Bernoulli(q)$ și notând $X = (X_1, \dots, X_{100})$, funcția de log-verosimilitate a datelor „observabile“ va fi scrisă astfel:

$$\ell(\pi) = \ln P(X|\pi) = \ln \left(C_{100}^{40} \underbrace{P(X_i = H)}_q^{40} \underbrace{P(X_i = T)}_{1-q}^{60} \right) = \ln \left(C_{100}^{40} q^{40} (1-q)^{60} \right),$$

unde

$$\begin{aligned} q &\stackrel{\text{not.}}{=} P(X = H) \stackrel{F.P.T.}{=} P(X = H, Z = 1) + P(X = H, Z = 0) \\ &= P(X = H|Z = 1) \cdot P(Z = 1) + P(X = H|Z = 0) \cdot P(Z = 0) \\ &= \frac{1}{2} \cdot \pi + \frac{1}{3} \cdot (1 - \pi) = \frac{1}{3} + \frac{\pi}{6}. \end{aligned}$$

Notând cu \tilde{z}_1 numărul (neobservabil) de apariții ale stemei pentru moneda 1 (din totalul celor $n_S = 40$ de aruncări) și cu \tilde{z}_0 numărul (neobservabil) de apariții ale banului pentru moneda 1 (din totalul celor $n_B = 60$ de aruncări), log-verosimilitatea datelor „complete“ va putea fi scrisă astfel:⁹⁵⁰

$$\begin{aligned} \ell_{compl}(\pi) &\stackrel{\text{def.}}{=} \ln P(X, Z|\pi) \stackrel{\text{indep.}}{=} \sum_{i=1}^{100} \ln (P(X_i|Z_i, \pi) \cdot P(Z_i|\pi)) \quad (390) \\ &= \tilde{z}_1 \ln(\pi p_1) + (n_S - \tilde{z}_1) \ln((1 - \pi)p_2) \\ &\quad + \tilde{z}_0 \ln(\pi(1 - p_1)) + (n_B - \tilde{z}_0) \ln((1 - \pi)(1 - p_2)) \\ &= \tilde{z}_1 (\ln \pi + \ln p_1) + (n_S - \tilde{z}_1) (\ln(1 - \pi) + \ln p_2) \\ &\quad + \tilde{z}_0 (\ln \pi + \ln(1 - p_1)) + (n_B - \tilde{z}_0) (\ln(1 - \pi) + \ln(1 - p_2)). \end{aligned}$$

Observație: În relația (390) am considerat variabilele aleatoare (neobservabile) Z_i luând valoarea 1 dacă „observația“ $X_i = x_i$ a fost realizată folosind moneda 1, și respectiv valoarea 0 dacă s-a folosit moneda 2.

Vom calcula funcția „auxiliară“ $Q(\pi|\pi^{(t)})$, folosind proprietatea de liniaritate a mediilor (vedeți pr. 9.a de la capitolul de *Fundamente*):

$$\begin{aligned} Q(\pi|\pi^{(t)}) &\stackrel{\text{def.}}{=} E[\ell_{compl}(\pi)] \\ &= E[\tilde{z}_1](\ln \pi + \ln p_1) + (n_S - E[\tilde{z}_1])(\ln(1 - \pi) + \ln p_2) \\ &\quad + E[\tilde{z}_0](\ln \pi + \ln(1 - p_1)) + (n_B - E[\tilde{z}_0])(\ln(1 - \pi) + \ln(1 - p_2)) \\ &= \underbrace{c}_{\text{const.}} + (E[\tilde{z}_1] + E[\tilde{z}_0]) \ln \pi + \underbrace{(n_S + n_B - E[\tilde{z}_1] - E[\tilde{z}_0])}_{100} \ln(1 - \pi). \quad (391) \end{aligned}$$

⁹⁵⁰Putem considera n_S și n_B ca fiind datele „observabile“.

Acum vom determina regulile de „actualizare“ de la cei doi pași (Pasul E și respectiv Pasul M) ai algoritmului EM:

Pasul E: Mediile $E[\tilde{z}_1]$ și $E[\tilde{z}_0]$ se calculează ținând cont că $\tilde{z}_1 \sim \text{Binomial}(40, q_1^{(t)})$ și $\tilde{z}_0 \sim \text{Binomial}(60, q_0^{(t)})$, unde

$$\begin{aligned} q_1^{(t)} &\stackrel{\text{not.}}{=} P(Z_i = 1 | X_i = 1, \pi^{(t)}) \stackrel{F.B., F.P.T.}{=} \frac{\pi^{(t)} p_1}{\pi^{(t)} p_1 + (1 - \pi^{(t)}) p_2} \\ q_0^{(t)} &\stackrel{\text{not.}}{=} P(Z_i = 1 | X_i = 0, \pi^{(t)}) \stackrel{F.B., F.P.T.}{=} \frac{\pi^{(t)} (1 - p_1)}{\pi^{(t)} (1 - p_1) + (1 - \pi^{(t)}) (1 - p_2)}, \end{aligned}$$

ceea ce implică⁹⁵¹

$$E[\tilde{z}_1] = 40q_1^{(t)} \text{ și } E[\tilde{z}_0] = 60q_0^{(t)}.$$

Pasul M: Pentru a determina regula de „actualizare“ de la acest pas, vom calcula derivata parțială a funcției „auxiliare“ $Q(\pi|\pi^{(t)})$ în raport cu π și apoi o vom egala cu 0:

$$\begin{aligned} \frac{\partial Q}{\partial \pi}(\pi|\pi^{(t)}) &\stackrel{(391)}{=} \frac{E[\tilde{z}_1] + E[\tilde{z}_0]}{\pi} - \frac{n_S + n_B - E[\tilde{z}_1] - E[\tilde{z}_0]}{1 - \pi} \\ \frac{\partial Q}{\partial \pi}(\pi) = 0 \Leftrightarrow (1 - \pi)(E[\tilde{z}_1] + E[\tilde{z}_0]) &= \pi(n_S + n_B - (E[\tilde{z}_1] + E[\tilde{z}_0])) \\ \Rightarrow \pi &= \frac{E[\tilde{z}_1] + E[\tilde{z}_0]}{n_S + n_B}. \end{aligned} \quad (392)$$

Se poate constata ușor că derivata a doua a funcției $Q(\pi|\pi^{(t)})$ în raport cu π este negativă pe tot domeniul ei de definiție, ceea ce implică faptul că soluția primei derive (care a fost calculată mai sus) este punct de maxim pentru $Q(\pi|\pi^{(t)})$.

Observați semnificația rezultatului (392): $E[n_S] + E[n_B]$ reprezintă câte aruncări au fost făcute (în medie) cu moneda 1 (la iterată t), deci este „natural“ ca π să devină (pentru iterată $t+1$) raportul dintre $E[n_S] + E[n_B]$ și numărul total de aruncări, $n_S + n_B$.

Concluzionând, regula de actualizare pentru Pasul M al algoritmului EM pentru rezolvarea problemei de estimare a parametrului π al distribuției probabiliste descrise în enunțul problemei este următoarea:

$$\pi^{(t+1)} = \frac{40q_1^{(t)} + 60q_0^{(t)}}{100} = \frac{2q_1^{(t)} + 3q_0^{(t)}}{5}.$$

b. Pseudo-codul algoritmului EM pentru rezolvarea problemei de tip MLE din enunț este următorul:

Initializare:

atribuie o valoare arbitrară $\pi^{(0)}$ în intervalul $(0, 1)$ pentru parametrul π ;

Corpul iterativ:

pentru $t = 0, \dots, T-1$ (cu T fixat în avans)

(sau: până când $|\pi^{(t)} - \pi^{(t+1)}| < \varepsilon$, cu ε fixat în avans) execută

Pasul E:

$$q_1^{(t)} = \frac{\pi^{(t)} p_1}{\pi^{(t)} p_1 + (1 - \pi^{(t)}) p_2};$$

⁹⁵¹Pentru formula de calcul a mediei distribuției binomiale, vedeti pr. 25.b de la capitolul de *Fundamente*.

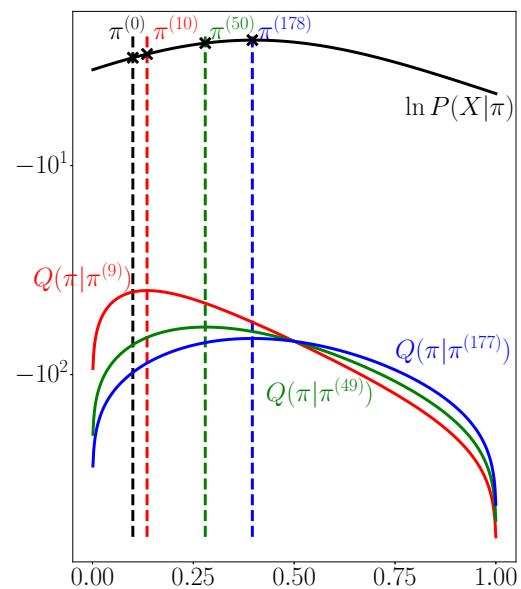
$$q_0^{(t)} = \frac{\pi^{(t)}(1-p_1)}{\pi^{(t)}(1-p_1) + (1-\pi^{(t)})(1-p_2)};$$

Pasul M:

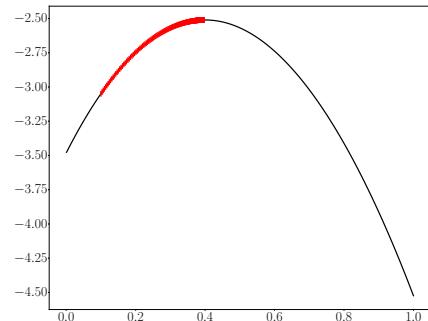
$$\pi^{(t+1)} = \frac{2q_1^{(t)} + 3q_0^{(t)}}{5};$$

Returnează $\pi^{(t+1)}$;

Graficul alăturat a fost făcut de către doctorandul Andi Munteanu. El pune în evidență funcția de log-verosimilitate a datelor „observabile“, precum și trei funcții „auxiliare“, corespunzătoare iterării $t = 9$, $t = 49$, și respectiv $t = 177$. Se observă convergența către valoarea optimă a funcției de log-verosimilitate a datelor „observabile“.



Graficul alăturat a fost făcut tot de către Andi Munteanu. El pune în evidență evoluția algoritmului EM. Pe axa Ox a fost reprezentată valoarea probabilității π , iar pe axa Oy funcția de verosimilitate. Punctul de plecare a fost $\pi^{(0)} = 0.1$. Numărul de iterări necesare pentru a ajunge la convergență (pentru $\epsilon = 10^{-4}$) a fost 178. Timpul necesar pentru executarea acestor iterări a fost de 0.02597 secunde.



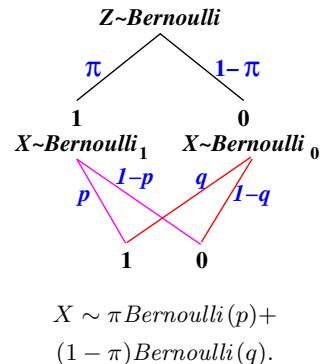
Observație: Pentru o comparație între numărul de iterări și timpul necesar pentru a se ajunge la convergență pentru metoda gradientului și metoda lui Newton, vedeți pr. 45 de la capitolul *Fundamente*.

6.

(Algoritmul EM: estimarea parametrilor unei mixturi de două distribuții Bernoulli (cazul general))

■ • CMU, 2008 fall, Eric Xing, HW4, pr. 1.4-7
CMU, "The EM Algorithm," Ajit Singh, November 20, 2005

Să presupunem că avem două monede. La aruncarea primei monede se obține față ‘stemă’ cu probabilitatea p , în vreme ce la aruncarea celei de-a două monede se obține ‘stemă’ cu probabilitatea q . Mai presupunem că se efectuează n aruncări, iar la fiecare aruncare se alege prima monedă cu probabilitatea π , iar moneda a două cu probabilitatea $1 - \pi$. Rezultatul fiecărei aruncări i este $x_i \in \{0, 1\}$, notația aceasta din urmă codificând mulțimea ordonată $\{T, H\} = \{\text{'tail'}, \text{'head'}\} = \{\text{'ban'}, \text{'stemă'}\}$.



Jocul pe care îl propunem este următorul:

Noi îți furnizăm doar rezultatul celor n aruncări, adică $x = \{x_1, x_2, \dots, x_n\}$, fără a-ți spune ce monedă am folosit pentru fiecare aruncare. Sarcina ta este următoarea: folosind algoritmul EM și disponând de datele „observabile“ x , va trebui să estimezi valorile [de verosimilitate maximă] pentru parametrii probabiliști p , q și π . Vom desemna ansamblul acestor parametri prin θ .

Pentru a calcula aceste estimări, se va considera $z = \{z_1, z_2, \dots, z_n\}$, cu $z_i \in \{0, 1\}$ variabilă „ascunsă“ indicând moneda utilizată la aruncarea i . Dacă, de exemplu, avem $z_2 = 1$, aceasta înseamnă că la aruncarea a două a fost folosită prima monedă.

- a. Arată că $E[z_i | x_i, \theta] = P(z_i = 1 | x_i, \theta)$.
- b. Folosește regula lui Bayes pentru a calcula $P(z_i = 1 | x_i, \theta)$ în funcție de x_i , z_i , p , q și π .
- c. Calculează log-verosimilitatea datelor „complete“, $\log P(x, z | \theta)$, ca funcție de x_i , z_i (pentru $i = 1, \dots, n$), p , q și π .
- d. *Pasul E:* Arată că media log-verosimilității datelor complete $Q(\theta | \theta^{(t)}) \stackrel{\text{not.}}{=} E_{P(z|x,\theta^{(t)})}[\log P(x, z | \theta)]$ este dată de expresia:

$$\begin{aligned} Q(\theta | \theta^{(t)}) &= \sum_{i=1}^n E[z_i | x_i, \theta^{(t)}] \cdot (\log \pi + x_i \log p + (1 - x_i) \log(1 - p)) + \\ &\quad + (1 - E[z_i | x_i, \theta^{(t)}]) \cdot (\log(1 - \pi) + x_i \log q + (1 - x_i) \log(1 - q)), \end{aligned}$$

unde $\theta^{(t)} \stackrel{\text{def.}}{=} \{p^{(t)}, q^{(t)}, \pi^{(t)}\}$ desemnează valorile celor trei parametri la iteratărea t a algoritmului EM.

- e. *Pasul M:* Elaborează formulele de calcul prin care la finalul iterăției t a algoritmului EM se obțin noile valori pentru parametri, $p^{(t+1)}$, $q^{(t+1)}$ și $\pi^{(t+1)}$.
- f. Scrie pseudo-codul algoritmului EM pentru rezolvarea acestui model de mixtură și apoi execută manual o iterăție pe inputul $x = \{1, 1, 0, 1, 0, 0, 1, 0,$

$0, 0, 1, 1\}$, cu valorile inițiale $1/3, 2/3$ și $1/2$ pentru parametrii p, q și respectiv π . Ce se poate observa în privința convergenței algoritmului EM?

Răspuns:

a. Este ușor de calculat $E[z_i | x_i, \theta]$, utilizând definiția mediei unei variabile aleatoare. (Stim că variabila z_i ia valori în mulțimea $\{0, 1\}$.)

$$\begin{aligned} E[z_i | x_i, \theta] &= \sum_{z \in \{0,1\}} z_i P(z_i | x_i, \theta) = 0 \cdot P(z_i = 0 | x_i, \theta) + 1 \cdot P(z_i = 1 | x_i, \theta) \\ &\Rightarrow E[z_i | x_i, \theta] = P(z_i = 1 | x_i, \theta). \end{aligned}$$

b. Folosind regula lui Bayes și probabilitățile din enunț, obținem:

$$\begin{aligned} P(z_i = 1 | x_i, \theta) &= \frac{P(x_i | z_i = 1, \theta)P(z_i = 1 | \theta)}{P(x_i | z_i = 1, \theta)P(z_i = 1 | \theta) + P(x_i | z_i = 0, \theta)P(z_i = 0 | \theta)} \\ &= \frac{p^{x_i} \cdot (1-p)^{1-x_i} \cdot \pi}{p^{x_i} \cdot (1-p)^{1-x_i} \cdot \pi + q^{x_i} \cdot (1-q)^{1-x_i} \cdot (1-\pi)}. \end{aligned}$$

La ultima egalitate am ținut cont că $x_i \in \{0, 1\}$. Observați că produsul

$$p^{x_i}(1-p)^{1-x_i} \quad (393)$$

exprimă simultan ambele cazuri posibile pentru $P(x_i | z_i = 1)$, adică $P(x_i = 1 | z_i = 1) = p$ și $P(x_i = 0 | z_i = 1) = 1 - p$. Similar pentru $q^{x_i}(1-q)^{1-x_i}$.

c. Pentru a exprima log-verosimilitatea datelor complete, ținem cont de independența celor n aruncări:⁹⁵²

$$\begin{aligned} \log P(x, z | \theta) &\stackrel{i.i.d.}{=} \log \prod_{i=1}^n P(x_i, z_i | \theta) = \log \prod_{i=1}^n P(x_i | z_i, \theta) \cdot P(z_i | \theta) \\ &= \log \prod_{i=1}^n (p^{x_i}(1-p)^{1-x_i}\pi)^{z_i} (q^{x_i}(1-q)^{1-x_i}(1-\pi))^{1-z_i} \\ &= \sum_{i=1}^n \log \left((p^{x_i}(1-p)^{1-x_i}\pi)^{z_i} (q^{x_i}(1-q)^{1-x_i}(1-\pi))^{1-z_i} \right) \\ &= \sum_{i=1}^n [z_i \log(p^{x_i}(1-p)^{1-x_i}\pi) + (1-z_i) \log(q^{x_i}(1-q)^{1-x_i}(1-\pi))]. \end{aligned}$$

La a treia egalitate de mai sus am folosit un „artificiu“ similar cu cel de la punctul c, ținând cont că $z_i \in \{0, 1\}$, iar valoarea $z_i = 1$ desemnează prima monedă în vreme ce valoarea $z_i = 0$ desemnează cea de-a doua monedă.

d. Pentru a calcula media log-verosimilității datelor complete în raport cu distribuția a posteriori $P(z | x, \theta^{(t)})$, vom ține cont de rezultatul de la punctul

⁹⁵²Din punct de vedere metodologic, pentru a calcula $P(x, z | \theta)$ ne putem inspira de la punctul precedent: observăm că la numitorul fracției care ne dă valoarea lui $P(z_i | x_i, \theta)$ avem $P(x_i, z_i = 1 | \theta)$ și $P(x_i, z_i = 0 | \theta)$. Pornind de la aceste două expresii, ne vom propune să le exprimăm în mod unitar, adică sub forma unei singure expresii, folosind *artificiul exponențierii* (engl., the exponentiation trick).

c și de proprietatea de liniaritate a mediilor.

$$\begin{aligned}
 Q(\theta | \theta^{(t)}) &\stackrel{\text{not.}}{=} E_{P(z|x, \theta^{(t)})}[\log P(x, z | \theta)] \\
 &= E_{P(z|x, \theta^{(t)})} \left[\sum_{i=1}^n [z_i \log(p^{x_i}(1-p)^{1-x_i}\pi) + \right. \\
 &\quad \left. (1-z_i) \log(q^{x_i}(1-q)^{1-x_i}(1-\pi))] \right] \\
 &= \sum_{i=1}^n \left[E[z_i | x_i, \theta^{(t)}] \cdot \log(p^{x_i}(1-p)^{1-x_i})\pi + \right. \\
 &\quad \left. + (1 - E[z_i | x_i, \theta^{(t)}]) \cdot \log(q^{x_i}(1-q)^{1-x_i}(1-\pi)) \right] \\
 &= \sum_{i=1}^n \left[E[z_i | x_i, \theta^{(t)}] \cdot (\log \pi + x_i \log p + (1-x_i) \log(1-p)) + \right. \\
 &\quad \left. + (1 - E[z_i | x_i, \theta^{(t)}]) \cdot (\log(1-\pi) + x_i \log q + (1-x_i) \log(1-q)) \right].
 \end{aligned}$$

e. Pentru a calcula relațiile de actualizare pentru $p^{(t+1)}$, $q^{(t+1)}$ și $\pi^{(t+1)}$, se maximizează funcția $Q(\theta | \theta^{(t)})$, care reprezintă media log-verosimilității datelor complete, în raport cu parametrii $\theta = (p, q, \pi)$. Mai exact, se folosesc derivatele parțiale de ordinul întâi în funcție de p , q și respectiv π . Pentru a spori claritatea calculelor de mai jos, vom nota cu $\mu_i^{(t)}$ media $E[z_i | x_i, \theta^{(t)}]$, conform calculelor de la punctele a și b:

$$\mu_i^{(t)} = E[z_i | x_i, \theta^{(t)}] \stackrel{a,b}{=} \frac{(p^{(t)})^{x_i} \cdot (1-p^{(t)})^{1-x_i} \cdot \pi^{(t)}}{(p^{(t)})^{x_i} \cdot (1-p^{(t)})^{1-x_i} \cdot \pi^{(t)} + (q^{(t)})^{x_i} \cdot (1-q^{(t)})^{1-x_i} \cdot (1-\pi^{(t)})}.$$

Cu această notație, funcția $Q(\theta | \theta^{(t)})$, a cărei expresie a fost calculată la punctul d devine:

$$\begin{aligned}
 Q(\theta | \theta^{(t)}) &= \sum_{i=1}^n \left[\mu_i^{(t)} (\log \pi + x_i \log p + (1-x_i) \log(1-p)) + \right. \\
 &\quad \left. + (1 - \mu_i^{(t)}) \cdot (\log(1-\pi) + x_i \log q + (1-x_i) \log(1-q)) \right].
 \end{aligned}$$

Regula de *actualizare* (engl., update) pentru parametrul p se obține astfel:

$$\begin{aligned}
 \frac{\partial Q(\theta | \theta^{(t)})}{\partial p} &= 0 \Leftrightarrow \sum_{i=1}^n \mu_i^{(t)} \left(\frac{x_i}{p} - \frac{1-x_i}{1-p} \right) = 0 \\
 \Leftrightarrow \frac{1}{p} \sum_{i=1}^n \mu_i^{(t)} x_i &= \frac{1}{1-p} \sum_{i=1}^n \mu_i^{(t)} (1-x_i) \Leftrightarrow (1-p) \sum_{i=1}^n \mu_i^{(t)} x_i = p \sum_{i=1}^n \mu_i^{(t)} (1-x_i) \\
 \Leftrightarrow \sum_{i=1}^n \mu_i^{(t)} x_i &= p \left(\sum_{i=1}^n \mu_i^{(t)} (1-x_i) + \sum_{i=1}^n \mu_i^{(t)} x_i \right) \Leftrightarrow \sum_{i=1}^n \mu_i^{(t)} x_i = p \sum_{i=1}^n \mu_i^{(t)} \\
 \Rightarrow p^{(t+1)} &= \frac{\sum_{i=1}^n \mu_i^{(t)} x_i}{\sum_{i=1}^n \mu_i^{(t)}} \in [0, 1]. \tag{394}
 \end{aligned}$$

Regula de actualizare pentru parametrul q este:

$$\frac{\partial Q(\theta | \theta^{(t)})}{\partial q} = 0 \Leftrightarrow \sum_{i=1}^n (1 - \mu_i^{(t)}) \left(\frac{x_i}{q} - \frac{1-x_i}{1-q} \right) = 0.$$

Se observă similaritatea cu derivata parțială în raport cu parametrul p , care a fost calculată mai sus. Așadar, rezultă:

$$q^{(t+1)} = \frac{\sum_{i=1}^n (1 - \mu_i^{(t)}) x_i}{\sum_{i=1}^n (1 - \mu_i^{(t)})} \in [0, 1]. \quad (395)$$

Regula de actualizare pentru parametrul π se obține astfel:

$$\begin{aligned} \frac{\partial Q(\theta | \theta^{(t)})}{\partial \pi} &= 0 \Leftrightarrow \sum_{i=1}^n \left(\frac{\mu_i^{(t)}}{\pi} - \frac{1 - \mu_i^{(t)}}{1 - \pi} \right) = 0 \\ &\Leftrightarrow \frac{1}{\pi} \sum_{i=1}^n \mu_i^{(t)} = \frac{1}{1 - \pi} \sum_{i=1}^n (1 - \mu_i^{(t)}) \Leftrightarrow (1 - \pi) \sum_{i=1}^n \mu_i^{(t)} = \pi \sum_{i=1}^n (1 - \mu_i^{(t)}) \\ &\Leftrightarrow \sum_{i=1}^n \mu_i^{(t)} = \pi \left(\sum_{i=1}^n (1 - \mu_i^{(t)}) + \sum_{i=1}^n \mu_i^{(t)} \right) \Leftrightarrow \sum_{i=1}^n \mu_i^{(t)} = \pi \sum_{i=1}^n 1 \Leftrightarrow \sum_{i=1}^n \mu_i^{(t)} = n\pi \\ &\Rightarrow \pi^{(t+1)} = \frac{1}{n} \sum_{i=1}^n \mu_i^{(t)} \in [0, 1]. \end{aligned} \quad (396)$$

Se observă că toate cele trei derivate parțiale de ordin secund ($\frac{\partial^2}{\partial p^2}$, $\frac{\partial^2}{\partial q^2}$ și $\frac{\partial^2}{\partial \pi^2}$) au doar valori negative pe domeniile lor de definiție. De asemenea, matricea hessiană (adică, matricea derivatelor parțiale de ordin secund) pentru funcția Q este diagonală și — datorită proprietății precedente — este negativ definită. În consecință, cele trei soluții aflate mai sus ($p^{(t+1)}$, $q^{(t+1)}$ și $\pi^{(t+1)}$) corespund punctului de maxim al funcției „auxiliare“ Q .

Observație: Se constată relativ ușor că formulele (394), (395) și (396) sunt variante probabiliste pentru formulele care ne-ar da estimările de verosimilitate maximă (MLE) pentru parametrii p , q și π în cazul în care valorile variabilelor z_i ar fi cunoscute.

f. Folosind rezultatele de la punctele b și e, putem scrie imediat pseudo-codul algoritmului EM pentru rezolvare de mixturi de distribuții Bernoulli.

Initializare:

atribuie valori arbitrarе ($\pi^{(0)}, p^{(0)}, q^{(0)}$ în intervalul $(0, 1)$)
pentru parametrii π , p și respectiv q ;

Corpul iterativ:

pentru $t = 0, \dots, T - 1$ (cu T fixat în avans)

(sau: până când log-verosimilitatea datelor observabile nu mai crește semnificativ),
(sau: până când $|\pi^{(t)} - \pi^{(t+1)}| < \varepsilon$, $|p^{(t)} - p^{(t+1)}| < \varepsilon$, $|q^{(t)} - q^{(t+1)}| < \varepsilon$,
cu ε fixat în avans)

execută

Pasul E: pentru $i = 1, \dots, n$, calculează

$$\mu_i^{(t)} = \frac{(p^{(t)})^{x_i} \cdot (1 - p^{(t)})^{1 - x_i} \cdot \pi^{(t)}}{(p^{(t)})^{x_i} \cdot (1 - p^{(t)})^{1 - x_i} \cdot \pi^{(t)} + (q^{(t)})^{x_i} \cdot (1 - q^{(t)})^{1 - x_i} \cdot (1 - \pi^{(t)})};$$

Pasul M:

$$p^{(t+1)} = \frac{\sum_{i=1}^n \mu_i^{(t)} x_i}{\sum_{i=1}^n \mu_i^{(t)}}, \quad q^{(t+1)} = \frac{\sum_{i=1}^n (1 - \mu_i^{(t)}) x_i}{\sum_{i=1}^n (1 - \mu_i^{(t)})}, \quad \pi^{(t+1)} = \frac{1}{n} \sum_{i=1}^n \mu_i^{(t)};$$

Returnează $\pi^{(t+1)}, p^{(t+1)}, q^{(t+1)}$;

La execuția primei iterării a algoritmului EM pe datele $x = \{1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1\}$, făcând calculele se constată ușor că la pasul E obținem

$$\mu_1^{(1)} = \mu_2^{(1)} = \mu_4^{(1)} = \mu_7^{(1)} = \mu_{11}^{(1)} = \mu_{12}^{(1)} = \frac{1}{3} \text{ și } \mu_3^{(1)} = \mu_5^{(1)} = \mu_6^{(1)} = \mu_8^{(1)} = \mu_9^{(1)} = \mu_{10}^{(1)} = \frac{2}{3},$$

iar la pasul M

$$p^{(1)} = \frac{1}{3} = p^{(0)}, q^{(1)} = \frac{2}{3} = q^{(0)} \text{ și } \pi^{(1)} = \frac{1}{2} = \pi^{(0)}.$$

Așadar, inițializând parametrii p, q și π cu valorile $p^{(0)}, q^{(0)}$ și respectiv $\pi^{(0)}$, algoritmul EM va bucla!

Observație: Sunt posibile două extensii relativ simple ale algoritmului EM care a fost dedus în această problemă. Vedeți problema 25.

7. (Mixturi de distribuții Bernoulli, cu și fără variabile neobservabile: un exemplu de rezolvare, folosind un set simplu de date; deducerea regulilor de actualizare pentru două variante ale algoritmului EM; aplicare)

■ □ L. Ciortuz, S. Ciobanu, 2020, folosind datele de la CMU, 2005 fall, T. Mitchell, A. Moore, midterm, pr. 1.3

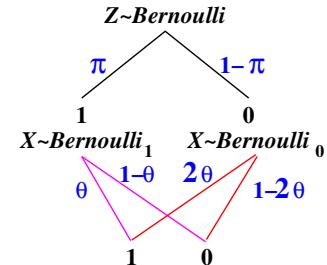
Comentariu:

[Estimarea parametrului unei mixturi de distribuții Bernoulli, în sens MLE]

În secțiunile A și B, care urmează mai jos, vom face o „revizitare“ a problemei 42 de la capitolul de *Fundamente* — ea oferă un exemplu simplu de estimare a unui parametru probabilist în sensul verosimilității maxime —, analizând acum datele *în mod explicit* din perspectiva unei mixturi de distribuții Bernoulli,⁹⁵³ și anume

$$X \sim \pi \text{Bernoulli}(\theta) + (1 - \pi) \text{Bernoulli}(2\theta), \quad (397)$$

sau, altfel spus,⁹⁵⁴



$(X|Z=1) \sim \text{Bernoulli}(\theta)$ și $(X|Z=0) \sim \text{Bernoulli}(2\theta)$, unde $Z \sim \text{Bernoulli}(\pi)$,

cu $\pi \in (0, 1)$ fixat. Este util să precizăm (și să comentăm) următoarele două situații posibile, în funcție de cum este variabila Z , *observabilă* sau *neobservabilă*.

A. [MLE, când toate datele sunt observabile]

Considerând variabila Z *observabilă*, putem scrie nu doar una ci două funcții de log-verosimilitate și apoi putem (de fapt, în consecință, trebuie) să arătăm care este relația dintre aceste două funcții.

	Moneda	Rezultat
i	Z_i	X_i
1	1	1 (stemă)
2	0	0 (ban)
3	0	0 (ban)
4	0	0 (ban)
5	0	1 (stemă)

⁹⁵³Problema 29 de la capitolul de *Fundamente* exemplifică noțiunea de *mixtură de distribuții categoriale*, iar problemele 113 și 114 de la același capitol exemplifică noțiunea de *mixtură de distribuții Bernoulli*.

⁹⁵⁴Veți observa că, pentru a obține o formă mai convenabilă la calculele care urmează, moneda a doua a fost asociată / desemnată cu $Z = 0$.

Funcția de log-verosimilitate a datelor comune („complete“) este următoarea:

$$\begin{aligned}\ell_{compl}(\theta) &\stackrel{not.}{=} \ln P(X, Z|\theta) \stackrel{indep.}{=} \ln \prod_{i=1}^5 P(X_i, Z_i|\theta) = \\ &\ln \prod_{i=1}^5 P(X_i|Z_i, \theta) \cdot P(Z_i|\theta) = \sum_{i=1}^5 [\ln P(X_i|Z_i, \theta) + \ln P(Z_i|\theta)] = \\ &\ln P(X_1 = 1|Z_1 = 1, \theta) + \ln P(Z_1 = 1|\theta) + \ln P(X_2 = 0|Z_2 = 0, \theta) + \ln P(Z_2 = 0|\theta) + \\ &\ln P(X_3 = 0|Z_3 = 0, \theta) + \ln P(Z_3 = 0|\theta) + \ln P(X_4 = 0|Z_4 = 0, \theta) + \ln P(Z_4 = 0|\theta) + \\ &\ln P(X_5 = 1|Z_5 = 0, \theta) + \ln P(Z_5 = 0|\theta) = \\ &\ln \theta + \ln \pi + 3(\ln(1 - 2\theta) + \ln(1 - \pi)) + \ln(2\theta) + \ln(1 - \pi) = \\ &\ln \theta + 3 \ln(1 - 2\theta) + \ln(2\theta) + \ln \pi + 4 \ln(1 - \pi).\end{aligned}$$

Funcția de log-verosimilitate a datelor condiționate este:

$$\ell_{cond}(\theta) \stackrel{not.}{=} \ln P(X|Z, \theta) \stackrel{indep.}{=} \ln \prod_{i=1}^5 P(X_i|Z_i, \theta) = \ln \theta + 3 \ln(1 - 2\theta) + \ln(2\theta).$$

identică cu funcția de log-verosimilitate care a fost calculată la punctul a al problemei 42 de la capitolul de *Fundamente*.

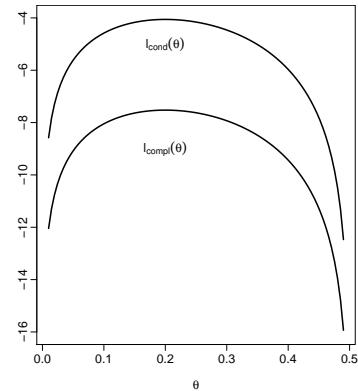
Presupunând că π este fixat (și nu depinde de θ), rezultă imediat că

$$\operatorname{argmax}_{\theta \in (0, 1/2)} \ell_{compl}(\theta) = \operatorname{argmax}_{\theta \in (0, 1/2)} \ell_{cond}(\theta).$$

pentru că $\ell_{compl}(\theta) = \ell_{cond}(\theta) + const.$

Prin urmare, în condițiile presupunerii de mai sus, $\hat{\theta}_{MLE}$, care este estimarea de verosimilitate maximă a parametrului θ poate fi calculat folosind oricare dintre cele două funcții de log-verosimilitate indicate mai sus.

Graficul alăturat a fost realizat pentru $\pi = 1/2$.



B. [MLE, în manieră analitică, în prezența unor date neobservabile⁹⁵⁵]

În cazul când variabila Z este *neobservabilă* (sau *latentă*, sau încă, *ascunsă*), vom opera cu log-verosimilitatea datelor „complete“ și cu log-verosimilitatea datelor „observabile“ (sau, „incomplete“). (*Observație*: Veți constata la secțiunea C că și algoritmul EM operează cu aceste două funcții de log-verosimilitate.)

	Moneda	Rezultat
i	Z_i	X_i
1	?	1 (stemă)
2	?	0 (ban)
3	?	0 (ban)
4	?	0 (ban)
5	?	1 (stemă)

Pentru prima dintre aceste două funcții de log-verosimilitate, deși definiția rămâne aceeași (ca mai sus), exprimarea concretă diferă, întrucât valorile variabilelor Z_i nu mai sunt observabile.⁹⁵⁶

⁹⁵⁵Ca și la secțiunea A, se va considera că valoarea lui π este fixată în intervalul $(0, 1)$.

⁹⁵⁶Observație: La cea de-a patra egalitate din deducerea expresiei pentru $\ell_{compl}(\theta)$ am folosit un mod de scriere compactă a p.m.f. pentru distribuția Bernoulli. Este vorba despre *artificiul ridicării la putere* (engl., the exponentiation trick).

$$\begin{aligned}
\ell_{compl}(\theta) &\stackrel{not.}{=} \ln P(X = (x_1, \dots, x_5), Z = (z_1, \dots, z_5) | \theta) \\
&\stackrel{indep.}{=} \ln \prod_{i=1}^5 P(X_i = x_i, Z_i = z_i | \theta) = \sum_{i=1}^5 \ln [P(X_i = x_i | Z_i = z_i, \theta) \cdot P(Z_i = z_i | \theta)] \\
&= \ln((\theta\pi)^{z_1}(2\theta(1-\pi))^{1-z_1}) + \ln(((1-\theta)\pi)^{z_2}((1-2\theta)(1-\pi))^{1-z_2}) + \\
&\quad \ln(((1-\theta)\pi)^{z_3}((1-2\theta)(1-\pi))^{1-z_3}) + \ln(((1-\theta)\pi)^{z_4}((1-2\theta)(1-\pi))^{1-z_4}) + \\
&\quad \ln((\theta\pi)^{z_5}(2\theta(1-\pi))^{1-z_5}) \\
&= z_1 \ln(\theta\pi) + (1-z_1) \ln(2\theta(1-\pi)) + \\
&\quad z_2 \ln((1-\theta)\pi) + (1-z_2) \ln((1-2\theta)(1-\pi)) + \\
&\quad z_3 \ln((1-\theta)\pi) + (1-z_3) \ln((1-2\theta)(1-\pi)) + \\
&\quad z_4 \ln((1-\theta)\pi) + (1-z_4) \ln((1-2\theta)(1-\pi)) + \\
&\quad z_5 \ln(\theta\pi) + (1-z_5) \ln(2\theta(1-\pi)). \tag{398}
\end{aligned}$$

Pe lângă faptul că această scriere a funcției de log-verosimilitate a datelor „complete“ este *generală* în raport cu diferite valori posibile ale variabilelor Z_i , observăm că ea mai are un avantaj: ea ne permite să scriem într-un mod foarte convenabil *media* funcției de log-verosimilitate a datelor „complete“, aplicând proprietatea de liniaritate a mediilor,⁹⁵⁷ adică pur și simplu înlocuind fiecare z_i din ultima expresie de mai sus cu $E[Z_i]$.⁹⁵⁸

În ce privește log-verosimilitatea datelor „observable“ (sau, „incomplete“) X_i , atunci când variabilele Z_i sunt *neobservable*, putem scrie:

$$\ell_{obs}(\theta) \stackrel{def.}{=} \ln P(X | \theta) = \ln \sum_Z P(X, Z | \theta),$$

înțelegând prin simbolul \sum_Z că se însumează probabilitățile $P(X, Z | \theta)$ obținute pentru toate asignările posibile ale variabilelor Z . Concret, pentru exemplul de mai sus (când, precizăm din nou, variabilele Z_i sunt neobservabile), log-verosimilitatea datelor observabile se calculează astfel:

$$\begin{aligned}
\ell_{obs}(\theta) &\stackrel{def.}{=} \ln P(X | \theta) \stackrel{indep.}{=} \ln \prod_{i=1}^5 P(X_i | \theta) = \sum_{i=1}^5 \ln P(X_i | \theta) = \sum_{i=1}^5 \ln \left(\sum_{Z_i \in \{0,1\}} P(X_i, Z_i | \theta) \right) \\
&= \sum_{i=1}^5 \ln \left(\sum_{Z_i \in \{0,1\}} P(X_i | Z_i, \theta) \cdot P(Z_i | \theta) \right) \\
&= 2 \ln(\pi\theta + (1-\pi)2\theta) + 3 \ln(\pi(1-\theta) + (1-\pi)(1-2\theta)).
\end{aligned}$$

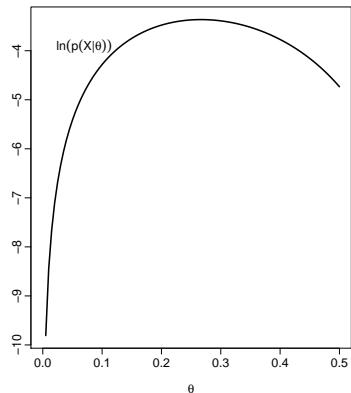
⁹⁵⁷Vedeți problema 9 de la capitolul de *Fundamente*.

⁹⁵⁸Vă vom preciza / indica la momentul corespunzător care anume este distribuția de probabilitate în raport cu care vom calcula această medie.

Valoarea parametrului θ pentru care se atinge minimul acestei funcții se obține rezolvând ecuația $\ell'_{obs}(\theta) = 0$, care depinde de valoarea care a fost atribuită lui π . De exemplu, pentru $\pi = 1/2$ graficul funcției $\ell_{obs}(\theta)$ este prezentat în figura alăturată și vom avea:

$$\begin{aligned}\ell'_{obs}(\theta) = 0 \Leftrightarrow \frac{2}{\theta} = \frac{9}{2 - 3\theta} \Leftrightarrow 4 - 6\theta = 9\theta \Leftrightarrow \\ \theta = \frac{4}{15} = 0.2(6).\end{aligned}$$

Se observă că această valoare aparține intervalului $(0, 1/2)$, deci este validă.



Similar, pentru $\pi = 1/5$, care este chiar probabilitatea (estimată în sensul verosimilității maxime) de alegere a monedei 1 în experimentul aleator din prima parte a acestei probleme, se obține soluția $\theta = 2/9$. Remarcăm faptul că ea *diferă!* ușor de $\hat{\theta}_{MLE} = 1/5$, estimarea de verosimilitate maximă care a fost obținută atunci când s-a considerat că variabilele Z_i sunt observabile.

Observație: Pentru aceste date — adică, pentru această mixtură de distribuții Bernoulli — am putut calcula în mod direct (folosind o *formulă analitică*) maximul funcției de log-verosimilitate a datelor „observabile“. Însă pentru foarte multe probleme de rezolvare de mixturi de distribuții probabiliste nu există o astfel de posibilitate. *Alternativa* este să folosim un algoritm de optimizare [iterativă], precum algoritmul EM (sau algoritmul gradientului ascendent dacă funcția ℓ_{obs} este derivabilă sau, încă, metoda lui Newton dacă funcția ℓ_{obs} este dublu derivabilă; vedeti pr. 80 de la capitolul de *Fundamente*).

C. [MLE, în prezența unor date neobservabile, folosind algoritmul EM]

Elaborați regulile de actualizare pentru pașii E și M ai algoritmului EM și apoi aplicați-l la rezolvarea unei de mixturi de distribuții Bernoulli,⁹⁵⁹ conform datelor din secțiunea B.

Vă cerem să elaborați *două variante* ale acestui algoritm EM:

- La acest punct veți considera că probabilitatea de selecție π este un număr fixat (în intervalul $(0, 1)$; aşadar, în acest caz singurul parametru (liber) al algoritmului EM este θ . Vă cerem să executați primele două iterații ale acestui algoritm, pentru cazul când $\pi = 1/2$, atribuind lui θ valoarea inițială $\theta^{(0)} = 0.45$. Veți reprezenta grafic (folosind în acest scop R, Matlab sau limbajul dumneavoastră preferat) atât funcția de log-verosimilitate a datelor observabile, $\ln P(x_1, \dots, x_5 | \theta)$, cât și funcțiile auxiliare $Q(\theta | \theta^{(0)})$ și $Q(\theta | \theta^{(1)})$.⁹⁶⁰ Marcați pe grafic punctele de maxim ale funcțiilor auxiliare.
- De adata aceasta veți considera că probabilitatea de selecție π este — și ea — parametru liber (de asemenea în intervalul $(0, 1)$; aşadar, în acest caz, parametrii algoritmului EM sunt θ și π .

⁹⁵⁹Puteți lucra folosind formularea schemei algoritmice EM din *Observația* (5) de la problema 1. Alternativ, puteți adapta algoritmul EM general pentru rezolvarea de mixturi de distribuții Bernoulli, al cărui pseudo-cod este dat la finalul soluției de la problema 6; în principiu, pentru ceea ce este necesar aici, veți ignora regulile de actualizare pentru parametrii q și π de acolo.

⁹⁶⁰Aceste două funcții auxiliare sunt mediile unor funcții de log-verosimilitate pentru datele „complete“ (adică, atât pentru datele „observabile“ cât și pentru datele „neobservabile“).

Răspuns:

Conform enunțului, vom lucra cu mixtura de distribuții Bernoulli

$$X \sim \pi \text{Bernoulli}(\theta) + (1 - \pi) \text{Bernoulli}(2\theta), \text{ cu } \theta \in (0, 1/2).$$

Altfel spus, considerând $Z \sim \text{Bernoulli}(\pi)$ **cu** $\pi \in (0, 1)$, **vom lucra cu** $X|(Z = 1) \sim \text{Bernoulli}(\theta)$ și $X|(Z = 0) \sim \text{Bernoulli}(2\theta)$.

a. La acest punct probabilitatea de selecție π **este considerată fixată.**

La pasul E calculăm mai întâi media variabilelor neobservabile Z_i în raport cu datele observabile x_i și cu valoarea curentă a parametrului θ , adică $\theta^{(t)}$.

$$\begin{aligned} E[z_i|x_i, \theta^{(t)}] &= P(z_i = 1|x_i, \theta^{(t)}) \\ &\stackrel{\text{F. Bayes}}{=} \frac{\pi(\theta^{(t)})^{x_i}(1 - \theta^{(t)})^{1-x_i}}{\pi(\theta^{(t)})^{x_i}(1 - \theta^{(t)})^{1-x_i} + (1 - \pi)(2\theta^{(t)})^{x_i}(1 - 2\theta^{(t)})^{1-x_i}} \stackrel{\text{not.}}{=} \mu_i^{(t)} \end{aligned} \quad (399)$$

Apoi calculăm log-verosimilitatea datelor complete, $\ell_{compl}(\theta)$, similar cu modul în care am procedat la relația (398).

$$\begin{aligned} \ell_{compl}(\theta) &\stackrel{\text{not.}}{=} \ln P(X, Z|\theta) \\ &= Z_1 \ln(\theta\pi) + (1 - Z_1) \ln(2\theta(1 - \pi)) + \\ &\quad Z_2 \ln((1 - \theta)\pi) + (1 - Z_2) \ln((1 - 2\theta)(1 - \pi)) + \\ &\quad Z_3 \ln((1 - \theta)\pi) + (1 - Z_3) \ln((1 - 2\theta)(1 - \pi)) + \\ &\quad Z_4 \ln((1 - \theta)\pi) + (1 - Z_4) \ln((1 - 2\theta)(1 - \pi)) + \\ &\quad Z_5 \ln(\theta\pi) + (1 - Z_5) \ln(2\theta(1 - \pi)). \end{aligned}$$

În sfârșit, vom calcula **funcția auxiliară** $Q(\theta|\theta^{(t)})$, folosind relația pe care tocmai am obținut-o și proprietatea de liniaritate a mediei (vedeți pr. 9.a de la capitolul de **Fundamente**).

$$\begin{aligned} Q(\theta|\theta^{(t)}) &\stackrel{\text{not.}}{=} E[\ell_{compl}(\theta)] \\ &= E[Z_1] \ln(\theta\pi) + (1 - E[Z_1]) \ln(2\theta(1 - \pi)) + \\ &\quad E[Z_2] \ln((1 - \theta)\pi) + (1 - E[Z_2]) \ln((1 - 2\theta)(1 - \pi)) + \\ &\quad E[Z_3] \ln((1 - \theta)\pi) + (1 - E[Z_3]) \ln((1 - 2\theta)(1 - \pi)) + \\ &\quad E[Z_4] \ln((1 - \theta)\pi) + (1 - E[Z_4]) \ln((1 - 2\theta)(1 - \pi)) + \\ &\quad E[Z_5] \ln(\theta\pi) + (1 - E[Z_5]) \ln(2\theta(1 - \pi)) \quad (400) \\ &\stackrel{\pi \text{ const.}}{=} \text{const.} + E[Z_1] \ln \theta + (1 - E[Z_1]) \ln \theta + \\ &\quad E[Z_2] \ln(1 - \theta) + (1 - E[Z_2]) \ln(1 - 2\theta) + \\ &\quad E[Z_3] \ln(1 - \theta) + (1 - E[Z_3]) \ln(1 - 2\theta) + \\ &\quad E[Z_4] \ln(1 - \theta) + (1 - E[Z_4]) \ln(1 - 2\theta) + \\ &\quad E[Z_5] \ln \theta + (1 - E[Z_5]) \ln \theta \\ &= \text{const.} + 2 \ln \theta + \\ &\quad (E[Z_2] + E[Z_3] + E[Z_4]) \ln(1 - \theta) + (3 - (E[Z_2] + E[Z_3] + E[Z_4])) \ln(1 - 2\theta). \end{aligned}$$

La pasul M, trebuie să calculăm noua valoare a parametrului θ :

$$\theta^{(t+1)} \stackrel{\text{def.}}{=} \underset{\theta \in (0, 1/2)}{\operatorname{argmax}} Q(\theta|\theta^{(t)}).$$

Pentru aceasta, vom calcula derivata funcției $Q(\theta|\theta^{(t)})$ în raport cu θ și o vom egala cu 0:

$$\begin{aligned} \frac{\partial}{\partial \theta} Q(\theta|\theta^{(t)}) = 0 &\Leftrightarrow \frac{2}{\theta} = \frac{E[Z_2] + E[Z_3] + E[Z_4]}{1 - \theta} + \frac{2[3 - (E[Z_2] + E[Z_3] + E[Z_4])]}{1 - 2\theta} \\ &\Leftrightarrow 2(1 - \theta)(1 - 2\theta) = (E[Z_2] + E[Z_3] + E[Z_4])\theta(1 - 2\theta) + \\ &\quad 2(3 - (E[Z_2] + E[Z_3] + E[Z_4]))\theta(1 - \theta) \\ &\Leftrightarrow 2(1 - \theta)(1 - 2\theta) = (\mu_2^{(t)} + \mu_3^{(t)} + \mu_4^{(t)})\theta(1 - 2\theta) + 2(3 - (\mu_2^{(t)} + \mu_3^{(t)} + \mu_4^{(t)}))\theta(1 - \theta) \\ &\Leftrightarrow a\theta^2 + b\theta + c = 0, \text{ unde } a = 10, b = -12 + \mu_2^{(t)} + \mu_3^{(t)} + \mu_4^{(t)}, c = 2 \in \mathbb{R}. \end{aligned} \quad (401)$$

Remarcați faptul că

$$\frac{\partial^2}{\partial \theta^2} Q(\theta|\theta^{(t)}) = -\frac{2}{\theta^2} - \overbrace{\frac{\mu_2^{(t)} + \mu_3^{(t)} + \mu_4^{(t)}}{(1 - \theta)^2}}^{\geq 0} - \overbrace{\frac{4(3 - (\mu_2^{(t)} + \mu_3^{(t)} + \mu_4^{(t)}))}{(1 - 2\theta)^2}}^{\geq 0} < 0.$$

Așadar, $Q(\theta|\theta^{(t)})$ este o funcție strict concavă și, prin urmare, ea admite un punct unic de maxim în intervalul $(0, 1/2)$.

Întrucât $\mu_i^{(t)} \in [0, 1]$ conform relației (399), rezultă că $b \in [-12, -9]$, deci discriminantul ecuației de gradul al doilea din relația (401) este $\Delta = b^2 - 4ac = b^2 - 80 \geq 1 > 0$. Prin urmare, rădăcinile acestei ecuații (notate cu θ_1 și θ_2) sunt din multimea numerelor reale. Folosind cunoștințe de analiză matematică din liceu, se poate demonstra ușor că θ_1 și θ_2 sunt pozitive.⁹⁶¹ Mai mult, se poate arăta că una dintre aceste rădăcini este în intervalul $(0, 1/2)$, iar cealaltă rădăcină este mai mare decât $1/2$ sau egală cu $1/2$. În consecință, această ecuație de gradul al doilea are o singură rădăcină în intervalul $(0, 1/2)$, și anume cea mai mică dintre θ_1 și θ_2 .⁹⁶²

Sintetizând, putem scrie:

$$\theta^{(t+1)} = \frac{-b - \sqrt{b^2 - 4ac}}{2a}. \quad (402)$$

Regulile de actualizare pentru această [primă] versiune a algoritmului EM sunt deci relațiile (399) și (402). Așadar, putem scrie acum pseudo-codul algoritmului EM pentru rezolvarea acestei mixturi de distribuții Bernoulli.

Inițializare:

atribuie o valoare arbitrară $\theta^{(0)}$ în intervalul $(0, 1/2)$ pentru parametrul θ ;

Corpul iterativ:

pentru $t = 0, \dots, T - 1$ (cu T fixat în avans)

(sau: până când $|\theta^{(t)} - \theta^{(t+1)}| < \varepsilon$, cu ε fixat în avans) execută

Pasul E:

$$\mu_i^{(t)} \stackrel{\text{not.}}{=} \frac{\pi(\theta^{(t)})^{x_i}(1 - \theta^{(t)})^{1-x_i}}{\pi(\theta^{(t)})^{x_i}(1 - \theta^{(t)})^{1-x_i} + (1 - \pi)(2\theta^{(t)})^{x_i}(1 - 2\theta^{(t)})^{1-x_i}};$$

Pasul M:

$$\theta^{(t+1)} = \frac{-b - \sqrt{b^2 - 4ac}}{2a}, \text{ cu}$$

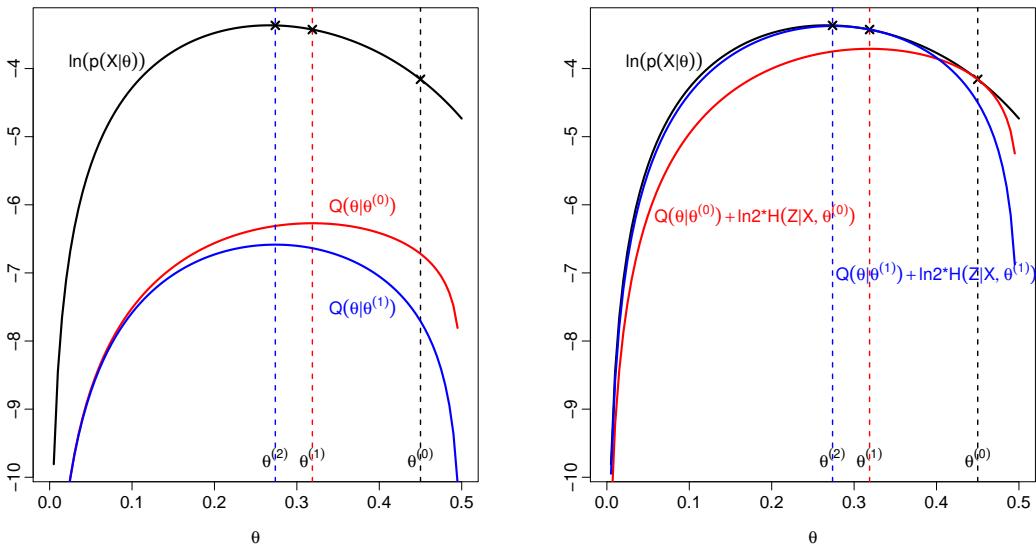
⁹⁶¹Fiindcă $S = -\frac{b}{a} > 0$ și $P = \frac{c}{a} > 0$, unde $S \stackrel{\text{not.}}{=} \theta_1 + \theta_2$ și $P \stackrel{\text{not.}}{=} \theta_1\theta_2$.

⁹⁶²Fiindcă se dovedește imediat că $P' \leq 0$, unde $P' \stackrel{\text{not.}}{=} (\theta_1 - \frac{1}{2}) \cdot (\theta_2 - \frac{1}{2})$.

$$a = 10, b = -12 + \mu_2^{(t)} + \mu_3^{(t)} + \mu_4^{(t)} \text{ și } c = 2;$$

Returnează $\theta^{(t+1)}$;

Graficul cerut în enunț — pentru primele două iterații ale acestui algoritm, pentru cazul când $\pi = 1/2$ și atribuind lui θ valoarea inițială $\theta^{(0)} = 0.45$ — este redat mai jos, în partea stângă. În graficul din partea dreaptă am adăugat la [fiecare dintre] funcțiile auxiliare entropia distribuției a posteriori a variabilelor neobservabile, conform proprietății (387) din problema 2.⁹⁶³



Menționăm că am obținut $\theta^{(1)} = 0.3187987$ și $\theta^{(2)} = 0.2738492$, iar valorile log-verosimilității datelor observabile ℓ pentru aceste valori ale lui θ sunt -3.426862 și respectiv -3.366261 (pentru $\theta^{(0)}$, ea a fost -4.157875). Pentru $t \geq 5$, se obține $\theta^{(t)} \approx 0.266 \approx 0.2(6)$, cât a fost calculat la secțiunea B!

Observație importantă (1):

Acum dispunem de un *model probabilist* (de tip MLE), mai exact o mixtură de distribuții Bernoulli, pentru datele din tabelul dat la începutul enunțului secțiunii B. Concret, acest model este cel obținut înlocuind în formula (397) parametrul θ cu $\theta^{(2)}$ și, bineînțeles, π cu $1/2$.

Ce putem face cu acest model?

Putem formula „cunoștințe“ noi, în locul semnelor de întrebare din tabelul menționat. Anume, pentru fiecare x_i , cu $i = 1, \dots, 5$, putem spune care este probabilitatea ca — în modelul obținut — x_i să fi fost generat de către prima distribuție Bernoulli (adică, prima monedă) din mixtura (397) și, complementar, care este probabilitatea ca x_i să fi fost generat de către a doua distribuție Bernoulli (adică, a doua monedă) din mixtura respectivă.

⁹⁶³Pe aceste două grafice se observă ușor „tinderea“ valorilor $\theta^{(t)}$ către abscisa punctului de maxim al funcției de log-verosimilitate a datelor „observabile“, $\ln P(X|\theta)$.

Așadar, pentru x_1 și x_5 (ambele fiind 1, adică 'stemă'), avem:

$$\begin{aligned} P(Z_i = 1|X_i = 1, \theta^{(2)}) &= E(Z_i|X_i = 1, \theta^{(2)}) = \mu_i^{(2)} \\ &= \frac{P(X_i = 1|Z_i = 1, \theta^{(2)}) \cdot \pi}{P(X_i = 1|Z_i = 1, \theta^{(2)}) \cdot \pi + P(X_i = 1|Z_i = 0, \theta^{(2)}) \cdot (1 - \pi)} = \frac{\theta^{(2)} \cdot \frac{1}{2}}{\theta^{(2)} \cdot \frac{1}{2} + 2\theta^{(2)} \cdot \frac{1}{2}} \\ &= \frac{1}{3}. \end{aligned}$$

Aceasta este probabilitatea ca x_1 și respectiv x_5 să fie în clusterul (de aruncări / „observații“) corespunzător primei monede.

Similar, pentru x_2 , x_3 și x_4 (toate fiind 0, adică 'ban'), avem:

$$\begin{aligned} P(Z_i = 1|X_i = 0, \theta^{(2)}) &= E(Z_i|X_i = 0, \theta^{(2)}) = \mu_i^{(2)} \\ &= \frac{P(X_i = 0|Z_i = 1, \theta^{(2)}) \cdot \pi}{P(X_i = 0|Z_i = 1, \theta^{(2)}) \cdot \pi + P(X_i = 0|Z_i = 0, \theta^{(2)}) \cdot (1 - \pi)} \\ &= \frac{(1 - \theta^{(2)}) \cdot \frac{1}{2}}{(1 - \theta^{(2)}) \cdot \frac{1}{2} + (1 - 2\theta^{(2)}) \cdot \frac{1}{2}} = \frac{1 - \theta^{(2)}}{2 - 3\theta^{(2)}} \approx 0.6162. \end{aligned}$$

Aceasta este probabilitatea ca x_2 , x_3 și respectiv x_4 să fie în clusterul (de „observații“) corespunzător primei monede.

Sumarizând, în tabelul de la începutul acestei secțiuni putem înlocui acum coloana Z_i (în care apar doar semnele '?') cu două coloane care corespund probabilităților $P(Z_i = 1|X_i = x_i, \theta^{(2)})$ și $P(Z_i = 0|X_i = x_i, \theta^{(2)})$. Aceste probabilități cuantifică numeric apartenența "soft" a instanțelor x_i la clusterul reprezentat de prima monedă și respectiv la clusterul reprezentat de cea de-a doua monedă.

i	$P(Z_i = 1 X_i, \theta^{(2)})$	$P(Z_i = 0 X_i, \theta^{(2)})$	X_i
1	1/3	2/3	1 (stemă)
2	0.6162	0.3838	0 (ban)
3	0.6162	0.3838	0 (ban)
4	0.6162	0.3838	0 (ban)
5	1/3	2/3	1 (stemă)

Dacă, în schimb, dorim să facem o asignare "hard" a instanțelor x_1, \dots, x_5 la aceste două clustere, atunci vom proceda folosind *regula de decizie* de la clasificarea Bayesiană și vom obține că primul cluster va fi format din instanțele x_2 , x_3 și x_4 , iar al doilea cluster va fi format din instanțele x_1 și x_5 .

b. Aşa cum s-a specificat în enunț, aici vom elabora — în vederea aplicării pe datele noastre — varianta de algoritm EM pentru rezolvarea unei mixturi de două distribuții Bernoulli pentru cazul când parametrul π , care reprezintă *probabilitatea de selecție* (a primei componente dintre cele două ale mixturii), este lăsat liber în intervalul $(0, 1)$. Similar cu modul în care am procedat la punctul a, vom elabora calculele necesare pentru obținerea *formulelor de actualizare* care vor fi folosite la cei doi pași ai algoritmului EM, în noile condiții.

Pasul E:

Calculăm media variabilelor neobservabile z_i în raport cu x_i și $\theta^{(t)}$.

$$\begin{aligned} E[z_i|x_i, \theta^{(t)}, \pi^{(t)}] &= P(z_i = 1|x_i, \theta^{(t)}, \pi^{(t)}) \\ &\stackrel{F. Bayes}{=} \frac{\pi^{(t)} (\theta^{(t)})^{x_i} (1 - \theta^{(t)})^{1-x_i}}{\pi^{(t)} (\theta^{(t)})^{x_i} (1 - \theta^{(t)})^{1-x_i} + (1 - \pi^{(t)}) (2\theta^{(t)})^{x_i} (1 - 2\theta^{(t)})^{1-x_i}} \stackrel{not.}{=} \mu_i^{(t)} \quad (403) \end{aligned}$$

și apoi funcția auxiliară $Q(\theta, \pi|\theta^{(t)}, \pi^{(t)})$, folosind proprietatea de liniaritate a mediei și obținând o relație similară cu relația (400):⁹⁶⁴

$$\begin{aligned} Q(\theta, \pi|\theta^{(t)}, \pi^{(t)}) &\stackrel{not.}{=} E[\ell_{compl}(\theta, \pi)] \\ &= E[Z_1] \ln(\theta\pi) + (1 - E[Z_1]) \ln(2\theta(1 - \pi)) + \\ &\quad E[Z_2] \ln((1 - \theta)\pi) + (1 - E[Z_2]) \ln((1 - 2\theta)(1 - \pi)) + \\ &\quad E[Z_3] \ln((1 - \theta)\pi) + (1 - E[Z_3]) \ln((1 - 2\theta)(1 - \pi)) + \\ &\quad E[Z_4] \ln((1 - \theta)\pi) + (1 - E[Z_4]) \ln((1 - 2\theta)(1 - \pi)) + \\ &\quad E[Z_5] \ln(\theta\pi) + (1 - E[Z_5]) \ln(2\theta(1 - \pi)). \end{aligned}$$

Pasul M:

$\theta^{(t+1)}$ va fi calculat folosind aceeași regulă de actualizare ca la punctul a , iar în ce privește $\pi^{(t+1)}$, reamintim că $\pi^{(t+1)} \stackrel{def.}{=} \text{argmax}_{\pi \in (0,1)} Q(\theta, \pi|\theta^{(t)}, \pi^{(t)})$, deci îl vom putea calcula egalând cu 0 derivata parțială a lui $Q(\theta, \pi|\theta^{(t)}, \pi^{(t)})$ în raport cu π .

$$\begin{aligned} \frac{\partial}{\partial \pi} Q(\theta, \pi|\theta^{(t)}, \pi^{(t)}) = 0 &\Leftrightarrow \\ \frac{\partial}{\partial \pi} [const. + (E[Z_1] + E[Z_2] + E[Z_3] + E[Z_4] + E[Z_5]) \ln \pi + \\ (5 - (E[Z_1] + E[Z_2] + E[Z_3] + E[Z_4] + E[Z_5])) \ln(1 - \pi)] &= 0 \Leftrightarrow \\ \frac{E[Z_1] + E[Z_2] + E[Z_3] + E[Z_4] + E[Z_5]}{\pi} - \frac{5 - (E[Z_1] + E[Z_2] + E[Z_3] + E[Z_4] + E[Z_5])}{1 - \pi} &= 0 \\ \Leftrightarrow E[Z_1] + E[Z_2] + E[Z_3] + E[Z_4] + E[Z_5] &= 5\pi \\ \Leftrightarrow \pi &= \frac{1}{5}(E[Z_1] + E[Z_2] + E[Z_3] + E[Z_4] + E[Z_5]). \end{aligned}$$

Așadar,

$$\pi^{(t+1)} = \frac{1}{5}(\mu_1^{(t)} + \mu_2^{(t)} + \mu_3^{(t)} + \mu_4^{(t)} + \mu_5^{(t)}) \in [0, 1]. \quad (404)$$

Remarcați faptul că $Q(\theta, \pi|\theta^{(t)}, \pi^{(t)})$ este funcție concavă în raport cu parametrul π .⁹⁶⁵ Remarcați de asemenea că formula (404) este varianta probabilistă a formulei care ne-ar da estimarea de verosimilitate maximă (MLE) pentru π în cazul în care toate datele ar fi observabile.

Regulile de actualizare pentru această nouă versiune (și anume, a doua) a algoritmului EM sunt relațiile (403) pentru pasul E, și (402) și (404) pentru

⁹⁶⁴Veți observa că mediile variabilelor neobservabile, $E[Z_i]$, se calculează aici în funcție de $\theta^{(t)}$ și $\pi^{(t)}$, în vreme ce la relația (400) ele se calculau doar în funcție de $\theta^{(t)}$.

⁹⁶⁵De fapt, mai mult, este ușor de arătat că matricea hessiană a lui $Q(\theta, \pi|\theta^{(t)}, \pi^{(t)})$ este negativ definită, deci Q admite un singur punct de maxim.

pasul M.⁹⁶⁶ Așadar, putem scrie acum pseudo-codul acestei noi variante a algoritmului EM pentru rezolvarea mixturii de distribuții Bernoulli din enunț.

Inițializare:

atribuie o valoare arbitrară $\theta^{(0)}$ în intervalul $(0, 1/2)$ pentru parametrul θ și o valoare arbitrară $\pi^{(0)}$ în intervalul $(0, 1)$ pentru parametrul π ;

Corpul iterativ:

pentru $t = 0, \dots, T - 1$ (cu T fixat în avans)

(sau: până când $|\theta^{(t)} - \theta^{(t+1)}| < \varepsilon$, cu ε fixat în avans) execută

Pasul E:

$$\mu_i^{(t)} \stackrel{\text{not.}}{=} \frac{\pi^{(t)} (\theta^{(t)})^{x_i} (1 - \theta^{(t)})^{1-x_i}}{\pi^{(t)} (\theta^{(t)})^{x_i} (1 - \theta^{(t)})^{1-x_i} + (1 - \pi^{(t)}) (2\theta^{(t)})^{x_i} (1 - 2\theta^{(t)})^{1-x_i}};$$

Pasul M:

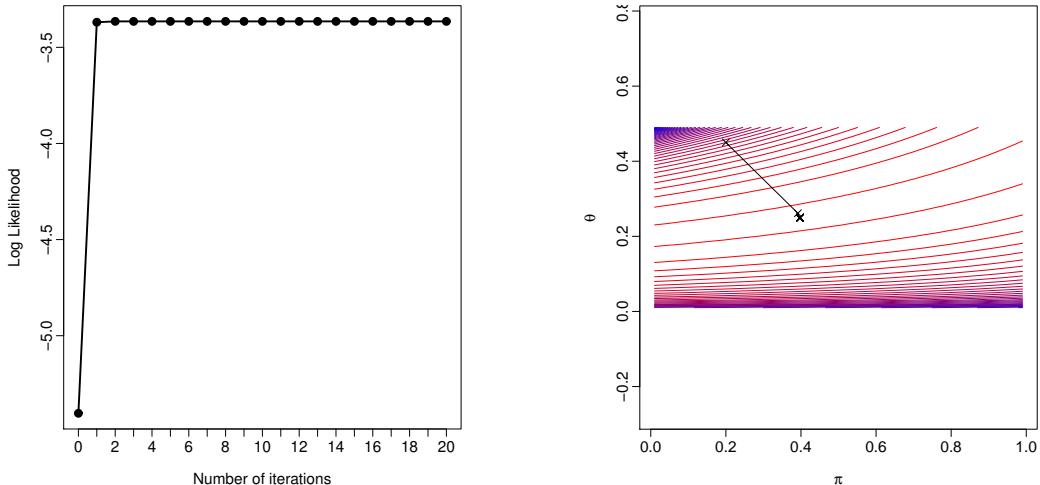
$$\theta^{(t+1)} = \frac{-b - \sqrt{b^2 - 4ac}}{2a}, \text{ cu}$$

$$a = 10, b = -12 + \mu_2^{(t)} + \mu_3^{(t)} + \mu_4^{(t)} \text{ și } c = 2;$$

$$\pi^{(t+1)} = \frac{1}{5}(\mu_1^{(t)} + \mu_2^{(t)} + \mu_3^{(t)} + \mu_4^{(t)} + \mu_5^{(t)}) \in [0, 1];$$

Returnează $\theta^{(t+1)}$ și $\pi^{(t+1)}$;

Graficele de mai jos reprezintă evoluția valorilor funcției de log-verosimilitate a datelor observabile în timpul primelor 20 de iterări ale acestei versiuni de algoritm EM, folosind aceleași date și aceleași inițializări ca la punctul precedent și, în plus, $\pi^{(0)} = 0.2$. În graficul din partea stângă se observă o creștere substanțială a valorilor acestei funcții chiar de la prima iterărie, după care, în câteva iterări, se atinge un platou. Graficul din partea dreaptă folosește curbe de izocontur pentru a arăta evoluția parametrilor θ și π în cursul execuției algoritmului EM.



Observație importantă (2) — pentru cazul când π este fixat

Datorită proprietății de *liniaritate a mediei*, urmează că media sumei unor variabile aleatoare este suma mediilor lor. Folosind această proprietate simplă, vom arăta acum că putem să reformulăm algoritmul nostru EM pentru

⁹⁶⁶Atenție, în legătură cu regula de actualizare (402) de folosit aici: ea a fost dedusă din ecuația (401), în care apar mediile $\mu_i^{(t)}$. Pentru prezenta versiune a algoritmului EM, aceste medii sunt calculate cu relația (403).

estimarea parametrilor unei mixturi de distribuții Bernoulli sub o *altă formă*, foarte convenabilă, folosind două distribuții binomiale, similare cu cele de la problema 5.

La rezolvarea dată la punctul *a* pentru *pasul E* (așadar, când probabilitatea de selecție, π , era considerată fixată), am obținut

$$\mu_i^{(t)} = P(z_i = 1|x_i, \theta^{(t)})^F \stackrel{\text{Bayes}}{=} \frac{\pi(\theta^{(t)})^{x_i}(1 - \theta^{(t)})^{1-x_i}}{\pi(\theta^{(t)})^{x_i}(1 - \theta^{(t)})^{1-x_i} + (1 - \pi)(2\theta^{(t)})^{x_i}(1 - 2\theta^{(t)})^{1-x_i}}.$$

Putem particulariza pentru *cazul* $x_i = 1$ (adică, pentru $x_i \equiv stemă$):

$$q_1^{(t)} \stackrel{\text{not.}}{=} P(z_i = 1|x_i = 1, \theta^{(t)}) = \frac{\pi \theta^{(t)}}{\pi \theta^{(t)} + (1 - \pi) 2\theta^{(t)}}, \quad (405)$$

ceea ce reprezintă probabilitatea a posteriori ca dacă la o aruncare oarecare a apărut fața *stemă*, ea să fi fost obținută cu prima monedă.

Similar, pentru *cazul* $x_i = 0$ (adică, pentru $x_i \equiv ban$):

$$q_0^{(t)} \stackrel{\text{not.}}{=} P(z_i = 1|x_i = 0, \theta^{(t)}) = \frac{\pi(1 - \theta^{(t)})}{\pi(1 - \theta^{(t)}) + (1 - \pi)(1 - 2\theta^{(t)})}. \quad (406)$$

ceea ce reprezintă probabilitatea a posteriori ca dacă la o aruncare oarecare a apărut fața *ban*, ea să fi fost obținută cu prima monedă.

Fie \tilde{z}_1 numărul (neobservabil) de apariții ale feței *stemă* pentru moneda 1, din totalul de $n_S \stackrel{\text{not.}}{=} 2$ apariții ale stemei în datele noastre.⁹⁶⁷

Similar, fie \tilde{z}_0 numărul (neobservabil) de apariții ale feței *ban* pentru moneda 1, din totalul de $n_B \stackrel{\text{not.}}{=} 3$ apariții ale banului în datele noastre.

Funcția de verosimilitate care corespunde datelor complete din tabelul prezentat la secțiunea B⁹⁶⁸ este:⁹⁶⁹

$$L_{compl}(\theta) = \theta^{\tilde{z}_1}(1 - \theta)^{\tilde{z}_0} \cdot (2\theta)^{n_S - \tilde{z}_1}(1 - 2\theta)^{n_B - \tilde{z}_0},$$

iar funcția de log-verosimilitate (tot a datelor complete) este

$$\ell_{compl}(\theta) = \tilde{z}_1 \ln \theta + \tilde{z}_0 \ln(1 - \theta) + (n_S - \tilde{z}_1) \ln(2\theta) + (n_B - \tilde{z}_0) \ln(1 - 2\theta).$$

Aceasta ne conduce imediat la *noua expresie a funcției auxiliare*:

$$\begin{aligned} Q(\theta|\theta^{(t)}) &\stackrel{\text{def.}}{=} E[\ell_{compl}(\theta)] \\ &\stackrel{\text{lin. med.}}{=} E[\tilde{z}_1] \ln \theta + E[\tilde{z}_0] \ln(1 - \theta) + (n_S - E[\tilde{z}_1]) \ln(2\theta) + (n_B - E[\tilde{z}_0]) \ln(1 - 2\theta). \end{aligned}$$

La *pasul E*, mediile

⁹⁶⁷Vă rugăm să remarcăți faptul că variabila \tilde{z}_1 este diferită de variabila z_1 care a fost folosită în rezolvările de la punctele *a* și *b*. Sperăm ca această „suprascrisere” (parțială) la nivel de notație să nu incomodeze cititorul. Am preferat / optat să păstrăm astfel convenția din problema 1, și anume: cu x se notează variabilele observabile și cu z variabilele neobservabile.

⁹⁶⁸Acum considerăm aceste date „sintetizate” sub forma următoare: două apariții ale stemei și trei apariții ale banului.

⁹⁶⁹Facem abstracție de un factor care depinde de probabilitatea de selecție π , dar nu depinde de θ :

$$C_{n_S}^{\tilde{z}_1} C_{n_B}^{\tilde{z}_0} \pi^{\tilde{z}_1 + \tilde{z}_0} (1 - \pi)^{n_S + n_B - (\tilde{z}_1 + \tilde{z}_0)}.$$

- $E[\tilde{z}_1]$, adică numărul așteptat (engl., expected number) de apariții ale feței *stema* pentru moneda 1 din totalul de $n_S \stackrel{\text{not.}}{=} 2$ apariții ale stemei în datele noastre, și
 - $E[\tilde{z}_0]$, adică numărul așteptat (engl., expected number) de apariții ale feței *ban* pentru moneda 1 din totalul de $n_B \stackrel{\text{not.}}{=} 3$ apariții ale banului în datele noastre,
- se calculează — la iterarea t — folosind fie proprietatea de liniaritate a mediilor, fie formula mediei distribuției binomiale:⁹⁷⁰

$$E[\tilde{z}_1] = n_S \cdot q_1^{(t)} \text{ și } E[\tilde{z}_0] = n_B \cdot q_0^{(t)},$$

fiindcă se poate observa relativ ușor că

$$\tilde{z}_1 \sim \text{Binomial}(n_S, q_1^{(t)}) \text{ și } \tilde{z}_0 \sim \text{Binomial}(n_B, q_0^{(t)}). \quad (407)$$

În această nouă formulare (a treia) a algoritmului EM pentru problema noastră, pasul E este constituit de formulele (405) și (406).

Pasul M se elaborează după cum urmează.

Stim că

$$\theta^{(t+1)} = \underset{\theta \in (0, 1/2)}{\operatorname{argmax}} Q(\theta | \theta^{(t)}). \quad (408)$$

Dacă maximul funcției $Q(\theta | \theta^{(t)})$ se situează în *interiorul* intervalului $(0, 1/2)$, atunci cu necesitate vom avea $\frac{\partial}{\partial \theta} Q(\theta | \theta^{(t)}) = 0$, deci

$$\begin{aligned} \frac{\partial}{\partial \theta} \left(E[\tilde{z}_1] \ln \theta + E[\tilde{z}_0] \ln(1 - \theta) + (n_S - E[\tilde{z}_1]) \ln(2\theta) + (n_B - E[\tilde{z}_0]) \ln(1 - 2\theta) \right) &= 0 \\ \Leftrightarrow \frac{E[\tilde{z}_1]}{\theta} - \frac{E[\tilde{z}_0]}{1 - \theta} + \frac{2(n_S - E[\tilde{z}_1])}{2\theta} - \frac{2(n_B - E[\tilde{z}_0])}{1 - 2\theta} &= 0 \\ \Leftrightarrow \frac{n_S \cdot q_1^{(t)}}{\theta} - \frac{n_B \cdot q_0^{(t)}}{1 - \theta} + \frac{2(n_S - n_S \cdot q_1^{(t)})}{2\theta} - \frac{2(n_B - n_B \cdot q_0^{(t)})}{1 - 2\theta} &= 0 \\ \Leftrightarrow \frac{n_S \cdot q_1^{(t)}}{\theta} - \frac{n_B \cdot q_0^{(t)}}{1 - \theta} + \frac{2n_S(1 - q_1^{(t)})}{2\theta} - \frac{2n_B(1 - q_0^{(t)})}{1 - 2\theta} &= 0 \\ \Leftrightarrow n_S(1 - \theta)(1 - 2\theta) - n_B q_0^{(t)} \theta(1 - 2\theta) - 2n_B(1 - q_0^{(t)}) \theta(1 - \theta) &= 0 \\ \Leftrightarrow n_S + \theta(-3n_S - \cancel{n_B q_0^{(t)}} - 2n_B + \cancel{2n_B q_0^{(t)}}) + \theta^2(2n_S + \cancel{2n_B q_0^{(t)}} + 2n_B - \cancel{2n_B q_0^{(t)}}) &= 0 \\ \Leftrightarrow 2(n_S + n_B)\theta^2 - (3n_S + 2n_B - n_B q_0^{(t)})\theta + n_S &= 0. \end{aligned} \quad (409)$$

Se constată ușor că ecuația (409) este de fapt aceeași cu ecuația (401) pe care am obținut-o la pasul M atunci când probabilitatea de selecție π era considerată fixată. Prin urmare, soluția problemei de optimizare (408) — adică, regula de actualizare a valorilor parametrului θ la pasul M pentru această nouă versiune a algoritmului EM — va avea exact forma (402), cu $a = 2(n_S + n_B)$, $b = -(3n_S + 2n_B - n_B q_0^{(t)})$ și $c = n_S$.

Din formula (409) se observă faptul că nu este necesar să mai calculăm $q_1^{(t)}$ la pasul E. Chiar și dacă ar fi trebuit să calculăm $q_1^{(t)}$, pasul E tot ar fi fost

⁹⁷⁰Vedeți problema 25.b de la capitolul de *Fundamente*.

mai eficient în noua versiune decât în versiunea „clasică“ a algoritmului EM pentru rezolvare de mixturi de distribuții Bernoulli, fiindcă acolo calculam n medii $\mu_i^{(t)}$. Acest fapt este foarte semnificativ pentru cazuri în care avem foarte multe instanțe x_i !

Putem scrie acum pseudo-codul algoritmului EM pentru rezolvarea acestei mixturi de distribuții Bernoulli.

Inițializare:

atribuie o valoare arbitrară $\theta^{(0)}$ în intervalul $(0, 1/2)$ pentru parametrul θ ;

Corpul iterativ:

pentru $t = 0, \dots, T - 1$ (cu T fixat în avans)

(sau: până când $|\theta^{(t)} - \theta^{(t+1)}| < \varepsilon$, cu ε fixat în avans) execută

Pasul E:

$$q_0^{(t)} = \frac{\pi (1 - \theta^{(t)})}{\pi (1 - \theta^{(t)}) + (1 - \pi) (1 - 2\theta^{(t)})};$$

Pasul M:

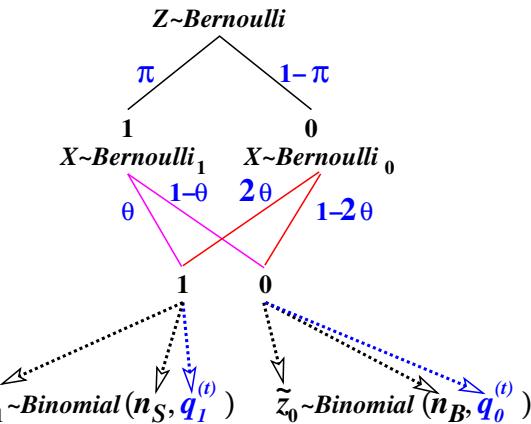
$$\theta^{(t+1)} = (-b - \sqrt{b^2 - 4ac})/(2a), \text{ cu}$$

$$a = 2(n_S + n_B), b = -(3n_S + 2n_B - n_B q_0^{(t)}), \text{ și } c = n_S;$$

Returnează $\theta^{(t+1)}$;

Remarcați următorul *fapt interesant*: Acest ultim tip de estimare a parametrului / parametrilor unei distribuții probabiliste cu ajutorul algoritmului EM este posibil datorită faptului că au loc relațiile (407). Ilustrăm această importantă „dependență“ în figura alăturată (pentru a facilita reținerea ei de către cititor!). Vă reamintim că n_S și n_B sunt datele / variabilele observabile, iar \tilde{z}_1 și \tilde{z}_0 sunt variabilele neobservabile.

Ideea aceasta a fost folosită la rezolvarea problemei 5 și va fi folosită și la rezolvarea problemei 30 și, într-o anumită măsură, la problemele 13, 14, 17, 31 și 32.



8.

(Un exemplu de rezolvare a unei mixturi de doi vectori de variabile Bernoulli independente și [respectiv] identic distribuite

A. când toate variabilele sunt observabile: MLE, în manieră directă;

B. când variabilele de selecție sunt neobservabile:

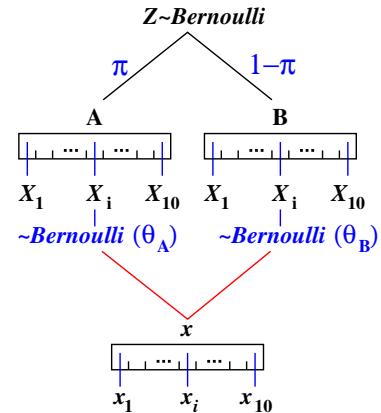
MLE, folosind algoritmul EM)

prelucrare de Liviu Ciortuz, 2015, după
 ■ □ “What is the expectation maximization algorithm?”,
 Chuong B. Do, Serafim Batzoglou,
 Nature Biotechnology, vol. 26, no. 8, 2008, pag. 897-899

Fie următorul experiment probabilist:

Dispunem de două monede, A și B . Efectuăm 5 serii de operațiuni de tipul următor:

- alegem în mod aleatoriu una dintre monedele A și B , cu probabilitate egală ($1/2$);
- aruncăm de 10 ori moneda care tocmai a fost aleasă (Z) și notăm rezultatul, summarizat ca număr (X) de fețe ‘head’ (rom., ‘stemă’) obținute în urma aruncării monedei respective.⁹⁷¹



Observație (1): Facem *presupunerea* că am reținut succesiunea [tuturor] rezultatelor obținute la aruncările celor două monede. Precizarea *de facto* a acestei succesiuni nu este esențială. Dacă nu s-ar face această *presupunere*, ar trebui să lucrăm cu *distribuția binomială*.

A. La acest punct vom considera că s-a obținut următorul rezultat pentru experimentul nostru:

i	Z_i	X_i
1	B	$5H$ ($5T$)
2	A	$9H$ ($1T$)
3	A	$8H$ ($2T$)
4	B	$4H$ ($6T$)
5	A	$7H$ ($3T$)

Semnificația variabilelor aleatoare Z_i și X_i pentru $i = 1, \dots, 5$ din tabelul de mai sus este imediată.

i. Calculați $\hat{\theta}_A$ și $\hat{\theta}_B$, probabilitățile de apariție a feței ‘head’ pentru cele două monede, folosind definiția clasică pentru probabilitatea evenimentelor aleatoare, și anume raportul dintre numărul de cazuri favorabile și numărul de cazuri posibile, relativ la întregul experiment.

ii. Calculați $L_1(\theta_A, \theta_B) \stackrel{not.}{=} P(X, Z | \theta_A, \theta_B)$, funcția de verosimilitate a datelor $X \stackrel{not.}{=} < X_1, \dots, X_5 >$ și $Z \stackrel{not.}{=} < Z_1, \dots, Z_5 >$ în raport cu parametrii θ_A și respectiv θ_B ai distribuțiilor (Bernoulli) care modelează aruncarea celor două monede.

iii. Calculați $\hat{\theta}_A \stackrel{not.}{=} \text{argmax}_{\theta_A} \log L_1(\theta_A, \theta_B)$ și $\hat{\theta}_B \stackrel{not.}{=} \text{argmax}_{\theta_B} \log L_1(\theta_A, \theta_B)$ cu ajutorul derivatelor parțiale de ordinul întâi. Baza logaritmului, lăsată aici nespecificată, se va considera supraunitară (de exemplu 2, e sau 10).

Observație (2): Lucrând corect, veți obține același rezultat ca la punctul *i*.

B. La acest punct se va relua experimentul de la punctul A, însă de data

⁹⁷¹ Observați că dacă în loc de 10 aruncări am considera o singură aruncare, atunci am fi practic în condițiile problemei 6, unde am prezentat algoritmul EM pentru mixturi de două distribuții Bernoulli.

Mixtura de aici poate fi exprimată în mod analitic astfel:

$$X = (x_1, \dots, x_{10}) \sim \pi \cdot \text{Bernoulli}(x_1; \theta_A) \dots \cdot \text{Bernoulli}(x_{10}; \theta_A) + (1-\pi) \cdot \text{Bernoulli}(x_1; \theta_B) \dots \cdot \text{Bernoulli}(x_{10}; \theta_B).$$

LC: Mulțumesc lui Sebastian Ciobanu pentru această observație.

aceasta vom considera că valorile variabilelor Z_i nu sunt cunoscute.

i	Z_i	X_i
1	?	5H (5T)
2	?	9H (1T)
3	?	8H (2T)
4	?	4H (6T)
5	?	7H (3T)

iv. Pentru conveniență, în locul variabilelor „neobservabile“ Z_i pentru $i = 1, \dots, 5$ vom considera variabilele-indicator (de asemenea neobservabile) $Z_{i,A}$, $Z_{i,B} \in \{0, 1\}$, cu $Z_{i,A} = 1_{\{Z_i=A\}}$ și $Z_{i,A} + Z_{i,B} = 1$.⁹⁷²

Folosind teorema lui Bayes, calculați mediile variabilelor neobservabile $Z_{i,A}$ și $Z_{i,B}$ condiționate de variabilele observabile X_i . Veți considera că parametrii acestor distribuții (Bernoulli) care modeleză aruncarea monedelor A și B au valorile $\theta_A^{(0)} = 0.6$ și respectiv $\theta_B^{(0)} = 0.5$.

Așadar, se cer: $E[Z_{i,A} | X_i, \theta_A^{(0)} = 0.6, \theta_B^{(0)} = 0.5] \text{ și } E[Z_{i,B} | X_i, \theta_A^{(0)} = 0.6, \theta_B^{(0)} = 0.5]$ pentru $i = 1, \dots, 5$. Ca și mai sus, probabilitățile a priori $P(Z_{i,A} = 1)$ și $P(Z_{i,B} = 1)$ se vor considera egale cu $1/2$.

v. Calculați media funcției de log-verosimilitate a datelor „complete“, X și Z :

$$L_2(\theta_A, \theta_B) \stackrel{\text{def.}}{=} E_{P(Z|X, \theta^{(0)})}[\log P(X, Z | \theta)],$$

unde $\theta = (\theta_A, \theta_B)$ și $\theta^{(0)} = (\theta_A^{(0)}, \theta_B^{(0)})$. Semnificația notației de mai sus este următoarea: funcția $L_2(\theta_A, \theta_B)$ este o medie a variabilei aleatoare reprezentată de log-verosimilitatea datelor complete (observabile și, respectiv, neobservabile), iar această medie se calculează în raport cu distribuția probabilistă condițională a datelor neobservabile, $P(Z | X, \theta^{(0)})$.

Observație (3): La elaborarea calculului, veți folosi mai întâi proprietatea de liniaritate a mediilor variabilelor aleatoare, și apoi rezultatele de la punctul iv.

vi. Calculați $\theta_A^{(1)} \stackrel{\text{not.}}{=} \operatorname{argmax}_{\theta_A} L_2(\theta_A, \theta_B)$ și $\theta_B^{(1)} \stackrel{\text{not.}}{=} \operatorname{argmax}_{\theta_B} L_2(\theta_A, \theta_B)$.

C. Formalizați pașii E și M ai algoritmului EM pentru estimarea parametrilor θ_A și θ_B în condițiile de la punctul B.⁹⁷³

D. Scrieți expresia funcției de log-verosimilitate a datelor observabile,

$$\ln P(X|\theta_A, \theta_B) \text{ unde } X \stackrel{\text{not.}}{=} (X_1, \dots, X_5).$$

Calculați valorile acestei funcții la finalul pasului de inițializare din algoritmul EM și apoi la finalul execuției primei iterații. Arătați că

$$P(X|\theta_A^{(0)}, \theta_B^{(0)}) \leq P(X|\theta_A^{(1)}, \theta_B^{(1)}),$$

ceea ce ilustrează proprietatea care a fost demonstrată la problema 2, și anume că de la o iterație la alta algoritmul EM crește (sau cel puțin păstrează ne-schimbătă) valoarea funcției de log-verosimilitate a datelor observabile.

⁹⁷²Evident, întrucât variabilele Z_i sunt aleatoare, rezultă că și variabilele $Z_{i,A}$ și $Z_{i,B}$ sunt aleatoare.

⁹⁷³Puteți lucra folosind formularea schemei algoritmice EM din *Observația (5)* de la problema 1.

Răspuns:

A. Acesta este un experiment probabilist în care toate variabilele sunt observabile.

i. Analizând datele din tabelul din enunț, rezultă imediat $\hat{\theta}_A = \frac{24}{24+6} = 0.8$ și $\hat{\theta}_B = \frac{9}{9+11} = 0.45$. Este de remarcat faptul că termenii 6 și respectiv 11 de la numitorii acestor fracții reprezintă numărul de fețe ‘tail’ (rom., ‘ban’) care au fost obținute la aruncarea monedei A și respectiv B: $6T = 1T + 2T + 3T$, $11T = 5T + 6T$.

Observație (4): Dacă în locul variabilelor binare $Z_i \in \{A, B\}$ pentru $i = 1, \dots, 5$ introducem în mod natural variabilele-indicator $Z_{i,A} \in \{0, 1\}$ și $Z_{i,B} \in \{0, 1\}$ tot pentru $i = 1, \dots, 5$, definite prin $Z_{i,A} = 1$ iff $Z_i = A$ și $Z_{i,B} = 1$ iff $Z_i = B$, atunci procesările necesare pentru calculul probabilităților / parametrilor $\hat{\theta}_A$ și $\hat{\theta}_B$ pot fi prezentate în mod sintetizat ca în tabelul de mai jos.⁹⁷⁴

i	$Z_{i,A}$	$Z_{i,B}$	X_i	$X_i \cdot Z_{i,A}$	$X_i \cdot Z_{i,B}$
1	0	1	$5H$	$0H (0T)$	$5H (5T)$
2	1	0	$9H$	$9H (1T)$	$0H (0T)$
3	1	0	$8H$	$8H (2T)$	$0H (0T)$
4	0	1	$4H$	$0H (0T)$	$4H (6T)$
5	1	0	$7H$	$7H (3T)$	$0H (0T)$
\Rightarrow				$\sum_{i=1}^5 X_i \cdot Z_{i,A} = 24H$	$\sum_{i=1}^5 X_i \cdot Z_{i,B} = 9H$
				$\sum_{i=1}^5 (10 - X_i) \cdot Z_{i,A} = 6T$	$\sum_{i=1}^5 (10 - X_i) \cdot Z_{i,B} = 11T$
				$\Rightarrow \begin{cases} \hat{\theta}_A = \frac{24}{24+6} = 0.8 \\ \hat{\theta}_B = \frac{9}{9+11} = 0.45 \end{cases}$	

ii. Calculul verosimilității datelor X și Z :

$$\begin{aligned}
 L_1(\theta_A, \theta_B) &\stackrel{def.}{=} P(X, Z_A, Z_B \mid \theta_A, \theta_B) = \prod_{i=1}^5 P(X_i, Z_{i,A}, Z_{i,B} \mid \theta_A, \theta_B) \\
 &\stackrel{indep.}{=} \prod_{i=1}^5 P(X_i \mid Z_{i,A}, Z_{i,B}, \theta_A, \theta_B) \cdot P(Z_{i,A}, Z_{i,B} \mid \theta_A, \theta_B) \\
 &= P(X_1 \mid Z_{B,1} = 1, \theta_B) \cdot 1/2 \cdot P(X_2 \mid Z_{A,2} = 1, \theta_A) \cdot 1/2 \cdot \\
 &\quad P(X_3 \mid Z_{A,3} = 1, \theta_A) \cdot 1/2 \cdot P(X_4 \mid Z_{B,4} = 1, \theta_B) \cdot 1/2 \cdot \\
 &\quad P(X_5 \mid Z_{A,5} = 1, \theta_A) \cdot 1/2 \\
 &= \theta_B^5 (1 - \theta_B)^5 \cdot \theta_A^0 (1 - \theta_A) \cdot \theta_A^8 (1 - \theta_A)^2 \cdot \theta_B^4 (1 - \theta_B)^6 \cdot \theta_A^7 (1 - \theta_A)^3 \cdot \frac{1}{2^5} \\
 &= \frac{1}{2^5} \theta_A^{24} (1 - \theta_A)^6 \theta_B^9 (1 - \theta_B)^{11}
 \end{aligned}$$

iii. Funcția de log-verosimilitate a datelor complete se exprimă astfel:

⁹⁷⁴Prezentăm acest „artificiu“ ca pregătire pentru rezolvarea (ulterioră a) punctului B al prezentei probleme.

$$\log L_1(\theta_A, \theta_B) = -5 \log 2 + 24 \log \theta_A + 6 \log(1 - \theta_A) + 9 \log \theta_B + 11 \log(1 - \theta_B)$$

Prin urmare, maximul acestei funcții în raport cu parametrul θ_A se calculează astfel:

$$\begin{aligned} \frac{\partial \log L_1(\theta_A, \theta_B)}{\partial \theta_A} = 0 &\Leftrightarrow \frac{\partial}{\partial \theta_A}[24 \log \theta_A + 6 \log(1 - \theta_A)] = 0 \\ \Leftrightarrow \frac{24}{\theta_A} - \frac{6}{1 - \theta_A} &= 0 \Leftrightarrow \frac{4}{\theta_A} = \frac{1}{1 - \theta_A} \Leftrightarrow 4 - 4\theta_A = \theta_A \Leftrightarrow \hat{\theta}_A = 0.8 \end{aligned}$$

Similar, se face calculul și pentru $\frac{\partial \log L_1(\theta_A, \theta_B)}{\partial \theta_B}$ și se obține $\hat{\theta}_B = 0.45$.⁹⁷⁵ Cele două valori obținute, $\hat{\theta}_A$ și $\hat{\theta}_B$ reprezintă estimarea de verosimilitate maximă a probabilităților de apariție a feței ‘head’ / stemă pentru moneda A și respectiv moneda B.

Observații:

5. Am arătat pe acest caz particular că metoda de calculare a probabilităților ($\hat{\theta}_A$ și $\hat{\theta}_B$) direct din datele observate (așa cum o știm din liceu) corespunde de fapt metodei de estimare în sensul verosimilității maxime (MLE).
6. La punctul B vom arăta cum anume se poate face estimarea acelorași parametri θ_A și θ_B în cazul în care o parte din date, și anume variabilele Z_i (pentru $i = 1, \dots, 5$) sunt neobservabile.

B. Algoritmul EM ne permite să facem în mod iterativ estimarea parametrilor θ_A și θ_B în funcție de valorile variabilelor observabile, X_i , și de valorile inițiale atribuite parametrilor (în cazul nostru, $\theta_A^{(0)} = 0.6$ și $\theta_B^{(0)} = 0.5$).

Notă:

Vom sintetiza calculele de la prima iterație a algoritmului EM — care vor fi detaliate la punctele iv, v și vi de mai jos — sub forma următoare, care seamănă (dar și diferă!) într-o anumită măsură de tabelele de la punctul A:

i	$Z_{i,A}$	$Z_{i,B}$	X_i	$E[Z_{i,A}]$	$E[Z_{i,B}]$	$X_i \cdot E[Z_{i,A}]$	$X_i \cdot E[Z_{i,B}]$
1	?	?	5H	0.45H	0.55H	2.2H (2.2T)	2.8H (2.8T)
2	?	?	9H	0.80H	0.20H	7.2H (0.8T)	1.8H (0.2T)
3	?	?	8H	0.73H	0.27H	5.9H (1.5T)	2.1H (0.5T)
4	?	?	4H	0.35H	0.65H	1.4H (2.1T)	2.6H (3.9T)
5	?	?	7H	0.65H	0.35H	4.5H (1.9T)	2.5H (1.1T)

$$\Rightarrow \begin{cases} \sum_{i=1}^5 X_i \cdot E[Z_{i,A}] = 21.3H \\ \sum_{i=1}^5 (10 - X_i) \cdot E[Z_{i,A}] = 8.7T \\ \sum_{i=1}^5 X_i \cdot E[Z_{i,B}] = 11.7H \\ \sum_{i=1}^5 (10 - X_i) \cdot E[Z_{i,B}] = 8.3T \end{cases} \Rightarrow \begin{cases} \hat{\theta}_A^{(1)} = \frac{21.3}{21.3 + 8.7} \approx 0.71 \\ \hat{\theta}_B^{(1)} = \frac{11.7}{11.7 + 8.3} \approx 0.58 \end{cases}$$

Vom arăta că

⁹⁷⁵Se verifică ușor faptul că într-addevăr rădăcinile derivatelor parțiale de ordinul întâi pentru funcția de log-verosimilitate reprezintă puncte de maxim. Pentru aceasta, se arată că matricea hessiană pentru funcția L_1 este negativ definită. Într-addevăr, derivatele de ordinul al doilea $\frac{\partial^2}{\partial \theta_A^2} L_1(\theta_A, \theta_B)$ și $\frac{\partial^2}{\partial \theta_B^2} L_1(\theta_A, \theta_B)$ au valori negative pe tot domeniul de definiție, în vreme ce $\frac{\partial^2}{\partial \theta_A \partial \theta_B} L_1(\theta_A, \theta_B) = 0$ și $\frac{\partial^2}{\partial \theta_B \partial \theta_A} L_1(\theta_A, \theta_B) = 0$.

- într-adevăr, este posibilă calcularea mediilor variabilelor neobservabile $Z_{i,A}$ și $Z_{i,B}$, condiționate de variabilele observabile X_i și în funcție de valorile asignate inițial pentru parametrii θ_A și θ_B ;
- față de tabloul de sinteză de la punctul precedent, când toate variabilele erau observabile și se calculau produsele $X_i \cdot Z_{i,A}$ și $X_i \cdot Z_{i,B}$, aici se înlocuiesc variabilele $Z_{i,A}$ și $Z_{i,B}$ cu mediile $E[Z_{i,A}]$ și $E[Z_{i,B}]$ în produsele respective. De fapt, în loc să se calculeze $\sum_i X_i \cdot Z_{i,A}$ se calculează media $E[\sum_i X_i \cdot Z_{i,A}]$, și similar pentru B .

Observație importantă (7): Pentru simplitate, în cele de mai sus (inclusiv în tabelele precedente), prin $E[Z_{i,A}]$ am notat $E[Z_{i,A} | X_i, \theta^{(0)}]$, iar prin $E[Z_{i,B}]$ am notat $E[Z_{i,B} | X_i, \theta^{(0)}]$, unde $\theta^{(0)} \stackrel{not.}{=} (\theta_A^{(0)}, \theta_B^{(0)})$.

iv. Întrucât variabilele $Z_{i,A}$ au valori booleene (0 sau 1), rezultă că

$$E[Z_{i,A} | X_i, \theta^{(0)}] = 0 \cdot P(Z_{i,A} = 0 | X_i, \theta^{(0)}) + 1 \cdot P(Z_{i,A} = 1 | X_i, \theta^{(0)}) = P(Z_{i,A} = 1 | X_i, \theta^{(0)})$$

Probabilitățile $P(Z_{i,A} = 1 | X_i, \theta^{(0)}) = P(Z_{i,A} = 1 | X_i, \theta^{(0)})$, pentru $i = 1, \dots, 5$ se pot calcula folosind teorema lui Bayes:

$$\begin{aligned} & P(Z_{i,A} = 1 | X_i, \theta^{(0)}) \\ &= \frac{P(X_i | Z_{i,A} = 1, \theta^{(0)}) \cdot P(Z_{i,A} = 1 | \theta^{(0)})}{P(X_i | Z_{i,A} = 1, \theta^{(0)}) \cdot P(Z_{i,A} = 1 | \theta^{(0)}) + P(X_i | Z_{i,A} = 0, \theta^{(0)}) \cdot P(Z_{i,A} = 0 | \theta^{(0)})} \\ &= \frac{P(X_i | Z_{i,A} = 1, \theta_A^{(0)})}{P(X_i | Z_{i,A} = 1, \theta_A^{(0)}) + P(X_i | Z_{i,B} = 1, \theta_B^{(0)})} \end{aligned}$$

S-a ținut cont că $P(Z_{i,A} = 1 | \theta^{(0)}) = P(Z_{i,B} = 1 | \theta^{(0)}) = 1/2$ (a se vedea enunțul).

De exemplu, pentru $i = 1$ vom avea:

$$\begin{aligned} E[Z_{A,1} | X_1, \theta^{(0)}] &= \frac{0.6^5(1 - 0.6)^5}{0.6^5(1 - 0.6)^5 + 0.5^5(1 - 0.5)^5} = \frac{0.24^5}{0.24^5 + 0.25^5} = \frac{1}{1 + \left(\frac{25}{24}\right)^5} \\ &\approx 0.45 \end{aligned}$$

Similar cu $E[Z_{A,1} | X_1, \theta^{(0)}]$ se calculează și celelalte medii $E[Z_{i,A} | X_i, \theta^{(0)}]$ pentru $i = 2, \dots, 5$ și $E[Z_{i,B} | X_i, \theta^{(0)}]$ pentru $i = 1, \dots, 5$.⁹⁷⁶ Am înregistrat aceste valori / medii în tabelul din mijloc din cadrul seriei formate din cele trei tabele din Nota de mai sus.

v. Media funcției de log-verosimilitate a datelor complete, $L_2(\theta_A, \theta_B)$, se calculează astfel:

$$\begin{aligned} L_2(\theta_A, \theta_B) &\stackrel{\text{def.}}{=} E_{P(Z|X, \theta^{(0)})} [\log P(X, Z | \theta)] \\ &\stackrel{\text{indep.}}{=} E_{P(Z|X, \theta^{(0)})} [\log \prod_{i=1}^5 P(X_i, Z_{i,A}, Z_{i,B} | \theta_A, \theta_B)] \\ &\stackrel{\text{reg. de mult.}}{=} E_{P(Z|X, \theta^{(0)})} [\log \prod_{i=1}^5 P(X_i | Z_{i,A}, Z_{i,B}; \theta_A, \theta_B) \cdot P(Z_{i,A}, Z_{i,B} | \theta_A, \theta_B)] \end{aligned}$$

⁹⁷⁶Se poate ține cont că, de îndată ce s-a calculat $E[Z_{i,A} | X_i, \theta^{(0)}]$, se poate obține imediat și $E[Z_{i,B} | X_i, \theta^{(0)}] = 1 - E[Z_{i,A} | X_i, \theta^{(0)}]$, fiindcă $Z_{i,A} + Z_{i,B} = 1$.

În continuare, omitând din nou distribuția probabilistă în raport cu care se calculează media aceasta întrucât ea poate fi subînțeleasă, vom scrie:

$$\begin{aligned}
 L_2(\theta_A, \theta_B) &= \\
 &= E \left[\log \prod_{i=1}^5 (\theta_A^{Z_{i,A}})^{X_i} \cdot [(1 - \theta_A)^{Z_{i,A}}]^{10-X_i} \cdot (\theta_B^{Z_{i,B}})^{X_i} \cdot [(1 - \theta_B)^{Z_{i,B}}]^{10-X_i} \cdot \frac{1}{2} \right] \\
 &= E \left[\sum_{i=1}^5 [X_i \cdot Z_{i,A} \cdot \log \theta_A + (10 - X_i) \cdot Z_{i,A} \cdot \log(1 - \theta_A) + \right. \\
 &\quad \left. X_i \cdot Z_{i,B} \cdot \log \theta_B + (10 - X_i) \cdot Z_{i,B} \cdot \log(1 - \theta_B) - \log 2] \right] \\
 &\stackrel{\text{lin.}}{=} \sum_{i=1}^5 [X_i \cdot E[Z_{i,A}] \cdot \log \theta_A + (10 - X_i) \cdot E[Z_{i,A}] \cdot \log(1 - \theta_A) + \\
 &\quad X_i \cdot E[Z_{i,B}] \cdot \log \theta_B + (10 - X_i) \cdot E[Z_{i,B}] \cdot \log(1 - \theta_B) - \log 2] \\
 &= \sum_{i=1}^5 \log[\theta_A^{X_i \cdot E[Z_{i,A}]} \cdot (1 - \theta_A)^{(10 - X_i) \cdot E[Z_{i,A}]} \cdot \\
 &\quad \theta_B^{X_i \cdot E[Z_{i,B}]} \cdot (1 - \theta_B)^{(10 - X_i) \cdot E[Z_{i,B}]} \cdot \frac{1}{2}] \\
 &= \log(\theta_A^{2.2} \cdot (1 - \theta_A)^{2.2} \cdot \theta_B^{2.8} \cdot (1 - \theta_B)^{2.8} \cdot \dots \cdot \theta_A^{4.5} \cdot (1 - \theta_A)^{1.9} \cdot \theta_B^{2.5} \cdot (1 - \theta_B)^{1.1} \cdot \frac{1}{2^5}).
 \end{aligned}$$

La ultima egalitate de mai sus, cantitățile fracționare provin din calculele simple $X_1 \cdot E[Z_{1,A} | X, \theta] \approx 2.2$, $X_1 \cdot E[Z_{1,B} | X, \theta] \approx 2.8$, ..., $X_5 \cdot E[Z_{1,A} | X, \theta] \approx 4.8$, $X_5 \cdot E[Z_{1,B} | X, \theta] \approx 2.5$ (a se vedea tabelele din cadrul *Notei* de mai sus).

Observație (8): Comparând funcția de [log-]verosimilitate de la partea A (vedeți punctul *ii.*) cu cea de aici, se constată că

- $X_i \cdot (Z_{i,A} + Z_{i,B})$ de acolo — mai precis: $X_i \cdot Z_{i,A}$ sau $X_i \cdot Z_{i,B}$, în funcție de valoarea lui Z_i — devine aici $X_i \cdot E[Z_{i,A} + Z_{i,B}] = X_i \cdot (E[Z_{i,A}] + E[Z_{i,B}])$,
- $\theta_A^{X_i \cdot Z_{i,A}} \cdot \theta_B^{X_i \cdot Z_{i,B}}$ — adică $\theta_A^{X_i}$ sau $\theta_B^{X_i}$, în funcție de valoarea lui Z_i — se înlocuiește cu $\theta_A^{X_i \cdot E[Z_{i,A}]} \cdot \theta_B^{X_i \cdot E[Z_{i,B}]}$.
- $(1 - \theta_A)^{(10 - X_i) \cdot Z_{i,A}} \cdot (1 - \theta_B)^{(10 - X_i) \cdot Z_{i,B}}$ — adică $(1 - \theta_A)^{10 - X_i}$ sau $(1 - \theta_B)^{10 - X_i}$, în funcție de valoarea lui Z_i — se înlocuiește cu $(1 - \theta_A)^{(10 - X_i) \cdot E[Z_{i,A}]} \cdot (1 - \theta_B)^{(10 - X_i) \cdot E[Z_{i,B}]}$.

Aceste înlocuiri / corespondențe, deși dau acum o *perspectivă intuitivă* asupra acestei instanțe particulare a schemei algoritmice EM — iar aceasta se va regăsi și în cazul altor instanțe ale aceleiași scheme algoritmice — se datorează în mod riguros proprietății de liniaritate a mediilor.

vi. Valorile parametrilor θ_A și θ_B pentru care se atinge maximul mediei funcției de log-verosimilitate a datelor complete se obțin cu ajutorul derivatelor parțiale de ordinul întâi:⁹⁷⁷

⁹⁷⁷ Este imediat că derivelele de ordinul al doilea $\frac{\partial^2}{\partial \theta_A^2} L_2(\theta_A, \theta_B)$ și $\frac{\partial^2}{\partial \theta_B^2} L_2(\theta_A, \theta_B)$ au valori negative pe tot domeniul de definiție, în vreme ce $\frac{\partial^2}{\partial \theta_A \partial \theta_B} L_2(\theta_A, \theta_B) = 0$ și $\frac{\partial^2}{\partial \theta_B \partial \theta_A} L_2(\theta_A, \theta_B) = 0$.

$$\begin{aligned}
\frac{\partial L_2(\theta_A, \theta_B)}{\partial \theta_A} &= 0 \\
\Rightarrow \frac{\partial}{\partial \theta_A} (2.2 \log \theta_A + 2.2 \log(1 - \theta_A) + \dots + 4.5 \log \theta_A + 1.9 \log(1 - \theta_A)) &= 0 \\
\Rightarrow \frac{2.2}{\theta_A} - \frac{2.2}{1 - \theta_A} + \dots + \frac{4.5}{\theta_A} - \frac{1.9}{1 - \theta_A} &= 0 \Rightarrow \dots \Rightarrow \theta_A^{(1)} \approx 0.71.
\end{aligned}$$

Similar, vom obține $\theta_B^{(1)} \approx 0.58$.

C. Formulele care se folosesc în cadrul algoritmului EM pentru rezolvarea problemei date (adică estimarea parametrilor θ_A și θ_B când variabilele Z_i sunt neobservabile) se deduc astfel:

Pasul E:

$$\begin{aligned}
E[Z_{i,A} | X_i, \theta] &= P(Z_{i,A} = 1 | X_i, \theta) = P(Z_{i,A} = 1 | X_i, \theta) \\
&= \frac{P(X_i | Z_{i,A} = 1; \theta) \cdot \overbrace{P(Z_{i,A} = 1 | \theta)}^{1/2}}{P(X_i | Z_{i,A} = 1; \theta) \cdot P(Z_{i,A} = 1 | \theta) + P(X_i | Z_{i,B} = 1; \theta) \cdot \overbrace{P(Z_{i,B} = 1 | \theta)}^{1/2}} \\
&= \frac{P(X_i | Z_{i,A} = 1; \theta)}{P(X_i | Z_{i,A} = 1; \theta) + P(X_i | Z_{i,B} = 1; \theta)} \\
&= \frac{\theta_A^{X_i} (1 - \theta_A)^{10 - X_i}}{\theta_A^{X_i} (1 - \theta_A)^{10 - X_i} + \theta_B^{X_i} (1 - \theta_B)^{10 - X_i}}
\end{aligned}$$

Procedând în mod similar, vom obține:

$$E[Z_{i,B} | X_i, \theta] = \frac{\theta_B^{X_i} (1 - \theta_B)^{10 - X_i}}{\theta_A^{X_i} (1 - \theta_A)^{10 - X_i} + \theta_B^{X_i} (1 - \theta_B)^{10 - X_i}}$$

Notând cu

- x_i valoarea variabilei X_i ,
- $\theta_A^{(t)}$ și respectiv $\theta_B^{(t)}$ estimările parametrilor θ_A și θ_B la iteratăia t a algoritmului EM,
- $p_{i,A}^{(t+1)}$ și respectiv $p_{i,B}^{(t+1)}$, mediile $E[Z_{i,A} | X_i, \theta_A^{(t)}]$ și $E[Z_{i,B} | X_i, \theta_B^{(t)}]$,

vom avea:

$$\begin{aligned}
p_{i,A}^{(t+1)} &= \frac{(\theta_A^{(t)})^{x_i} (1 - \theta_A^{(t)})^{10 - x_i}}{(\theta_A^{(t)})^{x_i} (1 - \theta_A^{(t)})^{10 - x_i} + (\theta_B^{(t)})^{x_i} (1 - \theta_B^{(t)})^{10 - x_i}} \\
p_{i,B}^{(t+1)} &= \frac{(\theta_B^{(t)})^{x_i} (1 - \theta_B^{(t)})^{10 - x_i}}{(\theta_A^{(t)})^{x_i} (1 - \theta_A^{(t)})^{10 - x_i} + (\theta_B^{(t)})^{x_i} (1 - \theta_B^{(t)})^{10 - x_i}}
\end{aligned}$$

Pasul M: Ca și mai înainte, în formulele de mai jos vom nota $E[Z_{i,A} | X_i, \theta^{(t)}]$ cu $E[Z_{i,A}]$ și $E[Z_{i,B} | X_i, \theta^{(t)}]$ cu $E[Z_{i,B}]$. Cu aceste notării, procedând similar cu calculul de la partea B, punctul v , vom avea:

$$L_2(\theta_A, \theta_B) = \log \prod_{i=1}^5 \theta_A^{x_i E[Z_{i,A}]} (1 - \theta_A)^{(10 - x_i) E[Z_{i,A}]} \theta_B^{x_i E[Z_{i,B}]} (1 - \theta_B)^{(10 - x_i) E[Z_{i,B}]}$$

Prin urmare,

$$\begin{aligned} \frac{\partial}{\partial \theta_A} L_2(\theta_A, \theta_B) = 0 &\Rightarrow \frac{1}{\theta_A} \sum_{i=1}^5 x_i E[Z_{i,A}] = \frac{1}{1-\theta_A} \sum_{i=1}^5 (10-x_i) E[Z_{i,A}] \\ \Rightarrow (1-\theta_A) \sum_{i=1}^5 x_i E[Z_{i,A}] &= \theta_A \sum_{i=1}^5 (10-x_i) E[Z_{i,A}] \Rightarrow \sum_{i=1}^5 x_i E[Z_{i,A}] = 10 \theta_A \sum_{i=1}^5 E[Z_{i,A}] \\ \Rightarrow \theta_A &= \frac{\sum_{i=1}^5 x_i E[Z_{i,A}]}{10 \sum_{i=1}^5 E[Z_{i,A}]} \quad \text{și, similar, } \theta_B = \frac{\sum_{i=1}^5 x_i E[Z_{i,B}]}{10 \sum_{i=1}^5 E[Z_{i,B}]} \end{aligned}$$

Așadar, la pasul M al algoritmului EM vom avea:

$$\theta_A^{(t+1)} = \frac{\sum_{i=1}^5 \frac{x_i}{10} p_{i,A}^{(t+1)}}{\sum_{i=1}^5 p_{i,A}^{(t+1)}} \quad \theta_B^{(t+1)} = \frac{\sum_{i=1}^5 \frac{x_i}{10} p_{i,B}^{(t+1)}}{\sum_{i=1}^5 p_{i,B}^{(t+1)}}$$

Observație (9): Implementând algoritmul EM cu relațiile obținute pentru pasul E și pasul M, după execuția a 10 iterări se vor obține valorile $\theta_A^{(10)} \approx 0.80$ și $\theta_B^{(10)} \approx 0.52$. Este interesant de observat că estimarea obținută pentru parametrul θ_A este acum la același nivel cu cea obținută prin metoda verosimilității maxime (MLE) în cazul observării tuturor variabilelor (0.80, vedeti rezolvarea de la partea A, punctul i), iar estimarea obținută pentru parametrul θ_B a coborât de la valoarea 0.58 care a fost obținută la prima iterare a algoritmului EM la o valoare (0.52) care este considerabil mai apropiată de estimarea prin metoda MLE (0.45).

D. Funcția de verosimilitate a datelor observabile se calculează astfel:

$$\begin{aligned} P(X|\theta_A, \theta_B) &= P(X_1, \dots, X_5|\theta_A, \theta_B) \stackrel{\text{indep.}}{=} \prod_{i=1}^5 P(X_i|\theta_A, \theta_B) \\ &\stackrel{\text{prob.}}{=} \text{marg} \prod_{i=1}^5 \left[\sum_{Z_i \in \{A, B\}} P(X_i, Z_i|\theta_A, \theta_B) \right] \\ &= \prod_{i=1}^5 [P(X_i, Z_i = A|\theta_A, \theta_B) + P(X_i, Z_i = B|\theta_A, \theta_B)] \\ &= \prod_{i=1}^5 [P(X_i|Z_i = A; \theta_A, \theta_B) \cdot \underbrace{P(Z_i = A|\theta_A, \theta_B)}_{1/2} + \\ &\quad P(X_i|Z_i = B; \theta_A, \theta_B) \cdot \underbrace{P(Z_i = B|\theta_A, \theta_B)}_{1/2}] \\ &= \frac{1}{2^5} \prod_{i=1}^5 [P(X_i|Z_i = A; \theta_A, \theta_B) + P(X_i|Z_i = B; \theta_A, \theta_B)]. \end{aligned}$$

În consecință, funcția de log-verosimilitate a datelor observabile este:

$$\begin{aligned} \ln P(X|\theta_A, \theta_B) &= \left(\sum_{i=1}^5 \ln [P(X_i|Z_i = A; \theta_A, \theta_B) + P(X_i|Z_i = B; \theta_A, \theta_B)] \right) - 5 \ln 2 \end{aligned}$$

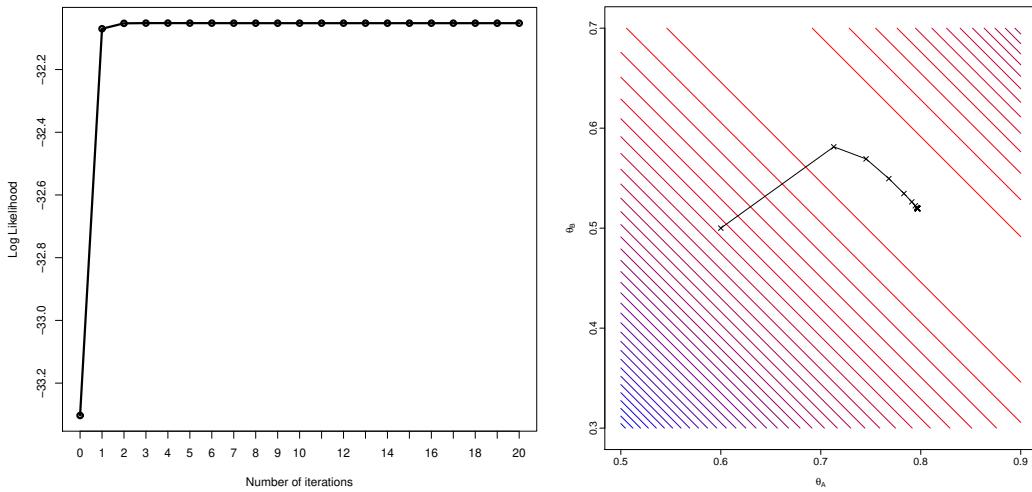
$$\begin{aligned}
&= \ln[\theta_A^5(1-\theta_A)^5 + \theta_B^5(1-\theta_B)^5] + \ln[\theta_A^9(1-\theta_A) + \theta_B^9(1-\theta_B)] + \\
&\quad \ln[\theta_A^8(1-\theta_A)^2 + \theta_B^8(1-\theta_B)^2] + \ln[\theta_A^4(1-\theta_A)^6 + \theta_B^4(1-\theta_B)^6] + \\
&\quad \ln[\theta_A^7(1-\theta_A)^3 + \theta_B^7(1-\theta_B)^3] - 5 \ln 2.
\end{aligned}$$

Folosind în această expresie valorile $\theta_A^{(0)} = 0.6$, $\theta_B^{(0)} = 0.5$ și respectiv $\theta_A^{(1)} = 0.71$, $\theta_B^{(1)} = 0.58$, după ce facem calculele obținem:

$$P(X|\theta_A^{(0)}, \theta_B^{(0)}) = -33.094 \text{ și } P(X|\theta_A^{(1)}, \theta_B^{(1)}) = -31.870.$$

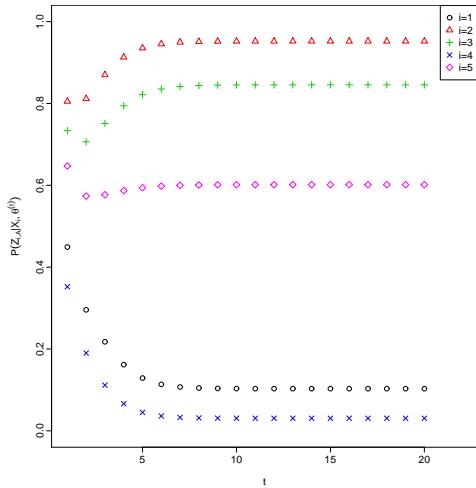
Prin urmare se verifică relația $P(X|\theta_A^{(0)}, \theta_B^{(0)}) \leq P(X|\theta_A^{(1)}, \theta_B^{(1)})$.

Observație (10): În graficul de mai jos, partea stângă sunt reprezentate valoare funcției de log-verosimilitate a datelor observabile care au fost obținute (folosind o implementare a algoritmului EM) la inițializare și apoi la finalul fiecareia din primele 20 iterări. În graficul din partea dreaptă avem reprezentarea sub formă de curbe de izocontur a valorilor log-verosimilității în funcție de cei doi parametri, θ_A și θ_B . Pe acest al doilea grafic a fost adăugat un „drum“ care pune în evidență succesiunea de valori pentru perechile $(\theta_A^{(t)}, \theta_B^{(t)})$ de-a lungul iterărilor executate de algoritm EM.⁹⁷⁸



⁹⁷⁸ Implementarea acestui algoritm EM, precum și graficele au fost realizate de către Sebastian Ciobanu.

Este interesant de asemenea să observăm care este evoluția valorilor mediilor variabilelor-indicator $Z_{i,A}$ și $Z_{i,B}$ (adică, a probabilităților a posteriori de apartenență la clustere a instanțelor X_i). În graficul alăturat prezentăm valorile $p_{i,A}^{(t)}$ not. $E[Z_{i,A} | X_i, \theta_A^{(t-1)}]$, pentru $i = 1, \dots, 5$. (Evident, $p_{i,B}^{(t)} = 1 - p_{i,A}^{(t)}$) Se observă că valorile acestea converg relativ repede și, dacă la convergență se face trecerea la apartenență de tip “hard” (adică se aplică o regulă de decizie de tipul celei folosite de algoritmul Bayes Naiv), atunci se regăsesc exact etichetele care au fost „sterse“ la începutul secțiunii B, adică (în ordine) $B, A, A, B, A!$



9.

(Algoritmul EM pentru rezolvarea unei mixturi de vectori de variabile Bernoulli independente; aplicare la clusterizarea cifrelor scrise de mâină)

□ • ○ CMU, 2015 fall, A. Smola, B. Poczos, HW2, pr. 1

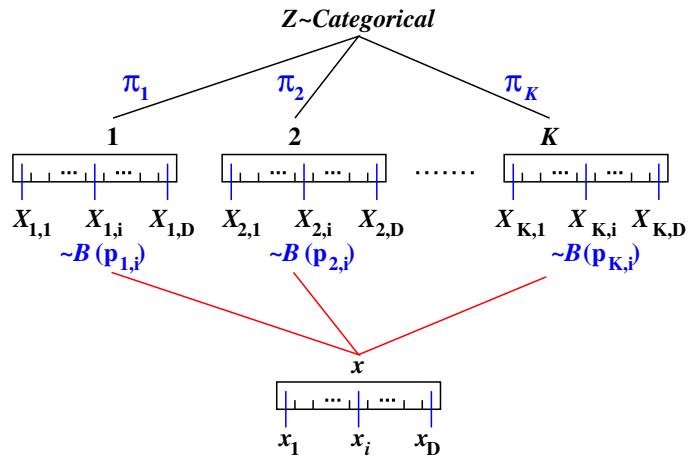
În acest exercițiu, veți crea un algoritm de tip EM (expectation-maximization) care să clusterizeze imagini alb-negru. Input-urile $x^{(i)}$ pot fi văzute ca vectori de valori binare corespunzătoare culorilor alb și negru. Obiectivul este să clusterizăm aceste imagini în [mai multe] grupuri. Pentru a rezolva această problemă, veți folosi un model de tip mixtură de vectori de distribuții Bernoulli independente.⁹⁷⁹

A. Mixtura de distribuții Bernoulli

- a. Considerăm un vector de variabile aleatoare binare, $x \in \{0, 1\}^D$. Presupunem că fiecare variabilă x_d urmează o distribuție Bernoulli(p_d), deci $P(x_d = 1) = p_d$. Fie $p \in (0, 1)^D$ vectorul rezultat de parametri Bernoulli. Scrieți expresia probabilității $P(x|p)$, considerând că variabilele x_d sunt independente între ele.

⁹⁷⁹Spre deosebire de modelul probabilist de la problema 8, aici variabilele Bernoulli dintr-un [același] vector nu vor mai fi identic distribuite.

b. Să presupunem acum că avem o mixtură de K vectori de distribuții Bernoulli: fiecare vector $x^{(i)}$ este generat folosind un vector de variabile Bernoulli independente, de parametru $p^{(k)}$ not. $(p_1^{(k)}, \dots, p_D^{(k)})$.



Presupunem de asemenea că dispunem de o distribuție aleatoare categorială $\pi \stackrel{\text{not.}}{=} (\pi_1, \dots, \pi_K)$, cu π_k indicând probabilitatea de selecție a setului de parametri Bernoulli $p^{(k)}$, pentru $k = 1, \dots, K$.⁹⁸⁰

Folosind formula probabilității totale, scrieți expresia probabilității $P(x^{(i)}|p, \pi)$, unde cu p notăm (de acum încolo) ansamblul de vectori de distribuții Bernoulli $(p^{(1)}, \dots, p^{(K)})$.

c. Presupunem că avem input-urile $X = \{x^{(i)}\}_{i=1, \dots, n}$. Folosind rezultatele de la punctele anterioare, scrieți expresia log-verosimilității datelor X , adică $\ln P(X|\pi, p)$.

B. Pasul de estimare (engl., expectation step)

d. Acum vom introduce variabilele latente pentru algoritmul EM. Fie $z^{(i)} \in \{0, 1\}^K$ un vector indicator, astfel încât $z_k^{(i)} = 1$ dacă vectorul $x^{(i)}$ a fost generat de vectorul de distribuții Bernoulli $(p^{(k)})$, și 0 în caz contrar. Fie $Z = \{z^{(i)}\}_{i=1, \dots, n}$. Cât este probabilitatea $P(z^{(i)}|\pi)$? Dar $P(x^{(i)}|z^{(i)}, p, \pi)$?

Indicație: Folosiți artificiul ridicării la putere (engl., exponentiation trick), ținând cont de faptul că $z_k^{(i)} \in \{0, 1\}$.

e. Folosind cele două probabilități calculate la punctul d, scrieți expresia verosimilității datelor complete, $P(X, Z|\pi, p)$.

f. Fie $\mu(z_k^{(i)}) \stackrel{\text{not.}}{=} E[z_k^{(i)}|x^{(i)}, \pi, p]$. Demonstrați că

$$\mu(z_k^{(i)}) = \frac{\pi_k \prod_{d=1}^D (p_d^{(k)})^{x_d^{(i)}} (1 - p_d^{(k)})^{1-x_d^{(i)}}}{\sum_{j=1}^K \pi_j \prod_{d=1}^D (p_d^{(j)})^{x_d^{(i)}} (1 - p_d^{(j)})^{1-x_d^{(i)}}}.$$

Fie \bar{p} și $\bar{\pi}$ noile valori ale parametrilor pe care dorim să le obținem prin maximizare, iar p și π valorile lor de la iterarea precedentă. Folosind aceste notății, arătați că funcția „auxiliară“, care este necesară pentru pasul M al algoritmului

⁹⁸⁰Mixtura aceasta poate fi exprimată în mod analitic astfel:

$$X = (x_1, \dots, x_D) \sim \pi_1 \cdot \text{Bernoulli}(x_1|p_{1,1}) \cdot \dots \cdot \text{Bernoulli}(x_D|p_{1,D}) + \\ \pi_K \cdot \text{Bernoulli}(x_1|p_{K,1}) \cdot \dots \cdot \text{Bernoulli}(x_D|p_{K,D}).$$

de estimare-maximizare, are expresia următoare:

$$\begin{aligned} E[\ln P(X, Z | \bar{p}, \bar{\pi}) | X, p, \pi] \\ = \sum_{i=1}^N \sum_{k=1}^K \mu(z_k^{(i)}) \left[\ln \bar{\pi}_k + \sum_{d=1}^D \left(x_d^{(i)} \ln \bar{p}_d^{(k)} + (1 - x_d^{(i)}) \ln (1 - \bar{p}_d^{(k)}) \right) \right]. \end{aligned}$$

C. Pasul de maximizare

g. Acum trebuie să maximizăm expresia funcției „auxiliare“ care a fost dedusă la punctul f, în raport cu $\bar{\pi}$ și \bar{p} . Mai întâi, arătați că valoarea \bar{p} care maximizează funcția „auxiliară“ este

$$\bar{p}^{(k)} = \frac{\sum_{i=1}^n \mu(z_k^{(i)}) x^{(i)}}{N_k},$$

unde $N_k = \sum_{i=1}^n \mu(z_k^{(i)})$.

h. Demonstrați că valoarea $\bar{\pi}$ care maximizează funcția „auxiliară“ este:

$$\bar{\pi}_k = \frac{N_k}{\sum_{k'} N_{k'}}.$$

Sugestie: Dat fiind faptul că este vorba de rezolvarea unei probleme de optimizare cu restricții, puteți apela la metoda multiplicatorilor Lagrange.⁹⁸¹

D. Folosiți algoritmul care a fost dedus mai sus pentru a clusteriza imagini ale cifrelor scrise de mână. Veți utiliza setul de date MNIST.⁹⁸² Fiecare input este o versiune simplificată (engl., flattened) a unei imagini cu 28×28 de pixeli, iar fiecare pixel este asociat unui număr binar, care corespunde culorilor alb sau negru.

Câteva sugestii:

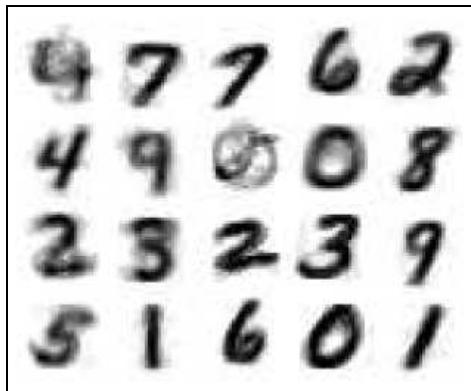
- Pentru a sigura stabilitatea calculelor numerice, folosiți operatorul \ln . În mod particular, fiți atenți la calcularea mediilor $\mu(z_k^{(i)})$.
- Va fi necesar să evitați ca π și p să primească valoarea 0; altfel veți obține $\ln 0 = -\infty$. Pentru „netezirea“ (engl., smoothing) acestor variabile, folosiți câte o distribuție a priori Dirichlet, de parametri α_1 și respectiv α_2 :

$$\bar{p}^{(k)} = \frac{\sum_{i=1}^N \mu(z_k^{(i)}) x^{(i)} + \alpha_1}{N_k + \alpha_1 D} \text{ și } \bar{\pi}_k = \frac{N_k + \alpha_2}{\sum_{k'} N_{k'} + \alpha_2 K}.$$

- Inițializați parametrii p în mod aleatoriu, eșantionând cf. distribuției $Uniform(0, 1)$ și normalizând fiecare $p^{(k)}$ pentru a avea suma egală cu unitatea, iar $\pi_k = 1/k$.
- Rulând algoritmul pe setul de date MNIST cu $K = 20$ (clustere) și $\alpha_1 = \alpha_2 = 10^{-8}$ timp de 20 de iterării, ar trebui să obțineți rezultate similare cu următoarele:

⁹⁸¹ Alternativ, puteți folosi notația de la familia de distribuții exponențiale. Vedeți problema 41 de la capitolul de *Fundamente*.

⁹⁸² Vedeți <https://profs.info.uaic.ro/~ciortuz/ML.ex-book/implementation-exercises/>
CMU.2015f.ASmola+BPoczos.HW2.pr2.EM4BMM.handwritten-digit-reco.data+code+sol/



Pentru fiecare cluster, plasați probabilitățile $p^{(d)}$ într-o matrice 28×28 și imprimați în nuanțe de gri imaginile care rezultă.⁹⁸³ Ce observați? Explicați. Folosind implementarea algoritmului dumneavoastră, clusterizați datele. Utilizând etichetele date în fișierul `yTrain`, precizați câte cifre [unice] sunt în fiecare cluster. Sunt clustere în care există exact [câte] o cifră?

Soluție:

a. Întrucât componentele vectorului x independente între ele, rezultă că $P(x|p) = \prod_{d=1}^D P(x_d|p_d) = \prod_{d=1}^D [p_d^{(x_d)} (1 - p_d)^{(1-x_d)}]$. Am folosit „artificiul exponentierii“.

b. Pentru fiecare $k \in \{1, \dots, K\}$ considerăm A_k evenimentul ca vectorul $x = x^{(i)}$ să fi fost generat de distribuția $Bernoulli(p^{(k)})$. Atunci,

$$P(x|p, \pi) = \sum_k P(x, A_k|p, \pi) = \sum_k P(x|A_k, p, \pi) P(A_k, p, \pi) = \sum_k P(x|p^{(k)}) \pi_k.$$

c. Tinând cont de independența producerii vectorilor $x^{(i)}$ pentru $i = 1, \dots, n$, rezultă că $\ln L(p, \pi) \stackrel{\text{def.}}{=} \ln P(X|p, \pi) = \sum_{i=1}^n \ln P(x^{(i)}|p, \pi)$.

d. Ca și la punctul b, pentru $k \in \{1, \dots, K\}$ considerăm A_k evenimentul ca vectorul $x = x^{(i)}$ să fi fost generat de distribuția $Bernoulli(p^{(k)})$. Atunci, tinând cont de faptul că $z_k^{(i)} \in \{0, 1\}$, putem scrie:

$$P(z^{(i)}|\pi) = \prod_{k=1}^K \pi_k^{z_k^{(i)}}$$

$$P(x^{(i)}|z^{(i)}, p, \pi) \stackrel{\text{indep.}}{=} \prod_{k=1}^K \left[P(x^{(i)}|z^{(i)}, p, \pi, A_k^{(i)}) \right]^{z_k^{(i)}} = \prod_{k=1}^K \left[P(x^{(i)}|p^{(k)}) \right]^{z_k^{(i)}}.$$

⁹⁸³Vedeți <https://www.gnu.org/software/octave/doc/interpreter/Representing-Images.html> pentru cum anume poate fi imprimată o astfel de matrice, sau folosiți funcția furnizată `show_clusters(p, a, b)` care va imprima mixturi din p pe un grid $a \times b$.

e. Verosimilitatea datelor „complete“ este:

$$\begin{aligned} P(X, Z|\pi, p) &= \prod_{i=1}^n P(x^{(i)}, z^{(i)}|p, \pi) = \prod_{i=1}^n P(x^{(i)}|z^{(i)}, p, \pi) P(z^{(i)}|\pi) \\ &= \prod_{i=1}^n \left[\prod_{k=1}^K [P(x^{(i)}|\pi^{(k)})]^{z_k^{(i)}} \right] \left[\prod_{k=1}^K \pi_k^{z_k^{(i)}} \right]. \end{aligned}$$

f. Pentru a calcula mediile $E[z_k^{(i)}|x^{(i)}, \pi, p]$, vom ține cont mai întâi de faptul că variabilele $z_k^{(i)}$ sunt variabile-indicator și apoi vom folosi formula lui Bayes.

$$\begin{aligned} \mu(z_k^{(i)}) &\stackrel{\text{not.}}{=} E[z_k^{(i)}|x^{(i)}, \pi, p] \\ &= P(z_k^{(i)} = 1|x^{(i)}, \pi, p) \\ &\stackrel{\text{F. Bayes}}{=} \frac{P(x^{(i)} = 1|z_k^{(i)} = 1, \pi, p) P(z_k^{(i)} = 1|\pi, p)}{\sum_{j=1}^K P(x^{(i)} = 1|z_j^{(i)} = 1, \pi, p) P(z_j^{(i)} = 1|\pi, p)} \\ &= \frac{\pi_k \prod_{d=1}^D (p_d^{(k)})^{x_d^{(i)}} (1 - p_d^{(k)})^{1-x_d^{(i)}}}{\sum_{j=1}^K \pi_j \prod_{d=1}^D (p_d^{(j)})^{x_d^{(i)}} (1 - p_d^{(j)})^{1-x_d^{(i)}}}. \end{aligned}$$

Folosind rezultatul de la punctul e, putem calcula expresia log-verosimilității datelor „complete“:

$$\begin{aligned} \ln P(X, Z|\bar{\pi}, \bar{p}) &= \sum_{i=1}^n \left[\sum_{k=1}^K z_k^{(i)} \ln [P(x^{(i)}|\bar{p}^{(k)})] \right] + \left[\sum_{k=1}^K z_k^{(i)} \ln \bar{\pi}_k \right] \\ &= \sum_{i=1}^n \sum_{k=1}^K z_k^{(i)} [\ln P(x^{(i)}|\bar{p}^{(k)}) + \ln \bar{\pi}_k] \\ &\stackrel{a.}{=} \sum_{i=1}^n \sum_{k=1}^K z_k^{(i)} \left[\ln \bar{\pi}_k + \ln \prod_{d=1}^D (\bar{p}_d^{(k)})^{x_d^{(i)}} (1 - \bar{p}_d^{(k)})^{1-x_d^{(i)}} \right] \\ &= \sum_{i=1}^n \sum_{k=1}^K z_k^{(i)} \left[\ln \bar{\pi}_k + \sum_{d=1}^D (x_d^{(i)} \ln \bar{p}_d^{(k)} + (1 - x_d^{(i)}) \ln (1 - \bar{p}_d^{(k)})) \right]. \end{aligned}$$

Aplicând acestei log-verosimilități operatorul medie (E), apoi ținând cont de proprietatea de liniaritate a mediei și, în final, înlocuind $E[z_k^{(i)}]$ cu $\mu(z_k^{(i)})$, vom ajunge la soluția indicată în enunț.

g. Vom egala cu 0 derivata parțială a funcței „auxiliare“ în raport cu $\bar{p}_d^{(k)}$:

$$\frac{\partial}{\partial \bar{p}_d^{(k)}} E[\ln P(X, Z|\bar{\pi}, \bar{p})] = \sum_{i=1}^n \mu(z_k^{(i)}) \left[\frac{x_d^{(i)}}{\bar{p}_d^{(k)}} - \frac{1 - x_d^{(i)}}{1 - \bar{p}_d^{(k)}} \right] = 0.$$

Aducând la același numitor, rezultă:

$$\sum_{i=1}^n \mu(z_k^{(i)}) \left[x_d^{(i)} (1 - \bar{p}_d^{(k)}) - (1 - x_d^{(i)}) \bar{p}_d^{(k)} \right] = \sum_{i=1}^n \mu(z_k^{(i)}) \left[x_d^{(i)} - \bar{p}_d^{(k)} \right] = 0.$$

Scoțându-l pe $\bar{p}_d^{(k)}$ din această ecuație, vom obține:

$$\bar{p}_d^{(k)} = \frac{\sum_{i=1}^n \mu(z_k^{(i)}) x_d^{(i)}}{\sum_{i=1}^n \mu(z_k^{(i)})} = \frac{\sum_{i=1}^n \mu(z_k^{(i)}) x_d^{(i)}}{N_k}.$$

Se poate demonstra relativ ușor că valorile $\bar{p}_d^{(k)}$ sunt corespunzătoare unui (unic) punct de maxim al funcției „auxiliare“.

h. Este suficient să minimizăm $-\sum_{i=1}^n \sum_{k=1}^K \mu(z_k^{(i)}) \ln \bar{\pi}_k$, fiindcă restul termenilor din expresia funcției „auxiliare“ nu depind de $\bar{\pi}$. Pentru a ne asigura că $\bar{\pi}$ corespunde unei distribuții probabiliste, trebuie să impunem restricția $\sum_k \bar{\pi}_k = 1$. Fie λ variabila Lagrange corespunzătoare acestei restricții. Lagrangeanul generalizat va avea expresia următoare:

$$L(\bar{\pi}, \lambda) = -\sum_{i=1}^n \sum_{k=1}^K \mu(z_k^{(i)}) \ln \bar{\pi}_k + \lambda \left(\sum_{k=1}^K \bar{\pi}_k - 1 \right).$$

Calculând derivata parțială în raport cu $\bar{\pi}_k$ și egalând-o apoi cu 0, vom obține:

$$\frac{\partial}{\partial \bar{\pi}_k} L(\bar{\pi}, \lambda) = -\sum_{i=1}^n \frac{\mu(z_k^{(i)})}{\bar{\pi}_k} + \lambda = 0.$$

Din această ecuație vom obține pentru $\bar{\pi}_k$ valoarea următoare:

$$\bar{\pi}_k = \frac{\sum_{i=1}^n \mu(z_k^{(i)})}{\lambda} = \frac{N_k}{\lambda}.$$

Înlocuind această valoare a lui $\bar{\pi}_k$ în expresia lagrangeanului generalizat $L(\bar{\pi}, \lambda)$, vom obține lagrangeanul dual:

$$L(\lambda) = -\sum_{i=1}^n \sum_{k=1}^K \mu(z_k^{(i)}) (\ln N_k - \ln \lambda) + \left(\sum_{k=1}^K N_k - \lambda \right).$$

Derivând această expresie în raport cu λ și egalând cu 0, vom avea:

$$\frac{1}{\lambda} \sum_{i=1}^n \sum_{k=1}^K \mu(z_k^{(i)}) - 1 = 0.$$

De aici, va rezulta următoarea valoare pentru λ :

$$\lambda = \sum_{i=1}^n \sum_{k=1}^K \mu(z_k^{(i)}) = \sum_{k=1}^K \sum_{i=1}^n \mu(z_k^{(i)}) = \sum_{k=1}^K N_k.$$

Înlocuind această valoare în relația obținută mai sus pentru $\bar{\pi}_k$, va rezulta

$$\bar{\pi}_k = \frac{N_k}{\sum_{k'} N_{k'}}.$$

Putem scrie acum pseudo-codul algoritmului EM pentru rezolvarea acestei mixturi de vectori de distribuții Bernoulli independente.

Inițializare:

atribuie o valori arbitrară în intervalul $(0, 1)$ pentru parametrii π_1, \dots, π_K și $\pi^{(1)}, \dots, \pi^{(K)}$;

Corpul iterativ:

pentru $t = 0, \dots, T - 1$ (cu T fixat în avans)

(sau: până când log-verosimilitatea datelor observabile nu mai crește semnificativ),
(sau...)

execută

Pasul E: pentru $i = 1, \dots, n$ și $k = 1, \dots, K$ calculează

$$\mu(z_k^{(i)}) = \frac{\pi_k \prod_{d=1}^D (p_d^{(k)})^{x_d^{(i)}} (1 - p_d^{(k)})^{1-x_d^{(i)}}}{\sum_{j=1}^K \pi_j \prod_{d=1}^D (p_d^{(j)})^{x_d^{(i)}} (1 - p_d^{(j)})^{1-x_d^{(i)}}};$$

Pasul M: calculează

$$\bar{p}^{(k)} = \frac{\sum_{i=1}^n \mu(z_k^{(i)}) x^{(i)}}{N_k}, \text{ cu } N_k = \sum_{i=1}^n \mu(z_k^{(i)}), \text{ pentru } k = 1, \dots, K;$$

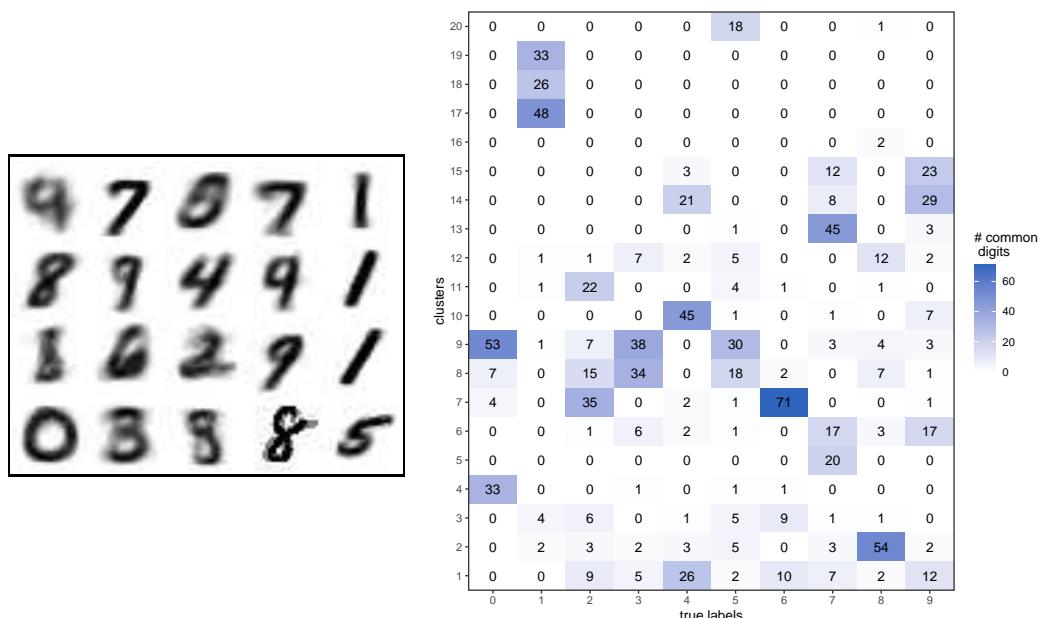
$$\bar{\pi}_k = \frac{N_k}{\sum_{k'} N_{k'}}, \text{ pentru } k = 1, \dots, K;$$

$$\pi_k \leftarrow \bar{\pi}_k, \text{ pentru } k = 1, \dots, K;$$

$$p^{(k)} \leftarrow \bar{p}^{(k)}, \text{ pentru } k = 1, \dots, K;$$

Returnează π_1, \dots, π_K și $\pi^{(1)}, \dots, \pi^{(K)}$;

D. După executarea a 100 de iterații, rezultatul clusterizării este cel ilustrat în figura de mai jos, partea stângă. (În această figură, imaginile corespunzătoare clusterelor sunt însiruite pe coloane.) În partea dreaptă este redată *matricea de confuzie*.⁹⁸⁴ Se poate observa că doar clusterul 3 corespunde unei singure cifre [unice]. Listăm aici în ordinea clusterelor, numărul de cifre [unice] prezente în fiecare cluster: 6, 8, 1, 8, 5, 5, 7, 2, 4, 2, 4, 5, 4, 4, 10, 6, 5, 6, 7, 6. În interiorul clusterelor, cifra care are cele mai multe apariții este (în ordine): 9, 1, 8, 3, 1, 5, 5, 1, 7, 7, 0, 4, 9, 3, 4, 9, 2, 2, 7, 6. Cele mai „confuze“ par a fi clusterele 4, 7, 10, 15 și 18. Pentru celelalte clustere se distinge destul de bine care este cifra care corespunde cel mai bine clusterului respectiv. Se observă că mai multe clustere pot corespunde aceleiași cifre; de exemplu, clusterele 2, 5 și 8 corespund unor modalități diferite de ortografiere a cifrei 1. Similar, clusterele 1, 13 și 16 corespund unor modalități diferite de ortografiere a cifrei 9.



⁹⁸⁴Graficele au fost realizate de către studentul Andi Munteanu (2022).

10. (Algoritmul EM pentru estimarea parametrilor unei mixturi de două distribuții probabiliste categoriale; elaborare / ilustrare într-un un caz particular⁹⁸⁵)
prelucrare de Alina Munteanu și Liviu Ciortuz, după CMU, 2009 spring, Ziv Bar-Joseph, HW5, pr. 1

Un student ia în fiecare dimineață autobuzul ca să vină la universitate. Dacă studentul ia autobuzul 71C, probabilitatea de a găsi un scaun liber lângă geam este μ_{11} , probabilitatea de a găsi un scaun liber lângă interval este μ_{12} , iar probabilitatea ca să călătorească în picioare este μ_{13} , unde $\mu_{11} + \mu_{12} + \mu_{13} = 1$. În schimb, dacă studentul ia autobuzul 500, probabilitățile corespunzătoare sunt $\mu_{21}, \mu_{22}, \mu_{23}$, cu $\mu_{21} + \mu_{22} + \mu_{23} = 1$. Probabilitatea de a lua autobuzul 71C este β_1 , iar probabilitatea de a lua autobuzul 500 este β_2 , cu $\beta_1 + \beta_2 = 1$.

În timpul a n deplasări pe care le-a făcut la universitate, studentul și-a notat de fiecare dată *poziția* p_i pe care a ocupat-o în autobuz — pe un scaun de lângă geam (1), pe un scaun de lângă interval (2) și respectiv în picioare (3) — văzută ca valoare a variabilei aleatoare Pos_i , dar a omis să-și noteze și *tipul* autobuzului $B_i \in \{1, 2\}$, corespunzând numerelor 71C respectiv 500.

În acest exercițiu vi se va cere să folosiți algoritmul EM pentru a estima parametrii $\theta = <\mu_{11}, \mu_{12}, \mu_{13}, \mu_{21}, \mu_{22}, \mu_{23}, \beta_1, \beta_2>$ ai acestui model.

- Expremați $\ln P(Pos, B | \theta)$, adică log-verosimilitatea datelor complete — observabile (Pos_i) și, respectiv, neobservabile (B_i) — în funcție de θ .
- Aplicând formula lui Bayes, calculați probabilitățile condiționate ale variabilelor neobservabile (B_i) în raport cu variabilele observabile (Pos_i) și cu $\theta^{(t)}$, care reprezintă valorile parametrilor θ la iterată t :

$$P(B_i | Pos_i, \theta^{(t)})$$

- Calculați expresia funcției „auxiliare“

$$Q(\theta | \theta^{(t)}) \stackrel{\text{def.}}{=} E_{P(B|Pos, \theta^{(t)})} [\ln P(Pos, B | \theta)]$$

care reprezintă media log-verosimilității datelor complete ($\ln P(Pos, B | \theta)$) în raport cu funcția de probabilitate condițională $P(B | Pos, \theta^{(t)})$.

- Calculați $\theta^{(t+1)} = (\mu_{11}^{(t+1)}, \mu_{12}^{(t+1)}, \mu_{13}^{(t+1)}, \mu_{21}^{(t+1)}, \mu_{22}^{(t+1)}, \mu_{23}^{(t+1)}, \beta_1^{(t+1)}, \beta_2^{(t+1)})$. Acestea sunt valorile parametrilor pentru care se atinge maximul expresiei $Q(\theta | \theta^{(t)})$:

$$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}} Q(\theta | \theta^{(t)}).$$

Indicație: Aceste valori pot fi calculate cu ajutorul metodei multiplicatorilor Lagrange. Metoda aceasta va fi aplicată problemei de optimizare constând din

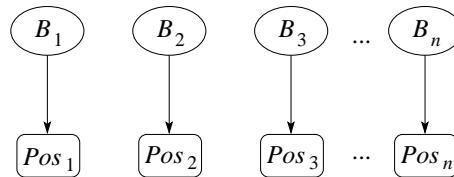
- funcția obiectiv $Q(\theta | \theta^{(t)})$
- restricțiile $\mu_{11} + \mu_{12} + \mu_{13} = 1$, $\mu_{21} + \mu_{22} + \mu_{23} = 1$ și $\beta_1 + \beta_2 = 1$, cu $\mu_{ij} \geq 0$ și $\beta_k \geq 0$.

⁹⁸⁵Pentru varianta generală a algoritmului EM pentru rezolvare de mixturi de K distribuții categoriale, vedeți problema 27.

e. Presupunem că, spre deosebire de situația din enunțul de mai sus, studentul a înregistrat numărul autobuzului B_i , dar nu a înregistrat poziția P_i pe care a avut-o în autobuz. Este oare posibil ca și în această variantă a problemei să folosim algoritmul EM pentru a calcula parametrii modelului? Justificați.

Răspuns:

a. Cele n deplasări făcute de student cu autobuzul sunt [considerate] independente între ele. Prin urmare, nu există nicio condiționare între diversele variabile B_i . Mai departe, de valoarea acestor variabile ascunse / neobservabile depinde *poziția* liberă observată în autobuz Pos_i . Așadar, putem asocia acestei probleme *modelul grafic* următor:



Probabilitatea comună a datelor complete (observabile și neobservabile) în funcție de θ este:

$$P(Pos, B | \theta) \stackrel{indep.}{=} \prod_{i=1}^n P(Pos_i, B_i | \theta)$$

Aplicând regula de înmulțire a probabilităților, obținem:

$$P(Pos, B | \theta) = \prod_{i=1}^n P(Pos_i | B_i, \theta) \cdot P(B_i | \theta)$$

Pentru $i \in \{1, \dots, n\}$, $j \in \{1, 2\}$ și $k \in \{1, 2, 3\}$ considerăm următoarele *variabile-indicator*:

$$b_{ij} = \begin{cases} 1 & \text{dacă } B_i = j \\ 0 & \text{altfel} \end{cases} \quad p_{ik} = \begin{cases} 1 & \text{dacă } Pos_i = k \\ 0 & \text{altfel} \end{cases}$$

Cu alte cuvinte b_{ij} este 1 dacă în ziua i studentul a mers cu autobuzul j , și 0 în caz contrar, cu convenția că valoarea $j = 1$ corespunde autobuzului 71C, iar $j = 2$ autobuzului 500. Similar, p_{ik} este 1 dacă în ziua i studentul a mers într-o poziție de tip k , unde $k = 1$ înseamnă pe scaun lângă geam, $k = 2$ – pe scaun lângă interval, iar $k = 3$ – în picioare.

Cu aceste notații, putem exprima (în manieră compactă) verosimilitatea datelor complete astfel:

$$\begin{aligned} P(Pos, B | \theta) &= \prod_{i=1}^n P(B_i | \theta) \cdot P(Pos_i | B_i, \theta) = \prod_{i=1}^n \left(\prod_{j=1}^2 \beta_j^{b_{ij}} \prod_{k=1}^3 (\mu_{jk})^{b_{ij} p_{ik}} \right) \\ &= \prod_{i=1}^n \prod_{j=1}^2 \left(\beta_j \prod_{k=1}^3 \mu_{jk}^{p_{ik}} \right)^{b_{ij}} \end{aligned}$$

În consecință, log-verosimilitatea datelor complete este:

$$\ln P(Pos, B | \theta) = \sum_{i=1}^n \sum_{j=1}^2 b_{ij} \left(\ln \beta_j + \sum_{k=1}^3 p_{ik} \ln \mu_{jk} \right)$$

b. Probabilitatea condiționată cerută este:

$$\begin{aligned} P(B_i = j \mid Pos_i = k, \theta^{(t)}) &\stackrel{T. Bayes}{=} \frac{P(Pos_i = k \mid B_i = j, \theta^{(t)})P(B_i = j \mid \theta^{(t)})}{\sum_{j'=1}^2 P(Pos_i = k \mid B_i = j', \theta^{(t)})P(B_i = j' \mid \theta^{(t)})} \\ &= \frac{\beta_j^{(t)} \prod_{k=1}^3 (\mu_{jk}^{(t)})^{p_{ik}}}{\sum_{j'=1}^2 \beta_{j'}^{(t)} \prod_{k=1}^3 (\mu_{j'k}^{(t)})^{p_{ik}}} \end{aligned}$$

Prin urmare, mediile variabilelor neobservabile b_{ij} la iterarea t vor avea și ele aceeași expresie:

$$E[b_{ij}^{(t)}] \stackrel{not.}{=} E[b_{ij} \mid Pos, \theta^{(t)}] = \frac{\beta_j^{(t)} \prod_{k=1}^3 (\mu_{jk}^{(t)})^{p_{ik}}}{\sum_{j'=1}^2 \beta_{j'}^{(t)} \prod_{k=1}^3 (\mu_{j'k}^{(t)})^{p_{ik}}}$$

c. Calculăm media log-verosimilității datelor complete combinând rezultatele de la punctele a și b și ținând cont de proprietatea de liniaritate a mediilor:

$$Q(\theta \mid \theta^{(t)}) = \sum_{i=1}^n \sum_{j=1}^2 E[b_{ij}^{(t)}] \left(\ln \beta_j + \sum_{k=1}^3 p_{ik} \ln \mu_{jk} \right)$$

d. Pentru a calcula $\theta^{(t+1)}$, trebuie maximizată funcția $Q(\theta \mid \theta^{(t)})$ în raport cu θ . Vom calcula separat valorile optime ale parametrilor, și anume mai întâi pentru β_j , cu $j \in \{1, 2\}$ și apoi pentru μ_{jk} , cu $j \in \{1, 2\}$ și $k \in \{1, 2, 3\}$.

Stim că variabilele β_j se supun restricției $\beta_1 + \beta_2 = 1$. Reținând din expresia funcției $Q(\theta \mid \theta^{(t)})$ doar termenii care conțin variabilele β_j , rezultă că vom avea de optimizat funcția

$$\sum_{i=1}^n (E[b_{i1}^{(t)}] \ln \beta_1 + E[b_{i2}^{(t)}] \ln \beta_2) = \sum_{i=1}^n (E[b_{i1}^{(t)}] \ln \beta_1 + E[b_{i2}^{(t)}] \ln (1 - \beta_1))$$

Derivând ultima expresie în raport cu β_1 și egalând apoi cu 0, obținem:⁹⁸⁶

$$\begin{aligned} \sum_{i=1}^n \left(\frac{E[b_{i1}^{(t)}]}{\beta_1} - \frac{E[b_{i2}^{(t)}]}{1 - \beta_1} \right) = 0 &\Leftrightarrow (1 - \beta_1) \sum_{i=1}^n E[b_{i1}^{(t)}] = \beta_1 \sum_{i=1}^n E[b_{i2}^{(t)}] \Leftrightarrow \\ \sum_{i=1}^n E[b_{i1}^{(t)}] = \beta_1 \sum_{i=1}^n (E[b_{i1}^{(t)}] + E[b_{i2}^{(t)}]) &\Leftrightarrow \sum_{i=1}^n E[b_{i1}^{(t)}] = \beta_1 \sum_{i=1}^n \underbrace{E[b_{i1}^{(t)} + b_{i2}^{(t)}]}_1 \Leftrightarrow \\ \sum_{i=1}^n E[b_{i1}^{(t)}] = n\beta_1 &\Leftrightarrow \beta_1 = \frac{1}{n} \sum_{i=1}^n E[b_{i1}^{(t)}] \end{aligned}$$

Este imediat că

$$\beta_2 = \frac{1}{n} \sum_{i=1}^n E[b_{i2}^{(t)}],$$

iar atât β_1 cât și β_2 se situează în intervalul $[0, 1]$, întrucât mediile $E[b_{ij}^{(t)}]$ sunt de fapt tot niște probabilități.

⁹⁸⁶Se demonstrează ușor că luând baza logaritmului supraunitară (așa cum am procedat de-a lungul întregului capitol), rezultă că soluțiile derivatelor parțiale de ordinul întâi — care sunt calculate aici și mai jos — reprezintă într-adevăr punctul de maxim al funcției $Q(\theta \mid \theta^{(t)})$.

Pentru a calcula valorile optime ale parametrilor μ_{jk} , cu indicii $j \in \{1, 2\}$ și $k \in \{1, 2, 3\}$ vom apela la metoda multiplicatorilor Lagrange întrucât stim că acești parametri se supun constrângerii $\mu_{j1} + \mu_{j2} + \mu_{j3} = 1$. Așadar, vom avea problema de optimizare cu obiectivul $\max Q(\theta | \theta^{(t)})$ și restricțiile $\sum_{k=1}^3 \mu_{jk} - 1 = 0$ și $\mu_{ij} \geq 0$. Funcția Lagrange asociată acestei probleme este:

$$\Lambda(\mu, \lambda) = \sum_{i=1}^n \sum_{j=1}^2 \sum_{k=1}^3 E[b_{ij}^{(t)}] p_{ik} \ln \mu_{jk} + \sum_{j=1}^2 \lambda_j \left(\sum_{k=1}^3 \mu_{jk} - 1 \right)$$

Egalând cu 0 derivata parțială a lui $\Lambda(\mu, \lambda)$ în raport cu μ_{jk} , obținem:

$$\frac{\partial \Lambda(\mu, \lambda)}{\partial \mu_{jk}} = 0 \Leftrightarrow \frac{1}{\mu_{jk}} \cdot \sum_{i=1}^n E[b_{ij}^{(t)}] p_{ik} + \lambda_j = 0 \Leftrightarrow \mu_{jk} = -\frac{1}{\lambda_j} \cdot \sum_{i=1}^n E[b_{ij}^{(t)}] p_{ik}, \text{ cu } j \in \{1, 2\}$$

Stim că

$$\frac{\partial \Lambda(\mu, \lambda)}{\partial \lambda_j} = 0 \Leftrightarrow \sum_{k=1}^3 \mu_{jk} - 1 = 0, \text{ pentru } j \in \{1, 2\}$$

Dacă în această relație vom înlocui μ_{jk} cu valorile obținute anterior, vom determina valoarea lui λ_j pentru $j \in \{1, 2\}$:

$$\begin{aligned} \sum_{k=1}^3 \left(-\frac{1}{\lambda_j} \cdot \sum_{i=1}^n E[b_{ij}^{(t)}] p_{ik} \right) - 1 &= 0 \Leftrightarrow \frac{1}{\lambda_j} \cdot \sum_{k=1}^3 \sum_{i=1}^n E[b_{ij}^{(t)}] p_{ik} = -1 \\ \Leftrightarrow \lambda_j &= -\sum_{k=1}^3 \sum_{i=1}^n E[b_{ij}^{(t)}] p_{ik}, \text{ cu } j \in \{1, 2\}. \end{aligned}$$

Prin urmare, valorile căutate pentru μ_{jk} sunt:

$$\begin{aligned} \mu_{jk}^{(t+1)} &= -\frac{1}{\lambda_j} \cdot \sum_{i=1}^n E[b_{ij}^{(t)}] p_{ik} = \frac{\sum_{i=1}^n E[b_{ij}^{(t)}] p_{ik}}{\sum_{l=1}^3 \sum_{i=1}^n E[b_{ij}^{(t)}] p_{il}} \\ &= \frac{\sum_{i=1}^n E[b_{ij}^{(t)}] p_{ik}}{\sum_{i=1}^n \sum_{l=1}^3 E[b_{ij}^{(t)}] p_{il}}, \text{ cu } j \in \{1, 2\} \text{ și } k \in \{1, 2, 3\} \end{aligned}$$

Este imediat că aceste valori se situează în intervalul $[0, 1]$.

- e. Așa cum se observă și din modelul grafic reprezentat la punctul a, dacă se cunoaște tipul autobuzului B_i , nu se pot estima parametrii [reprezentând distribuția] variabilei care reprezintă poziția Pos_i ocupată de student în autobuz. Altfel spus, variabilele observabile B_i nu oferă nicio informație relevantă pentru a putea calcula variabilele neobservabile Pos_i .

11. (Rezolvarea unei mixturi de doi vectori de distribuții categoriale, folosind presupoziția de independentă condițională de tip Bayes Naiv,⁹⁸⁷ aka algoritmul Bayes Naiv nesupervizat)

*prelucrare de Liviu Ciortuz, după
■ □ • ○ * CMU, 2014 spring, A. Singh, B. Poczos, HW3, pr. 2.1*

Fie setul de date etichetate $\mathcal{D} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$.

Atunci când se folosește presupoziția de independentă condițională specifică clasificatorului Bayes Naiv, putem scrie verosimilitatea [comună a] datelor \mathcal{D} astfel:

$$P(\mathcal{D}) \stackrel{i.i.d.}{=} \prod_{i=1}^n P(X_i, Y_i) = \prod_{i=1}^n (P(X_i|Y_i) \cdot P(Y_i)) = \prod_{i=1}^n \left(P(Y_i) \cdot \prod_{j=1}^d P(X_i^j|Y_i) \right)$$

Notația X_i^j , unde $j \in \{1, \dots, d\}$, desemnează atributul de intrare j al instanței etichetate (X_i, Y_i) .

Observație: În cele ce urmează vom nota variabilele aleatoare cu majuscule, iar valorile lor cu minuscule.⁹⁸⁸

Presupunem că $y_i \in \{0, 1\} \forall i$ și $x_i^j \in \{1, 2, \dots, V\} \forall i, j$. Așadar, considerăm (doar pentru simplitate) că V , numărul de valori, este același pentru toate atribuțele X^j , cu $j \in \{1, \dots, d\}$ și, în plus, mulțimile de valori ale acestor atrbute coincid. Vom nota $\pi_0 = P(Y = 0)$, $\pi_1 = P(Y = 1)$ și $\pi = (\pi_0, \pi_1)$. Similar, pentru fiecare $j \in \{1, \dots, d\}$, vom nota parametrii distribuțiilor $P(X^j|Y)$ astfel: $\beta_0^j = (\beta_{10}^j, \beta_{20}^j, \dots, \beta_{V0}^j)$ și $\beta_1^j = (\beta_{11}^j, \beta_{21}^j, \dots, \beta_{V1}^j)$, deci $\beta_{k0}^j \stackrel{\text{not.}}{=} P(X^j = k|Y = 0)$ și $\beta_{k1}^j \stackrel{\text{not.}}{=} P(X^j = k|Y = 1)$, pentru fiecare $k \in \{1, 2, \dots, V\}$. În sfârșit, $\beta \stackrel{\text{not.}}{=} (\beta_0^1, \beta_1^1, \dots, \beta_0^d, \beta_1^d)$ și $\theta \stackrel{\text{not.}}{=} (\pi, \beta)$.

Evident, în cazul în care dispunem de date etichetate — adică variabila Y_i este „observabilă“, pentru toate valorile lui i —, vom putea obține ușor estimări (în sensul MLE) pentru parametrii π și β , folosind procedeul clasic de enumerare (engl., counting) și normalizare.

În acest exercițiu vom presupune însă că nu dispunem de date etichetate. Scopul nostru este de a arăta că și în situația aceasta putem estima parametrii θ (adică π și β^j , cu $j = 1, \dots, d$), și anume folosind algoritmul de estimare-maximizare (EM).

a. Se știe — vedeti problema 1 — că funcția de log-verosimilitate a datelor observabile ($\log P(x_1, \dots, x_n|\theta)$) este minorată de o funcție de două argumente $F(q, \theta)$, unde q este o distribuție probabilistică oarecare definită peste variabilele-indicator z_i .⁹⁸⁹

Vă cerem⁹⁹⁰

⁹⁸⁷În cazul $d = 1$ (vedeți notațiile din problemă), se obține varianta de algoritm EM pentru o mixtură de două distribuții categoriale, care a fost deja ilustrată la problema 10. Pentru mixturi de K distribuții categoriale, vedeți problema 27.

Pentru un exercițiu (inclusând implementare) în care face clusterizare de caractere scrise de mână, aplicând un algoritm EM de tipul celui prezentat în problema de față, însă lucrând cu variabile Bernoulli în locul variabilelor categoriale, vedeți exercițiul 9.

⁹⁸⁸De exemplu, Y_i este o variabilă aleatoare, iar y_i desemnează o instanțiere a lui Y_i la una din valorile ei posibile.

⁹⁸⁹Observație: Veți ține cont că rolul celor variabile z_i (de la problema 1) va fi jucat aici de către y_i .

⁹⁹⁰Sugestie: Procedați în mod similar cu rezolvarea dată la problema 1 punctul a.

- să scrieți expresia funcției de (log-)verosimilitate pe care urmărește să o maximizeze algoritmul EM în contextul prezentului exercițiu;⁹⁹¹
- să introduceți apoi probabilitățile $q(y_i)$ în mod forțat — însă într-o manieră convenabilă — în expresia scrisă anterior pentru funcția de log-verosimilitate;
- și, în sârșit, folosind inegalitatea lui Jensen, să obțineți expresia marginii inferioare $F(q, \theta)$ pentru funcția de log-verosimilitate care a fost menționată mai sus.

Observație importantă: Spre deosebire de cum am procedat la toate exercițiile de până acum legate de algoritmul EM, aici nu vom [mai] lucra cu funcția „auxiliară“ Q ,⁹⁹² ci direct cu funcția F .

- b. Pasul E: Calculați probabilitatea fiecărei asignări posibile [la o anumită clasă] pentru fiecare instanță, date fiind valorile actuale ale parametrilor π și β . Concret, veți determina valorile optime pentru $q(y_i)$ (i.e., $q(Y_i = 0)$ și $q(Y_i = 1)$), folosind derivele partiale ale funcției $F(q(y_i), \theta)$ care a fost dedusă la punctul precedent.
- c. Pasul M: Deduceți regulile pentru actualizarea parametrilor π și β în funcție de probabilitățile de asignare [la o anumită clasă] pentru fiecare instanță, $q(y_i)$, care au fost calculate la pasul E.
- d. Schițați o strategie convenabilă pentru inițializarea parametrilor (π și β) și indicați motivele care au stat la baza alegерii / conceperii respectivei strategii.

Răspuns:

- a. Algoritmul EM urmărește să maximizeze funcția de log-verosimilitate a datelor observabile, $\log P(x_1, \dots, x_n | \theta)$. Fără a mai menționa parametrul θ în calculul care urmează, putem scrie:

$$\begin{aligned}
\log P(x_1, \dots, x_n) &\stackrel{i.i.d.}{=} \log \prod_{i=1}^n P(x_i) = \log \prod_{i=1}^n \sum_{y_i \in \{0,1\}} P(x_i, y_i) \\
&= \sum_{i=1}^n \log \sum_{y_i \in \{0,1\}} P(x_i, y_i) = \sum_{i=1}^n \log \sum_{y_i \in \{0,1\}} P(x_i | y_i) P(y_i) \\
&= \sum_{i=1}^n \log \sum_{y_i \in \{0,1\}} P(y_i) \prod_{j=1}^d P(x_i^j | y_i) \\
&= \sum_{i=1}^n \log \sum_{y_i \in \{0,1\}} q(y_i) \frac{P(y_i) \prod_{j=1}^d P(x_i^j | y_i)}{q(y_i)} \\
&\stackrel{\text{def.}}{=} \sum_{i=1}^n \log E_q \left[\frac{P(y_i) \prod_{j=1}^d P(x_i^j | y_i)}{q(y_i)} \right] \\
&\geq \sum_{i=1}^n E_q \left[\log \frac{P(y_i) \prod_{j=1}^d P(x_i^j | y_i)}{q(y_i)} \right] \quad (\text{Jensen})
\end{aligned}$$

⁹⁹¹Indicație: Verosimilitatea unei instanțe neeticatate, $P(x_i | \theta)$ se poate obține însumând probabilitățile corespunzătoare tuturor asignărilor posibile y_i ale variabilei latente / neobservabile Y_i .

⁹⁹²Vedeți finalul rezolvării problemei 1 de la acest capitol.

$$\begin{aligned}
&\stackrel{\text{def.}}{=} \sum_{i=1}^n \sum_{y_i \in \{0,1\}} q(y_i) \log \frac{P(y_i) \prod_{j=1}^d P(x_i^j | y_i)}{q(y_i)} \\
&= \sum_{i=1}^n \sum_{y_i \in \{0,1\}} q(y_i) \left[\log P(y_i) + \left(\sum_{j=1}^d \log P(x_i^j | y_i) \right) - \log q(y_i) \right].
\end{aligned}$$

Așadar, reintroducând parametrul θ , marginea inferioară a funcției de log-verosimilitate a datelor observabile x_1, \dots, x_n este:

$$F(q, \theta) = \sum_{i=1}^n \sum_{y_i \in \{0,1\}} q(y_i) \left[\log P(y_i | \theta) + \left(\sum_{j=1}^d \log P(x_i^j | y_i, \theta) \right) - \log q(y_i) \right]. \quad (410)$$

b. La pasul E, pentru $i = 1, \dots, n$, vom actualiza $q(y_i)$ în funcție de valorile curente pentru parametrii π și β , mai exact în funcție de $P_\pi(y_i) \stackrel{\text{not.}}{=} P(y_i | \pi)$ și $P_\beta(x_i^j | y_i) \stackrel{\text{not.}}{=} P(x_i^j | y_i, \beta)$.⁹⁹³ Calculând derivata parțială a funcției F în raport cu $q(y_i)$, obținem:

$$\log P_\pi(y_i) + \sum_{j=1}^d \log P_\beta(x_i^j | y_i) - \log q(y_i) - 1,$$

iar apoi egalându-o cu 0, deducem ușor următoarele reguli de actualizare:

$$\begin{aligned}
y_i = 0 : \quad q(Y_i = 0) &\propto P_\pi(Y_i = 0) \prod_{j=1}^d P_\beta(x_i^j | Y_i = 0) \\
y_i = 1 : \quad q(Y_i = 1) &\propto P_\pi(Y_i = 1) \prod_{j=1}^d P_\beta(x_i^j | Y_i = 1),
\end{aligned}$$

unde simbolul \propto înseamnă *proporțional cu*. Se poate observa ușor că factorul de proporționalitate în precedentele două relații este același. Așadar, ținând cont de restricția $q(y_i = 0) + q(y_i = 1) = 1$, putem scrie:

$$\begin{aligned}
q(Y_i = 0) &= \frac{P_\pi(Y_i = 0) \prod_{j=1}^d P_\beta(x_i^j | Y_i = 0)}{\sum_{\ell \in \{0,1\}} P_\pi(Y_i = \ell) \prod_{j=1}^d P_\beta(x_i^j | Y_i = \ell)} \\
q(Y_i = 1) &= \frac{P_\pi(Y_i = 1) \prod_{j=1}^d P_\beta(x_i^j | Y_i = 1)}{\sum_{\ell \in \{0,1\}} P_\pi(Y_i = \ell) \prod_{j=1}^d P_\beta(x_i^j | Y_i = \ell)}
\end{aligned}$$

c. La pasul M, cunoaștem valorile $q(y_i)$ și vom maximiza funcția [obiectiv] $F(q, \theta)$ în raport cu parametrul θ (i.e., π și β). După calcule relativ simple, cu

⁹⁹³Pentru a determina valorile optime pentru $q(y_i = 0)$ și $q(y_i = 1)$, am putea proceda ca și la exercițiile precedente:

- i. fie ținând cont de relația $q(y_i = 0) + q(y_i = 1) = 1$, $\forall i \in \{1, \dots, n\}$ și înlocuind, de exemplu, $q(y_i = 1)$ în funcție de $q(y_i = 0)$ în relația (410), după care calculăm rădăcina derivatei $\frac{\partial F(q, \theta)}{\partial q(y_i = 0)}$ și verificăm dacă această rădăcină aparține intervalului $[0, 1]$;
- ii. fie folosim metoda multiplicatorilor Lagrange pentru a rezolva (pentru fiecare $i \in \{1, \dots, n\}$) problema de optimizare care are funcția obiectiv (410) și restricția $q(y_i = 0) + q(y_i = 1) = 1$.

Însă rezolvarea pentru care am optat aici este de un alt tip, care într-o anumită măsură seamănă cu metoda i, dar inversează cei doi pași.

care de acum suntem obișnuiți, vom obține următoarele reguli de actualizare pentru parametrii π :⁹⁹⁴

$$\begin{cases} \pi_0 \stackrel{\text{not.}}{=} P(Y = 0) \propto \sum_{i=1}^n q(Y_i = 0) \\ \pi_1 \stackrel{\text{not.}}{=} P(Y = 1) \propto \sum_{i=1}^n q(Y_i = 1) \end{cases} \Rightarrow \begin{cases} \pi_0 = \frac{1}{n} \sum_{i=1}^n q(Y_i = 0) \\ \pi_1 = \frac{1}{n} \sum_{i=1}^n q(Y_i = 1) \end{cases}$$

și apoi pentru parametrii β :⁹⁹⁵

$$\begin{aligned} & \begin{cases} \beta_{v0}^j \stackrel{\text{not.}}{=} P(X^j = v | Y = 0) \propto \sum_{i=1}^n q(Y_i = 0) \cdot 1_{\{x_i^j=v\}} \\ \beta_{v1}^j \stackrel{\text{not.}}{=} P(X^j = v | Y = 1) \propto \sum_{i=1}^n q(Y_i = 1) \cdot 1_{\{x_i^j=v\}}, \end{cases} \\ & \Rightarrow \begin{cases} \beta_{v0}^j = \frac{\sum_{i=1}^n q(Y_i = 0) \cdot 1_{\{x_i^j=v\}}}{\sum_{\ell \in \{0,1\}} \sum_{i=1}^n q(Y_i = \ell) \cdot 1_{\{x_i^j=v\}}} \\ \beta_{v1}^j = \frac{\sum_{i=1}^n q(Y_i = 1) \cdot 1_{\{x_i^j=v\}}}{\sum_{\ell \in \{0,1\}} \sum_{i=1}^n q(Y_i = \ell) \cdot 1_{\{x_i^j=v\}}} \end{cases} \end{aligned}$$

Notăția $1_{\{\dots\}}$ folosită mai sus desemnează o *funcție-indicator*. Așa cum știm, ea ia valoarea 1 atunci când condiția specificată [între acolade] este satisfăcută și valoarea 0 în caz contrar.

Este interesant *de observat* că relațiile obținute la acest pas al algoritmului EM sunt foarte similare cu relațiile corespunzătoare folosite [pentru estimarea parametrilor] de către clasificatorul Bayes Naiv (supervizat).

d. O strategie convenabilă pentru inițializarea parametrilor mixturii ar trebui să țină cont de cunoștințele / informațiile pe care le avem în legătură cu o soluție plauzibilă. Dacă avem motive să credem că o anumită clasă este mai probabilă decât cealaltă clasă, vom putea „injecta“ această informație în valoarea folosită la inițializarea lui π . Pentru parametrii β , dacă intuim că o anumită valoare (k) a atributului X^j se coreleză mai bine cu (adică este mai „indicativă“ în raport cu) o anumită clasă ℓ , atunci vom da o valoare (probabilitate) mai mare parametrului $\beta_{k\ell}^j$, care corespunde respectivei valori a atributului X^j . Dacă nu avem niciun fel de informații / cunoștințe a priori, putem să inițializăm acești parametri în mod aleator.

⁹⁹⁴Pentru actualizarea parametrilor π_0 și π_1 , putem folosi oricare dintre metodele *i* și *ii* menționate la nota de subsol precedentă.

În cazul *i*, reținând din expresia (410) doar termenii care-l conțin pe $P(Y = y_i | \theta)$, vom avea:

$$\frac{\partial}{\partial P_\pi(Y = 0)} \left(\sum_{i=1}^n (q(y_i = 0) \log P_\pi(Y = 0) + q(y_i = 1) \log(1 - P_\pi(Y = 0))) \right) = 0 \Leftrightarrow \text{ș.a.m.d.}$$

În cazul *ii*, lagrangeanul de optimizat va fi

$$\begin{aligned} \mathcal{L}(\lambda) &= F(q, \theta) + \lambda(1 - P_\pi(Y = 0) - P_\pi(Y = 1)), \text{ deci} \\ \frac{\partial \mathcal{L}(\lambda)}{\partial P_\pi(Y = 0)} &= 0 \Leftrightarrow \frac{1}{P_\pi(Y = 0)} \sum_{i=1}^n q(Y_i = 0) = \lambda \Leftrightarrow P_\pi(Y = 0) = \frac{1}{\lambda} \sum_{i=1}^n q(Y_i = 0) \end{aligned}$$

și similar pentru $P_\pi(Y = 1)$.

⁹⁹⁵Folosim metoda multiplicatorilor lui Lagrange (*ii*), din cauza restricțiilor $\sum_{v=1}^V \beta_{v0}^j = 1$ și $\sum_{v=1}^V \beta_{v1}^j = 1$. Detaliile de calcul sunt similare cu cele din ultima parte de la nota de subsol precedentă.

12. (Algoritmul EM pentru mixturi de distribuții categoriale;
 [aplicare la] identificarea domeniilor semantice
 asociate cuvintelor dintr-un document-text)
- □ • ○ CMU, 2012 fall, E. Xing, A. Singh, HW3, pr. 3

În acest exercițiu vi se va cere să deduceți relațiile corespunzătoare pașilor E și M din corpul iterativ al algoritmului EM pentru „modelarea“ variabilelor „neobservabile“ care desemnează *domeniile semantice* (engl., topics) implicate în generarea unui document oarecare de tip text.

Vom considera că fiecare *cuvânt* din documentul dat este reprezentat de o variabilă aleatoare w care poate lua valorile $1, \dots, V$ relativ la un *vocabular* dat. De fapt, în cele ce urmează vom desemna fiecare cuvânt w printr-un *vector-indicator* format din V componente astfel încât $w(i) = 1$ dacă w ia valoarea cuvântului de pe poziția i din vocabular și 0 în caz contrar. Așadar, $\sum_{i=1}^V w(i) = 1$.

Dată fiind un *document* constituit din cuvintele w_j , $j = 1, \dots, N$, unde N este lungimea documentului, vom presupune că aceste cuvinte sunt generate de către o mixtură de K distribuții categoriale:

$$\begin{aligned} P(w_j) &= \sum_{s=1}^K \pi_s P(w_j | \beta_s) \\ P(w_j | \beta_s) &= \prod_{i=1}^V P(w_j(i) = 1 | Z_j = s) = \prod_{i=1}^V \beta_s(i)^{w_j(i)}, \end{aligned}$$

unde

$\pi_s \stackrel{\text{not.}}{=} P(Z = s)$ este probabilitatea (a priori) ca variabila latentă Z , care desemnează domeniul semantic asociat unui cuvânt oarecare, să ia valoarea s ;

$\beta_s \stackrel{\text{not.}}{=} (\beta_s(1), \dots, \beta_s(i), \dots, \beta_s(V))$, cu $\beta_s(i) \geq 0$ pentru $i = 1, \dots, V$ și $\sum_{i=1}^V \beta_s(i) = 1$, pentru fiecare $s = 1, \dots, K$. Vectorul β_s determină distribuția categorială asociată cu domeniul semantic s ;

$\beta_s(i) \stackrel{\text{not.}}{=} P(w_j(i) = 1 | Z_j = s)$ desemnează probabilitatea ca $w_j \stackrel{\text{not.}}{=} (w_j(1), \dots, w_j(V))$ să fi fost generat de către domeniul semantic s .

Observație: Remarcați faptul că acest model nu ține cont de ordinea / asocierea cuvintelor din documentul considerat. Deși faptul aceasta constituie o încălcare evidentă a proprietăților definitoare ale unui text, modelul propus aici are totuși o utilitate practică dovedită.⁹⁹⁶

a. Referitor la pasul E al algoritmului EM, pentru fiecare cuvânt w_j , calculați $q_j(s) \stackrel{\text{not.}}{=} P(Z_j = s | w_j; \theta')$, probabilitatea (a posteriori) ca, dat un cuvânt w_j , acesta să corespundă unui anumit domeniu semantic s din ansamblul celor K domenii semantice considerate. Prin θ desemnăm ansamblul parametrilor din modelul de mixtură considerat: π_s și β_s cu $s = 1, \dots, K$, iar θ' notează valoarea acestor parametri la pasul de inițializare [al algoritmului EM], respectiv valoarea obținută la iterația precedentă.

⁹⁹⁶Vedeți capitolul *Word Sense Disambiguation* din cartea *Foundations of Statistical Natural Language Processing*, Christopher Manning, Hinrich Schütze, MIT Press, 2002, pag 252-256.

b. Pentru pasul M, determinați valoarea parametrului θ care maximizează marginea inferioară (vedeți problema 1) pentru log-verosimilitatea datelor observabile, i.e., a cuvintelor din documentul dat:

$$\ell(w|\theta) \stackrel{not.}{=} \ln \prod_{j=1}^N P(w_j|\theta).$$

Indicație: Procedând în mod clasic, adică sumând în raport cu variabila latentă s care reprezintă tematica / domeniul semantic, putem scrie funcția de log-verosimilitate $\ell(w|\theta)$ astfel:

$$\ell(w|\theta) = \sum_{j=1}^N \ln \sum_s P(w_j, s|\theta) = \sum_{j=1}^N \ln \sum_s q_j(s) \frac{P(w_j, s|\theta)}{q_j(s)},$$

unde cantitățile $q_j(s)$ cu $j \in \{1, \dots, V\}$ și $s \in \{1, \dots, K\}$ au fost calculate la pasul E al iterației curente. Mai departe, folosind inegalitatea lui Jensen (vedeți pr. 79 de la capitolul *Fundamente*), obținem:

$$\begin{aligned} \ell(w|\theta) &\geq \sum_{j=1}^N \sum_s q_j(s) \ln \frac{P(w_j, s|\theta)}{q_j(s)} = \\ &= \sum_{j=1}^N \sum_s q_j(s) \ln P(w_j, s|\theta) - \sum_{j=1}^N \sum_s q_j(s) \ln q_j(s) = \\ &= \sum_{j=1}^N \sum_s q_j(s) \ln P(w_j, s|\theta) + \sum_{j=1}^N H(q_j). \end{aligned}$$

Am notat cu $H(q_j)$ entropia $\sum_s q_j(s) \ln q_j(s) = \sum_s P(s|w_j; \theta') \ln P(s|w_j; \theta')$.⁹⁹⁷ Evident, aceasta nu depinde de θ (deși depinde de θ'). În concluzie, la pasul M veți calcula acea valoare a lui θ pentru care se atinge maximul *funcției auxiliare*

$$Q(\theta|\theta') \stackrel{not.}{=} \sum_{j=1}^N \sum_s q_j(s) \ln P(w_j, s|\theta).$$

Răspuns:

a. La pasul E, calculăm probabilitățile a posteriori $q_j(s) \stackrel{not.}{=} P(Z_j = s|w_j; \theta')$. Pentru conveniență, în rezolvarea / calculul care urmează vom renunța la a folosi accentul ('') pentru componentele din setul de parametri $\theta' = (\pi', \beta'_1, \dots, \beta'_K)$ indicat în enunț — ar fi trebuit, de fapt, să scriem $\theta^{(t)} = (\pi^{(t)}, \beta_1^{(t)}, \dots, \beta_K^{(t)})$ — deoarece notațiile ar deveni prea complicate. (Cititorul atent va ști din context la ce anume se referă notația $\theta = (\pi, \beta_1, \dots, \beta_K)$.)

$$\begin{aligned} q_j(s) &\stackrel{not.}{=} P(Z_j = s|w_j; \theta) \stackrel{\text{Bayes T.}}{=} \frac{P(w_j|Z_j = s; \theta) P(Z_j = s|\theta)}{P(w_j|\theta)} \\ &= \frac{\pi_s P(w_j|\beta_s)}{\sum_{s'=1}^K \pi_{s'} P(w_j|\beta_{s'})} = \frac{\pi_s \prod_{l=1}^V \beta_s(l)^{w_j(l)}}{\sum_{s'=1}^K \pi_{s'} \prod_{l=1}^V \beta_{s'}(l)^{w_j(l)}}. \end{aligned}$$

⁹⁹⁷Pentru simplitate, am presupus că atât funcția de log-verosimilitate cât și entropia se calculează folosind pentru logaritm o aceeași bază, supraunitară.

b. Conform *Indicației* din enunț, calculăm mai întâi funcția auxiliară asociată iterației curente (t) :

$$\begin{aligned}
 Q(\theta|\theta') &\stackrel{\text{not.}}{=} \sum_{j=1}^N \sum_{s=1}^K q_j(s) \ln P(w_j, Z_j = s|\theta) \\
 &= \operatorname{argmax}_{\theta} \sum_{j=1}^N \sum_{s=1}^K q_j(s) \ln P(w_j|Z_j = s; \theta) P(Z_j = s|\theta) \\
 &= \operatorname{argmax}_{\theta} \sum_{j=1}^N \sum_{s=1}^K q_j(s) \ln \left(\pi_s \prod_{l=1}^V \beta_s(l)^{w_j(l)} \right) \\
 &= \operatorname{argmax}_{\theta} \sum_{j=1}^N \sum_{s=1}^K [q_j(s) \ln \pi_s + q_j(s) \sum_{l=1}^V \ln \beta_s(l)^{w_j(l)}] \\
 &= \operatorname{argmax}_{\theta} \sum_{j=1}^N \sum_{s=1}^K [q_j(s) \ln \pi_s + q_j(s) \sum_{l=1}^V w_j(l) \ln \beta_s(l)]. \tag{411}
 \end{aligned}$$

Se știe că $\theta^{(t+1)} \stackrel{\text{def.}}{=} \operatorname{argmax}_{\theta} Q(\theta|\theta')$. Problema aceasta de optimizare se poate rezolva maximizând separat $Q(\theta|\theta')$ în raport cu parametrii $\beta_s(l)$ și respectiv în raport cu parametrii π_s (unde $s \in \{1, \dots, K\}$).⁹⁹⁸

În ce privește optimizarea lui $Q(\theta|\theta')$ în raport cu $\beta_s(l)$:

După ce eliminăm din expresia (411) termenii care sunt constanți în raport cu β_s , obținem:

$$\sum_{j=1}^N q_j(s) \sum_{l=1}^V w_j(l) \ln \beta_s(l).$$

Apoi, deoarece $\beta_s \stackrel{\text{not.}}{=} (\beta_s(1), \dots, \beta_s(i), \dots, \beta_s(V))$, cu restricțiile $\beta_s(i) \geq 0$ pentru $i = 1, \dots, V$ și $\sum_{i=1}^V \beta_s(i) = 1$, vom folosi metoda multiplicatorilor lui Lagrange pentru a ne asigura că cea de-a doua restricție este satisfăcută.⁹⁹⁹ Așadar, introducând variabila / multiplicatorul Lagrange $\lambda \in \mathbb{R}$, pornind de la expresia anterior obținută vom scrie funcția „lagrangeană“

$$\mathcal{L}(\beta_s(l)) = \sum_{j=1}^N q_j(s) \sum_{l=1}^V w_j(l) \ln \beta_s(l) + \lambda \left(\sum_{l=1}^V \beta_s(l) - 1 \right),$$

pe care o vom optimiza – prin derivare – în raport cu $\beta_s(l)$:

$$\begin{aligned}
 \frac{\partial}{\partial \beta_s(l)} \mathcal{L}(\beta_s(l)) &= 0 \Leftrightarrow \sum_{j=1}^N q_j(s) \frac{w_j(l)}{\beta_s(l)} + \lambda = 0 \\
 \Leftrightarrow \frac{1}{\beta_s(l)} \sum_{j=1}^N q_j(s) w_j(l) + \lambda &= 0 \Leftrightarrow \frac{1}{\beta_s(l)} = \frac{-\lambda}{\sum_{j=1}^N q_j(s) w_j(l)} \\
 \Leftrightarrow \beta_s(l) &= \frac{\sum_{j=1}^N q_j(s) w_j(l)}{-\lambda}. \tag{412}
 \end{aligned}$$

⁹⁹⁸Cititorul pedant va putea observa după calculele de mai jos că matricea hessiană a funcției auxiliare $Q(\theta|\theta')$ este diagonală și negativ definită, deci $Q(\theta|\theta')$ este funcție concavă, așadar ea admite o singură valoare maximă, care este globală.

⁹⁹⁹Veți vedea că soluția obținută ulterior va satisface și prima restricție.

Întrucât $\sum_{l=1}^V \beta_s(l) = 1$, din relația (412) va rezulta

$$\sum_{l=1}^V \frac{\sum_{j=1}^N q_j(s) w_j(l)}{-\lambda} = 1 \Leftrightarrow -\lambda = \sum_{l=1}^V \sum_{j=1}^N q_j(s) w_j(l)$$

Substituind această valoare a lui $-\lambda$ în relația (412), obținem ceea ce va constitui noua valoare a parametrului $\beta_s(l)$, pentru iterată $t + 1$:

$$\begin{aligned} \beta_s(l) &= \frac{\sum_{j=1}^N q_j(s) w_j(l)}{\sum_{l=1}^V \sum_{j=1}^N q_j(s) w_j(l)} = \frac{\sum_{j=1}^N q_j(s) w_j(l)}{\sum_{j=1}^N \sum_{l=1}^V q_j(s) w_j(l)} \\ &= \frac{\sum_{j=1}^N q_j(s) w_j(l)}{\underbrace{\sum_{j=1}^N q_j(s) \sum_{l=1}^V w_j(l)}_1} = \frac{\sum_{j=1}^N q_j(s) w_j(l)}{\sum_{j=1}^N q_j(s)}. \end{aligned}$$

În mod intuitiv, expresia pe care tocmai am obținut-o (și care este, evident, nenegativă) poate fi interpretată ca fiind „ponderea“ care îi revine cuvântului de pe poziția l din vocabular — fiindcă $w(l) = 1$ — în ansamblul [clusterului corespunzător] domeniului semantic s .

În mod similar, pentru a optimiza funcția auxiliară $Q(\theta|\theta')$ în raport cu probabilitatea de selecție π_s , vom începe prin a elimina din expresia (411) termenii care sunt constanți în raport cu π_s . Expresia care rezultă este:

$$\sum_{j=1}^N q_j(s) \ln \pi_s.$$

După aceea, folosind din nou metoda multiplicatorilor lui Lagrange pentru a impune satisfacerea restricției $\sum_{s=1}^K \pi_s = 1$, vom scrie funcția „lagrangeană“

$$\mathcal{L}(\pi_s) = \sum_{j=1}^N q_j(s) \ln \pi_s + \lambda \left(\sum_{s=1}^K \pi_s - 1 \right).$$

Valoarea optimă pentru π_s va fi obținută cu ajutorul derivatei parțiale:

$$\begin{aligned} \frac{\partial}{\partial \pi_s} \mathcal{L}(\pi_s) = 0 &\Leftrightarrow \sum_{j=1}^N \frac{q_j(s)}{\pi_s} + \lambda = 0 \Leftrightarrow \frac{1}{\pi_s} \sum_{j=1}^N q_j(s) = -\lambda \\ &\Leftrightarrow \pi_s = \frac{\sum_{j=1}^N q_j(s)}{-\lambda}. \end{aligned} \tag{413}$$

Întrucât $\sum_{s=1}^K \pi_s = 1$, vom avea:

$$\begin{aligned} \sum_{s=1}^K \frac{\sum_{j=1}^N q_j(s)}{-\lambda} = 1 &\Leftrightarrow \frac{1}{-\lambda} \sum_{s=1}^K \sum_{j=1}^N q_j(s) = 1 \\ &\Leftrightarrow -\lambda = \sum_{s=1}^K \sum_{j=1}^N q_j(s). \end{aligned}$$

Substituind această valoare a lui $-\lambda$ în expresia (413), obținem *noua valoare a parametrului π_s* :

$$\pi_s = \frac{\sum_{j=1}^N q_j(s)}{\sum_{s=1}^K \sum_{j=1}^N q_j(s)} = \frac{\sum_{j=1}^N q_j(s)}{\underbrace{\sum_{j=1}^N \sum_{\substack{s=1 \\ 1}}^K q_j(s)}_1} = \frac{\sum_{j=1}^N q_j(s)}{N}.$$

În mod intuitiv, această ultimă expresie (care este, evident, nenegativă) poate fi interpretată ca fiind „ponderea“ care revine clusterului s din ansamblul celor N cuvinte din vocabular.

Sumarizând, *regulile de actualizare* de la iterată t a algoritmului EM sunt

Pasul E:

$$q_j(s) = \frac{\pi_s \prod_{l=1}^V \beta_s(l)^{w_j(l)}}{\sum_{s'=1}^K \pi_{s'} \prod_{l=1}^V \beta_{s'}(l)^{w_j(l)}} \text{ pentru } j = 1, \dots, N \text{ și } s = 1, \dots, K,$$

unde π_s , $\pi_{s'}$, β_s și $\beta_{s'}$ provin de la pasul M al iterăției $t - 1$ atunci când $t > 1$ (și respectiv de la inițializare, în cazul iterăției $t = 1$);

Pasul M:

$$\beta_s(l) = \frac{\sum_{j=1}^N q_j(s) w_j(l)}{\sum_{j=1}^N q_j(s)} \text{ pentru } s = 1, \dots, K \text{ și } l = 1, \dots, V$$

$$\pi_s = \frac{\sum_{j=1}^N q_j(s)}{N} \text{ pentru } s = 1, \dots, K.$$

8.1.3 Distribuții binomiale / multinomiale

13. (Algoritmul EM: „învățarea“ unei distribuții probabiliste multinomiale (și implicit a unei distribuții categoriale), determinate de un singur parametru (de estimat), în condițiile existenței unei variabile neobservabile)

prelucrare de Liviu Ciortuz, după CMU, 2008 fall, Eric Xing, final exam, pr. 6

Considerăm un curs de învățare automată la care probabilitatea ca un student să fie notat cu calificativul A este $P(A) = 1/2$, cu B este $P(B) = \theta$, cu C este $P(C) = 2\theta$, iar cu D este $P(D) = 1/2 - 3\theta$, unde $\theta \in (0, 1/6)$ este un parametru. Ni se mai spune că un număr de c studenți au luat calificativul C , iar d studenți au luat calificativul D . Nu știm cu exactitate căți studenți au luat calificativul A sau căți studenți au luat calificativul B , dar știm că în total h studenți au luat fie calificativul A , fie calificativul B .

Ne propunem să utilizăm algoritmul Expectation-Maximization pentru a obține o estimare de verosimilitate maximă a parametrului θ .

Observații:

1. Remarcăm faptul că în această problemă nu „rezolvăm“ o mixtură de distribuții, ci

învățăm o distribuție (categorială), a cărei definiție depinde de un parametru (θ).

2. Am specificat distribuția de învățat în tabelul alăturat (mai precis, în primele două coloane din acest tabel). Evident, dacă am cunoaște valoarea lui a , am putea calcula imediat estimarea de verosimilitate maximă (MLE) a lui θ .

	prob.	count-uri
A	1/2	$a = ?$
B	θ	$b = h - a$
C	2θ	c
D	$1/2 - 3\theta$	d

De exemplu, folosind ultimele trei linii din tabel, am putea scrie:

$$\frac{3\theta}{2} = \frac{b+c}{b+c+d} \Leftrightarrow 6\theta = \frac{b+c}{b+c+d} \Leftrightarrow \theta = \frac{1}{6} \cdot \frac{b+c}{b+c+d} \Leftrightarrow \theta = \frac{1}{6} \cdot \frac{h-a+c}{h-a+c+d}$$

3. Conform formalizării generale a algoritmului EM, datele (observabile și neobservabile) ar trebui să fie (în contextul acestei probleme) x_1, \dots, x_{h+c+d} , fiecare x_i aparținând mulțimii $\{A, B, C, D\}$. Datele observabile ar fi acele instanțe x_i care aparțin mulțimii $\{C, D\}$, iar datele neobservabile restul ($x_i \in \{A, B\}$). Cum ordinea instanțelor x_i nu contează (întrucât operăm cu distribuția categorială, nu cea multinomială), se constată că este suficient să considerăm ca date observabile h , c și d . Variabila neobservabilă va fi considerată a . (Corespunzător, vom avea $b = h - a$.) Se observă că toate cele patru linii din tabel vor fi necesare la scrierea funcției „auxiliare“ pentru algoritmul EM, care reprezintă media log-verosimilității datelor complete.

4. În mod normal, algoritmul EM face la pasul E calculul unei medii (care, în cazul mixturilor, este o probabilitate de forma $P(z_i = j|x_i, \theta^{(t)})$). În problema noastră, cunoscând $\theta^{(t)}$, valoarea parametrului θ la iterarea t , vom putea calcula valoarea medie / „așteptată“ a lui a , pe care o vom nota cu \hat{a} .

- a. **Pasul de calculare a mediilor (E):** Care dintre formulele următoare reprezintă valorile „așteptate“ (engl., expected values) pentru variabilele a și b (văzute ca variabile aleatoare), exprimate în funcție de parametrul θ ?

$$(i) \quad \hat{a} = \frac{\frac{1}{2}}{\frac{1}{2} + h} \theta \quad \hat{b} = \frac{\theta}{\frac{1}{2} + h} \theta \quad (ii) \quad \hat{a} = \frac{\frac{1}{2}}{\frac{1}{2} + \theta} h \quad \hat{b} = \frac{\theta}{\frac{1}{2} + \theta} h$$

$$(iii) \quad \hat{a} = \frac{\theta}{\frac{1}{2} + \theta} h \quad \hat{b} = \frac{\frac{1}{2}}{\frac{1}{2} + \theta} h \quad (iv) \quad \hat{a} = \frac{\frac{1}{2}}{1 + \theta^2} h \quad \hat{b} = \frac{\theta}{1 + \theta^2} h$$

- b. **Pasul de maximizare (M):** Având mediile pentru variabilele a și b (notate cu \hat{a} și respectiv \hat{b} ca mai sus), care dintre formulele următoare reprezintă estimarea de verosimilitate maximă a parametrului θ ?

$$(i) \quad \hat{\theta} = \frac{h - \hat{a} + c}{6(h - \hat{a} + c + d)} \quad (ii) \quad \hat{\theta} = \frac{h - \hat{a} + d}{6(h - 2\hat{a} - d)}$$

$$(iii) \quad \hat{\theta} = \frac{h - \hat{a}}{6(h - 2\hat{a} + c)} \quad (iv) \quad \hat{\theta} = \frac{2(h - \hat{a})}{3(h - \hat{a} + c + d)}$$

Răspuns:

- a. Valoarea medie pentru a se calculează astfel:

Notând $P(A \cup B)$ probabilitatea ca un student oarecare să ia fie calificativul A fie calificativul B, rezultă:¹⁰⁰⁰

$$\hat{a} = P(A | A \cup B) \cdot h \stackrel{\text{def.}}{=} \frac{P(A)}{P(A) + P(B)} \cdot h = \frac{\frac{1}{2}}{\frac{1}{2} + \theta} \cdot h$$

La cea de-a doua egalitate am ținut cont și de faptul că A și B, văzute ca mulțimi, sunt disjuncte.

Procedând în mod similar, vom obține $\hat{b} = E[b | h, \theta] = \frac{\theta}{\frac{1}{2} + \theta} \cdot h$. Deci răspunsul corect este (ii).

b. La pasul M, algoritmul EM maximizează media log-verosimilității datelor complete în funcție de θ . Ținând cont de independența datelor, funcția de verosimilitate a datelor complete se exprimă astfel:¹⁰⁰¹

$$P(a, b, c, d | \theta) = \left(\frac{1}{2}\right)^a (\theta)^b (2\theta)^c \left(\frac{1}{2} - 3\theta\right)^d.$$

Așadar,

$$\ln P(a, b, c, d | \theta) = a \ln \frac{1}{2} + b \ln \theta + c \ln(2\theta) + d \ln \frac{1 - 6\theta}{2}$$

Înlocuind în această expresie variabilele necunoscute / neobservabile a și b cu mediile lor, obținem

$$E[\ln P(a, b, c, d | \theta)] = \hat{a} \ln \frac{1}{2} + \hat{b} \ln \theta + c \ln(2\theta) + d \ln \frac{1 - 6\theta}{2}$$

Maximizarea acestei medii (în raport cu parametrul θ) se realizează cu ajutorul derivatei de ordinul întâi:¹⁰⁰²

$$\begin{aligned} \frac{\partial E[\ln p(a, b, c, d | \theta)]}{\partial \theta} &= 0 \Leftrightarrow \frac{\hat{b}}{\theta} + \frac{2c}{2\theta} - \frac{3d}{1 - 6\theta} = 0 \Leftrightarrow \frac{\hat{b}}{\theta} + \frac{c}{\theta} - \frac{6d}{1 - 6\theta} = 0 \\ \Leftrightarrow \frac{\hat{b} + c}{\theta} &= \frac{6d}{1 - 6\theta} \Leftrightarrow 6d\theta = (\hat{b} + c)(1 - 6\theta) \Leftrightarrow 6\theta(d + \hat{b} + c) = \hat{b} + c \\ \Rightarrow \hat{\theta} &= \frac{\hat{b} + c}{6(\hat{b} + c + d)} = \frac{h - \hat{a} + c}{6(h - \hat{a} + c + d)} \end{aligned}$$

Se poate verifica imediat că $\hat{\theta} \in (0, 1/6)$. Prin urmare, răspunsul corect este (i).

Observații:

¹⁰⁰⁰Valoarea „așteptată“ pentru a (numărul de studenți care au obținut calificativul A) este dată de media distribuției binomiale de parametri h și $\frac{P(A)}{P(A) + P(B)}$. Se știe că această medie este exact produsul celor doi parametri. Vedeți problema 25.b de la capitolul de *Fundamente*.

¹⁰⁰¹În egalitatea următoare omitem factorul $\frac{(a + b + c + d)!}{a! b! c! d!}$ fiindcă nu depinde de θ .

¹⁰⁰²Este imediat că derivata a doua este negativă pe tot domeniul ei de definiție, deci funcția de log-verosimilitate este strict concavă și, în consecință, admite un singur punct de maxim.

5. Regula de actualizare care tocmai a fost obținută la punctul b este un corespondent natural al formulei de estimare a parametrului θ corespunzătoare cazului când nu există date ascunse. (Vedeți *Observația 2* de mai sus.)
 6. Folosind notația ușuală pentru algoritmul EM, putem spune că mediile $\hat{a} \stackrel{\text{not.}}{=} a^{(t+1)}$ și $\hat{b} \stackrel{\text{not.}}{=} b^{(t+1)}$ au fost calculate în funcție de „ipoteza“ curentă $\theta^{(t)}$, iar apoi expresia $E[\ln P(a, b, c, d | \theta)]$ a fost calculată în funcție de θ , și \hat{b} (deci în funcție de θ și $\theta^{(t)}$).

14. (Algoritmul EM pentru „învățarea“ unei distribuții multinomiale (și implicit a unei distribuții categoriale); o aplicație în domeniul bioinformaticii)

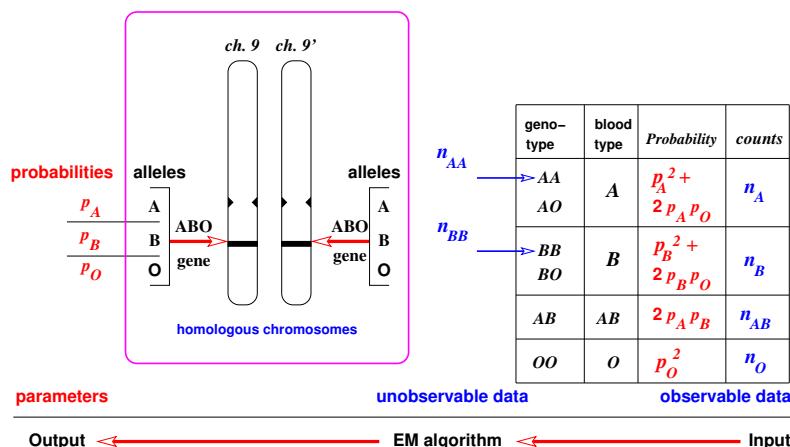
*Liviu Ciortuz, 2017, pornind de la
 ■ □ ○ Ex. 20.10 din “Probability for Statistics and Machine Learning”
 Anirban DasGupta, Springer, 2011*

Grupele sangvine ale oamenilor sunt determinate de variantele („alelele“) unei gene situate pe cromozomul 9 (mai exact, la poziția 9q34.2), numită gena *ABO*. Se știe că fiecare dintre noi dispunem de câte o pereche de astfel de cromozomi, deci de câte două copii ale genei *ABO*, și anume una moștenită de la tată și cealaltă moștenită de la mamă.

Se notează cu A , B și O cele trei tipuri de alele ale genei *ABO*. Alelele A și B sunt *dominante* în raport cu alela O . (Altfel spus, alela O este *recesivă* în raport cu fiecare dintre alelele A și B .) Alelele A și B sunt *codominante*. Prin urmare, grupele sangvine pot fi specificate conform tabelului alăturat.

grupe sangvine	alele moștenite
A	AA AO
B	BB BO
AB	AB
O	OO

Presupunem că, într-o populație oarecare, fiecare dintre alelele A , B și O are la bărbați aceeași frecvență ca și la femei. În cele ce urmează, aceste frecvențe (notate respectiv cu p_A , p_B și p_O) vor fi considerate a priori necunoscute, dar urmează să le determinăm.



- a. Considerăm că într-un eșantion de populație format din n persoane sunt n_A persoane cu grupa sangvină A , n_B persoane cu grupa sangvină B , n_{AB}

persoane cu grupa sangvină AB și n_O persoane cu grupa sangvină O . (Evident, $n = n_a + n_B + n_{AB} + n_O$.) Pornind de la aceste date „observabile“, să se deriveze algoritmul EM pentru determinarea probabilităților p_A , p_B și p_O . (Evident, este suficient să se determine două dintre ele, fiindcă suma celor trei probabilități este 1.)

b. Implementați algoritmul EM pe care l-ați conceput la punctul a și rulați-l pentru inputul $n_A = 186$, $n_B = 38$, $n_{AB} = 13$, $n_O = 284$ ($n = 521$). Ca valori inițiale pentru parametri, veți lucra mai întâi cu $p_A = p_B = p_O = \frac{1}{3}$, iar apoi cu $p_A = p_O = 0.01$ și $p_B = 0.98$. Pentru oprire, veți cere ca $p_A^{(t)}$, $p_B^{(t)}$ și $p_O^{(t)}$ să nu difere (fiecare în parte) cu mai mult de 10^{-4} față de valorile calculate la iterația precedentă. Comparați rezultatele obținute pentru fiecare din cele două inițializări.

Indicații:

1. Presupunând că procesul de asociere a alelor moștenite de către un individ de la părinții lui respectă proprietățile specifice evenimentelor aleatoare independente, rezultă că probabilitățile de realizare a combinațiilor (în termeni genetici: „fenotipurile“) AA , AO , BB , BO , AB și OO într-o populație oarecare sunt p_A^2 , $2p_A p_O$, p_B^2 , $2p_B p_O$, $2p_A p_B$, și respectiv p_O^2 .
2. Considerând $n_A = n_{AA} + n_{AO}$ și $n_B = n_{BB} + n_{BO}$, unde semnificațiile numerelor n_{AA} , n_{AO} , n_{BB} și n_{BO} sunt similare cu semnificațiile numerelor n_A , n_B , n_{AB} , și n_O care au fost precizate mai sus, este natural ca în formularea algoritmului EM datele n_{AA} și n_{BB} să fie considerate „neobservabile“. Ca parametri ai modelului, se vor considera probabilitățile p_A , p_B și p_O .
3. La pasul E al algoritmului EM veți calcula \hat{n}_{AA} și \hat{n}_{BB} , care reprezintă numărul „așteptat“ de apariții ale combinației de alele AA și respectiv numărul „așteptat“ de apariții ale combinației de alele BB în populația dată. Veți scrie apoi funcția de log-verosimilitate a datelor complete („observabile“ și „neobservabile“), exprimată cu ajutorul distribuției $Multinomial(n; p_A^2, 2p_A p_O, p_B^2, 2p_B p_O, 2p_A p_B, p_O^2)$.
4. La pasul M al algoritmului EM, pornind de la media funcției de log-verosimilitate care a fost calculată la pasul E, veți stabili regulile de actualizare pentru probabilitățile p_A , p_B și p_O . Atenție: suma acestor probabilități fiind 1, problema de optimizare pe care va trebui să o rezolvați la pasul M este una cu restricții. În acest sens, *metoda multiplicatorilor lui Lagrange* vă poate fi de folos.

Răspuns:

Vom folosi următoarele notății:

- setul de parametri: $p = \{p_A, p_B, p_O\}$;
de fapt, va fi suficient să estimăm doar p_A și p_B , fiindcă $p_A + p_B + p_O = 1$
- variabilele observabile: $n_{obs} = \{n_A, n_B, n_{AB}, n_O\}$
- variabilele neobservabile: $n_{unobs} = \{n_{AA}, n_{AO}, n_{BB}, n_{BO}\}$;
de fapt, vom reduce setul de variabile n_{unobs} la $\{n_{AA}, n_{BB}\}$, pentru că $n_{AO} = n_A - n_{AA}$ și $n_{BO} = n_B - n_{BB}$;
- datele complete: $n_{compl} = n_{obs} \cup n_{unobs}$
- $n = n_A + n_B + n_{AB} + n_O$, $n_A = n_{AA} + n_{AO}$, $n_B = n_{BB} + n_{BO}$.

Pornind de la definiția funcției masă de probabilitate (p.m.f.) pentru *distribuția multinomială*, rezultă că expresia funcției de verosimilitate a datelor complete din problema noastră este următoarea:¹⁰⁰³

$$\begin{aligned} L(p) &\stackrel{\text{not.}}{=} P(n_{\text{compl}}|p) = \\ &= \frac{n!}{n_{AA}! n_{AO}! n_{BB}! n_{BO}! n_{AB}! n_O!} \cdot \\ &\quad (p_A^2)^{n_{AA}} \cdot (2 p_A p_O)^{n_{AO}} \cdot (p_B^2)^{n_{BB}} \cdot (2 p_B p_O)^{n_{BO}} \cdot (2 p_A p_B)^{n_{AB}} \cdot (p_O^2)^{n_O} \end{aligned}$$

Logaritmând această expresie, obținem funcția de log-verosimilitate a datelor complete:

$$\begin{aligned} \ell(p) &\stackrel{\text{def.}}{=} \ln L(p) \\ &= c + n_{AA} \ln(p_A^2) + n_{AO} \ln(2 p_A p_O) + n_{BB} \ln(p_B^2) + n_{BO} \ln(2 p_B p_O) + \\ &\quad n_{AB} \ln(2 p_A p_B) + n_O \ln(p_O^2) \\ &= c' + 2 n_{AA} \ln p_A + n_{AO} (\ln p_A + \ln p_O) + \\ &\quad 2 n_{BB} \ln p_B + n_{BO} (\ln p_B + \ln p_O) + n_{AB} (\ln p_A + \ln p_B) + 2 n_O \ln p_O \\ &= c' + 2 n_{AA} \ln p_A + (n_A - n_{AA}) (\ln p_A + \ln p_O) + \\ &\quad 2 n_{BB} \ln p_B + (n_B - n_{BB}) (\ln p_B + \ln p_O) + n_{AB} (\ln p_A + \ln p_B) + 2 n_O \ln p_O, \end{aligned}$$

unde c și c' sunt constante care nu depind de parametrul p .

Funcția auxiliară se obține din expresia acestei funcții de log-verosimilitate, prin aplicarea operatorului E , care se referă la media variabilelor neobservabile:

$$\begin{aligned} Q(p|p^{(t)}) &\stackrel{\text{def.}}{=} E[\ell(p)|n_{\text{obs}}; p^{(t)}] \\ &= c' + 2 \hat{n}_{AA} \ln p_A + (n_A - \hat{n}_{AA}) (\ln p_A + \ln p_O) + \\ &\quad 2 \hat{n}_{BB} \ln p_B + (n_B - \hat{n}_{BB}) (\ln p_B + \ln p_O) + n_{AB} (\ln p_A + \ln p_B) + 2 n_O \ln p_O, \end{aligned}$$

unde

$$\begin{aligned} \hat{n}_{AA} &\stackrel{\text{not.}}{=} E[n_{AA}|n_{\text{obs}}; p^{(t)}] = E[n_{AA}|n_A, n_B, n_{AB}, n_O; p_A^{(t)}, p_B^{(t)}, p_O^{(t)}] \\ \hat{n}_{BB} &\stackrel{\text{not.}}{=} E[n_{BB}|n_{\text{obs}}; p^{(t)}] = E[n_{BB}|n_A, n_B, n_{AB}, n_O; p_A^{(t)}, p_B^{(t)}, p_O^{(t)}]. \end{aligned}$$

Pasul E:

După cum se observă în expresia funcției auxiliare $Q(p|p^{(t)})$, acum trebuie să calculăm mediile \hat{n}_{AA} și \hat{n}_{BB} , adică numărul „așteptat“ (engl., expected number) de persoane care au *fenotipul* (adică, perechea de alele) *AA*, și respectiv *BB*. Tinând cont că n_{AA} (sau, echivalent, fenotipul *AA*), văzut ca variabilă aleatoare, urmează *distribuția binomială* de parametri n_A și $\frac{(p_A^{(t)})^2}{(p_A^{(t)})^2 + 2 p_A^{(t)} p_O^{(t)}}$, rezultă că media sa este:¹⁰⁰⁴

$$\begin{aligned} \hat{n}_{AA} &\stackrel{\text{not.}}{=} E[n_{AA}|n_A, n_B, n_{AB}, n_O; p_A^{(t)}, p_B^{(t)}, p_O^{(t)}] \\ &= \frac{(p_A^{(t)})^2}{(p_A^{(t)})^2 + 2 p_A^{(t)} p_O^{(t)}} \cdot n_A. \end{aligned} \tag{414}$$

¹⁰⁰³Observați că factorul $\frac{n!}{n_{AA}! n_{AO}! n_{BB}! n_{BO}! n_{AB}! n_O!}$ generalizează C_n^k (combinări de n luate câte k), ceea ce este natural, întrucât distribuția multinomială este o generalizare a distribuției binomiale. (Vedeți problema 25 de la capitolul de *Fundamente*.)

¹⁰⁰⁴Vedeți problema 25.b de la capitolul de *Fundamente*.

În mod similar,

$$\begin{aligned}\hat{n}_{BB} &\stackrel{not.}{=} E[n_{BB}|n_A, n_B, n_{AB}, n_O; p_A^{(t)}, p_B^{(t)}, p_O^{(t)}] \\ &= \frac{(p_B^{(t)})^2}{(p_B^{(t)})^2 + 2 p_B^{(t)} p_O^{(t)}} \cdot n_B.\end{aligned}\quad (415)$$

Pasul M:

Întrucât probabilitățile p_A , p_B și p_O trebuie să satisfacă restricția $p_A + p_B + p_O = 1$, vom aplica metoda multiplicatorilor lui Lagrange și vom introduce variabila / „multiplicatorul“ $\lambda \in \mathbb{R}$ cu scopul de rezolvă problema de optimizare următoare:

$$\begin{aligned}p^{(t+1)} &\stackrel{not.}{=} (p_A^{(t+1)}, p_B^{(t+1)}, p_O^{(t+1)}) \\ &= \underset{p_A, p_B, p_O}{\operatorname{argmax}} [Q(p_A, p_B, p_O | p_A^{(t)}, p_B^{(t)}, p_O^{(t)}) + \lambda(1 - (p_A + p_B + p_O))] \\ &= \underset{p_A, p_B, p_O}{\operatorname{argmax}} [c' + 2 \hat{n}_{AA} \ln p_A + (n_A - \hat{n}_{AA})(\ln p_A + \ln p_O) + 2 \hat{n}_{BB} \ln p_B + \\ &\quad (n_B - \hat{n}_{BB})(\ln p_B + \ln p_O) + n_{AB}(\ln p_A + \ln p_B) + 2 n_O \ln p_O + \\ &\quad \lambda(1 - (p_A + p_B + p_O))].\end{aligned}$$

Calculând derivatele parțiale ale *funcției obiectiv* din această problemă de optimizare în raport cu p_A , p_B și p_O și egalându-le apoi cu 0, vom avea:

$$\begin{aligned}\frac{1}{p_A}(2 \hat{n}_{AA} + n_A - \hat{n}_{AA} + n_{AB}) - \lambda &= 0 \Rightarrow \hat{p}_A = \frac{1}{\lambda}(\hat{n}_{AA} + n_A + n_{AB}) \\ \frac{1}{p_B}(2 \hat{n}_{BB} + n_B - \hat{n}_{BB} + n_{AB}) - \lambda &= 0 \Rightarrow \hat{p}_B = \frac{1}{\lambda}(\hat{n}_{BB} + n_B + n_{AB}) \\ \frac{1}{p_O}(n_A - \hat{n}_{AA} + n_B - \hat{n}_{BB} + 2 n_O) - \lambda &= 0 \Rightarrow \hat{p}_O = \frac{1}{\lambda}(n_A - \hat{n}_{AA} + n_B - \hat{n}_{BB} + 2 n_O).\end{aligned}$$

În continuare, aplicând restricția $\hat{p}_A + \hat{p}_B + \hat{p}_O = 1$, rezultă:

$$\begin{aligned}\frac{1}{\lambda}(\hat{n}_{AA} + n_A + n_{AB} + \hat{n}_{BB} + n_B + n_{AB} + n_A - \hat{n}_{AA} + n_B - \hat{n}_{BB} + 2 n_O) &= 1 \Leftrightarrow \\ \frac{2}{\lambda}(n_A + n_B + n_{AB} + n_O) &= 1.\end{aligned}$$

Întrucât $n = n_A + n_B + n_{AB} + n_O$, obținem imediat $\frac{1}{\lambda}2n = 1$ și, în consecință, $\lambda = 2n$. Înlocuindu-l pe λ cu această valoare ($2n$) în expresiile pe care le-am obținut anterior pentru \hat{p}_A , \hat{p}_B și \hat{p}_O , vom obține următoarele *reguli de actualizare* pentru pasul M al algoritmului EM:

$$\hat{p}_A^{(t+1)} = \frac{1}{2n}(\hat{n}_{AA} + n_A + n_{AB}) \quad (416)$$

$$\hat{p}_B^{(t+1)} = \frac{1}{2n}(\hat{n}_{BB} + n_B + n_{AB}) \quad (417)$$

$$\hat{p}_O^{(t+1)} = \frac{1}{2n}(n_A - \hat{n}_{AA} + n_B - \hat{n}_{BB} + 2 n_O) = \frac{1}{2n}(n_{AO} + n_{BO} + 2 n_O). \quad (418)$$

Se observă ușor că toate aceste valori sunt în intervalul $[0, 1]$.

Sumarizând, relațiile de actualizare pentru algoritmul EM sunt (414) și (415) pentru pasul E, și respectiv (416), (417) și (418) pentru pasul M.

b. Rulând o implementare a algoritmului EM pe datele specificate în enunț și făcând inițializările indicate acolo (și menționate în următoarele tabele), am obținut rezultate:

Valori inițiale:

$$p_A = p_B = p_O = 1/3.$$

Iterații:

t	p_A	p_B	p_O
1	0.2505	0.0611	0.6884
2	0.2185	0.0505	0.7311
3	0.2142	0.0502	0.7357
4	0.2137	0.0501	0.7362
5	0.2136	0.0501	0.7363
6	0.2136	0.0501	0.7363

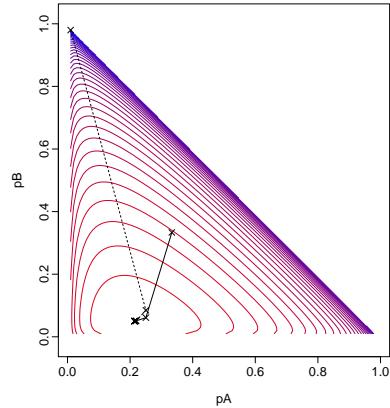
Valori inițiale:

$$p_A = p_O = 0.01, p_B = 0.98.$$

Iterații:

t	p_A	p_B	p_O
1	0.2505	0.0847	0.6648
2	0.2193	0.0511	0.7296
3	0.2143	0.0502	0.7355
4	0.2137	0.0501	0.7362
5	0.2136	0.0501	0.7363
6	0.2136	0.0501	0.7363

Se constată că rezultatele finale care au fost obținute de către algoritmul EM pornind de la cele două inițializări diferite sunt identice. Acest fapt era de așteptat, întrucât funcția de log-verosimilitate a datelor observabile este strict concavă și are un singur punct de maxim, aşa cum se observă și în graficul alăturat, realizat de către Sebastian Ciobanu.



8.1.4 Sume de variabile aleatoare

15.

(Algoritmul EM: estimarea parametrilor pentru o sumă de două distribuții exponentiale)

■ • CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW2, pr. 2.2

Considerăm Z_1 și Z_2 două variabile aleatoare independente, distribuite exponentional, de parametri λ_1 și respectiv λ_2 , cu $\lambda_1 \neq \lambda_2$.¹⁰⁰⁵ Vă reamintim că funcția densitate de probabilitate a unei variabile aleatoare exponentiale Z de parametru $\lambda > 0$ este definită astfel:¹⁰⁰⁶

$$f_\lambda(z) = \begin{cases} \lambda \cdot e^{-\lambda z} & \text{pentru } z \geq 0 \\ 0 & \text{pentru } z < 0. \end{cases}$$

¹⁰⁰⁵ Notație: $Z_1 \sim \exp\left(\frac{1}{\lambda_1}\right)$ și $Z_2 \sim \exp\left(\frac{1}{\lambda_2}\right)$.

¹⁰⁰⁶ Pentru graficul acestei funcții, a se vedea problema 31.a de la capitolul de *Fundamente*.

Se consideră variabila aleatoare $X = Z_1 + Z_2$. Se dau instanțele x_1, x_2, \dots, x_n independente și identic distribuite conform distribuției probabiliste a lui X .

a. Să se calculeze funcția densitate de probabilitate a lui X în funcție de parametrii λ_1 și λ_2 .

Sugestie: Calculați mai întâi funcția de distribuție cumulativă a lui X . Pentru aceasta, puteți folosi formula $F(x) = \int_0^x \int_0^{x-z_1} f_{\lambda_1}(z_1) \cdot f_{\lambda_2}(z_2) dz_2 dz_1$.¹⁰⁰⁷

b. Elaborați în detaliu cei doi pași ai algoritmului EM pentru estimarea parametrilor λ_1 și λ_2 . Precizați la final care sunt ecuațiile de actualizare pentru acești doi parametri, pornind de la setul de date $\{x_1, x_2, \dots, x_n\}$.

Răspuns:

a. Funcția de distribuție cumulativă a lui X se calculează astfel:

$$\begin{aligned}
 F(x) &= P(Z_1 + Z_2 < x) = \int_0^x \int_0^{x-z_1} f_{\lambda_1}(z_1) \cdot f_{\lambda_2}(z_2) dz_2 dz_1 \\
 &= \int_0^x \int_0^{x-z_1} \lambda_1 e^{-\lambda_1 z_1} \cdot \lambda_2 e^{-\lambda_2 z_2} dz_2 dz_1 \\
 &= \int_0^x (-\lambda_1) e^{-\lambda_1 z_1} \left(\int_0^{x-z_1} (-\lambda_2) e^{-\lambda_2 z_2} dz_2 \right) dz_1 \\
 &= \int_0^x (-\lambda_1) e^{-\lambda_1 z_1} \left(e^{-\lambda_2 z_2} \Big|_0^{x-z_1} \right) dz_1 \\
 &= \int_0^x (-\lambda_1) e^{-\lambda_1 z_1} \left(e^{-\lambda_2(x-z_1)} - e^{-\lambda_2 \cdot 0} \right) dz_1 \\
 &= \int_0^x (-\lambda_1) e^{-\lambda_1 z_1} e^{-\lambda_2(x-z_1)} dz_1 - \int_0^x (-\lambda_1) e^{-\lambda_1 z_1} dz_1 \\
 &= \frac{-\lambda_1}{\lambda_2 - \lambda_1} e^{-\lambda_2 x} \int_0^x (\lambda_2 - \lambda_1) e^{(\lambda_2 - \lambda_1) z_1} dz_1 - \int_0^x (-\lambda_1) e^{-\lambda_1 z_1} dz_1 \\
 &= -\frac{\lambda_1}{\lambda_2 - \lambda_1} e^{-\lambda_2 x} \left(e^{(\lambda_2 - \lambda_1) z_1} \Big|_0^x \right) - e^{-\lambda_1 z_1} \Big|_0^x \\
 &= -\frac{\lambda_1}{\lambda_2 - \lambda_1} e^{-\lambda_2 x} \left(e^{(\lambda_2 - \lambda_1)x} - 1 \right) - (e^{-\lambda_1 x} - 1) \\
 &= -\frac{\lambda_1}{\lambda_2 - \lambda_1} e^{-\lambda_1 x} + \frac{\lambda_1}{\lambda_2 - \lambda_1} e^{-\lambda_2 x} - e^{-\lambda_1 x} + 1 \\
 &= 1 - \frac{1}{\lambda_2 - \lambda_1} (\lambda_1 e^{-\lambda_1 x} - \lambda_1 e^{-\lambda_2 x} + \lambda_2 e^{-\lambda_1 x} - \lambda_1 e^{-\lambda_1 x}) \\
 &= 1 - \frac{\lambda_2 e^{-\lambda_1 x} - \lambda_1 e^{-\lambda_2 x}}{\lambda_2 - \lambda_1}
 \end{aligned}$$

¹⁰⁰⁷Deducerea acestei formule este destul de simplă:

$$\begin{aligned}
 F(x) &\stackrel{\text{def.}}{=} P(Z_1 + Z_2 < x) \\
 &= \int_0^x \left(\int_0^{x-z_2} f_{\lambda_1}(z_1) dz_1 \right) \cdot f_{\lambda_2}(z_2) dz_2 \\
 &= \int_0^x \left(\int_0^{x-z_1} f_{\lambda_2}(z_2) dz_2 \right) \cdot f_{\lambda_1}(z_1) dz_1 = \int_0^x \int_0^{x-z_1} f_{\lambda_1}(z_1) \cdot f_{\lambda_2}(z_2) dz_2 dz_1.
 \end{aligned}$$

Cunoscând funcția de distribuție cumulativă a lui X , vom calcula funcția de densitate de probabilitate a lui X :

$$p(x) = \frac{\partial F(x)}{\partial x} = -\frac{1}{\lambda_2 - \lambda_1} (-\lambda_1 \lambda_2 e^{-\lambda_1 x} + \lambda_1 \lambda_2 e^{-\lambda_2 x}) = \frac{\lambda_1 \lambda_2}{\lambda_2 - \lambda_1} (e^{-\lambda_1 x} - e^{-\lambda_2 x})$$

b. Vom prezenta două variante / metode de elaborare a algoritmului EM pentru estimarea parametrilor λ_1 și λ_2 :

- Prima metodă constă în aplicarea clasicea schemei algoritmice EM:

Pasul E (Expectation): Se calculează funcția „auxiliară“ (media log-verosimilității datelor complete) pentru iterată t :

$$Q(\lambda | \lambda^{(t)}) = E_{P(Z|X, \lambda^{(t)})} [\log P(X, Z | \lambda)]$$

Pasul M (Maximization): Se maximizează media log-verosimilității datelor complete, calculată la pasul E, în raport cu λ :

$$\lambda^{(t+1)} = \underset{\lambda}{\operatorname{argmax}} Q(\lambda | \lambda^{(t)})$$

Notăriile de mai sus au următoarele semnificații:

$$\begin{aligned} \lambda &= (\lambda_1, \lambda_2) \\ \lambda^{(t)} &= \text{valoarea parametrului } \lambda \text{ la iterată } t \\ X &= \text{variabila observabilă, cu instanțele } x_1, x_2, \dots, x_n \\ Z &= (Z_1, Z_2) \text{ variabilele ascunse / neobservabile} \\ &\quad \text{având valorile } z_{1j}, z_{2j}, \dots, z_{nj} \text{ cu } j \in \{1, 2\}, \\ &\quad \text{așa încât } x_i = z_{i1} + z_{i2}. \end{aligned}$$

În continuare vom elabora calculele corespunzătoare celor doi pași (E și M).

(Pasul E) Expresia funcției „auxiliare“ este:

$$\begin{aligned} Q(\lambda | \lambda^{(t)}) &= E_{P(Z|X, \lambda^{(t)})} \left[\log \prod_{i=1}^n p(x_i, z_{i1}, z_{i2} | \lambda) \right] \\ &= E_{P(Z|X, \lambda^{(t)})} \left[\log \prod_{i=1}^n f_{\lambda_1}(z_{i1}) \cdot f_{\lambda_2}(z_{i2}) \right] \\ &= E_{P(Z|X, \lambda^{(t)})} \left[\sum_{i=1}^n \log (\lambda_1 e^{-\lambda_1 z_{i1}} \cdot \lambda_2 e^{-\lambda_2 z_{i2}}) \right] \\ &= E_{P(Z|X, \lambda^{(t)})} \left[\sum_{i=1}^n (\log \lambda_1 - \lambda_1 z_{i1} + \log \lambda_2 - \lambda_2 z_{i2}) \right] \\ &= E_{P(Z|X, \lambda^{(t)})} \left[n \log \lambda_1 + n \log \lambda_2 - \sum_{i=1}^n (\lambda_1 z_{i1} + \lambda_2 z_{i2}) \right] \\ &= n \log \lambda_1 + n \log \lambda_2 - \lambda_1 \sum_{i=1}^n E_{p(z_{i1}|x_i, \lambda^{(t)})}[z_{i1}] - \lambda_2 \sum_{i=1}^n E_{p(z_{i2}|x_i, \lambda^{(t)})}[z_{i2}] \end{aligned}$$

Ultima egalitate de mai sus are loc datorită proprietății de liniaritate a mediilor.

Vom trece acum la calcularea mediilor variabilelor neobservabile. Media variabilei z_{i1} în raport cu probabilitatea condiționată $p(z_{i1} | x_i, \lambda^{(t)})$ este:

$$E_{p(z_{i1}|x_i, \lambda^{(t)})}[z_{i1}] \stackrel{\text{def.}}{=} \int_0^{x_i} z_{i1} \cdot p(z_{i1} | x_i, \lambda^{(t)}) dz_{i1}$$

Așadar, pentru a calcula această medie trebuie să obținem mai întâi expresia probabilității condiționate

$$p(z_{i1} | x_i, \lambda^{(t)}) \stackrel{\text{def.}}{=} \frac{p(z_{i1}, x_i | \lambda^{(t)})}{p(x_i | \lambda^{(t)})}$$

Probabilitatea de la numitor a fost calculată la punctul a , iar pentru cea de la numărător vom folosi relația $x_i = z_{i1} + z_{i2}$. Prin urmare,

$$p(z_{i1} | x_i, \lambda^{(t)}) = \frac{p(z_{i1}, z_{i2} | \lambda^{(t)})}{p(x_i | \lambda^{(t)})} = \frac{p(z_{i1} | \lambda_1^{(t)}) \cdot p(z_{i2} | \lambda_2^{(t)})}{p(x_i | \lambda^{(t)})},$$

ultima egalitate având loc datorită faptului că variabilele aleatoare Z_1 și Z_2 sunt independente. În consecință,

$$\begin{aligned} p(z_{i1} | x_i, \lambda^{(t)}) &= \frac{f_{\lambda_1^{(t)}}(z_{i1}) \cdot f_{\lambda_2^{(t)}}(x_i - z_{i1})}{\frac{\lambda_1^{(t)} \lambda_2^{(t)}}{\lambda_2^{(t)} - \lambda_1^{(t)}} \left(e^{-\lambda_1^{(t)} x_i} - e^{-\lambda_2^{(t)} x_i} \right)} \\ &= (\lambda_2^{(t)} - \lambda_1^{(t)}) \cdot \frac{\lambda_1^{(t)} e^{-\lambda_1^{(t)} z_{i1}} \cdot \lambda_2^{(t)} e^{-\lambda_2^{(t)} (x_i - z_{i1})}}{\lambda_1^{(t)} \lambda_2^{(t)} \left(e^{-\lambda_1^{(t)} x_i} - e^{-\lambda_2^{(t)} x_i} \right)} \\ &= (\lambda_2^{(t)} - \lambda_1^{(t)}) \cdot \frac{e^{(\lambda_2^{(t)} - \lambda_1^{(t)}) z_{i1}}}{\left(e^{-\lambda_1^{(t)} x_i} - e^{-\lambda_2^{(t)} x_i} \right) \cdot e^{\lambda_2^{(t)} x_i}} = (\lambda_2^{(t)} - \lambda_1^{(t)}) \cdot \frac{e^{(\lambda_2^{(t)} - \lambda_1^{(t)}) z_{i1}}}{e^{(\lambda_2^{(t)} - \lambda_1^{(t)}) x_i} - 1} \end{aligned}$$

Așadar, media variabilei z_{i1} în raport cu probabilitatea condiționată $p(z_{i1} | x_i, \lambda^{(t)})$ este:

$$\begin{aligned} E_{p(z_{i1}|x_i, \lambda^{(t)})}[z_{i1}] &\stackrel{\text{def.}}{=} \int_0^{x_i} z_{i1} \cdot p(z_{i1} | x_i, \lambda^{(t)}) dz_{i1} \\ &= \int_0^{x_i} z_{i1} \cdot (\lambda_2^{(t)} - \lambda_1^{(t)}) \cdot \frac{e^{(\lambda_2^{(t)} - \lambda_1^{(t)}) z_{i1}}}{e^{(\lambda_2^{(t)} - \lambda_1^{(t)}) x_i} - 1} dz_{i1} \\ &= \frac{1}{e^{(\lambda_2^{(t)} - \lambda_1^{(t)}) x_i} - 1} \int_0^{x_i} z_{i1} \cdot (\lambda_2^{(t)} - \lambda_1^{(t)}) \cdot e^{(\lambda_2^{(t)} - \lambda_1^{(t)}) z_{i1}} dz_{i1} \end{aligned}$$

Vom rezolva ultima integrală de mai sus utilizând formula de integrare prin părți:

$$\int f \cdot g' = f \cdot g - \int f' \cdot g$$

Așadar,

$$\begin{aligned}
& \int_0^{x_i} z_{i1} \cdot (\lambda_2^{(t)} - \lambda_1^{(t)}) \cdot e^{(\lambda_2^{(t)} - \lambda_1^{(t)})z_{i1}} dz_{i1} \\
&= \int_0^{x_i} z_{i1} \cdot \frac{\partial}{\partial z_{i1}} \left(e^{(\lambda_2^{(t)} - \lambda_1^{(t)})z_{i1}} \right) dz_{i1} \\
&= \left(z_{i1} \cdot e^{(\lambda_2^{(t)} - \lambda_1^{(t)})z_{i1}} \right) \Big|_0^{x_i} - \int_0^{x_i} e^{(\lambda_2^{(t)} - \lambda_1^{(t)})z_{i1}} dz_{i1} \\
&= \left(x_i \cdot e^{(\lambda_2^{(t)} - \lambda_1^{(t)})x_i} - 0 \right) - \frac{1}{\lambda_2^{(t)} - \lambda_1^{(t)}} \cdot e^{(\lambda_2^{(t)} - \lambda_1^{(t)})z_{i1}} \Big|_0^{x_i} \\
&= x_i \cdot e^{(\lambda_2^{(t)} - \lambda_1^{(t)})x_i} - \frac{e^{(\lambda_2^{(t)} - \lambda_1^{(t)})x_i} - 1}{\lambda_2^{(t)} - \lambda_1^{(t)}}
\end{aligned}$$

Prin urmare,

$$\begin{aligned}
E_{p(z_{i1}|x_i, \lambda^{(t)})}[z_{i1}] &= \frac{1}{e^{(\lambda_2^{(t)} - \lambda_1^{(t)})x_i} - 1} \cdot \left(x_i \cdot e^{(\lambda_2^{(t)} - \lambda_1^{(t)})x_i} - \frac{e^{(\lambda_2^{(t)} - \lambda_1^{(t)})x_i} - 1}{\lambda_2^{(t)} - \lambda_1^{(t)}} \right) \\
&= \frac{x_i \cdot e^{(\lambda_2^{(t)} - \lambda_1^{(t)})x_i}}{e^{(\lambda_2^{(t)} - \lambda_1^{(t)})x_i} - 1} - \frac{1}{\lambda_2^{(t)} - \lambda_1^{(t)}}
\end{aligned}$$

Pentru a calcula media lui z_{i2} în raport cu probabilitatea condiționată $p(z_{i2} | x_i, \lambda^{(t)})$, vom utiliza relația $x_i = z_{i1} + z_{i2}$ și media lui z_{i1} calculată anterior:

$$\begin{aligned}
E_{p(z_{i2}|x_i, \lambda^{(t)})}[z_{i2}] &= E_{p(z_{i2}|x_i, \lambda^{(t)})}[x_i - z_{i1}] = x_i - E_{p(z_{i2}|x_i, \lambda^{(t)})}[z_{i1}] \\
&= x_i - E_{p(z|x_i, \lambda^{(t)})}[z_{i1}] = x_i - E_{p(z_{i1}|x_i, \lambda^{(t)})}[z_{i1}] \\
&= x_i - \frac{x_i \cdot e^{(\lambda_2^{(t)} - \lambda_1^{(t)})x_i}}{e^{(\lambda_2^{(t)} - \lambda_1^{(t)})x_i} - 1} + \frac{1}{\lambda_2^{(t)} - \lambda_1^{(t)}} \\
&= \frac{1}{\lambda_2^{(t)} - \lambda_1^{(t)}} - \frac{x_i}{e^{(\lambda_2^{(t)} - \lambda_1^{(t)})x_i} - 1}
\end{aligned}$$

(Pasul M) Etapa de maximizare: Folosind expresia obținută pentru funcția auxiliară Q , vom calcula

$$\begin{aligned}
\lambda^{(t+1)} &= \underset{\lambda_1 > 0, \lambda_2 > 0}{\operatorname{argmax}} \left(n \log \lambda_1 + n \log \lambda_2 - \lambda_1 \sum_{i=1}^n E_{p(z_{i1}|x_i, \lambda^{(t)})}[z_{i1}] - \right. \\
&\quad \left. - \lambda_2 \sum_{i=1}^n E_{p(z_{i2}|x_i, \lambda^{(t)})}[z_{i2}] \right)
\end{aligned}$$

Valorile parametrilor λ_1 și λ_2 corespunzătoare maximului acestei funcții se obțin folosind metoda derivatelor parțiale:

$$\frac{\partial}{\partial \lambda_1} Q(\lambda | \lambda^{(t)}) = 0 \Leftrightarrow \frac{n}{\lambda_1} = \sum_{i=1}^n E_{p(z_{i1}|x_i, \lambda^{(t)})}[z_{i1}] \Rightarrow \lambda_1^{(t+1)} = \frac{n}{\sum_{i=1}^n E_{p(z_{i1}|x_i, \lambda^{(t)})}[z_{i1}]} > 0$$

$$\frac{\partial}{\partial \lambda_2} Q(\lambda | \lambda^{(t)}) = 0 \Leftrightarrow \frac{n}{\lambda_2} = \sum_{i=1}^n E_{p(z_{i2}|x_i, \lambda^{(t)})}[z_{i2}] \Rightarrow \lambda_2^{(t+1)} = \frac{n}{\sum_{i=1}^n E_{p(z_{i2}|x_i, \lambda^{(t)})}[z_{i2}]} > 0$$

Se verifică imediat că aceste soluții încrătează $Q(\lambda | \lambda^{(t)})$.

- O a doua metodă de estimare a parametrilor λ_1 și λ_2 constă în aplicarea schemei algoritmice EM pentru funcția / variabila $X = Z_1 + Z_2$, urmând ideile de bază ale acestei scheme, dar făcând la pasul M [în locul calculelor] analogia uzuială cu soluția metodei de estimare directă (MLE) pentru parametrul unei distribuții exponențiale. Ca și mai înainte, variabilele Z_1 și Z_2 pot fi văzute ca variabile ascunse.

Pasul E (Expectation):

Se calculează mediile variabilelor ascunse z_{i1} și z_{i2} :

$$E_{p(z_{i1}|x_i, \lambda^{(t)})}[z_{i1}] \text{ și } E_{p(z_{i2}|x_i, \lambda^{(t)})}[z_{i2}]$$

Calculele efective sunt cele realizate mai sus.

Pasul M (Maximization):

Se calculează / actualizează parametrul $\lambda \stackrel{\text{not.}}{=} (\lambda_1, \lambda_2)$ în funcție de valorile medii obținute la pasul precedent:¹⁰⁰⁸

Deoarece $Z_1 \sim \exp\left(\frac{1}{\lambda_1}\right)$, iar estimarea în sensul verosimilității maxime (MLE) a parametrului distribuției exponențiale este inversul mediei setului de exemple (vedeți problema 49 de la capitolul de *Fundamente*), aplicând operatorul E (media) și ținând cont de proprietatea de liniaritate a mediilor, rezultă:

$$\frac{1}{\lambda_1} = \frac{1}{n} \sum_{i=1}^n E_{p(z_{i1}|x_i, \lambda^{(t)})}[z_{i1}] \Rightarrow \lambda_1^{(t+1)} = \frac{n}{\sum_{i=1}^n E_{p(z_{i1}|x_i, \lambda^{(t)})}[z_{i1}]}$$

Similar, pentru $Z_2 \sim \exp\left(\frac{1}{\lambda_2}\right)$ obținem:

$$\frac{1}{\lambda_2} = \frac{1}{n} \sum_{i=1}^n E_{p(z_{i2}|x_i, \lambda^{(t)})}[z_{i2}] \Rightarrow \lambda_2^{(t+1)} = \frac{n}{\sum_{i=1}^n E_{p(z_{i2}|x_i, \lambda^{(t)})}[z_{i2}]}$$

16.

(Algoritmul EM pentru învățarea parametrilor a două distribuții gaussiene, pornind de la instanțe generate de suma a două variabile care urmează aceste distribuții)

■ □ • ○ *Stanford, 2016 fall, A. Ng, J. Duchi, HW4, pr. 2*

La o conferință de învățare automată au fost trimise spre recenzare și, eventual, publicare P lucrări (engl., papers). Comitetul de recenzare a lucrărilor este format din R recenzori. Fiecare dintre acești R recenzori va citi toate cele P lucrări și va da fiecărei lucrări un scor, indicând astfel cât de bună crede el că este lucrarea respectivă. Vom nota cu x_{pr} scorul pe care recenzorul r îl atribuie lucrării p . Dacă scorul este mare, înseamnă că recenzorului i-a plăcut lucrarea respectivă; acest scor reprezintă o recomandare din partea recenzorului ca lucrarea respectivă să fie acceptată pentru prezentare la conferință. Dacă scorul este mic, înseamnă că recenzorului nu i-a plăcut lucrarea respectivă.

¹⁰⁰⁸Este imediat că estimarea parametrului λ_1 poate fi făcută în mod independent de estimarea parametrului λ_2 (până la distribuția probabilistă în raport cu care s-au calculat mediile de la pasul E).

Vom presupune că fiecare lucrare p are o anumită valoare „intrinsecă“, pe care o vom nota cu μ_p ; atunci când valoarea aceasta este mare înseamnă că lucrarea respectivă este bună. Fiecare recenzor încearcă să estimeze, în urma citirii / analizării lucrării p , cât este μ_p . În mod concret, scorul x_{pr} , care este raportat de către recenzorul r , reprezintă încercarea acestuia de a ghici cât este μ_p .

Există tot soiul de factori aleatori care influențează procesul de recenzare a lucrărilor. Din acest motiv, în această problemă vom folosi / propune un model care incorporează mai multe surse de „zgomot“ / perturbații (engl., noise).

Unii recenzori sunt înclinați să credă că toate lucrările sunt bune; ei tind să acorde tuturor lucrărilor scoruri mari. Alți recenzori sunt, din contră, foarte severi și tind să acorde scoruri mici tuturor lucrărilor. În mod similar, este foarte posibil ca scorurile pe care doi recenzori diferiți le acordă lucrărilor pe care le recenzează să manifeste varianțe / dispersii foarte diferite, ceea ce înseamnă că unii recenzori sunt mai credibili decât alții.

Vom nota cu ν_r bias-ul recenzorului r ; astă înseamnă că scorurile puse de acest recenzor tind în general să fie cu cantitatea ν_r mai mari decât ar trebui să fie.

Așadar, din punct de vedere formal vom presupune că scorurile puse de recenzori sunt generate de către un proces aleatoriu, care este definit astfel:

$$\begin{aligned} y_{pr} &\sim \mathcal{N}(\mu_p, \sigma_p^2), \\ z_{pr} &\sim \mathcal{N}(\nu_r, \tau_r^2), \\ x_{pr}|y_{pr}, z_{pr} &\sim \mathcal{N}(y_{pr} + z_{pr}, \sigma^2). \end{aligned}$$

Variabilele y_{pr} și z_{pr} sunt independente; variabilele comune (x, y, z) care corespund unor perechi diferite de tip lucrare-recenzor sunt de asemenea independente. Pe lângă aceasta, trebuie precizat că noi observăm doar valorile variabilelor x_{pr} ; așadar, toate variabilele y_{pr} și z_{pr} sunt latente / neobservabile.

Ceea ce dorim este să estimăm valorile parametrilor μ_p , σ_p^2 , ν_r și τ_r^2 pentru $p = 1, \dots, P$ și $r = 1, \dots, R$. Din rațiuni de simplitate, vom trata σ^2 (varianța condițională a variabilei x_{pr} în raport cu y_{pr} și z_{pr}) ca și cum ar fi o constantă cunoscută, fixată. Dacă vom obține estimări bune pentru valorile „intrinseci“ (μ_p) ale lucrărilor, atunci aceste estimări vor putea fi folosite pentru fundamentarea deciziilor care trebuie luate în privința acceptării / respingerii lucrărilor pentru [prezentare la] conferință.

Vom estima valorile parametrilor μ_p , σ_p^2 , ν_r și τ_r^2 maximizând verosimilitatea datelor observabile $\{x_{pr} ; p = 1, \dots, P, r = 1, \dots, R\}$. În această problemă, aşa cum am precizat deja, variabilele latente sunt y_{pr} și z_{pr} , iar maximizarea verosimilității nu se poate face în mod direct (engl., in closed form). Prin urmare, vom folosi algoritmul EM.

Sarcina dumneavoastră va fi să derivați regulile de actualizare specifice celor doi pași (E și M) ai algoritmului EM. Aceste reguli vor putea să conțină doar operatori de adunare, scădere, înmulțire, împărțire, log, exp și extragere de rădăcină pătrată ($\sqrt{\cdot}$) din constante reale, precum și adunare, scădere, înmulțire, inversare de matrice de numere reale și calculul de determinanți.

a. La acest punct veți obține relațiile de actualizare pentru pasul E:

i. Distribuția comună $p(y_{pr}, z_{pr}, x_{pr})$ este de tip gaussian multidimensional. Găsiți vectorul de medii asociat acestei distribuții, precum și matricea de

covarianță corespunzătoare, în funcție de parametrii $\mu_p, \sigma_p^2, \nu_r, \tau_r^2$ și σ^2 .

Sugestie: x_{pr} poate fi scris ca $x_{pr} = y_{pr} + z_{pr} + \varepsilon_{pr}$, unde $\varepsilon_{pr} \sim \mathcal{N}(0, \sigma^2)$ este un „zgomot”, reprezentat de o variabilă de tip gaussian, independentă în raport cu y_{pr} și z_{pr} .

ii. Deducreți expresia distribuției condiționale $q_{pr}(y_{pr}, z_{pr}) \stackrel{\text{def.}}{=} p(y_{pr}, z_{pr} | x_{pr})$ de la pasul E, folosind regulile pentru condiționare în raport cu submulțimi de variabile aleatoare gaussiene comune.¹⁰⁰⁹

b. Deducreți regulile de actualizare de la pasul M pentru parametrii μ_p, ν_r, σ_p^2 și τ_r^2 .

Sugestie: S-ar putea să vă fie de folos să exprimați marginea inferioară a log-verosimilității datelor observabile ca o medie (engl., expectation) pentru valorile (y_{pr}, z_{pr}) generate de o distribuție aleatoare având funcția de densitate $q_{pr}(y_{pr}, z_{pr})$.

Comentariu: În articolul *Learning from the Wisdom of Crowds by Minimax Entropy* (NIPS, 2012), autorii Dengyong Zhou, John C. Platt,¹⁰¹⁰ Sumit Basu și Yi Mao au descris implementarea unei metode care este destul de asemănătoare cu cea pe care am prezentat-o în acest exercițiu, cu scopul de a estima valorile „intrinseci“ ale lucrărilor, μ_p . În acel articol, problema este ceva mai complicată, fiindcă nu toți recenzorii evaluează fiecare lucrare, însă ideile de bază sunt în esență aceleași cu cele din acest exercițiu. Întrucât modelul acesta încearcă să estimeze și să corecteze biasurile recenzorilor (ν_r), estimările obținute pentru μ_p sunt mult mai folosoare pentru fundamentarea deciziilor de acceptare / respingere a lucrărilor decât scorurile „brute“ acordate de recezori.

Răspuns:

Vom desemna cu θ întregul set de parametri pe care îi estimăm.

Pășii algoritmului EM pentru problema noastră se scriu — la un nivel înalt — astfel:

Pasul E: Pentru fiecare pereche p, r vom calcula

$$q_{pr}^{(t)}(y_{pr}, z_{pr}) \stackrel{\text{not.}}{=} p(y_{pr}, z_{pr} | x_{pr}; \theta^{(t)});$$

Pasul M: $\theta^{(t+1)} = \operatorname{argmax}_{\theta} \sum_{p=1}^P \sum_{r=1}^R E_{q_{pr}^{(t)}(y_{pr}, z_{pr})} \ln p(x_{pr}, y_{pr}, z_{pr}; \theta).$

Chestiunea care se pune acum este cum se calculează în mod concret aceste reguli de actualizare.

a. Pentru pasul E, facem următoarea *observație importantă*:

Dacă la calcularea probabilităților $p(y_{pr}, z_{pr} | x_{pr}; \theta^{(t)})$ am folosi regula lui Bayes combinată cu formula probabilității totale, atunci la numitor ar apărea integralele unor funcții de densitate gaussiene, care n-ar fi deloc ușor de calculat. În schimb, putem observa că funcția densitate de probabilitate

$$\begin{aligned} p(y_{pr}, z_{pr}, x_{pr}) &= p(y_{pr}, z_{pr}) \cdot p(x_{pr} | y_{pr}, z_{pr}) \\ &= p(y_{pr}) \cdot p(z_{pr}) \cdot p(x_{pr} | y_{pr}, z_{pr}) \end{aligned}$$

¹⁰⁰⁹Vedeți relația (38) din *Observația* finală de la rezolvarea problemei 38 de la capitolul de *Fundamente*.

¹⁰¹⁰John C. Platt este autorul algoritmului SMO (Sequential Minimal Optimization), care este „înima“ implementărilor pentru mașinile cu vectori-suport (SVM). Vedeți problemele 23 și 52 de la capitolul *Mașini cu vectori-suport*.

este un produs de alte trei funcții de densitate gaussiene, deci este și ea o densitate gaussiană multidimensională.¹⁰¹¹

Ca o consecință, distribuția [comună și] condiționată $y_{pr}, z_{pr} | x_{pr}$ va fi, la rândul ei, tot o distribuție gaussiană, pe care o vom putea calcula folosind *regulile pentru condiționarea gaussienelor*.

Pentru a obține expresia funcției de densitate pentru distribuția comună (y_{pr}, z_{pr}, x_{pr}) , vom ține cont de faptul că o distribuție gaussiană multidimensională este parametrizată [complet] de către vectorul medie și matricea de covarianță.

Pentru a calcula *vectorul medie* al distribuției comune (y_{pr}, z_{pr}, x_{pr}) , vom descrie x_{pr} în forma următoare:

$$x_{pr} = y_{pr} + z_{pr} + \varepsilon_{pr},$$

unde variabila $\varepsilon_{pr} \sim \mathcal{N}(0, \sigma^2)$ reprezintă un „zgomot“ (engl., noise) care urmează o distribuție gaussiană și este independentă de variabilele y_{pr} și z_{pr} .¹⁰¹² Așadar,

$$\begin{aligned} E[y_{pr}] &= \mu_p \\ E[z_{pr}] &= \nu_p \\ E[x_{pr}] &= E[y_{pr} + z_{pr} + \varepsilon_{pr}] = E[y_{pr}] + E[z_{pr}] + E[\varepsilon_{pr}] \\ &= \mu_p + \nu_r + 0 = \mu_p + \nu_r. \end{aligned}$$

Pentru a calcula *matricea de covarianță* a distribuției comune (y_{pr}, z_{pr}, x_{pr}) , observăm că $Var(y_{pr}) = \sigma_p^2$, $Var(z_{pr}) = \tau_r^2$, iar $Cov(y_{pr}, z_{pr}) = Cov(z_{pr}, y_{pr}) = 0$ întrucât y_{pr} și z_{pr} sunt independente.¹⁰¹³ De asemenea, fiindcă y_{pr} , z_{pr} și ε_{pr} sunt independente, vom avea:¹⁰¹⁴

$$\begin{aligned} Var(x_{pr}) &= Var(y_{pr} + z_{pr} + \varepsilon_{pr}) = Var(y_{pr}) + Var(z_{pr}) + Var(\varepsilon_{pr}) \\ &= \sigma_p^2 + \tau_r^2 + \sigma^2. \end{aligned}$$

În fine,

$$\begin{aligned} Cov(y_{pr}, x_{pr}) &= Cov(x_{pr}, y_{pr}) = Cov(y_{pr} + z_{pr} + \varepsilon_{pr}, y_{pr}) \\ &= Cov(y_{pr}, y_{pr}) + Cov(z_{pr}, y_{pr}) + Cov(\varepsilon_{pr}, y_{pr}) \\ &= \sigma_p^2 + 0 + 0 = \sigma_p^2, \end{aligned}$$

unde antipenultima egalitate se verifică ușor,¹⁰¹⁵ iar penultima egalitate rezultă din independența variabilelor y_{pr} , z_{pr} și ε_{pr} . În mod similar obținem

¹⁰¹¹Vedeți problema 34 de la capitolul de *Fundamente*, corroborată cu explicația de la nota de subsol 1012.

Se poate demonstra că în general (deci nu doar pentru distribuții gaussiene independente), produsul a două distribuții gaussiene multidimensionale este tot o distribuție gaussiană multidimensională. Vedeți Appendix A.2 în cartea *Gaussian processes in machine learning*, Karl Rasmussen, Christopher Williams, MIT Press, 2006.

¹⁰¹²Pentru a vedea de ce aceasta rezultă din definiția dată în enunțul problemei, observați că probabilitatea ca $\varepsilon_{pr} = x_{pr} - y_{pr} - z_{pr}$ să ia o anumită valoare ε este

$$\begin{aligned} p(\varepsilon_{pr} = \varepsilon | y_{pr}, z_{pr}) &= p(x_{pr} - y_{pr} - z_{pr} = \varepsilon | y_{pr}, z_{pr}) \\ &= p(x_{pr} = \varepsilon + y_{pr} + z_{pr} | y_{pr}, z_{pr}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\varepsilon^2\right), \end{aligned}$$

care nu depinde nici de y_{pr} și nici de z_{pr} . Astfel, ε_{pr} poate fi privit ca un „zgomot“ independent, gaussian, cu media 0 și varianță σ^2 .

¹⁰¹³Vedeți problema 10 de la capitolul de *Fundamente*.

¹⁰¹⁴Vedeți problema 23 de la capitolul de *Fundamente*, corelată cu problema 10 de la același capitol.

¹⁰¹⁵Se poate demonstra proprietatea $Cov(X, Y + Z) = Cov(Y + Z, X) = Cov(Y, X) + Cov(Z, X)$ pentru orice variabile aleatoare X , Y și Z . În locul lui $Y + Z$ se poate apoi considera orice sumă finită de variabile aleatoare.

$Cov(z_{pr}, x_{pr}) = Cov(x_{pr}, x_{pr}) = \tau_r^2$. Așadar, putem scrie:

$$y_{pr}, z_{pr}, x_{pr} \sim \mathcal{N} \left(\begin{bmatrix} \mu_p \\ \nu_r \\ \mu_p + \nu_r \end{bmatrix}, \begin{bmatrix} \sigma_p^2 & 0 & \sigma_p^2 \\ 0 & \tau_r^2 & \tau_r^2 \\ \sigma_p^2 & \tau_r^2 & \sigma_p^2 + \tau_r^2 + \sigma^2 \end{bmatrix} \right)$$

Acum vom folosi un rezultat clasic de la *condiționarea subseturilor de variabile* din distribuțiile gaussiene multidimensionale — și anume, relația (38) de la problema 38 de la capitolul de *Foundamente* — pentru a obține expresia căutată pentru $p(y_{pr}, z_{pr}|x_{pr}; \theta^{(t)})$:

$$q_{pr}^{(t)}(y_{pr}, z_{pr}) \stackrel{\text{not.}}{=} p(y_{pr}, z_{pr}|x_{pr}; \theta) = \mathcal{N} \left(\begin{bmatrix} \mu_{pr,Y} \\ \mu_{pr,Z} \end{bmatrix}, \begin{bmatrix} \Sigma_{pr,YY} & \Sigma_{pr,ZY} \\ \Sigma_{pr,YZ} & \Sigma_{pr,ZZ} \end{bmatrix} \right), \quad (419)$$

unde¹⁰¹⁶

$$\mu_{pr} \stackrel{\text{not.}}{=} \begin{bmatrix} \mu_{pr,Y} \\ \mu_{pr,Z} \end{bmatrix} = \begin{bmatrix} \mu_p + \frac{\sigma_p^2}{\sigma^2 + \sigma_p^2 + \tau_r^2}(x_{pr} - \mu_p - \nu_r) \\ \nu_r + \frac{\tau_r^2}{\sigma^2 + \sigma_p^2 + \tau_r^2}(x_{pr} - \mu_p - \nu_r) \end{bmatrix} \quad (420)$$

$$\Sigma_{pr} \stackrel{\text{not.}}{=} \begin{bmatrix} \Sigma_{pr,YY} & \Sigma_{pr,ZY} \\ \Sigma_{pr,YZ} & \Sigma_{pr,ZZ} \end{bmatrix} = \frac{1}{\sigma^2 + \sigma_p^2 + \tau_r^2} \begin{bmatrix} \sigma_p^2(\tau_r^2 + \sigma^2) & -\sigma_p^2\tau_r^2 \\ -\sigma_p^2\tau_r^2 & \tau_r^2(\sigma_p^2 + \sigma^2) \end{bmatrix}. \quad (421)$$

Am omis detaliile de calcul, dar ele nu sunt dificil de elaborat.

b. Pentru pasul M, este bine să ne readucem aminte că distribuția a posteriori $q^{(t)}$ este calculată în funcție de $\theta^{(t)}$, în vreme ce parametrii pe care vrem să-i calculăm pentru pasul următor sunt $\theta^{(t+1)}$. Aceasta înseamnă că parametrii din distribuția $q^{(t)}$ sunt constanți în raport cu acei parametri în funcție de care se face maximizarea de la acest pas. Desemnând — într-o formă ușor simplificată — cu $E_q[\cdot]$ media calculată în raport cu distribuția de probabilitate $q_{pr}^{(t)}(y_{pr}, z_{pr})$ pentru fiecare p și r , vom face maximizarea media log-verosimilității datelor observabile:

$$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}} Q(\theta|\theta^{(t)}),$$

unde

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= \sum_{p=1}^P \sum_{r=1}^R E_q \ln p(x_{pr}, y_{pr}, z_{pr}|\theta) \\ &\stackrel{i.i.d.}{=} \sum_{p=1}^P \sum_{r=1}^R E_q \ln \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x_{pr}-y_{pr}-z_{pr})^2} \cdot \frac{1}{\sqrt{2\pi}\sigma_p} e^{-\frac{1}{2\sigma_p^2}(y_{pr}-\mu_p)^2} \cdot \frac{1}{\sqrt{2\pi}\tau_r} e^{-\frac{1}{2\tau_r^2}(z_{pr}-\nu_r)^2} \right] \\ &= \sum_{p=1}^P \sum_{r=1}^R E_q \left[\ln \frac{1}{(\sqrt{2\pi})^{3/2}\sigma\sigma_p\tau_r} - \frac{1}{2\sigma^2}(x_{pr}-y_{pr}-z_{pr})^2 \right. \\ &\quad \left. - \frac{1}{2\sigma_p^2}(y_{pr}-\mu_p)^2 - \frac{1}{2\tau_r^2}(z_{pr}-\nu_r)^2 \right] \end{aligned}$$

¹⁰¹⁶Pentru a crește lizibilitatea, în expresiile următoare am renunțat la a mai scrie la exponent numărul iterației (t) pentru parametrii μ , σ^2 , τ^2 și Σ .

$$\begin{aligned}
& \stackrel{\theta=(\sigma_p, \tau_r)}{=} \sum_{p=1}^P \sum_{r=1}^R E_q \left[\ln \frac{1}{\sigma_p \tau_r} - \frac{1}{2\sigma_p^2} (y_{pr} - \mu_p)^2 - \frac{1}{2\tau_r^2} (z_{pr} - \nu_r)^2 \right] \\
& = \sum_{p=1}^P \sum_{r=1}^R E_q \left[\ln \frac{1}{\sigma_p \tau_r} - \frac{1}{2\sigma_p^2} ((y_{pr})^2 - 2y_{pr}\mu_p + \mu_p^2) \right. \\
& \quad \left. - \frac{1}{2\tau_r^2} ((z_{pr})^2 - 2z_{pr}\nu_r + \nu_r^2) \right] \\
& \stackrel{\text{lin. med.}}{=} \sum_{p=1}^P \sum_{r=1}^R \left[\ln \frac{1}{\sigma_p \tau_r} - \frac{1}{2\sigma_p^2} (E_q[(y_{pr})^2] - 2E_q[y_{pr}]\mu_p + \mu_p^2) \right. \\
& \quad \left. - \frac{1}{2\tau_r^2} (E_q[(z_{pr})^2] - 2E_q[z_{pr}]\nu_r + \nu_r^2) \right] \\
& \stackrel{\text{lin. med.}}{=} \sum_{p=1}^P \sum_{r=1}^R \left[\ln \frac{1}{\sigma_p \tau_r} - \frac{1}{2\sigma_p^2} (\Sigma_{pr,YY} + \mu_{pr,Y}^2 - 2\mu_{pr,Y} \cdot \mu_p + \mu_p^2) \right. \\
& \quad \left. - \frac{1}{2\tau_r^2} (\Sigma_{pr,ZZ} + \mu_{pr,Z}^2 - 2\mu_{pr,Z} \cdot \nu_r + \nu_r^2) \right].
\end{aligned}$$

Ultima egalitate este o consecință a faptului că $E_q[y_{pr}] = \mu_{pr,Y}$ și $E_q[(y_{pr})^2] = (E_q[(y_{pr})^2] - E_q[y_{pr}]^2) + E_q[y_{pr}]^2 = \Sigma_{pr,YY} + \mu_{pr,Y}^2$. Relații similare se scriu și pentru $E_q[z_{pr}]$ și $E_q[(z_{pr})^2]$.

Calculând derivatele parțiale ale funcției „auxiliare“ $Q(\theta|\theta^{(t)})$ în raport cu parametrii μ_p , ν_r , σ_p și respectiv τ_r și egalându-le apoi cu 0, vom obține:

$$-\frac{1}{2\sigma_p^2} \sum_{r=1}^R (2\mu_p - 2\mu_{pr,Y}) = 0 \Rightarrow \mu_p^{(t+1)} = \frac{1}{R} \sum_{r=1}^R \mu_{pr,Y} \quad (422)$$

$$-\frac{1}{2\tau_r^2} \sum_{p=1}^P (2\nu_r - 2\mu_{pr,Z}) = 0 \Rightarrow \nu_r^{(t+1)} = \frac{1}{P} \sum_{p=1}^P \mu_{pr,Z} \quad (423)$$

$$\sum_{r=1}^R \left[-\frac{1}{\sigma_p} + \frac{1}{\sigma_p^3} (\Sigma_{pr,YY} + \mu_{pr,Y}^2 - 2\mu_{pr,Y} \mu_p + \mu_p^2) \right] = 0 \Rightarrow$$

$$(\sigma_p^2)^{(t+1)} = \frac{1}{R} \sum_{r=1}^R (\Sigma_{pr,YY} + \mu_{pr,Y}^2 - 2\mu_{pr,Y} \mu_p + \mu_p^2) \quad (424)$$

$$\sum_{p=1}^P \left[-\frac{1}{\tau_r} + \frac{1}{\tau_r^3} (\Sigma_{pr,ZZ} + \mu_{pr,Z}^2 - 2\mu_{pr,Z} \nu_r + \nu_r^2) \right] = 0 \Rightarrow$$

$$(\tau_r^2)^{(t+1)} = \frac{1}{P} \sum_{p=1}^P (\Sigma_{pr,ZZ} + \mu_{pr,Z}^2 - 2\mu_{pr,Z} \nu_r + \nu_r^2) \quad (425)$$

Folosind acum rezultatele pe care le-am obținut, putem reformula pașii E și M ai algoritmului EM astfel:

Pasul E: Pentru fiecare p și r , calculează $q_{pr}^{(t)}(y_{pr}, z_{pr})$, conform relației (419), folosind μ_{pr} și Σ_{pr} din relațiile (420) și respectiv (421);

Pasul M: Pentru fiecare p și r , calculează $\mu_p^{(t+1)}$, $\nu_r^{(t+1)}$, $(\sigma_p^2)^{(t+1)}$ și $(\tau_r^2)^{(t+1)}$ folosind relațiile (422), (423), (424) și respectiv (425).

8.1.5 Alte instanțe ale schemei algoritmice EM

17.

(Algoritmul EM pentru învățarea parametrului distribuției Poisson, când se consideră că o parte din date lipsesc)

■ □ *Ex. 20.8 din “Probability for Statistics and Machine Learning”, Anirban DasGupta, Springer, 2011*

Presupunem că vrem să modelăm statistic numărul de accidente ușoare care s-au produs în n locații într-un anumit interval de timp, să zicem o săptămână. În acest scop, vom folosi o distribuție Poisson de parametru λ .¹⁰¹⁷ La sfârșitul perioadei de timp respective, ni se transmite de la m din cele n locații câte o „înregistrare“ (notată cu x_i), reprezentând numărul de accidente ușoare produse în locația i .

În mod implicit, ar trebui să considerăm că în cele $n-m$ locații de la care n-am primit înregistrări nu s-a produs niciun accident. Însă, în urma „inspectării“ datelor suntem determinați să luăm în considerare *presupunerea* că în unele din aceste $n-m$ locații s-a produs câte un [singur] accident ușor, care a fost „trecut sub tăcere“ la raportare.

Așadar, formalizând, vom considera datele „neobservabile“ $z_1 = 0, \dots, z_{n_0} = 0, z_{n_0+1} = 1, \dots, z_{n_0+n_1} = 1$, cu $n_0 + n_1 = n - m$, alături de datele observabile $x_1, \dots, x_m \in \mathbb{N}$. Toate aceste date sunt produse de variabile aleatoare urmând distribuția Poisson de [același] parametru λ .

Elaborați algoritmul EM (în speță pasul E și pasul M) pentru estimarea parametrului λ . Sugestie: În loc să se lucreze cu z_j , cu $j = 1, \dots, m$, va fi suficient să considerați ca date neobservabile n_0 și n_1 . Mai mult, considerând numerele n și m cunoscute, va fi suficient să lucrați doar cu n_1 ca dată neobservabilă (bineînțeles, pe lângă datele observabile x_i).

Răspuns:

Tinând cont de faptul că datele $z_1, \dots, z_{n_0}, z_{n_0+1}, \dots, z_{n_0+n_1}, x_1, \dots, x_m$ urmează distribuția Poisson de parametru λ , a cărei funcție de densitate este $P(x|\lambda) = \frac{1}{e^\lambda} \cdot \frac{\lambda^x}{x!}$, putem scrie expresia care ne dă verosimilitatea datelor „complete“:

$$\begin{aligned}
L(\lambda) &\stackrel{\text{def.}}{=} P(z_1, \dots, z_{n_0}, z_{n_0+1}, \dots, z_{n_0+n_1}, x_1, \dots, x_m | \lambda) \\
&\stackrel{i.i.d.}{=} \prod_{j=1}^{n_0+n_1} P(z_j | \lambda) \cdot \prod_{i=1}^m P(x_i | \lambda) \\
&= \prod_{j=1}^{n_0} \frac{1}{e^\lambda} \frac{\lambda^0}{0!} \cdot \prod_{j=n_0+1}^{n_0+n_1} \frac{1}{e^\lambda} \frac{\lambda^1}{1!} \cdot \prod_{i=1}^m \frac{1}{e^\lambda} \frac{\lambda^{x_i}}{x_i!} \\
&= \frac{1}{(e^\lambda)^{n_0+n_1+m}} \cdot \lambda^{n_1} \cdot \prod_{i=1}^m \frac{\lambda^{x_i}}{x_i!} = \frac{1}{(e^\lambda)^n} \cdot \lambda^{n_1} \cdot \prod_{i=1}^m \frac{\lambda^{x_i}}{x_i!} \\
&= e^{-n\lambda} \cdot \lambda^{n_1} \cdot \lambda^{\sum_{i=1}^m x_i} \cdot \frac{1}{\prod_{i=1}^m x_i!} = e^{-n\lambda} \cdot \lambda^{n_1 + \sum_{i=1}^m x_i} \cdot \frac{1}{\prod_{i=1}^m x_i!}
\end{aligned}$$

¹⁰¹⁷Am precizat deja la pr. 46 de la capitolul de *Fundamente* faptul că distribuția Poisson este utilă la modelarea fenomenelor rare.

Log-verosimilitatea datelor complete este:

$$\ell(\lambda) \stackrel{\text{def}}{=} \ln L(\lambda) = -n\lambda + \left(n_1 + \sum_{i=1}^m x_i \right) \ln \lambda - \sum_{i=1}^m \ln x_i!$$

Funcția „auxiliară“ va fi scrisă cu ajutorul distribuției a posteriori a datelor „neobservabile“ (n_1) în raport cu datele observabile (x_i , cu $i = 1, \dots, m$) și cu $\lambda^{(t)}$, care desemnează valoarea parametrului λ la iterată t . Așadar,

$$Q(\lambda|\lambda^{(t)}) \stackrel{\text{def}}{=} E_{n_1|x_i, \lambda^{(t)}}[\ell(\lambda)] = -n\lambda + \left(E[n_1|x_i, \lambda^{(t)}] + \sum_{i=1}^m x_i \right) \ln \lambda - \sum_{i=1}^m \ln x_i!.$$

Pasul E:

$E[n_1|x_i, \lambda^{(t)}]$ este numărul „așteptat“ de instanțe neobservabile z_j care au valoarea 1, din totalul de $n - m$ instanțe neobservabile. În cazul distribuției Poisson, definiția funcției de densitate implică $P(x = 1|\lambda) = \frac{1}{e^\lambda} \cdot \frac{\lambda^1}{1!} = \frac{1}{e^\lambda} \cdot \lambda$ și $P(x = 0|\lambda) = \frac{1}{e^\lambda} \cdot \frac{\lambda^0}{0!} = \frac{1}{e^\lambda}$. Rezultă că $E[n_1|x_i, \lambda^{(t)}]$ este chiar media distribuției binomiale $\text{Bin}\left(n - m; \frac{\lambda^{(t)}}{1 + \lambda^{(t)}}\right)$,¹⁰¹⁸ deci

$$E[n_1|x_i, \lambda^{(t)}] = (n - m) \frac{\lambda^{(t)}}{1 + \lambda^{(t)}}.$$

Pasul M:

Tinând cont de expresia care tocmai a fost obținută la pasul E, vom descrie funcția auxiliară sub forma următoare:

$$Q(\lambda|\lambda^{(t)}) = -n\lambda + \left[(n - m) \frac{\lambda^{(t)}}{1 + \lambda^{(t)}} + \sum_{i=1}^m x_i \right] \ln \lambda - \sum_{i=1}^m \ln x_i!.$$

Derivatele întâi și a doua ale acestei funcții în raport cu λ sunt:

$$\begin{aligned} \frac{\partial}{\partial \lambda} Q(\lambda|\lambda^{(t)}) &= -n + \left[(n - m) \frac{\lambda^{(t)}}{1 + \lambda^{(t)}} + \sum_{i=1}^m x_i \right] \frac{1}{\lambda} \\ \frac{\partial^2}{\partial \lambda^2} Q(\lambda|\lambda^{(t)}) &= - \left[(n - m) \frac{\lambda^{(t)}}{1 + \lambda^{(t)}} + \sum_{i=1}^m x_i \right] \frac{1}{\lambda^2} \end{aligned}$$

Conform enunțului, $n > m$ și $\sum_{i=1}^m x_i \geq 0$. La pasul de inițializare al algoritmului EM, parametrul λ î se asignează o valoare $(\lambda^{(0)})$ pozitivă, ceea ce implică (prin inducție completă) că $\lambda^{(t)} > 0$ la orice iteratie $t > 0$, întrucât valoarea pentru care se anulează derivata întâi este

$$\lambda^{(t+1)} = \frac{1}{n} \left[(n - m) \frac{\lambda^{(t)}}{1 + \lambda^{(t)}} + \sum_{i=1}^m x_i \right].$$

Rezultă că derivata a doua a funcției auxiliare $Q(\lambda|\lambda^{(t)})$ este totdeauna negativă, ceea ce înseamnă că această funcție admite un maxim, și anume exact pentru $\lambda^{(t+1)}$.

¹⁰¹⁸Vedeți problema 25.b de la capitolul de *Fundamente*.

18.

(Algoritmul EM: estimarea probabilității de selecție a unei componente din cadrul unei mixturi – combinație liniară de două distribuții probabiliste oarecare)

■ • CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW2, pr. 2.1
CMU, 2006 spring, Carlos Guestrin, final exam, pr. 8

Se consideră variabila aleatoare X având funcția densitate de probabilitate sub forma unei mixturi de două componente:

$$p_\alpha(x) = \alpha \cdot p_1(x) + (1 - \alpha) \cdot p_2(x)$$

Se cunosc componentele $p_1(x)$ și $p_2(x)$ care reprezintă modele / funcții de densitate de probabilitate (considerate nespecificate aici), dar nu se cunoaște valoarea parametrului $\alpha \in [0, 1]$.

Cunoscând exemplele $\{x_1, x_2, \dots, x_n\}$ care au fost generate în mod independent și sunt identic distribuite conform distribuției variabilei X , formulați algoritmul EM pentru estimarea parametrului α . Descrieți în mod explicit cei doi pași — pasul E (pentru calculul mediilor) și pasul M (de maximizare) — ai algoritmului.

Observație: Se poate face în mod natural extensia la mixturi cu un număr arbitrar (dar fixat) de componente probabiliste.

Răspuns:

Pentru a aplica algoritmul EM considerăm variabilele ascunse $\{z_1, z_2, \dots, z_n\}$, cu $z_i \in \{1, 2\}$, pentru $i = \overline{1, n}$. Pentru fiecare valoare (fixată) a lui i , dacă $z_i = 1$, atunci exemplul x_i a fost generat de componenta $p_1(x)$ a mixturii, iar dacă $z_i = 2$, atunci exemplul x_i a fost generat de componenta $p_2(x)$.

Pasul E al algoritmului EM constă în calcularea mediei log-verosimilității datelor complete:

$$Q(\alpha | \alpha^{(t)}) \stackrel{\text{def.}}{=} E_{P(Z|X, \alpha^{(t)})} [\log P(X, Z | \alpha)]$$

Deoarece fiecare variabilă z_i depinde doar de x_i , iar datele x_i sunt independente între ele, putem scrie:

$$\begin{aligned} Q(\alpha | \alpha^{(t)}) &\stackrel{\text{def., i.i.d.}}{=} E_{P(Z|X, \alpha^{(t)})} \left[\sum_{i=1}^n \log p(x_i, z_i | \alpha) \right] \\ &\stackrel{\text{lin. med.}}{=} \sum_{i=1}^n E_{p(z_i|x_i, \alpha^{(t)})} [\log p(x_i, z_i | \alpha)] \\ &\stackrel{\text{def. E}}{=} \sum_{i=1}^n \left[\sum_{z_i \in \{1, 2\}} p(z_i | x_i, \alpha^{(t)}) \cdot \log p(x_i, z_i | \alpha) \right] \\ &= \sum_{i=1}^n \left[p(z_i = 1 | x_i, \alpha^{(t)}) \cdot \log p(x_i, z_i = 1 | \alpha) + \right. \\ &\quad \left. + p(z_i = 2 | x_i, \alpha^{(t)}) \cdot \log p(x_i, z_i = 2 | \alpha) \right] \\ &= \sum_{i=1}^n \left[p(z_i = 1 | x_i, \alpha^{(t)}) \cdot \log(\alpha p_1(x_i)) + \right. \\ &\quad \left. + p(z_i = 2 | x_i, \alpha^{(t)}) \cdot \log((1 - \alpha)p_2(x_i)) \right] \end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^n p(z_i = 1 | x_i, \alpha^{(t)}) \cdot \log(\alpha) + \sum_{i=1}^n p(z_i = 1 | x_i, \alpha^{(t)}) \cdot \log p_1(x_i) + \\
&\quad \sum_{i=1}^n p(z_i = 2 | x_i, \alpha^{(t)}) \cdot \log(1 - \alpha) + \sum_{i=1}^n p(z_i = 2 | x_i, \alpha^{(t)}) \cdot \log p_2(x_i)
\end{aligned}$$

Probabilitățile condiționate implicate în această formulă pot fi calculate cu ajutorul teoremei lui Bayes:

$$\begin{aligned}
p(z_i = 1 | x_i, \alpha^{(t)}) &\stackrel{T.B.}{=} \frac{p(x_i | z_i = 1, \alpha^{(t)}) \cdot p(z_i = 1 | \alpha^{(t)})}{p(x_i, \alpha^{(t)})} \\
&= \frac{p_1(x_i) \cdot \alpha^{(t)}}{p_1(x_i) \cdot \alpha^{(t)} + p_2(x_i) \cdot (1 - \alpha^{(t)})} \\
p(z_i = 2 | x_i, \alpha^{(t)}) &\stackrel{T.B.}{=} \frac{p(x_i | z_i = 2, \alpha^{(t)}) \cdot p(z_i = 2 | \alpha^{(t)})}{p(x_i, \alpha^{(t)})} \\
&= \frac{p_2(x_i) \cdot (1 - \alpha^{(t)})}{p_1(x_i) \cdot \alpha^{(t)} + p_2(x_i) \cdot (1 - \alpha^{(t)})}
\end{aligned}$$

Pasul M al algoritmului EM constă în determinarea valorii parametrului α pentru care se obține maximul expresiei $Q(\alpha | \alpha^{(t)})$, care a fost deja calculată mai sus. Pentru a determina $\alpha^{(t+1)} \stackrel{\text{def.}}{=} \operatorname{argmax}_{\alpha} Q(\alpha | \alpha^{(t)})$ vom folosi derivata de ordinul întâi în raport cu α :

$$\begin{aligned}
\frac{\partial Q(\alpha | \alpha^{(t)})}{\partial \alpha} = 0 &\Leftrightarrow \frac{1}{\alpha} \sum_{i=1}^n p(z_i = 1 | x_i, \alpha^{(t)}) - \frac{1}{1-\alpha} \sum_{i=1}^n p(z_i = 2 | x_i, \alpha^{(t)}) = 0 \\
&\Leftrightarrow (1-\alpha) \sum_{i=1}^n p(z_i = 1 | x_i, \alpha^{(t)}) - \alpha \sum_{i=1}^n p(z_i = 2 | x_i, \alpha^{(t)}) = 0 \\
&\Leftrightarrow \sum_{i=1}^n p(z_i = 1 | x_i, \alpha^{(t)}) - \alpha \left(\sum_{i=1}^n p(z_i = 1 | x_i, \alpha^{(t)}) + \sum_{i=1}^n p(z_i = 2 | x_i, \alpha^{(t)}) \right) = 0 \\
&\Leftrightarrow \sum_{i=1}^n p(z_i = 1 | x_i, \alpha^{(t)}) - \alpha \sum_{i=1}^n \underbrace{(p(z_i = 1 | x_i, \alpha^{(t)}) + p(z_i = 2 | x_i, \alpha^{(t)}))}_{=1} = 0 \\
&\Leftrightarrow \sum_{i=1}^n p(z_i = 1 | x_i, \alpha^{(t)}) - n \cdot \alpha = 0 \Rightarrow \alpha = \frac{1}{n} \sum_{i=1}^n p(z_i = 1 | x_i, \alpha^{(t)}).
\end{aligned}$$

Se observă ușor că această soluție corespunde unui punct de maxim (de fapt, unicul punct de maxim) al funcției Q . Prin urmare, ținând cont de expresia probabilității $p(z_i = 1 | x_i, \alpha^{(t)})$ care a fost calculată mai sus (vedeți pasul E), formula de actualizare pentru parametrul α este:

$$\alpha^{(t+1)} = \frac{1}{n} \sum_{i=1}^n \frac{\alpha^{(t)} \cdot p_1(x_i)}{\alpha^{(t)} \cdot p_1(x_i) + (1 - \alpha^{(t)}) \cdot p_2(x_i)}$$

Comentariu: Pentru o aplicare relativ simplă a acestui algoritm EM (extins pentru mai multe componente în mixtură) vedeți paginile 9-11 din documentul *Handout 12*, de Brani Vidakovic, de la Georgia Institute of Technology, *Bayesian Statistics course (ISyE 8843A)*, 2004.

19.

(Adevărat ori Fals?)

a.

CMU, 2002 fall, Andrew Moore, final exam, pr. 1.e

Spre deosebire de metoda gradientului, care se poate bloca în poziția unui optim local, algoritmul EM (Expectation-Maximization) identifică întotdeauna optimul global.

b.

CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh, midterm, pr. 1.8.b

Algoritmul EM nu micșorează valoarea funcției obiectiv de la o iterare la alta.

Răspuns:

a. Fals. Ambele metode se pot bloca într-un optim local.

b. Adevărat. La problema 1 s-a arătat că algoritmul EM maximizează o margine inferioară ($F(q(z), \theta)$) pentru funcția de log-verosimilitate a datelor observabile ($\log P(x|\theta) = \log \sum_z P(x, z|\theta)$). Metoda folosită pentru maximizare este una iterativă (“coordinate ascent”). La problema 2 s-a demonstrat că valoarea funcției de log-verosimilitate a datelor observabile nu se micșorează de la o iterare la alta a algoritmului EM.

8.2 Schema algoritmică EM — Probleme propuse

8.2.1 Fundamente teoretice

20. (ELBO pentru funcția obiectiv a algoritmului EM)

• CMU, 2020 fall, E. Xing, Z. Bar-Joseph, HW4, pr. 4.3

Considerăm datele observabile $X = \{x_i\}_{i=1}^n$, împreună cu variabilele neobservabile / ascunse $Z = \{z_i\}_{i=1}^n$, unde $z_i \in \{1, \dots, K\}$. Parametrii modelului [de tip mixtura de distribuții probabiliste pe care vrem să-l asociem acestor date] sunt desemnați prin simbolul θ , iar cu $\theta^{(t)}$ vom nota valoarea estimată pentru acești parametri la pasul t al algoritmului EM. Vrem să maximizăm log-verosimilitatea datelor „incomplete“, $\ell(X; \theta) \stackrel{\text{not.}}{=} \ln p(X|\theta)$.

Care dintre următoarele margini inferioare (engl., lower-bounds) pentru log-verosimilitatea datelor „incomplete“ sunt valide, adică $F(q, \theta) \leq \ell(X; \theta)$?

- $\sum_{i=1}^n \ln p(x_i|\theta)$
- $\sum_{i=1}^n \ln \left(\sum_{k=1}^K p(x_i, z_i = k|\theta) \right)$
- $\sum_{i=1}^n \sum_{k=1}^K p(z_i = k|x_i, \theta^{(t)}) \ln \frac{p(x_i, z_i = k|\theta)}{p(z_i = k|x_i, \theta^{(t)})}$
- $\frac{1}{K} \sum_{i=1}^n \sum_{k=1}^K \ln p(x_i, z_i = k|\theta) + n \ln K$
- niciuna dintre cele de mai sus.

21. (MAP EM: o întrebare)

• ○ CMU, 2012 fall, T. Mitchell, Z. Bar-Joseph, midterm, pr. 1.j

Să zicem că vi se dă un algoritm EM care maximizează verosimilitatea [datelor observabile] folosind un model [probabilist] cu variabile latente. Vi se cere să modificați algoritmul EM astfel încât să maximizeze [nu verosimilitatea, ci] probabilitatea a posteriori (MAP) a datelor observabile. Care pas (sau pași) ai algoritmului EM trebuie modificați:

- A. pasul E (expectation)
- B. pasul M (maximization)
- C. ambii pași
- D. nu este necesară nicio modificare.

22.

(Algoritmul EM semi-supervizat)

• Stanford, 2020 summer, Andrew Ng, HW3, pr. 4.a

EM (Expectation Maximization) este un exemplu clasic de algoritm folosit pentru *învățare nesupervizată* (adică, învățare în care sunt folosite variabile „ascunse“ sau latente). În această problemă vom explora una dintre modalitățile în care algoritmul EM poate fi adaptat pentru a realiza *învățare semi-supervizată*, caz în care avem atât exemple de antrenament etichetate cât și exemple neetichetate.

În cadrul standard al *învățării nesupervizate*, avem $n \in \mathbb{N}$ exemple neetichetate $\{x^{(1)}, \dots, x^{(n)}\}$ și dorim să învățăm parametrii distribuției $p(x, z; \theta)$ pornind de la date, însă variabilele $z^{(i)}$ nu sunt observabile. *Algoritmul EM clasic*, care a fost conceput exact pentru acest scop, maximizează în mod indirect $p(x; \theta)$, verosimilitatea datelor observabile — a realiza această sarcină în mod direct este în general dificil, eventual chiar nerealizabil —, executând în mod iterativ pasul E și pasul M și maximizând de fiecare dată o *margine inferioară* (engl., lower bound) pentru $p(x; \theta)$ (ceea ce în general este o sarcină fezabilă). Funcția obiectiv poate fi scrisă în mod concret astfel:

$$\ell_{\text{unsup}}(\theta) = \sum_{i=1}^n \ln p(x^{(i)}; \theta) = \sum_{i=1}^n \ln \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta)$$

Aici vom urmări să construim o extensie / variantă a algoritmului EM care să funcționeze pentru *cadrul semi-supervizat*. Vom presupune că avem $\tilde{n} \in \mathbb{N}$ exemple *aditionale* etichetate, $\{(\tilde{x}^{(1)}, \tilde{z}^{(1)}), \dots, (\tilde{x}^{(\tilde{n})}, \tilde{z}^{(\tilde{n})})\}$, unde atât $\tilde{x}^{(i)}$ cât și $\tilde{z}^{(i)}$ sunt observabile. Dorim să maximizăm simultan *verosimilitatea marginală* (engl., marginal likelihood) a parametrilor în raport cu *exemplile neetichetate*, precum și *verosimilitatea totală* (engl., full likelihood) a parametrilor în raport cu *exemplile etichetate*, și anume optimizând suma lor ponderată (cu ajutorul unui anumit hiperparametru α). Mai precis, funcția obiectiv semi-supervizată $\ell_{\text{semi-sup}}(\theta)$ poate fi scrisă astfel:

$$\ell_{\text{semi-sup}}(\theta) = \ell_{\text{unsup}}(\theta) + \alpha \ell_{\text{sup}}(\theta),$$

unde

$$\ell_{\text{sup}}(\theta) = \sum_{i=1}^{\tilde{n}} \ln p(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta).$$

a. Formulele de actualizare de la pașii E și M ai algoritmului EM pentru cazul semi-supervizat se pot determina folosind aceeași abordare ca și în cazul nesupervizat. Arătați că obținem:

Pasul E (semi-supervizat): pentru $i \in \{1, \dots, n\}$, calculează

$$q^{(t)}(z^{(i)}) = p(z^{(i)} | x^{(i)}; \theta^{(t)}).$$

Pasul M (semi-supervizat):

$$\theta^{(t+1)} = \arg \max_{\theta} \left[\sum_{i=1}^n \left(\sum_{z^{(i)}} q^{(t)}(z^{(i)}) \ln \frac{p(x^{(i)}, z^{(i)}; \theta)}{q^{(t)}(z^{(i)})} \right) + \alpha \left(\sum_{i=1}^{\tilde{n}} \ln p(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta) \right) \right].$$

b. [Convergență.]

Demonstrați că acest algoritm converge. Pentru aceasta, este suficient să arătați că $\ell_{\text{semi-sup}}(\theta)$, adică funcția obiectiv pentru cazul semi-supervizat, crește monoton la fiecare iterație a noului algoritm EM. Mai specific, arătați că $\ell_{\text{semi-sup}}(\theta^{(t+1)}) \geq \ell_{\text{semi-sup}}(\theta^{(t)})$, unde cu $\theta^{(t)}$ am notat valorile parametrilor θ obținute la sfârșitul iterării t a acestui algoritm EM.

Observație: pentru două „instanțe“ ale acestei scheme algoritmice de tip EM semi-supervizat, puteți vedea problema 70 de la capitolul de *Clusterizare* (EM/GMM semi-supervizat) și problema 37 de la prezentul capitol (EM/BernoulliMM semi-supervizat).

23.

(Algoritmul “hard” EM; exemplificare pentru rezolvarea unei mixturi de vectori de distribuții categoriale, folosind presupoziția de independentă condițională de tip Bayes Naiv, cu asignare “hard” a instanțelor la clustere)

*prelucrare de Liviu Ciortuz, după
□ • ○ CMU, 2014 spring, A. Singh, B. Poczos, HW3, pr. 2.2*

Fie un set de instanțe neetichetate $\mathcal{D} = \{x_1, \dots, x_n\}$. La problema 1.a (fundamentarea teoretică a algoritmului EM), am folosit la scrierea funcției de log-verosimilitate ($\log P(\mathcal{D}|\theta) = \sum_i \log P(x_i|\theta)$) faptul că $P(x_i|\theta)$ poate fi rescris / exprimat însumând probabilitățile comune (engl., joint probabilities) corespunzătoare lui x_i (pe de o parte) și fiecarei asignări posibile a variabilelor latente / neobservabile (pe de altă parte).

În acest exercițiu ne interesează să vedem ce se întâmplă atunci când în loc să facem sumarea despre care am vorbit mai sus, vom seta variabile neobservabile la valorile lor cele mai probabile în raport cu estimările / valorile actuale ale parametrilor. Această strategie este numită adeseori *algoritmul “hard” EM*. În anumite cazuri, el funcționează bine.¹⁰¹⁹

a. Arătați că și în situația în care în loc să procedăm la sumarea probabilităților pentru toate asignările posibile ale variabilelor latente — $P(X_i = x_i|\pi, \beta) = \sum_z P(X_i = x_i, Z = z|\pi, \beta)$ — facem maximizare, are loc o optimizare a unei margini inferioare pentru funcția de log-verosimilitate a datelor observabile.

b. Elaborați pasul E pentru varianta “hard” a algoritmului EM.

c. Elaborați pasul M pentru varianta “hard” a algoritmului EM într-un cadru particular, și anume atunci când *obiectivul* nostru este să rezolvăm problema de *învățare nesupervizată de tip Bayes Naiv*, însă nu folosind asignare “soft” a instanțelor la clustere cum am procedat la exercițiul 11, ci folosind asignare “hard”.

¹⁰¹⁹Spre exemplu, putem obține o variantă a algoritmului K-means aplicând algoritmul “hard” EM unei mixturi de distribuții gaussiene.

24.

(Algoritmul EM generalizat)

Stanford, 2007 fall, Andrew Ng, HW3, pr. 5

Uneori, atunci când încercăm să rulăm algoritmul EM, se poate să avem dificultăți în a executa pasul M în manieră exactă. (Vă reamintim că adeseori trebuie să implementăm proceduri numerice de optimizare ca să efectuăm maximizarea, ceea ce poate fi costisitor.) Așadar, în loc să găsim maximul global pentru marginea inferioară a funcției de log-verosimilitate, ar fi suficient să obținem o anumită creștere pe această margine inferioară (văzută ca funcție), executând (de exemplu) o iterație a algoritmului gradientului ascendent. Această variantă a algoritmului EM este cunoscută îndeobște sub numele de *algoritmul EM generalizat* (engl., Generalized EM, GEM).

Ca să ne exprimăm într-o manieră ceva mai formală, vă readucem aminte că la pasul M din algoritmul EM standard se execută maximizarea

$$\theta \leftarrow \operatorname{argmax}_{\theta} \sum_i \sum_{z^{(i)}} q_i(z^{(i)}) \ln \frac{p(x^{(i)}, z^{(i)}; \theta)}{q_i(z^{(i)})}.$$

În schimb, algoritmul GEM folosește la pasul M următoarea regulă de actualizare:

$$\theta \leftarrow \theta + \alpha \nabla_{\theta} \left(\sum_i \sum_{z^{(i)}} q_i(z^{(i)}) \ln \frac{p(x^{(i)}, z^{(i)}; \theta)}{q_i(z^{(i)})} \right),$$

unde α este rata de învățare, despre care presupunem că este aleasă suficient de mică, în aşa fel încât funcția obiectiv să nu descrească atunci când se execută acest pas [al gradientului ascendent].

a. Demonstrați că algoritmul GEM descris mai sus converge. Aceasta revine la a arăta că funcția de verosimilitate este monoton crescătoare, ca și în cazul algoritmului EM. Așadar, demonstrați că $\ell(\theta^{(t+1)}) \geq \ell(\theta^{(t)})$.

b. Să presupunem că în loc să folosim algoritmul EM am vrea să aplicăm algoritmul gradientului ascendent pentru a maximiza funcția de log-verosimilitate în mod direct. Cu alte cuvinte, încercăm să maximizăm funcția (non-convexă)

$$\ell(\theta) = \sum_i \ln \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta),$$

folosind regula de actualizare

$$\theta \leftarrow \theta + \alpha \nabla_{\theta} \left(\sum_i \ln \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta) \right).$$

Demonstrați că această regulă de actualizare este de fapt aceeași cu regula de actualizare de la algoritmul GEM descris mai sus.

8.2.2 Mixturi de distribuții Bernoulli / categoriale

25. (Două extensii ale algoritmului EM pentru estimarea parametrilor unei mixturi de două distribuții Bernoulli (cazul general))

• Liviu Ciortuz, 2023

a. Modificați algoritmul EM de la problema 6 astfel încât să putem rezolva mixturi de două distribuții Bernoulli folosind (pentru eficiență) două variabile aleatoare de tip binomial, \tilde{z}_1 și \tilde{z}_0 . Aceste două variabile aleatoare vor avea următoarea semnificație: \tilde{z}_1 va desemna numărul (necunoscut) de aruncări — din totalul celor n_s aruncări (număr cunoscut) la care s-a obținut față *stemă* — la care s-a folosit prima monedă; similar, \tilde{z}_0 va desemna numărul (necunoscut) de aruncări — din totalul celor n_B aruncări (număr cunoscut) la care s-a obținut față *ban* — la care s-a folosit cea de-a doua monedă.

Implementați algoritmul EM pentru rezolvarea acestui model de mixtură și apoi rulați-l pe inputul $x = \{1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1\}$, cu valorile inițiale $1/3$, $2/3$ și $1/2$ pentru parametrii p , q și respectiv π . Ce se observă în privința convergenței? Dar dacă se inițializează parametrii p , q și π cu alte valori?

b. Modificați algoritmul EM de la problema 6 astfel încât să putem rezolva mixturi de K (deci nu doar două) distribuții Bernoulli. Parametrii acestor distribuții vor fi notați cu p_1, \dots, p_K , iar probabilitățile de selecție [ale respectivei distribuții] cu π_1, \dots, π_K , unde $p_i \in [0, 1]$ și $\pi_i \in [0, 1]$ pentru $i = 1, \dots, K$, iar $\pi_1 + \dots + \pi_K = 1$.

Sugestie: La acest punct ar trebui ca variabilele neobservabile Z_i să ia valori în mulțimea $\{1, \dots, K\}$. Însă pentru a putea aplica ușor artificiul exponentierii (engl., the exponentiation trick), este mai convenabil ca în locul fiecărei astfel de variabile Z_i să folosiți variabilele-indicator Z_{i1}, \dots, Z_{iK} , cu proprietatea $Z_{ij} = 1$ dacă instanța de antrenament x_i a fost generată de către distribuția j și $Z_{ij} = 0$ în cazul contrar. De asemenea, întrucât avem restricția $\pi_1 + \dots + \pi_K = 1$, va trebui să folosiți metoda multiplicatorilor lui Lagrange. Vă puteți inspira din rezolvarea problemei 9 (sau 27).

Observație: Este posibil să combinați cele două extensii — care au fost descrise la punctele a și b — într-o singură, însă este ceva mai dificil.¹⁰²⁰

26. (Algoritmul EM pentru mixturi de distribuții Bernoulli: executarea manuală a unei iterări)

prelucrare de Liviu Ciortuz, după

• CMU, 2020 fall, E. Xing, Z. Bar-Joseph, HW4, pr. 5

Presupunem că dispunem de două monede, C_1 și C_2 , pentru care probabilitățile [respective] de apariție a feței *stemă* sunt necunoscute. Vom nota aceste două probabilități cu p și respectiv q . Moneda C_1 este aleasă cu probabilitatea π ,

¹⁰²⁰Cu titlu de *Sugestie*, vă sfătuim să folosiți $2K$ variabile aleatoare de tip binomial, și anume \tilde{z}_{1j} și \tilde{z}_{0j} , cu $j = 1, \dots, K$. Variabila \tilde{z}_{1j} va desemna numărul (necunoscut) de aruncări — din totalul celor n_s aruncări la care s-a obținut față *stemă* — la care s-a folosit moneda j ; similar, variabila \tilde{z}_{0j} va desemna numărul (necunoscut) de aruncări — din totalul celor n_B aruncări la care s-a obținut față *ban* — la care s-a folosit moneda j .

iar moneda C_2 este aleasă cu probabilitatea $1 - \pi$. Aruncăm moneda aleasă o dată, iar apoi notăm rezultatul. Acest experiment probabilist este repetat de cinci ori, iar rezultatul obținut în final este $X = \{H, H, T, H, T\}$.

Obiectivul acestei probleme este să estimăm parametrii $\theta \stackrel{\text{not.}}{=} (p, q, \pi)$ folosind algoritmul EM. Facem *convenția* următoare: dacă variabila neobservabilă Z_i ia valoarea 1, atunci înseamnă că la aruncarea cu numărul i a fost folosită moneda C_1 , iar dacă $Z_i = 0$ atunci a fost folosită moneda C_2 . Veți inițializa parametrii cu următoarele valori: $p^{(0)} = \frac{1}{4}$, $q^{(0)} = \frac{2}{3}$ și $\pi^{(0)} = \frac{1}{2}$.

Observație: Rezolvarea acestei probleme necesită în prealabil scrierea algoritmului EM/BMM de la problema 6 în sensul *Observației importante* (2) de la finalul rezolvării problemei 7.

Note: Pentru întrebările $b - h$ de mai jos, vă cerem să scrieți răspunsul sub forma unei *fracții ireductibile* de forma $\frac{a}{b}$, unde a și b sunt numere naturale prime între ele. De exemplu, dacă răspunsul este 0.4125, va trebui să-l scrieți sub forma $\frac{33}{80}$.

- Faceți o *repräsentare grafică* pentru mixtura descrisă în enunțul problemei.
- Pasul E:** Calculați $r_1^{(0)} \stackrel{\text{not.}}{=} P(z_i = 1|x_i = H; \theta^{(0)})$, unde $\theta^{(0)} \stackrel{\text{not.}}{=} (p^{(0)}, q^{(0)}, \pi^{(0)})$.
- Notăm cu \tilde{z}_1 numărul de fețe *stemă* (engl., head, H) produse de moneda C_1 . Observați faptul că \tilde{z}_1 este o variabilă aleatoare (latentă / ascunsă / neobservabilă). Calculați media (sau, valoarea „așteptată“; engl., expected value) pentru variabila \tilde{z}_1 la prima iterație a algoritmului EM. Justificați răspunsul în mod detaliat.
- Pasul E:** Calculați $r_0^{(0)} \stackrel{\text{not.}}{=} P(z_i = 1|x_i = T, \theta^{(0)})$.
- Notăm cu \tilde{z}_0 numărul de fețe *ban* (engl., tail, T) produse de moneda C_1 . Observați faptul că \tilde{z}_0 este o variabilă aleatoare (de asemenea latentă). Calculați media (sau, valoarea „așteptată“; engl., expected value) a lui \tilde{z}_0 la prima iterăție a algoritmului EM. Justificați răspunsul în mod detaliat.
- Pasul M:** Calculați $\pi^{(1)}$, $p^{(1)}$, $q^{(1)}$.
- Scriți expresia funcției $L(\theta^{(t)}) \stackrel{\text{not.}}{=} P(X|\theta^{(t)})$, verosimilitatea datelor observabile la iterăția t . Apoi calculați $L(\theta^{(1)})$ și $L(\theta^{(0)})$. La final, verificați că $L(\theta^{(1)}) \geq L(\theta^{(0)})$.

27.

(Algoritmul EM: estimarea parametrilor unei mixturi de distribuții categoriale: cazul general¹⁰²¹⁾

■ □ • ○ CMU, 2015 spring, T. Mitchell, N. Balcan, HW6, pr. 1

La acest exercițiu vom lucra cu modele de mixturi de distribuții *categoriale*.¹⁰²²

Fie un set de date $x = \{x_1, \dots, x_n\}$, unde fiecare $x_i \in \{v_1, \dots, v_M\}$ este generat [în mod independent de ceilalți x_j , cu $j \neq i$] de către una din K distribuții categoriale posibile, notată cu X_i . Vom considera că distribuția care l-a generat pe x_i a fost desemnată de către o altă variabilă aleatoare categorială, luând valori în multimea $\{1, \dots, K\}$ și având vectorul de parametri $\pi \in [0, 1]^K$, cu $\sum_k \pi_k = 1$. Vom nota cu $\theta_k \in \mathbb{R}^M$ parametrul distribuției categoriale asociate cu componenta k din mixtură, deci $\theta_k \in (0, 1)^M$ și $\sum_{j=1}^M \theta_{kj} = 1$ pentru $k = 1, \dots, K$.

Procesul generativ pentru un *model de mixturuă categorială* poate fi sumarizat astfel: $Z_i \sim \text{Categorial}(\pi)$, $X_i \sim \text{Categorial}(\theta_{Z_i})$. Pentru acest model, în care observăm variabilele X dar nu și variabilele Z , obiectivul este să învățăm parametrii $\Theta = \{\pi, \theta_1, \dots, \theta_K\}$. Veți folosi algoritmul EM pentru a realiza acest obiectiv.

Observație: Atunci când lucrăm cu distribuții categoriale, este utilă folosirea funcțiilor indicator [în locul variabilelor-indicator]. Prin definiție, funcția-indicator $1_{\{x=j\}}$ are valoarea 1 dacă $x = j$ și 0 în caz contrar.¹⁰²³

- Calculați *distribuția comună* a datelor observabile (X) și neobservabile (Z): $P(X, Z; \Theta)$.
- Calculați *probabilitățile a posteriori* corespunzătoare *variabilelor latente*, $P(Z_i = k | X_i; \Theta)$.
- Calculați *media log-verosimilități* datelor complete,

$$Q(\Theta | \Theta') \stackrel{\text{def.}}{=} E_{Z|X; \Theta'} [\log P(X, Z; \Theta)].$$

- Deducreți regulile de actualizare pentru parametrii Θ . Altfel spus, care este valoarea lui Θ care maximizează funcția $Q(\Theta | \Theta')$ calculată la punctul c?

Indicație: Aveți grijă că soluția pe care o veți obține trebuie să satisfacă restricția ca parametrii distribuțiilor categoriale trebuie să se sumeze la valoarea 1. Metoda multiplicatorilor Lagrange este o cale foarte convenabilă pentru rezolvarea unor astfel de probleme de optimizare cu restricții.

¹⁰²¹Problema aceasta reprezintă o generalizare în raport cu problema 6, unde am prezentat algoritmul EM pentru o mixturuă de două distribuții Bernoulli. Generalizarea se referă la faptul că în locul distribuțiilor Bernoulli aici se consideră distribuții categoriale, iar pe de altă parte, numărul de distribuții cu care se lucrează poate fi oricare ($K \geq 1$).

Problema 10 este un caz particular (sau, o ilustrare) pentru problema de față. Pe de altă parte, atunci când mixtura are $K = 2$ componente, problema 11 reprezintă o generalizare a problemei de față.

¹⁰²²La problema 44 de la capitolul de *Fundamente* am arătat cum se face estimarea în sens MLE pentru parametrii unei distribuții categoriale. Aici vom estima parametrii unei mixturi de K distribuții categoriale.

¹⁰²³De exemplu, dacă vom considera o variabilă aleatoare Y care ia valori în multimea $\{1, \dots, N\}$ și urmează o distribuție categorială (notație: $Y \sim \text{Categorial}(\phi)$, unde $\phi \in (0, 1)^N$), cu $\phi \stackrel{\text{not.}}{=} (\phi_1, \dots, \phi_N)$, vom putea să exprimăm probabilitatea ca Y să ia o anunită valoare în felul următor:

$$P(Y) = \prod_{i=1}^N \phi_i^{1_{\{Y=i\}}}.$$

28. (EM pentru combinarea unei clasificări de tip Bayes Naiv cu o mixtură de vectori de distribuții Bernoulli, folosind și un termen de penalizare / regularizare. Aplicație: recunoașterea cifrelor scrise de mână)

• ○ *U. Toronto, Radford Neal,
“Statistical Methods for Machine Learning and Data Mining” course,
2014 spring, HW2*

Obiectivul dumneavoastră în această problemă va fi să clasificați cifre scrise de mână, cu ajutorul unor modele de mixturi [de distribuții probabiliste] care vor fi obținute prin maximizarea verosimilității datelor însotită de un termen de penalizare, folosind algoritmul EM.

Datele pe care le veți folosi constau din 800 de imagini de antrenament și 1000 de imagini de test pentru cifre scrise de mână (din codurile poștale de pe plicuri din SUA). Aceste imagini au fost extrase din bine-cunoscutul set de date MNIST, prin selectare aleatorie dintr-un total de 60000 de imagini de antrenament furnizate, reducând rezoluția acestor imagini de la 28×28 de pixeli la 14×14 de pixeli, făcând media valorilor pixelilor pentru fiecare bloc de pixeli de dimensiune 2×2 , iar apoi folosind niște praguri pentru aceste medii în așa fel încât să obținem valori binare. Pe pagina web a acestei culegeri¹⁰²⁴ este furnizat un fișier de date de antrenament având 800 de linii, fiecare linie conținând 196 de valori pentru pixeli și anume, fie 0, fie 1. Tot acolo este furnizat și un alt fișier care conține etichetele corespunzătoare acestor 800 de cifre (de la 0 to 9). În mod similar, vă punem la dispoziție și un fișier cu 1000 de imagini de test, precum și un fișier conținând etichetele acestor 1000 de imagini de test. (Nu veți avea voie să vă uitați la etichetele acestor imagini de test decât la sfârșit de tot, și anume atunci când va trebui să analizați cât de bine se comportă algoritmii de învățare automată.)

Veți clasifica aceste imagini pentru cifre scrise de mână folosind un *model generativ*, din care, dată fiind imaginea unei astfel de cifre, veți deriva probabilitățile pentru cele 10 clase posibile. *Clasa asociată unei cifre de test este cea care are probabilitatea maximă.*

A. Modelul generativ pe care îl vom folosi estimează *probabilitățile claselor* ca [fiind] *frecvențele* lor din setul de date de antrenament (ceea ce va determina o distribuție aproximativ uniformă — însă nu exact uniformă — peste cele 10 cifre). De asemenea, acest model estimează *distribuțile de probabilitate ale imaginilor din cadrul fiecărei clase* cu ajutorul unui *model de mixtură cu K componente*, fiecare componentă modelând valorile celor 196 de pixeli ca [fiind] variabile aleatoare *independente*. Va fi convenabil să combinăm toate aceste 10 modele de mixturi [de distribuții probabiliste] într-un singur *model de mixtură cu 10K componente*, care modelează atât valorile pixelilor cât și etichetele claselor. Totuși, *probabilitățile [conditionate ale] componentelor* din cadrul modelului fiecărei clase vor fi *fixate*, astfel încât K componente vor atribui probabilitatea condiționată 1 cifrei 0, alte K componente vor atribui probabilitatea condiționată 1 cifrei 1, alte K vor atribui probabilitatea condiționată 1 cifrei 2 și.m.d.

Așadar, modelul probabilist pentru *distribuția comună* asupra etichetelor y_i , precum și asupra valorilor pixelilor $x_{i,1}, \dots, x_{i,196}$ pentru $i = 0, \dots, 9$ poate fi

¹⁰²⁴Vedeți <https://profs.info.uaic.ro/~ciortuz/ML.ex-book/implementation-exercises/UToronto.2014s.RNeal.HW2.EM-for-BernoulliMM-using-the-NBayes-assumption.handwritten-char-reco.data+R-code+sol/>

exprimat sub forma următoare:

$$P(y_i, x_i | \pi, \theta) = \sum_{k=1}^{10K} \left(\pi_k q_{k,y_i} \prod_{j=1}^{196} \theta_{k,j}^{x_{i,j}} (1 - \theta_{k,j})^{1-x_{i,j}} \right).$$

Instanțele etichetate (x_i, y_i) se presupun a fi independente.

Parametrii acestui model sunt *probabilitățile de mixare / selecție*, π_1, \dots, π_{10K} , precum și probabilitățile $\theta_{k,j}$ asociate pixelilor (mai precis, pentru ca aceștia să fie setați la valoarea 1), pentru fiecare componentă în parte, adică pentru $k = 1, \dots, 10K$ și $j = 1, \dots, 196$. Probabilitățile condiționate $q_{k,y}$ ale componentelor fiecărei clase sunt fixate astfel:

$$q_{k,y} = \begin{cases} 1 & \text{atunci când } k \in \{Ky + 1, \dots, Ky + K\} \\ 0 & \text{în caz contrar,} \end{cases}$$

pentru $k = 1, \dots, 10K$ și $y = 0, \dots, 9$.

La sfârșitul acestui exercițiu veți implementa în Python / R / Matlab o funcție care identifică valorile parametrilor care maximizează log-verosimilitatea datelor de antrenament plus un *termen de penalizare* (engl., penalty term).¹⁰²⁵

Algoritmul EM poate fi ușor adaptat pentru a găsi [în locul estimării verosimilității maxime] estimarea verosimilității maxime penalizate. În raport cu versiunea generală a algoritmului EM [care a fost prezentată la curs], pasul E se păstrează neschimbăt, însă pasul M va maximiza acum o funcție care reprezintă media log-verosimilității penalizate a datelor complete, $E_Q[\ln P(x, z | \theta) + G(\theta)]$, unde $G(\theta)$ este termenul de penalizare.

Scopul penalizării este acela de a evita ca estimările probabilităților asociate pixelilor să fie 0 sau aproape de 0, situații care pot cauza probleme la clasificarea instanțelor de test. (Este imediat că dacă o astfel de probabilitate este 0, rezultă că probabilitatea oricărei instanțe de test va fi 0 pentru clasa respectivă.) Termenul de penalizare care se sumează la log-verosimilitate trebuie să fie

$$G(\theta) = \alpha \sum_{k=1}^{10K} \sum_{j=1}^{196} [\log(\theta_{k,j}) + \log(1 - \theta_{k,j})],$$

unde constanta α controlează „magnitudinea“ penalizării.¹⁰²⁶

- a. Demonstrați că într-adevăr [regula de actualizare de la] pasul E rămâne la fel ca în cazul versiunii generale a algoritmului EM.
- b. La pasul M, reestimarea probabilităților de selecție π va rămâne de asemenea neschimbătă [în raport cu cazul general],¹⁰²⁷ însă pentru reestimarea parametrilor θ va trebui să luati în considerare și termenul de penalizare. Demonstrați că formula pentru actualizarea parametrilor $\theta_{k,j}$ la pasul M este

$$\hat{\theta}_{k,j} = \frac{\alpha + \sum_{i=1}^n r_{i,k} x_{i,j}}{2\alpha + \sum_{i=1}^n r_{i,k}},$$

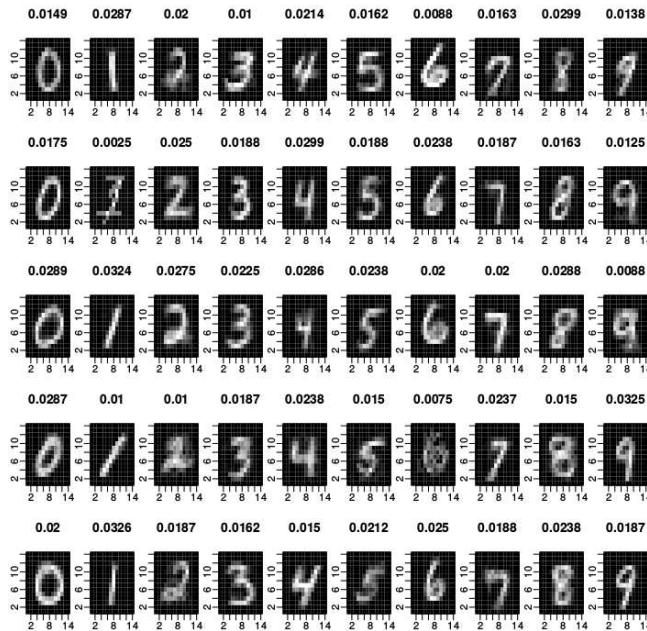
¹⁰²⁵Cu cât valorile acestui termen de penalizare vor fi mai mari, cu atât va fi mai bine.

¹⁰²⁶Pentru acest exercițiu, α va putea fi fixat la valoarea 0.05, deși în aplicații reale el va trebui probabil să fie determinat folosind, de exemplu, metoda cross-validation.

¹⁰²⁷Va trebui să vă convingeți singuri de acest fapt.

unde $r_{i,k}$ este probabilitatea ca instanța i să fi fost generată de către componentă k a mixturii, probabilitate care a fost estimată la pasul E.¹⁰²⁸

Comentariu: Pentru $K = 5$, putem ilustra sub forma unor imagini valorile parametrilor θ care au fost învățate pentru cele 50 de componente ale mixturii:



Este clar că în general cele cinci componente corespunzătoare fiecărei cifre au identificat diferite variante rezonabile de caligrafie a cifrelor, exceptând probabil câteva cazuri, cărora le corespund probabilități de selecție destul de mici (indicate sub formă numerică, deasupra imaginilor), așa cum este caracterul “1” de pe cel de-al doilea rând de imagini.

B. Funcția / procedura în cadrul căreia veți implementa algoritmul EM va trebui să ia ca argumente [numele fișierelor care conțin] imaginile din setul de date de antrenament și etichetele asociate acestor imagini, numărul de componente ale mixturii (acest număr, K , este același pentru toate clasele / cifrele), „magnitudinea“ penalizării (α), precum și numărul de iterații pe care trebuie să le execute algoritmul EM. Procedura aceasta va trebui să returneze o listă formată din estimările parametrilor (π și θ), precum și „responsabilitățile“ (r). Va trebui să startați algoritmul cu anumite valori inițiale pentru „responsabilități“, iar apoi să intrați în pasul M. „Responsabilitatea“ pentru componenta k a instanței / imaginii i trebuie să fie 0 în cazul în care pentru componenta k avem $q_{k,y_i} = 0$. Petru cazul $q_{k,y_i} = 1$, va trebui să setați în mod aleatoriu $r_{i,k}$ conform distribuției continue uniforme definite pe intervalul $[1, 2]$, iar apoi să rescalăți aceste valori astfel încât pentru fiecare valoare $i \in \{1, \dots, n\}$ să avem $\sum_{k=1}^K r_{i,k} = 1$.

c. După fiecare iterație, procedura dumneavoastră [care implementează algoritmul] EM va trebui să imprime valoarea log-verosimilității, precum și valoarea sumei dintre log-verosimilitate și termenul de penalizare. Suma aceasta

¹⁰²⁸Ca să obțineți această formulă, ați putea să porniți de la demonstrația care a fost prezentată la problema 9, modificând-o în aşa fel încât să includă și termenul de penalizare.

nu trebuie să descrească de la o iterare la alta — în cazul în care ea descrește, înseamnă că aveți o eroare în procedura EM. Va trebui să executați suficiente interații pentru ca aceste valori să se devină [aproape] stabilizate atunci când se ajunge la ultima iterare.

d. Veți implementa de astemenea o procedură care să ia ca input valorile parametrilor care au fost returnate de către procedura de antrenare EM și le folosește pentru a prezice clasa asociată unei imagini oarecare de test. Această procedură va trebui să folosească regula lui Bayes pentru a calcula probabilitățile ca imaginea de test respectiv să fi fost generată de fiecare dintre cele $10K$ componente ale mixturii, iar apoi să adune probabilitățile corespunzătoare celor K componente asociate fieacării clase / cifre, obținând astfel probabilitățile ca imaginea să reprezinte respectiv fiecare cifră de la 0 la 9. Procedura va trebui să returneze aceste probabilități, iar ele vor putea fi folosite ulterior pentru a decide cărei cifre îi corespunde imaginea respectivă de test, identificând cifra cu cea mai mare probabilitate.

e. Veți rula procedura EM pe datele de antrenament și apoi procedura de predicție pe toate datele de test pentru $K = 1$; aceasta din urmă ar trebui să producă exact aceleași rezultate ca și algoritmul Bayes Naiv.¹⁰²⁹

Apoi veți executa 10 rulări pentru $K = 5$, folosind diferite numere [ca “seed”-uri] pentru initializări aleatorii, iar după aceea veți afișa [ca rezultat] acuratețea la predicție / testare pentru fiecare dintre aceste 10 rulări.

Pentru fiecare imagine de test, veți calcula media probabilităților de apartență la fiecare clasă care au fost obținute la cele 10 rulări, iar apoi veți folosi aceste probabilități medii pentru a clasifica imaginile de test.

Veți încheia comparând mai întâi acuratețea produsă de către Bayes Naiv ($K = 1$) cu cea produsă folosind mixtura (cu $K = 5$), iar apoi acuratețea produsă de către „ansamblul“ de predicții cu acuratețile produse folosind câte o singură rulare a algoritmului EM și, în sfârșit, cu acuratețea produsă de către cea mai bună rulare a algoritmului EM în raport cu log-verosimilitatea datelor de antrenament (cu și, respectiv, fără termenul de penalizare).

29.

(Modele de mixturi pentru analiza semantică a documentelor de tip text:
modelul K-bag-of-words și modelul domeniilor semantice)

• ○ * CMU, 2011 fall, Eric Xing, HW5, pr. 2

Introducere:

În acest exercițiu veți compara două modele de analiză semantică pentru documente de tip text: modelul *K-bag-of-words* (în secțiunea A de mai jos) și modelul mixturii domeniilor semantice (în secțiunea B).¹⁰³⁰ Ambele modele sunt bazate pe *variable latente / neobservable*: valorile anumitor variabile sunt necunoscute, iar noi suntem interesați să găsim aceste valori. În modelul *K-bag-of-words*, vrem să aflăm pentru fiecare document (i) dintr-o colecție de documente date care este domeniul semantic „ascuns“ (t_i) asociat documentului respectiv (în manieră “soft”), dintr-o mulțime de K domenii. În

¹⁰²⁹ Veți observa că algoritmul EM converge imediat în cazul $K = 1$.

¹⁰³⁰ O versiune simplificată a modelului mixturii domeniilor semantice a fost prezentată la problema 12.

modelul domeniilor semantice, suntem interesați mai ales de distribuțiile „as-cunse“ ($\theta_i \stackrel{\text{not.}}{=} (\theta_i^1, \dots, \theta_i^K)$, cu $\theta_i^k \in [0, 1]$ și $\sum_{k=1}^K \theta_i^k = 1$) asociate documentului i și de domeniu semantic (sau: sensul) asociat (și aici în manieră “soft”) fiecărui cuvânt (z_{ij}).

A. Modelul K -bag-of-words — un sens pentru fiecare document

În cadrul primei părți a acestui exercițiu veți face cunoștință cu o mixtură (oarecum simplificată) de distribuții categoriale care este folosită pentru a modela [generarea de] cuvinte în documente de tip text. Acest model este cunoscut sub numele de *modelul celor K urne de cuvinte* (engl., *K -bag-of-words*).

Punctul de plecare îl constituie K distribuții categoriale care sunt asociate „urnelor“ de cuvinte. Fiecare dintre aceste K distribuții categoriale este definită peste cuvintele dintr-un (același!) vocabular. Numărul de cuvinte din vocabular este V . Parameterii acestor distribuții sunt desemnați respectiv prin β_1, \dots, β_K . Fiecare β_k (cu $k = 1, \dots, K$) este un vector V -dimensional, ale cărui componente sunt numere nenegative care, însumate dau valoarea 1.

Vom considera N documente, numerotate cu $1, \dots, N$, fiecare document i având M_i cuvinte. Remarcați faptul că documentele pot conține un număr diferit de cuvinte. Pentru fiecare document i , cuvântul cu numărul de ordine j va fi desemnat prin $w_{ij} \in \{1, \dots, V\}$. Așadar, vom folosi numere întregi (reprezentând *indicele poziției* corespunzătoare din vocabular) pentru a desemna cuvintele din documente.

Cuvintele (de fapt, *aparițiile* cuvintelor) din cele N documente sunt modelate după cum urmează:

Se consideră [încă] o distribuție de probabilitate de tip discret, [la care ne vom referi cu termenul *a priori* și] care este reprezentată prin vectorul π . Pentru fiecare document i se alege în mod aleatoriu, conform distribuției menționate, un *indicator* [de domeniu semantic / tematic] $t_i \in \{1, \dots, K\}$. Indicatorul [tematic] t_i ne spune care dintre cele K distribuții categoriale generează cuvintele din documentul i . Apoi, extragem (adică, generăm) fiecare cuvânt w_{ij} al documentului i în conformitate cu parametrii β_{t_i} ai distribuției categoriale t_i , extragerile făcându-se în mod independent unele de altele.¹⁰³¹

Acet design corespunde următorului *proces generativ*:

$$\begin{aligned} t_i &\sim \text{Categorial}(\pi) \text{ pentru } i \in \{1, \dots, N\} \\ w_{ij} &\sim \text{Categorial}(\beta_{t_i}) \text{ pentru } i \in \{1, \dots, N\} \text{ și } j \in \{1, \dots, M_i\}. \end{aligned}$$

Remarcați faptul că în acest model, în raport cu modelul clasic de mixtură de K distribuții categoriale, diferența este faptul că se consideră seturi de instanțe / date „observate“ care au în comun același indicator de mixtură t_i (și anume, cuvintele w_{ij} , cu $j = 1, \dots, M_i$).

Indicație: Vă cerem ca, în rezolvarea acestui exercițiu, să folosiți indici superioiri (engl., superscripts) pentru a desemna componentele vectorilor. De exemplu, prin notația π^k veți desemna elementul de pe poziția k din vectorul π . De asemenea, variabila-indicator t_i care a fost introdusă mai sus va fi rescrisă

¹⁰³¹Așadar, nu se ține cont de *ordinea* de apariție a cuvintelor în document.

sub forma vectorului-indicator $t_i \stackrel{\text{not.}}{=} (t_i^1, t_i^2, \dots, t_i^K)$. Deci prin t_i^k veți indica elementul de pe poziția k din vectorul-indicator t_i . Precizăm că, prin convenție, în acest vector doar una dintre componente este 1,¹⁰³² toate celelalte fiind 0.

a. Vectorul de probabilități de selecție π se notează cu $\pi = (\pi^1, \dots, \pi^K)$. Am precizat deja că probabilitatea selectării unui domeniu semantic $t_i \in \{1, \dots, K\}$ este π^{t_i} . Scriem formal acest fapt astfel: $P(t_i|\pi) = \pi^{t_i}$.

Vă cerem să exprimați într-o manieră unitară probabilitățile $P(t_i|\pi)$ pentru $i = 1, \dots, N$, folosind un artificiu de genul celui pe care l-am introdus la rezolvarea mixturilor de distribuții Bernoulli (vedeți relația (393) de la problema 6), extins acum la distribuția categorială.

b. Exprimați probabilitatea $P(w_{ij}|t, \beta, \pi)$ într-o formă cât mai simplă, ținând cont de relațiile de (in)dependentă. Altfel spus, va trebui să scrieți o egalitate de forma $P(w_{ij}|t, \beta, \pi) = P(w_{ij}|\dots)$, unde simbolul '...' desemnează o anumită submulțime din $\{t, \beta, \pi\}$. Simbolii t și β scriși fără indici inferiori (engl., subscripts) reprezintă ansamblul tuturor vectorilor t_1, \dots, t_N , respectiv β_1, \dots, β_K .

Sugestie: Încercați să reprezentați grafic modelul de mixtură de tip *K-bag-of-words*. Probabilitatea pe care o veți indica în răspunsul final va trebui să fie condiționată în raport cu toți vectorii β_1, \dots, β_K .

c. Explicați probabilitatea „simplificată” pe care ați indicat-o în răspunsul de la punctul b, folosind definiția distribuției categoriale.

B. Modelul mixturii domeniilor semantice — *un sens pentru fiecare cuvânt*

În cea de-a doua parte a exercițiului nostru vom explora „mixtura de mixturi” care formează baza aşa-numitului *model al [mixturii] domeniilor semantice* (engl., topic model) pentru clusterizare de documente de tip text.¹⁰³³ Putem gândi acest model ca fiind obținut prin operarea a două modificări asupra modelului (mai simplu) *K-bag-of-words* care a fost prezentat în secțiunea A:

În primul rând, în loc ca documentului i să-i fie asignat un singur *domeniu semantic*¹⁰³⁴ t_i , acum îi vom permite acestuia să fie caracterizat / modelat de o *mixtură de domenii semantice*. Vom asocia acestei mixturi un vector $\theta_i \stackrel{\text{not.}}{=} (\theta_i^1, \dots, \theta_i^K)$, cu $\theta_i^k \in [0, 1]$ și $\sum_{k=1}^K \theta_i^k = 1$. Mai precis, θ_i trebuie văzut ca o distribuție de probabilitate definită peste K domenii semantice reprezentate respectiv de tuplurile de parametri β_1, \dots, β_K . Ca și în secțiunea A, pentru fiecare $k \in \{1, \dots, K\}$, β_k este un vector, $(\beta_k^1, \dots, \beta_k^V)$, cu $\beta_k^v \in [0, 1]$ și $\sum_{v=1}^V \beta_k^v = 1$, unde V este mărimea vocabularului. Aceasta din urmă este comun pentru toate documentele ($i = 1, \dots, N$). Distribuția probabilistă desemnată prin $\theta_i = (\theta_i^1, \dots, \theta_i^K)$ este una de tip *Dirichlet* și va fi prezentată mai jos.

În al doilea rând, introducem pentru fiecare cuvânt w_{ij} câte un *indicator* $z_{ij} \in \{1, \dots, K\}$ care determină domeniul semantic asociat cuvântului w_{ij} . Remarcați cum diferă acest design de modelul *K-bag-of-words*: acum permitem fiecărui cuvânt să aibă propriul său domeniu semantic, în loc să-i impunem să „adopte” domeniul semantic al documentului (t_i) .¹⁰³⁵ Desigur, vom genera va-

¹⁰³²Si anume, poziția corespunzătoare *variabilei*-indicator t_i , care a fost introdusă anterior.

¹⁰³³Pentru mai multe informații, vă recomandăm articolul *Latent Dirichlet Allocation*, Blei et al, 2003.

¹⁰³⁴Sau: tematică.

¹⁰³⁵Cele două abordări / principii sunt cunoscute în domeniul lingvisticii computaționale sub forma *un sens la fiecare apariție* (engl., “one sense per occurrence”), respectiv *un /singur/ sens pe întregul document* (engl., “one sense per document”).

lorile variabilei z_{ij} conform distribuției $\theta_i = (\theta_i^1, \dots, \theta_i^K)$ asociate documentului i .

Acstea două schimbări dau naștere următorului *proces generativ*:

$$\theta_i = (\theta_i^1, \dots, \theta_i^K) \sim \text{Dirichlet}(\alpha) \text{ pentru } i \in \{1, \dots, N\}$$

$$z_{ij} \in \{1, \dots, K\} \sim \text{Categorial}(\theta_i) \text{ pentru } i \in \{1, \dots, N\} \text{ și } j \in \{1, \dots, M_i\},$$

$$w_{ij} \in \{1, \dots, V\} \sim \text{Categorial}(\beta_{z_{ij}}) \text{ pentru } i \in \{1, \dots, N\} \text{ și } j \in \{1, \dots, M_i\},$$

unde M_i este lungimea documentului i , iar $\alpha > 0$ este un parametru scalar pentru distribuția (simetrică) Dirichlet. Distribuția Dirichlet este definită sub forma următoare:

$$P(\theta_i | \alpha) = \frac{\Gamma(K\alpha)}{[\Gamma(\alpha)]^K} \prod_{k=1}^K (\theta_i^k)^{\alpha-1},$$

simbolul Γ desemnând funcția Gamma.¹⁰³⁶

Observați că acest model este într-adevăr o „mixtură de mixturi“ de distribuții categoriale: fiecărui document i i se asociază o mixtură θ_i peste domeniile β_1, \dots, β_K , și există N astfel de mixturi, $\theta_1, \dots, \theta_N$, care constituie împreună mixtura de distribuții categoriale.

- d. Exprimăți probabilitatea $P(z_{ij} | \theta, \alpha, \beta)$ într-o formă cât mai simplă, ținând cont de relațiile de (in)dependență. Simbolii θ și β scriși fără indici (engl., subscripts) reprezintă ansamblurile de parametri $\theta_1, \dots, \theta_N$ și respectiv β_1, \dots, β_K .
- e. Explicitați probabilitatea „simplificată“ pe care ati indicat-o ca răspuns la punctul precedent, folosind definiția distribuției categoriale.
- f. Exprimăți probabilitatea $P(w_{ij} | z, \theta, \alpha, \beta)$ într-o formă cât mai simplă, ținând cont de relațiile de (in)dependență.
- g. Explicitați probabilitatea „simplificată“ pe care ati indicat-o ca răspuns la punctul precedent, folosind definiția distribuției categoriale.

C. Comparație

Punctele / întrebările următoare vă vor ajuta să faceți o *comparație* între modelul K -bag-of-words și modelul mixturii domeniilor semantice pe care le-am prezentat la secțiunile A și respectiv B.

- h. Atât indicatorul t_i din modelul K -bag-of-words, cât și distribuția θ_i din modelul mixturii domeniilor semantice spun ceva anume despre conținutul tematic (sau: domeniu semantic) al documentului i . Formulați într-o singură frază diferența principală dintre t_i și θ_i .
- i. Discutați implicațiile răspunsului de la punctul h. În ce fel este mai utilă modelarea de tipul mixturii domeniilor semantice decât *modelarea* de tip K -bag-of-words?
- j. Atât în modelul K -bag-of-words cât și în modelul mixturii domeniilor semantice, parametrii β_k reprezintă vocabulare pentru fiecare domeniu semantic k . Nu discutăm aici despre învățarea valorilor parametrilor β , dar se poate

¹⁰³⁶Vedeți problema 46 de la capitolul de *Fundamente*, precum și http://en.wikipedia.org/wiki/Gamma_function.

arăta că anumite strategii de învățare clasice (de exemplu, *Algoritmul EM*¹⁰³⁷ și *Gibbs sampling*) vor „produce“ uneori domenii semantice care folosesc în comun (engl., share) anumite cuvinte — altfel spus, putem avea $\beta_k^v > 0$ și $\beta_l^v > 0$ pentru un cuvânt v și două domenii semantice distincte k și l . De ce este oare util aşa ceva (ne referim la *word sharing*)?

8.2.3 Distribuții binomiale / multinomiale

30. (Estimarea în sens MLE a parametrilor unor [mixturi de distribuții Bernoulli, cu ajutorul unor] distribuții binomiale: rezolvare folosind 1. metode clasice de optimizare și 2. algoritmul EM)

*prelucrare de Liviu Ciortuz, după
□ • ○ U. Toronto, 2015 fall, Statistical Computation course,
Radford Neal, HW1 and HW2, pr. 1*

A. Să presupunem că un student de la Universitatea din Toronto este interesat să afle ce procentaj din populația adultă din Toronto joacă jocul Minecraft în mod regulat. În acest scop, el a trimis chestionare la 130 de adulți pe care i-a selectat în mod uniform aleatoriu dintre toți adulții din Toronto. În mod surprinzător, el a primit respunsuri de la toți cei cărora le-a trimis chestionare.¹⁰³⁸ Dintre cele $n = 130$ persoane care au fost incluse în acest sondaj, $x = 75$ au spus că joacă Minecraft.

Estimarea de verosimilitate maximă (MLE) pentru procentajul de adulți din Toronto care joacă Minecraft se poate obține ușor: $x/n = 75/130$. Din păcate, studentul își dă seama doar acum că nu doar lucrul acesta îl interesează, ci el ar dori să cunoască și procentajele de bărbați și respectiv de femei [din Toronto] care joacă Minecraft. Însă el nu cunoaște genul persoanelor care au răspuns la chestionar.

Pentru a rezolva această chestiune, studentul decide să trimită încă un chestionar la 25 de bărbați selectați în mod aleatoriu dintre bărbații din Toronto și la 25 de femei selectate în mod aleatoriu dintre femeile din Toronto. (El nu poate să-și permită să lucreze cu un eșantion mai mare, fiindcă nu mai are multe fonduri disponibile.) Din fericire, norocul este în continuare de partea lui și toate persoanele cărora le-a adresat acest al doilea chestionar îi răspund. Dintre cei $m_1 = 25$ de bărbați care la care a fost trimis acest al doilea chestionar, $x_1 = 20$ spun că joacă Minecraft, iar dintre cele $m_2 = 25$ de femei la care a fost trimis acest chestionar, $x_2 = 6$ spun că joacă Minecraft.

În cele două secțiuni care urmează, vă vom cere să aflați (în două maniere diferite) estimările de verosimilitate maximă pentru procentajul de bărbați din Toronto care joacă Minecraft (p_1) și respectiv procentajul de femei din Toronto care joacă Minecraft (p_2), folosind datele de mai sus. Veți folosi presupoziția că populația orașului Toronto este alcătuită în mod egal din bărbați și femei

¹⁰³⁷Vedeți problema 12 de la acest capitol.

¹⁰³⁸Desigur, în cazul sondajelor reale, nereturnarea răspunsurilor la chestionare este o mare problemă, pe care însă noi o vom ignora în cazul de față.

și veți presupune, de asemenea, că faptul că o persoană joacă Minecraft este independent de faptul că o altă persoană joacă sau nu Minecraft.

B. Funcția de verosimilitate a datelor din secțiunea A poate fi scrisă ținând cont de următoarele fapte:

- x urmează distribuția binomială de parametri n și $(p_1 + p_2)/2$,
- x_1 urmează distribuția binomială de parametri m_1 și p_1 , iar
- x_2 urmează distribuția binomială de parametri m_2 și p_2 .

Așadar, putem să exprimăm funcția de verosimilitate astfel:

$$\begin{aligned} L(p_1, p_2) = & C_n^x ((p_1 + p_2)/2)^x (1 - (p_1 + p_2)/2)^{n-x} \\ & \times C_{m_1}^{x_1} p_1^{x_1} (1 - p_1)^{m_1 - x_1} \times C_{m_2}^{x_2} p_2^{x_2} (1 - p_2)^{m_2 - x_2} \end{aligned}$$

Obiectivul în această secțiune a problemei va fi să implementați proceduri care să identifice estimările de verosimilitate maximă pentru procentajele p_1 și p_2 pentru datele din secțiunea A folosind mai multe metode clasice de optimizare, iar apoi să comparați cum anume se comportă ele. În toate aceste metode veți maximiza log-verosimilitatea, nu verosimilitatea datelor.

Metodele de optimizare pe care va trebui să le testați sunt:

- Metoda gradientului ascendent.
- Metoda Newton-Raphson.¹⁰³⁹
- Metoda maximizării alternante [pe coordonate], folosind metoda bisecției pentru a găsi în mod alternativ maximul în raport cu p_1 și respectiv în raport cu p_2 .¹⁰⁴⁰

Veți testa aceste metode pe datele care au fost descrise mai sus și veți discuta cât de bine lucrează ele. În particular, veți analiza cel puțin următoarele aspecte:

- i. cât de sensibile sunt aceste metode la inițializările aleatorii;
- ii. dacă aceste metode produc același răspuns (cu condiția unei / unor inițializări aleatorii corespunzătoare), iar în cazul contrar, care dintre răspunsuri este cel mai bun (adică, se apropiie cel mai mult de maximul log-verosimilității);
- iii. cât de rapid converg aceste metode și, în mod specific dacă ele converg în timp liniar [LC: în raport cu numărul de instanțe de antrenament], pătratic sau de o altă natură;
- iv. cât de ușor sunt de implementat aceste metode.

Va trebui să calculați funcția de log-verosimilitate, vectorul ei gradient și matricea hessiană corespunzătoare. De asemenea, veți indica estimările pe care le-ați obținut, precum și orice alte reprezentări grafice sau textuale care credeți că sunt relevante pentru analiza pe care ați făcut-o.

C. În această secțiune, veți concepe, veți implementa și veți testa un algoritm EM pentru a găsi estimarea de verosimilitate maximă tot pentru datele care au fost prezentate în secțiunea A, însă acum veți considera că datele respective sunt modelate folosind o mixtură de distribuții probabiliste, iar ce componentă anume a mixturii a generat o anumită instantă poate uneori să fie observabil,

¹⁰³⁹Pentru o ușoară introducere în această metodă, vedeti problema 80 de la capitolul de *Fundamente*.

¹⁰⁴⁰Vedeți paginile 26-27 din <http://cs229.stanford.edu/notes2020fall/notes2020fall/cs229-notes3.pdf>.

iar alteori nu. Datele „neobservable” manipulate cu ajutorul algoritmului EM vor fi variabilele-indicator care arată ce componentă a mixturii a generat o instanță oarecare, atunci când această componentă nu este cunoscută.

Așadar, aici veți găsi estimările de verosimilitate maximă pentru procentajele p_1 și p_2 folosind algoritmul EM. Veți furniza rezultatele pentru aceleași date ca mai sus, și anume $n = 130$, $m_1 = 25$, $m_2 = 25$, $x = 75$, $x_1 = 20$ și $x_2 = 6$.

Datele „neobservable” din această problemă sunt genul celor x persoane care au răspuns *da* la primul chestionar, precum și genul celor $n - x$ persoane care au răspuns *nu* la acel chestionar.

a. Pentru pasul E al algoritmului, veți calcula valorile medii / „așteptate” (engl., expected values) pentru aceste două numere, ținând cont pe de o parte de datele observate și pe de altă parte de estimările curente pentru p_1 și p_2 .

b. La pasul M, veți maximiza valoarea medie (engl., expected value) pentru cât ar trebui să fie log-verosimilitatea datelor dacă am cunoaște aceste numere, această medie fiind calculată în raport cu distribuția probabilistă care a fost calculată la pasul E (și care este fixată pe durata execuției pasului M).

Observație: La pasul E, puteți să calculați valorile „așteptate” ale datelor neobservabile doar în cazurile în care ele sunt neapărat necesare pentru pasul M (întrucât este posibil ca doar mediile anumitor cantități să fie necesare pentru calculele de la pasul M).

c. Calculați expresia log-verosimilității datelor „observate” și verificați la fiecare iterație a algoritmului EM că ea nu descrește niciodată (probabil cu foarte puține excepții, datorate unor erori de rotunjire).

d. Comentați succint cât este (sau nu) de ușor să se folosească algoritmul EM pentru această problemă, și cât de repede se ajunge la convergență.

31. (Algoritmului EM pentru „învățarea” unei distribuții multinomiale care este definită cu ajutorul unui [singur] parametru; o aplicație în domeniul bioinformaticii)

*Liviu Ciortuz, 2017, după
■ □ ○ Georgia Institute of Technology,
Bayesian Statistics course (ISyE 8843A),
Brani Vidakovic, 2004, Handout 12, sec. 1.2.1*

Introducere: La problema 14 am arătat cum se pot estima parametrii unei distribuții categoriale (și, în mod implicit, ai unei distribuții multinomiale) în cazul unei probleme de bioinformatică, în care este implicată o genă (numită ABO) care are trei alele. Distribuția categorială din acea problemă avea trei parametri (de fapt doi, fiindcă suma celor trei parametri trebuie să fie 1). În problema de față, vom folosi un singur parametru (ψ) pentru a estima probabilitățile unei alte distribuții multinomiale, care este specifică unei alte probleme de bioinformatică, unde de această dată sunt considerate două perechi de gene, fiecare genă având câte două alele, iar probabilitățile de „recombinare” pentru aceste alele sunt diferite în cazul bărbaților, față de cazul femeilor.¹⁰⁴¹

¹⁰⁴¹Pentru mai multe explicații de tip biologic / genetic pentru această problemă, redăm aici secțiunea 1.2.1 din documentul indicat mai sus: *Handout 12*, de la cursul de Statistică bayesiană (ISyE 8843A) care a fost ținut

A. Fie o variabilă aleatoare X care ia valori în mulțimea $\{v_1, v_2, v_3, v_4\}$ și urmează distribuția $Multinomial\left(n; \frac{2+\psi}{4}, \frac{1-\psi}{4}, \frac{1-\psi}{4}, \frac{\psi}{4}\right)$, unde ψ este un parametru cu valori în intervalul $(0, 1)$. Cele patru probabilități listate în definiția acestei distribuții multinomiale sunt în corespondență directă cu valorile v_1, v_2, v_3 și respectiv v_4 .

a. Presupunem că n_1, n_2, n_3 și n_4 sunt numărul de „realizări“ ale valorilor v_1, v_2, v_3 și v_4 în totalul celor n „observații“. (Pentru fixarea ideilor, vom considera $n_1 = 125$, $n_2 = 18$, $n_3 = 20$ și $n_4 = 34$.) Calculați în manieră analitică estimarea de verosimilitate maximă (MLE) a parametrului ψ .

b. Fie acum variabila aleatoare $X' \sim Multinomial\left(n; \frac{1}{2}, \frac{\psi}{4}, \frac{1-\psi}{4}, \frac{1-\psi}{4}, \frac{\psi}{4}\right)$. Valorile luate de variabila X' , corespunzător acestor cinci probabilități, sunt (în ordine) v_1, v_1, v_2, v_3 și v_4 .¹⁰⁴² Vom considera n_{11}, n_{12}, n_2, n_3 și respectiv n_4 numărul de „realizări“ ale acestor valori, însă de data aceasta n_{11} și n_{12} vor fi neobservabile. În schimb, vom furniza suma $n_1 = n_{11} + n_{12}$ ca dată „observabilă“, alături de celelalte date observabile, n_2, n_3 și n_4 .

Să se estimeze parametrul ψ folosind algoritmul EM. Cum este această esti-

de către profesorul Brani Vidakovic la Georgia Institute of Technology, în anul 2004.

Vom considera două locusuri bi-alelice legate, A și B , având alele A și a , respectiv B și b , unde alela A este dominantă în raport cu alela a , iar alela B este dominantă în raport cu alela b .

Un heterozigot dublu $AaBb$ va produce gameti de patru tipuri: AB , Ab , aB și ab . Întrucât locusurile sunt legate, tipurile AB și ab vor apărea cu o frecvență diferită de frecvența tipurilor Ab și aB , să zicem $1 - r$ și respectiv r la bărbați, și $1 - r'$ și respectiv r' la femei.

Aici vom presupune că originea parentală a acestor heterozigoți este cauzată de împerecherea $AABB \times aabb$, astfel încât r și r' sunt probabilitățile / ratele de recombinare ale celor două locusuri la bărbați și respectiv la femei.

Problema este să estimăm probabilitățile r și r' , dacă este posibil, din descendenții imediați ai heterozigoților dubli.

Întrucât gameții AB , Ab , aB și ab sunt produși în proporții de $(1-r)/2$, $r/2$, $r/2$ și $(1-r)/2$ de către părinți masculini, respectiv $(1-r')/2$, $r'/2$, $r'/2$ și $(1-r')/2$ de către părinți feminini, zigotii care au genotipurile $AABB$, $AaBB$, ... etc., sunt produși cu frecvențele $(1-r)(1-r')/4$, $(1-r)r'/4$, etc.

Problema este următoarea: deși există 16 genotipuri distincte de descendenți, luând în considerare originea parentală, relațiile de dominare implică faptul că noi observăm doar 4 fenotipuri distincte, pe care le desemnăm prin $A * B*$, $A * b*$, $a * B*$ și $a * b*$. Aici $A *$ (respectiv $B *$) desemnează ceea ce este dominant, în vreme ce $a *$ (respectiv $b *$) desemnează fenotipurile recesive determinante de alele lui A (respectiv B).

Astfel, indivizi care au genotipurile $AABB$, $AaBB$, AAb sau $AaBb$, care reprezintă 9/16 din combinațiile de gameți (verificați!), manifestă fenotipul $A * B*$, adică alternativa dominantă în raport cu ambele caractere, în vreme ce indivizi care au genotipurile $AAbb$ sau $Aabb$ (3/16) manifestă fenotipul $A * b*$, indivizi cu genotipurile $aaBB$ și $aaBb$ (3/16) manifestă fenotipul $a * B*$ și, în fine, indivizi cu genotipul dublu recesiv $aabb$ (1/16) manifestă fenotipul $a * b*$.

Este cumva surprinzător faptul că probabilitățile celor patru clase fenotipice pot fi definite folosind parametrul ψ not.¹⁰⁴² $(1-r)(1-r')$, după cum urmează: $a * b*$ are probabilitatea $\psi/4$ (este ușor de verificat), $a * B*$ și $A * b*$ au ambele probabilitățile $(1-\psi)/4$, în vreme ce $A * B*$ are probabilitatea 1 minus suma probabilităților precedente, adică $(2+\psi)/4$.

Presupunem acum că avem un eșantion aleatoriu format din n descendenți imediați ai heterozigoților noștri dubli. Astfel, cele 4 clase fenotipice vor fi reprezentate grosso-modo proporțional cu probabilitățile lor teoretice, distribuția lor comună fiind de tip multinomial:

$$Multinomial\left(n; \frac{2+\psi}{4}, \frac{1-\psi}{4}, \frac{1-\psi}{4}, \frac{\psi}{4}\right).$$

Observați faptul că aici niciuna dintre probabilitățile r și r' nu poate fi estimată în mod separat din aceste date, ci doar produsul $(1-r)(1-r')$.

¹⁰⁴² Observați că, în raport cu distribuția multinomială de la punctul a, am „descompus“ probabilitatea $\frac{2+\psi}{4}$

în două probabilități, $\frac{1}{2}$ și $\frac{\psi}{4}$, ca și cum ele ar corespunde unor evenimente disjuncte.

mare față de valoarea obținută la punctul a ?

Indicație: Veți elabora formulele corespunzătoare pasului E și pasului M. Apoi veți face o implementare și veți rula algoritmul EM (pornind, de exemplu, cu valoarea inițială 0.5 pentru ψ) până când valorile acestui parametru până la cea de-a șasea zecimală nu se mai modifică.

c. Calculați — și apoi reprezentați grafic pe intervalul $(0, 1)$ — funcția de log-verosimilitate a datelor observabile

$$\ell_c(\psi) \stackrel{\text{def.}}{=} \ln \sum_{n_{11}+n_{12}=n_1} g_c(\underbrace{n_{11}, n_{12}}_{\text{neobs.}}, \underbrace{n_2, n_3, n_4}_{\text{obs.}}, \psi)$$

unde

$$\begin{aligned} g_c(n_{11}, n_{12}, n_2, n_3, n_4, \psi) &= \frac{n!}{n_{11}! n_{12}! n_2! n_3! n_4!} \left(\frac{1}{2}\right)^{n_{11}} \left(\frac{\psi}{4}\right)^{n_{12}} \left(\frac{1-\psi}{4}\right)^{n_2} \left(\frac{1-\psi}{4}\right)^{n_3} \left(\frac{\psi}{4}\right)^{n_4} \\ &= \frac{n!}{n_{11}! n_{12}! n_2! n_3! n_4!} \left(\frac{1}{2}\right)^{n_{11}} \left(\frac{\psi}{4}\right)^{n_{12}+n_4} \left(\frac{1-\psi}{4}\right)^{n_2+n_3} \end{aligned}$$

cu $n = n_{11} + n_{12} + n_2 + n_3 + n_4$.

d. Pentru a pune în evidență convergența algoritmului EM pe datele de mai sus, la iterațiile $t = 0, 1, 2$ ale algoritmului EM, la pasul M, calculați $\psi^{(t+1)}$ și trasați graficul funcției auxiliare

$$Q(\psi|\psi^{(t)}) = \ln \frac{n!}{\hat{n}_{11}! \hat{n}_{12}! n_2! n_3! n_4!} + \hat{n}_{11} \ln \frac{1}{2} + (\hat{n}_{12} + n_4) \ln \frac{\psi}{4} + (n_2 + n_3) \ln \frac{1-\psi}{4}.$$

B. Vom presupune acum că parametrul necunoscut ψ urmează o distribuție a priori, și anume distribuția Beta¹⁰⁴³

$$p(\psi) = \frac{1}{B(\nu_1, \nu_2)} \psi^{\nu_1-1} (1-\psi)^{\nu_2-1},$$

unde $B(\nu_1, \nu_2) \stackrel{\text{def.}}{=} \int_0^1 \psi^{\nu_1-1} (1-\psi)^{\nu_2-1} d\psi \stackrel{\text{calcul}}{=} \frac{\Gamma(\nu_1)\Gamma(\nu_2)}{\Gamma(\nu_1 + \nu_2)}$, iar Γ este funcția Gamma a lui Euler.¹⁰⁴⁴ Distribuția Beta este o conjugată naturală pentru distribuții ale datelor lipsă (engl., missing data), ca în cazul variabilei n_{12} , care urmează distribuția Binomial $\left(n_1, \frac{\psi/4}{1/2 + \psi/4}\right)$.¹⁰⁴⁵

e. Demonstrați că expresia log-verosimilității condiționale — ignorând anumiți termeni, care se însumează și nu depind de ψ — este următoarea:

$$\ln L(\psi) + \ln p(\psi) = (n_{12} + n_4 + \nu_1 - 1) \ln \psi + (n_2 + n_3 + \nu_2 - 1) \ln(1 - \psi).$$

f. Scrieți o explicație succintă pentru faptul că pasul E al algoritmului EM pentru această problemă de estimare în sens MAP coincide cu pasul E din

¹⁰⁴³Distribuția Beta este un caz particular al distribuției Dirichlet, care poate fi folosită ca distribuție a priori pentru parametrii distribuției categoriale. Vedeți ex. 128 și ex. 129.d de la capitolul de *Fundamente*.

¹⁰⁴⁴Vedeți ex. 31.b de la capitolul de *Fundamente*.

¹⁰⁴⁵La ex. 43 și ex. 129 de la capitolul de *Fundamente* aveți două exemple de folosire a distribuției Beta ca distribuție a priori pentru parametrul distribuției Bernoulli și respectiv al distribuției geometrice.

algoritmul EM standard. Așadar, el revine la a înlocui variabila n_{12} cu media sa condiționată, $n_1 \frac{\psi^{(t)}/4}{1/2 + \psi^{(t)}/4}$.

Apoi demonstrați că regula de actualizare de la pasul M este

$$\psi^{(t+1)} = \frac{n_{12}^{(t)} + n_4 + \nu_1 - 1}{n_{12}^{(t)} + n_2 + n_3 + n_4 + \nu_1 + \nu_2 - 2}.$$

Observație: Atunci când distribuția Beta este uniformă (adică, pentru $\nu_1 = \nu_2 = 1$), soluția MAP coincide cu soluția de verosimilitate maximă (engl., maximum likelihood) standard.

g. Adaptați implementarea pe care ați realizat-o la punctul b pentru a lua în considerare și distribuția a priori Beta asupra parametrului ψ .

32. (Algoritmul EM pentru „învățarea“ parametrilor unei distribuții multinomiale; aplicare pe datele unui „studiu“ epidemiologic)
- · University of Chicago, 2004 spring, Statistics, course, Michael Eichler, HW3, pr. 2

A. Datele din această problemă au fost preluate dintr-un studiu epidemiologic care a fost realizat în urma înregistrării mai multor cazuri de intoxicații alimentare care au apărut după un prânz organizat pentru personalul unei companii de asigurări. În total, 419 persoane au fost contactate pentru acest studiu epidemiologic. În următorul tabel apar două variabile, *Illness*, care indică dacă persoana respectivă a suferit de intoxicație alimentară, precum și *Fish*, care indică dacă persoana respectivă a mâncat pește la felul principal de la acel prânz:

		Fish		
Illness	Yes(1)	Yes(1)	No(0)	Unknown
		142	4	62
No(0)	104	54	53	

O parte din aceste date sunt necunoscute (engl., unknown / missing) fiindcă doar $m = 304$ persoane au răspuns la solicitările realizatorilor acest studiu. Totuși, folosind datele companiei a fost posibil să aflăm cine anume a avut intoxicație alimentară după acel prânz.

Pentru $k = 1, \dots, n = 419$, vom considera

$$X_k = \begin{cases} 1 & \text{dacă persoana } k \text{ a suferit de intoxicație alimentară,} \\ 0 & \text{în cazul contrar} \end{cases}$$

$$Y_k = \begin{cases} 1 & \text{dacă persoana } k \text{ a mâncat pește,} \\ 0 & \text{în cazul contrar} \end{cases}$$

și vom presupune că valorile variabilelor Y_{m+1}, \dots, Y_n sunt necunoscute. Mai departe, vom nota

$$N_{ij}^{(o)} = \sum_{k=1}^m 1_{\{X_k=i \text{ and } Y_k=j\}} \quad \text{și} \quad N_{ij}^{(m)} = \sum_{k=m+1}^n 1_{\{X_k=i \text{ and } Y_k=j\}},$$

pentru $i, j \in \{0, 1\}$.¹⁰⁴⁶

Datele observate sunt

$$N_{obs} = (N_{00}^{(o)}, N_{10}^{(o)}, N_{01}^{(o)}, N_{11}^{(o)}, N_{0+}^{(m)}, N_{1+}^{(m)}),$$

cu $N_{i+}^{(m)} = N_{i0}^{(m)} + N_{i1}^{(m)}$ pentru $i \in \{0, 1\}$.

Fie

$$N_{ij} = N_{ij}^{(o)} + N_{ij}^{(m)} \text{ pentru } i, j \in \{0, 1\}$$

datele complete. Presupunem că $N = (N_{ij})$ este distribuit multinomial, cu parametrul $\theta = (\theta_{ij})$,

$$N \sim Multinomial(n, \theta_{00}, \theta_{10}, \theta_{01}, \theta_{11}).$$

a. Formulați un algoritm EM pentru acest model. În acest scop, va trebui să specificați care anume sunt datele neobservabile și apoi să deduceți regulile de actualizare de la pasul E și respectiv pasul M.

b. Implementați algoritmul pe care l-ați conceput la punctul precedent (în Python, R, Matlab sau limbajul de programare pe care îl preferați) și aplicați-l pe datele de mai sus. Veți inițializa parametrii $\theta_{00}, \dots, \theta_{11}$ cu valorile estimărilor de verosimilitate maximă (MLE) care se obțin din primele două coloane ale tabelului din enunț, adică $\theta_{00}^{(0)} = \frac{54}{304}, \dots, \theta_{11}^{(0)} = \frac{142}{304}$.

B. Pentru a identifica posibilele cauze ale intoxicației alimentare, vrem să stim dacă îmbolnăvirile sunt datorate alegerii felului principal de mâncare. În primul rând, va trebui să antrenăm (engl., fit) modelul folosind „ipoteza nulă“ că X_k și Y_k sunt independente. Cu această presupozitie, parametrii vor satisface proprietatea

$$\theta_{ij} = (\theta_{i0} + \theta_{i1})(\theta_{0j} + \theta_{1j}).$$

Așadar, modelul va putea fi reparametrizat cu noii parametri

$$\alpha = \theta_{10} + \theta_{11} \quad \text{și} \quad \beta = \theta_{01} + \theta_{11}.$$

c. Folosiți o nouă variantă a algoritmului EM pentru a antrena (engl., fit) modelul folosind „ipoteza nulă“. Explicați de ce pasul E este identic cu cel [de la algoritmul EM] de mai sus și apoi formulați pasul M.

d. Implementați noul algoritm EM (în Python, R, Matlab sau limbajul de programare pe care îl preferați) și apoi aplicați-l de datele de mai sus. Ca și la punctul b, veți inițializa parametrii α și β cu valorile estimărilor de verosimilitate maximă (MLE) care se obțin din primele două coloane ale tabelului din enunț, adică $\alpha^{(0)} = \frac{142 + 4}{304}$ și $\beta^{(0)} = \frac{142 + 104}{304}$.

e. Verificați dacă „ipoteza nulă“ este adevărată. Puteți realiza aceasta folosind testul raportului de verosimilitate (engl., likelihood ratio test). Testați dacă cele două variabile sunt independente. (Sugestie: Folosiți statistică χ^2 .)

¹⁰⁴⁶Vă readucem aminte că notația $1_{\{\cdot\}}$ desemnează o *funcție-indicator*, a cărei valoare este 1 atunci când este înăplinită condiția scrisă între acolade, și 0 în cazul contrar.

8.2.4 Alte instanțe ale schemei algoritmice EM

33.

(Algoritmul EM: estimarea parametrilor unei mixturi de distribuții Poisson)

• *University of Chicago, 2004 spring, Statistics course, Michael Eichler, HW3, pr. 3*

Următorul tabel conține [într-o manieră compactă] numărul de decese ale femeilor în vîrstă de 80 de ani sau mai mult care au fost raportate în fiecare zi în cursul anilor 1910-1912.

Numărul de decese (k)	0	1	2	3	4	5	6	7	8	9
Frecvența observată (n_k)	162	267	271	185	111	61	27	8	3	1

- a. Dacă decesele sunt independente unele de altele și frecvența lor nu se schimbă în timp, atunci pentru a modela aceste date este natural să folosim distribuția Poisson. Antrenați (engl., fit) o distribuție Poisson pe aceste date. Analizând rezultatul pe care l-ați obținut, considerați că datele sunt într-adevăr distribuite conform distribuției Poisson?
- b. Este posibil ca vara și respectiv iarna să apară pattern-uri diferite cu privire la numărul de decese înregistrate, din care cauză o mixtură de două distribuții Poisson ar putea să constituie un model mai bun [pentru datele de mai sus] decât o singură distribuție Poisson.
Considerând că se folosește o mixtură de două distribuții Poisson, scrieți expresia verosimilității datelor observabile.
- c. Formulați un algoritm EM pentru acest model de mixtură. Pentru aceasta, va trebui să specificați care anume sunt datele complete și datele lipsă și apoi să deduceți regulile de actualizare de la pasul E și respectiv pasul M.
- d. Implementați în Python / R / Matlab algoritmul EM pe care l-ați conceput la punctul precedent și aplicați-l pe datele de mai sus. Cât de bine se potrivește acest model cu datele, comparativ cu distribuția Poisson de la punctul a? Care dintre cele două modele descrie mai bine datele?

34.

(Algoritmul EM: estimarea parametrilor unei mixturi de vectori de distribuții Poisson i.[i.]d.)

• *Univ. of Utah, 2008 fall, Hal Daumé III, HW9, pr. 1*

Am făcut cunoștință cu distribuția Poisson la problemele 27 și 46 de la capitolul de *Fundamente*. Vă reamintim că această distribuție este definită peste numerele naturale pozitive, și anume: dat fiind parametrul λ , funcția masă de probabilitate (p.m.f.) a distribuției Poisson este dată de expresia

$$p(x|\lambda) = \frac{1}{e^\lambda} \cdot \frac{\lambda^x}{x!}, \text{ pentru orice } x \in \mathbb{N}.$$

În problema 46 menționată mai sus, am văzut că, dată fiind o secvență de numere naturale pozitive x_1, \dots, x_n , estimarea de verosimilitate maximă (MLE)

pentru parametrul λ este $\frac{1}{n} \sum_{i=1}^n x_i$, adică exact media [aritmetică a] numerelor date.

În acest exercițiu vom considera o generalizare a acestei distribuții: modelul mixturii de [vectori de] distribuții Poisson. Nu știm dacă ați aflat [sau nu] până acum, însă acest model este folosit la monitorizarea serverelor de internet: numărul de cereri de acces care sunt adresate unui server de internet într-o unitate de timp (de exemplu, un minut) urmează de obicei o distribuție Poisson.

Considerăm că avem n servere de internet și că le monitorizăm pe fiecare pe o durată de M minute. Așadar, vom obține $n \times M$ numere pozitive (counturi); vom nota cu $x_{i,m}$ numărul de cereri adresate serverului i în minutul m . Scopul nostru este să *clusterizăm* serverele de internet în funcție de frecvența cererilor adresate lor [în timp].¹⁰⁴⁷

Presupunem că dorim să formăm K clustere. Definiți un *model* de tip *mixtura de [vectori de] distribuții Poisson* pentru această problemă. Veți nota cu z_i variabila latență care desemnează cărui cluster de servere de internet îi aparține serverul i (mai precis, indicând un număr de la 1 la K). De asemenea, veți nota cu λ_k parametrul distribuției Poisson care este urmată de fiecare componentă a vectorilor de variabile Poisson din clusterul k , iar cu π_k probabilitatea de selecție a respectivului cluster.

Elaborați algoritmul EM pentru rezolvarea acestui de tip *mixtura de [vectori de] distribuții Poisson*. Stabiliti pentru pasul E relațiile de calcul pentru mediiile variabilelor latente z_i , iar pentru pasul M relațiile de actualizare pentru parametrii λ și probabilitățile de selecție π pentru distribuțiile din acest model.

Indicație: Verosimilitatea datelor complete va arăta astfel:

$$\begin{aligned} L(\lambda, \pi) &\stackrel{\text{def.}}{=} P(\bar{x}, z | \lambda, \pi) \stackrel{i.i.d.}{=} \prod_{i=1}^n P(x_i, z_i | \lambda, \pi) = \prod_{i=1}^n P(x_i | z_i, \lambda, \pi) \cdot P(z_i | \lambda, \pi) \\ &= \prod_{i=1}^n \prod_{k=1}^K \left[\underbrace{\pi_k}_{\pi_k} \prod_{m=1}^M p(x_{i,m} | \lambda_k) \right]^{1_{\{z_i=k\}}}, \end{aligned}$$

unde

$$\begin{aligned} \bar{x} &\stackrel{\text{not.}}{=} (x_1, \dots, x_n), \text{ cu } x_i \stackrel{\text{not.}}{=} (x_{i,1}, \dots, x_{i,M}) \text{ pentru } i = 1, \dots, n, \\ z &\stackrel{\text{not.}}{=} (z_1, \dots, z_K), \pi \stackrel{\text{not.}}{=} (\pi_1, \dots, \pi_K), \lambda \stackrel{\text{not.}}{=} (\lambda_1, \dots, \lambda_K), \text{ iar} \\ 1_{\{z_i=k\}} &\text{ este funcția-indicator; ea ia valoarea 1 dacă } z_i = k \text{ și 0 în caz contrar.} \end{aligned}$$

¹⁰⁴⁷În cazul (limită!) în care $M = 1$, modelul nostru va deveni modelul unei mixturi de K distribuții de tip Poisson (nu *vectori* de distribuții de tip Poisson). Pentru cazul (și mai restrictiv!) $K = 2$, vedeți pr. 33.

35. (Algoritmul EM: pasul E pentru estimarea parametrilor unei mixturi de distribuții Gamma)

prelucrare de Liviu Ciortuz, după

■ □ • ○ CMU, 2012 fall, T. Mitchell, Z. Bar-Joseph, final exam, pr. 2

Considerăm instanțele $X_i \in \mathbb{R}^+$, ($i = 1, \dots, n$) produse de către următoarea mixtură de distribuții:

$$\begin{aligned} Z_i &\sim \text{Categorical}(\pi_1, \pi_2, \dots, \pi_K) \\ X_i &\sim \text{Gamma}(2, \beta_{Z_i}) \end{aligned}$$

Funcția de densitate a distribuției de probabilitate $\text{Gamma}(2, \beta)$ este definită astfel: $P(X = x) = \beta^2 x e^{-\beta x}$.

- Vom considera numărul de distribuții din mixtură $K = 3$ și parametrii (secunzi) ai acestor distribuții, $\beta_1 = 1, \beta_2 = 2, \beta_3 = 4$. Calculați $P(Z_i = 1 | X_i = 1)$.
- Elaborați pasul E al algoritmului EM pentru estimarea parametrilor acestei mixturi, scriind câte o formulă / relație matematică pentru fiecare expresie / cantitate care trebuie calculată la acest pas.
- Credeți că acest model de mixtură de distribuții Gamma poate opera cu clustere non-disjuncte (engl., overlapping), similar cu modelul mixturii de distribuții gaussiene?

36. (EM pentru învățare supervizată: cazul mixturilor de regresori liniari)

□ • ○ Stanford, 2007 fall, Andrew Ng, HW4, pr. 1

Algoritmul EM, aşa cum a fost prezentat la curs, se folosește pentru învățare nesupervizată. În particular, el servește pentru a reprezenta distribuția instanțelor de antrenament $p(x)$ prin marginalizare în raport cu o variabilă aleatoare latentă z :

$$p(x) = \sum_z p(x, z) = \sum_z p(x|z) p(z).$$

Însă algoritmul EM poate fi aplicat de asemenea și pentru învățare supervizată, iar în această problemă vom discuta *modelul unei mixturi de regresori liniari* (engl., mixture of linear regressors).¹⁰⁴⁸

Urmărind să reprezentăm distribuția condiționată $p(y|x)$, unde $x \in \mathbb{R}^n$ și $y \in \mathbb{R}$, vom introduce din nou o variabilă aleatoare latentă z discretă:

$$p(y|x) = \sum_z p(y, z|x) = \sum_z p(y|x, z) p(z|x).$$

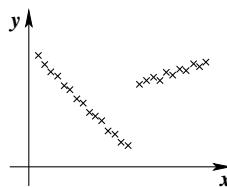
Pentru simplitate vom presupune că z are valori binare, $p(y|x, z)$ este o funcție densitate de probabilitate (p.d.f) gaussiană, iar distribuția $p(z|x)$ este produsă de către un model de regresie logistică. În mod formal, vom scrie:

$$\begin{aligned} p(z|x; \phi) &= g(\phi^\top x)^z (1 - g(\phi^\top x))^{1-z} \\ p(y|x, z = i; \theta_i) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y - \theta_i^\top x)^2}{2\sigma^2}\right), \quad i = 1, 2, \end{aligned}$$

¹⁰⁴⁸Mixtura de regresori liniari este un caz particular pentru ceea ce se numește *modelul mixturii ierarhice de experți* (engl., the Hierarchical Mixture of Experts model).

unde g este funcția logistică / sigmoidală, σ este un parametru cunoscut, iar ϕ, θ_0 și $\theta_1 \in \mathbb{R}^n$ sunt parametrii modelului.¹⁰⁴⁹

În mod intuitiv, procesul stochastic care corespunde acestui model poate fi imaginat după cum urmează. Date fiind o instanță x , folosind un model de regresie logistică, vom determina mai întâi dacă instanța respectivă aparține uneia sau celeilalte dintre cele două clase „ascunse“, $z = 0$ sau $z = 1$. Apoi, determinăm y ca o funcție liniară de x (avem funcții liniare distincte pentru diferitele valori ale lui z) și adăugăm zgomotul gaussian, exact ca în modelul standard de regresie liniară. De exemplu, următorul set de date poate fi reprezentat bine cu ajutorul unui astfel de model, însă nu poate fi reprezentat bine cu regresia liniară standard.



a. Presupunem că x , y și z sunt observabile, așa că obținem un set de date de antrenament $\{(x^{(1)}, y^{(1)}, z^{(1)}), \dots, (x^{(m)}, y^{(m)}, z^{(m)})\}$. Scrieți log-verosimilitatea acestor date în funcție de parametrii ϕ, θ_0 și θ_1 , iar apoi calculați estimările de verosimilitate maximă pentru acești parametri. Remarcați că datorită faptului că $p(z|x)$ este un model de regresie logistică, nu va exista o formulă analitică (engl., closed form formula) pentru estimarea lui ϕ . Așadar, calculați vectorul-gradient și matricea hessiană pentru funcția de verosimilitate a datelor în raport cu ϕ ; în practică, acestea pot fi folosite pentru a calcula numeric estimarea de verosimilitate maximă (MLE) a parametrilor.

b. Acum vom presupune că z este o variabilă aleatoare latentă (neobservabilă). Scrieți log-verosimilitatea datelor în funcție de parametri, iar apoi deduceți un algoritm EM care identifică maximul acestei log-verosimilități. Formulați în mod clar pasul E și pasul M. (Precizăm din nou că pasul M necesită o rezolvare numerică, așadar va trebui să calculați vectorii-gradient și matricele hessiene corespunzătoare.)

37.

(Algoritmul EM semi-supervizat:
cazul mixturilor de distribuții Bernoulli;
aplicare)

*formulare de Liviu Ciortuz, după
□ • Stanford, 2020 summer, Andrew Ng, HW3, pr. 4.bc
CMU, 2012 spring, Ziv Bar-Joseph, HW4, pr. 2*

La problema 22 am prezentat o formă generală a algoritmului EM semi-supervizat, precum și proprietățile sale de bază: forma regulilor de actualizare pentru pasul E și pasul M, precum și proprietatea de convergență. Aici vom reveni asupra modelului de mixturi de distribuții Bernoulli (engl., Bernoulli Mixture Model, BMM), și-i vom aplica algoritmul EM semi-supervizat.

¹⁰⁴⁹Aici folosim indicele $i \in \{1, 2\}$ pentru θ ca să desemnăm doi vectori de parametri distincți, nu pentru a indexa în acești vectori o anumită componentă.

Vom considera că datele sunt generate de către $K \in \mathbb{N}^*$ distribuții Bernoulli, având probabilitățile necunoscute $p_j \in [0, 1]$, cu $j \in \{1, \dots, K\}$. Avem n instanțe $x^{(i)} \in \{0, 1\}$, cu $i \in \{1, \dots, n\}$, iar fiecarei instanțe îi este asociată o variabilă latentă (ascunsă / necunoscută) $z^{(i)} \in \{1, \dots, K\}$, indicând ce distribuție a generat instanța $x^{(i)}$. În mod specific, $z^{(i)} \sim \text{Categorical}(\pi)$, cu $\sum_{j=1}^K \pi_j = 1$ și $\pi_j \geq 0$ pentru orice j , iar $(x^{(i)}|z^{(i)}) \sim \text{Bernoulli}(p_{z^{(i)}})$ sunt i.i.d. Așadar, p și π , unde $p \stackrel{\text{not.}}{=} (p_1, \dots, p_K)$ și $\pi \stackrel{\text{not.}}{=} (\pi_1, \dots, \pi_K)$, sunt parametrii modelului.

Vom mai considera \tilde{n} instanțe suplimentare, $\tilde{x}^{(i)} \in \mathbb{R}^d$ cu $i \in \{1, \dots, \tilde{n}\}$, precum și variabilele *observeate* asociate $\tilde{z}^{(i)} \in \{1, \dots, K\}$, fiecare $\tilde{z}^{(i)}$ indicând distribuția care a generat instanța $\tilde{x}^{(i)}$. Remarcați faptul că $\tilde{z}^{(i)}$ sunt constante cunoscute (în contrast cu $z^{(i)}$, care sunt variabile aleatoare necunoscute). Ca și mai înainte, vom presupune că $(\tilde{x}^{(i)}|\tilde{z}^{(i)}) \sim \text{Bernoulli}(p_{z^{(i)}})$ sunt i.i.d.

Rezumând, avem $n + \tilde{n}$ exemple, dintre care n sunt instanțe neetichetate $x^{(i)}$, asociate cu variabilele neobservabile $z^{(i)}$, iar \tilde{n} sunt instanțe etichetate $\tilde{x}^{(i)}$, având asociate în mod corespunzător etichetele observabile $\tilde{z}^{(i)}$. Algoritmul EM tradițional este conceput să ia ca input doar n exemple neetichetate și învață un model al cărui parametri sunt p și π .

Acum va trebui să aplicați algoritmul EM semi-supervizat la modelul de mixturi Bernoulli (BMM) pentru a folosi și cele \tilde{n} instanțe etichetate suplimentare, și să obțineți regulile de actualizare specifice pentru pașii E și M în varianta semi-supervizată.

Pentru *fixarea ideilor*, puteți începe să lucrați folosind *setul de date* din tabelul alăturat, care reprezintă rezultatul unui experiment probabilist în care sunt utilizate două monede, care modelează distribuții Bernoulli.¹⁰⁵⁰ (Remarcați faptul că în acest caz $K = 2$, ceea ce face problema semnificativ mai ușoară.)

Veți presupune că în coloana *moneda* sunt indicate valoările $z^{(i)}$ (respectiv $\tilde{z}^{(i)}$), și că $H = 1$ și $T = 0$.

<i>moneda</i>	$x^{(i)}/\tilde{x}^{(i)}$
?	<i>H</i>
?	<i>H</i>
?	<i>T</i>
?	<i>H</i>
?	<i>T</i>
1	<i>H</i>
2	<i>T</i>
2	<i>T</i>
2	<i>T</i>
2	<i>H</i>

a. [Pasul E, semi-supervizat]

Specificați în mod clar care sunt variabilele latente care trebuie să fie reestimate la pasul E. Deducreți formulele de actualizare de la pasul E pentru reestimarea tuturor variabilelor latente pe care le-ați menționat. Expresia finală de la pasul E trebuie să conțină doar x, z, p, π și constantă universale.

b. [Pasul M, semi-supervizat]

Specificați în mod clar care sunt parametrii care trebuie să fie recalculați la pasul M. Deducreți formulele de actualizare pentru toți parametrii pe care i-ați menționat. Mai precis, deducreți formule analitice (engl., closed form expressions) pentru actualizarea parametrilor $p_j^{(t+1)}$ și $\pi_j^{(t+1)}$, pornind de la funcția-obiectiv semi-supervizată.

c. [Aplicare]

Execuțați o iterație a algoritmului EM semi-supervizat pe datele din tabelul

¹⁰⁵⁰Datele supervizate sunt luate de la pr. 42 de la capitolul de *Fundamente*. Datele nesupervizate sunt luate de la CMU, 2020 fall, E. Xing, Z. Bar-Joseph, HW4, pr. 5.

de mai sus, luând $\alpha = 1$ și folosind ca valori inițiale pentru parametrii p_1, p_2, π_1 și π_2 estimările de verosimilitate maximă (MLE) care pot fi calculate din instanțele complet observabile (adică, etichetate).

38. (Algoritmul EM: chestiuni metodologice)

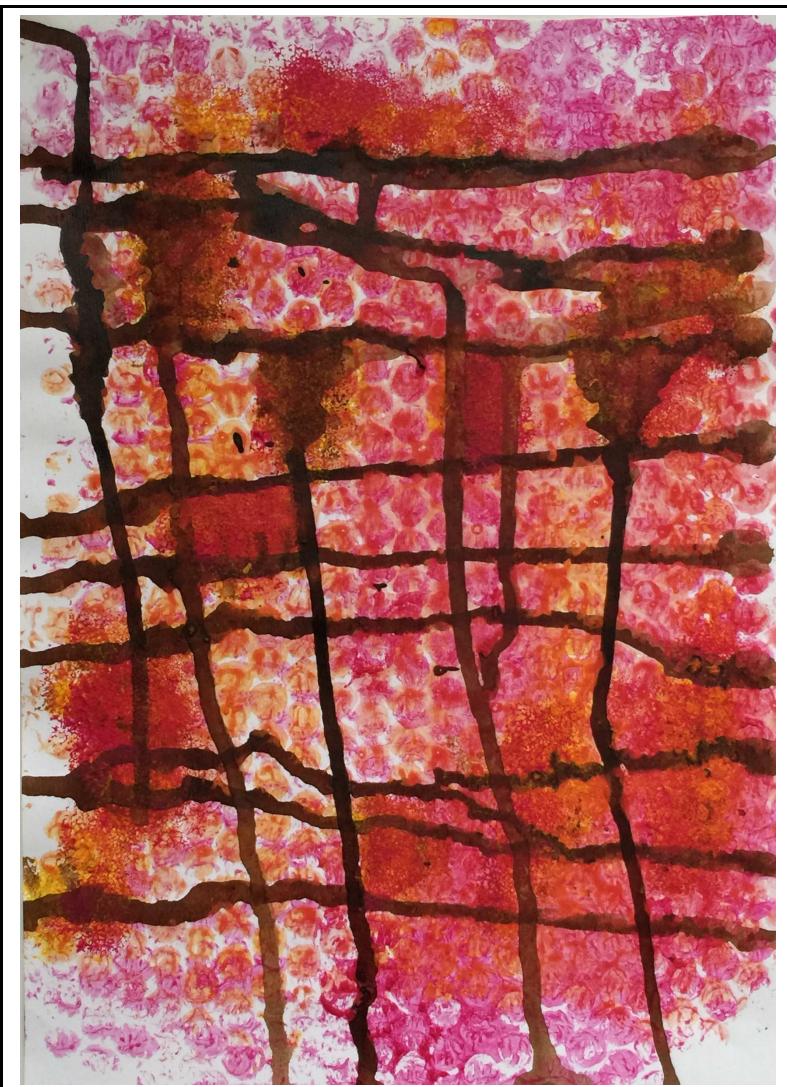
• ○ *CMU, 2014 spring, A. Singh, B. Poczos, HW3, pr. 2.3*

Algoritmul EM converge în general la un optim local. Legat de această chestiune, formulați în mod succint câteva strategii care ar putea fi folosite pentru a se obține totuși estimări acceptabil de bune ale parametrilor, atunci când se utilizează algoritmul EM în practică.

39. (Adevărat ori Fals?)

* *CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh, midterm exam., pr. 1.8.ab*

- a. Algoritmul EM optimizează o margine inferioară (engl., lower bound) a funcției sale obiectiv, $\ln \prod_i P(x_i | \theta)$, unde x_1, \dots, x_n sunt datele observate, iar θ sunt parametrii modelului asociat acestor date.
- b. Funcția obiectiv optimizată de algoritmul EM poate fi optimizată și cu metoda gradientului, care va găsi optimul global, în vreme ce EM găsește soluția sa mai rapid dar poate returna doar un optim local. Adevărat sau fals? Justificați.



© M. Romanică

9 Modele Markov ascunse

Sumar

- Noțiuni preliminare: programare dinamică, schema algoritmică EM.
- Verificarea înțelegerii unor noțiuni de bază în referitoare la modelele Markov ascunse (engl., Hidden Markov Models, HMM): ex. 1-4, ex. 15-17.
- Model Markov vizibil: exemplificare, legătura cu HMM: ex. 5.
- HMM ca model probabilist total: ex. 14.
- Algoritmul Forward:¹⁰⁵¹
exemple de aplicare: ex. 6, ex. 7, ex. 9.a;
demonstrarea formulei de la pasul inductiv: ex. 18.a;
calcularea probabilității de emitere a unei secvențe, folosind probabilitățile Forward: ex. 18.b.
- Algoritmul Backward:¹⁰⁵²
demonstrarea formulei de la pasul inductiv: ex. 8;
exemplu de aplicare: ex. 9.a.
- Algoritmul Viterbi:¹⁰⁵³
exemplu de aplicare: ex. 9.b;
determinarea căii celei mai probabile de generare a unei secvențe: ex. 9.c;
determinarea probabilității ca un anumit simbol [din secvența de semnale] să fi fost generat într-o anumită stare: ex. 9.d;
o variație pe tema algoritmului Viterbi: ex. 20.
- Algoritmul Forward-Backward / Baum-Welch / EM pentru HMM:¹⁰⁵⁴
exemplu de aplicare: ex. 11.b;
demonstrarea necesare pentru calcularea mediilor variabilelor neobservabile care corespund tranzițiilor: ex 10;
demonstrarea faptului că algoritmul Forward-Backward lasă neschimbate probabilitățile de tranziție sau de emisie care sunt nule [la inițializare]: ex 12.
- Demonstrarea unei formule alternative — în raport cu formulele bazate pe probabilitățile Forward și respectiv probabilitățile Backward — pentru calcularea probabilității de emitere a unei secvențe: ex. 19.
- HMM cu emisii gaussiene: ex. 13, ex. 22.

¹⁰⁵¹Pentru calculul probabilităților Forward, notează cu $\alpha_i(t) = P(O_1, \dots, O_t, X_{t+1} = S_i)$.

¹⁰⁵²Pentru calculul probabilităților Backward, notează cu $\beta_i(t) = P(O_t O_{t+1} \dots O_T | X_t = S_i)$.

¹⁰⁵³Pentru calculul cantităților $\delta_i(t) = \max_{X_1 \dots X_{t-1}} P(X_1 \dots X_{t-1}, O_1 \dots O_{t-1}, X_t = s_i)$.

¹⁰⁵⁴Pentru „învățarea“ parametrilor (i.e., a probabilităților de tranziție și respectiv de emisie) ai / ale unui HMM.

9.1 Modele Markov ascunse — Probleme rezolvate

1. (Verificarea înțelegерii unor noțiuni de bază)
CMU, 2001 fall, Andrew Moore, final, pr. 7

Fie un model Markov ascuns (engl., Hidden Markov Model, HMM) cu trei stări s_1, s_2, s_3 și trei simboli de emisie observabili X, Y, Z . Probabilitățile de start, de tranziție și respectiv de emisie sunt definite de următoarele matrice:

$$\begin{aligned} \pi &= \begin{bmatrix} \pi_1 = 1 \\ \pi_2 = 0 \\ \pi_3 = 0 \end{bmatrix} \\ A &= \begin{bmatrix} a_{11} = 1/2 & a_{12} = 1/2 & a_{13} = 0 \\ a_{21} = 0 & a_{22} = 1/2 & a_{23} = 1/2 \\ a_{31} = 0 & a_{32} = 0 & a_{33} = 1 \end{bmatrix} \\ B &= \begin{bmatrix} b_1(X) = 1/2 & b_1(Y) = 1/2 & b_1(Z) = 0 \\ b_2(X) = 1/2 & b_2(Y) = 0 & b_2(Z) = 1/2 \\ b_3(X) = 0 & b_3(Y) = 1/2 & b_3(Z) = 1/2 \end{bmatrix} \end{aligned}$$

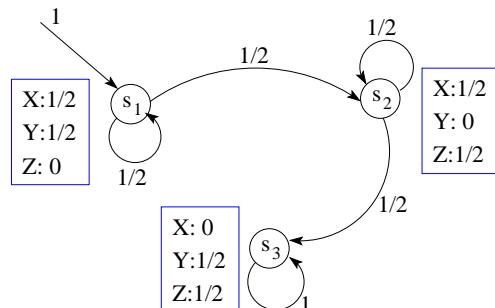
- Desenați automatul nedeterminist corespunzător acestui model Markov.
- Presupunând că acest model a emis secvența de semnale $XZXYYZYZZ$, identificați cea mai probabilă secvență de stări corespunzătoare, precum și probabilitatea acestei emisii. Se consideră că emisiile au loc la intrarea în stările respective.¹⁰⁵⁵ (Nu este necesară aplicarea algoritmului Viterbi.)

Răspuns:

- Automatul este cel din figura alăturată:
- Deoarece în acest caz particular, în fiecare stare a automatului este o probabilitate de emisie nulă, putem identifica relativ ușor din structura sa cea mai probabilă secvență de stări pentru a fi emisă secvența de semnale $XZXYYZYZZ$, și anume:
 $s_1s_2s_2s_3s_3s_3s_3s_3s_3$.

Probabilitatea acestei emisii se poate calcula astfel:

$$\begin{aligned} P(X) &= \pi_1 \cdot b_1(X) = \frac{1}{2} \\ P(XZ) &= P(X) \cdot a_{12}b_2(Z) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2^3} \\ P(XZX) &= P(XZ) \cdot a_{22}b_2(X) = \frac{1}{2^3} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2^5} \\ P(XZXY) &= P(XZX) \cdot a_{23}b_3(Y) = \frac{1}{2^7} \end{aligned}$$



¹⁰⁵⁵Atenție! La fel vom proceda și în toate problemele care urmează, dacă nu se specifică altfel, în mod explicit.

$$P(XZXYY) = P(XZXY) \cdot a_{33}b_3(Y) = \frac{1}{2^8}$$

$$P(XZXYYZ) = P(XZXYY) \cdot a_{33}b_3(Z) = \frac{1}{2^9}$$

$$P(XZXYYZY) = P(XZXYYZ) \cdot a_{33}b_3(Y) = \frac{1}{2^{10}}$$

$$P(XZXYYZY) = P(XZXYYZY) \cdot a_{33}b_3(Z) = \frac{1}{2^{11}}$$

$$P(XZXYYZYZZ) = P(XZXYYZY) \cdot a_{33}b_3(Z) = \frac{1}{2^{12}}$$

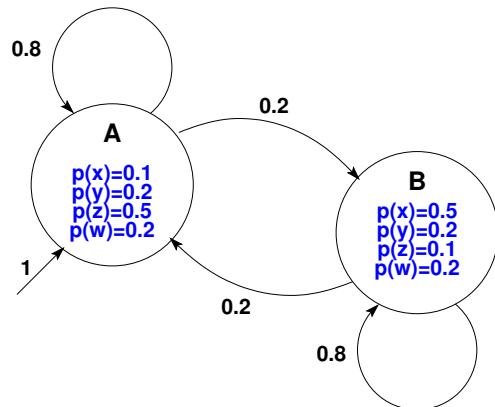
2.

(Verificarea înțelegерii unor noțiuni de bază
CMU, 2002 fall, Andrew Moore, final, pr. 3

Considerăm modelul Markov ascuns ilustrat în figura alăturată. Notație: q_t și O_t sunt starea și respectiv semnalul observat la momentul t .

Calculați:

- a. $P(q_2 = A)$
- b. $P(O_2 = x)$
- c. $P(q_2 = A | O_2 = x)$
- d. $P(O_{100} = w)$
- e. Care este cea mai probabilă secvență de stări dacă $O_1 = O_2 = O_3 = O_4 = O_5 = x$. (Nu este necesară aplicarea algoritmului Viterbi.)



Răspuns:

- a. Probabilitatea ca la momentul 2 automatul să fie în starea A este:

$$P(q_2 = A) = 1 \cdot 0.8 = 0.8$$

- b. Probabilitatea ca la momentul 2 automatul să emită simbolul x este:

$$\begin{aligned} P(O_2 = x) &= P(O_2 = x, q_2 = A) + P(O_2 = x, q_2 = B) \\ &= 0.8 \cdot 0.1 + 0.2 \cdot 0.5 = 0.08 + 0.1 = 0.18 \end{aligned}$$

- c. Probabilitatea ca automatul să fie în starea A la momentul 2, știind că al doilea simbol emis este x :

$$P(q_2 = A | O_2 = x) = \frac{P(q_2 = A, O_2 = x)}{P(O_2 = x)} = \frac{0.1 \cdot 0.8}{0.18} = \frac{0.08}{0.18} = \frac{4}{9}$$

- d. Probabilitatea cerută este:

$$\begin{aligned} P(O_{100} = w) &= P(q_{100} = A, O_{100} = w) + P(q_{100} = B, O_{100} = w) \\ &= 0.2 \cdot P(q_{100} = A) + 0.2 \cdot P(q_{100} = B) \\ &= 0.2 \cdot [P(q_{100} = A) + P(q_{100} = B)] = 0.2 \cdot 1 = 0.2 \end{aligned}$$

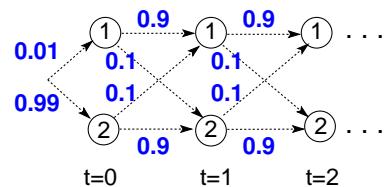
e. Cea mai probabilă secvență de stări este $ABBBB$, întrucât probabilitatea de a emite x este mai mare în starea B decât în starea A . Se intră inițial în starea A deoarece $\pi_A = 1$, dar apoi orice rămânere în starea A conduce la penalizare.

3.

(Verificarea înțelegerii unor noțiuni de bază)

CMU, 2003 fall, T. Mitchell, A. Moore, final, pr. 10

Considerăm un model Markov ascuns care modelează aruncarea unei monede imperfekte. Figura alăturată ilustrează tranzițiile împreună cu probabilitățile asociate lor în acest model.



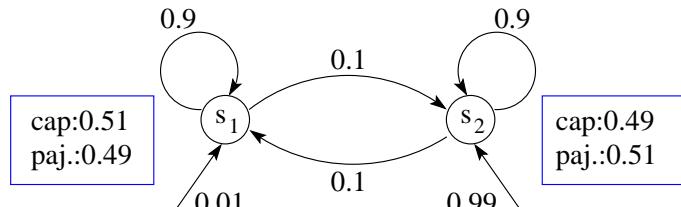
Probabilitatea ca în urma aruncării monedei să se obțină *cap* respectiv *pajură* este:

$$\begin{aligned} P(x = \text{cap} | s = 1) &= 0.51, \quad P(x = \text{pajură} | s = 1) = 0.49 \\ P(x = \text{cap} | s = 2) &= 0.49, \quad P(x = \text{pajură} | s = 2) = 0.51 \end{aligned}$$

- a. Presupunând că la trei aruncări succesive ale monedei s-a obținut de fiecare dată *cap*, care este cea mai probabilă secvență de stări? (Nu este necesar să se folosească algoritmul Viterbi.)
- b. Care este cea mai probabilă secvență de stări dacă la un număr foarte mare (de exemplu: 10^6) de aruncări consecutive ale monedei obținem un același rezultat: *cap*.

Răspuns:

Automatul este:



a. Cea mai probabilă secvență de stări este 2, 2, 2. Probabilitățile de a obține simbolul *cap* sunt aproape identice pentru cele două stări, este foarte probabil ca sistemul să înceapă cu starea 2 și să rămână aici. Se observă că se pierde mult din probabilitate dacă se trece în starea 1. (Emiterea acestei secvențe de simboluri în starea 2 se face cu probabilitatea $0.99 \cdot 0.49 \cdot 0.9 \cdot 0.49 \cdot 0.9 \cdot 0.49$.)

b. Când crește foarte mult numărul de aruncări și de obținere a rezultatului *cap*, crește și presiunea sistemului de a merge în starea 1, aceasta având un mic avantaj. La un moment dat se va face trecerea în starea 1, iar de aici nu există niciun avantaj de a merge înapoi în starea 2. Costul tranziției de la starea 2 la starea 1 este același indiferent de moment, însă realizarea tranziției mai devreme e mai favorabilă, întrucât crește probabilitatea obținerii unei serii lungi de *cap*. Așadar, cel mai bine este să se înceapă cu s_2 și să se continue imediat cu s_1 . Se obține astfel secvența de stări: 2, 1, 1, ..., 1 și probabilitatea $0.99 \cdot 0.49 \cdot 0.1 \cdot 0.51 \cdot 0.9 \cdot 0.51 \cdot 0.9 \dots 0.51 \cdot 0.9$.

4.

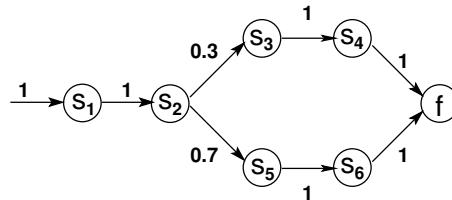
(Verificarea înțelegerei unor noțiuni de bază)

CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, final, pr. 5

Fie modelul Markov ascuns definit de matricea de tranzitii și emisii din tabelul următor:

	0	s_1	s_2	s_3	s_4	s_5	s_6	f	A	C	G	T
0	0	1	0	0	0	0	0	0				
s_1	0	0	1	0	0	0	0	0	0.5	0.3	0	0.2
s_2	0	0	0	0.3	0	0.7	0	0	0.1	0.1	0.2	0.6
s_3	0	0	0	0	1	0	0	0	0.2	0	0.1	0.7
s_4	0	0	0	0	0	0	0	1	0.1	0.3	0.4	0.2
s_5	0	0	0	0	0	0	1	0	0.1	0.3	0.3	0.3
s_6	0	0	0	0	0	0	0	1	0.2	0.3	0	0.5

Automatul este cel din figura următoare:



La fiecare din punctele de mai jos, puneți unul din semnele $<$, $,$, $=$ între cei doi termeni specificați.

- a. $P(O_1 = A, O_2 = C, O_3 = T, O_4 = A, X_1 = s_1, X_2 = s_2)$
 $P(O_1 = A, O_2 = C, O_3 = T, O_4 = A \mid X_1 = s_1, X_2 = s_2).$
- b. $P(O_1 = A, O_2 = C, O_3 = T, O_4 = A, X_3 = s_3, X_4 = s_4)$
 $P(O_1 = A, O_2 = C, O_3 = T, O_4 = A \mid X_3 = s_3, X_4 = s_4).$
- c. $P(O_1 = A, O_2 = C, O_3 = T, O_4 = A, X_3 = s_3, X_4 = s_4)$
 $P(O_1 = A, O_2 = C, O_3 = T, O_4 = A, X_3 = s_5, X_4 = s_6).$
- d. $P(O_1 = A, O_2 = C, O_3 = T, O_4 = A)$
 $P(O_1 = A, O_2 = C, O_3 = T, O_4 = A, X_3 = s_3, X_4 = s_4).$
- e. $P(O_1 = A, O_2 = C, O_3 = T, O_4 = A)$
 $P(O_1 = A, O_2 = C, O_3 = T, O_4 = A \mid X_3 = s_3, X_4 = s_4).$
- f. $P(O_1 = A, O_2 = C, O_3 = T, O_4 = A)$
 $P(O_1 = A, O_2 = T, O_3 = T, O_4 = G).$

Răspuns:

- a. $p_1 = P(O_1 = A, O_2 = C, O_3 = T, O_4 = A, X_1 = s_1, X_2 = s_2),$
 $p_2 = P(O_1 = A, O_2 = C, O_3 = T, O_4 = A \mid X_1 = s_1, X_2 = s_2).$

Se observă că $p_1 = p_2 \cdot P(X_1 = s_1, X_2 = s_2) = p_2 \cdot 1 \cdot 1$ ($\pi_1 = 1$ și $a_{s_1 s_2} = 1$). Deci rezultă $p_1 = p_2$.

- b. $p_1 = P(O_1 = A, O_2 = C, O_3 = T, O_4 = A, X_3 = s_3, X_4 = s_4),$
 $p_2 = P(O_1 = A, O_2 = C, O_3 = T, O_4 = A \mid X_3 = s_3, X_4 = s_4).$

Putem scrie că $p_1 = p_2 \cdot P(X_3 = s_3, X_4 = s_4) = p_2 \cdot 0.3 \cdot 1$. **Așadar,** $p_1 < p_2$.

$$\text{c. } p_1 = P(O_1 = A, O_2 = C, O_3 = T, O_4 = A, X_3 = s_3, X_4 = s_4), \\ p_2 = P(O_1 = A, O_2 = C, O_3 = T, O_4 = A, X_3 = s_5, X_4 = s_6).$$

În acest caz se vor calcula cele două probabilități:

$$\begin{aligned} p_1 &= P(O_1 = A, O_2 = C, O_3 = T, O_4 = A \mid X_3 = s_3, X_4 = s_4) \cdot P(X_3 = s_3, X_4 = s_4) \\ &= P(O_1 = A, O_2 = C) \cdot P(O_3 = T, O_4 = A, X_3 = s_3, X_4 = s_4) \\ &= 0.5 \cdot 0.1 \cdot 0.7 \cdot 0.1 \cdot 0.3 = 0.05 \cdot 0.021 \\ p_2 &= P(O_1 = A, O_2 = C, O_3 = T, O_4 = A \mid X_3 = s_5, X_4 = s_6) \cdot P(X_3 = s_5, X_4 = s_6) \\ &= P(O_1 = A, O_2 = C) \cdot P(O_3 = T, O_4 = A, X_3 = s_5, X_4 = s_6) \\ &= 0.5 \cdot 0.1 \cdot 0.3 \cdot 0.2 \cdot 0.7 = 0.05 \cdot 0.042 \end{aligned}$$

Prin urmare, $p_1 < p_2$.

$$\text{d. } p_1 = P(O_1 = A, O_2 = C, O_3 = T, O_4 = A), \\ p_2 = P(O_1 = A, O_2 = C, O_3 = T, O_4 = A, X_3 = s_3, X_4 = s_4).$$

Se observă că $p_1 = p_2 + P(O_1 = A, O_2 = C, O_3 = T, O_4 = A, X_3 = s_5, X_4 = s_6)$, iar cel de-al doilea termen al doilea al acestei sume este strict pozitiv (conform calculelor de mai sus). **În consecință,** $p_1 > p_2$.

$$\text{e. } p_1 = P(O_1 = A, O_2 = C, O_3 = T, O_4 = A), \\ p_2 = P(O_1 = A, O_2 = C, O_3 = T, O_4 = A \mid X_3 = s_3, X_4 = s_4).$$

$$\begin{aligned} p_1 &= P(O_1 = A, O_2 = C, O_3 = T, O_4 = A, X_3 = s_3, X_4 = s_4) + \\ &\quad + P(O_1 = A, O_2 = C, O_3 = T, O_4 = A, X_3 = s_5, X_4 = s_6) \\ &= p_2 \cdot P(X_3 = s_3, X_4 = s_4) + \\ &\quad + P(O_1 = A, O_2 = C, O_3 = T, O_4 = A \mid X_3 = s_5, X_4 = s_6) \cdot P(X_3 = s_5, X_4 = s_6) \end{aligned}$$

Rezultă că $p_1 < p_2$.

$$\text{f. } p_1 = P(O_1 = A, O_2 = C, O_3 = T, O_4 = A), \\ p_2 = P(O_1 = A, O_2 = T, O_3 = T, O_4 = G).$$

Se calculează imediat

$$p_1 = 0.5 \cdot 0.1 \cdot (0.3 \cdot 0.7 \cdot 0.1 + 0.7 \cdot 0.3 \cdot 0.2) = 0.5 \cdot 0.1 \cdot 0.7 \cdot (0.03 + 0.06) = 0.5 \cdot 0.1 \cdot 0.7 \cdot 0.09$$

$$\text{și } p_2 = 0.5 \cdot 0.6 \cdot (0.3 \cdot 0.7 \cdot 0.4 + 0.7 \cdot 0.3 \cdot 0) = 0.5 \cdot 0.6 \cdot 0.7 \cdot 0.3 \cdot 0.4.$$

$$p_1 < p_2 \Leftrightarrow 0.1 \cdot 0.09 < 0.6 \cdot 0.12 \text{ (A). În concluzie, } p_1 < p_2.$$

5.

(Model Markov vizibil: exemplificare; legătura cu un model Markov ascuns)

Heléène Touzet, Université de Lille 1, France

Se consideră secvența de nucleotide a unui organism imaginär *Bizarrus Examinus*, care are următoarele caracteristici:

- Nucleotida actuală este cu probabilitate de 25% din bazele A, C, G sau T, dacă precedentele două nucleotide sunt identice.

- Altfel (adică: în cazul în care precedentele două nucleotide sunt diferite), probabilitatea ca nucleotida actuală să fie C sau G este de două ori mai mare decat A sau T. Mai mult, între C și G (respectiv A și T), *purinele* (A și G) vor apărea mai frecvent decât *pirimidinile* (C sau T), și anume în 60% din cazuri.

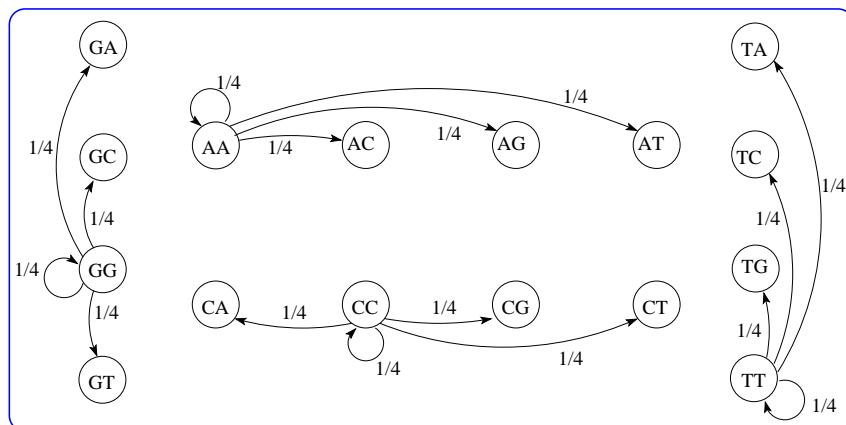
Arătați că secvența *Bizarrus Examinus* poate fi modelată folosind un model Markov vizibil; veți indica stările și probabilitățile de tranziție.

Răspuns:

Stările modelului Markov cerut sunt toate combinațiile posibile de câte 2 nucleotide, care înseamnă ultimele două nucleotide produse, fiind deci posibilă tranziția dintr-o stare în alta în cazul în care ultima nucleotidă a primei stări este egală cu prima nucleotidă din starea a două.

Observație: Fiecare stare a modelului Markov are 4 posibile tranziții. Cum există $4 \cdot 4 = 16$ stări, vor fi 72 de tranziții.

Să începem cu stările cu nucleotide identice. Din enunț obținem că probabilitățile de tranziție din aceste stări sunt $\frac{1}{4}$. Putem reprezenta aceste tranziții astfel:



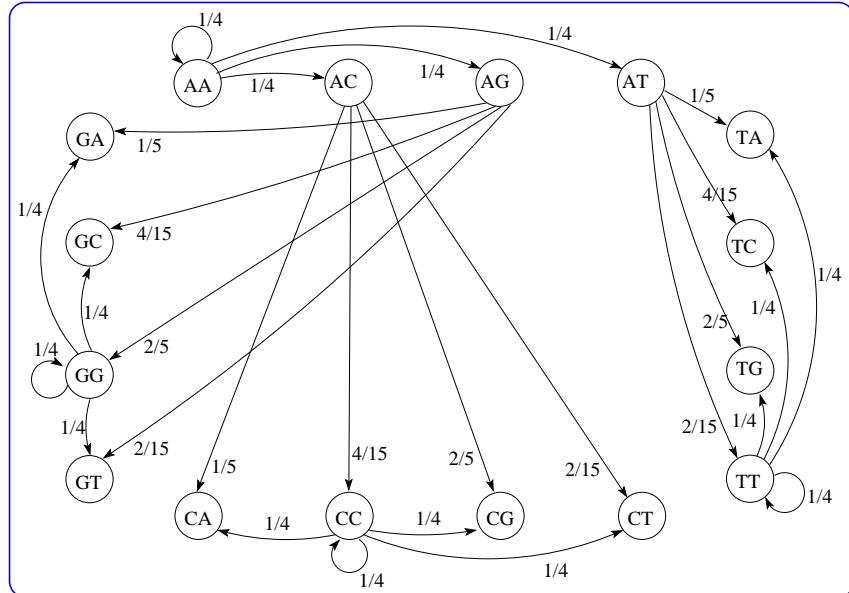
În continuare ne vom ocupa de tranzițiile din stările cu nucleotide diferite. Să notăm cu pq prima stare, unde $p \neq q$, și probabilitățile vor fi:

$$\begin{aligned} P(X_{i-1} = pq, X_i = qA) &= \frac{1}{3} \cdot \frac{6}{10} = \frac{1}{5} \\ P(X_{i-1} = pq, X_i = qC) &= \frac{2}{3} \cdot \frac{4}{10} = \frac{4}{15} \\ P(X_{i-1} = pq, X_i = qG) &= \frac{2}{3} \cdot \frac{6}{10} = \frac{2}{5} \\ P(X_{i-1} = pq, X_i = qT) &= \frac{1}{3} \cdot \frac{4}{10} = \frac{2}{15} \end{aligned}$$

Altfel scris, probabilitățile de tranziție dintr-o stare cu nucleotide diferite sunt în funcție de nucleotida actuală:

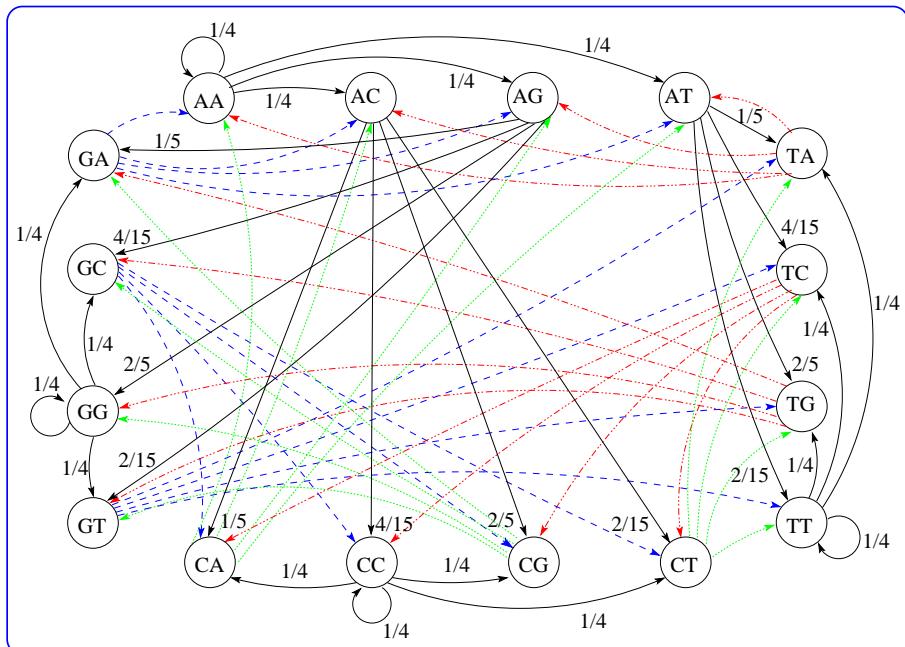
$$A : \frac{1}{5}, \quad C : \frac{4}{15}, \quad G : \frac{2}{5}, \quad T : \frac{2}{15}$$

Dacă trasăm tranzițiile din stările AC , AG și AT obținem:



În mod similar vor fi trasate și tranzițiile din celelalte stări (GA , GC , GT , CA , CG , CT , TA , TC și TG).

Rezultatul este un model Markov de forma:



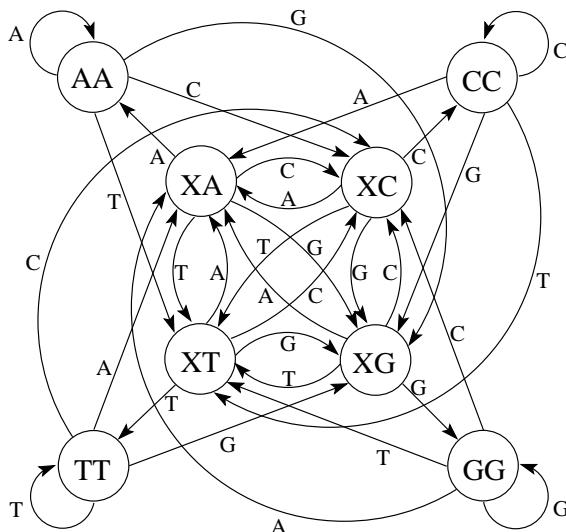
Simplificând, putem asocia acestui model Markov vizibil un model Markov ascuns cu

- 8 stări: AA , X_A (unde X semnifică literă diferită de A), CC , X_C (cu X diferit de C), GG , X_G (cu X diferit de G), TT , X_T (cu X diferit de T);

- emisii de căte un caracter, conform desenului de mai jos, având probabilitatea 1;
- tranziții având probabilitățiile (calculate mai sus):

AA, AA	$1/4$	GG, XA	$1/4$
AA, XC	$1/4$	GG, XC	$1/4$
AA, XG	$1/4$	GG, GG	$1/4$
AA, XT	$1/4$	GG, XT	$1/4$
CC, XA	$1/4$	TT, XA	$1/4$
CC, CC	$1/4$	TT, XC	$1/4$
CC, XG	$1/4$	TT, XG	$1/4$
CC, XT	$1/4$	TT, TT	$1/4$

XA, AA	$1/5$	XG, XA	$1/5$
XA, XC	$4/15$	XG, XC	$4/15$
XA, XG	$2/5$	XG, GG	$2/5$
XA, XT	$2/15$	XG, XT	$2/15$
XC, XA	$1/5$	XT, XA	$1/5$
XC, CC	$4/15$	XT, XC	$4/15$
XC, XG	$2/5$	XT, XG	$2/5$
XC, XT	$2/15$	XT, TT	$2/15$



6.

(Algoritmul Forward: aplicare)

CMU, 2006 fall, E. Xing, T. Mitchell, final exam, pr. 6

Fie un model Markov ascuns cu stările $X_t \in \{S_1, S_2, S_3\}$, observațiile $O_t \in \{A, B, C\}$ și parametrii

$\pi_1 = 1$	$a_{11} = 1/2$	$a_{12} = 1/4$	$a_{13} = 1/4$	$b_1(A) = 1/2$	$b_1(B) = 1/2$	$b_1(C) = 0$
$\pi_2 = 0$	$a_{21} = 0$	$a_{22} = 1/2$	$a_{23} = 1/2$	$b_2(A) = 1/2$	$b_2(B) = 0$	$b_2(C) = 1/2$
$\pi_3 = 0$	$a_{31} = 0$	$a_{32} = 0$	$a_{33} = 1$	$b_3(A) = 0$	$b_3(B) = 1/2$	$b_3(C) = 1/2$

a. Calculați $P(X_5 = S_3)$.

Pentru punctele b , c și d , vom presupune că observăm secvența $AABCABC$. Se consideră că emisia simbolului O_t are loc la ieșirea din starea X_t .

b. Calculați $P(X_5 = S_3 | O = AABCABC)$.

c. Completați următorul tabel, presupunând observația $O = AABCABC$. Notație: $\alpha_i(t) = P(O_1, \dots, O_t, X_{t+1} = S_i)$.

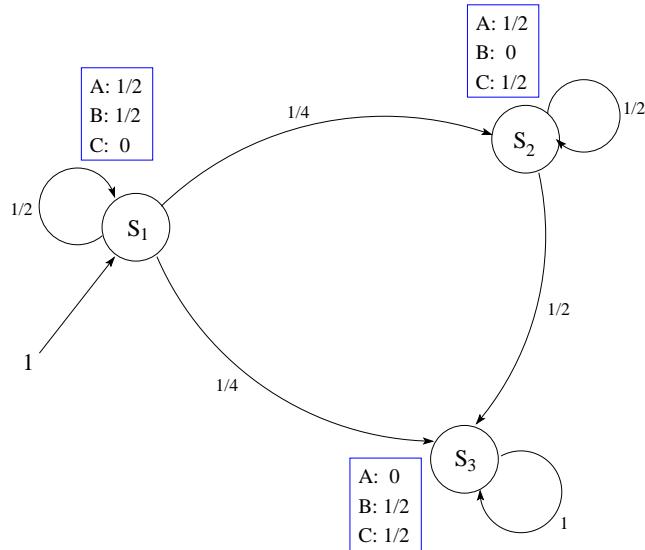
t	1	2	3	4	5	6	7	8
$\alpha_1(t)$								
$\alpha_2(t)$								
$\alpha_3(t)$								

d. Știind că a fost emisă secvența $O = AABCABC$, scrieți secvența de stări X care are probabilitate maximă a posteriori, $P(X|O)$.

Cât este această probabilitate a posteriori?

Răspuns:

Modelul Markov este cel din figura alăturată:



a. A calcula $P(X_5 = S_3)$ enumerând toate posibilitățile pentru X_1, X_2, X_3 și X_4 conduce la prea multe variante. Probabilitatea evenimentului contrar, $P(X_5 \neq S_3) = P(X_5 = S_1 \text{ sau } X_5 = S_2)$ este însă mult mai ușor de calculat. Așadar, vom folosi formula $P(X_5 = S_3) = 1 - P(X_5 = S_1) - P(X_5 = S_2)$.

Pentru ca la momentul 5 automatul să fie în starea 1, se observă că există o singură posibilitate, și anume ca automatul să intre în starea 1 și să rămână acolo, deci:

$$P(X_5 = S_1) = 1 \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{16}$$

Pentru $P(X_5 = S_2)$ avem patru posibilități, corespunzătoare celor patru momente în care se poate face trecerea din starea 1 în starea 2, deci:

$$\begin{aligned} P(X_5 = S_2) &= 4 \cdot 1 \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8} \\ P(X_5 = S_3) &= 1 - \frac{1}{16} - \frac{1}{8} = \frac{13}{16} \end{aligned}$$

b. Întrucât la momentul 5 s-a emis simbolul A , iar probabilitatea de emitere a lui A în starea 3 este 0, rezultă $P(X_5 = S_3 | O = AABCABC) = 0$.

c. Tabelul completat arată astfel:

t	1	2	3	4	5	6	7	8
O_t	A	A	B	C	A	B	C	
$\alpha_1(t)$	1	$\frac{1}{2^2}$	$\frac{1}{2^4}$	$\frac{1}{2^6}$	0	0	0	0
$\alpha_2(t)$	0	$\frac{1}{2^3}$	$\frac{1}{2^4}$	$\frac{1}{2^7}$	$\frac{1}{2^9}$	$\frac{1}{2^{11}}$	0	0
$\alpha_3(t)$	0	$\frac{1}{2^3}$	$\frac{1}{2^4}$	$\frac{5}{2^7}$	$\frac{11}{2^9}$	$\frac{1}{2^{11}}$	$\frac{1}{2^{12}}$	$\frac{1}{2^{13}}$

d. Secvența de stări este $X = S_1S_1S_1S_2S_2S_3S_3$ și are probabilitatea a posteriori $P(X | O) = 1$ datorită probabilităților specifice de emisie ale acestui model Markov.

7.

(Algoritmul Forward: aplicare)

CMU, 2008 spring, Eric Xing, HW4, pr. 3.1

Fie modelul Markov ascuns pentru cazinoul necinstit.¹⁰⁵⁶ Probabilitatea de a începe cu un zar corect / cinstit este 0.99. $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$ pentru un zar cinstit, iar $P(1) = P(2) = P(3) = P(4) = P(5) = 1/10$ și $P(6) = 1/2$ pentru un zar măsluit. Probabilitatea de rămânere în starea „cinstit“ este 0.95, iar cea de rămînere în starea „măsluit“ este 0.90. Se consideră că emisia unui simbol (oarecare) are loc la ieșirea dintr-o stare.

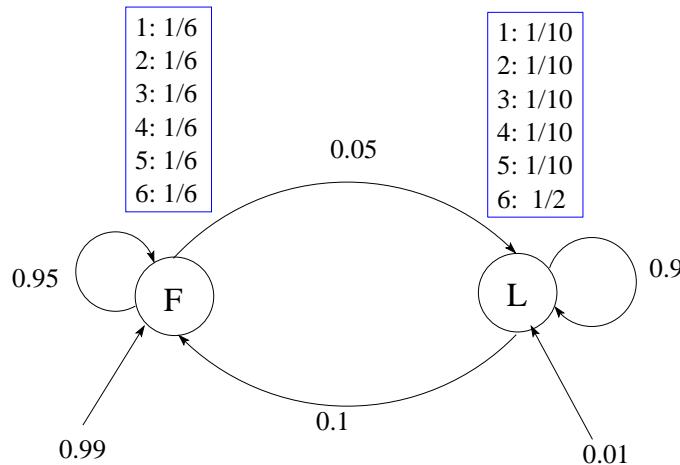
a. Calculați probabilitatea ca în patru aruncări de zar, primele trei să provină de la un zar cinstit, cu rezultatele 3, 1 și 5, iar cea de-a patra aruncare de zar să provină de la un zar măsluit, cu rezultatul 6, aşadar $P(O_1 = 3, O_2 = 1, O_3 = 5, O_4 = 6, X_1 = F, X_2 = F, X_3 = F, X_4 = L)$.

b. Calculați $P(O_1 = 3, O_2 = 1, O_3 = 5)$ folosind algoritmul Forward.

Răspuns:

Modelul Markov este următorul (se folosesc notațiile: F – Fair dice = zar cinstit, L – Loaded dice = zar măsluit):

¹⁰⁵⁶A se vedea *Biological sequence analysis*, R. Durbin et al, 1998, pag. 54-55, 56-57.



a. Probabilitatea cerută este:

$$\begin{aligned}
 P(O_1 = 3, O_2 = 1, O_3 = 5, O_4 = 6 | X_1 = F, X_2 = F, X_3 = F, X_4 = L) &= \\
 &= P(X_1 = F) \cdot P(O_1 = 3 | X_1 = F, X_2 = F) \cdot P(X_2 = F | X_1 = F) \cdot \\
 &\quad P(O_2 = 1 | X_2 = F, X_3 = F) \cdot P(X_3 = F | X_2 = F) \cdot P(O_3 = 5 | X_3 = F, X_4 = L) \cdot \\
 &\quad P(X_4 = L | X_3 = F) \cdot P(O_4 = 6 | X_4 = L) \\
 &= \pi_F \cdot b_{FF3} \cdot a_{FF} \cdot b_{FF5} \cdot a_{FF} \cdot b_{FL1} \cdot a_{FL} \cdot b_{LL6} \\
 &= 0.99 \cdot \frac{1}{6} \cdot 0.95 \cdot \frac{1}{6} \cdot 0.95 \cdot \frac{1}{6} \cdot 0.05 \cdot \frac{1}{2} = 0.000103411
 \end{aligned}$$

b. Probabilitățile Forward se calculează folosind un tabel corespunzător:

Output		3		1		5	
t	1		2		3		4
$\alpha_F(t)$	0.99		0.15685		0.024926083		0.003967936
$\alpha_L(t)$	0.01		0.00915		0.002130583		0.00039947
$P(O_1 \dots O_{t-1})$	1		0.166		0.027056666		0.004367406

Deci $P(O_1 = 3, O_2 = 1, O_3 = 5) = 0.004367406$.

8.

(Algoritmul Backward; demonstrație – pasul inductiv)

Liviu Ciortuz

a. Redăm mai jos demonstrația formulei de recurență de la pasul inductiv al algoritmului Backward pentru HMM. Pentru fiecare egalitate din cadrul demonstrației, precizați motivul/motivele pentru care are loc egalitatea respectivă. Acolo unde este cazul, detaliați.

$$\begin{aligned}
 \beta_i(t) &= P(O_t O_{t+1} \dots O_T | X_t = i, \mu) \\
 &= \sum_{j=1}^N P(O_t O_{t+1} \dots O_T | X_t = i, X_{t+1} = j, \mu) P(X_{t+1} = j | X_t = i, \mu) \\
 &= \sum_{j=1}^N P(O_{t+1} \dots O_T | O_t, X_t = i, X_{t+1} = j, \mu) P(O_t | X_t = i, X_{t+1} = j, \mu) a_{ij}
 \end{aligned}$$

$$\begin{aligned}
&= \sum_{j=1}^N P(O_{t+1} \dots O_T | X_{t+1} = j, \mu) b_{ijO_t} a_{ij} \\
&= \sum_{j=1}^N \beta_j(t+1) b_{ijO_t} a_{ij}
\end{aligned} \tag{426}$$

b. Vă reamintim că la pasul de initializare al algoritmului Backward se fixează $\beta_i(T+1) = 1$ pentru $i = 1, \dots, N$, unde N este numărul de stări din modelul Markov ascuns μ . Cum se poate calcula $P(O_1 O_2 \dots O_T | \mu)$, probabilitatea de emisie a unei secvențe de semnale $O_1 O_2 \dots O_T$ de către modelul μ , folosind probabilitățile Backward?

Răspuns:

a. Demonstrăm relația (1): Din definiție, $\beta_i(t)$ este probabilitatea de a emite secvența de simboluri de la O_t la O_T cu condiția ca la momentul de timp t să fie în starea s_i .

Demonstrăm că (2)=(1): Pentru probabilitățile din această egalitate se aplică definiția probabilității condiționate, obținându-se:

$$(2) = \frac{\sum_{j=1}^N P(O_t \dots O_T, X_t = i, X_{t+1} = j, \mu)}{P(X_t = i, X_{t+1} = j, \mu)} \cdot \frac{P(X_{t+1} = j, X_t = i, \mu)}{P(X_t = i, \mu)}$$

Simplificând această relație și apoi renunțând la sumă (folosind formula probabilității totale), se obține probabilitatea din relația (1):

$$(2) = \frac{P(O_t O_{t+1} \dots O_T, X_t = i | \mu)}{P(X_t = i | \mu)} = P(O_t O_{t+1} \dots O_T | X_t = i, \mu) = (1)$$

Demonstrăm egalitatea (3)=(2): Se aplică definiția probabilității de tranziție din starea i în starea j , $a_{ij} = P(X_{t+1} = j | X_t = i, \mu)$, apoi definiția probabilităților condiționate și se simplifică:

$$\begin{aligned}
(3) &= \sum_{j=1}^N P(O_{t+1} \dots O_T | O_t, X_t = i, X_{t+1} = j, \mu) P(O_t | X_t = i, X_{t+1} = j, \mu) a_{ij} \\
&= \sum_{j=1}^N \frac{P(O_t \dots O_T, X_t = i, X_{t+1} = j, \mu)}{P(O_t, X_t = i, X_{t+1} = j, \mu)} \cdot \frac{P(O_t, X_t = i, X_{t+1} = j, \mu)}{P(X_t = i, X_{t+1} = j, \mu)} \cdot \\
&\quad P(X_{t+1} = j | X_t = i, \mu) \\
&= \sum_{j=1}^N \frac{P(O_t \dots O_T, X_t = i, X_{t+1} = j, \mu)}{P(X_t = i, X_{t+1} = j, \mu)} \cdot P(X_{t+1} = j | X_t = i, \mu) \\
&= \sum_{j=1}^N P(O_t \dots O_T | X_t = i, X_{t+1} = j, \mu) P(X_{t+1} = j | X_t = i, \mu) = (2)
\end{aligned}$$

Demonstrăm egalitatea (4)=(3): b_{ijO_t} reprezintă probabilitatea de emitere a simbolului O_t la trecerea de la starea i la starea j , și deci are formula $b_{ijO_t} = P(O_t | X_t = i, X_{t+1} = j, \mu)$. Probabilitatea $P(O_{t+1} \dots O_T | O_t, X_t = i, X_{t+1} = j, \mu)$

aflată într-o sumă după j , nu depinde nici de O_t , nici de $X_t = i$, deci este egală cu $P(O_{t+1} \dots O_T | X_{t+1} = j, \mu)$. Egalitatea este astfel evidentă.

Egalitatea (4)=(5) este o chestiune de notație, $\beta_j(t+1)$ fiind probabilitatea de a emite secvența $O_{t+1} \dots O_T$ cu condiția ca la momentul $t+1$ să se afle în starea j .

b. Probabilitatea de emisie a unei secvențe de semnale $O = O_1 O_2 \dots O_T$ folosind probabilitățile Backward este:

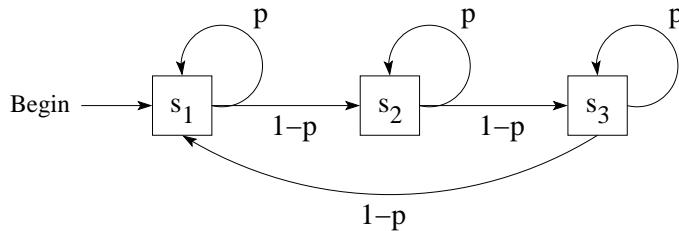
$$P(O | \mu) = \sum_{i=1}^N \pi_i \beta_i(1).$$

9.

(Algoritmii Forward, Backward, Viterbi: aplicare)

Marita Olsson, Univ. of Stockholm

Fie modelul Markov ascuns (HMM) de mai jos.



Fiecare stare/nod reprezintă o urnă care conține un număr de bile, dintre care unele sunt roșii (R) iar celelalte sunt albastre (B). O secvență observabilă x este generată de acest HMM în modul următor:

- se alege la întâmplare o bilă din prima urnă, după care se notează culoarea ei (R sau B), apoi bila este repusă în urnă;
- procedura de mai sus se repetă de mai multe ori, cu singura diferență că bila se alege cu probabilitatea p din aceeași urnă ca la pasul precedent și cu probabilitatea $1 - p$ din urna următoare.

În prima urnă sunt 3 bile roșii și 2 bile albastre, a doua urnă conține o bilă roșie și 4 bile albastre, iar în a treia urnă sunt 4 bile roșii și o bilă albastră. Se consideră $p = 1/3$.

a. Calculați $P(x = RBB)$.

În continuare, se consideră secvența $y = RRB BB$. Pentru a determina π^* , calea cea mai probabilă pentru generarea acestei secvențe, s-a rulat algoritmul Viterbi și s-au obținut rezultatele de mai jos:

O	R	R	B	B	B	
t	1	2	3	4	5	6
$\delta_1(t)$	1	$1/5$	$1/25$	$2/375$	$32/5625$	$128/84375$
$\delta_2(t)$	0	$2/5$	$2/25$	$8/375$	$32/5625$	$128/84375$
$\delta_3(t)$	0	0	$4/75$	$16/375$	$64/5625$	$256/84375$
$\psi_1(t)$	1	1	1	1	3	3
$\psi_2(t)$	–	1	1	2	2	1 sau 2
$\psi_3(t)$	–	–	2	2	2	2

b. Arătați cum a fost calculat $\delta_2(3)$.

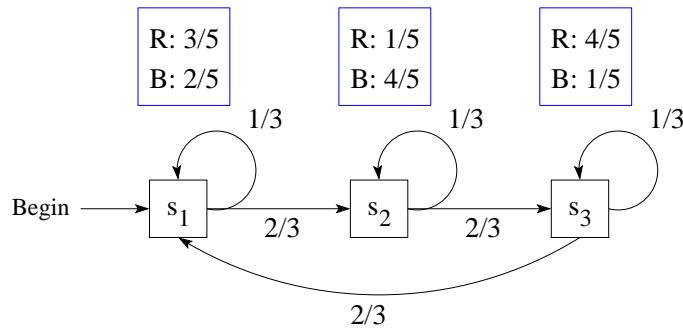
Observație: Am notat cu $\delta_t(i)$ probabilitatea calculată de către algoritmul Viterbi: cea mai mare dintre probabilitățile de generare a secvenței de semnale $O_1 \dots O_t$, cu ajungere (la finalul generării secvenței) în starea i .

c. Determinați π^* .

d. Presupunând că se dă secvența $O = O_1 O_2 \dots O_{100}$, precizați cum anume se poate calcula probabilitatea ca O_{77} să fie generat în starea s_3 .

Răspuns:

Modelul Markov ascuns (HMM) corespunzător datelor problemei este:



a. Probabilitatea cerută $P(x = RBB)$ se poate calcula în mai multe moduri:

Metoda 1 (Forță brută): Pentru producerea secvenței de semnale RBB sunt posibile 4 drumuri: s_1, s_1, s_1 ; s_1, s_1, s_2 ; s_1, s_2, s_2 ; sau s_1, s_2, s_3 . Se poate calcula probabilitatea pentru fiecare dintre acestea:

$$\begin{aligned}
 p_1 &= P(x = RBB, X = 111) = \frac{3}{5} \cdot \frac{1}{3} \cdot \frac{2}{5} \cdot \frac{1}{3} \cdot \frac{2}{5} = \frac{2^2}{5^3 \cdot 3} \\
 p_2 &= P(x = RBB, X = 112) = \frac{3}{5} \cdot \frac{1}{3} \cdot \frac{2}{5} \cdot \frac{2}{3} \cdot \frac{4}{5} = \frac{2^4}{5^3 \cdot 3} \\
 p_3 &= P(x = RBB, X = 122) = \frac{3}{5} \cdot \frac{2}{3} \cdot \frac{4}{5} \cdot \frac{1}{3} \cdot \frac{4}{5} = \frac{2^5}{5^3 \cdot 3} \\
 p_4 &= P(x = RBB, X = 123) = \frac{3}{5} \cdot \frac{2}{3} \cdot \frac{4}{5} \cdot \frac{2}{3} \cdot \frac{1}{5} = \frac{2^4}{5^3 \cdot 3}
 \end{aligned}$$

Probabilitatea secvenței de semnale RBB este suma acestor probabilități:

$$P(x = RBB) = \sum_{i=1}^4 p_i = \frac{4 + 16 + 32 + 16}{375} = \frac{68}{375}$$

Metoda 2 (Algoritmul Forward): Se calculează probabilitățile Forward $\alpha_i(t)$:¹⁰⁵⁷

¹⁰⁵⁷ $\alpha_i(t) \stackrel{\text{def.}}{=} P(O_1 \dots O_t, X_{t+1} = s_i)$, cu $\alpha_i(1) = \pi_i$, pentru $i = 1, 2, 3$, iar pentru $t \leq T$, $\alpha_i(t+1) = \sum_{j=1}^N \alpha_j(t) \cdot a_{ji} \cdot b_{ji} O_t$, unde N este numărul de stări, a_{ji} este probabilitatea de tranziție din starea s_i în starea s_j , iar $b_{ji} O_t$ este probabilitatea de emisie a simbolului O_t la trecerea din starea s_i în starea s_j .

x	R	B	B	B
t	1	2	3	4
$\alpha_1(t)$	1	1/5	2/75	36/1125
$\alpha_2(t)$	0	2/5	4/25	56/1125
$\alpha_3(t)$	0	0	16/75	112/1125
$P(x)$	1	3/5	2/5	68/375

$$\alpha_1(2) = 1 \cdot 1/3 \cdot 3/5 = 1/5$$

$$\alpha_2(2) = 1 \cdot 2/3 \cdot 3/5 = 2/5$$

$$\alpha_3(2) = 0$$

$$\alpha_1(3) = 1/5 \cdot 1/3 \cdot 2/5 = 2/75$$

$$\alpha_2(3) = 1/5 \cdot 2/3 \cdot 2/5 + 2/5 \cdot 1/3 \cdot 4/5 = 4/75 + 8/75 = 12/75 = 4/25$$

$$\alpha_3(3) = 2/5 \cdot 2/3 \cdot 4/5 = 16/75$$

$$\alpha_1(4) = 2/75 \cdot 1/3 \cdot 2/5 + 16/75 \cdot 2/3 \cdot 1/5 = \frac{36}{3^2 \cdot 5^3}$$

$$\alpha_2(4) = 2/75 \cdot 2/3 \cdot 2/5 + 4/25 \cdot 1/3 \cdot 4/5 = \frac{56}{3^2 \cdot 5^3}$$

$$\alpha_3(4) = 4/25 \cdot 2/3 \cdot 4/5 + 16/75 \cdot 1/3 \cdot 1/5 = \frac{112}{3^2 \cdot 5^3}$$

*Metoda 3 (Algoritmul Backward): Se calculează probabilitățile Backward $\beta_i(t)$:*¹⁰⁵⁸

x	R	B	B	B	t
					1
$\beta_1(t)$	68/375		4/15	2/5	1
$\beta_2(t)$	34/1125		8/25	4/5	1
$\beta_3(t)$	4/25		1/15	1/5	1
$P(x)$	68/375				

$$\beta_1(3) = 1 \cdot 1/3 \cdot 2/5 \cdot 1 + 1 \cdot 2/3 \cdot 2/5 = 2/5$$

$$\beta_2(3) = 1 \cdot 1/3 \cdot 4/5 + 1 \cdot 2/3 \cdot 4/5 = 4/5$$

$$\beta_3(3) = 1 \cdot 1/3 \cdot 1/5 + 1 \cdot 2/3 \cdot 1/5 = 1/5$$

$$\beta_1(2) = 2/5 \cdot 1/3 \cdot 2/5 + 4/5 \cdot 2/3 \cdot 2/5 = 4/15$$

$$\beta_2(2) = 4/5 \cdot 1/3 \cdot 4/5 + 1/5 \cdot 2/3 \cdot 4/5 = 8/25$$

$$\beta_3(2) = 2/5 \cdot 2/3 \cdot 1/5 + 1/5 \cdot 1/3 \cdot 1/5 = 1/15$$

$$\beta_1(1) = 4/15 \cdot 1/3 \cdot 3/5 + 8/25 \cdot 2/3 \cdot 3/5 = \frac{4}{3 \cdot 5^2} + \frac{16}{5^3} = \frac{68}{3 \cdot 5^3}$$

$$\beta_2(1) = 8/25 \cdot 1/3 \cdot 1/5 + 1/15 \cdot 2/3 \cdot 1/5 = \frac{8}{3 \cdot 5^3} + \frac{2}{3^2 \cdot 5^2} = \frac{34}{3^2 \cdot 5^3}$$

$$\beta_3(1) = 4/15 \cdot 2/3 \cdot 4/5 + 1/15 \cdot 1/3 \cdot 4/5 = \frac{32}{3^2 \cdot 5^2} + \frac{4}{3^2 \cdot 5^2} = \frac{4}{5^2}$$

¹⁰⁵⁸ $\beta_i(t) \stackrel{\text{def.}}{=} P(O_t O_{t+1} \dots O_T | X_t = s_i)$, cu $\beta_i(T+1) = 1$, pentru $i = 1, 2, 3$, iar pentru $t \leq T$, conform problemei 8, $\beta_i(t) = \sum_{j=1}^N \beta_j(t+1) \cdot a_{ji} \cdot b_{ji O_t}$.

b. $\delta_2(3)$ se calculează după formula:

$$\begin{aligned}\delta_2(3) &= \max(\delta_1(2) \cdot a_{12} \cdot b_1(R); \delta_2(2) \cdot a_{22} \cdot b_2(R); \delta_3(2) \cdot a_{32} \cdot b_3(R)) \\ \Rightarrow \delta_2(3) &= \max\left(\frac{1}{5} \cdot \frac{2}{3} \cdot \frac{3}{5}; \frac{2}{5} \cdot \frac{1}{3} \cdot \frac{1}{5}; 0\right) \Rightarrow \delta_2(3) = \max\left(\frac{2}{25}; \frac{2}{75}\right) \Rightarrow \delta_2(3) = \frac{2}{25}\end{aligned}$$

c. Pentru a determina π^* folosind algoritmul Viterbi se procedează astfel: Se alege maximul de pe ultima coloană (din primele 3 rânduri - adică dintre $\delta_i(T)$). Aceasta este obținut pe linia 3, deci $\delta_3(6)$. Așadar, ultima stare a drumului va fi s_3 . În continuare se trece în tabel la $\psi_3(6)$, unde este valoarea 2. Această valoare indică penultima stare ca fiind s_2 și de asemenea linia pe care se trece în continuare pentru a reconstitui calea cea mai probabilă de generare a secvenței respective. Se trece pe linia 2; de aici se continuă tot pe linia 2 de două ori, apoi se trece la linia 1. Se obține în final drumul: $s_1 s_1 s_2 s_2 s_2 s_3$.

O	R	R	B	B	B	
t	1	2	3	4	5	6
δ_1	1	1/5	1/25	2/375	32/5625	128/84375
δ_2	0	2/5	2/25	8/375	32/5625	128/84375
δ_3	0	0	4/75	16/375	64/5625	256/84375
Ψ_1	1	1	1	1	3	3
Ψ_2	-	1	1	2	2	1 sau 2
Ψ_3	-	-	2	2	2	2

d. Știm că $P(O, X_t = s_i) = \alpha_i(t) \cdot \beta_i(t)$, unde $\alpha_i(t)$ și $\beta_i(t)$ sunt probabilitățile Forward și respectiv Backward. Așadar:

$$P(X_{77} = s_3 | O) = \frac{P(X_{77} = s_3, O_1 \dots O_{100})}{P(O | \mu)} = \frac{\alpha_3(77) \cdot \beta_3(77)}{\sum_{i=1}^3 \alpha_i(101)}$$

10. (Demonstrația formulei necesare pentru calculul mediilor variabilelor neobservabile corespunzătoare tranzitiei în algoritmul Forward-Backward)

Liviu Ciortuz

Fie un model Markov ascuns având probabilitățile de tranzitie $a_{ij} = P(X_{t+1} = j | X_t = i, \mu)$ și probabilitățile de emisie $b_{ijO_t} = P(O_t | X_{t+1} = j, X_t = i, \mu)$.

Demonstrați formula:

$$P(X_t = i, X_{t+1} = j | O, \mu) = \frac{\alpha_i(t) a_{ij} b_{ijO_t} \beta_j(t+1)}{P(O | \mu)}$$

unde $\alpha_i(t) \stackrel{not.}{=} P(O_1 O_2 \dots O_{t-1}, X_t = i | \mu)$ și $\beta_i(t) \stackrel{not.}{=} P(O_t O_{t+1} \dots O_T | X_t = i, \mu)$.

Răspuns:

Se pornește de la membrul stâng al relației și se aplică formula probabilității condiționate de mai multe ori (fie în sens direct fie în sens invers):

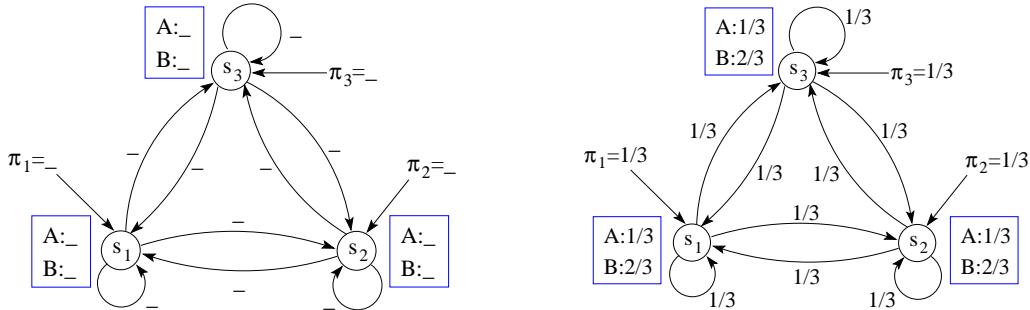
$$\begin{aligned}
 P(X_t = i, X_{t+1} = j | O, \mu) &= \frac{P(X_t = i, X_{t+1} = j, O | \mu)}{P(O | \mu)} \\
 &= \frac{1}{P(O, \mu)} \cdot P(O_1 O_2 \dots O_{t-1} O_t O_{t+1} \dots O_T, X_t = i, X_{t+1} = j) \\
 &= \frac{1}{P(O, \mu)} P(O_{t+1} \dots O_T | X_t = i, X_{t+1} = j, O_1 \dots O_t) P(X_t = i, X_{t+1} = j, O_1 \dots O_t) \\
 &= \frac{1}{P(O, \mu)} P(O_{t+1} \dots O_T | X_{t+1} = j) P(X_t = i, X_{t+1} = j, O_1 \dots O_t) \\
 &= \frac{1}{P(O, \mu)} \cdot \beta_j(t+1) \cdot P(X_t = i, X_{t+1} = j, O_1 \dots O_t) \\
 &= \frac{1}{P(O, \mu)} \cdot \beta_j(t+1) \cdot P(O_1 \dots O_{t-1} O_t, X_t = i, X_{t+1} = j) \\
 &= \frac{1}{P(O, \mu)} \cdot \beta_j(t+1) P(O_1 \dots O_{t-1} | O_t, X_t = i, X_{t+1} = j) P(O_t, X_t = i, X_{t+1} = j) \\
 &= \frac{1}{P(O, \mu)} \cdot \beta_j(t+1) \cdot P(O_1 \dots O_{t-1} | X_t = i) \cdot P(O_t, X_t = i, X_{t+1} = j) \\
 &= \frac{1}{P(O, \mu)} \cdot \beta_j(t+1) \cdot \frac{P(O_1 \dots O_{t-1}, X_t = i)}{P(X_t = i)} \cdot P(O_t, X_t = i, X_{t+1} = j) \\
 &= \frac{1}{P(O, \mu)} \cdot \beta_j(t+1) \cdot \alpha_i(t) \cdot \frac{P(O_t, X_t = i, X_{t+1} = j)}{P(X_t = i)} \\
 &= \frac{1}{P(O, \mu)} \cdot \beta_j(t+1) \cdot \alpha_i(t) \cdot \frac{P(O_t | X_t = i, X_{t+1} = j) \cdot P(X_t = i, X_{t+1} = j)}{P(X_t = i)} \\
 &= \frac{1}{P(O, \mu)} \cdot \beta_j(t+1) \cdot \alpha_i(t) \cdot \alpha_i(t) \cdot b_{ijO_t} \cdot \frac{P(X_t = i, X_{t+1} = j)}{P(X_t = i)} \\
 &= \frac{1}{P(O, \mu)} \cdot \beta_j(t+1) \cdot \alpha_i(t) \cdot b_{ijO_t} \cdot P(X_{t+1} = j | X_t = i) \\
 &= \frac{1}{P(O, \mu)} \cdot \beta_j(t+1) \cdot \alpha_i(t) \cdot b_{ijO_t} \cdot a_{ij} = \frac{\alpha_i(t) a_{ij} b_{ijO_t} \beta_j(t+1)}{P(O | \mu)}
 \end{aligned}$$

11. (HMM: noțiuni de bază; algoritmul Forward-Backward)

CMU, 2003 fall, T. Mitchell, A. Moore, final, pr. 10.cd

- a. Se consideră modelul Markov ascuns de mai jos, în partea stângă, cu trei stări s_1 , s_2 și s_3 . Probabilitățile de intrare în fiecare dintre stări sunt π_1 , π_2 , respectiv π_3 .

Stabiliți un set de probabilități astfel încât modelul Markov rezultat să maximizeze probabilitatea obținerii secvenței *ABA*.



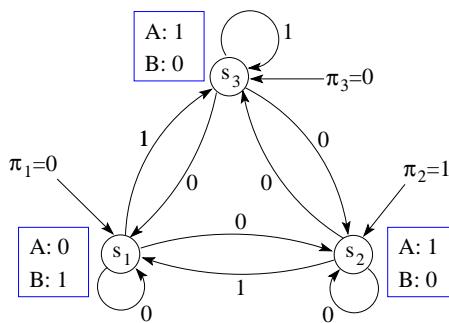
b. Se va folosi algoritmul EM (cunoscut pentru HMM și sub numele Forward-Backward sau Baum-Welch¹⁰⁵⁹) pentru estimarea parametrilor modelului Markov ascuns. Înaintea primei iterării s-au realizat inițializările prezentate în figura de mai sus din partea dreaptă. Pentru aceste valori initiale, va converge cu succes algoritmul EM la un model care maximizează probabilitatea secvenței ABA?

Răspuns:

a. Pentru ca modelul Markov ascuns obținut să fie valid, trebuie ca parametrii acestuia, adică probabilitățile, să respecte următoarele ecuații:

$$\begin{aligned}\pi_1 + \pi_2 + \pi_3 &= 1 \\ a_{i1} + a_{i2} + a_{i3} &= 1, \forall i \in \{1, 2, 3\} \\ b_i(A) + b_i(B) &= 1, \forall i \in \{1, 2, 3\}\end{aligned}$$

Sunt mai multe modele Markov ascunse pentru care probabilitatea obținerii secvenței ABA să fie 1, adică maximă. Unul dintre acestea este următorul:



Într-adevăr, în acest model, secvența ABA poate fi obținută pe următoarea cale: $q_1 = s_2$, $q_2 = s_1$, $q_3 = s_3$.

¹⁰⁵⁹Fizicianul Ruslan Stratonovich este cel dintâi care are prezentat algoritmul Forward-Backward, în anul 1960, în articolul *Conditional Markov Processes* din revista *Theory of Probability and Its Applications*, 5(2):156–178, în limba rusă. Procesele Markov poartă numele matematicianului rus Andrei Markov (1856-1922). Andrei Markov și fratele său Vladimir Markov (1871-1896) sunt cei care au demonstrat inegalitatea probabilistă cunoscută sub numele de inegalitatea [fraților] Markov; vedeti pr. 21.a de la capitolul de *Fundamente*. Vladimir Markov a murit de tuberculoză la vîrstă de doar 25 de ani.

Leonard Baum (1931 – 2017) și Lloyd Welch (1927-) au descris algoritmul Forward-Backward (și formalismul teoretic al modelelor Markov ascunse) în limba engleză, în cadrul unor cercetări întreprinse la sfârșitul anilor '60 și începutul anilor '70 din secolul trecut, la *Institute for Defence Analyses* din S.U.A. (Cf. Lawrence Rabiner, *First-Hand: The Hidden Markov Model*, https://ethw.org/First-Hand:The_Hidden_Markov_Model.)

b. Algoritmul EM (Forward-Backward) constă în repetarea iterativă a următorilor doi pași:

Estimare: Folosind probabilități Forward și Backward, $\alpha_i(t)$ și respectiv $\beta_i(t)$, se calculează $p_t(i, j) = \frac{\alpha_i(t) a_{ij} b_{ij}(O_t) \beta_j(t+1)}{\sum_{m=1}^N \alpha_m(t) \beta_m(t)}$ pentru $t \in \{1, 2, 3\}$ și $i, j \in \{1, 2, 3\}$.¹⁰⁶⁰

Maximizare: Se recalculează probabilitățile inițiale ale modelului

$$\hat{\pi}_i = \sum_{j=1}^N p_1(i, j) \quad \hat{a}_{ij} = \frac{\sum_{t=1}^T p_t(i, j)}{\sum_{l=1}^N \sum_{t=1}^T p_t(i, l)} \quad \hat{b}_{ijk} = \frac{\sum_{t: O_t=k, 1 \leq t \leq T} p_t(i, j)}{\sum_{t=1}^T p_t(i, j)}$$

Vom aplica prima iterație a algoritmului EM. Pentru aceasta trebuie calculate probabilitățile Forward și Backward. Datorită simetriei probabilităților celor trei stări s_1 , s_2 și s_3 în modelul considerat, vom avea $\alpha_1(t) = \alpha_2(t) = \alpha_3(t)$ și $\beta_1(t) = \beta_2(t) = \beta_3(t)$, $\forall t$ fixat. Aceste probabilități sunt:

Output		A		B		A	
t	1		2		3		4
$\alpha_i(t)$, $i \in \{1, 2, 3\}$	1/3		1/3 ²		2/3 ³		2/3 ⁴
$\beta_i(t)$, $i \in \{1, 2, 3\}$	2/3 ³		2/3 ²		1/3		1

Se observă că probabilitatea de a obține secvența ABA este foarte mică:

$$P(ABA) = \sum_i \alpha_i(4) = 3 \cdot \frac{2}{3^4} = \frac{2}{3^3} \approx 0.074$$

Pasul de estimare:

$$\begin{aligned} p_1(i, j) &= \frac{\alpha_i(1)a_{ij}b_{ij}(A)\beta_j(2)}{\sum_{m=1}^3 \alpha_m(1)\beta_m(1)} = \frac{\frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{2}{3^2}}{3 \left(\frac{1}{3} \cdot \frac{2}{3^3} \right)} = \frac{1}{3^2} \\ p_2(i, j) &= \frac{\alpha_i(2)a_{ij}b_{ij}(B)\beta_j(3)}{\sum_{m=1}^3 \alpha_m(2)\beta_m(2)} = \frac{\frac{1}{3^2} \cdot \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{1}{3}}{3 \left(\frac{1}{3^2} \cdot \frac{2}{3^2} \right)} = \frac{1}{3^2} \\ p_3(i, j) &= \frac{\alpha_i(3)a_{ij}b_{ij}(A)\beta_j(4)}{\sum_{m=1}^3 \alpha_m(3)\beta_m(3)} = \frac{\frac{2}{3^3} \cdot \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{3}}{3 \left(\frac{2}{3^3} \cdot \frac{1}{3} \right)} = \frac{1}{3^2} \end{aligned}$$

Pasul de maximizare:

$$\hat{\pi}_i = \sum_{j=1}^3 p_1(i, j) = 3 \cdot \frac{1}{3^2} = \frac{1}{3} \text{ pentru } i \in \{1, 2, 3\}$$

¹⁰⁶⁰ Numitorul $\sum_{m=1}^N \alpha_m(t) \beta_m(t)$ este probabilitatea generării secvenței de semnale $O_1 \dots O_T$. (Formula aceasta este una dintre posibilitățile de calculare a acestei probabilități.) Numărătorul $\alpha_i(t) a_{ij} b_{ij}(O_t) \beta_j(t+1)$ este probabilitatea ca la momentul t să se treacă din starea i în starea j , știind că a fost emisă secvența $O_1 \dots O_T$. Raportul acestor două probabilități, adică $p_t(i, j)$, poate fi interpretat ca fiind media unei variabile (aleatoare) indicator neobservabile, $X_t(i, j)$, care ia valoarea 1 dacă la momentul t se face trecerea din starea i în starea j (și se emite semnalul O_t), și 0 în caz contrar.

$$\begin{aligned}\hat{a}_{ij} &= \frac{\sum_{t=1}^3 p_t(i, j)}{\sum_{l=1}^3 \sum_{t=1}^3 p_t(i, l)} = \frac{3 \cdot \frac{1}{3^2}}{3 \cdot 3 \cdot \frac{1}{3^2}} = \frac{1}{3} \text{ pentru } i, j \in \{1, 2, 3\} \\ \hat{b}_{ij}(A) &= \frac{\sum_{t \in \{1, 3\}} p_t(i, j)}{\sum_{t=1}^3 p_t(i, j)} = \frac{2 \cdot \frac{1}{3^2}}{3 \cdot \frac{1}{3^2}} = \frac{2}{3} \text{ pentru } i, j \in \{1, 2, 3\} \\ \hat{b}_{ij}(B) &= \frac{\sum_{t \in \{2\}} p_t(i, j)}{\sum_{t=1}^3 p_t(i, j)} = \frac{1 \cdot \frac{1}{3^2}}{3 \cdot \frac{1}{3^2}} = \frac{1}{3} \text{ pentru } i, j \in \{1, 2, 3\}\end{aligned}$$

Se observă că nu se modifică probabilitățile de intrare π_i și nici probabilitățile de tranziție a_{ij} . De asemenea, deși se modifică probabilitățile de emisie, acestea își păstrează simetria: $b(A)$, și respectiv, $b(B)$ sunt egale pentru toate cele 3 stări. (De fapt, s-au inversat valorile precedente pentru $b(A)$ și $b(B)$.)

Probabilitățile Forward și Backward recalculate sunt:

Output		A		B		A	
t	1		2		3		4
$\alpha_i(t), i \in \{1, 2, 3\}$	1/3		2/3 ²		2/3 ³		4/3 ⁴
$\beta_i(t), i \in \{1, 2, 3\}$	4/3 ³		2/3 ²		2/3		1

iar probabilitatea de a obține secvența ABA s-a dublat:

$$P(ABA) = \sum_i \alpha_i(4) = 3 \cdot \frac{4}{3^4} = \frac{4}{3^3} \approx 0.148$$

Se poate constata ușor (prin calcule) că orice iterare ulterioară a algoritmului EM nu va aduce nicio modificare, toate probabilitățile rămânând neschimbate.

În concluzie, pentru valorile inițiale date în enunț, algoritmul EM nu converge la un model care maximizează probabilitatea secvenței ABA. (Vă reamintim că la punctul a am demonstrat că această probabilitate maximă este 1.)

12. (O proprietate a algoritmului Forward-Backward)

*Christopher Manning, Hinrich Schütze,
Foundations of statistical natural language processing, 1999*

Demonstrați că dacă într-un HMM o emisie/tranziție are probabilitatea 0, atunci algoritmul Forward-Backward o lasă neschimbată.

Răspuns:

Algoritmul Forward-Backward are ca date de intrare probabilitățile modelului Markov: π_i , a_{ij} și b_{ijk} . Folosind probabilități Forward și Backward, $\alpha_i(t)$ și respectiv $\beta_i(t)$, algoritmul recalculează probabilitățile inițiale, obținând $\hat{\pi}_i$, \hat{a}_{ij} și \hat{b}_{ijk} .

Arătăm că dacă $a_{ij} = 0$ atunci $\hat{a}_{ij} = 0$, iar dacă $b_{ijk} = 0$ atunci $\hat{b}_{ijk} = 0$.

Algoritmul Forward-Backward constă din mai multe iterări, fiecare iterare constând în doi pași:

- **Estimare:** Se calculează $p_t(i, j) = \frac{\alpha_i(t)a_{ij}b_{ijk}\beta_j(t+1)}{\sum_{m=1}^N \alpha_m(t)\beta_m(t)}$.

Dacă oricare dintre a_{ij} sau b_{ijk} (unde $O_t = k$) este 0, rezultă că $p_t(i, j) = 0$, indiferent de valoarea lui t .

- **Maximizare:**

$$\hat{a}_{ij} = \frac{\sum_{t=1}^T p_t(i, j)}{\sum_{l=1}^N \sum_{t=1}^T p_t(i, l)}$$

Dacă $a_{ij} = 0$, atunci $p_t(i, j) = 0$, deci $\hat{a}_{ij} = 0$.

$$\hat{b}_{ijk} = \frac{\sum_{t:O_t=k, 1 \leq t \leq T} p_t(i, j)}{\sum_{t=1}^T p_t(i, j)}$$

Dacă $b_{ijk} = 0$, atunci $p_t(i, j) = 0$ pentru orice t pentru care $O_t = k$, deci $\hat{b}_{ijk} = 0$.

Am demonstrat astfel că algoritmul Forward-Backward lasă aceste probabilități neschimbate, dacă ele sunt inițial nule.

13.

(Un model Markov ascuns cu emisii gaussiene)

• CMU, 2008 fall, Eric Xing, final, pr. 8

Fie un model Markov ascuns cu emisii continue, având K stări. Notăm cu x_i emisia de la momentul i , iar cu z_i starea ascunsă corespunzătoare. Emisiile stării $k \in \{1, \dots, K\}$ au o distribuție gaussiană de medie μ_k și deviație standard σ_k .¹⁰⁶¹ Prin urmare, probabilitățile de emisie sunt $P(x_i | z_i = k, \theta) = \mathcal{N}(x_i | \mu_k, \sigma_k)$. θ reprezintă multimea parametrilor modelului, adică probabilitățile inițiale π , matricea probabilităților de tranziție a , mediile μ_1, \dots, μ_K și deviațiile standard $\sigma_1, \dots, \sigma_K$.

- Calculați log-verosimilitatea secvenței de emisii (x_1, \dots, x_n) dacă succesiunea de stări în care se produc aceste emisii este (z_1, \dots, z_n) .
- Calculați formulele de update pentru probabilitățile Forward și Backward pentru acest HMM și explicați pe scurt diferența față de cele studiate în cazul discret.
- Presupunem că ni se dau secvența de observații $X = (x_1, \dots, x_n)$ și secvența de stări corespunzătoare, $Z = (z_1, \dots, z_n)$. Vrem să găsim valorile parametrilor θ pentru acest model Markov ascuns, folosind un algoritm de tipul EM / Forward-Backward.

Sunt oare relațiile de actualizare pentru probabilitățile a_{ij} și π_i diferite de cele ale modelului Markov ascuns cu emisii discrete? Explicați de ce da, sau de ce nu.

Care vor fi relațiile de actualizare pentru parametrii μ_k și σ_k ai distribuțiilor gaussiene care modelează emisiile? Sugestie: Nu este necesar să derivați aceste relații. Date fiind stările, emisiile sunt independente unele de altele, iar fiecare emisie urmează distribuția gaussiană asociată stării în care se produce emisia respectivă.

¹⁰⁶¹Se va considera că aceste emisii depind doar de starea k , nu și de stările în care se poate trece pornind din starea k .

Pentru punctele d și e de mai jos, vom presupune că ni se dă doar secvența de observații $X = (x_1, \dots, x_n)$. Vrem să găsim valorile parametrilor θ pentru acest model Markov ascuns.

d. Algoritmul de învățare nesupervizată EM / Forward-Backward optimizează la fiecare iterație media log-verosimilității datelor complete. Oare de ce este aceasta o funcție obiectiv rezonabilă?

Scrieți expresia acestei funcții (notată cu $E[\ell_c(x, y; \theta)]$) pentru modelul Markov ascuns cu emisii gaussiene. Nu este necesar să demonstrați cum anume se obține această expresie.

e. Presupunem că vrem să găsim estimările de verosimilitate maximă ($\hat{\mu}_k$ și $\hat{\sigma}_k$) pentru parametrii μ_k și σ_k . Vor avea oare formulele [pentru obținerea] acestor estimări exact aceeași formă ca și în cazul mixturilor de distribuții gaussiene? Justificați pe scurt. Sugestie: Identificați acei termeni din expresia funcției $E[\ell_c(x, y; \theta)]$ care sunt relevanți pentru optimizare (adică, termenii care conțin parametrii μ_k și / sau σ_k).

Răspuns:

a. Log-verosimilitatea condițională cerută este:

$$\begin{aligned} \ln P(x_1, \dots, x_n \mid z_1, \dots, z_n) &= \ln \prod_{i=1}^n P(x_i \mid z_i) = \ln \prod_{i=1}^n \mathcal{N}(x_i \mid \mu_{z_i}, \sigma_{z_i}) \\ &= \sum_{i=1}^n \ln \mathcal{N}(x_i \mid \mu_{z_i}, \sigma_{z_i}) = \sum_{i=1}^n \left[-\ln(\sqrt{2\pi}\sigma_{z_i}) - \frac{(x_i - \mu_{z_i})^2}{2\sigma_{z_i}^2} \right] \end{aligned}$$

b. Probabilitățile Forward se calculează astfel:

$$\begin{aligned} \alpha_k(t+1) &\stackrel{\text{def.}}{=} P(x_1, \dots, x_t, z_{t+1} = k) \\ &= \sum_i P(x_1, \dots, x_t, z_{t+1} = k, z_t = i) \\ &= \sum_i P(x_1, \dots, x_{t-1}, z_t = i) \cdot P(x_t, z_{t+1} = k \mid x_1, x_2, \dots, x_{t-1}, z_t = i) \\ &= \sum_i \alpha_i(t) \cdot P(x_t, z_{t+1} = k \mid z_t = i) \\ &= \sum_i \alpha_i(t) \cdot P(x_t \mid z_{t+1} = k, z_t = i) \cdot P(z_{t+1} = k \mid z_t = i) \\ &= \sum_i \alpha_i(t) \cdot \mathcal{N}(x_t \mid \mu_k, \sigma_k) \cdot a_{ik} = \mathcal{N}(x_t \mid \mu_k, \sigma_k) \sum_i \alpha_i(t) \cdot a_{ik} \end{aligned}$$

Similar, probabilitățile Backward se calculează astfel:

$$\begin{aligned} \beta_k(t) &\stackrel{\text{def.}}{=} P(x_t, \dots, x_T \mid z_t = k) \\ &= \sum_i P(x_t, \dots, x_T, z_{t+1} = i \mid z_t = k) \\ &= \sum_i P(x_t, \dots, x_T \mid z_{t+1} = i, z_t = k) \cdot \underbrace{P(z_{t+1} = i \mid z_t = k)}_{=a_{ki}} \\ &= \sum_i P(x_{t+1}, \dots, x_T \mid x_t, z_{t+1} = i, z_t = k) \cdot P(x_t \mid z_{t+1} = i, z_t = k) \cdot a_{ki} \end{aligned}$$

$$\begin{aligned}
&= \sum_i P(x_{t+1}, \dots, x_T \mid z_{t+1} = i) \cdot \mathcal{N}(x_t \mid \mu_i, \sigma_i) \cdot a_{ki} \\
&= \sum_i \beta_i(t+1) \cdot \mathcal{N}(x_t \mid \mu_i, \sigma_i) \cdot a_{ki}
\end{aligned}$$

Așadar, formulele de recurență pentru un model Markov ascuns cu emisii continue au o formă similară cu cele din cazul discret,¹⁰⁶² singura diferență constând în probabilitățile de emisie, care în cazul continuu sunt modelate de distribuții probabiliste gaussiene.

c. Relațiile de actualizare pentru probabilitățile a_{ij} și π_i sunt aceleași cu cele din cazul emisiilor discrete, fiindcă ele implică doar numărarea tranzițiilor dintre stări și sunt independente de forma aleasă pentru probabilitățile de emisie.

Relațiile de actualizare pentru parametrii μ_k și σ_k ai distribuțiilor gaussiene care modelează emisiile sunt următoarele:

$$\begin{aligned}
\mu_k &= \frac{\sum_i 1_{\{z_i=k\}} x_i}{\sum_i 1_{\{z_i=k\}}} \\
\sigma_k &= \frac{\sum_i 1_{\{z_i=k\}} (x_i - \mu_k)^2}{\sum_i 1_{\{z_i=k\}}}
\end{aligned}$$

pentru $k = 1, \dots, K$.¹⁰⁶³ (Vă reamintim că notația $1_{\{\cdot\}}$ desemnează funcția indicator.)

d. Media log-verosimilității datelor complete (X și Z) este o margine inferioară pentru funcția de log-verosimilitate a datelor observabile.¹⁰⁶⁴ Algoritmul EM / Forward-Backward converge la un optim local al funcției de log-verosimilitate a datelor observabile, așadar este rezonabil să folosim ca funcție obiectiv [la fiecare iterație a acestui algoritm] media log-verosimilității datelor complete.

$$E[\ell_c(\theta; x, y)] = \sum_n E[y_{n,1}^i] \pi_i + \sum_n \sum_{t=2}^T E[y_{n,t-1}^i y_{n,t}^j] \ln a_{ij} + \sum_n \sum_{t=1}^T E[y_{n,t}^i] \ln \mathcal{N}(x_n \mid \mu_i, \sigma_i),$$

unde n indexează stările modelului Markov ascuns,¹⁰⁶⁵ iar $y_{n,t}^i$ este o variabilă aleatoare care ia valoarea 1 dacă la momentul de timp t modelul se află în stareea cu indicele i și respectiv valoarea 0 în cazul contrar.¹⁰⁶⁶

e. Da, formulele pentru estimările de verosimilitate maximă pentru parametrii μ_k și σ_k au aceeași formă în ambele cazuri. În expresia funcției $E[\ell_c(\theta; x, y)]$ termenul relevant este (doar) ultimul termen, care seamănă foarte mult cu cel din expresia mediei log-verosimilității datelor complete de la mixtura de distribuții gaussiene.¹⁰⁶⁷

¹⁰⁶²Pentru formulele de calcul ale probabilităților Forward și respectiv Backward în cazul emisiilor discrete, vedeți problemele 18.a și 8.a, relațiile (427) și (426).

¹⁰⁶³Vedeți problemele 50.a și respectiv 51.a de la capitolul de *Fundamente*.

¹⁰⁶⁴Vedeți problema 1 de la capitolul *Schema algoritmică EM*.

¹⁰⁶⁵Nu are nicio legătură cu indicele n din notația secvenței $X = (x_1, \dots, x_n)$.

¹⁰⁶⁶Pentru fiecare valoare fixată pentru $t \in \{1, \dots, T\}$, numai una dintre variabilele $y_{n,t}^i$ ia valoare diferită de 0.

¹⁰⁶⁷Vedeți problema 18 de la capitolul *Clusterizare*.

9.2 Modele Markov ascunse — Probleme propuse

14. (Modelul Markov – model probabilistic total)

*prelucrare de Liviu Ciortuz, după
Problems and solutions in biological sequence analysis,
M. Borodovsky, S. Ekinsheva, 2006, pr. 3.1-3.3*

Fie μ un model Markov de ordinul întâi.

a. Suma probabilităților tuturor secvențelor de stări de lungime L se poate scrie ca

$$\sum_x P(x) = \sum_{x_1} \sum_{x_2} \dots \sum_{x_L} P(x_1) \prod_{i=2}^L a_{x_{i-1} x_i}$$

Arătați că această sumă este 1.

b. Presupunem că acest model are o stare finală ‘silent’ notată *End*, și că tranzițiile din orice stare în starea *End* au aceeași probabilitate, τ . Arătați că suma probabilităților tuturor secvențelor de lungime L (și care se termină prin tranziție în starea *End*) este $\tau(1 - \tau)^{L-1}$.

c. Arătați că suma probabilităților tuturor secvențelor posibile (de orice lungime) este 1. Aceasta dovedește că un model/lanț Markov este într-adevăr o distribuție de probabilitate peste spațiul tuturor secvențelor definite peste alfabetul respectiv.

15. (Verificarea înțelegerii unor noțiuni de bază)

* CMU, 2003 fall, T. Mitchell, A. Moore, HW7, pr. 2.a

Fie un model Markov ascuns cu două simboluri observate: X și Y . Se consideră următoarea probabilitate (cu notațiile tradiționale pentru HMM):

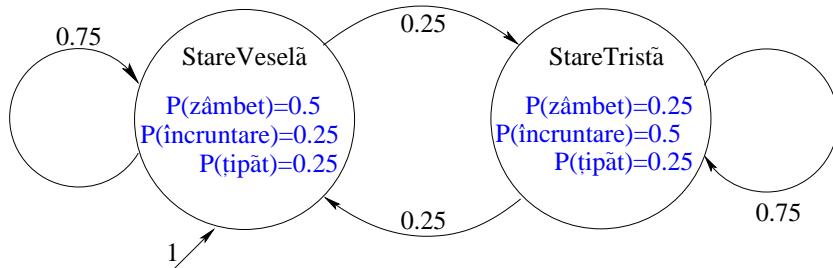
$$P(q_{t+2} = s_i \wedge q_{t+3} = s_j \wedge O_{t+3} = X \mid q_t = s_k \wedge q_{t+1} = s_m \wedge O_t = X)$$

Calculați valoarea acestei probabilități, în funcție de elementele corespunzătoare din matricea de tranziție A și din matricea de emisie B .

16. (Verificarea înțelegerii unor noțiuni de bază)

* CMU, 2003 fall, T. Mitchell, A. Moore, HW7, pr. 2.b

Vulpela Foxy duce o viață simplă. În unele zile e veselă iar în alte zile e tristă. Dar își ascunde starea emoțională, și poți să observi doar dacă zâmbește, se încrustă sau țipă. Începem în prima zi cu starea veselă și continuăm cu o tranziție în fiecare zi. Vom reprezenta evoluția stărilor emoționale a lui Foxy cu ajutorul unui model Markov ascuns, și anume:



Dacă notăm prin q_t starea din ziua t , iar prin O_t observația din ziua t , realizați următoarele cerințe:

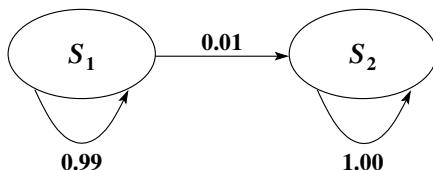
- Calculați $P(q_2 = \text{Tristă})$.
- Calculați $P(O_2 = \text{zâmbet})$.
- Calculați $P(q_2 = \text{Tristă} | O_2 = \text{zâmbet})$.
- Cât este $P(O_2 = \text{tipăt})$.
- Notăm $\phi_t = P(q_t = \text{Veselă})$. Este evident că $\phi_1 = 1$. Se poate da o definiție inductivă a lui ϕ_{t+1} în funcție de ϕ_t printr-o expresie de forma: $\phi_{t+1} = X + Y \cdot \phi_t$. Calculați valorile numerice ale lui X și Y .
- Se consideră $O_1 = \text{încruntare}$, $O_2 = \text{încruntare}$, $O_3 = \text{încruntare}$, $O_4 = \text{încruntare}$ și $O_5 = \text{încruntare}$. Care este cea mai probabilă secvență de stări? (Nu este necesar să aplicați algoritmul Viterbi.)

17.

(Verificarea înțelegerii unor noțiuni de bază)

CMU, 2010 fall, Aarti Singh, HW4, pr. 5

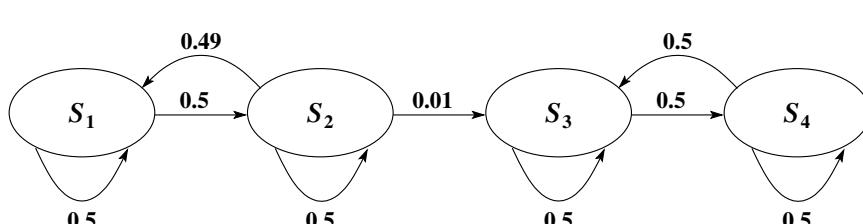
Figura de mai jos reprezintă un model Markov ascuns cu două stări. Probabilitățile de tranziție sunt înscrise pe diagramă. Probabilitățile de emisie corespunzătoare fiecărei stări sunt definite pe alfabetul de simboli $\{1, 2, 3, 4\}$ și apar în tabelul de lângă diagramă. Se consideră că probabilitățile de intrare sunt aceleași pentru ambele stări.



	S_1	S_2
$P(X = 1)$	0	0.1
$P(X = 2)$	0.199	0
$P(X = 3)$	0.8	0.7
$P(X = 4)$	0.001	0.2

- Dați un exemplu de secvență de ieșire de lungime 2 care nu poate fi generată de modelul Markov dat. Justificați răspunsul.
- Am generat o secvență de 10701^{2010} de simboli cu acest HMM și am observat că ultimul simbol din secvență a fost 3. Care este starea cea mai probabilă în care a fost generat acest simbol?
- Considerăm secvență de semnale $\{3, 3\}$. Care este cea mai probabilă secvență de stări corespunzătoare acestei secvențe de semnale? Faceți calculele.

- d. Acum vom considera secvența de semnale $\{3, 3, 4\}$. Care sunt primele două stări ale celei mai probabile secvențe de stări corespunzătoare secvenței de semnale considerate? Faceți calculele.
- e. Putem încerca să extindem întrucâtva capacitatea de modelare a acestui HMM prin „duplicarea“ fiecărei din cele două stări. Urmând această idee, am creat diagrama de mai jos. Putem stabili probabilitățile de intrare și cele de emisie în așa fel încât acest model cu 4 stări, având probabilitățile de tranziție indicate în diagramă, să fie echivalent cu modelul original cu 2 stări (adică, pentru orice secvență de semnale O , $P(O | \text{HMM}_{2\text{states}}) = P(O | \text{HMM}_{4\text{states}})$)? Dacă da, cum anume? Dacă nu, justificați.



18. (Algoritmul Forward; demonstrație – pasul inductiv)
Liviu Ciortuz, 2020

Fie un model Markov ascuns μ .

- a. La pasul de inițializare al algoritmului Forward se fixează $\alpha_i(1) = \pi_i$ pentru $i = 1, \dots, N$, unde N este numărul de stări din modelul Markov ascuns μ , iar π_i este probabilitatea de intrare în starea i . Formula de recurență de la pasul inductiv al algoritmului Forward pentru HMM este următoarea:

$$\alpha_j(t+1) = \sum_{i=1}^N \alpha_i(t) a_{ij} b_{ij O_t}, \text{ pentru } j = 1, \dots, N \text{ și } t = 1, \dots, T. \quad (427)$$

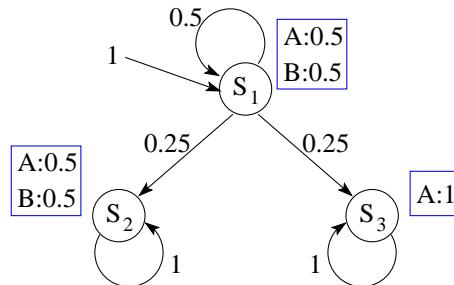
Faceți demonstrația acestei formule de recurență. Pentru fiecare egalitate din cadrul demonstrației, precizați motivul/motivele pentru care are loc egalitatea respectivă.

- b. Cum se poate calcula $P(O_1 O_2 \dots O_T | \mu)$, probabilitatea de emisie a unei secvențe de semnale $O_1 O_2 \dots O_T$ de către modelul μ , folosind probabilitățile Forward (α)?

19. (HMM, calculul probabilității secvențelor emise:
o formulă alternativă; demonstrație și aplicare)
** CMU, 2009 spring, Ziv Bar-Joseph, final, pr. 5*

Pentru calcularea probabilității unei secvențe de observații, $P(O_1, \dots, O_T)$, suntem obișnuiți cu folosirea probabilităților Forward sau a probabilităților Backward (vedeți problemele 18.b și respectiv 8.b). În această problemă veți folosi alte probabilități pentru a determina o formulă de calcul pentru $P(O_1, \dots, O_T)$.

Apoi veți aplica noua formulă pentru secvențe de observații corespunzătoare următorului model Markov ascuns:



a. Fie următoarele notării:

$$\nu_i^t = P(O_1, \dots, O_T \mid q_t = s_i) \text{ și } p_t(i) = P(q_t = s_i).$$

Determinați o formulă pentru $P(O_1, \dots, O_T)$ care să folosească doar ν_i^t și $p_t(i)$.

b. Folosind formula găsită la punctul precedent, calculați $P(O_1 = B, \dots, O_{200} = B)$ pentru modelul dat, adică probabilitatea observării a 200 de emisii *B* consecutive. Pentru aceasta trebuie să găsiți un moment *t* convenabil, să calculați ν_i^t și $p_t(i)$ corespunzători și apoi să folosiți aceste valori pentru a calcula probabilitatea cerută. (*Indicație*: pentru calcularea $p_t(i)$ se observă că tranzițiile din și spre stările S_2 și S_3 sunt simetrice pentru orice *t*, deci $p_t(S_2) = p_t(S_3)$.)

c. Folosind formula găsită la punctul *a*, calculați $P(O_1 = A, \dots, O_{200} = A)$ pentru modelul dat, adică probabilitatea observării a 200 de emisii *A* consecutive. Trebuie să găsiți din nou un moment *t* convenabil, dar de data aceasta este permisă calcularea în funcție de ν_1^t , pentru care nu trebuie obținută valoarea numerică. Toate celelalte probabilități implicate în formulă trebuie însă calculate. (*Indicație*: momentul *t* ales la acest punct poate fi diferit de cel ales la punctul *b*.)

20.

(HMM: o variație pe tema algoritmului Viterbi)

*York University, UK, 2001 spring,
Univ. of York (UK), J. Cussens, D. Kudenko, open exam., pr. 1.1*

Trebuie să călătorești între două puncte, A și B. Aceste puncte sunt despărțite de *n* râuri. Pe fiecare din cele *n* râuri sunt construite *m* poduri. Poți trece peste un râu doar dacă folosești unul din cele *m* poduri de pe râul respectiv. Considerăm că se cunosc distanțele dintre orice două poduri care sunt situate pe râuri vecine.

- a. Definește un algoritm eficient care găsește cel mai scurt drum dintre A și B și returnează (și) lungimea acestui drum. *Observație*: Nu trebuie să implementezi acest algoritm, ci doar să descrii în pseudo-cod cum lucrează.
- b. Compară algoritmul rezultat la punctul *a* cu algoritmul Viterbi.

21.

(Algoritmul Forward-Backward: implementare)

* Liviu Ciortuz

Pentru un model Markov ascuns se consideră date (ca variabile globale):

- N (numărul de stări) și o secvență O de lungime T peste un alfabet de K simboli;
- vectorul de probabilități de intrare $\pi[N]$, matricea de probabilități de tranziție $a[N,N]$ și matricea probabilități de emisie $b[N,N,K]$;
- matricele $\alpha[N,T+1]$ și $\beta[N,T+1]$ care se presupun a fi calculate de algoritmii Forward și respectiv Backward;
- matricea $p[T+1][N][N]$ pentru claculul mediilor variabilelor neobservabile (reprezentând, fiecare, tranziția de la starea i la starea j la momentul t).

Definiți în limbajul Python / C două funcții, expectation() și maximization(), care servesc la implementarea algoritmului EM/Forward-Backward pentru HMM. Funcția expectation va calcula matricea $p[T+1,N,N]$, iar funcția maximization va înlocui π , și b cu noile valori calculate de algoritmul EM.

Vă reamintim algoritmul EM/Forward-Backward:

• Estimare:

Dată o secvență (observație) O , se definește probabilitatea traversării unui anumit arc la momentul de timp t :

$$\begin{aligned} p_t(i,j) &\stackrel{\text{def.}}{=} P(X_t = i, X_{t+1} = j | O, \mu) \\ &= \frac{P(X_t = i, X_{t+1} = j, O | \mu)}{P(O | \mu)} = \frac{\alpha_i(t) a_{ij} b_{ijO_t} \beta_j(t+1)}{\sum_{m=1}^N \alpha_m(t) \beta_m(t)} \\ &= \frac{\alpha_i(t) a_{ij} b_{ijO_t} \beta_j(t+1)}{\sum_{m=1}^N \sum_{n=1}^N \alpha_m(t) a_{mn} b_{mnO_t} \beta_n(t+1)} \end{aligned}$$

Sumând după t , obținem:

$\sum_{t=1}^T p_t(i,j) =$ media (valoarea așteptată) a numărului de tranziții de la s_i la s_j în timpul emisiei lui O ;

$\sum_{j=1}^N \sum_{t=1}^T p_t(i,j) =$ media (valoarea așteptată) a numărului de tranziții din s_i în timpul emisiei lui O .

• Maximizare:

Folosind $\mu = (A, B, \Pi)$, se calculează $\hat{\mu} = (\hat{A}, \hat{B}, \hat{\Pi})$:

$$\begin{aligned} \hat{\pi}_i &= \frac{\sum_{j=1}^N p_1(i,j)}{\sum_{l=1}^N \sum_{j=1}^N p_1(l,j)} = \sum_{j=1}^N p_1(i,j) = \gamma_i(1) \\ \hat{a}_{ij} &= \frac{\sum_{t=1}^T p_t(i,j)}{\sum_{l=1}^N \sum_{t=1}^T p_t(i,l)} \\ \hat{b}_{ijk} &= \frac{\sum_{t:O_t=k, 1 \leq t \leq T} p_t(i,j)}{\sum_{t=1}^T p_t(i,j)}. \end{aligned}$$

22.

(Un HMM cu emisii gaussiene: varianta normalizată pentru probabilitățile Forward, Backward și pentru algoritmul EM/Forward-Backward; implementare și aplicare)

• o CMU, 2009 spring, Ziv Bar-Joseph, HW4, pr. 3

Robert merge să înnoate zilnic. El folosește fie culoarul 1, fie culoarul 2 din piscină. Dacă piscina este aglomerată, atunci el alege să înnoate pe culoarul 2, în caz contrar alege culoarul 1. Își notează întotdeauna timpul dus-întors de înot. Descoperă astfel că performanța sa pe culoarul 2 corespunde unei distribuții normale în jurul a 60 de secunde cu deviația standard 1, iar pe culoarul 1 unei distribuții normale în jurul a 63 de secunde cu deviația standard 1. Robert observă de asemenea un tipar al utilizării piscinei: dacă într-o zi este aglomerat, atunci probabilitatea ca în ziua următoare să fie de asemenea aglomerat este de 0.3, iar dacă nu este aglomerat, atunci probabilitatea ca în ziua următoare să fie aglomerat este de 0.6.

Deși își notează zilnic timpul efectuat, Robert nu își notează pe ce culoar înnoată. Astfel în însemnările sale se găsesc următoarele date:

Ziua (n)	Timpul (x) în secunde
1	60.0
2	62.0
3	59.0
4	63.0
5	62.0
6	64.0
7	61.0

- Desenați un *model grafic* care să descrie cât mai bine măsurătorile făcute de Robert (noteate sub forma x_n) precum și variabilele ascunse (noteate cu z_n).
- Calculați distribuția condiționată a variabilelor z_n în raport cu x_1, \dots, x_n . Vă cerem să elaborați calculul/demonstrația în întregime.

Indicație: Ați putea formula $\alpha(z_n) = P(z_n | x_1, \dots, x_n)$ în funcție de $\alpha(z_{n-1})$, într-o formă parametrizată.

- Derivați ecuația de recursie înapoi (engl., backward equation) pentru probabilitatea a posteriori $P(z_n | x_1, \dots, x_N)$, notată cu $\beta(z_n)$, unde n ia valori în $\{1, \dots, N\}$. Faceți calculul/demonstrația în întregime.

Indicație: Ați putea formula $\beta(z_n)$ în funcție de $\beta(z_{n+1})$ și $\alpha(z_n)$, într-o formă parametrizată.

- Cât este media condiționată pentru tranziția z_n, z_{n+1} dată fiind secvența de emisii x_1, \dots, x_N , adică $E(z_n z_{n+1}^T | x_1, \dots, x_N)$? Faceți calculul / demonstrația în întregime.

Indicație: Tineți cont că aveți variabile ascunse discrete (de tip categorial) pentru stări și variabile continue (de tip gaussian) pentru observații. Așadar, $E(z_n z_{n+1}^T | x_1, \dots, x_N)$ va fi o matrice 2×2 în acest caz.

Pentru punctele următoare ($e - h$), este necesar ca în prealabil să faceți implementarea formulelor pe care le-ați determinat la punctele $b - d$.

e. Folosind datele din tabelul de mai sus, stabiliți dacă piscina este aglomerată sau nu în ziua a patra.

f. Robert nu este sigur în legătură cu estimările sale pentru culoarele 1 și 2. Ați putea să faceți o re-estimare? În mod concret, puteți presupune că performanțele lui pe culoarele 1 și 2 corespund distribuțiilor normale $\mathcal{N}(\mu_1, \sigma_1^2)$ și respectiv $\mathcal{N}(\mu_2, \sigma_2^2)$. Puteți estima parametrii $\mu_1, \sigma_1, \mu_2, \sigma_2$?

Sugestie: Veți putea apela la un algoritm de tip EM care folosește probabilitățile $\alpha(z_n)$ și $\beta(z_n)$ determinante anterioare. Veți considera că probabilitățile de tranziție, precum și valorile inițiale ale parametrilor $\mu_1, \sigma_1, \mu_2, \sigma_2$ nu se re-estimează aici.

g. Pe de altă parte, Robert nu este sigur în legătură cu pattern-ul zilelor aglomerate de la piscină, adică nu știe dacă probabilitățile de tranziție au fost corect estimate ca fiind 0.3 și 0.6. Ați putea să faceți o re-estimare a acestor probabilități folosind observațiile de mai sus?

Sugestie: Veți putea apela la un [alt] algoritm de tip EM care folosește probabilitățile $\alpha(z_n)$ și $\beta(z_n)$ determinante anterioare. Veți considera că valorile inițiale ale probabilităților de tranziție nu se re-estimează aici. Veți folosi ca valori pentru parametrii celor două distribuții gaussiene mediile $\mu_1 = 60, \mu_2 = 63$ și varianțele $\sigma_1^2 = \sigma_2^2 = 1$.

h. Să presupunem că Robert nu este sigur nici în legătură cu parametrii $\mu_1, \sigma_1, \mu_2, \sigma_2$ și nici cu parametrii zilelor aglomerate de la piscină. Ați putea să faceți o re-estimare a tuturor acestor parametri folosind observațiile de mai sus?

Vă solicităm să implementați algoritmul dumneavoastră de învățare automată în limbajul de programare preferat. Indicați rezultatele obținute în urma estimării parametrilor folosind datele din tabel.

„Mai bun este sfârșitul unui lucru decât începutul lui...“

Cartea Eclesiastului 7:8.a

Alcătuirea acestei culegeri a fost pentru noi, autorii ei, asemenea unui drum, pe care l-am parcurs sub îndrumarea unor profesori și asistenți de departe. La capătul acestui drum, ne dăm seama că lungimea drumului și greutățile prin care am trecut n-au fost determinante, ci cel mai mult au contat „peisajele“ care se desfășoară de o parte și de alta a drumului.

Dragă „cititorule“, sperăm că după ce vei fi parcurs, la rândul tău, alături de noi o parte din acest drum (care pe noi ne-a încântat!), ți-am devenit și noi călăuze, dar despărțindu-ne aici, îți vei continua propriul tău drum, fie aplicând cele arătate aici, fie parcurgând alte capitole. Sau, pur și simplu, folosind abilitățile pe care le-ai dobândit *antrenându-te* cu astfel de exerciții, vei putea să le *generalizezi* în propriile tale domenii de interes. Îți dorim sincer *Mult succes!*

Autorii



© M. Romanică