Domaće zadaće iz strojnog učenja

Ak god 2019./2020.

Planira li netko ovo nadopunjavati tijekom ljetnog semestra? Ako da, ima nas još kojima je ostalo strojno pa se možemo dogovoriti da ne mora sve jedan. :)

Tesko preko ljetnog semestra, ali najvjerojatnije preko ljeta.

I

1.NAPOMENE

Napomene: Podebljana, *nakrivljena* i <u>podcrtana</u> slova označavaju vektore.
(Tako se označavaju u 2. poglavlju **zasada**!)
Npr.
$\underline{\boldsymbol{X}} = [x_1\ x_2\ x_3\ x_3\ x_4\ x_5]$

Bilo bi super staviti nekakve formulice iz Linearne algebre ili sitna obješnjena iz statistike u odvojeno poglavlje ili kao napomena u istom poglavlju za ljude koje ili nemaju osnove ili im se ne da kopati po Šnajdijevoj nedovršenoj skripti pa ništa ne naći nakon pola sata traženja.

Jos par dana imamo, jos samo grupiranje :)

Todo:

- Urediti poglavlja kao što je to u ishodima
- Napisati broj zadaće iznad svake (kao što je u prvoj)
- Nadam se da netko može pomoći :)

Domaća zadaća 1

2. Osnovni koncepti

1. **Objasnite razliku izmedu klasifikacije i regresije. Koji je od ta dva pristupa prikladan za: (a) filtriranje nezeljene e-poste (spam), (b) predvidanje kretanja dionica, (c) rangiranje rezultata trazilice? Kako biste u ovim slucajevima definirali ciljne oznake y?**
   Razlika je u tome što želimo dobiti - kod **klasifikacije** želimo dobiti jedno od *n* klasa, gdje je *n* prirodan broj dok kod **regresije** želimo dobiti izlaz koji je *realan broj*. Malo formalnije, kodomena svih hipoteza modela kod regresije je **R**, dok je kodomena svih hipoteza modela kod klasifikacije **N**. (boldani N i R su oznake skupova)
   1. Filtriranje spam-a je klasifikacija, ulaz je jedna e-poruka, a ciljna oznaka *y* je oznaka je li ulazna poruka spam ili nije.
   2. Predviđanje kretanja dionica je regresija. Ulaz bi mogao biti zadnjih *n* vrijednosti dionice, a oznaka *y* je iduca vrijednost dionice.
   3. Rangiranje rezultata tražilice također moze biti regresija. Ulaz u model može biti, primjerice, korisnikova ključna riječ, sam korisnik, njegova povijest pretraživanja, navike, oglasi koje je pogledao te je li ili nije unutar zadnjih 24 sata sam ili u društvu spomenuo 'piletinu', a izlaz je nekakva ocjena koliko će ona stranica, koja odgovara ključnoj riječi, biti zanimljiva korisniku. Zatim sortiramo vraćene rezultate po toj ocjeni, DESC.

2. **Hipoteza, model, prostor inacica**
   1. Hipoteza je funkcija koja preslikava **ulazni prostor** u **izlazni prostor**, definirana do na **parametre**. Model je **skup** hipoteza, indeksiranih **parametrima** θ . Model takoder nazivamo prostorom **hipoteza**, a dimenzija tog prostora jednaka je _____. Ucenje modela odgovara pretrazivanju **prostora hipoteza** u potrazi za **optimalnom** hipotezom. To je ona hipoteza koja **najbolje** klasificira oznacene primjere, sto procjenjujemo pomocu **empirijske pogreske** mjerene na **skupu za testiranje**. Drugim rijecima, ucenje modela svodi se na **trazenje** parametara modela s **indikatorskom funkcijom** kao kriterijskom funkcijom.
   2. **Rjesavamo problem binarne klasifikacije u prostoru primjera $\chi = \{0, 1\}^2$. Definirajte linearan model koji ce primjere odvajati pravcem.** $h(x ; \theta) = x_1 \theta_1 + x_2 \theta_2 + \theta_0$
   3. **Koja je dimenzija prostora parametara? Koliko razlicitih hipoteza postoji u H?**
      Nemam pojma!
   4. **Neka je skup oznacenih primjera sljedeci:** $D = \{(x(i), y(i))\} = \{((0, 0), 0), ((1, 1,), 0), ((1, 0), 1), ((0, 1), 1)\}$.
      Odredite konkretnu hipotezu koja ima najmanju empirijsku pogresku. Postoji vise hipoteza sa najmanjom empirijskom pogreskom. Neke od njih su $h(x ; \boldsymbol{\theta}) = 1 \cdot x_1 + 1 \cdot x_2 - 0.5 \geq 0$ ili $h(x ; \boldsymbol{\theta}) = 1 \cdot x_1 + 1 \cdot x_2 - 1.5 \leq 0$
   5. **Definirajte prostor inacica $VS_{H,D}$. Odredite $VS_{H,D}$ za ovaj konkretni problem.**
      Prostor inacica VS H,D je podskup modela H u koji ulaze sve one hipoteze koje ispravno

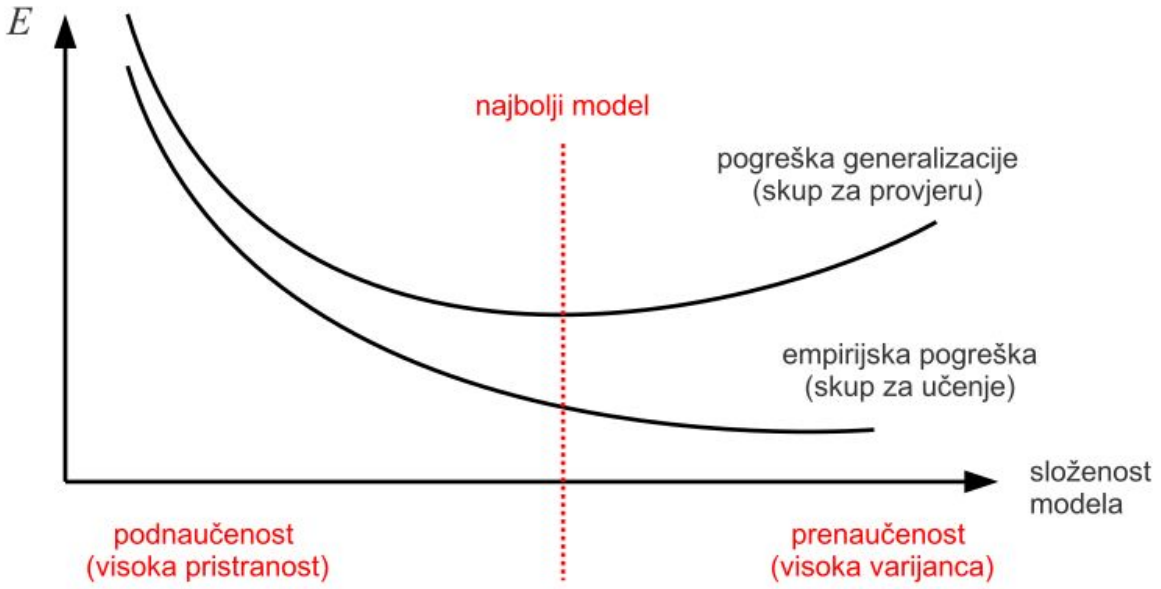
klasificiraju **sve** primjere iz skupa za ucenje D. U gornjem primjeru, VS H ,D =∅ jer ne postoji pravac 2D ravnine koji bi tocno klasificirao sve tocke iz D.

3. Induktivna pristranost

1. **Definirajte induktivnu pristranost (neformalno i formalno). Koje su dvije vrste pristranosti koje sačinjavaju induktivnu pristranost?**
   Induktivna pristranost je skup pretpostavki, dodatne informacije, koje imamo o problemu, s kojima onda možemo zaključiti o dosad neviđenom primjeru. To su upravo pretpostavke s kojima generaliziramo iznad skupa za ucenje. Induktivna pristranost se sastoji od pristranosti ogranicavanjem (odabir modela, poput linearnog ili skupa pravokutnika) i od pristranosti preferencijom (kad hipoteze pretražujemo određenim redoslijedom).
2. Raspolazemo skupom oznacenih primjera u ulaznom prostoru $\chi = \{0 , 1\}^3$ D={(x (i) , y (i))}={((0 , 0 ,0) , 0),((1 , 0 , 0) , 1),((1 ,0 ,1),1), ((0 , 1 , 0), 1) ,((0 , 1 ,1) ,1)}
   Koja je klasifikacija nevidenih primjera?
   Nepoznata!!!
3. **Definirajte linearan model H za** $\chi = \{0 , 1\}^3$ **. Koja je to vrsta pristranosti?** Pristranost ograničavanjem. $h(x ; \boldsymbol{\theta}) = x_1 \theta_1 + x_2 \theta_2 + x_3 \theta_3 + \theta_0$
4. **Možete li odrediti klasifikaciju neviđenih primjera uz odabrani model H? Je li pristranost koja proizlazi iz odabira modela dovoljna za jednoznačnu klasifikaciju primjera iz D? Odredite skup prostora $VS_{H,D}$.**
   Postoje dvije moguce klasifikacije, u obje su točke (1, 1, 1) i (1, 1, 0) oznacene sa 1. U jednoj od mogucih klasifikacija, tocka (0, 0, 1) oznacena je sa 0, a u drugoj je oznacena sa 1, i u obje klasifikacije moguce je naci ravninu koja odjeljuje pozitivne od negativnih primjera te je za obje klasifikacije skup
   $VS_{H,D} \neq \varnothing$ te zbog toga svega klasifikacija svih primjera iz X još uvijek nije jednoznačna (moguca je za sve primjere iz skupa D).
5. **Definirajte (neformalno) neku dodatnu nepristranost takvu da klasifikacija svakog primjera slijedi jednoznačno na temelju skupa primjera D. Koje je vrste ta dodatna pristranost?**
   Mozemo definirati da ce sve tocke s prvom i trecom koordinatom jednakom nuli biti oznacene sa 0, a sve ostale sa 1, čime jednoznačno određujemo jednu od prethodnih klasifikacija. Ako želimo drugu, mozemo reći da je jedina tocka koja je negativno oznacena (0, 0, 0). Obje vrsta pristranosti su vrste pristranosti ograničavanjem.

4. **Osnovne komponente**
    1. **Nabrojite tri osnovne komponente algoritma strojnog učenja.**
    Model (skup hipoteza), funkcija gubitka (informacija koliko je predviđanje modela daleko od stvarnosti) i optimizacijski postupak (postupak kojim nalazimo one parametre za hipotezu koji minimiziraju funkciju gubitka).
    2. **Identificirajte uz koje se komponente veze koja vrsta induktivne pristranosti.**
    Pristranost ograničavanjem veze se uz model, dok se pristranost preferencijom veže uz funkciju gubitka i optimizacijski postupak

5. **Funkcija gubitka, empirijska pogreska**
    1. **Pogreska hipoteze je ocekivanje funkcije gubitka L. Nad kojom distribucijom je definirano to ocekivanje? Koji je problem s takvom definicijom u praksi?**
    Pogreska hipoteze je ocekivanje funkcije gubitka L definirano nad distribucijom skupa za ucenje D. To u praksi ponekad predstavlja problem kad se skupljeni skup za ucenje po distribuciji znacajno razlikuje od stvarne distribucije (skupljeni podaci imaju svoju pristranost).
    2. **Definirajte empirijsku pogresku preko funkcije gubitka L. Koja je pretpostavka implicitno ugradena u tu definiciju?**
    Empirijska pogreska definirana je kao ocekivanje funkcije gubitka nad primjerima iz skupa za ucenje. Implicitno je ugradena pretpostavka da su svi primjeri skupljeni u skupu za ucenje D jednako vjerojatni (faktor 1/N).
    $$E(\theta|D)=\frac{1}{N}\sum_{i=1}^{N} L\left(y^{(i)}, h(\boldsymbol{x}^{(i)}|\theta)\right)$$
    3. **Kod asimetričnih gubitaka, funkciju L mozemo definirati preko matrice gubitaka. Definirajte takvu matricu za problem klasifikacije neželjene e-poste te izracunajte funkciju pogreske za slucaj pet pogresno negativnih i dvije pogresno pozitivne klasifikacije od ukupno deset (N=10) primjera.**
    S obzirom da je propuštanje važnog maila jer je otisao u spam (osobno, moze biti drugi broj) 10 puta gore od dobivanja spam poste u inbox, matrica guibitka izgleda ovako :
    $$\begin{bmatrix} 0 & 1 \\ 10 & 0 \end{bmatrix}$$
    gdje red i stupac 0 oznacavaju "poruka nije spam". S tom definicijom racunamo funkciju pogreske:
    $$\frac{1\cdot 2+10\cdot 5}{10}=5.2$$



6. **Intuicija oko odabira modela**
   1. **Skicirajte krivulje pogreske ucenja i ispitne pogreske u ovisnosti o slozenosti modela. Naznacite podrucja prenaucenosti i podnaucenosti.**
      Evo slikica:

   2. **Objasnite zasto pogreska ucenja s povecanjem slozsnosti modela teži prema nuli.**
      Zato sto se kompleksniji model moze vise prilagoditi podacima iz skupa za ucenje D, stoga broj krivo klasificiranih tocaka pada.
   3. **Raspolazemo modelom Ha koji ima hiperparametar a kojim se moze ugadati slozenost modela. Za odabrani a, naucili smo hipotezu koja minimizira empirijsku pogresku. Unakrsnom provjerom utvrdili smo da je ispitna greska znatno veca od pogreske ucenja. Je li nas odabir hiperparametra suboptimalan?**
      Naravno. Velika razlika izmedu greske na skupu za treniranje i skupu za testiranje nam snazno hinta na to da se model pretrenirao na ispitne podatke i sum u njima, i nije uspio generalizirati.
   4. **Raspolazemo modelm Ha s hiperparametrom a (veci a daje slozeniji model). Raspolazemo dvama optimizacijskim algoritmima L1 i L2.Algoritam L2 losiji je od algoritma L1, u smislu da L2 nalazi parametre koji su losiji od parametara koje pronalazi L1, odnosno E(t2|D) > E(t1|D). Neka a*1 oznacava optimalnu vrijednost hiperparametra za Ha ucenog algoritmom L1, a a*2 optimalnu vrijednost za Ha ucenog algoritmom L2. Nacinite skicu analognu onom u (a) zadatku i naznacite vrijednosti pogresaka za modele Ha*1 i Ha*2.**
      Nemam pojma!!!

5. **Moze li model ucen losijim algoritmom L2 imati manju ispitnu pogresku od modela koji je ucen boljim algoritmom L1, ali nije optimalan? Skicirajte takvu situaciju na prethodnoj skici.**
   S obzirom da ne znam prethodni, ne znam ni ovaj. Pretpostavljam da moze s obzirom da traze skicu ¯_(ツ)_/ ¯

Domaća zadaća 2

3. Regresija

1. **[Svrha: Razumjeti matrično rješenje za neregulariziranu i regulariziranu regresiju i izvježbati potrebnu matematku. Razumjeti kako loša kondicija matrice utječe na stabilnost rješenja i kako regularizacija to popravlja.]**

   (a) Izvedite u matričnom obliku rješenje za vektor w za neregularizirani linearni model regresije uz kvadratnu funkciju gubitka.

$$h(x \mid w) = w^T * x$$

$$E(h|D) = \frac{1}{2}\sum(h^{(i)}(x) - y^{(i)})^2 = \frac{1}{2}\sum(w^T x^{(i)} - y^{(i)})^2$$

$$E(h|D) = \frac{1}{2}(w^T x^{(i)} - y^{(i)})^T (w^T x^{(i)} - y^{(i)}) = \frac{1}{2}(xw - y)^T (xw - y)$$

MINIMIZACIJA POGREŠKE: $\nabla_w E(w|D) = 0$

$$E(h|D) = \frac{1}{2}((xw)^T xw - (xw)^T y - y^T xw + y^T y) =$$
$$= \frac{1}{2}((xw)^T xw - y^T(xw) - y^T xw + y^T y)$$
$$= \frac{1}{2}((xw)^T xw - 2y^T(xw) + y^T y) = \frac{1}{2}(w^T x^T xw - 2y^T xw + y^T)$$

Ovo je najbolje što sam našao za znak parcijalne derivacije E po vektoru težina w.

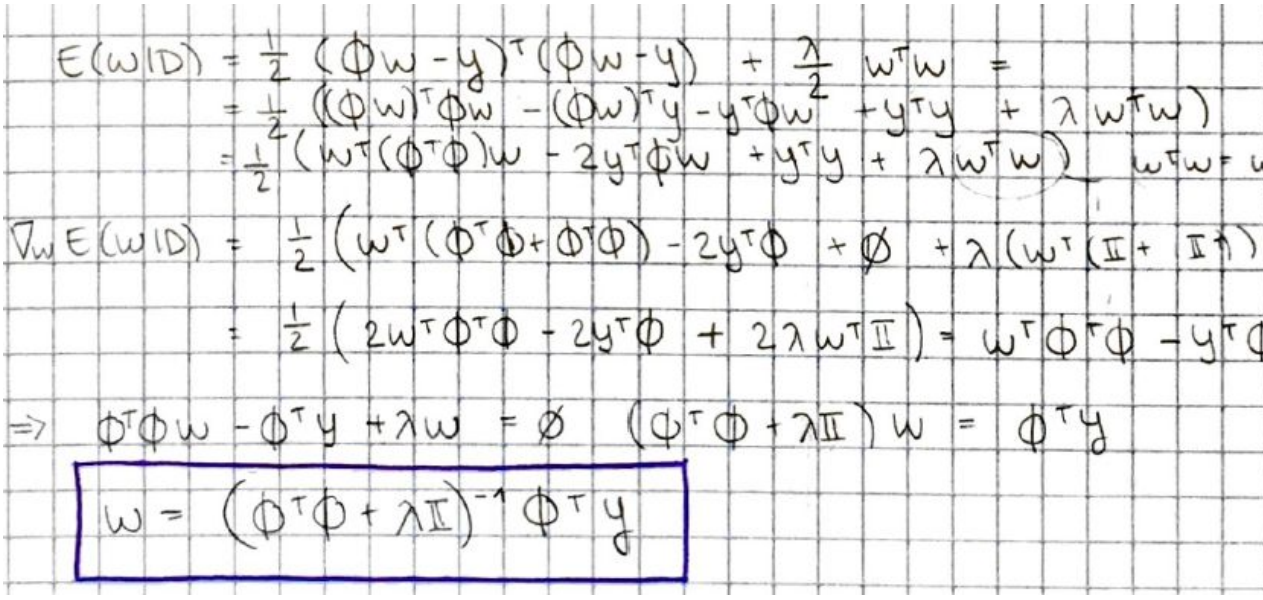
$$\nabla_w E(w|D) = \frac{1}{2}(w^T(A + A^T) - 2y^T x + 0)$$
$$= \frac{1}{2}(w^T(x^T x + x^T x) - 2y^T x) = 0$$

$$w^T x^T x = y^T x \rightarrow x^T xw = x^T y \rightarrow w = (x^T x)^{-1} x^T y$$

Pseudoinverz od x: $(x^T x)^{-1} x^T$

$$w = X^+ y = (X^T X)^{-1} x^T y$$

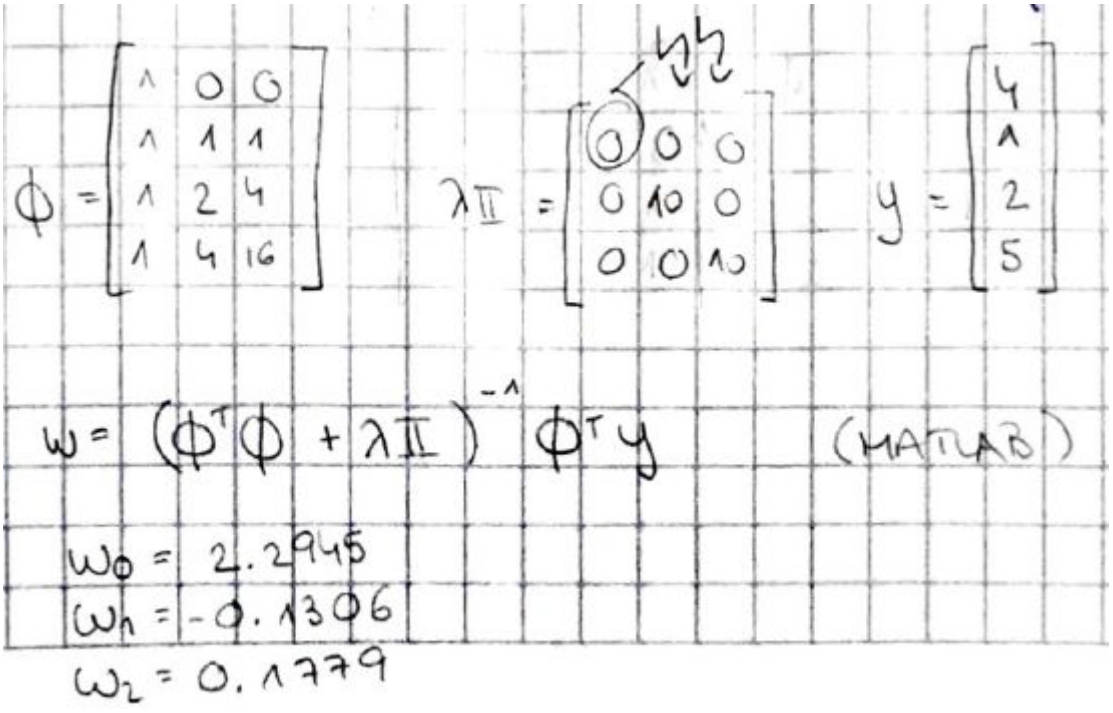
(b) Izvedite to rješenje za L2-regularizirani model.



(c) Raspolažemo sljedećim skupom primjera za učenje:

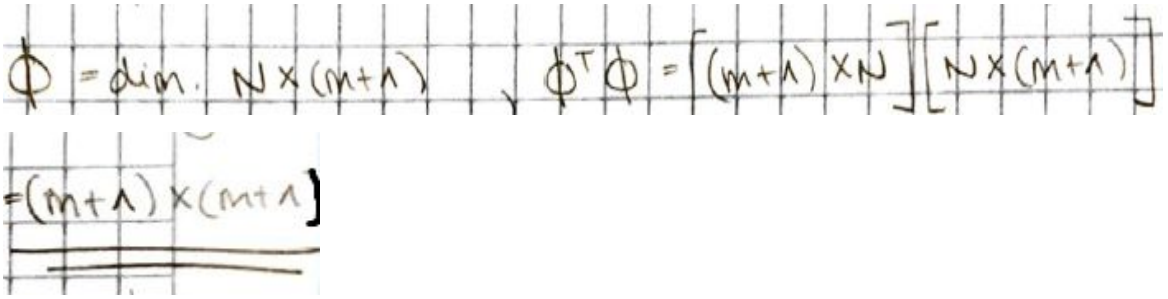
$$\mathcal{D} = \left\{(x^{(i)}, y^{(i)})\right\}_{i=1}^{4} = \left\{(0,4),(1,1),(2,2),(4,5)\right\}.$$

Podatke želimo modelirati polinomijalnom regresijskom funkcijom $h(x) = w_0 + w_1 * x + w_2 * x^2$, uz regularizacijski faktor $\lambda = 10$. Napišite kako bi u ovome konkretnom slučaju izgleda jednadžba iz zadatka (b) (Ne morate ju izračunavati, samo ju napišite.)



(d) Jednadžba iz zadatka (b) daje rješenje u zatvorenoj formi, međutim njezin izračun ponekad može biti računalno zahtjevan. Što konkretno predstavlja problem? Je li problem izražen kada imamo mnogo primjera za učenje ili kada imamo mnogo značajki? Obrazložite odgovor.

Računanje inverza matrice velikih dimenzija računski je izrazito zahtjevno. Problem je izražen kad imamo mnogo značajki n zbog:



A kako je od tog faktora **(+λ*I)** tražimo inverz, računska je složenost veća što imamo veći broj značajki.

(e) Rješenje jednadžbe iz zadatka (a) može biti numerički nestabilno. Što to znači kada i će to biti slučaj? Postoji li takav problem u slučaju formulacije problema iz zadatka (b)? Obrazložite.

Rješenje može biti nestabilno kada postoji singularitet. To će biti slučaj kada je matrica $\Phi^T\Phi$ loše kondicionirana, odnosno kada dva stupca ili retka te matrice jako koreliraju.

U slučaju formulacije problema iz (b) to neće biti problem jer matrici $\Phi^T\Phi$ koja može biti problematična, dodajemo faktor **λ*I** (koji je dijagonalna matrica) što rješava problem. Na taj način smanjujemo korelaciju između stupaca i/ili redaka, odnosno smanjujemo kondicijski broj.

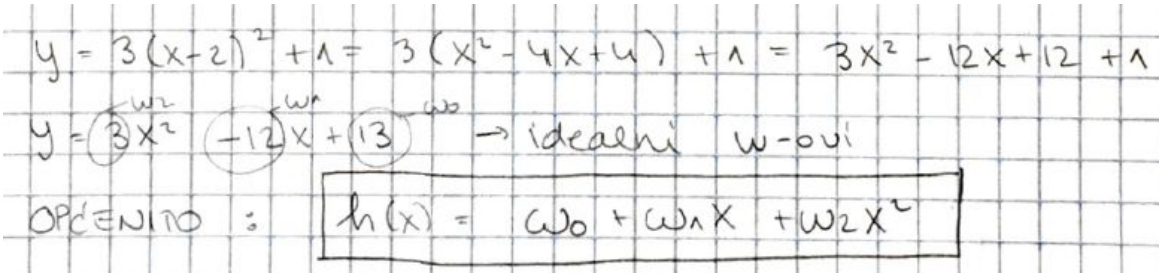
2. **[Svrha: Shvatiti kako se nelinarna funkcija u ulaznom prostoru funkcija preslikava u linearnu funkciju odnosno (hiper)ravninu u prostoru znacajki.]**

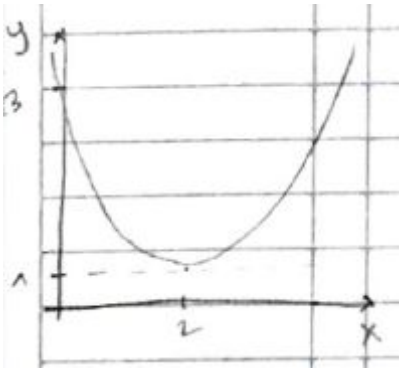
   (a) Regresijom zelimo aproksimirati funkciju jedne varijable:

   $y = 3 * (x - 2)^2 + 1$

   Skicirajte tu funkciju u ulaznome prostoru. Denirajte linearan model h(x) uz funkciju preslikavanja $\phi(x) = (1, x, x^2)$.

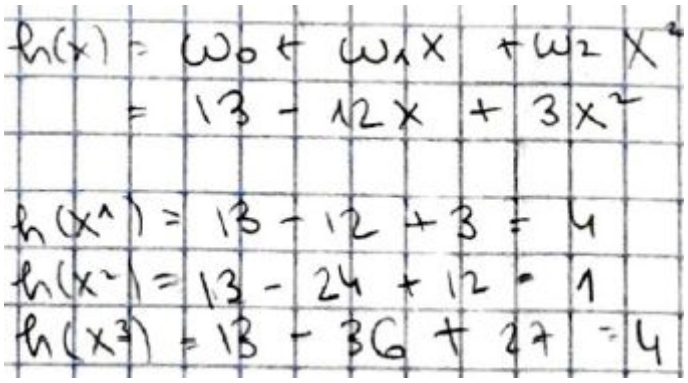
   Odredite vektor tezina **w =** (w0, w1, w2) tog modela.

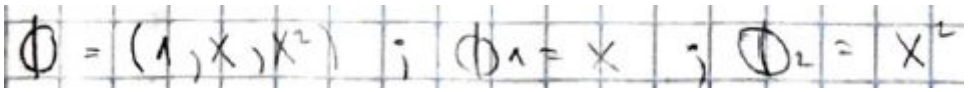


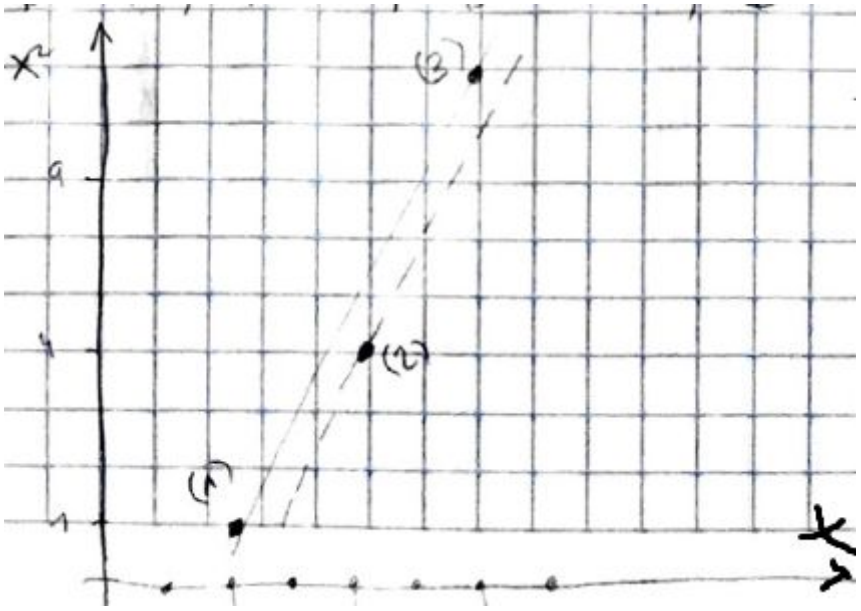


(b) Skicirajte u prostoru (x_1, x_2) izokonture funkcije y. Naznacite u tom prostoru tocke u koje se preslikavaju primjeri $x^{(1)} = 1$, $x^{(2)} = 2$ i $x^{(3)} = 3$.
Koja je vrijednost od h(x) za navedene primjere?

Vrijednosti od h(x) za navedene primjere:

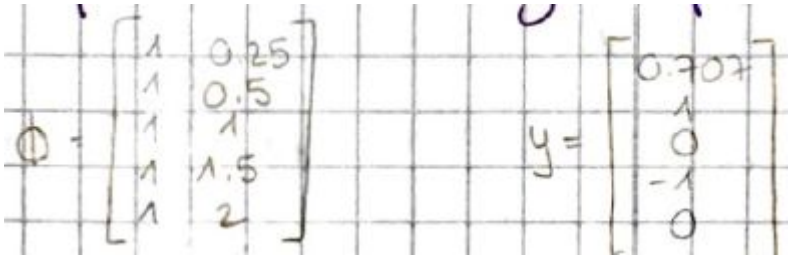


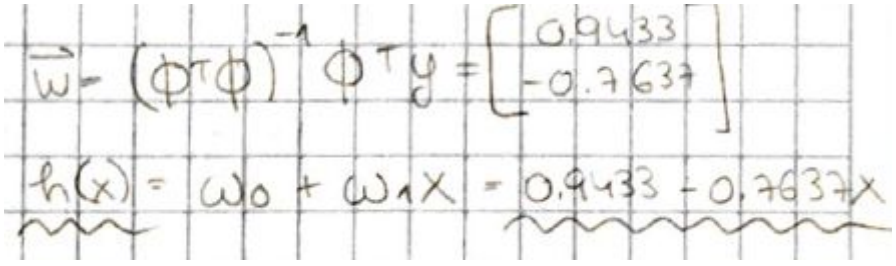


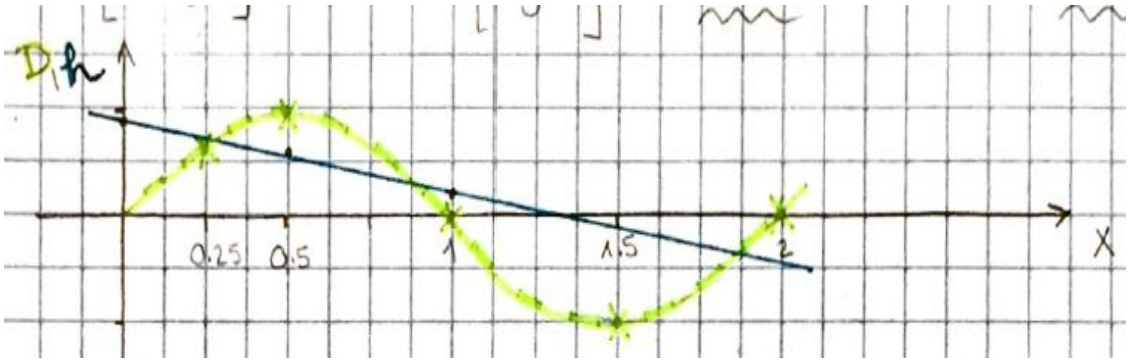


3. **[Svrha: Isprobati izracun regresijskog modela s razlicitim funkcijama preslikavanja i razviti intuiciju o tome kako funkcija preslikavanja odreduje slozenost hipoteze u ulaznome prostoru.]**
   Linearnim modelom univarijatne regresije zelimo aproksimirati jednu periodu funkcije $f(x) = \sin(\pi x)$.
   Raspolažemo skupom primjera za učenje:
   $$\mathcal{D} = \{(0.25, 0.707), (0.5, 1), (1, 0), (1.5, -1), (2, 0)\}.$$

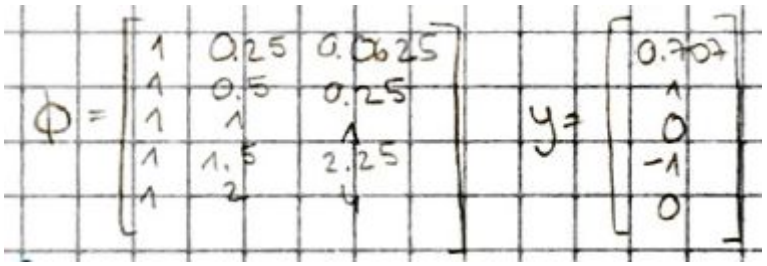
   (a) Izracunajte parametre linearnog modela regresije u izvornom prostoru primjera, tj. s funkcijom preslikavanja deniranom kao $\phi(x) = (1, x)$. Skicirajte dobivenu regresijsku funkciju.

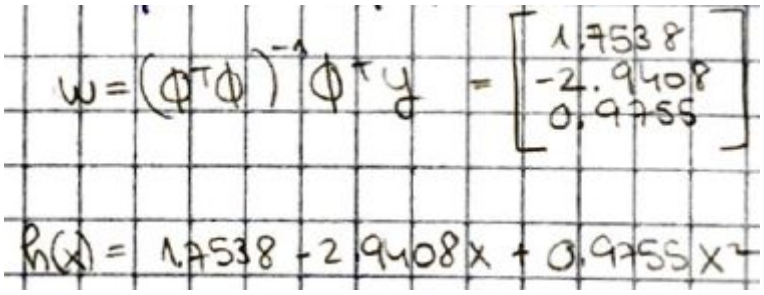


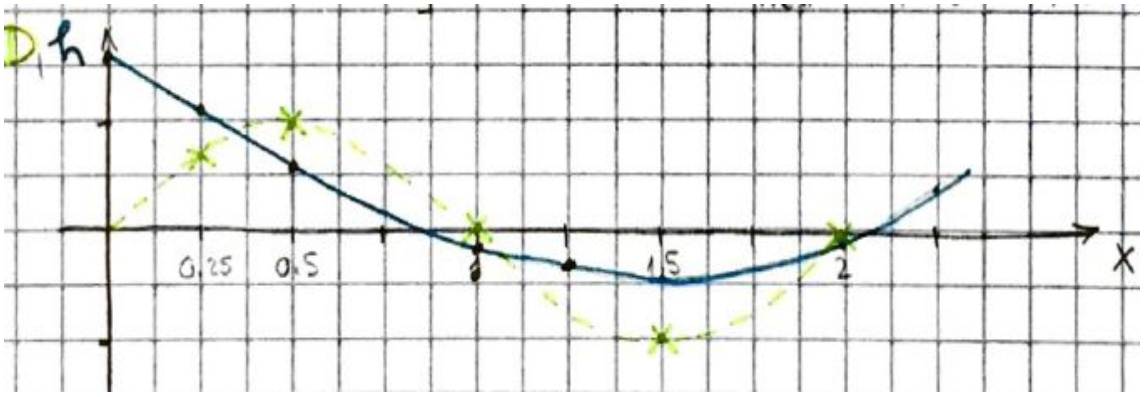




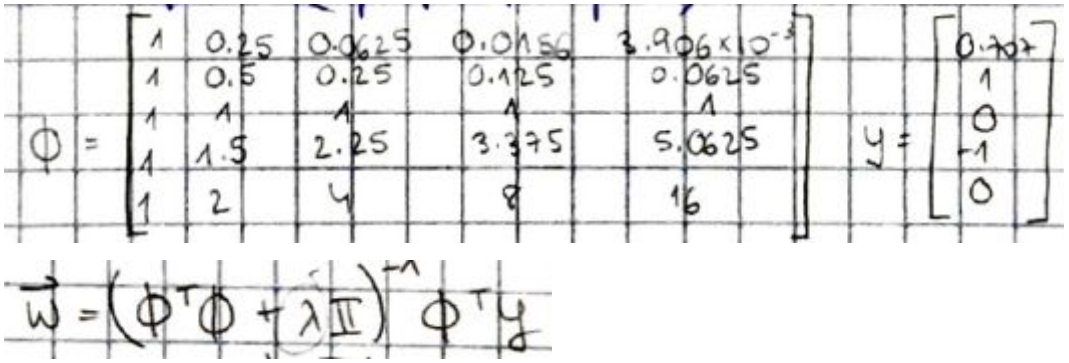
(b) Izracunajte parametre modela jednostruke polinomijalne regresije drugog stupnja, tj. modela koji koristi funkciju preslikavanja $\phi(x) = (1, x, x^2)$.
Skicirajte dobivenu regresijsku funkciju.



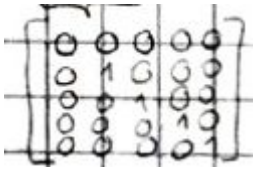


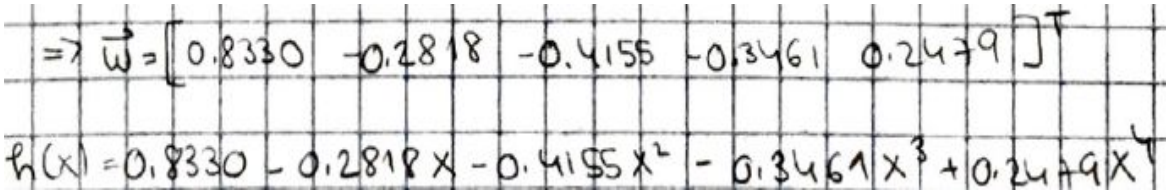


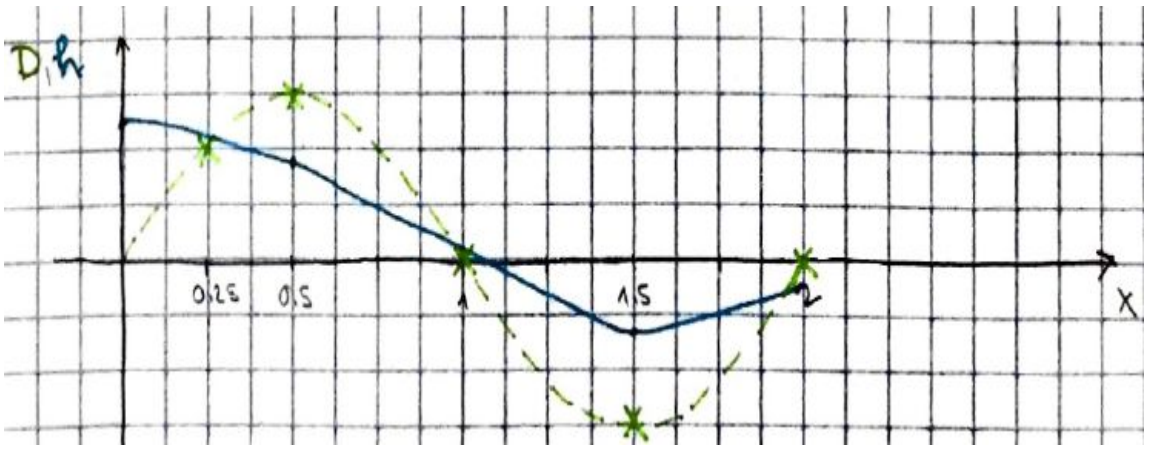
(c) Izracunajte parametre modela visestruke polinomijalne regresije cetvrtog stupnja, tj. modela koji koristi funkciju preslikavanja $\phi(x) = (1, x, x^2, x^3, x^4)$, uz L2-regularizaciju (**λ** = 1). Skicirajte dobivenu regresijsku funkciju.



λ: (Zapamtite da regularizacija ne obuhvaća x_0, stoga je 0 na poziciji (0,0) unutar matrice)





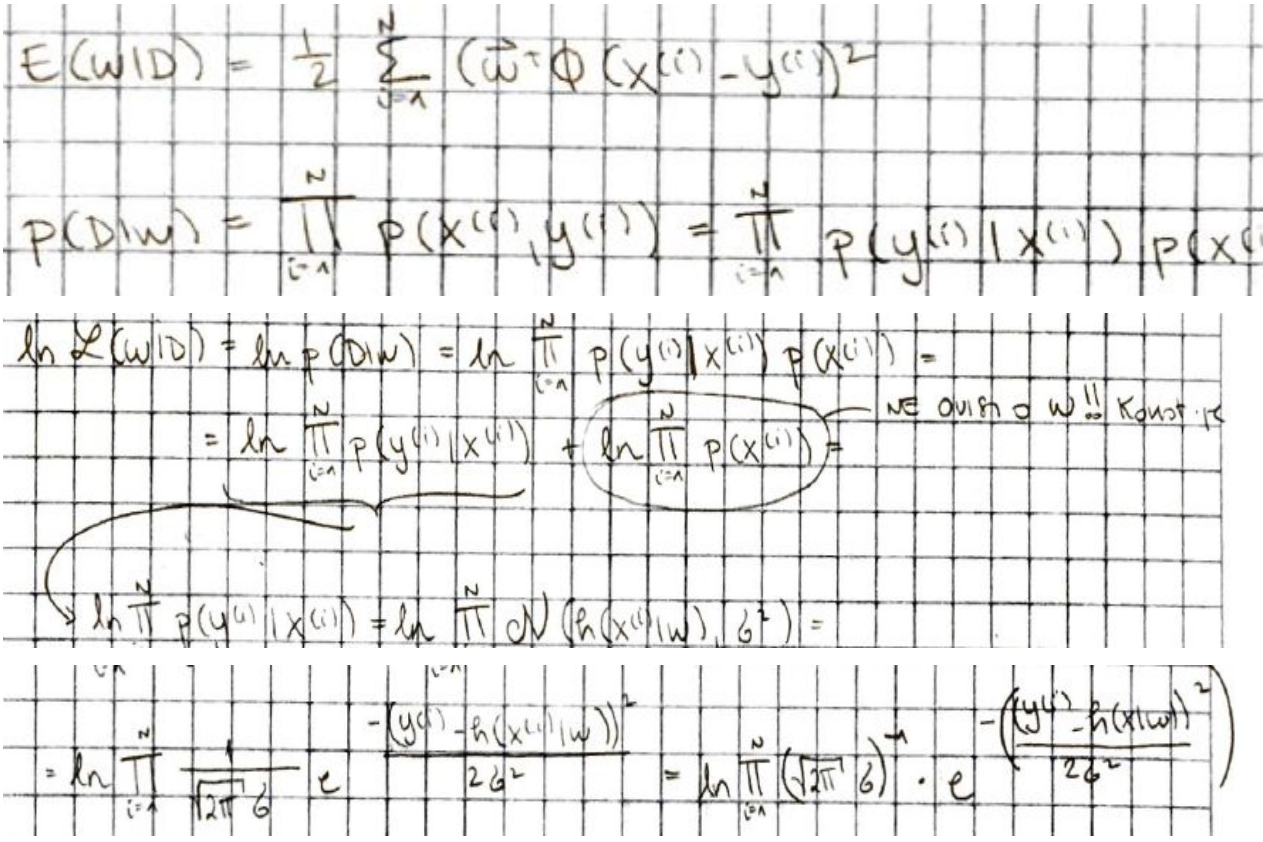


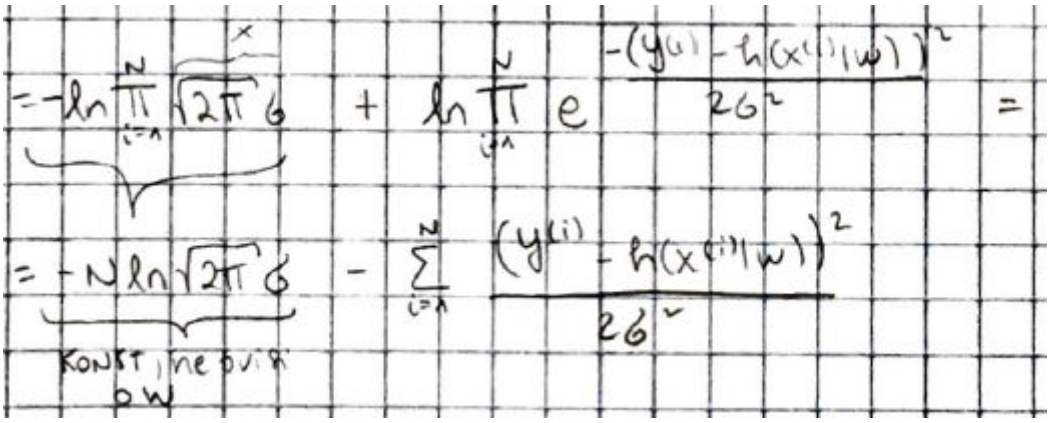
(d) Koji je model u ovom slucaju najprikladniji? Zasto?
Najprikladiniji je model pod **(c)** s obzirom da ima najmanju kvadratnu pogrešku i oblikom najbolje prati željenu funkciju $f(x) = \sin(\pi x)$.

4. **[Svrha: Uvjeriti se da, uz odredene pretpostavke, kvadratna pogreska ima probabilisticko tumacenje i opravdanje.]**
Kod postupka najmanjih kvadrata empirijska je pogreška denirana kao:

$$E(\mathbf{w}|\mathcal{D}) = \frac{1}{2}\sum_{i=1}^{N}\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}^{(i)}) - y^{(i)}\right)^2.$$

Pokazite da je minimizacija gornjeg izraza istovjetna maksimizaciji log-izglednosti $\ln P(\mathcal{D}|\mathbf{w})$ (odnosno minimizaciji negativne log-izglednosti) uz pretpostavku normalno distribuiranog suma $\mathcal{N}(h(\mathbf{x}|\mathbf{w}), \sigma^2)$.





5. **[Svrha: Razumjeti vezu izmedu faktora regularizacije i slozenosti modela.]**

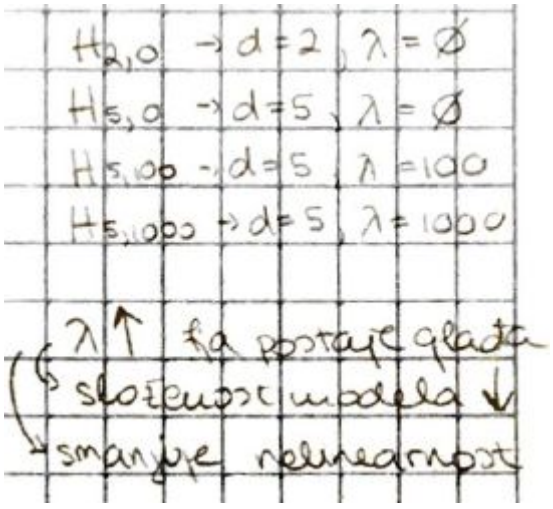
   Neka $\mathcal{H}_{d,\lambda}$ oznacava model polinomijalne regresije stupnja d s L2-regularizacijskim faktorom λ. Razmatramo cetiri modela:

   $\mathcal{H}_{2,0},\ \mathcal{H}_{5,0},\ \mathcal{H}_{5,100},\ \mathcal{H}_{5,1000}$

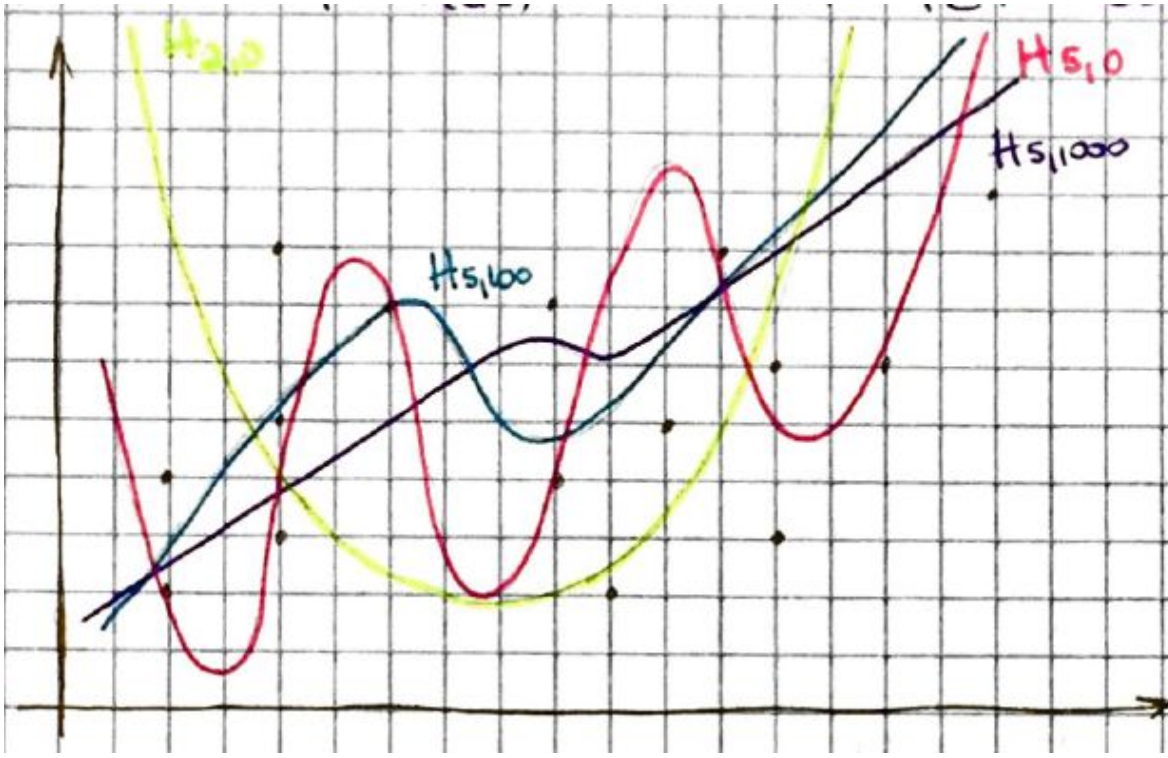
   u ulaznome prostoru $\mathcal{X} = \mathbb{R}$.

   Pretpostavimo da su podatci u stvarnosti generirani funkcijom koja je polinom treceg stupnja (d 3). Pretpostavite da imamo razmjerno malo podataka i da je šum u podatcima razmjerno velik. Na dva odvojena crteza skicirajte

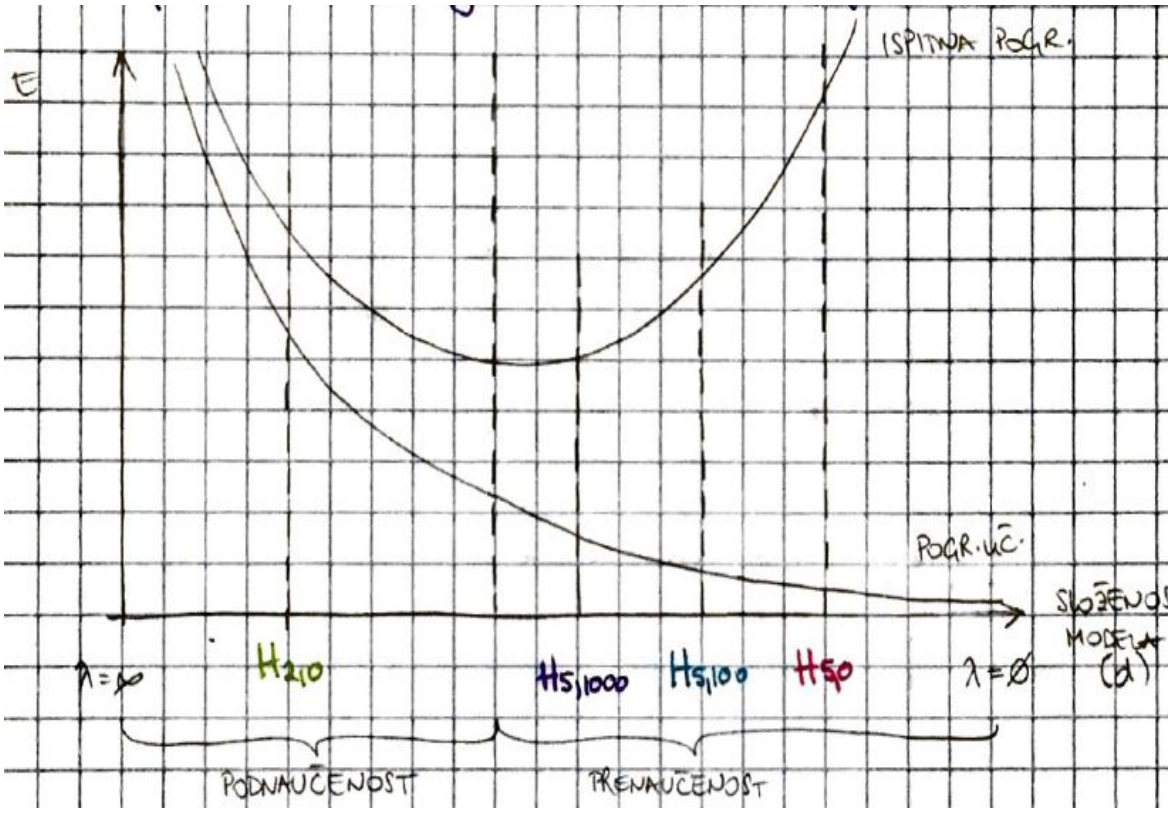
   (a) regresijsku funkciju h(x) za svih pet modela te



(Žuto je $H_{2,0}$)



(b) pogresku ucenja i ispitnu pogresku za svih pet modela.



$H_{2,0}$ ima sigurno najveću pogrešku, stupanj d je premali i nema regularizacije. $H_{5,0}$ će imati najmanju pogrešku učenja (model smo odlično istrenirali na primjerima), ali će imati najveću pogrešku generalizacije (jer nemamo uključenu regularizaciju i model se prilagodio šumu).

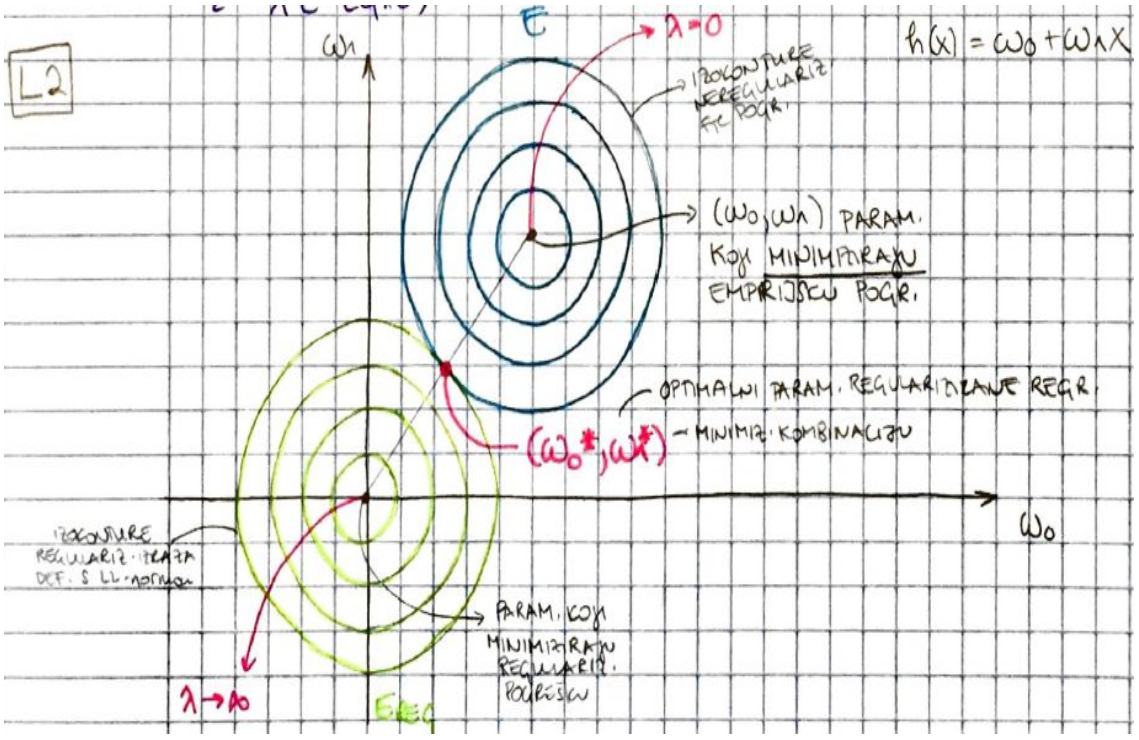
$H_{5,100}$ ima veću pogrešku generalizacije od $H_{5,1000}$ zbog manjeg λ., ali manje od $H_{5,0}$ jer ima veći λ.. Pogreška učenja od $H_{5,100}$ je manja od $H_{5,1000}$, ali veću od $H_{5,0}$ jer se manje prilagođava na šum zbog regularizacije.

6. **[Svrha: Shvatiti kako regularizacija utjece na optimizaciju. Shvatiti geometrijski argument zasto L1-regularizacija rezultira rijetkim modelima.]**
   (a) Objasnite koja je svrha regularizacije i na kojoj se pretpostavci temelji.
   SVRHA regularizacije je sprječavanje prenaučenosti modela ograničavanjem rasta vrijednosti parametara.

   TEMELJI se na pretpostavci da što je linearni model složeniji, to ima veće vrijednosti parametara. Iz tog razloga kažnjavamo hipoteze s visokim vrijednostima parametara! (u funkciju pogreške ugrađujemo mjeru složenosti modela)

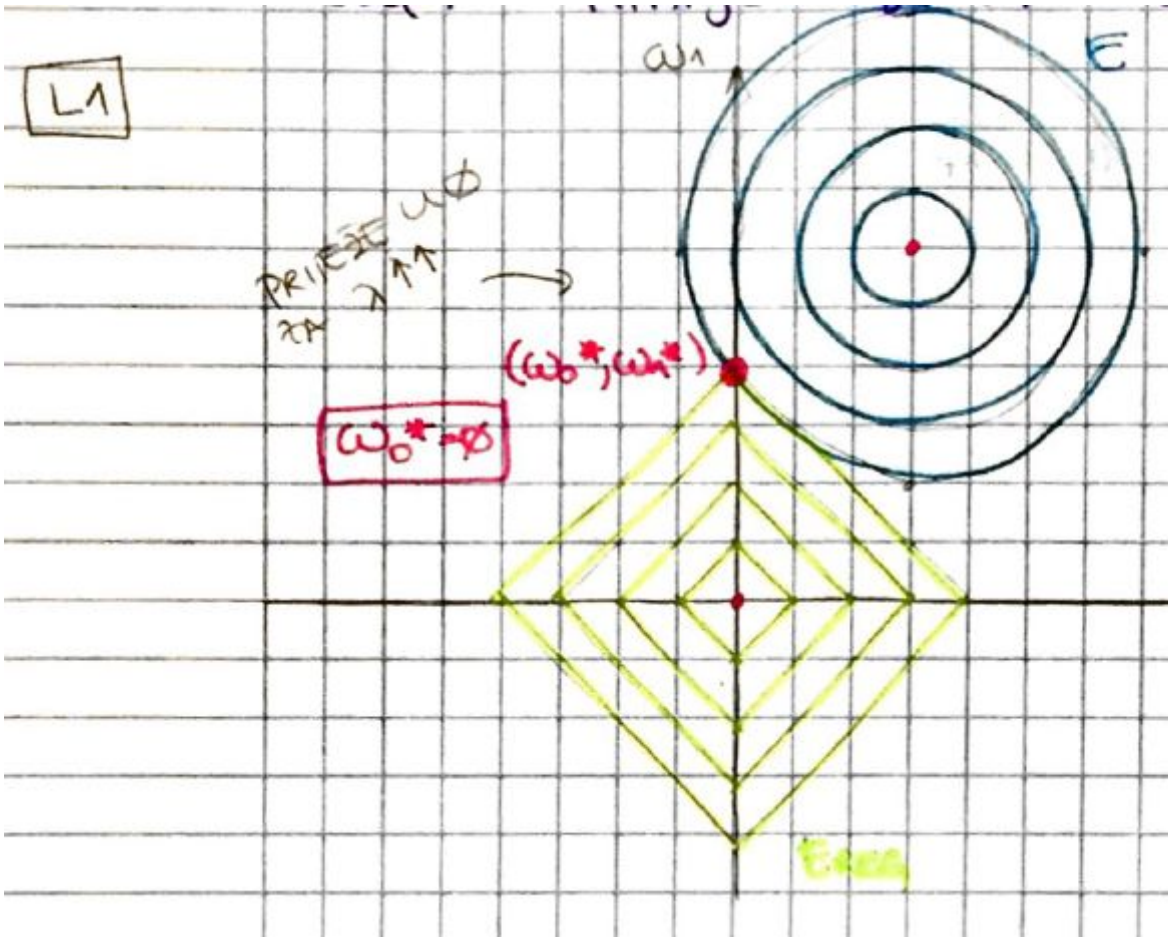
   (b) Koja je prednost regulariziranog modela u odnosu na neregularizirani? Dolazi li ta prednost vise do izrazaja u slucajevima kada imamo puno primjera za ucenje ili kada ih imamo malo?
   Prednost regulariziranog modela u odnosu na neregularizirani je ta da je regrulazirirani model teže pr enaučiti! To dolazi izražaja kad imamo malo primjera za učenje.
   Druga prednost je poboljšanje numeričke stabilnosti.

   (c) Razmatramo visestruku regresiju, $h(x) = w_0 + w_1x_1 + w_2x_2$. Skicirajte izokonture neregularizirane funkcije pogreske u ravnini $\mathbb{R}^2$ koju definiraju parametri w_1 i w_2 (napomena: funkcija pogreške je konveksna). Zatim skicirajte izokonture regularizacijskog izraza deniranog L2-normom vektora tezina (i ova je funkcija konveksna). Pomocu ove skice objasnite na koji nacin regularizacija utjece na izbor optimalnih parametara (w_1^*, w_2^*).
   Skicirajte krivulju mogucih rjesenja za $\lambda \in [0, \infty)$.



Optimalni parametar uvijek se nalazi na spojnici izokontura.
Optimalni parametri su manji od parametara koji minimiziraju funkciju pogreške (kod neregularizirane regresije).
L2-regresija ne može rezultirati rijetkim modelima, za to bi se optimalni parametar morao nalaziti na w_1 osi . Plus, L2 kažnmjava težine proporcionalno njihovom iznosu.
L2 daje rješenje u zatvorenoj formi.

(d) Ponovite prethodnu skicu, ali ovog puta sa L1-regularizacijom. Na temelju ove skice pokusajte odgovoriti na pitanje zasto L1-regularizacija daje rjede modele od L2-regularizacije.



Izokonture regularizacijskog izaraza kod L1-reguarizacije imaju KVADRATAN OBLIK te je vjerojatnost da će se s izokonturom neregulariziranog funkcije pogreške sjeći blizu (ili na) osi kvadratnog sustava (w_0, w_1) puno veća nego kod izokonture L2 regularizacijskog izraza, rezultirajući time rjeđim modelima.

7. **[Svrha: Shvatiti vezu izmedu tezine znacajki, vaznosti znacajki i slozenosti modela.]**

(a) Treniramo model regresije uz nelinearnu funkciju preslikavanja $\phi : \mathbb{R}^n \to \mathbb{R}^m$, gdje je m > n, uz L2-regularizaciju. Optimalan regularizacijski faktor λ odredili smo unakrsnom provjerom. Kako, nakon treniranja modela, mozemo provjeriti (1) koje su znacajke nebitne i (2) je li izvorni model preslozen?

(1) One značajke čije su vrijednosti nakon treniranja modela regresije uz L2 regeularizaciju jako prigušene (odnosno parametri w_j uz te značajke su približno nula) su nebitne!

Odnosno manje su bitne od značajki koje nisu puno prigušene. (zato što L2 teže kažnjava značajke većeg iznosa ($w_j >> 0$))

(2) Izvorni model je presložen kad su parametri w_j uz "najsloženije" značajke (najviše potencije kod polinoma) blizu nule! To znači da je L2 kaznio, odnosno prigušio te značajke.

(b) Kako bi se u ovom slucaju ponasao L1-regularizirani model?

L1-regularizacijski model bi vjerojatno te značajke u potpunosti pritegnuo na nulu, rezultirajući rijetkim modelom. Ili bi ih u najmanju ruku još više prigušio.

(c) Pretpostavite da u skupu postoji skup multikolinearnih znacajki koje su, osim sto su redundantne, takoder i irelevantne, odnosno zavisna varijabla u stvarnosti uopce ne ovisi o tim varijablama. Ako model nije regulariziran, koje su ocekivane tezine tih znacajki?

NEMAM RJEŠENO

Domaća zadaća 3

4. Linearni diskriminativni modeli

Link na zadaću: https://www.fer3.net/attachments/36338/

Domaća zadaća 4

5. Logistička regresija

Ima netko možda ove DZ kojih nema u docsu? :)

Domaća zadaća 5

5. Stroj potpornih vektora

6. **Jezgrene i neparametarske metode**

7. **Procjena parametara**

Neka je zajednicka vjerojatnost P(X,Y) varijabli sljedeća: P(1, 1) = 0.2, P(1, 2)=0.05, P(1, 3) = 0.3, P(2, 1) = 0.05, P(2, 2) = 0.3, P(2, 3) = 0.1.

(a) Izracunajte ocekivanje E[X], varijancu Var(X), kovarijancu Cov(X, Y), koeficijent korelacije $\rho_{X,Y}$ i kovarijacijsku matricu Σ.

- $X = [[1,\ 2],\ [0.55,\ 0.45]],\ Y = [[1,\ 2,\ 3],\ [[0.25\ 0.35\ 0.4]]$
- $E[X] = \sum_{i=1}^{2} p(x_i)x_i = 1.45,\ E[Y] = \sum_{i=1}^{3} p(y_i)y_i = 2.15$
- $Var(X) = E[(X - E[X])^2 = 0.55 \cdot (-0.45)^2 + 0.45 \cdot 0.55^2 = 0.2475$
  Drugi nacin (treba se dobit isto):
  $Var(X) = E[X^2] - (E[X])^2 = 0.55 \cdot 1 + 0.45 \cdot 4 - (1.45)^2 = 0.2475$
  $Var(Y) = 0.6275$, dobiveno bilo kojom formulom
- $Cov(X,\ Y) = \sum_{i=1}^{2}\sum_{j=1}^{3} p_{ij}(x_i, y_j)(x_i - E[X])(y_j - E[Y]) =$
  $0.2 \cdot (1 - 1.45)$(*Dokažite da su nezavisne varijable linearno nekorelirane*.$1 - 2.15) + 0.05 \cdot (1 -$
  $+ 0.05 \cdot (2 - 1.45)(1 - 2.15) + 0.3 \cdot (2 - 1.45)(2 - 2.15) + 0.1 \cdot (2 - 1.45)(3 - 2.15) =$
  $= -7/400 = -0.0175$
  Takoder, mozemo iskoristiti $Cov(X, Y) = E[XY] - E[X]E[Y]$ na nacin da prvo pomnozimo sve moguce vrijednosti X sa Y i vidimo vjerojatnosti dobivenih umnozaka, pa izracunamo ocekivanje E[XY], a a onda samo oduzmemo ocekivanje pojedinih
  $XY = [[1,\ 2,\ 3,\ 4,\ 6], [0.2,\ 0.1,\ 0.3,\ 0.3,\ 0.1]]$
  $E[XY] = 0.2 \cdot 1 + 0.1 \cdot 2 + 0.3 \cdot 3 + 0.3 \cdot 4 + 0.1 \cdot 6 = 3.1$
  $Cor(X, Y) = 3.1 - 1.45 \cdot 2.15 = -0.0175$
- $\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{-0.0175}{\sqrt{0.2475 \cdot 0.6275}} =$ -0.0444
- $\Sigma = [[0.2475,\ -0.0175], [-0.0175,\ 0.6275]]$
- Hint za poslije: koristi numpy ko covjek, a ne Casio (osim na ispitu)

Razumjeti kako podaci odreduju izglednost parametara putem funkcije izglednosti

(b) Definirajte nezavisnost slucajnih varijabli (preko zajednicke vjerojatnosti i preko uvjetne vjerojatnosti).

- $P(X, Y) = P(X) \cdot P(Y)$. Alternativno:
  $P(A \cap B) = P(A) \cdot P(B) \Rightarrow P(A) = \frac{P(A \cap B)}{P(B)} = P(A|B)$, odnosno uvjetna vjerojatnost A, uz dani B, je ista kao vjerojatnost A (isto je i obratno), dakle B ne uvjetuje nista oko A, to jest, A i B su nezavisni.

(c) Sudeci po iznosu koeficijenta $\rho_{X,Y}$, jesu li X i Y linearno zavisne ili nezavisne?

- A ono, kao jesu nesto malo negativno linearno zavisne, u praksi bi to bilo cak i zanemarivo. Nezavisne ne mozemo reci, mozda imaju neku drugu zavisnost, korelacija mjeri samo linearnu zavisnost.

(d) Za koje od sljedecih varijabli ocekujete da su zavisne, a za koje da je ta zavisnost linearna: (i) dobi i velicina cipela, (ii) dob i sati spavanja, (iii) razina buke i udaljenost od izvora buke, (iv) dobi i prihodi?

- (i) Zavisno, i to linearno do neke dobi a onda stane.
- (ii) Klinci spavaju puno, odrasli manje, bake i dedeki opet puno. Zavisnost postoji ali nije linearna.
- (iii) Intenzitet zvucnog vala pada kvadratno s obzirom na udaljenost od izvora (Fiz2), ali tlak pada obrnuto proporcionalno udaljenosti, a razina zvuka pada logaritamski sa udaljenosti. Nema linearne zavisnosti.
- (iv) Klinci ne dobivaju nista, odrasli dobivaju puno, bake i dedeki dobivaju… Skoro pa nista haha :(. Nema linearne zavisnosti

(e) Dokažite da su nezavisne varijable linearno nekorelirane.

- Po definiciji, varijable su nezavisne ako im je ocekivanje zdruzene vjerojatnosti jednako umnosku ocekivanja, te iz tog trivijalno slijedi da je kovarijacija, a time i korelacija 0.
  Dokaz:

$$\rho_{X,Y} = \frac{\mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y)}{\sqrt{\mathrm{E}(X^2) - \mathrm{E}(X)^2} \cdot \sqrt{\mathrm{E}(Y^2) - \mathrm{E}(Y)^2}}$$

  Ako je ocekivanje E(X,Y) jednako umnosku ocekivanja E(X)E(Y) onda brojnik pada na 0.

Razumjeti kako podaci određuju izglednost parametara putem funkcije izglednosti.

(f) Definirajte funkciju izglednosti L(Θ|D). Na kojoj se pretpostavci o skupu D temelji ta definicija?

- Temelji se na pretpostavci da su sve tocke nezavisne i da dolaze iz iste distribucije (independent and identically distributed, i.i.d.)
- $L(\theta|D) = \prod_{i=1}^{N} p(x^{(i)}|\theta) = p(D|\theta)$

(g) Raspolazemo skupom (neoznacenih) primjera D={-2, -1, 1, 3, 5, 7}. Pretpostavljamo da se primjeri pokoravaju Gaussovoj distribuciji. Napisite funkciju izglednosti L(μ, σ^2|D). Koliko iznosi izglednost parametra μ=0 i σ^2=1, a koliko vjerojatnost uzorka D uz te parametre?

- $L(\mu,\ \sigma^2|D) = \prod_{i=1}^{N} p(x^{(i)}|\theta) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^6 exp(-(6\mu^2 - 26\mu + 89)/2\sigma^2)$

  Zasto ne -N/2 - N ln sigma - suma (xi - mi)^2 / 2 sigma^2 (MLE formula raspisana), ispadne -50.01

Zato jer pise da racunamo L, a ne ln(L). Prepisana formula odgovara logaritmu gubitka, a to se u zadatku ne trazi.

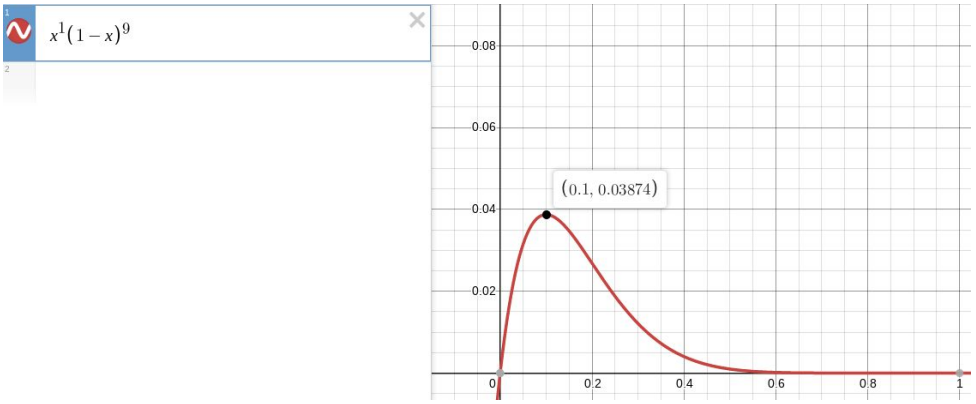
- Izglednosti parametra $L(\theta|D) = p(D|\theta)$
- Vjerojatnost uzorka uz te parametre: $p(D|\theta) = \left(\frac{1}{\sqrt{2\pi}}\right)^6 exp(-44.5) = 1.903e-22$. Dost malo, jer to nisu optimalni parametri.

(h) Novcic bacamo N puta, pri cemu smo m puta dobili glavu, a N-m puta pismo. Ishodi bacanja novcica sacinjavaju nas uzorak D. Napisite izraz za funkciju izglednosti parametra μ Bernoullijeve varijable, parametriziranu s N i m, tj L(μ|N, m)

- $L(\mu|N,m) = \prod_{i=1}^{N} \mu^x(1-\mu)^{1-x} = \mu^m(1-\mu)^{N-m}$

(i) Skicirajte funkciju izglednosti za slucaj N=10 i m=1. Koja vrijednost parametra μ je najizglednija? Uz koju je vrijednost μ skup D najvjerojatniji?

➢ Skrinano sa desmos.com/calculator



- Najizglednija je vrijednost μ=0.1, te je uz tu vrijednost skup D najvjerojatniji s vjerojatnosti dataseta od 0.03874

Osvjeziti znanje iz matematike potrebno za izvođenje MLE procjenitelja

(j) Definirajte MLE procjenitelj Θ_{MLE}.

- $\Theta_{MLE} = argmax_\Theta(L(\Theta|D)$
- S obzirom da je ln(x) uvijek rastuća funkcija, možemo ju iskoristiti za pojednostavljenje naših izraza, pa izraz gore postaje: $\Theta_{MLE} = argmax_\Theta(ln(L(\Theta|D))$
- Mi, ljudi, jako volimo matematičku jednostavnost

(k) Izvedite MLE-procjenitelj za parametar μ Bernoullijeve razdiobe P(x|μ)

- Izvod iz Snajderovog videa:

$$\ln \mathcal{L}(\mu|\mathcal{D}) = \ln \prod_{i=1}^{N} P(x|\mu) = \ln \prod_{i=1}^{N} \mu^{x^{(i)}} (1-\mu)^{1-x^{(i)}}$$
$$= \sum_{i=1}^{N} x^{(i)} \ln \mu + \left(N - \sum_{i=1}^{N} x^{(i)}\right) \ln(1-\mu)$$

$$\frac{\mathrm{d} \ln \mathcal{L}}{\mathrm{d}\mu} = \frac{1}{\mu} \sum_{i=1}^{N} x^{(i)} - \frac{1}{1-\mu}\left(N - \sum_{i=1}^{N} x^{(i)}\right) = 0$$
$$\Rightarrow \quad \hat{\mu}_{\mathrm{MLE}} = \frac{1}{N} \sum_{i=1}^{N} x^{(i)} = \frac{m}{N}$$

Nasa Likelihood funkcija je umnožak vjerojatnosti pojavljivanja pojedinih primjera, pa samo uvrstimo za vjerojatnost formulu za Bernoullijevu razdiobu
Nakon toga, uzmemo ln koji nam simpatično olakša život pretvarajući umnožak u sumu. Dobiveni izraz deriviramo i izjednačimo s 0 (tražimo maksimum jel) i dobije se samo jedno rješenje, a to je relativna frekvencija, odnosno, MLE procjenitelj za učestalost klase 1 u binarnom datasetu je udio klase 1 u svim primjerima.

(l) Isto ovo gore, ali za multinulijevu razdiobu. Koristi Lagrangeove multiplikatore.

- Nema sanse, nisam ih ni naucio na mat2. Hvala bogu, ljudi su to radili prije nas… ova preza, slajd 38-52 ili ovdje, strana 12. dno, 13. Vrh
  Tutorial putem primjera za lagrange1980s: How Donald Trump Created Donald Trump | NBC News
  :
  https://www.youtube.com/watch?v=ry9cgNx1QV8

(m)Izvedite parametre za μ i σ^2 univarijatne gaussove razdiobe.

- Ovdje, strana 2 (pozor, vecina koraka je preskocena)
- Takoder, Snajder:

$$\ln \mathcal{L}(\mu, \sigma^2|\mathcal{D}) = \ln \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x^{(i)}-\mu)^2}{2\sigma^2}\right)$$
$$= -\frac{N}{2}\ln(2\pi) - N\ln\sigma - \frac{\sum_i (x^{(i)}-\mu)^2}{2\sigma^2}$$
$$\frac{\partial \ln \mathcal{L}}{\partial \mu} = 0 \;\Rightarrow\; \hat{\mu}_{\mathrm{MLE}} = \frac{1}{N}\sum_{i=1}^{N} x^{(i)}$$
$$\frac{\partial \ln \mathcal{L}}{\partial \sigma^2} = 0 \;\Rightarrow\; \hat{\sigma}^2_{\mathrm{MLE}} = \frac{1}{N}\sum_{i=1}^{N} (x^{(i)} - \hat{\mu}_{\mathrm{MLE}})^2$$

(n) TODO: prepisati ove postupke u latex ili nesto, ndms sad a msm da ce biti dobra vjezba za ZI / ZIR
Ne da mi se ni sad lol

Isprobati izračun pristranosti procjenitelja i shvatiti da MLE moze biti pristran, tj da najveća izglednost ne jamči nepristranost.

(o) Dokažite da je μ_{MLE} nepristran, a σ^2_{MLE} pristran.
- Ovdje je za varijancu (btw varijanca == disperzija)
- Za μ_{MLE} istim postupkom dobijemo da je unbiased

(p) Je li ta pristranost u praksi problematična?
- Ovisi o količini podataka. Ako imamo mali broj uzoraka, korištenje biased umjesto unbiased estimatora može dovesti do značajnijih pogrešaka, pogotovo uz puno različitih računa. Ako imamo puno podataka, razlika je nezamjetna i više manje je svejedno.
- Navodno je već nakon 30-ak okej koristiti biased estimator, al ono, možda bolje samo stalno korigirati.

Izvježbati izračun procjene parametara multivarijatne Gaussove razdiobe. Uočiti da multikolinearnost značajki dovodi do problema.
Raspolažemo uzorkom D za koji pretpostavljamo da dolazi iz multivarijatne Gaussove razdiobe.
D = [(9.5, -0.7, -2.8), (8.8, -0.8, -3.2), (6.5, -0.2, -0.8), (2.3, 0.3, 1.2), (2.2, 0.0, 0.0), (3.6, 0.3, 1.2)]

(q) Izracunajte MLE procjenu srednje vrijednosti i MLE-procjenu kovarijacijske matrice.
(r) Izracunajte gustocu vjerojatnosti za primjer x = (-2, 1, 0). Je li ta gustoca dobro definirana? Zasto?
(s) Matrica kovarijacije Sigma mora biti pozitivno definitna, a da bi imala pozitivnu determinantu i inverz. Multikolinearnost znacajki je jedan od mogucih razloga zasto matrica nije pozitivno definitna. Izracunajte Pearsonov koeficijent korelacije ro izmedu svih parova varijabli. Izbacite varijablu koja je previse korelirana s nekom drugom varijablom te zatim u tako smanjenome ulaznom prostoru pokusajte ponovo izracunati funkciju gustoce za primjer x.

- NumPy to the rescue:

```
D:
[[ 9.5 -0.7 -2.8]
 [ 8.8 -0.8 -3.2]
 [ 6.5 -0.2 -0.8]
 [ 2.3  0.3  1.2]
 [ 2.2  0.   0. ]
 [ 3.6  0.3  1.2]]

u_MLE = np.mean(D, axis=0)
u_MLE:
[ 5.48333333 -0.18333333 -0.73333333]

Sigma_MLE = np.cov(D, rowvar=False, bias=True)
Sigma_MLE:
[[ 8.77138889 -1.19805556 -4.79222222]
 [-1.19805556  0.19138889  0.76555556]
 [-4.79222222  0.76555556  3.06222222]]

det(Sigma_MLE) = 0.0 - Ne valja! Necemo ni pokusavati provlaciti
primjer [[-2.  1.  0.]] kroz funkciju distribucije jer nam treba
inverz, a on ne postoji. Idemo istraziti nase varijable...

R = np.corrcoef(D, rowvar=False)
R:
[[ 1.         -0.92466499 -0.92466499]
 [-0.92466499  1.          1.        ]
 [-0.92466499  1.          1.        ]]
Oh my. Varijabla 2 i varijabla 3 su savrseno linearno zavisne (x_3 = 4*x_2)
Damn you, Snajdi.
Okej, rjesavamo se varijable 3...

D = D[:, 0:2]
D:
[[ 9.5 -0.7]
 [ 8.8 -0.8]
 [ 6.5 -0.2]
 [ 2.3  0.3]
 [ 2.2  0. ]
 [ 3.6  0.3]]

u2_MLE = np.mean(D, axis=0)
u2_MLE:
[ 5.48333333 -0.18333333]
n = 2

Sigma2_MLE = np.cov(D, rowvar=False, bias=True)
Sigma2_MLE:
[[ 8.77138889 -1.19805556]
 [-1.19805556  0.19138889]]

p(primjer) = np.exp(-0.5 * (u - primjer).T @ np.linalg.inv(Sigma2_MLE) @ (u - primjer) / (2pi det(Sigma2_MLE)) =
0.0006602702635529042
```

Razumjeti MAP procjenitelj i način njegovog izračuna za Bernoullijevu varijablu (beta-Bernoullijev model). Uočiti kako svojstvo konjugatnosti olakšava izračun aposteriorne distribucije.

(t) Definirajte MAP-procjenitelj theta_{MAP} i objasnite zašto je on bolji od MLE-procjenitelja.

- MLE procjenitelj je sklon overfittanju. Njegova je pretpostavka da je dataset na kojem je obavljena procjena reprezentativan, no ako smo dobili neki edge case od distribucije, moš se slikat.
  Problem leži u uzimanje procjene - koristimo iskljucivo dataset, a nikakvo prethodno znanje o mogućoj distribuciji (a priori). Tu dolazi MAP koji kombinira prior knowledge o distribuciji sa podacima za bolju procjenu od MLE.
- $\Theta_{MAP} = \text{argmax}_{\Theta}(p(D|\Theta)p(\Theta)) = \text{argmax}_{\Theta}(L(\Theta|D)p(\Theta))$

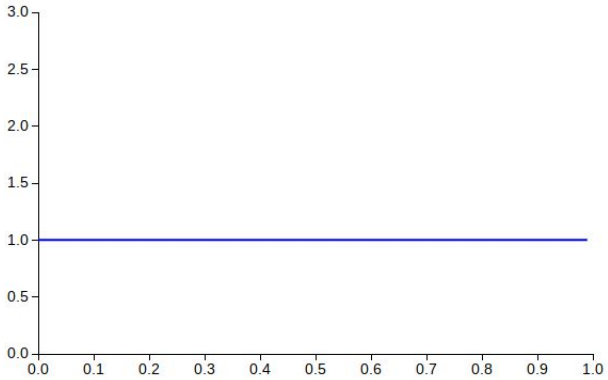
(u) Objasnite što je (1) konjugatna distribucija i (2) konjugatna apriorna distribucija. Zašto nam je svojstvo konjugatnosti bitno?

- (1) Konjugatne distribucije su one distribucije koje pripadaju istoj familiji.
- (2) Konjugatna apriorna distribucija je ona distribucija, koja pomnožena sa distribucijom podataka, daje istu vrstu distribucije kao posterior.
  Matematicki: $p(\Theta|D) \propto p(D|\Theta)p(\Theta)$, $p(\Theta)$ i $p(\Theta|D)$ su iste vrste

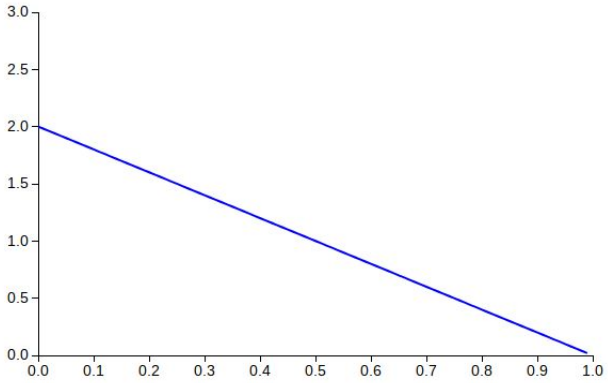
- Bitno je jer umnozak dviju distribucija ne mora dat neku lijepu distribuciju s kojom znamo raditi i naci joj maksimum, nego nesto bezveze
  Takoder, cak i ako nam da lijepu distribuciju, jos bi ljepse bilo kad bismo dobili isti tip distribucije, jer onda mozemo online uciti, odnosno prilagodavati dalje model kako dolaze novi podaci.

(v) Skica beta distribucije:

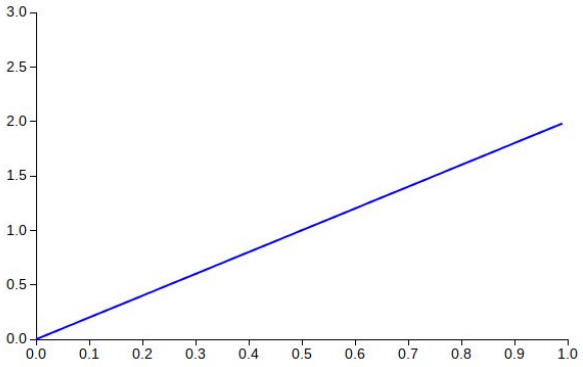
- (1) α = 1 = β



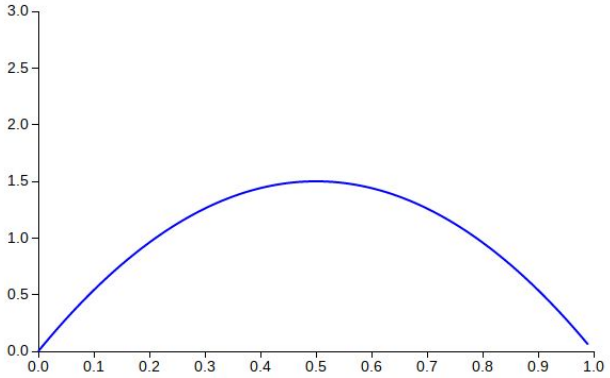
- (2) α = 1, β = 2



- (3) α = 2, β = 1



- (4) α = 2 = β



(w) Izvedite izraz za aposteriornu distribuciju parametra p(mi|N, m, alpha, beta)

- Iz Snajdijevih videa/natuknica s predavanja...

$$
\begin{aligned}
p(\mu|\mathcal{D},\alpha,\beta) &= \mu^m(1-\mu)^{N-m}\frac{1}{B(\alpha,\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1}\frac{1}{p(\mathcal{D})} \\
&= \mu^{m+\alpha-1}(1-\mu)^{N-m+\beta-1}\frac{1}{B(\alpha,\beta)p(\mathcal{D})} \\
&= \mu^{\alpha'-1}(1-\mu)^{\beta'-1}\frac{1}{B(\alpha',\beta')}
\end{aligned}
$$

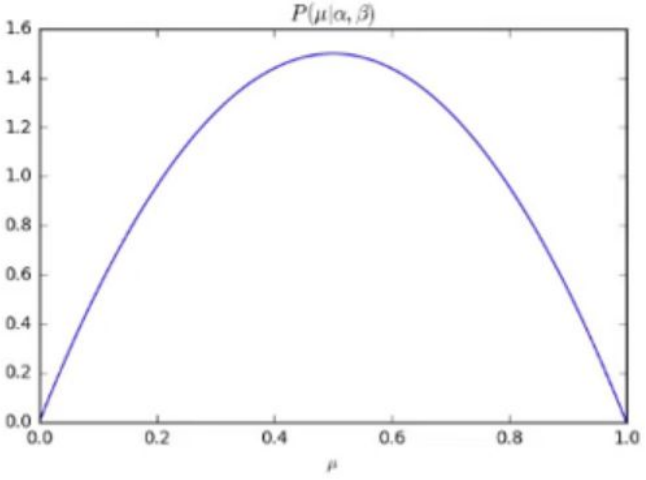
gdje $\alpha' = m + \alpha$ i $\beta' = N - m + \beta$

- MAP-procjenitelj odgovara modu aposteriorne beta-distribucije:

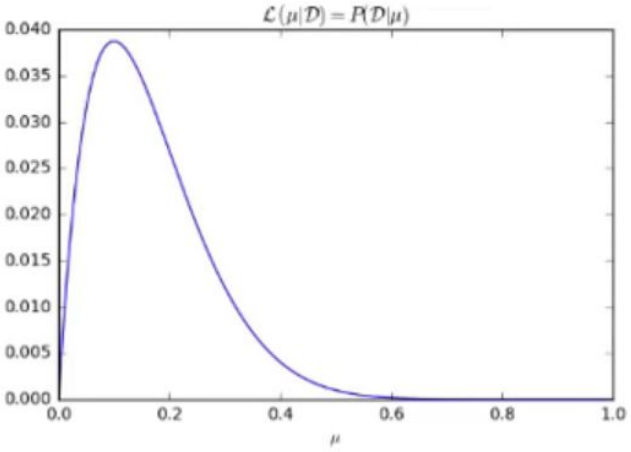
$$\hat{\mu}_{\mathrm{MAP}} = \frac{\alpha' - 1}{\alpha' + \beta' - 2} = \frac{m + \alpha - 1}{\alpha + N + \beta - 2}$$

(x) Recimo da vjerujemo da je novcic pravedan, ali da u to nismo 100% uvjereni. To mozemo modelirati beta-distribucijom sa alpha=2=beta. Zatim smo u N=10 bacnja novcica samo m=1 puta dobili glavu. Skicirajte apriornu gustocu, funkciju izglednosti te njihov umnozak.

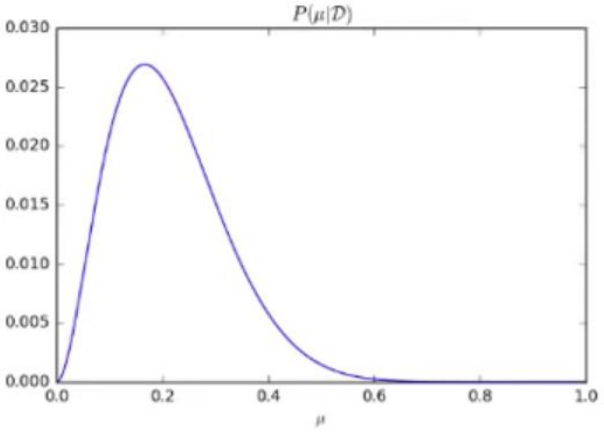
- A priori (najvece su sanse oko 0.5):



- Funkcija izglednosti nad podacima:



- Umnozak:



(y) Izracunajte MLE i MAP procjenitelj te komentirajte razliku. Kako bi porast broja primjera N utjecao na ovu razliku?

- μ_{MLE} = 1/10=0.100
- μ_{MAP} = (1+2-1) / (2 + 10 + 2 - 2) = 0.1666
- Vidljivo je kako se vrh pomaknuo prema 0.5. Daljnim povecanjem N bi se razlika smanjivala, odnosno umnozak bi opet tezio ka MLE.
  PROVJERITI!!
  pa da, ako samo dobivas jednu g ylavu, no vjerojatnije je da bi dobili vise glava i onda bi isli prema 0.5

(z) Pokazite da se MAP procjenitelj za parametar μ Bernoullijeve varijable svodi na Laplaceov procjenitelj, ako se apriorna distribucija parametra modelira beta distribucijom te ako se odaberu odgovarajuci parametra alpha i beta.

MAP-procjenitelj odgovara modu aposteriorne beta-distribucije:

$$\hat{\mu}_{\text{MAP}} = \frac{\alpha' - 1}{\alpha' + \beta' - 2} = \frac{m + \alpha - 1}{\alpha + N + \beta - 2}$$

- Snajdi kaze:

Laplaceovo zagladivanje (Laplace smoothing) – MAP sa $\alpha = \beta = 2$:

$$\hat{\mu}_{\text{MAP}} = \frac{m + 1}{N + 2}$$

- Wikipedia pak kaze da je formula:

$$p_{i,\ \alpha\text{-smoothed}} = \frac{x_i + \alpha}{N + \alpha d},$$

Gdje je alpha neki odmak a d broj klasa. Ako postavimo broj klasa na 2 (koliko ih i ima u bernoulijevoj sluc. var.) i alpha=1 dobijemo Laplace “Add One” smoothing. Quod erat demonstrandum

Razumjeti MAP procjenitelj i način njegovog izračuna za kategorijsku varijablu (Dirichlet-kategorijski model).

(aa) Definirajte Dirichletovu distribuciju.

- Dirichletova distribucija je poopćenje Beta distribucije. Koristimo ju jer je konjugatni prior multinoullijeve varijable.
- Definira se kao umnožak očekivanja za svaku klasu, dignutih na neku potenciju, koji su parametri distribucije, s time da suma svih pojedinih očekivanja klasa mora biti jednaka 1.

(bb) Definirajte Dirichlet-kategorijski model i izvedite MAP procjenitelj za ak = 2.

1. Izvod je jako jako slican izvodu za Bernoulijevu razdiobu:

- MAP-procjenitelj odgovara modu Dirichletove distribucije:

$$\hat{\mu}_{k,\text{MAP}} = \frac{\alpha'_k - 1}{\sum_{k=1}^{K} \alpha'_k - K}$$

gdje $\alpha'_k = N_k + \alpha_k$ i $N_k = \sum_i x_k^{(i)}$ (broj nastupanja k-te vrijednosti)

Razumjeti vezu između probabilističkih modela i poopćenih linearnih modela preko veze između MLE procjenitelja i minimizacije empirijske pogreške. Razumjeti vezu izmedu MAP procjenitelja i minimizacije L2 regularizirane empirijske pogreške.

(cc) Pokazite da je MLE procjena za parametre w kod linearne regresije (uz pretpostavku Gaussovog šuma) ekvivalentna postupku najmanjih kvadrata

- (Dosl smo to radili u zadaci 3, samo smo tad to zvali drukcije)

(dd) Pokažite da je MLE procjena za parametre w kod logisticke regresije (uz pretpostavku Bernoullieve distribucije oznaka) ekvivalentna minimizaciji pogreške unakrsne entropije.

- (Dosl smo to radili u zadaci 5, samo smo tad to zvali drukcije)

(ee) Gornja dva zadatka demonstriraju vezu između MLE procjenitelja i minimizacije empirijske pogreške. Postoji analogna veza izmedu MAP procjenitelja i minimizacije L2 regularizirane empirijske pogreške. Razmotrimo konkretno linearnu regresiju. Ako se apriorna vjerojatnost razdiobe težina w definira kao p(w) = N(0, 1/a * I), tj. Kao multivarijatna normalna razdioba sa središtem u ishodištu i izotropnom kovarijacijskom matricom, pomnožena nekim hiperparametrom 1/a, onda je MAP procjenitelj ekvivalentan L2 regulariziranoj kvadratnoj pogrešci. Dokažite to.

- Huh?
- Odgovor je ovdje, slajdovi 30-32
- Mental note to self: read more books

(ff) Je li u prethodnom zadatku bilo ključno to sto je Gaussova distribucija samokonjugatna? Možemo li isti princip primijeniti i kod modela gdje izglednost nije Gaussova, npr kod logističke regresije (i drugih poopćenih linearnih modela)?

- *blank stare*

8. Bayesov klasifikator

Razumjeti model Bayesovog klasifikatora i njegove komponente

(a) Definirajte model Bayesovog klasifikatora i navedite sve veličine koje se pojavljuju u definiciji modela. Objasnite zašto faktoriziramo brojnik. Objasnite ulogu nazivnika i objasnite kada ga možemo zanemariti.

- $h_j(x;theta) = P(y = j \mid x) = \frac{p(x|y=j)P(y=j)}{\sum_k p(x|y=k)P(y=k)}$
- Dijelovi:
  - ➢ P(y=j) - apriorna vjerojatnost. Vjerojatnost da je neki primjer u nekoj klasi onda kad ništa ne znamo o tom primjeru. Npr ako znamo da otprilike 0.8% ljudi ima rak, kad nam dođe pacijent, prije ikakvog pitanja znamo da je vjerojatnost da ima rak 0.008
  - ➢ p(x|y=j) - izglednost klase. Vjerojatnost da ćemo vidjeti neki primjer za neku klasu. Kao, sličnost između primjera kojeg imamo i najkarakterističnijeg primjera. Npr. ako znamo da rak pluća karakterizira kašalj, bol u prsima, krvava slina i pušenje, a dode nam pacijent sa svime navedenim, osim boli u prsima, onda možemo reći da je sličnost ~90%.
  - ➢ P(y=j|x) je aposteriorna vjerojatnost. To je udio izglednosti klase j među izglednostima svih klasa skupa.
- Posve općenito, zajednička distribucija moze biti vrlo složena, pogotovo jer kombiniramo diskretne distribucije sa kontinuiranim. Zato, faktoriziramo vjerojatnost i svaki dio modeliramo jednostavijom distribucijom.

- Kada nas ne zanima uopće vjerojatnost klase, već samo želimo klasificirati primjer, onda nam nazivnik nije bitan i maksimiziramo brojnik.

(b) Je li taj model parametarski ili neparametarski?

- Parametarski model. Broj parametara ne ovisi o broju primjera, te algoritam pretpostavlja da se primjeri pokoravaju nekoj vjerojatnosnoj distribuciji.

(c) Objasnite zašto Bayesov model nazivamo generativnim i opišite generativnu priču Bayesovskog klasifikatora.

- Jednom kad naučimo distribucije klasa i primjera koji pripadaju klasama, možemo “okrenuti pilu naopako” i generirati umjetne podatke koji bi mogli pripadati nekoj klasi. Npr kad bi naš model dobro prepoznavao ima li pacijent rak, mogli bismo ga pitati da nam iz distribucije uzoraka neki primjer koji bi mogao imati rak (ajmorec pluca) pa bi nam taj model vratio, recimo, (kašalj, bol u prsima, pušenje).

(d) Objasnite razliku između generativnih i diskriminativnih modela te navedite prednosti jednih i drugih.

- Diskriminativni modeli izravno modeliraju aposteriori vjerojatnost. Od njih možemo direktno dobiti granicu između klasa. Oni zahtjevaju puno podataka

- Generativni modeli modeliraju zajedničku distribuciju klasa i primjera. Kod njih granica između klasa ispada implicitno (tamo di su vjerojatnosti jednake). Također, u generativne modele je moguće ugraditi stručno znanje, lakša je interpretacija rezultata, te detekcija outliera, no ponekad su nepotrebno složeni i trebaju puno više podataka od diskriminativnih modela :D

Razmotrimo problem klasifikacije nezeljenih emailova u klase spam (y=1), important(y=2) i normal (y=3). Neka su apriorne vjerojatnosti tih klasa P(y=1) = 0.2, P(y=2)=0.05 i P(y=3)=0.75. Za neku poruku x, izglednosti iznose p(x|y=1)=0.8, p(x|y=2)=p(x|y=3)=0.5. Izracunajte aposteriorne vjerojatnosti za svaku od klasa te maksimalnu aposteriornu hipotezu za primjer x.

(e) p(x|y=1)*P(y=1) = 0.16
p(x|y=2)*P(y=2) = 0.025
p(x|y=3)*P(y=3) = 0.375
Sve skupa: 0.56

(f) p(y=1|x) = 0.16/0.56 = 0.2857
p(y=2|x) = 0.025/0.56 = 0.0447
p(y=3|x) = 0.375/0.56 = 0.6696

(g) Model kaze da je primjeni email najvj normalan

Razumjeti faktorizaciju zajednicke vjerojatnosti uz pretpostavku uvjetne nezavisnosti te povezanost toga s induktivnom pristranoscu i, posljedicno, brojem znacajki modela.

(h) Definirajte naivan Bayesov klasifikator i pretpostavku na kojoj se temelji.

- Pretpostavljamo uvjetnu nezavisnost P(x|y, z) = P(x|y), cime se onda brojnik iz Bayesovog pravila jednostavno faktorizira:

$$P(x|y) = \prod_{k=1}^{n} p(x_k|\, y)$$

(i) Zasto nam treba ova induktivna pretpostavka i kojoj vrsti ona pripada?

- Treba nam jer nam smanjuje slozenost modela. Kad ne bi bilo te pretpostavke, nas model bi morao imati parametar za svaku meduovisnost, a broj parametara raste eksponencijalno sa brojem znacajki.
- Ta pretpostavka je induktivna pristranost ogranicavanjem (nisam siguran, ali ima smisla)

(j) Naivan Bayesov klasifikator koristimo za klasifikaciju rukom pisanih znamenki u jednu od 10 klasa. Znamenke su prikazane kao vektor binarnih znacajki (b&w pixeli) u matrici s rezolucijom 32x32. Odredite ukupan broj parametara naivnog bayesovog klasifikatora.

- Za svaki piksel moramo imat vjerojatnost P(x=1|y=k), odnosno ako je 1, on pripada klasi k. (Ako je 0, i pripada, ta vjerojatnost se racuna kao 1 - p(x=1|y=k). Piksela ima 32x32, a klasa 10, dakle treba nam tocno 32x32x10 = 10240 parametara (vjerojatnosti).

Isprobati Naivnog Bayesa na konkretnom primjeru. (TODO pretipkati prvi primjer u latex!!!!)

(k) Trenirati N.B. po tablici:

i	Mjesto	Otok	Smještaj	Prijevoz	$y^{(i)}$
1	Istra	da	privatni	auto	da
2	Kvarner	ne	kamp	bus	ne
3	Dalmacija	da	hotel	avion	da
4	Dalmacija	ne	privatni	avion	ne
5	Istra	ne	privatni	auto	da
6	Kvarner	ne	kamp	bus	ne
7	Dalmacija	da	hotel	auto	da

Izracunati MLE procjene i klasificirati primjere $x^{(1)}$=(Istra, ne, kamp, bus) i $x^{(2)}$=(Dalmacija, da, hotel, bus)

- Prvo, idu prior vjerojatnosti za klase:
  P(y=da) = 4/7
  P(y=ne) = 3/7
- Zatim posebno gledamo redove gdje je y=da i y=ne
- Prva varijabla, za svaku mogucu vrijednost:
  $P(x_1$=Istra | y=da) = 0.50 (2 / 4)
  $P(x_1$=Istra | y=ne) = 0.0 (0 / 3)
  $P(x_1$=Kvarner | y=da) = 0.0 (0 / 4)
  $P(x_1$=Kvarner | y=ne) = 0.666 (2 / 3)
  $P(x_1$=Dalmacija | y=da) = 0.50 (2 / 4)
  $P(x_1$=Dalmacija | y=ne) = 0.333 (1 / 3)
- Druga varijabla:
  $P(x_2$=da | y=da) = 0.75 (3 / 4)
  $P(x_2$=da | y=ne) = 0.0 (0 / 3)
  $P(x_2$=ne | y=da) = 0.25 (1 / 4)
  $P(x_2$=ne | y=ne) = 1.0 (3 / 3)
- Treca varijabla:
  $P(x_3$=privatni | y=da) = 0.50 (2 / 4)
  $P(x_3$=privatni | y=ne) = 0.33 (1 / 3)
  $P(x_3$=kamp | y=da) = 0.0 (0 / 4)
  $P(x_3$=kamp | y=ne) = 0.666 (2 / 3)
  $P(x_3$=hotel | y=da) = 0.50 (2 / 4)
  $P(x_3$=hotel | y=ne) = 0.00 (0 / 3)
- Cetvrta varijabla:
  $P(x_4$=auto | y=da) = 0.75 (3 / 4)
  $P(x_4$=auto | y=ne) = 0.0 (0 / 3)

P(x_4=bus | y=da) = 0.0 (0 / 4)
P(x_4=bus | y=ne) = 0.666 (2 / 3)
P(x_4=avion | y=da) = 0.25 (1 / 4)
P(x_4=avion | y=ne) = 0.333 (2 / 3) krivo ! (⅓)

- x=(Istra, ne, kamp, bus)
  P(x|da)P(y=da) = (0.5*0.25*0.0*0.0)*(4/7) = 0
  Uzeli smo ovdje istra|da y da * otok ne|da * kamp dada * bus dada * y da
  P(x|ne)P(y=ne) = (0.0*1.0*0.666*0.666)*(3/7) = 0
  excuse_me_wtf.jpg
  x=(Dalmacija, da, hotel, bus)
  P(x|da)P(da) = 0
  P(x|ne)P(ne) = 0
  Aww_hell_naw.png

(l) Sad isto to sve samo sa Laplacian smoothing.

- Umro sam dok sam sve ovo pretipkao u LaTeX: #MADRESPECT
  Objasnjenje: brojnik je uvijek + 1; nazivnik += broj klasa (broj razlicitih vrijednosti za varijablu)

Priors:
$P(y = da) = 5/9$
$P(y = ne) = 4/9$

Prva varijabla:
$P(x_1 = Istra|y = da) = 0.4286 \quad (\frac{2+1}{4+3})$
$P(x_1 = Istra|y = ne) = 0.1667 \quad (\frac{0+1}{3+3})$
$P(x_1 = Kvarner|y = da) = 0.1430 \quad (\frac{0+1}{4+3})$
$P(x_1 = Kvarner|y = ne) = 0.5000 \quad (\frac{2+1}{3+3})$
$P(x_1 = Dalmacija|y = da) = 0.4286 \quad (\frac{2+1}{4+3})$
$P(x_1 = Dalmacija|y = ne) = 0.3333 \quad (\frac{1+1)}{3+3})$

Druga varijabla:
$P(x_2 = da|y = da) = 0.6667 \quad (\frac{3+1}{4+2})$
$P(x_2 = da|y = ne) = 0.2000 \quad (\frac{0+1}{3+2})$
$P(x_2 = ne|y = da) = 0.3333 \quad (\frac{1+1}{4+2})$
$P(x_2 = ne|y = ne) = 0.8000 \quad (\frac{3+1}{3+2})$

Treca varijabla:
$P(x_3 = privatni|y = da) = 0.4286 \quad (\frac{2+1}{4+3})$
$P(x_3 = privatni|y = ne) = 0.3333 \quad (\frac{1+1}{3+3})$
$P(x_3 = kamp|y = da) = 0.1430 \quad (\frac{0+1}{4+3})$
$P(x_3 = kamp|y = ne) = 0.5000 \quad (\frac{2+1}{3+3})$
$P(x_3 = hotel|y = da) = 0.4286 \quad (\frac{2+1}{4+3})$
$P(x_3 = hotel|y = ne) = 0.1667 \quad (\frac{0+1}{3+3})$

Cetvrta varijabla:
$P(x_4 = auto|y = da) = 0.5714 \quad (\frac{3+1}{4+3})$
$P(x_4 = auto|y = ne) = 0.1667 \quad (\frac{0+1}{3+3})$
$P(x_4 = bus|y = da) = 0.1430 \quad (\frac{0+1}{4+3})$
$P(x_4 = bus|y = ne) = 0.5000 \quad (\frac{2+1}{3+3})$
$P(x_4 = avion|y = da) = 0.2857 \quad (\frac{1+1}{4+3})$
$P(x_4 = avion|y = ne) = 0.5000 \quad (\frac{2+1}{3+3})$

- x = (Istra, ne, kamp, bus)
  P(x|y=da)P(t=da) =
  = (0.4286*0.3333*0.1430*0.1430) * (0.5556) =
  = 1.623e-3
  P(x|y=ne)P(y=ne) =
  = (0.1667*0.8000*0.5000*0.5000)*(0.4444) =
  = 14.818e-3
  P(y=da|x) = 0.0982
  P(y=ne|x) = 0.9018

x=(Dalmacija, da, hotel, bus)
P(x|y=da)*P(y=da) =
= (0.4286*0.6667*0.4286*0.1430) * (0.5556) =
= 9.730e-3
P(x|y=ne)*P(y=ne) =
= (0.3333*0.2000*0.1667*0.5000) * (0.4444) =
= 2.469e-3
p(y=da|x) = 0.7976
p(y=ne|x) = 0.2024

Razumjeti definiciju uzajamne informacije i nacin njezina izracuna.

(m)Krenuvsi od definicija za entropiju i relativnu entropiju, izvedite mjeru uzajamne informacije I(X,Y) kao KL divergenciju izmedu zajednicke razdiobe, P(X,Y), i zajednicke razdiobe uz pretpostavku nezavisnosti P(X)*P(Y)

- S wiki:

$$H = -\sum_{i} p_i \log_2 (p_i)$$

$$\begin{aligned} \mathrm{I}(X;Y) &\equiv \mathrm{H}(X) - \mathrm{H}(X|Y) \\ &\equiv \mathrm{H}(Y) - \mathrm{H}(Y|X) \end{aligned}$$

$$H(p, q) = -\sum_{x \in \mathcal{X}} p(x) \log q(x) \quad \text{(Eq.1)}$$

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right).$$

$$\begin{aligned} D_{\mathrm{KL}}(P \parallel Q) &= -\sum_{x \in \mathcal{X}} p(x) \log q(x) + \sum_{x \in \mathcal{X}} p(x) \log p(x) \\ &= \mathrm{H}(P, Q) - \mathrm{H}(P) \end{aligned}$$

$$\mathrm{I}(X;Y) = D_{\mathrm{KL}} \left(p_{(X,Y)} \parallel p_X p_Y \right)$$

- Jos jedan zanimljivi (subjektivno) teorem:

$$\mathrm{I}(X;Y) = \mathbb{E}_Y \left[D_{\mathrm{KL}} \left(p_{X|Y} \parallel p_X \right) \right]$$

Dakle ne samo zajednicka vs nezavisna, vec uvjetna vs obicna

(n) Izracunajte mjeru uzajamne informacije I(X;Y) za varijable X i Y s razdiobom definiranom u zadatku 1 u ovoj zadaci. Biste li, temeljem vrijednosti uzajamne informacije, rekli da su varijable X i Y nezavisne?

```
>>> import numpy as np
>>> X = np.array([[0.55, 0.45]])
>>> Y = np.array([[0.25, 0.35, 0.4]])
>>> X_Y = np.array([
... [0.2, 0.05, 0.3],
... [0.05, 0.3, 0.1]])
>>> X.T @ Y    # Ovo je nezavisna distribucija P(X,Y) = P(X) * P(Y)
array([[0.1375, 0.1925, 0.22  ],
       [0.1125, 0.1575, 0.18  ]])
>>> I = -np.sum(X_Y / np.log(X.T @ Y))
>>> I
0.57278960345526
```

- X = [[1, 2], [0.55, 0.45]], Y = [[1, 2, 3], [[0.25 0.35 0.4]]
- P(1, 1) = 0.2, P(1, 2)=0.05, P(1, 3) = 0.3, P(2, 1) = 0.05, P(2, 2) = 0.3, P(2, 3) = 0.1.

- Hvala python3/numpy
- Cini se da su varijable zavisne.
- Zna tko kako bi ovo na papiru?

(o) Uzajamna informacija nije odozgo ogranicena, ali je ogranicena odozdo. Primjenom Jensenove nejednakosti, dokazite da vrijedi I(X,Y) >= 0.

- Pisano u Latexu jer mi se sve manje svida ovaj gdocs

Jensenova nejednakost (za teoriju vjerojatnosti) kaze sljedece: ($\phi(x)$ je konveksna funkcija)

$$\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)]$$
$$\mathbb{E}[X] = \sum_x p(x)x$$
$$\mathbb{E}[\phi(x)] = \sum_x p(x)\phi(x)$$

Funkcija $f(x)$ je konveksna na nekom intervalu I ako vrijedi:

$$\frac{d^2}{dx^2} f(x) \geq 0, \forall x \in I$$

Pretpostavimo da je nasa $\phi(x) = -ln(x)$.
$\phi(x)$ je konveksna:

$$\frac{d^2}{dx^2}(-ln(x)) = \frac{d}{dx}(-\frac{1}{x}) = \frac{1}{x^2} > 0, \forall x \in < 0, \infty >$$

E sad s tim znanjem napadnemo transinformaciju. Krenimo od definicije:

$$-D_{KL}(P||Q) = \sum_x p(x) ln\left(\frac{q(x)}{p(x)}\right)$$

Minus ispred D smo progutali u zamjenu poretka u logaritmu. Dalje, iskoristimo $TheNejednakost^{TM}$:

$$\sum_x p(x) ln\left(\frac{q(x)}{p(x)}\right) \leq ln\left(\sum_x p(x)\frac{q(x)}{p(x))}\right) = ln\left(\sum_x q(x)\right) = ln(1) = 0$$

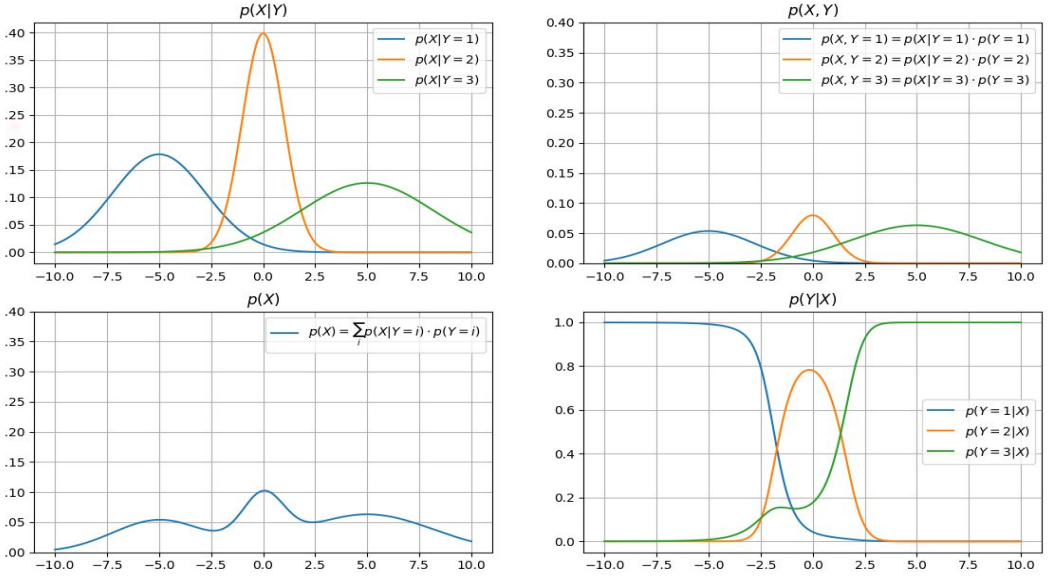
Dakle imamo da je $-D_{KL}(P||Q) \leq 0$, odnosno

$$D_{KL}(P||Q) \geq 0, \forall P, Q$$

Quod erat demonstratum.

Razviti intuiciju za model kontinuiranog Bayesovog klasifikatora.
Izradujemo Bayesov model za klasifikaciju primjera iz X=R u 3 klase. Imamo sljedece parametre: P(y=1) = 0.3, P(y=2)=0.2, u1=-5, u2=0, u3 = 5, s21 = 5, s22=1 s23=10. Skicirajte p(x|y), p(x, y), p(x) i p(y|x).

- Slikica:



- Kod koji to crta:
  Gaussian je obicna implementacija gaussove distribucije
  Same plot funkcije nisu bitne :) bitno je kak se izracunavaju funkcije distribucija

```
P_1 = 0.3
P_2 = 0.2
P_3 = 1 - P_1 - P_2

u1 = -5
u2 = 0
u3 = 5

sigma2_1 = 5
sigma2_2 = 1
sigma2_3 = 10

p_x_given_y1 = gaussian(u1, sigma2_1)
p_x_given_y2 = gaussian(u2, sigma2_2)
p_x_given_y3 = gaussian(u3, sigma2_3)

p_x_y1 = lambda x: P_1*p_x_given_y1(x)
p_x_y2 = lambda x: P_2*p_x_given_y2(x)
p_x_y3 = lambda x: P_3*p_x_given_y3(x)

p_x = lambda x: p_x_y1(x) + p_x_y2(x) + p_x_y3(x)

p_y1_given_x = lambda x: p_x_y1(x) / p_x(x)
p_y2_given_x = lambda x: p_x_y2(x) / p_x(x)
p_y3_given_x = lambda x: p_x_y3(x) / p_x(x)

plot_x_given_y([p_x_given_y1, p_x_given_y2, p_x_given_y3])
plot_x_y([p_x_y1, p_x_y2, p_x_y3])
plot_x(p_x)
plot_y_given_x([p_y1_given_x, p_y2_given_x, p_y3_given_x])
plt.show()
```

Razumjeti izvod modela kontinuiranog bayesovog klasifikatora i osvjeziti potrebno znanje matematike

(p) Krenuvsi od izraza (4.29) iz skripte, izvedite model visedimenzijskog Bayesovog klasifikatora s kontinuiranim ulazima s djeljenom i dijagonalnom kovarijacijskom matricom.

- (4.29):

$$h_j(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j - \frac{1}{2}\boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j + \ln P(\mathcal{C}_j)\,.$$

Preskace se korak di se mnoze vjerojatnosti svake varijable posebno, pa se onda umnozak eksponencijala svede na zbroj eksponenata.

2. pojednostavljenje: dijagonalna kovarijacijska matrica

Daljnje pojednostavljenje modela moguće je uz pretpostavku da varijable nisu korelirane odnosno da su nezavisne.[5] U tom slučaju koristimo dijagonalnu kovarijacijsku matricu, $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_i^2)$. Matrica $\boldsymbol{\Sigma}^{-1}$ onda je također dijagonalna te vrijedi $\boldsymbol{\Sigma}^{-1} = \mathrm{diag}(1/\sigma_i^2)$ i $|\boldsymbol{\Sigma}| = \prod_i \sigma_i$. Multivarijatna Gaussova gustoća (4.24) degenerira u produkt univarijatnih Gaussovih razdiobi:

$$p(\mathbf{x}|\mathcal{C}_j) = \frac{1}{(2\pi)^{n/2}\prod_{i=1}^{n}\sigma_i}\exp\Big\{-\frac{1}{2}\sum_{i=1}^{n}\Big(\frac{x_i-\mu_{ij}}{\sigma_i}\Big)^2\Big\} \qquad (4.30)$$

$$= \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma_i}\exp\Big\{-\frac{1}{2}\Big(\frac{x_i-\mu_{ij}}{\sigma_i}\Big)^2\Big\}\,.$$

(q) Napisite broj parametara ovog modela.

- Ako je K broj klasa, a n dimenzija vektora x, treba nam Kn srednjih vrijednosti (svaka klasa ima n srednjih vrijednosti) i n varijanci (svaka dimenzija je neovisna o drugoj i ima svoju varijancu) i K-1 (P(Cj)).

(r) Objasnite zasto je izglednost faktorizirana u produkt univarijatnih razdioba, sto odgovara pretpostavci o uvjetnoj nezavisnosti, premda znacajke mogu biti nelinearno zavisne.

Nije bitno, nisu linearno zavisne. (HELP WANTED)

Zato da model bude jednostavniji, a time i manje sklon prenaučenosti. Umjesto O(n^2) dobijamo O(n)

8. [Svrha: Razviti intuiciju za složenost modela kontinuiranog Bayesovog klasifikatora i shvatiti kako se problem u konačnici svodi na odabir optimalnog modela.]
Želimo izgraditi klasifikator za klasifikaciju brucoša u jednu od dvije klase:

y = 1 => “Završava FER u roku” i y = 2 “Produljuje studij”.

Svaki je primjer opisan sa šest ulaznih varijabli: prosjek ocjena 1.–4. razreda (četiri varijable), bodovi državne mature iz matematike te bodovi državne mature iz fizike. Raspolažemo trima modelima: modelom H1 s dijeljenom kovarijacijskom matricom, modelom H2 s dijagonalnom (i dijeljenom) kovarijacijskom matricom i modelom H3 s izotropnom kovarijacijskom matricom.

(a) Koliko svaki od ova tri modela ima parametara?

- Prvi broj predstavlja broj parametara za kov. matricu, drugi broj parametara za vektore prototipa, a treći (K-1) broj par. za apriorne vjerojatnosti klasa
- H1: $\frac{n}{2}(n+1) + K*n + K - 1 = 34$
- H2: $n + K*n + K - 1 = 19$
- H3: $1 + K*n + K - 1 = 14$
- Može netko napisati koje su vrijednosti n a koje K ovdje?

(b) Za koji od ova tri modela očekujete da će najbolje generalizirati u ovom konkretnom slučaju (uzmite u obzir prirodu problema i očekivane odnose između značajki)? Zašto?

- H1 modelira međukolinearnost, a budući da su značajke kolinearne (netko tko je imao dobre ocjene vjerojatno će dobro napisati i maturu) moglo bi doci do loše kondicioniranosti matrice Σ^{-1} . Također ima puno parametara i sklon je prenaučenosti
- H2 ne modelira međukolinearnost zbog čega ga je teže prenaučiti nego H1, a koristi normaliziranje euklidske udaljenosti zbog čega je neosjetljiv na razlike u varijanci između pojedinih dimenzija, za razliku od H3
- H3 sve varijance modelira jednakima što može lako dovesti do podnaučenosti. Velika je razlika u skalama ocjena (1-5) i bodova na maturi (1-100)
- Vjerojatno će H2 najbolje generalizirati

(c) Nacrtajte skicu funkcije empirijske pogreške i pogreške generalizacije i naznačite na njoj točke koje označavaju navedenim trima modelima.

- Ona uobičajena skica za očekivanja, H3 na podnaučenosti, H2 optimum, H1 prenaučenost

(d) Kako biste u praksi odredili koji ćete model upotrijebiti?

- Unakrsnom provjerom

9. [Svrha: Razumjeti vezu izmedu Bayesovog klasifikatora i logističke regresije, odnosno probabilističku interpretaciju logističke regresije. Razumjeti razliku u broju parametara izmedu diskriminativnog i generativnog modela te utjecaj broja klasa i broja primjera na taj odnos.]

(a) Izvedite model logističke regresije krenuvši od generativne definicije za P(y =1|x). Izvod napravite korak po korak te se uvjerite da možete obrazložiti svaki korak u izvodu. Napišite sve pretpostavke koje ste ugradili u izvod.

- Ideja: pokazati da logistička regresija i Bayesov klasifikator izračunavaju isti $P(y|\mathbf{x})$
- Model **logističke regresije**:

$$h(\mathbf{x};\mathbf{w}) = P(y=1|\mathbf{x}) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x})$$

- Aposteriorna vjerojatnost za **kontinuirani Bayesov klasifikator** (za dvije klase):

$$P(y=1|\mathbf{x}) = \frac{p(\mathbf{x}|y=1)P(y=1)}{p(\mathbf{x}|y=1)P(y=1)+p(\mathbf{x}|y=2)P(y=2)} = \frac{1}{1+\frac{p(\mathbf{x}|y=2)P(y=2)}{p(\mathbf{x}|y=1)P(y=1)}} =$$
$$= \frac{1}{1+\exp\left(\ln\frac{p(\mathbf{x}|y=2)P(y=2)}{p(\mathbf{x}|y=1)P(y=1)}\right)} = \frac{1}{1+\exp(-\alpha)} = \sigma(\alpha)$$

gdje

$$\alpha = \ln\frac{p(\mathbf{x}|y=1)P(y=1)}{p(\mathbf{x}|y=2)P(y=2)} = \underbrace{\ln p(\mathbf{x}|y=1)P(y=1)}_{h_1(\mathbf{x})} - \underbrace{\ln p(\mathbf{x}|y=2)P(y=2)}_{h_2(\mathbf{x})}$$

- Možemo li α prikazati kao linearnu kombinaciju težina, $\alpha = \mathbf{w}^{\mathrm{T}}\mathbf{x}$?
- Da, ako pretpostavimo **dijeljenu kovarijacijsku matricu**:

$$\alpha = h_1(\mathbf{x}) - h_2(\mathbf{x})$$
$$= \mathbf{x}^{\mathrm{T}}\underbrace{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}_{\mathbf{w}} \underbrace{-\frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\frac{P(y=1)}{P(y=2)}}_{w_0} = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$$

$\Rightarrow$ logistička regresija istovjetna je Bayesovom klasifikatoru s dijeljenom $\boldsymbol{\Sigma}$

(b) Model logističke regresije koristimo za binarnu klasifikaciju primjera s n = 100 značajki. Odredite broj parametara modela logističke regresije te njemu odgovarajućeg generativnog modela.

- Logistička regresija: $n + 1 = 101$
- Bayesov klasifikator: $\frac{n}{2}(n+1) + Kn + K - 1 = 5251$

(c) Izračunajte broj parametara za isti slučaj, ali sa K = 5 klasa.

- Logistička regresija uz OVR shemu višeklasne klasifikacije: $K(n+1) = 505$
- Bayesov klasifikator: $\frac{n}{2}(n+1) + Kn + K - 1 = 5554$

(d) Pretpostavite da klasificiramo u K = 10 klasa. Izračunajte koliko velika mora biti dimenzija prostora značajki n, a da bi se logistička regresija isplatila jer ima manje parametara od odgovarajućega generativnog modela.

- To je moguće samo ako radimo OVO shemu višeklasne klasifikacije:

  $\frac{10*9}{2}(n+1) = \frac{n}{2}(n+1) + nK + K - 1$

  Rješavanje ovog sustava dovodi do rješenja n = 70.

9. Probabilisticki grafički modeli

1. Razumjeti sto je to probabilisticki graficki model. Shvatiti specificnosti modela Bayesove mreze te kako taj model predstavlja zajednicku distribuciju. Shvatiti koje induktivne pretpostavke ovakva reprezentacija koristi.

- Navedite tri osnovna aspekta svakog probabilistickog grafickog modela (PGM).

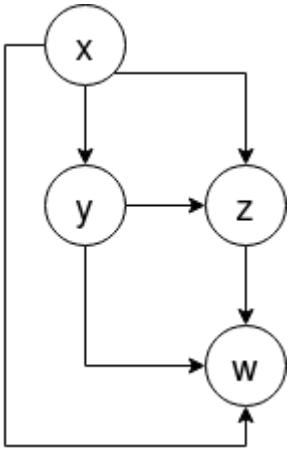
  Reprezentacija (je li model usmjereni graf (Bayes) ili neusmjeren (Markov)
  Zakljucivanje
  Ucenje

- Je li PGM parametarski ili neparametarski model? Je li generativni ili diskriminativni? Obrazlozite odgovore.

  Parametarski jer mu broj parametara ne ovisi o broju primjera vec o grafu kojeg sami postavimo i ostaje fiksan.
  Generativni s obzirom da ne uci decizijsku granicu vec uci modelirati zajednicku distribuciju.

- Pretpostavite zajednicku distribuciju cetiriju varijabli xyzw. Faktorizirajte ovu distribuciju primjenom osnovnih pravila vjerojatnosti te skicirajte Bayesovu mrezu koja odgovara toj faktorizaciji.

  P(x, y, z, w) = P(x)P(y|x)P(z|x, y)P(w|x,y,z)

- Ponovite isto, ali ovaj put pretpostavljajuci $x \perp z|w$ i $y \perp z|w$. Kojoj vrsti induktivnih pretpostavki odgovaraju ove pretpostavke o nezavisnosti? Obrazlozite motivaciju za uvodenjem dodatnih pretpostavki u model.

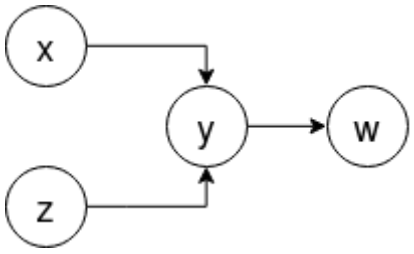
  P(x,y,z,w) = P(w)P(z|w)P(x| z,w)P(y|x,z,w) = P(w)P(z|w)P(x|w)P(y|x,w)
  Pretpostavke o nezavisnosti pripadaju pristranosti ograničenja (?)
  Dodatne pretpostavke se uvode u model kako bi mu se smanjila složenost

- Formalno definirajte uredajno Markovljevo svojstvo i topoloski uredaj cvorova mreze. Primjenom uredajnog Markovljevog svojstva izvedite uvjetne nezavisnosti kodirane Bayesovom mrezom koja odgovara faktorizaciji $P(x,y,z,w) = P(x)P(y|x,z)P(z)P(w|y)$.

  Uređajno Markovljevo svojstvo: x_k je uvjetno nezavisna od svih varijabli koje joj prethode, osim "roditeljskih" varijabli, odnosno direktnih prethodnika: $x_k \perp (pred(x_k) \setminus parent(x_k)) \mid parent(x_k)$

  Topološki uređaj: čvorovi poredani tako da roditelji dolaze prije djece

  Graf:

  Možemo odabrati jedan od logičnih poredaka: x, z, y, w ili z, x, y, w -> odabiremo x, z, y, w

  $x \perp (\{\} \setminus \{\}) \mid \{\}$

  $z \perp (\{x\} \setminus \{\}) \mid \{\} \rightarrow z \perp x \mid \{\}$

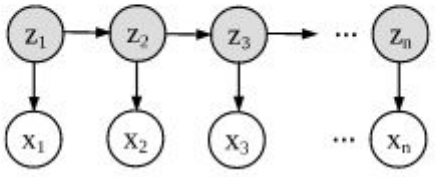
  $y \perp (\{x,z\} \setminus \{x,z\}) \mid \{x,z\}$

  $w \perp (\{x,z,y\} \setminus \{y\}) \mid \{y\}$ -> $w \perp x \mid y$ i $w \perp z \mid y$

- Nacrtajte Bayesovu mrezu HMM i napisite pripadnu faktorizaciju zajednicke vjerojatnosti p(x,z). Koja je svrha latentnih varijabli z i koje su uvjetne nezavisnosti kodirane ovom mrezom?

  Slikica:

  $$p(\mathbf{x},\mathbf{z}) = p(z_1)p(x_1|z_1)\prod_{k=2}^{n} p(z_k|z_{k-1})p(x_k|z_k)$$

  Latentne varijable indirektno modeliraju dulje zavisnosti

  Kao poredak uzimamo: $z_1, x_1, z_2, x_2, \ldots, z_n, x_n$

  $z_1 \perp (\{\} \setminus \{\}) \mid \{\}$

  $x_1 \perp (\{z_1\} \setminus \{z_1\}) \mid \{z_1\}$

  $z_2 \perp (\{z_1,x_1\} \setminus \{z_1\}) \mid \{z_1\}$ -> $z_2 \perp x_1 \mid z_1$

  $x_2 \perp (\{z_1,x_1,z_2\} \setminus \{z_2\}) \mid \{z_2\}$ -> $x_2 \perp z_1 \mid z_2$, $x_2 \perp x_1 \mid z_2$

…

Odnosno, $z_n \perp x_1, ..., x_{n-1}, z_1, ..., z_{n-2} \mid z_{n-1}$ i $x_n \perp x_1, ..., x_{n-1}, z_1, ..., z_{n-1} \mid z_n$

2. Gradimo Bayesovu mrežu koja predviđa hoće li student/ica uspješno položiti SU. Mreža sadrži pet varijabli: pohađa li osoba konzultacije (x_1), je li osoba dobra u pythonu (x_2), rješava li osoba samostalno domaće zadaće i labose (x_3), ocjenu iz predmeta UI (x_4) te varijablu koja govori je li osoba položila SU (y). Pritom vrijedi x_1, x_2, x_3, x_5 ε {T, F}, i x_1 = {2,3,4,5}. P(x_1 =T) = 0.2, P(x_2 =T)=0.6. Dane su i tablice uvjetnih vjerojatnosti:

x_1	x_2	$P(x_3 = \top)$
$\perp$	$\perp$	0.3
$\perp$	$\top$	0.5
$\top$	$\perp$	0.8
$\top$	$\top$	0.9

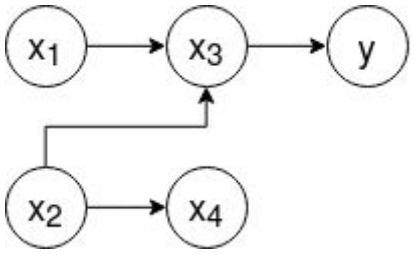
x_3	$P(y = \top)$
$\perp$	0.2
$\top$	0.9

x_2	$P(x_4 = 2)$	$P(x_4 = 3)$	$P(x_4 = 4)$	$P(x_4 = 5)$
$\perp$	0.4	0.2	0.3	0.1
$\top$	0.2	0.1	0.1	0.6

-

- Skicirajte bayesovu mrežu ako je faktorizacija zajedničke distribucije sljedeća $P(x_1, x_2, x_3, x_4, y) = P(x_1)P(x_2)P(x_3|x_1, x_2)P(x_4|x_2)P(y|x_3)$.

  Cak i bez faktorizacije se moze dobit graf

- Koji je ukupan broj parametara ove mreze?

  Kada bi se računala zajednička vjerojatnost, broj parametara bi bio:
  2*2*2*4*2 = 64
  U slučaju mreže :
  x_1 (2) + x_2 (2) + x_3 (4 kombinacije x_1, x_2 * 2) + (2 kombinacije x_2 * 4) + (4 kombinacije x_3 * 2) = 28

- Postupkom egzaktnog zakljucivanja izračunajte P(y=T| x_1 =T, x_4 =3).
  - Zna li itko?
- Koja je razlika između posteriornog i MAP upita? O kakvom tipu upita se radi u prošlom zadatku? Obrazložite.

  Razlika je u (programeri bi rekli) povratnom tipu. Posteriori vraca cijelu distribuciju a MAP upit vraca onu vrijednost za koju je vjerojatnost najveća.

To je bio posterori upit jer nas je zanimala vjerojatnost. Da nas je zanimala "Koja vrijednost y je najveca ako imamo x1=T,x4=4", to bi bio MAP.

- Utjece li broj varijabli u mrezi na ucinkovitost zakljucivanja? Zasto?
  Utječe, što je veći n, to je teže egzaktno zaključivati te dolazimo do kombinatorne eksplozije broja varijabli -> NP- složen problem
- Objasnite ideju približnog zaključivanja uzorkovanjem. Koja je prednost tog postupka? U kratkim crtama objasnite kako biste uzorkovali P(x_1, x_2, x_3, x_4, y) koristeći unaprijedno uzorkovanje.
  - PRIBLIŽNO ZAKLJUČIVANJE UZROKOVANJEM
    Uzorkujemo varijablu x iz njezine distribucije P(x) i onda kad imamo takav uzorak veličine N, možemo izračunati očekivanje bilo koje vrijednosti varijable x kao relativnu frekvenciju te vrijednosti u uzorku:

$$P(\mathbf{x} = x) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{\mathbf{x} = x\}$$

3.1 Unaprijedno uzorkovanje

To je, dakle, osnovna ideja. U našem slučaju $P(\mathbf{x})$ je definirana pomoću Bayesove mreže. Sada je pitanje: kako možemo uzorkovati iz tako prikazane distribucije? Prva ideja koja se nameće jest da jednostavno uzorkujemo po strukturi Bayesove mreže, krenuvši od roditeljskih čvorova prema čvorovima djeci. Konkretno, krenemo od prvog čvora u mreži (prvih po topološkom uređaju) i slučajno generiramo, sukladno distribuciji tog čvora, jednu vrijednost za tu varijablu. Zatim idemo na iduću varijablu po topološkom uređaju, generiramo slučajnu vrijednost za nju. Ako je ta varijabla uvjetovana prvom varijablom, onda naravno uzimamo to u obzir kod generiranja, tj. kod čvorova djece generiramo iz uvjetne distribucije s pravilno postavljenim vrijednostima varijabli u čvorovima roditelja. Budući da idemo topološkim uređajem, imamo zajamčeno da ćemo u svakom čvoru već imati generirane vrijednosti za sve varijable roditelja. Naposlijetku, kada tako obiđemo cijelu mrežu, imat ćemo jedan slučajan vektor iz zajedničke distribucije. Ponavljanjem ovog postupka generiramo niz ovakvih vektora, i oni čine naš uzorak. Ovaj postupak zove se **unaprijedno uzorkovanje (engl. forward sampling)**.

3. Razumjeti ideju D-odvojivosti i kako se ona moze provesti grafički. Shvatiti motivaciju iza ispitivanja uvjetne nezavisnosti parova varijabli.
   - Zašto bismo htjeli znati koji parovi varijabli su uvjetno nezavisni? Nije li ta informacija već kodirana unutar strukture mreze? Objasnite.
     - Želimo znati koji parovi varijabli su uvjetno nezavisni jer nam to smanjuje složenost modela. Ta informacija je već implicitno kodirana u strukturu mreže.

- Formalno definirajte d-odvojivost i objasnite koji uvjeti (i kada) moraju vrijediti da bi neke dvije varijable bile uvjetno nezavisne

  Raspolažemo skupom varijabli E koje su opažene.

  Za **stazu** P od čvora x do čvora y kažemo da je **d-odvojena (d-separated)** akko vrijedi **barem jedno** od sljedećeg:

  - P sadrži **lanac** $x \to z \to y$ ili $x \leftarrow z \leftarrow y$ i $z \in E$
  - P sadrži **račvanje** $x \leftarrow z \to y$ i $z \in E$
  - P sadrži **sraz** $x \to z \leftarrow y$ i varijabla z **nije** u E i nijedan sljedbenik od z nije u E

  Za **par čvorova** x i y kažemo da su čvorovi x i y d-separirani za dani E ako su **sve staze** između ta dva čvora d-separirane za dani E.

  Čvorovi x i y su d-separirani za dani E **akko** su uvjetno nezavisni za dani E.

- Na temelju Bayesove mreze iz zadatka 2., odredite pod kojim uvjetima su varijable prolaza SU(y) i ocjene iz predmeta UI uvjetno nezavisne
  - Traži se nezavisnost između x_4 i y. Te varijable su uvj. nezavisne ako vrijedi:
    (1) Postoji lanac x_4 -> ? -> y ili obrnuto: NE
    (2) Postoji račvanje x_4 <- ? -> y: postoji, ali preko x_3 pa mislim da NE
    (3) Postoji sraz x_4 -> ? <- y: NE
    - Mislim da ovo nije dobro. Ima li netko možda logičnije objašnjenje?
    - Ovo je možda isto pogrešno razmišljanje, ali mislim da na putu između y i x4 imamo jedno račvanje (x3 <- x2 -> x4) i jedan lanac (x2 -> x3 -> y) i da bi y i x4 bile nezavisne, moramo opažati x2 i moramo opažati x3 jer su nam onda ovo račvanje i lanac d-separirani
- Svojim riječima objasnite efekt objašnjavanja (explaining away) koristeći za primjer varijable x_1 , x_2 i x_3
  - Uzmimo za primjer slučaj da imamo visoku temperaturu (x_3) i kao uzrok promatramo mogućnosti da je u pitanju mononukleoza (x_1) ili upala grla (x_2): $P(x_1, x_2, x_3) = P(x_3 \mid x_1, x_2)$. Ukoliko znamo da imamo upalu grla odmah je manja vjerojatnost da imamo mononukleozu jer objašnjava visoku temperaturu. Zato vrijedi:

    $$p(x|z) \neq p(x|y,z) \;\Leftrightarrow\; x \not\perp y|z$$

    , x_1 = x, x_2 = y, x_3 = z

4. Razumjeti učenje Bayesovih mreza i njegovu povezanost s procjenom parametara. Znati kako pristupiti učenju modela ako su podaci nepotpuni.
   - Što su parametri Bayesove mreže i na koji način ih učimo iz podataka?
     - Parametar je θ, parametar distribucije koja opisuje podatke. Učenje se svodi na procjenu tog parametra što se radi procjeniteljima, npr. MAP ili MLE.

- Izvedite log-izglednost (proizvoljne) Bayesove mreze. Objasnite zašto je moguće procjenjivati parametre svakog čvora zasebno.

$$
\begin{aligned}
\ln \mathcal{L}(\boldsymbol{\theta}|\mathcal{D}) &= \ln p(\mathcal{D}|\boldsymbol{\theta}) \\
&= \ln p(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}|\boldsymbol{\theta}) \\
&= \ln \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\boldsymbol{\theta}) \\
&= \ln \prod_{i=1}^{N} \prod_{k=1}^{n} p(x_k^{(i)}|\mathrm{pa}(x_k^{(i)}), \boldsymbol{\theta}_k) \\
&= \ln \prod_{k=1}^{n} \prod_{i=1}^{N} p(x_k^{(i)}|\mathrm{pa}(x_k^{(i)}), \boldsymbol{\theta}_k) \\
&= \sum_{k=1}^{n} \sum_{i=1}^{N} \ln p(x_k^{(i)}|\mathrm{pa}(x_k^{(i)}), \boldsymbol{\theta}_k)
\end{aligned}
$$

Vidimo da se log-izglednost **dekomponirala** prema strukturi grafa Bayesove mreže, a to je dobro jer to znači da možemo procijeniti parametre svakog čvora mreže nezavisno.

- Objasnite sto to znači da neki model ima skrivene (latentne) varijable. Kako one utječu na postupak učenja modela?
  - Znači da ima varijable koje nisu opažene. One mogu biti korisne (varijable upita, query, x_q) ili beskorisne (varijable smetnje, nuissance, x_n).

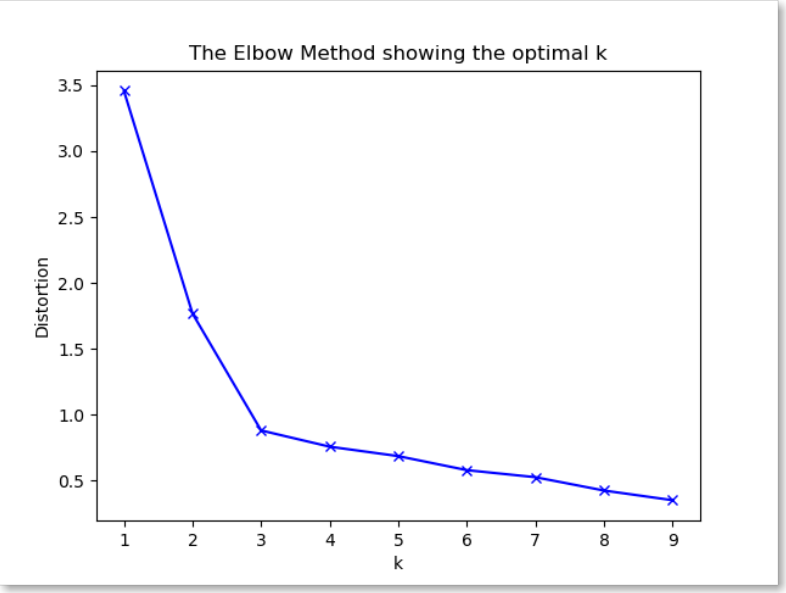
- Modeli sa skrivenim varijablama (npr., HMM, GMM) ⇒ tzv. **nepotpuni podatci**
  - Log-izglednost se ne dekomponira po strukturi grafa

  ⇒ MLE nema rješenje u zatvorenoj formi
  - Učenje pomoću **algoritma maksimizacije očekivanja** ili **gradijentnim usponom**

9. Grupiranje

Algoritam k-sredina minimizira kriterij pogreske J(u1, …, uk | D). Vrijedniost tog kriterija ovisi o broju grupa K, koji je unaprijed postavljen, te o polozajima sredista, koja se mjenjaju kroz iteracije.

(a) Nacrtajte skicu vrijednosti kriterija pogreske J kao funkcije broja grupa K. Koja je minimalna vrijednost funkcije J i zasto?

- 

  Minimalna vrijednost, 0.0, se postize kad je broj klasa K jednak broju primjera. Tada je svaka tocka sama sebi centroid i svaka udaljenost je 0.

(b) Izaberite na skici iz zadatka (a) tri vrijednosti za K i skicirajte na jednom grafikonu vrijednost kriterija pogreske J kao funkcije broja iteracija (tri krivulje).

- Uzmimo npr K = {3, 5, 7}. Graf ce izgledati kao u a), no imat ce tri krivulje: najvisu K=3, srednju K=5 i donju K=7. S brojem iteracija grupacija primjera ce biti sve bolja i zato ce s brojem iteracija vrijednost pogreske J opadati i doci do neke vrijednosti ($J_3^* > J_5^* > J_7^*$)

(c) Izaberite na skici iz zadatka (a) jednu vrijednost za K. Skicirajte na jednom grafikonu vrijednosti kriterija pogreske J kao funkcije broja iteracija, ali ovaj put uzevsi u obzir stohasticnost uslijed slucajnog odabira pocetnih sredista (nacrtajte nekoliko mogucih krivulja na istom grafikonu). Koje od tih kivulja su izglednije za algoritam kmeans++?

- Zbog toga što algoritam K-sredina ovisi o početnoj točki, neće nužno pronaći globalni optimum, ali će uvijek završiti u nekom lokalnom optimumu. K-means++ ne odabire točke nasumično, već nakon nasumičnog odabira prve točke za svaki sljedeći centroid bira točku koja ima maksimalnu udaljenost do najbližeg primjera. Na taj način puno bolje odabire početne centroide i zbog toga i sam algoritam brže

konvergira. Od ovih krivulja bi za k-means++ bile izglednije one koje prije dođu do optimuma i čiji optimumi imaju nižu pogrešku.

Isprobati rad algoritma KMeans i KMedioids na konkretnom primjeru. Shvatiti da je složenost ovog drugog puno nepovoljnija.

Raspolazemo skupom neoznačenih primjera {a(5, 2), b(7, 1), c(1, 4), d(6 2), e(2, 8), f(3, 6), g(0, 4)}.

Udaljenost je euklidska (ovdje ju kvadriram jer mi se ne da korijenovat)

(d) Izvedite jedan korak algoritma k-sredina uz K=3. Za pocetna sredista odaberite b, c i e.

- D(b, a) = 5, D(b, d) = 2, D(b, f) = 41, D(b, g) = 58
- D(c, a) = 20, D(c, d) = 29, D(c, f) = 29, D(c, g) = 1.
- D(e, a) = 45, D(e, d) = 52, D(e, f) = 5, D(e, g) = 20.
- a i d su najblizi b, f je najbliza e, g je najbliza c.
- Novi centroidi (prosjek komponenti po klasteru): u1=(6, 1.666), u2=(2.5, 7), u3=(0.5, 4)

(e) Izvedite jedan korak algoritma k-medioida uz K=3. Pocetna sredista su opet b, c i e.

- Prvi korak je gotovo isti kao kod k-means. Za svaku točku koja nije medoid se u vektor b zapisuje kojem medoidu je najbliža (one-hot encoding). Razlika u algoritmima je u tome da se u drugom koraku umjesto računanja novog centroida točaka koje su zajedno grupirane, za svaki od K medoida računa koja bi od (N-K) točaka mogla sa ostalih (N-K) točaka dati manju sumu mjere. Ukoliko je takva točka pronađena, trenutni medoid se zamjenjuje tom točkom.

(f) Usporedite računalnu složenost ova dva algoritma!

- K-means u prvom koraku mora za svaku od N točaka provjeriti za svaku od n značajki izračunati udaljenost do centroida svake od K klasa, a u drugom za svaku od N točaka za svaku od n značajki izračunati vrijednost sljedećeg centroida. $T(O(nNK) + O(nN)) = O(TnNK)$, gdje je T broj iteracija
- K-medoids u prvom koraku mora za svaku od N-K točaka koje nisu medoidi provjeriti kojoj od K klasa je najbliže, a u drugom za svaki od K medoida računa koja bi od (N-K) točaka mogla sa ostalih (N-K) točaka dati manju sumu mjere. $T(O(K(N-K) + O(K(N-K)^2)) = O(TK(N-K)^2)$, gdje je T broj iteracija

(g) Sto su prednosti, a sto nedostaci algoritma k-medioida?

- Prednost je sto rade na podacima koji ne moraju biti iz vektorskog prostora (dakle grafovi, stringovi n shit) i mogu koristiti različite mjere za udaljenost do medoida, a nedostaci to što nema efikasne implementacije pa je složenost velika.

3. [Svrha: Razumjeti model miješane gustoće i razlog zašto maksimizacija logizglednost nije analitički rješiva. Razumjeti kako uvođenje latentnih varijabli rješava taj problem. Razumjeti, na općenitoj razini, E-korak i M-korak. Razumjeti rad algoritma kao maksimizacije log-izglednosti i razumjeti kako ishod ovisi o broju grupa i početnoj inicijalizaciji.]

Algoritam maksimizacije očekivanja (EM-algoritam), kada se koristi za grupiranje, zapravo je poopćenje algoritma k-sredina.

(a) Sto je prednost, a što nedostatak, algoritma maksimizacije očekivanja u odnosu na algoritam k-sredina?

- EM-algoritam koristi meko grupiranje gdje imamo probabilistički izlaz i zato primjer može pripadati u više grupa. Zbog toga je moguće računati potpunu log-izglednosti, ali je značajno složeniji u odnosu na algoritam K-means.

(b) Napišite izraz za gustoću p(x) za model miješane gustoće (bez latentnih varijabli) i izraz za pripadnu (nepotpunu) log-izglednost.

- $$p(\mathbf{x}) = \sum_{k=1}^{K} p(\mathbf{x}, y = k) = \sum_{k=1}^{K} P(y = k) p(\mathbf{x}|y = k) = \sum_{k=1}^{K} \pi_k \, p(\mathbf{x}|\boldsymbol{\theta}_k)$$
- $$\ln \mathcal{L}(\boldsymbol{\theta}|\mathcal{D}) = \ln \prod_{i=1}^{N} p(\mathbf{x}^{(i)}) = \ln \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k \, p(\mathbf{x}^{(i)}|\boldsymbol{\theta}_k) = \sum_{i=1}^{N} \ln \sum_{k=1}^{K} \pi_k \, p(\mathbf{x}^{(i)}|\boldsymbol{\theta}_k)$$

(c) Napišite izraz za mješavinu s latentnim varijablama i izvedite izraz za (potpunu) log-izglednost tog modela. Možemo li dalje raditi izravno s tom log-izglednošću? Zašto?

- $$p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) = P(\mathbf{z}) p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) = \prod_{k=1}^{K} \pi_k^{z_k} \prod_{k=1}^{K} p(\mathbf{x}|\boldsymbol{\theta}_k)^{z_k} = \prod_{k=1}^{K} \pi_k^{z_k} p(\mathbf{x}|\boldsymbol{\theta}_k)^{z_k}$$
- $$\begin{aligned} \ln \mathcal{L}(\boldsymbol{\theta}|\mathcal{D}, \mathbf{Z}) &= \ln \prod_{i=1}^{N} p(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}|\boldsymbol{\theta}) = \ln \prod_{i=1}^{N} \prod_{i=1}^{K} \pi_k^{z_k^{(i)}} \, p(\mathbf{x}|\boldsymbol{\theta}_k)^{z_k^{(i)}} \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K} z_k^{(i)} \Big(\ln \pi_k + \ln p(\mathbf{x}^{(i)}|\boldsymbol{\theta}_k) \Big) \end{aligned}$$
- Kada bi $z^{(i)}$ bile poznate, maksimizacija ove log-izglednosti imala bi analitičko rješenje i mogli bismo dalje izravno raditi s tom log-izglednošću. $z^{(i)}$ su nepoznate, no možemo uz fiksirane π_k i Θ_k izračunati očekivanje izglednosti

(d) Definirajte E-korak i M-korak algoritma maksimizacije očekivanja primijenjenog na Gaussovu mješavinu.

- E - korak: Izračun očekivanja potpune izglednosti uz fiksirane parametre u iteraciji t:

$$\begin{aligned} \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \mathbb{E}_{\mathbf{Z}|\mathcal{D},\boldsymbol{\theta}^{(t)}} \Big[\sum_{i=1}^{N} \sum_{k=1}^{K} z_k^{(i)} \big(\ln \pi_k + \ln p(\mathbf{x}^{(i)}|\boldsymbol{\theta}_k) \big) \Big] \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K} \underbrace{\mathbb{E}[z_k^{(i)}|\mathcal{D}, \boldsymbol{\theta}^{(t)}]}_{=h_k^{(i)}} \big(\ln \pi_k + \ln p(\mathbf{x}^{(i)}|\boldsymbol{\theta}_k) \big) \end{aligned}$$

- M - korak: Izračun parametara za iteraciju (t+1) koji maksimiziraju očekivanje

$$\nabla_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = 0$$

$$\nabla_{\pi_k}\Big(\sum_{i=1}^{N}\sum_{k=1}^{K} h_k^{(i)} \ln \pi_k + \lambda\big(\sum_k \pi_k - 1\big)\Big) = 0 \quad \Rightarrow \quad \pi_k^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} h_k^{(i)}$$

$$\nabla_{\boldsymbol{\theta}_k} \sum_{i=1}^{N} h_k^{(i)} \ln p(\mathbf{x}^{(i)}|\boldsymbol{\theta}_k) = 0 \quad \Rightarrow \quad \boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_i h_k^{(i)} \mathbf{x}^{(i)}}{\sum_i h_k^{(i)}}$$

$$\Rightarrow \quad \boldsymbol{\Sigma}_k^{(t+1)} = \frac{\sum_i h_k^{(i)} (\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{(t+1)})(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{(t+1)})^{\mathrm{T}}}{\sum_i h_k^{(i)}}$$

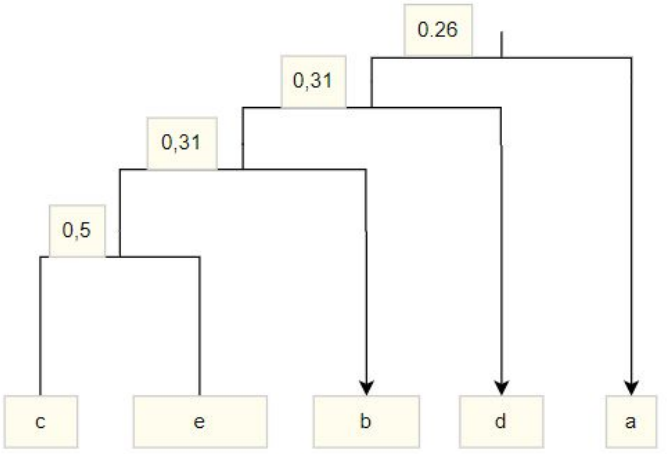
(e) Skicirajte vrijednost log-izglednosti $lnL(\theta|D)$ modela Gaussove mješavine kao funkcije broja iteracija, i to za tri različite vrijednosti parametra K (broj grupa): K = 1, K = 10 i K = 100. Na istom grafikonu skicirajte krivulju za K = 10 kada se za inicijalizaciju središta koristi algoritam k-sredina.

-

4. [Svrha: Isprobati rad algoritma hijerarhijskog aglomerativnog grupiranja (HAC) na konkretnom primjeru, za slučaj kada primjeri nisu vektori. Uočiti razliku izmedu udaljenosti i sličnosti te razliku izmedu jednostruke i potpune povezanosti.]
Jednako kao i algoritam k-medoida, algoritam hijerarhijskog algomerativnog grupiranja može se primijeniti u slučajevima kada primjeri nisu prikazani kao vektori značajki te kada umjesto mjere udaljenosti između vektora raspolažemo općenitijom mjerom sličnosti (ili različitosti). Neka je sličnost primjera iz D definirana sljedećom matricom sličnosti:

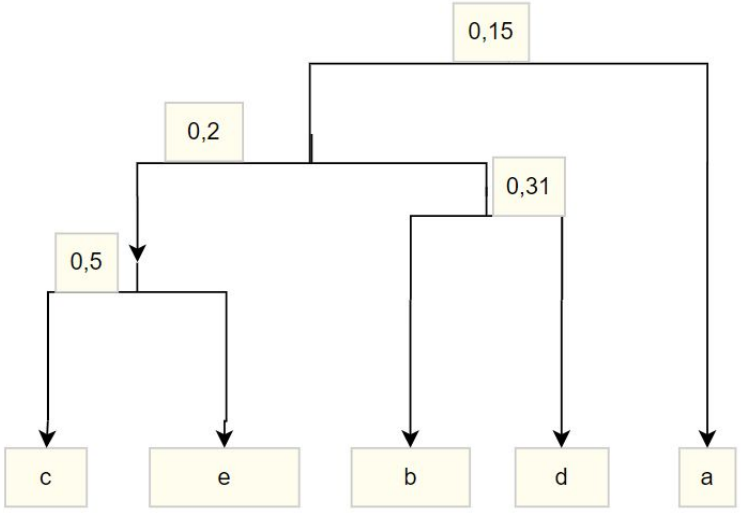
$$S = \begin{array}{c} \\ a \\ b \\ c \\ d \\ e \end{array} \begin{array}{c} \begin{array}{ccccc} a & b & c & d & e \end{array} \\ \begin{pmatrix} 1.00 & 0.26 & 0.15 & 0.20 & 0.17 \\ 0.26 & 1.00 & 0.24 & 0.31 & 0.31 \\ 0.15 & 0.24 & 1.00 & 0.20 & 0.50 \\ 0.20 & 0.31 & 0.20 & 1.00 & 0.24 \\ 0.17 & 0.31 & 0.50 & 0.24 & 1.00 \end{pmatrix} \end{array}$$

(a) Izgradite dendrogram uporabom jednostrukog povezivanja. Kada bi bilo potrebno napraviti particiju grupa, na kojoj biste razini presjekli taj dendrogram?



-
- Šnajder je rekao da bi dobro bilo sjeći tamo gdje su grane dugačke jer su tamo primjeri prirodno grupirani i razdvojeni, pa bih rekao da treba presjeći odmah nakon što se grupiraju c i e

(b) Izgradite dendrogram uporabom potpunog povezivanja. Kada bi bilo potrebno napraviti particiju grupa, na kojoj biste razini presljekli taj dendrogram?



-
- Ako imamo dvije iste vrijednosti pri odabiru minimalne, je li svejedno koju odaberemo jer npr u ovom slučaju možemo prvo spojiti b,d grupu s a pa onda tek a,b,d grupu s c,e grupom? Mislim da je svejedno
- Najveći razmak je i ovdje nakon spajanja c i e. Može netko provjerit?

5. [Svrha: Razumjeti kako se unutarnji kriterij algoritma grupiranja može (pokušati) upotrijebiti za provjeru grupiranja (odabir optimalnog broja grupa). Razumjeti da Akaikeov kriterij u stvari oponaša regulariziranu funkciju pogreške, koja pak aproksimira pogrešku generalizacije. Isprobati izračun Randovog indeksa na konkretnom primjeru.] Nedostatak svih algoritama

grupiranja koje smo razmotrili jest što se broj grupa K mora zadati unaprijed. Osim u rijetkim slučajevima kada nam je taj broj unaprijed poznat, to predstavlja problem.

(a) Skicirajte krivulju log-izglednosti kod EM-algoritma kao funkciju broja grupa K. Obrazložite izgled krivulje. Možete li temeljem ove krivulje odrediti optimalan broj grupa? Kako?

- Krivulja će rasti kako raste broj grupa po platoima, odnosno bit će mjesta na kojima raste strmije, a bit će i mjesta gdje će se činiti da funkcija uopće ne raste. Na početku tih platoa očito je došlo do neke kvalitetnije grupacije i zato će na mjestima početka platoa biti kvalitetniji (optimalni) broj grupa.

(b) Optimizacija broja grupa K može se provesti nekim kriterijem koji kombinira funkciju pogreške (odnosno log-izglednost) i složenost modela. Takav kriterij odgovara strukturnome riziku modela, koji je minimalan za optimalan broj grupa. Jedan takav kriterij jest Akaikeov informacijski kriterij (AIC):

$$K^* = \underset{K}{\operatorname{argmin}} \left(-2 \ln \mathcal{L}(K) + 2q(K) \right)$$

(1) gdje je $-lnL(K)$ negativna log-izglednost podataka za K grupa, a q(K) je broj parametara modela s K grupa. Pretpostavite da podatci D u stvarnosti dolaze iz K = 5 grupa. Podatke grupiramo dvjema varijantama EM-algoritma: standardni algoritam i preinačeni algoritam s dijeljenom kovarijacijskom matricom (zajednička kovarijacijska matrica procijenjena nad čitavim skupom primjera D na početku izvodenja algoritma). Skicirajte za ta dva algoritma funkciju koju minimizira Akaikeov minimizacijski kriterij.

- Graf će izgledati slično onome za pogrešku ispitivanja s obzirom na složenost modela. Kako raste broj klasa, prvo će opadati zbog porasta negativne zlog-izglednosti, a u nekom trenutku će se taj porast smanjiti te će funkcija početi rasti zbog porasta broja klasa. Preinačeni algoritam s dijeljenom kovarijacijskom matricom će uvijek biti malo ispod običnog jer će imati manje parametara.

(c) Kada su primjeri ili podskup primjera označeni, kvaliteta grupiranja (uključivo i broj grupa K) može se procijeniti Randovim indeksom. Randov indeks zapravo izračunava točnost s kojom ćemo par jednako označenih primjera smjestiti u istu grupu, odnosno par različito označenih primjera u različitu grupu. Izračunajte Randov indeks za sljedeću pariticiju označenih primjera (podskupovi su grupe dobivene grupiranje, a brojke su oznake klasa primjera):

$$\{\{0,0,1,2\},\{1,1\},\{2,2,2,1,0\}\}.$$

- a - broj isto označenih parova primjera u grupi
- b - broj suprotno označenih primjera u suprotnim grupama
- a = 1 + 1 + 3 = 5
- b = (1-2): 2*2 + 1*2 + (1-3): 2*3 + 2*1 + 1*3 + 1*1 + 1*1 + 1*1 + (2-3): 2*3 + 2 = 28
- $R = \frac{a+b}{\frac{11*10}{2}} = 0,6$

(d) Skicirajte vrijednost Randovog indeksa kao funkcije broja grupa K.

- Randov indeks će kontinuirano brzo rasti i usporiti na kraju kako se približava 1 jer će se nazivnik povećavati, a -> 0, b -> $\sum_{i=1}^{K-1} i$ kad svaki primjer bude u zasebnoj klasi

(e) Randov indeks možemo koristiti samo ako su podatci označeni ili je podskup podataka označen. Medutim, čini se da to onda ujedno podrazumijeva da je unaprijed poznat broj grupa K. Imamo li koristi od Randovog indeksa čak i onda kada unaprijed znamo broj grupa? Možemo li ikako upotrijebiti Randov indeks, a da nam broj grupa nije unaprijed poznat?

- Možemo ga upotrijebiti za uspoređivanje rezultata grupiranja algoritama

10. Vrednovanje modela

Raspolazemo skupom od 11 ispitnih primjera koje zelimo klasificirati u tri klase. Oznaka $y^{(i)}$ i izlaz modela $h(x^{(i)})$ za svaki od 11 primjera.

{(y, h(x)} = {(1, 1), (0, 2), (2, 2), (1, 2), (1, 1), (0, 0), (1, 1), (2, 1), (0, 1), (2, 0), (2, 1)}
Izracunajte prezicnost, odziv i F_1 mjeru, i to mikro i makro varijante.

- Global confusion matrix

klase	0	1	2
0	1	0	1
1	1	3	2
2	1	1	1

Class 0

1	1
2	7

Class 1

3	3
1	4

Class 2

1	2
3	5

Micro

5	6
6	16

- Mikro:
  Accuracy = (5+16) / 33 = 0.636363
  Precision = 5 / 11 = 0.454545
  Recall = 5 / 11 = 0.454545 (ovo je full slucajno ispalo isto lmao)
  F1 = 2 * 5/11 * 5/11 * (5 / 11 + 5 / 11) = 0.4545454
- Makro
  Accuracy = ⅓ (8/11 + 7 / 11 + 6 / 11) = 21/33 = 0.636363 (i ovo je slucajno isto kao mikro avg)
  Precision = ⅓(½ + 3/6 + ⅓) = 0.444444
  Recall = ⅓ (⅓ + 3/4 +¼) = 0.444444
  F1 = 2 * 0.444444 * 0.444444 * (0.444444 + 0.444444) = 0.444444
- nas model je onak uzasan.

Raspolazemo sa 1000 oznacenih primjera. Za vrednovanje SVM-a s hiperparamterima C i gamma koristimo ugnijezdenu unakrsnu provjeru sa po 5 ponavljanja u obje petlje. Hiperparametre optimiramo resetkastim pretrazivanjem u rasponima C E {2e-5 .. 2e15} i gamma E {2e-15 .. 2e3}.

(b) Koliko cemo puta ukupno trenirati model?

- Broj C = 21, broj gamma = 19, broj_kombinacija_parametara = 21*19, k = l = 5, 5*(21*19*(5) + 1)= 9980

(c) Koliko cemo primjera u svakoj od iteracija koristiti za treniranje, koliko za provjeru a koliko za ispitivanje?
- Prvo se dataset dijeli na train_val i test, posto ima 5 foldova, dakle 200 na test, 800 na train_val
- Zatim se train_val dijeli na train i val, opet 5 foldova, dakle 160 na val i 640 na train
- Na 640 primjera treniramo, na 160 primjera ocjenimo tocnost, i najboljem zatim ocjenimo finalnu tocnost na 200 primjera

(d) Kako glase odgovori na prethodna pitanja, ako bismo u vanjskoj petlji umjesto peterostruke unakrsne provjere koristili unakrsnu provjeru izdvoji jednoga?
- Sad nam vanjska petlja ima 1000 ponavljanja, pa bi model trenirali 1000*(21*19*(5) + 1)= 1996000 puta
- Vanjski preklopi bi nam sad svaki sadržavao 1 primjer, pa bi onda Dtrain imao 999 primjera, a Dtest 1 primjer
- Onda bi još Dtrain podijelili na unutarnje preklope, od kojih bi svaki sadržavao 999/5 = 200 primjera cca i to znači da bi nam onda Dtrain' imao 799 primjera, a Dvalidate 200 primjera

(e) Klasifikator svm je posebno osjetljiv na razlike u rasponima izmedu znacajki, pa se preporuca standardizirati znacajke. Sto to tocno znaci i kako biste standardizaciju znacajki ugradili u unakrsnu provjeru?
- Standardiziramo sa srednjom vrijednosti i varijancom izracunatom nad training setom, i onda te vrijednosti koristimo za validation set.
- Ne znam sto se dogada sa test setom - pretpostavljam da bi i njih trebalo standardizirati?

(f) Gdje biste u ugnijezdenu petlju ugradili odabir znacajki modela i optimizaciju praga po mjeri AUC?
- *blank stare*

F_1 mjeru klasifikatora procjenjujemo deseterostrukom unakrsnom provjerom (10-fold CV), pri cemu smo dobili sljedece vrijednosti: 0.68, 0.74, 0.71, 0.66, 0.58, 0.75, 0.76, 0.62, 0.78, 0.68

(g) Izracunajte 95% i 99% interval pouzdanosti. Sto pritom morate pretpostaviti i zasto?
- Pretpostavljamo da su brojevi iz Gaussove distribucije
- Ne znamo varijancu - procjenjujemo ju iz uzorka
- Po kojim formulama se sve ovo računa? $\bar{x}$ je aritmeticka sredina,

  $\widehat{\sigma^2} = \frac{1}{N-1}\sum_{x_i}(x_i - \bar{x})^2$

  Ide pristrani procjenitelj varijance
- $\bar{x}$=0.696, $\widehat{\sigma}$=0.064326

- Zbog toga što je sigma procijenjen iz uzorka, ne koristi se z-statistika, nego t-statistika i to za vrijednosti $\frac{\alpha}{2}$, odnosno Studentova T distribucija sa N=9 stupnjeva slobode - kriticne vrijednosti su 2.262 za 95% i 3.25 za 99%(kako se došlo do ovih kritičnih vrijednosti? Iščitavanje iz tablice)
- Interval pouzdanosti $\bar{x}$ ± **t** * σ/sqrt(N) (**t** je kriticna vrijednost)
  - ➢ 95% [0.64999, 0.742] -> može postupak računanja brojeva? Ne bi li ovdje trebalo biti 95% [0.696, 0.046] ?
  - ➢ 99% [0.62989, 0.7621]

(h) Bi li interval bio siri ili uzi da smo ove procjene za tocnost i standarnu devijaciju dobili na temelju peterostruke, a ne deseterostruke unakrsne provjere?
- Definiticno siri jer 1/sqrt(5) > 1/sqrt(10).
- Cak i bez toga, kriticne vrijednosti od t za N=4 su vece nego za N=9.

(i) *Bismo li na isti nacin mogli izracunati interval pouzdanosti za unakrsnu privjeru leave-one-out?

(j) *Bismo li na isti nacin mogli izracunati interval pouzdanosti procjene F_1 mjere da niste radili unakrsnu provjeru, vec samo ispitivali klasifikator na jednom ispitnom skupu (holdout method). Zasto?

Trenirali smo model h_2 i zelimo provjeriti je li njegova tocnost bolja od baseline modela h_1 koji sve primjere klasificira u vecinsku klasu. Oba modela vrednujemo deseterostrukom unakrsnom provjerom na ukupno N=1000 primjera te racunamo tocnosti oba modela na svakom od deset preklopa. Rezultati su sljedeci:

acc(h_1)	0.52	0.56	0.44	0.58	0.49	0.39	0.47	0.57	0.55	0.43
acc(h_2)	0.60	0.54	0.58	0.58	0.46	0.58	0.50	0.67	0.61	0.55
diffs	-0.08	0.02	-0.14	0.0	0.03	-0.19	-0.03	-0.10	-0.06	-0.12

(k) Primjenite upareni t-test i provjerite hipotezu da je razlika u tocnosti statisticki znacajna na razini znacajnosti 5% (dvostrani test). Iskazite zakljucak.
- H0: μ=0
  H1: μ != 0
- $\bar{d}$ = -0.067
  $\hat{\sigma}_d$ = 0.068855
  t = $\frac{\bar{d}-0}{\hat{\sigma}_d/sqrt(9)}$ = -2.91918
  Treba provjeriti (samo jedno pozitivno je dovoljno):
  - Varijanta a) $|t| \geq t_{\alpha/2}$? -> 2.91918 > 2.262 DA
  - Varijanta b) $p(|X| > t) \leq \alpha$? -> Piše u nastavnim materijalima da je ovo provjera p-vrijednosti, odnosno vjerojatnosti da smo opazili $\bar{d}$ ili ekstremnije, ako je hipoteza istinita. Pretpostavljam da bi to bilo 5/10 i to zaista je $\leq$ 0.95. Može netko ovo provjerit?

- Zakljucak: mozemo odbaciti H0 i prihvatiti H1 na razini znacajnosti 5% (modeli su razliciti)

(l) Je li razlika znacajna za 1%?
- Ne. Kriticna vrijednost za 1% za N=9 stupnjeva slobode je -3.25

(m) Sad isto ko (a) ali jednostrani test.
- Kriticna vrijednost za N=9 jednostrano na razini 5% je -1.8331
- Mozemo odbaciti H0 i zakljuciti da je model 2 bolji od modela 1.

(n) Koje pretpostavke moraju vrijediti da biste uopce mogli primjeniti t-test? Vrijede li te pretpostavke u gornjim slucajevima? Igra li ukupan broj primjera N ikakvu ulogu pri statistickom testiranju.
- Da bismo mogli uopce koristiti t-test, podaci moraju dolaziti iz Gaussove distribucije.
  - ➢ Ta pretpostavka uglavnom vrijedi za tocnosti jer se one, kao srednje vrijednosti, pokoravaju Gaussu
- Ukupan broj primjera igra veliku ulogu - sto vise primjera, bolji smo u procjeni