

Învățare automată

— Licență, anul III, 2021-2022, examenul parțial II —

Nume student:

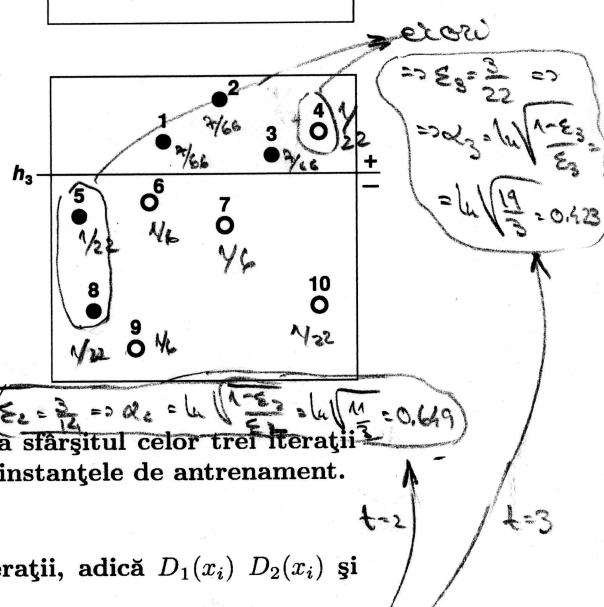
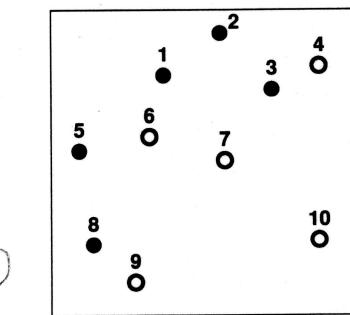
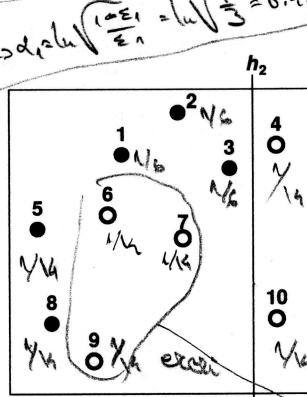
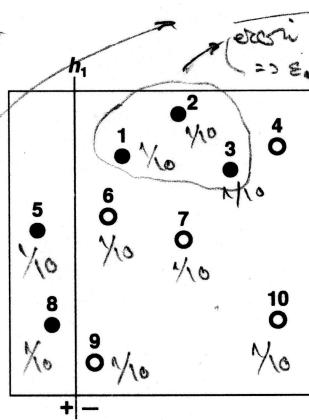
Grupa:

1.

(Algoritmul AdaBoost: aplicare pe un set de date din \mathbb{R}^2)

Se consideră că aplicăm algoritmul AdaBoost pe dataset-ul din figura alăturată. (Pentru ușurință exprimării la calcule, am notat pe figură indicii instanțelor de antrenament, în imediata apropiere a acestora.)

La primele trei iterații ale algoritmului au fost selectați compasii de decizie h_1, h_2 și h_3 (în această ordine), așa cum se indică în figurile de mai jos.



Obiectivul acestei probleme este să determinăm dacă la sfârșitul celor trei iterații algoritmul AdaBoost reușește să clasifice perfect toate instanțele de antrenament.

a. Calculați

- distribuțiile probabiliste corespunzătoare celor 3 iterații, adică $D_1(x_i)$, $D_2(x_i)$ și $D_3(x_i)$, pentru $i = 1, \dots, 10$;
- pentru fiecare dintre cele 3 iterații ($t = 1, 2, 3$): eroarea ponderată la antrenare (ε_t) produsă de compasul de decizie h_t , și ponderea (α_t) asociată ipotezei / compasului de decizie h_t .

Veți completa tabelele următoare și veți indica succint(!) modul în care ați procedat pentru a ajunge la rezultatele respective.

i	1	2	3	4	5	6	7	8	9	10
$D_1(x_i)$	<u>$1/10$</u>	<u>$1/10$</u>	<u>$1/10$</u>	$1/10$	<u>$1/10$</u>	<u>$1/10$</u>	<u>$1/10$</u>	<u>$1/10$</u>	<u>$1/10$</u>	<u>$1/10$</u>
$D_2(x_i)$	$1/6$	<u>$1/6$</u>	<u>$1/6$</u>	$1/14$	<u>$1/14$</u>	<u>$1/14$</u>	<u>$1/14$</u>	<u>$1/14$</u>	<u>$1/14$</u>	<u>$1/14$</u>
$D_3(x_i)$	$7/66$	<u>$7/66$</u>	<u>$7/66$</u>	$1/22$	<u>$1/22$</u>	$1/6$	<u>$1/6$</u>	<u>$1/6$</u>	<u>$1/6$</u>	<u>$1/6$</u>

Justificare: D_1 este distribuția uniformă (echitară)

Justificare ④
x₂, x₃ au același probabilitate ca și x₁, fiindcă - testele sunt echivalente clarifică de către h₂
- $D_2(1)=D_2(2)=D_2(3)$

Justificare ⑤
x₁ nu este clarificat de către h₁, fiindcă testele 3 (și deasupra) sunt clarificate de către h₁, iar la iterare t=1 testele au avut același probabilitate (1/10)
iar $D_1(1)=D_1(2)=D_1(3)$

Justificare ⑥ pag. 1
x₅, ..., x₁₀ au același probabilitate ca și x₄, fiindcă (ele sunt) mult clarificate de către h₂, iar la iterare t=1 testele au avut același probabilitate (1/10)
 $D_3(6)=D_3(7)=D_3(8)$ - $D_2(6)=D_2(7)=D_2(8)$...

Justificare ⑦
x₅, ..., x₁₀ au același probabilitate ca și x₄, fiindcă - testele sunt clarificate de către h₃
- $D_2(4)=D_2(5)=D_2(6)$ - $D_2(6)=D_2(7)=D_2(8)$...

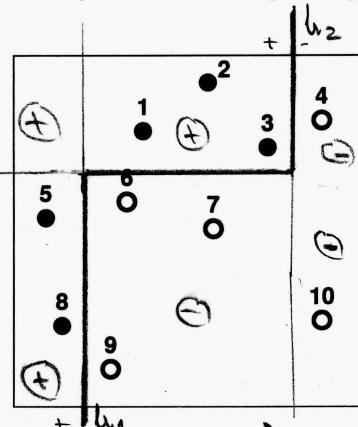
t	1	2	3
ε_t	$3/10$	$3/14$	$3/22$
α_t	$\ln(7/3)$ 0.423	$\ln(14/3)$ 0.649	$\ln(19/3)$ 0.923

Atenție! Pentru a vă ușura munca, am completat noi câteva dintre elementele primului tabel. Vă puteți baza pe valorile indicate de noi, ca să vă simplificați operațiunile / calculele!

b. Folosind ipoteza combinată obținută de algoritmul AdaBoost la finalul celei de-a treia iterării, stabiliți

i. eroarea la antrenare produsă (pentru calcularea ei, puteți folosi tabelul de mai jos),

ii. zonele de decizie corespunzătoare acestui clasificator (veți justifica modul în care ați procedat!). Veți indica aceste zone de decizie, precum și granițele de decizie, pe desenul de alăturat.



t	α_t	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1	$\alpha_1 = 0.423$	○ (-)	○ (-)	○ (-)	+	+	-	-	+	-	-
2	$\alpha_2 = 0.649$	+	+	+	-	+	(+) (+)	(+) (+)	+	(+) (-)	-
3	$\alpha_3 = 0.923$	+	+	+	(+) (-)	(-)	-	-	(-)	-	-
$H_3(x_i)$											

$$\Rightarrow \text{err}(H_3) = 0$$

Atenție! Pentru a vă ușura munca, vă furnizăm noi următoarele valori numerice: $\ln \sqrt{7/3} \approx 0.423$, $\ln \sqrt{14/3} \approx 0.649$, $\ln \sqrt{19/3} \approx 0.923$.

$$\text{err}(H_3) = 0$$

\Rightarrow Zonele de decizie au renunțat instanțelor situate în interiorul zonelor kerperii
(le-am marcat - încă e dată - cu (+) și (-))
Granițele de decizie le-am marcat împreună.

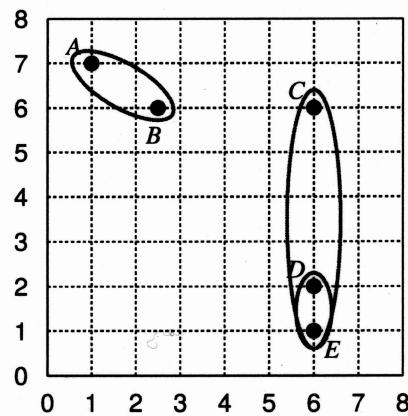
Pe linia acesta în figura, se poate vedea că
zona din față în zonă
kerp h_1, h_2, h_3
se poate

pe 2

2. (Clusterizare ierarhică aglomerativă: un exemplu simplu de aplicare, cu single-, complete- și average-linkage, pe date din \mathbb{R}^2)

Considerăm punctele $A(1, 7)$, $B(2.5, 6)$, $C(6, 6)$, $D(6, 2)$ și $E(6, 1)$ din planul euclidian. Pe acest dataset veți aplica algoritmul de clusterizare ierarhică aglomerativă (i.e., bottom-up), folosind pe rând (separat) funcțiile de similaritate single-linkage, complete-linkage și average-linkage.

Care dintre aceste funcții de similaritate va conduce după executarea a trei iterări consecutive la ierarhia aplatizată prezentată în figura alăturată?



Indicație:

Veți folosi grid-urile de mai jos.

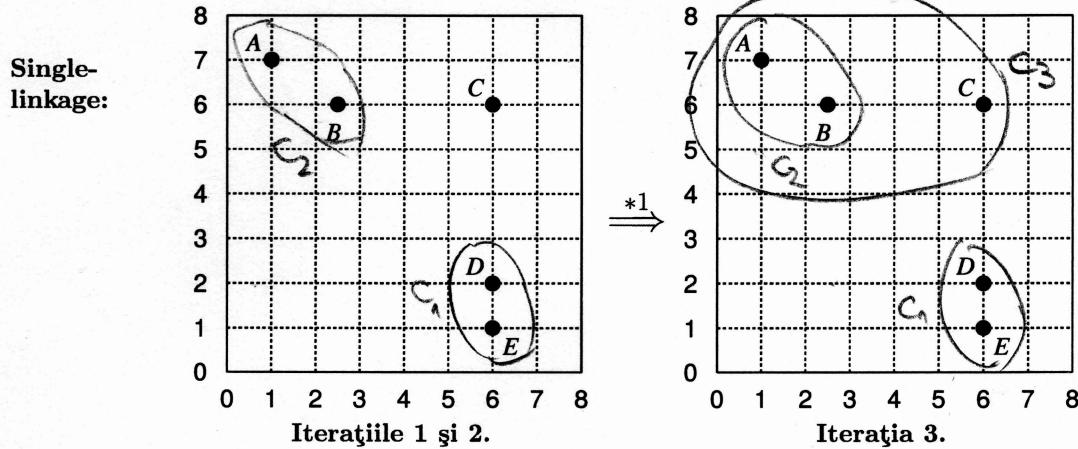
Pentru fiecare dintre cele 3 tipuri de măsuri de similaritate, veți justifica riguros trecerea de la iterăria 2 la iterăria 3 determinând (numerical!) minimul distanțelor dintre perechile de clustere care au fost formate la iterăria 2.

De exemplu, dacă în figura de mai sus notăm $C_1 = \{D, E\}$, $C_2 = \{A, B\}$, atunci justificarea acestui rezultat ar putea fi scrisă sub forma

$$\underbrace{d(C_1, C)}_{= \dots} < \underbrace{d(C_2, C)}_{= \dots} \text{ și } d(C_1, C) < \underbrace{d(C_1, C_2)}_{= \dots}, \quad (*)$$

unde d va fi una dintre „măsurile de distanță“ d_{SL} , d_{CL} sau d_{AL} , corespunzătoare măsurilor de similaritate single-linkage, complete-linkage și respectiv average-linkage.

Atenție!
Nu cum ești
= justificarea
pt. cum ne
dă rezultatul
de la iterăriile
n. 1 și 2



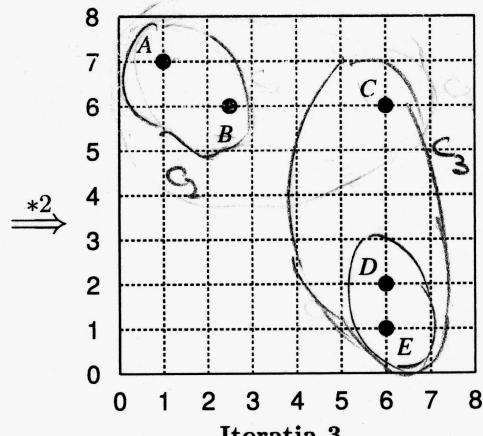
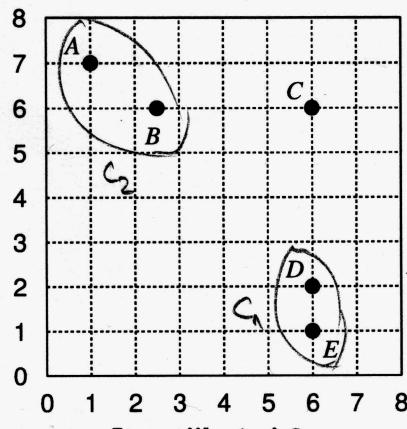
*1: $d_{SL}(C_1, C_2) = 4 > d_{SL}(C_2, C) = 3.5$, iar $d_{SL}(C_1, C_2) = \text{dist}(B, D) > 4 \rightarrow \text{min. este } d_{SL}(C_2, C) = 3.5$

$\frac{3.5}{\sqrt{4^2 + 3.5^2}} = 5.31$

Pag. 3

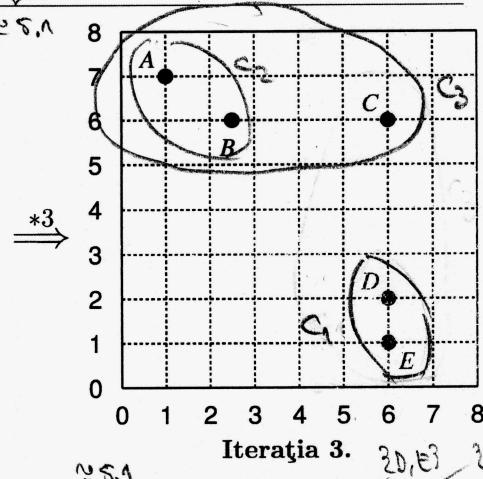
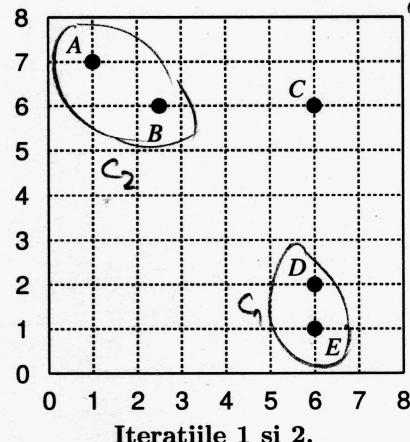
$\frac{3.5}{\sqrt{4^2 + 3.5^2}} = 5.31$

Complete-linkage:



$$*2: d_{CL}(C_1, C) = 5 < d_{CL}(C_2, C) = \sqrt{5^2 + 1^2}, \text{ iar } d_{CL}(C_1, C_2) = \text{dist}(A, B) > 5 \Rightarrow \text{nu merge } d(C_1, C_2) = \sqrt{5^2 + 6^2} \approx 7,81$$

Average-linkage:



$$*3: d_{AL}(C_1, C) = 4,5 > d_{AL}(C_2, C) = \frac{1}{2}(3,5 + \sqrt{5^2 + 1^2}), \text{ iar } d_{AL}(C_1, C_2) \text{ nu merge} > 4,5$$

$\{D, E\}$

$\{A, B\}$

$$\approx \frac{3,5 + 5,1}{2} = 4,3$$

$$\frac{1}{2}(\dots)$$

fiecare

$$\begin{aligned} d(C, D) &< d(B, D) \\ &< d(A, D) \\ \text{și } d(C, E) &< d(B, E) \\ &< d(A, E) \end{aligned}$$

\Rightarrow minimum este $d_{AL}(C_2, C)$

Concluzie: deoarece pentru complete-linkage se obține

după 3 iteratii ierarhia aplatizată din emne

numerice:

$$\begin{aligned} &\frac{1}{2}(\sqrt{3,5^2 + 4^2} \\ &+ \sqrt{3,5^2 + 5^2} \\ &+ \sqrt{2,5^2} \\ &+ \sqrt{5^2 + 6^2}) \\ &= 6,57 \end{aligned}$$

3.

(Algoritmul K-means: chestiuni legate de valoarea criteriului J)

Vă readucem aminte că la clusterizare cu algoritmul K -means obiectivul este să găsim pozițiile celor K centroizi ai clusterelor, notați cu $c_j \in \mathbb{R}^d, j \in \{1, \dots, K\}$, astfel încât suma pătratelor distanțelor dintre fiecare instanță x_i și cel mai apropiat centroid să fie minimizată. Așadar, funcția obiectiv este

$$\sum_{i=1}^n \min_{j \in \{1, \dots, K\}} \|x_i - c_j\|^2, \quad (1)$$

unde n este numărul de instanțe de clusterizat. Altfel spus, încercăm să găsim c_1, \dots, c_k astfel încât să fie minimizată expresia (1). Pentru aceasta, efectuăm mai multe iterații în cadrul cărora asignăm fiecare instanță x_i la cel mai apropiat centroid și apoi actualizăm poziția fiecărui centroid c_j la media instanțelor asignate la clusterul j .

Insă prietenul tău Ionuț, în loc să mențină numărul de clustere K fixat, încearcă să minimizeze valoarea expresiei (1) variindu-l pe K . Tu ești de părere că această idee nu este bună.

În mod concret, tu îl convingi pe Ionuț dându-i două valori: α , care reprezintă minimul valorilor posibile pentru expresia (1), și β , care este valoarea lui K atunci când expresia (1) își atinge valoarea minimă.

În mod concret, tu îl convingi pe Ionuț dându-i două valori: α , care reprezintă minimul valorilor posibile pentru expresia (1), și β , care este valoarea lui K atunci când expresia (1) își atinge valoarea minimă.

a. Cât este valoarea expresiei $\alpha + \beta$ în cazul în care $n = 100$?

b. Vrem să vedem cum lucrează algoritmul K -means atunci când datele de clusterizat sunt situate pe axa reală. Considerăm că avem 4 instanțe, și anume $x_1 = 1, x_2 = 2, x_3 = 5$ și $x_4 = 7$, iar K are valoarea 3. Cât este în acest caz valoarea optimă a funcției obiectiv (1)?

a. Expressia 1 este chiar criteriul J . Dacă fixăm K , putem scrie $J_K = \underline{J}_K$.
 Stiu de la curs că $\underline{J}_K = 0$, unde \underline{J}_K denotă cea mai mică valoare posibilă pt J_K , atunci cind K este fixat
 și $K \geq n$, dacă toate clusterile sunt singlete
 (dacă este posibil și poate ca să nu
 există, pt $K \geq n$ avem $\underline{J}_K = 0$)
 $\Rightarrow \underline{J} = 0, \beta = n \Rightarrow \underline{J} + \beta = n = 100$

b.

x_1	x_2	x_3	x_4
1	2	5	7

Nu se poate atinge valoarea 0 pt α în acest caz fiindcă $n=4 > k=3$
 și toate instanțele sunt distincte

\Rightarrow minimul lui J_3 se atinge pt cău în care 2 clusteră sunt singlete,
 iar cel de-al 3-lea cluster este format din cele mai apropiate 2 instanțe

$$\Rightarrow \left\{ \begin{array}{l} C_1 = \{x_1, x_2\} \Rightarrow \mu_1 = 1.5 \\ C_2 = \{x_3\} \Rightarrow \mu_2 = x_3 = 5 \\ C_3 = \{x_4\} \Rightarrow \mu_3 = x_4 = 7 \end{array} \right. \Rightarrow \underline{J} = J_3 = 2 \cdot \left(\frac{1}{2} \right)^2 = \frac{1}{2} \quad x_1, x_2$$

4

5. (Distribuția Bernoulli: estimare în sensul verosimilității maxime)

La curs am spus că la estimarea de verosimilitate maximă (engl., Maximum Likelihood Estimation, MLE), definim

$$\hat{\theta}_{MLE} = \arg \max_{\theta} P(\mathcal{D}|\theta) = \arg \max_{\theta} \ln P(\mathcal{D}|\theta),$$

unde \mathcal{D} este setul de date pe care se estimează parametrul θ al distribuției de probabilitate P .

Imaginează-ți că ești un *data scientist* care lucrează pentru o firmă de publicitate (engl., advertisement). Această firmă a făcut recent o campanie de publicitate [care constă în publicarea unui anunț publicitar pe ecranul calculatorului / telefonului unor anumite persoane], iar tăie să cere să evaluezi succesul acestei campanii de publicitate.

În cadrul acestei campanii, au fost targetate / vizate N persoane. Variabila aleatoare Y_i primește valoarea $y_i = 1$ dacă persoana i a făcut "click" pe anunțul publicitar și $y_i = 0$ în cazul contrar. În total, $\sum_{i=1}^N y_i = k$ persoane au decis să facă "click" pe anunțul publicitar. Vom presupune că probabilitatea ca persoana i să facă "click" pe anunțul publicitar respectiv este θ , iar probabilitatea ca persoana i să nu facă "click" pe anunțul publicitar este $1 - \theta$.

Desigur, $P(\mathcal{D}|\theta) = P(Y_1, \dots, Y_N|\theta)$.

a. Dedu expresia analitică pentru calcularea lui $\hat{\theta}_{MLE}$.

b. Presupunem că $N = 100$ și $k = 10$. Calculează $\hat{\theta}_{MLE}$.

Soluție:

$$\begin{aligned}
 \text{(1)} \quad & a. \quad P(\mathcal{D}|\theta) = \theta^k (1-\theta)^{N-k} \stackrel{\text{Bernoulli}}{=} L(\theta) = \text{funcția de verosimilitate (def. pe } [0,1]) \\
 \text{(2)} \quad & l(\theta) = \ln L(\theta) = k \ln \theta + (N-k) \ln (1-\theta) = \rightarrow \text{log-verosimilitate (def. pe } [0,1]) \\
 \text{(3)} \quad & \Rightarrow l'(\theta) = \frac{k}{\theta} - \frac{N-k}{1-\theta}, \quad \forall \theta \in (0,1) \\
 \text{(4)} \quad & l'(\theta) = 0 \Leftrightarrow \frac{k}{\theta} = \frac{N-k}{1-\theta} \Leftrightarrow k(1-\theta) = (N-k)\theta \Leftrightarrow N = 2k \Leftrightarrow \theta = \frac{k}{N} \in [0,1] \\
 \text{(5)} \quad & l''(\theta) = -\frac{k}{\theta^2} - \frac{(N-k)}{(1-\theta)^2} < 0 \quad (\text{în cazurile particulari } \theta \in \{0,1\} \Rightarrow L(\theta)=0, \text{ în rest } L(\theta) > 0) \\
 \text{(6)} \quad & \Rightarrow l - f. concavă \quad \Rightarrow \hat{\theta}_{MLE} = \frac{k}{N} \quad (*) \\
 & \quad (\text{adm } l'' \text{ max}) \\
 b. \quad & N=100, \quad k=10 \quad \Rightarrow \hat{\theta}_{MLE} = \frac{10}{100} = 0,1
 \end{aligned}$$

nu nu aștept ca
studentul să trateze
ac. ca particular,
deci arătă nu se
de-punctează
(Se punctează după
menționată de către cineva
tratată ac. ca.)

56.

(O problemă à la C. Do și S. Batzoglou:
rezolvarea unei mixturi de doi vectori de distribuții Bernoulli i.i.d.
identificarea parametrilor și a variabilelor neobservabile;
scrierea funcției de log-verosimilitate)

Să zicem că avem două monede, A și B . Sarcina ta este aceea de a afla (a „estima“) θ_A și θ_B , probabilitățile de apariție a feței *stemă* (engl., head) pentru fiecare dintre cele două monede. Însă eu sunt cam răutacios și nu-ți dau voie să arunci tu monedele. În schimb, decid să procedez astfel: voi arunca eu însuși monedele și după aceea îți voi comunica ție rezultatele aruncărilor. În mod concret, îți voi spune ceva de forma următoare: am ales una dintre cele două monede (nu-ți spun care anume), am aruncat-o de 10 ori și am obținut de 7 ori stema și de 3 ori banul. Apoi am ales din nou una dintre cele două monede (poate aceeași cu cea dinainte, poate nu), am aruncat-o de 10 ori și am obținut de 5 ori stema și de 5 ori banul. În total, îți comunic de N ori câte o astfel de informație. (Așadar, la final vei dispune de rezultatele a $10N$ aruncări ale monedelor.)

Formulează aceasta ca pe o problemă de tip EM (Expectation-Maximization).

- a. i. Care sunt datele observabile?
ii. Care sunt parametrii modelului (adică parametrii distribuțiilor probabiliste folosite)?
iii. Cine sunt variabilele ascunse / latente (engl., hidden variables)?
iv. Scrie expresia funcției de *log-verosimilitate a datelor complete* pentru această problemă.
(Atenție: Nu ti se cere să rezolvi efectiv problema EM.)

(Atenție: Nu îți se cere să rezolvi efectiv problema EM.)

Solutio

४

Date (e.g. intuitive/internal)

Rezia 1: $\boxed{\dots}$ TH, ST
 Rezia 2: $\boxed{\dots}$ SH, ST
 ; ; ;
 men N: $\boxed{\dots}$ TH, TH

Mr. de Fete Hin
Peria N

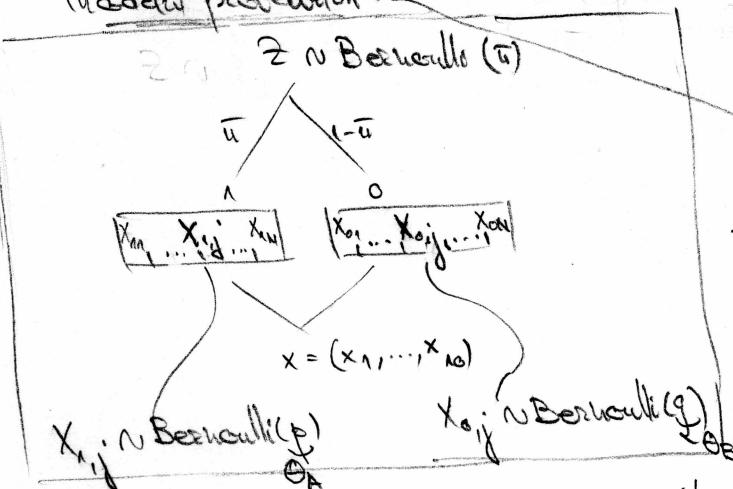
derivative
in Date (dpv form):

$$x_i = (x_{i1}, \dots, x_{iL})$$

(Q2: Deși suntem cerut în cinci), ar fi de dorit ca studenții

na furnizata reprezentarea grafică pt.

Model probabilist



Mixtura de
vectori de var al.
Bernoulli indep.
ni identic distib.
(relativ la "clasa"
 $Z = 1$ korp $Z = 0$)

- calc 10 outcome - w/ "observed" la media (de amuncări) i,
 Tete $\hat{p}_i = 1, \dots, N$

Locau, rezumativ; nu de ^{H?} de
nu de ti

"generat" de $X_{1,j}$ } după care $\tau_i = 1$ sau
 respectiv $X_{0,j}$ } $\tau_i = 0$

ii. Variabilele aruncă: $z_i \in \{0,1\}$ - deci următoarele sunt a patra clasă (a secundă)

iii. Parameterneindeutig: $\Theta \stackrel{\text{vert}}{=} (\Pi, p, q)$

b. log-verosimilitatea unei (singure) date "complete" la iteratia t.

$$\ln P(X_i, z_i | \theta^{(t)}) = \ln P(X_i | z_i; \theta^{(t)}) \cdot P(z_i | \theta^{(t)}) \quad (1)$$

$$\text{indep.} \rightarrow \ln \left\{ \left[(p^{(t)})^{\#H_i=x_i} \cdot (1-p^{(t)})^{\#T_i-x_i} \cdot \bar{\pi}^{(t)} \right]^{z_i} \right\} \quad (2)$$

$$(p^{(t)}, p^{(t)}, q^{(t)}) \text{ the "exponentiation trick"} \quad \left[(q^{(t)})^{\#H_i=x_i} \cdot (1-q^{(t)})^{\#T_i-x_i} \cdot (1-\bar{\pi}^{(t)})^{1-z_i} \right] \quad (2)$$

$$\text{prop. log.} \rightarrow z_i [\#H_i=x_i \ln p^{(t)} + \#T_i-x_i \ln (1-p^{(t)}) + \ln \bar{\pi}^{(t)}] + (1-z_i) [\#H_i=x_i \ln q^{(t)} + \#T_i-x_i \ln (1-q^{(t)}) + \ln (1-\bar{\pi}^{(t)})] \quad (3)$$

log-verosimilitatea datelor "complete":

$$\ln P(X_1, \dots, X_N, z_1, \dots, z_N | \theta^{(t)}) \stackrel{\text{ind}}{=} \ln \prod_{i=1}^N P(X_i, z_i | \theta^{(t)}) = \quad (4)$$

$$(X_1, \dots, X_N) \quad \text{prop. log.} \sum_{i=1}^N \ln P(X_i, z_i | \theta^{(t)}) \quad (5)$$

$$\stackrel{(3)}{=} \sum_{i=1}^N \left\{ z_i [\#H_i=x_i \ln p^{(t)} + (N - H_i) \ln (1-p^{(t)}) + \ln \bar{\pi}^{(t)}] + (1-z_i) [N \ln q^{(t)} + (N - H_i) \ln (1-q^{(t)}) + \ln (1-\bar{\pi}^{(t)})] \right\} \quad (6)$$

in categorie X_i (probabil cei mai multi studenti
se vor inscrie din verosimilarea aceea;
incertitudine, difara de "mitaxa")