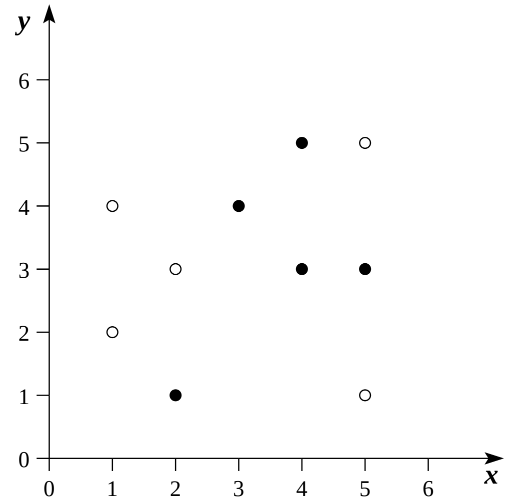
Învăţare automată

— Licenţă, anul III, 2018-2019, examenul parţial II —

Nume student: **Grupa:**

1. **(Comparaţie între algoritmii 1-NN: zone şi suprafeţe de decizie; calculul erorii la CV cu metoda "leave-one-out")**

a. Pe setul de date alăturat desenaţi *graniţele de decizie* produse de către algoritmul 1-NN (veţi obţine deci *diagrama Voronoi*). Apoi haşuraţi *suprafaţa de decizie* corespunzătoare clasei +, marcată prin eticheta / simbolul •.



b. Pe acelaşi set de date, calculaţi eroarea produsă la cross-validare cu metoda "leave-one-out" (CVLOO) de către algoritmul 1-NN. Veţi exprima această eroare sub forma unei *fracţii*.

***Atenţie*: În cazul în care veţi obţine ,,paritate" de voturi, veţi calcula eroarea CVLOO folosind algoritmul 3-NN. Dacă nici atunci nu reuşiţi să eliminaţi paritatea, aplicaţi algoritmul 5-NN ş.a.m.d.**

Răspuns:

Data	Eticheta	Vecinătate	Clasificare la CVLOO	Eroare?
$A(1,2)$		$\{\ldots\}$		
$B(1,4)$				
$C(2,1)$				
$D(2,3)$				
$E(3,4)$				
$F(4,3)$				
$G(4,5)$				
$H(5,1)$				
$I(5,3)$				
$J(5,5)$				

***Observaţie*: La completarea coloanei *Vecinătate*, veţi folosi litere în loc de coordonatele punctelor. (Pentru aceasta, este recomandabil ca, în prealabil, pe desenul de mai sus să puneţi literele corespunzătoare punctelor, conform primei coloane din tabel.)**

$$err_{CVLOO}(\mathit{1\text{-}NN}) = \ \ldots$$

2. **(Distribuţia gaussiană uni-variată: estimarea parametrilor; calcularea unor probabilităţi a posteriori)**

X	Y
0	A
2	A
3	B
4	B
5	B
6	B
7	B

Presupunem că dispunem de setul de date de antrenament din tabelul alăturat; singurul atribut de intrare (X) ia valori reale, iar atributul de ieşire (Y) este de tip Bernoulli, deci ia două valori, notate cu A şi respectiv B.

a. Pornind de la acest set de date şi presupunând că instanţele din clasa A au fost generate de o distribuţie gaussiană, iar instanţele din clasa B au fost generate de o altă gaussiană, estimaţi *parametrii* acestor gaussiene, prin metoda verosimilităţii maxime (MLE).

Atenţie!

Veţi enunţa mai întâi formulele corespunzătoare din capitolul *Estimarea parametrilor; metode de regresie*.

Centralizaţi rezultatele, completând tabelul de mai jos.

$\mu_A =$	$\sigma_A^2 =$	$P(Y = A) =$
$\mu_B =$	$\sigma_B^2 =$	$P(Y = B) =$

b. Notăm $\alpha = p(X = 2|Y = A)$ şi $\beta = p(X = 2|Y = B)$.

– **Cât este $p(X = 2, Y = A)$ în funcţie de α?**

– **Cât este $p(X = 2, Y = B)$ în funcţie de β?**

– **Cât este $p(X = 2)$ în funcţie de α şi β?**

– **Cât este $p(Y = A|X = 2)$ în funcţie de α şi β?**

c. [Bonus]
Cum va clasifica varianta gaussiană a algoritmului Bayes Naiv (pe care am prezentat-o în mod succint la curs) punctul $X = 2$?
Veţi exprima răspunsul mai întâi în funcţie de α şi β. Apoi veţi veţi face calculele folosind *valorile* lui α şi β, determinate în funcţie de *parametrii* calculaţi la punctul precedent.

Răspuns:

3. **(Algoritmul K-means – varianta care foloseşte variabile-indicator γ_{ij}: aplicare pe date din $\mathbb{R}$; Algoritmul EM/GMM, cazul $\sigma_1^2 = \sigma_2^2 = 1$, $\pi_1 = \pi_2$: executarea unei iteraţii, pe date din $\mathbb{R}$)**

A. **Fie $X = \{x_1, x_2, \ldots, x_n\}$ o mulţime de instanţe de clusterizat, iar K numărul de clustere cu care vom lucra.**

Veţi folosi următoarea variantă a algoritmului K-means:

- **Se iniţializează în mod arbitrar centroizii $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K$ şi se ia $C = \{1, \ldots, K\}$.**
- **Repetă:**

Pasul 1:

Calculează matricea γ (de dimensiune $n \times K$ şi având elemente din mulţimea $\{0, 1\}$) astfel:

$$\gamma_{ij} \leftarrow \begin{cases} 1, & \textbf{dacă } \|x_i - \boldsymbol{\mu}_j\|^2 \leq \|x_i - \boldsymbol{\mu}_{j'}\|^2, \ \forall j' \in C, \\ 0, & \textbf{în caz contrar.} \end{cases}$$

În caz de egalitate, alege în mod arbitrar cărui cluster (dintre cele eligibile) să-i aparţină instanţa x_i.

Pasul 2:

Recalculează $\boldsymbol{\mu}_j$ folosind matricea γ actualizată:
Pentru fiecare $j \in C$, dacă $\sum_{i=1}^n \gamma_{ij} > 0$, asignează

$$\boldsymbol{\mu}_j \leftarrow \frac{\sum_{i=1}^n \gamma_{ij} x_i}{\sum_{i=1}^n \gamma_{ij}}.$$

Altfel, menţine neschimbat centroidul μ_j.

până când matricea γ nu se mai schimbă de la o iteraţie la alta.

În continuare se va considera că $n = 2$, $x_1 = 0.5$ şi $x_2 = 2$, iar valorile iniţiale pentru *centroizii* $\boldsymbol{\mu}_1$ şi $\boldsymbol{\mu}_2$ sunt 1 şi respectiv 2. (Notaţie: $\boldsymbol{\mu}_1^{(0)} = 1$, $\boldsymbol{\mu}_2^{(0)} = 2$.)

Aplicaţi algoritmul K-means (în varianta de mai sus!) pe aceste date.

<u>Răspuns:</u>

Iniţializare:

$\boldsymbol{\mu}_1^{(0)} = 1$, $\boldsymbol{\mu}_2^{(0)} = 2$

Iteraţia 1:

Iteraţia 2:

Iteraţia 3:

...

B. Fie un model de mixtură gaussiană (engl., Gaussian mixture model, GMM) cu două componente având varianţe cunoscute şi probabilităţi *a priori* egale pentru selecţia celor două distribuţii:

$$\frac{1}{2}\mathcal{N}(x;\mu_1,1)+\frac{1}{2}\mathcal{N}(x;\mu_2,1),\ x\in\mathbb{R}.$$

În continuare se va considera (din nou) că $n=2$, $x_1=0.5$ şi $x_2=2$, iar valorile iniţiale pentru *mediile* μ_1 şi μ_2 sunt 1 şi respectiv 2. (Notaţie: $\mu_1^{(0)}=1$, $\mu_2^{(0)}=2$.)

Executaţi în mod manual o iteraţie a algoritmului EM, versiunea prezentată la curs (preluată din cartea *Machine Learning* a lui Tom Mitchell, pag. 193) pe aceste date, astfel:

a. Pasul E — estimarea probabilităţilor pentru variabilele ,,neobservabile":

Pentru $i\in\{1,2\}$ şi $j\in\{1,2\}$, calculaţi $P(Z_{ij}=1|X=x_i;\mu_1^{(0)},\mu_2^{(0)})$, probabilităţile a posteriori de apartenenţă a datelor observate (x_1 şi x_2) la cele două componente ale mixturii. (Vă readucem aminte că $Z_{ij}=1$ dacă instanţa x_i a fost generată de către gaussiana cu media μ_j, iar $Z_{ij}=0$ în cazul contrar.) Justificaţi în mod detaliat!

***Indicaţie*: În vederea efectuării calculelor, pentru convenienţă puteţi considera valorile distribuţiei normale / gaussiene standard $\mathcal{N}(x;\mu=0,\sigma^2=1)$ în punctele 0, 0.5, 1, 1.5 şi 2 ca fiind respectiv 0.4, 0.35, 0.24, 0.13 şi 0.05.**

b. Pasul E (continuare) — calcularea funcţiei ,,auxiliare":

Definiţia funcţiei ,,auxiliare" la iteraţia 1 este următoarea:

$$Q(\mu_1,\mu_2|\mu_1^{(0)},\mu_2^{(0)})=E[\ln P(Y|\mu_1,\mu_2)],$$

unde

> **$Y\stackrel{not.}{=}\{y_1,y_2\}$, cu $y_i\stackrel{not.}{=}(x_i,Z_{i1},Z_{i2})$, pentru $i\in\{1,2\}$,**
>
> **μ_1 şi μ_2 sunt din $\mathbb{R}$ şi sunt consideraţi parametri liberi, iar**
>
> **media E se calculează în funcţie de mediile variabilelor ,,neobservable" Z_{ij} calculate [folosind probabilităţile a posteriori deduse] la punctul precedent.**

Determinaţi formula (formulele) de calcul pentru

$$\mu^{(1)}\stackrel{not.}{=}\arg\max_{\mu\in\mathbb{R}^2}Q(\mu|\mu^{(0)}).$$

Am folosit notaţiile $\mu=(\mu_1,\mu_2)$ şi $\mu^{(t)}=(\mu_1^{(t)},\mu_2^{(t)})$.

c. Pasul M — maximizarea funcţiei ,,auxiliare" Q:

Calculaţi valorile parametrilor μ_1 şi μ_2 la iteraţia 1 (adică $\mu_1^{(1)}$ şi $\mu_2^{(1)}$), în funcţie de probabilităţile calculate la primul punct. (Justificaţi în mod detaliat!) Care credeţi că va fi tendinţa de mişcare a mediilor la următoarele iteraţii?

C. La curs am enunţat un *rezultat teoretic* (demonstrat în carte), care afirmă că în anumite condiţii, algoritmul EM pentru mixturi de distribuţii gaussiene se comportă *la limită* asemenea algoritmului K-means. Concret, la ce anume se referă această trecere la limită?

Folosind acest rezultat teoretic, propuneţi o *schimbare* [minimală!] relativă la *setarea iniţială* a parametrilor distribuţiilor gaussiene astfel încât, pe datele $X=\{x_1,x_2\}$ de la punctele A şi B de mai sus, *la limită* poziţiile finale ale mediilor obţinute de către EM să coincidă cu centroizii obţinuţi de către algoritmul K-means la convergenţă.

<u>Răspuns:</u>

a.

$P(Z_{11}=1|x_1;\mu_1^{(0)},\mu_2^{(0)})=\ldots$

$P(Z_{12}=1|x_1;\mu_1^{(0)},\mu_2^{(0)})=\ldots$

$P(Z_{21} = 1|x_2; \mu_1^{(0)}, \mu_2^{(0)}) = \ldots$

$P(Z_{22} = 1|x_2; \mu_1^{(0)}, \mu_2^{(0)}) = \ldots$

b.

$y_1 \overset{not.}{=} (x_1, Z_{11}, Z_{12}) \Rightarrow \ln P(y_1|\mu_1, \mu_2) = \ldots$

$y_2 \overset{not.}{=} (x_2, Z_{21}, Z_{22}) \Rightarrow \ln P(y_2|\mu_1, \mu_2) = \ldots$

$Y \overset{not.}{=} \{y_1, y_2\} \Rightarrow \ln P(Y|\mu_1, \mu_2) = \ldots$

$\Rightarrow Q(\mu_1, \mu_2|\mu_1^{(0)}, \mu_2^{(0)})] \overset{def.}{=} E[\ln P(Y|\mu_1, \mu_2)] = \ldots$

c.

Formula de calcul pentru $\mu_1^{(1)}$:

Justificare (demonstraţie!):

Aşadar, valoarea lui $\mu_1^{(1)}$ este:

Formula de calcul pentru $\mu_2^{(1)}$:

Justificare (demonstraţie!):

Aşadar, valoarea lui $\mu_2^{(1)}$ este:

C.