Învăţare automată

— Licenţă, anul III, 2021-2022 —
examen „suplimentar“ I

Nume student: **Grupa:**

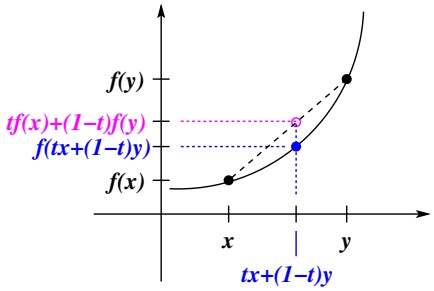
1. **(Inegalitatea lui Jensen şi câteva consecinţe ale ei)**

Liviu Ciortuz, 2019

Dacă $f : \mathbb{R} \to \mathbb{R}$ **este o *funcţie convexă*, atunci, conform *definiţiei*,**[1] **pentru orice** $t \in [0,1]$ **şi orice** $x_1, x_2 \in \mathbb{R}$ **urmează**

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2). \quad (1)$$

Dacă f **este funcţie strict convexă, atunci egalitatea are loc doar dacă** $x_1 = x_2$**.**



a. Folosind definiţia de mai sus, demontraţi *inegalitatea lui Jensen*:[2]
Pentru orice $a_i \geq 0,\ i = 1, \ldots, n$ **cu** $\sum_i a_i = 1$ **şi orice** $x_i \in \mathbb{R},\ i = 1, \ldots, n$**, dacă** f **este funcţie convexă,**[3] **atunci**

$$f\big(\sum_i a_i x_i\big) \leq \sum_i a_i f(x_i). \quad (2)$$

Mai general, pentru orice $a'_i \geq 0$**, cu** $i = 1, \ldots, n$ **şi** $\sum_i a'_i \neq 0$ **avem**

$$f\left(\frac{\sum_i a'_i x_i}{\sum_j a'_j}\right) \leq \frac{\sum_i a'_i f(x_i)}{\sum_j a'_j}. \quad (3)$$

***Observaţii*:**
1. Dacă f **este strict convexă, atunci în relaţiile de mai sus egalitatea are loc doar dacă** $x_1 = \ldots = x_n$**.**
2. Evident, rezultate similare cu cele de mai sus pot fi formulate şi pentru funcţii concave, înlocuind în relaţiile (2) **şi** (3) **semnul** $\leq$ **cu** $\geq$**.**

b. Demonstraţi *inegalitatea mediilor* folosind inegalitatea lui Jensen:[4]

$$\frac{x_1 + x_2 + \ldots + x_n}{n} \geq \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} \quad \textbf{pentru orice } x_i \geq 0,\ i = 1, \ldots, n.$$

c. În contextul teoriei probabilităţilor, inegalitatea lui Jensen este exprimată astfel: dacă X **este o variabilă aleatoare şi** f **este o funcţie convexă, atunci** $f(E[X]) \leq E[f(X)]$**. (Similar, dacă** f **este funcţie concavă, atunci** $f(E[X]) \geq E[f(X)]$**.)**
Demontraţi această inegalitate în cazul în care X **este variabilă aleatoare discretă cu un număr finit de valori (adică,** $|Val\,(X)| < \infty$**).**

---

[2] Johan Jensen, inginer şi matematician danez (1859–1925).

[3] Dacă numerele a_i, cu $i = 1, \ldots, n$, satisfac cele două proprietăţi indicate, spunem că suma $\sum_i a_i x_i$ este o *combinaţie convexă* a numerelor $x_1, \ldots, x_n$. Atunci când se foloseşte doar prima condiţie ($a_i \geq 0,\ i = 1, \ldots, n$), suma respectivă se numeşte *combinaţie canonică*. În fine, dacă se renunţă la ambele condiţii, suma se numeşte *combinaţie afină*. (Cf. https://en.wikipedia.org/wiki/Linear_combination.)

[4] Primul membru al *inegalităţii mediilor* este media aritmetică, iar cel de-al doilea este media geometrică.

2. **(Variabile aleatoare: proprietăţi de bază pentru medii, varianţă, covarianţă)**

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

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

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

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

Arătaţi că pentru orice două variabile aleatoare W şi Z de acelaşi tip (adică fie ambele discrete fie ambele continue), având acelaşi domeniu de definiţie (Ω), avem

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

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

$$E[aX] = aE[X]. \tag{4}$$

Notaţi că aX este o variabilă aleatoare definită pe acelaşi domeniu (Ω) ca şi variabila X, cu proprietatea că $(aX)(\omega) \stackrel{def.}{=} a\,X(\omega)$ pentru orice $\omega \in \Omega$.

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

$$E[a_1X_1 + \ldots + a_nX_n] = a_1E[X_1] + \ldots + a_nE[X_n].$$

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

b. Fie X o variabilă aleatoare. Notăm $\bar{X} = E[X]$. *Varianţa* lui X se defineşte ca fiind $Var\,(X) = E[(X - \bar{X})^2]$. Arătaţi că:

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

***Observaţie importantă*: Veţi vedea că această proprietate este adeseori folosită (în locul definiţiei varianţei) în diverse demonstraţii care vor urma.**

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

$$Var\,(aX) = a^2\,Var\,(X). \tag{5}$$

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

***Indicaţie*: La acest punct nu este necesar să faceţi demonstraţiile separat pentru cele două cazuri, discret şi respectiv continuu.**

c. *Covarianţa* a două variabile aleatoare X şi Y care au acelaşi domeniu de definiţie (Ω) se defineşte astfel: $Cov\,(X, Y) = E[(X - E[X])(Y - E[Y])]$, unde $E[X]$ este media lui X. (Este imediat faptul că noţiunea de covarianţă generalizează noţiunea de varianţă.)

Demonstraţi egalitatea:

$$Cov\,(X,Y) = E[XY] - E[X]\cdot E[Y]. \tag{6}$$

Consecinţă imediată **(din relaţiile** (4) **şi** (6)**):**

$$Cov\,(aX,bY) = ab\,Cov\,(X,Y)\ \forall a,b\in\mathbb{R}. \tag{7}$$

Observaţii**:**

1. Este imediat faptul că proprietatea (7) **generalizează proprietatea** (5)**.**

2. Spre deosebire de varianţă, care poate lua doar valori mai mari sau egale cu 0 **(ceea ce decurge imediat din definiţia de la punctul** b**), covarianţa poate lua şi valori negative. Mai mult, se poate demonstra că atunci când** $Var\,(X)\neq 0$ **şi** $Var\,(Y)\neq 0$**, avem următoarele margini (una inferioară şi cealaltă superioară) pentru** $Cov\,(X,Y)$**:**

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

3. **(Distribuţia gaussiană unidimensională: verificarea condiţiilor de definiţie pentru p.d.f., calculul mediei şi al varianţei)**

Considerăm o variabilă aleatoare X **care urmează distribuţia normală (gaussiană):**

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

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

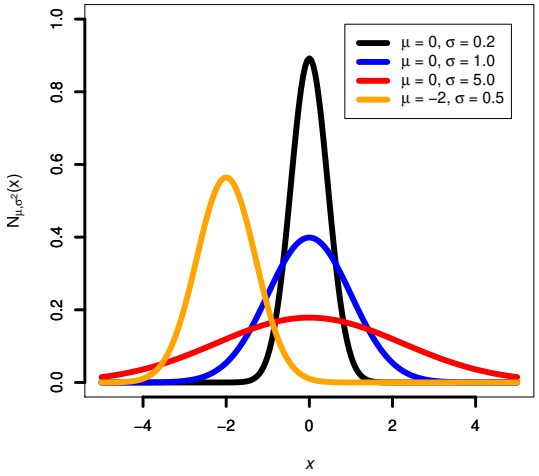
Arătaţi că:

a. $\mathcal{N}$ **este într-adevăr o funcţie de densitate de probabilitate (p.d.f.), adică** $\int_{-\infty}^{+\infty}\mathcal{N}(x)\,dx = 1$**.**

b. $E[X]=\mu$**.**

c. $Var[X]=\sigma^2$**.**

Distribuţia gaussiană: p.d.f.



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

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

Pentru demonstrarea ultimei egalităţi vă recomandăm să arătaţi că

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

deci

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

Ultima egalitate poate fi demonstrată prin trecerea din sistemul de coordonate cartezian în sistemul de coordonate polare. În mod concret, $(x,y)\mapsto(r,\theta)$, unde $r\in[0,+\infty)$ şi $\theta\in[0,2\pi)$, cu $x=r\cos\theta$ şi $y=r\sin\theta$, ceea ce constituie o corespondenţă bijectivă. Vă reamintim că *regula de schimbare de variabilă* pentru cazul vectorial, formulată (aici) pentru cazul probabilist, este următoarea:[5]
Presupunem că $V=[V_1\dots V_n]^\top\in\mathbb{R}^n$ este un vector de variabile aleatoare având funcţia de densitate de probabilitate comună $f_V:\mathbb{R}^n\to\mathbb{R}$. Dacă definim un alt vector de variabile aleatoare, Z, obţinut prin compunerea $Z=H\circ V$, unde $H:\mathbb{R}^n\to\mathbb{R}^n$ este o funcţie bijectivă şi derivabilă [pe componente] în raport cu fiecare dintre argumentele sale, atunci Z va avea funcţia de densitate de probabilitate comună $f_Z:\mathbb{R}^n\to\mathbb{R}$, unde

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

Matricea al cărei determinant este calculat în expresia de mai sus se numeşte matrice *jacobiană*, iar determinantul respectiv se numeşte *determinant jacobian*, de la numele matematicianului german Carl Gustav Jacob Jacobi (1804 – 1851).

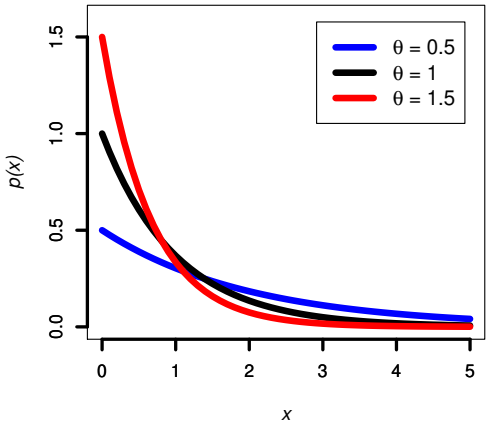
4. **(Distribuţia exponenţială şi distribuţia Gamma: verificarea condiţiilor de definiţie pentru p.d.f., calculul mediilor şi al varianţelor)**

a. Distribuţia *exponenţială* este o distribuţie continuă, care are funcţia densitate de probabilitate

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

unde $\theta>0$ este un parametru real. Arătaţi mai întâi că funcţia $p(\ )$ este într-adevăr funcţie densitate de probabilitate (engl., probability density function, p.d.f.) şi apoi că media şi respectiv varianţa distribuţiei exponenţiale sunt $\frac{1}{\theta}$ şi respectiv $\frac{1}{\theta^2}$.

Distribuţia exponenţială: p.d.f.



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

Formula de definiţie a acestei funcţii este următoarea:

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

[5] Cf. *The Multivariate Gaussian Distribution*, Chuong Do, Stanford University, 2008.

Demontraţi următoarele proprietăţi ale funcţiei Γ:

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

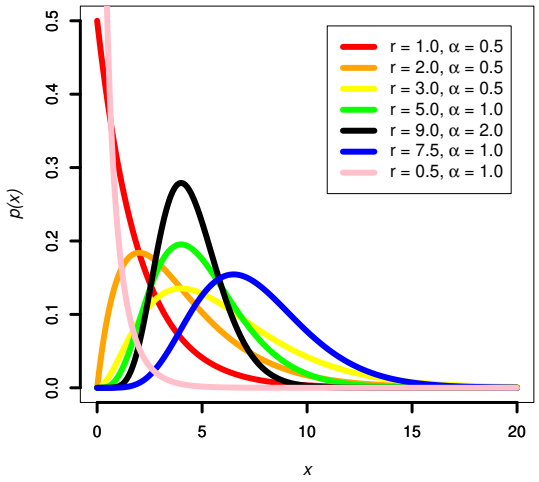
ii. $\Gamma(1) = 1$ **şi** $\Gamma(1/2) = \sqrt{\pi}$.[6]

c. Distribuţia *Gamma* este o distribuţie continuă, de parametri $r > 0$ (care dă *forma* distribuţiei, engl., shape) şi $\alpha > 0$ (numit *rata*, engl., rate), cu funcţia densitate de probabilitate definită pe $\mathbb{R}^+$ prin expresia următoare:

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

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

Distribuţia Gamma: p.d.f.



Demonstraţi mai întâi că funcţia $p(\,)$ este într-adevăr p.d.f. şi apoi că media distribuţiei Gamma este $\frac{r}{\alpha}$, iar varianţa ei este $\frac{r}{\alpha^2}$.

***Sugestie*: Puteţi folosi următoarea proprietate, care este demonstrată la problema 3:**

$$\int_{-\infty}^{\infty} v^2\, e^{-\frac{v^2}{2}}\, dv = 2 \int_{0}^{\infty} v^2\, e^{-\frac{v^2}{2}}\, dv = \sqrt{2\pi}. \tag{8}$$

***Observaţie*: Se poate vedea imediat că $Gamma(x|1,\alpha) = \alpha e^{-\alpha x}$ pentru orice $x \geq 0$, ceea ce corespunde [definiţiei funcţiei de densitate a] distribuţiei exponenţiale (vedeţi punctul a). Aşadar, se poate spune că distribuţia exponenţială este membru al familiei de distribuţii Gamma.**

5. **(Distribuţii de tip Bernoulli; calculul verosimilităţii datelor; estimarea parametrilor în sensul MLE)**

CMU, 2005 fall, T. Mitchell, A. Moore, midterm, pr. 1.3

Avem două monede. Probabilitatea de apariţie a stemei este θ în cazul primei monede şi 2θ în cazul celei de-a doua monede.
Presupunem că aruncăm aceste două monede de mai multe ori, în mod independent una de cealaltă, şi obţinem rezultatele din tabelul alăturat.

Moneda	Rezultat
1	stemă
2	ban
2	ban
2	ban
2	stemă

a. Care este log-verosimilitatea acestor date în funcţie de θ?

b. Cât este estimarea / valoarea de verosimilitate maximă (engl., maximum likelihood, ML) a lui θ?

[6] Din relaţia de recurenţă *i.* şi din relaţia $\Gamma(1) = 1$, rezultă $\Gamma(r+1) = r \cdot \Gamma(n) = r \cdot (r-1) \cdot \Gamma(r-1) = \ldots = r \cdot (r-1) \cdot \ldots \cdot 2 \cdot 1 = r!$. Aşadar, într-adevăr funcţia Γ a lui Euler generalizează noţiunea de *produs factorial*.

6. **(Entropie, entropie comună, entropie condiţională, câştig de informaţie: definiţii şi proprietăţi imediate)**

Fie X şi Y variabile aleatoare discrete. Dăm pe scurt următoarele *definiţii*:

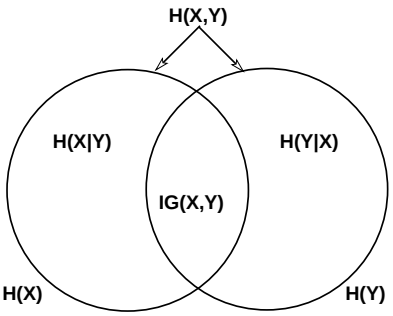
- **Entropia variabilei X:**
  $H(X) \stackrel{def.}{=} -\sum_i P(X = x_i) \log P(X = x_i) \stackrel{not.}{=} E_X[-\log P(X)]$.
  Prin *convenţie*, dacă $p(x) = 0$ atunci vom considera $p(x) \log p(x) = 0$.
- **Entropia condiţională specifică a variabilei Y în raport cu valoarea x_k a variabilei X:**
  $H(Y \mid X = x_k) \stackrel{def.}{=} -\sum_j P(Y = y_j \mid X = x_k) \log P(Y = y_j \mid X = x_k)$
  $\stackrel{not.}{=} E_{Y|X=x_k}[-\log P(Y \mid X = x_k)]$.
- **Entropia condiţională medie a variabilei Y în raport cu variabila X:**
  $H(Y \mid X) \stackrel{def.}{=} \sum_k P(X = x_k) H(Y \mid X = x_k) \stackrel{not.}{=} E_X[H(Y \mid X)]$.
- **Entropia comună a variabilelor X şi Y:**
  $H(X, Y) \stackrel{def.}{=} -\sum_i \sum_j P(X = x_i, Y = y_j) \log P(X = x_i, Y = y_j)$
  $\stackrel{not.}{=} E_{X,Y}[-\log P(X, Y)]$.
- ***Câştigul de informaţie* al variabilei X în raport cu variabila Y (sau invers), numit de asemenea *informaţia mutuală* a variabilelor X şi Y:**
  $IG(X, Y) \stackrel{not.}{=} MI(X, Y) \stackrel{def.}{=} H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$.
  (Observaţie: ultima egalitate de mai sus are loc datorită rezultatului de la punctul c de mai jos.)

Arătaţi că:

a. $H(X) \geq 0$. **În particular,** $H(X) = 0$ **dacă şi numai dacă variabila X este constantă.**

b. $H(Y \mid X) = -\sum_i \sum_j P(X = x_i, Y = y_j) \log P(Y = y_j \mid X = x_i)$.

c. $H(X, Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y)$.
Mai general: $H(X_1, \ldots, X_n) = H(X_1) + H(X_2 \mid X_1) + \ldots + H(X_n \mid X_1, \ldots, X_{n-1})$ **(regula de înlănţuire).**
***Observaţie*: Relaţia precedentă, precum şi relaţia de definiţie pentru câştigul de informaţie sunt ilustrate în figura alăturată.**



7. **(Entropie, entropie condiţională specifică, câştig de informaţie: exemplificare)**

Problema aceasta se referă la aruncarea a două zaruri perfecte, cu 6 feţe.

a. Calculează distribuţia probabilistă a sumei numerelor de pe cele două feţe care au fost obţinute / „observate“ în urma aruncării zarurilor.

În continuare, suma aceasta va fi asimilată cu o variabilă aleatoare, notată cu S.

b. Cantitatea de *informaţie* obţinută (sau: *surpriza* pe care o resimţim) la „observarea“ producerii valorii x a unei variabile aleatoare X oarecare este prin *definiţie*

$$\mathit{Information}(P(X = x)) = \mathit{Surprise}(P(X = x)) = \log_2 \frac{1}{P(X = x)} = -\log_2 P(X = x).$$

Această cantitate este exprimată (numeric) în *biţi de informaţie*.

Cât de surprins vei fi atunci când vei „observa“ $S = 2$, respectiv $S = 11$, $S = 5$ şi $S = 7$? (Vei exprima de fiecare dată rezultatul în biţi. Puteţi folosi $\log_2 3 = 1.584962501$.)

c. Calculează entropia variabilei S.

d. Să presupunem acum că vei arunca aceste două zaruri pe rând, iar la aruncarea primului zar se obţine numărul 4. Cât este entropia lui S în urma acestei „observaţii“? S-a pierdut, ori s-a câştigat informaţie în acest proces? Calculează cât de multă informaţie (exprimată în biţi) s-a pierdut ori s-a câştigat.

8. (Câştigul de informaţie / informaţia mutuală, o aplicaţie: selecţia de trăsături)

În tabelul următor se dă un set de opt observaţii / instanţe, reprezentate ca tupluri de valori ale variabilelor aleatoare binare de „intrare“ X_1, X_2, X_3, X_4, X_5 şi ale variabilei aleatoare binare de „ieşire“ Y.

Am dori să reducem spaţiul de trăsături $\{X_1, X_2, X_3, X_4, X_5\}$ folosind o metodă de selecţie de tip *filtru*.

a. Calculaţi câştigul de informaţie / informaţia mutuală $MI(X_i, Y)$ pentru fiecare i.

b. Ţinând cont de rezultatul de la punctul precedent, alegeţi cel mai mic subset de trăsături în aşa fel încât cel mai bun clasificator antrenat pe acest spaţiu (redus) de trăsături să fie cel puţin la fel de bun ca şi cel mai bun clasificator antrenat pe întreg spaţiul de trăsături. Justificaţi alegerea pe care aţi făcut-o.

X_1	X_2	X_3	X_4	X_5	Y
0	1	1	0	1	0
1	0	0	0	1	0
0	1	0	1	0	1
1	1	1	1	0	1
0	1	1	0	0	1
0	0	0	1	1	1
1	0	0	1	0	1
1	1	1	0	1	1

9. (O margine superioară pentru valorea entropiei unei variabile aleatoare discrete)

***Comentariu*:** La problema 6.a am demonstrat că entropia oricărei variabile aleatoare discrete este nenegativă ($H(X) \geq 0$).[7] La acest exerciţiu veţi demonstra — tot pentru cazul discret — că există şi o margine superioară pentru $H(X)$.

Aşadar, fie X o variabilă aleatoare discretă care ia n valori şi urmează distribuţia probabilistă P. Conform definiţiei, entropia lui X este

$$H(X) = -\sum_{i=1}^{n} P(X = x_i) \log_2 P(X = x_i).$$

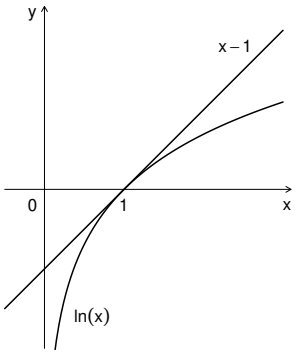
[7] Extensia acestei proprietăţi la cazul variabilelor aleatoare continue este imediată.

Arătaţi că $H(X) \leq \log_2 n$.

***Sugestie* (1): Puteţi folosi inegalitatea $\ln x \leq x - 1$, care are loc pentru orice $x > 0$.**

***Sugestie* (2): Puteţi folosi inegalitatea lui Jensen (vedeţi problema 1).**

***Sugestie* (3): Puteţi folosi metoda multiplicatorilor lui Lagrange.**



10. **(Entropia comună: forma particulară a relaţiei de „înlănţuire“ în cazul variabilelor aleatoare independente)**

□ ***prelucrare de Liviu Ciortuz, după CMU, 2012 spring, Roni Rosenfeld, HW2, pr. 7.b***

Conform problemei 6.c, *formula de înlănţuire* a entropiilor pentru cazul general (adică, indiferent dacă X şi Y sunt sau nu independente) este:

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y). \tag{9}$$

Demonstraţi că dacă X şi Y sunt variabile aleatoare independente discrete, atunci $H(X,Y) = H(X) + H(Y)$, şi reciproc: atunci când are loc egalitatea $H(X,Y) = H(X) + H(Y)$ rezultă că variabilele X şi Y sunt independente.

11. **(Nenegativitatea câştigului de informaţie; o condiţie necesară şi suficientă pentru anularea lui)**

Definiţia *câştigului de informaţie* (sau: a *informaţiei mutuale*) al unei variabile aleatoare X în raport cu o altă variabilă aleatoare Y este $IG(X,Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$.[8] La problema 15 s-a demonstrat — pentru cazul în care X şi Y sunt discrete — că $IG(X,Y) = KL(P_{X,Y}||P_X\, P_Y)$, unde KL desemnează *entropia relativă* (sau: *divergenţa Kullback-Leibler*), P_X şi P_Y sunt distribuţiile variabilelor X şi, respectiv, Y, iar $P_{X,Y}$ este distribuţia comună a acestor variabile. Tot la problema 15 s-a arătat că divergenţa KL este întotdeauna nenegativă. În consecinţă, $IG(X,Y) \geq 0$ pentru orice X şi Y.

a. Aici vă cerem să demonstraţi inegalitatea $IG(X,Y) \geq 0$ în manieră directă, plecând de la prima definiţie dată mai sus, fără a mai apela la divergenţa Kullback-Leibler.

b. Arătaţi tot într-o manieră directă că $IG(X,Y) = 0$ dacă şi numai dacă X şi Y sunt independente. (Într-o manieră indirectă, acest rezultat a fost demonstrat la problema 15.c.)

***Sugestie*: Puteţi folosi următoarea formă [particulară] a inegalităţii lui Jensen:[9]**

$$\sum_{i=1}^{n} a_i \log x_i \leq \log \left(\sum_{i=1}^{n} a_i x_i \right)$$

[8] Vedeţi problema 6.
[9] Vedeţi problema 1.

unde baza logaritmului se consideră supraunitară, $a_i \geq 0$ pentru $i = 1, \ldots, n$ şi $\sum_{i=1}^{n} a_i = 1$.[10] Egalitatea are loc dacă şi numai dacă $x_1 = x_2 = \ldots = x_n$.

12. (Entropia distribuţiei continue uniforme şi a distribuţiei gaussiene uni-dimensionale)

a. Calculaţi entropia distribuţiei uniforme continue definite pe intervalul $[a, b]$.

b. Calculaţi entropia distribuţiei gaussiene unidimensionale de parametri $\mu \in \mathbb{R}$ şi $\sigma^2 \in \mathbb{R}_+$, pentru care funcţia de densitate de probabilitate (p.d.f.) este:

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ pentru } x \in \mathbb{R}.$$

13. (Calcularea entropiei unor variabile aleatoare continue: cazul distribuţiei exponenţiale şi al distribuţiei Gamma)

Pentru o variabilă aleatoare X care urmează o distribuţie continuă cu funcţia densitate de probabilitate (p.d.f.) p, entropia se defineşte astfel:

$$H(X) = \int_{-\infty}^{+\infty} p(x) \log_2 \frac{1}{p(x)} dx \overset{p(x)\neq 0}{=} -\int_{-\infty}^{+\infty} p(x) \log_2 p(x) dx.$$

***Indicaţie*: Dacă $p(x) = 0$, veţi presupune că $-p(x) \log_2 p(x) = 0$.**

a. Calculaţi entropia *distribuţiei* continue *exponenţiale* de parametru $\lambda > 0$. Vă reamintim că definiţia p.d.f.-ului acestei distribuţii este următoarea:[11]

$$p(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{dacă } x \geq 0; \\ 0, & \text{dacă } x < 0. \end{cases}$$

b. Calculaţi entropia distribuţiei Gamma, pentru care funcţia de densitate de probabilitate este:[12]

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

cu $r > 0$ (forma), $\alpha > 0$ (rata) şi $\Gamma(r) \overset{def.}{=} \int_0^{+\infty} x^{r-1} e^{-x}\, dx$ pentru orice $r > 0$ (funcţia lui Euler). (Vedeţi ex. 4.b.)

***Indicaţie*:**

Este posibil să fie nevoie să folosiţi o proprietate specială, numită *regula de derivare sub semnul de integrală* (engl., differentiation under the integral sign), care constituie obiectul pentru următoarea

***Teoremă*: Dacă $f(r, t)$ este o funcţie cu valori reale, continuă şi derivabilă în raport**

[10] Avantajul la această problemă, comparativ cu problema 15.a, este că aici se lucrează cu o singură distribuţie (p), nu cu două distribuţii (p şi q). Totuşi, demonstraţia de aici va fi mai laborioasă.

[11] La ex. 4.a puteţi vedea graficul acestei funcţii de densitate pentru câteva valori ale parametrului (λ).

[12] La ex. 4.c puteţi vedea graficul acestei funcţii de densitate pentru câteva valori ale parametrilor r şi α.

cu r pe intervalul (a, b), iar derivata parţială $\dfrac{\partial}{\partial r} f(r, t)$ este de asemenea continuă pe intervalul (a, b), atunci[13]

$$\int_a^b \frac{\partial}{\partial r} f(r, t)\, dt = \frac{\partial}{\partial r} \int_a^b f(r, t)\, dt.$$

Menţionăm că regula aceasta, datorată lui Gottfried Leibnitz (1646-1716), se generalizează — sub o formă semnificativ mai elaborată — la cazul când limitele de integrare a şi b depind de r, adică sunt de forma $a(r)$ şi respectiv $b(r)$, iar aceste două funcţii sunt continue şi au derivatele continue.

14. (Redescoperirea definiţiei entropiei pornind de la un set de *proprietăţi dezirabile* ale ei)

Prin definiţie, *entropia* (în sens Shannon) a unei variabile aleatoare discrete X ale cărei valori sunt luate cu probabilităţile $p_1, p_2, \ldots, p_n$ este $H(X) = -\sum_i p_i \log p_i$. Însă legătura dintre această definiţie formală şi obiectivul avut în vedere — şi anume, acela de a exprima gradul de *incertitudine* cu care se produc valorile unei astfel de variabile aleatoare — nu este foarte intuitivă.

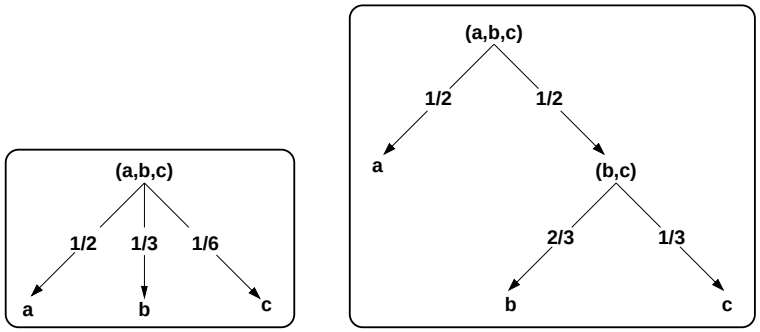
Scopul acestui exerciţiu este de a arăta că orice funcţie $\psi_n(p_1, \ldots, p_n)$ care satisface trei proprietăţi dezirabile pentru entropie este în mod necesar de forma $-K \sum_i p_i \log p_i$ unde K este o constantă reală pozitivă. Iată care sunt aceste *proprietăţi*:[14]

- **A1. Funcţia $\psi_n(p_1, \ldots, p_n)$ este continuă în fiecare din argumentele ei şi *simetrică*. Din punct de vedere formal, în acest caz simetria se traduce prin egalitatea $\psi_n(p_1, \ldots, p_i, \ldots, p_j, \ldots, p_n) = \psi_n(p_1, \ldots, p_j, \ldots, p_i, \ldots, p_n)$ pentru orice $i \neq j$. Informal spus, dacă două dintre valorile care sunt luate de variabila aleatoare X (şi anume x_i şi x_j) îşi schimbă între ele probabilităţile (p_i şi respectiv p_j), valoarea entropiei lui X nu se schimbă.**
- **A2. Funcţia $\psi_n(1/n, \ldots, 1/n)$ este monoton *crescătoare* în raport cu n. Altfel spus, dacă toate evenimentele sunt echiprobabile, atunci entropia creşte odată cu numărul de evenimente posibile.**
- **A3. Dacă faptul de a alege între mai multe evenimente posibile poate fi realizat prin mai multe alegeri succesive, atunci $\psi_n(p_1, \ldots, p_n)$ trebuie să se poată scrie ca o sumă ponderată a entropiilor calculate la fiecare stadiu / alegere.**
  De *exemplu*, dacă evenimentele (a, b, c) se produc respectiv cu probabilităţile $(1/2, 1/3, 1/6)$, atunci acest fapt poate fi echivalat cu
  - **a alege mai întâi cu probabilitate de $1/2$ între a şi (b, c),**
  - **urmat de a alege între b şi c cu probabilităţile $2/3$ şi $1/3$ respectiv. (A se vedea imaginile de mai jos, Encoding 1 şi Encoding 2.)**

  Din punct de vedere formal, proprietatea A3 impune ca, pe acest exemplu, $\psi_3(1/2, 1/3, 1/6)$ să fie egal cu $\psi_2(1/2, 1/2) + 1/2 \cdot \psi_2(2/3, 1/3)$.

[13] Pentru demonstraţie, vedeţi „Proof of basic form“ pe site-ul https://en.wikipedia.org/wiki/Leibniz_integral_rule#Alternative_Proof_of_General_Form_with_Variable_Limits,_using_the_Chain_Rule (accesat la data de 15.12.2020).

[14] LC: Deşi nu se specifică în enunţul original al problemei, este necesar / natural să considerăm şi proprietatea următoare: [A0.] $\psi_n(p_1, \ldots, p_n) \geq 0$ pentru orice $n \in \mathbb{N}^*$ şi orice $p_1, \ldots, p_n \in [0, 1]$ astfel încât $\sum_i p_i = 1$, pentru că ψ_n este văzută că măsură a *dezordinii*; de asemenea, $\psi_1(1) = 0$ pentru că în acest caz particular nu există niciun fel de dezordine.

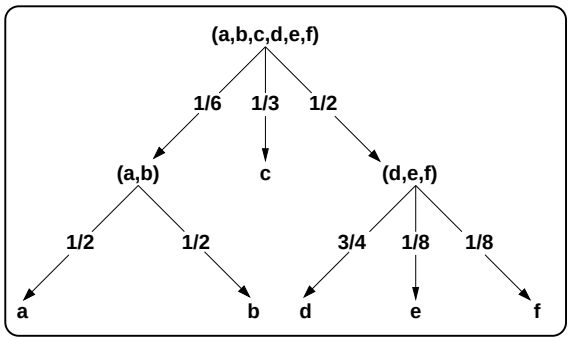


Aşadar, în acest exerciţiu vi se cere să arătaţi că dacă o funcţie de n variabile $\psi_n(p_1, \ldots, p_n)$ satisface proprietăţile A1, A2 şi A3 de mai sus, atunci există $K \in \mathbb{R}^+$ astfel încât $\psi_n(p_1, \ldots, p_n) = -K \sum_i p_i \log p_i$ unde $K \in \mathbb{R}_+$ este o constantă.[15]

***Indicaţie*:**
Veţi face rezolvarea acestei probleme în mod gradual, parcurgând următoarele puncte (dintre care primele două puncte au rolul de a vă acomoda cu noţiunile din enunţ):

a. Arătaţi că $H(1/2, 1/3, 1/6) = H(1/2, 1/2) + \frac{1}{2} H(2/3, 1/3)$. Altfel spus, verificaţi faptul că definiţia clasică a entropiei, $H(X) = \sum_i p_i \log 1/p_i$, satisface proprietatea A3 pe *exemplul* care a fost dat mai sus.

b. Calculaţi entropia în cazul distribuţiei / „codificării" din figura de mai jos, folosind din nou proprietatea A3.



Următoarele întrebări tratează cazul particular $A(n) \overset{not.}{=} \psi(1/n, 1/n, \ldots, 1/n)$.

c. Arătaţi că

$$A(s^m) = m\, A(s) \text{ pentru orice } s, m \in \mathbb{N}^*. \tag{10}$$

La punctele $d - g$ de mai jos, pentru orice pereche de numere $s, m \in \mathbb{N}^*$ fixate, cu $s > 1$, vom considera $t, n \in \mathbb{N}^*$ astfel încât

$$s^m \leq t^n \leq s^{m+1}. \tag{11}$$

[15] În demonstraţie, vom vedea, va rezulta $K = \frac{1}{\log s} \psi_s\left(\frac{1}{s}, \ldots, \frac{1}{s}\right)$ pentru un număr oarecare $s \in \mathbb{N}^* \backslash \{1\}$, fixat.

d. Verificaţi că, prin logaritmarea acestei duble inegalităţi şi apoi prin rearanjare, obţinem $\frac{m}{n} \leq \frac{\log t}{\log s} \leq \frac{m}{n} + \frac{1}{n}$ **pentru** $s \neq 1$**, şi deci**

$$\left| \frac{m}{n} - \frac{\log t}{\log s} \right| \leq \frac{1}{n}. \tag{12}$$

e. Explicaţi de ce $A(s^m) \leq A(t^n) \leq A(s^{m+1})$**.**

f. Combinând ultima inegalitate de mai sus cu egalitatea (10)**, avem** $A(s^m) \leq A(t^n) \leq A(s^{m+1}) \Rightarrow m\,A(s) \leq n\,A(t) \leq (m+1)\,A(s)$**. Verificaţi că**

$$\left| \frac{m}{n} - \frac{A(t)}{A(s)} \right| \leq \frac{1}{n} \text{ pentru } s \neq 1. \tag{13}$$

g. Combinând inegalităţile (12) **şi** (13)**, arătaţi că**

$$\left| \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \right| \leq \frac{2}{n} \text{ pentru } s \neq 1 \tag{14}$$

şi, în consecinţă

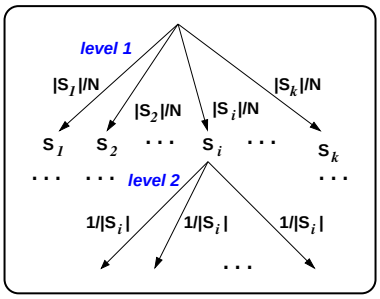
$$A(t) = K \log t \text{ cu } K > 0 \text{ (din cauza proprietăţii A2)}. \tag{15}$$

***Observaţie*:**

Rezultatul de mai sus ($A(t) = K \log t \Leftrightarrow \psi_t(1/t, \ldots, 1/t) = Kt\frac{1}{t}\log t$) **se generalizează uşor la cazul** $\psi(p_1, \ldots, p_k)$ **cu** $p_i \in \mathbb{Q}$ **pentru** $i = 1, \ldots, k$**.**

Considerăm o mulţime de N **evenimente echiprobabile. Fie** $\mathcal{P} = (S_1, S_2, \ldots, S_k)$ **o partiţionare a acestei mulţimi de evenimente. Notăm** $p_i = \mid S_i \mid /N$**.**

Propunem *codificarea* din figura alăturată. Vom alege mai întâi S_i**, una dintre submulţimile din partiţia** $\mathcal{P}$**, în funcţie de probabilităţile** $p_1, \ldots, p_k$**. Extragem apoi unul dintre elementele mulţimii** S_i**, cu probabilitate uniformă.**



Conform egalităţii (15)**, avem** $A(N) = K\,logN$**. Folosind proprietatea A3 şi *codificarea* în doi paşi propusă mai sus, rezultă că**

$$A(N) = \psi_k(p_1, \ldots, p_k) + \textstyle\sum_i p_i A(\mid S_i \mid).$$

Aşadar,

$$K \log N = \psi_k(p_1, \ldots, p_k) + K \textstyle\sum_i p_i \log \mid S_i \mid.$$

Prin urmare,

$$\begin{aligned} \psi_k(p_1, \ldots, p_k) &= K[\,\log N - \sum_i p_i \log \mid S_i \mid\,] = K[\,(\log N)\sum_i p_i - \sum_i p_i \log \mid S_i \mid\,] \\ &= -K \sum_i p_i \log \frac{\mid S_i \mid}{N} = -K \sum_i p_i \log p_i \end{aligned}$$

Putem trata în sfârşit şi cazul $p_i \in \mathbb{R}$**, cu** $\sum_i p_i = 1$**.**[16] **Considerăm pentru fiecare** $p_i,\ i = 1, \ldots, k-1$ **un şir de aproximări succesive** $\{q_i^{(n)}\}_{n \geq 1} \subset \mathbb{Q}$ **astfel încât** $\lim_{n \to \infty} q_i^{(n)} = p_i$**.**

[16] Cf. Ştefan Bălăucă, student, FII, 2020f.

(De exemplu, putem defini $q_i^{(n)}$ = numărul format luând în considerare doar din primele n zecimale ale lui p_i.)

Definim de asemenea şirul $\{q_k^{(n)}\}_{n\geq 1} \subset \mathbb{Q}$ astfel: $q_k^{(n)} = 1 - \sum_{i=1}^{k-1} q_i^{(n)}$. Trecând la limită în această relaţie de definiţie, obţinem:

$$\lim q_k^{(n)} = 1 - \sum_{i=1}^{k-1} \lim q_i^{(n)} = 1 - \sum_{i=1}^{k-1} p_i = p_k.$$

Aşadar, avem $\lim\limits_{n\to\infty} q_i^{(n)} = p_i,\ \forall i = 1, \ldots, k$. Deoarece $q_i^{(n)} \in \mathbb{Q}$, pentru $i = 1, \ldots, k$, cu $k \geq 1$, este valabilă relaţia

$$\psi(q_1^{(n)}, \ldots, q_k^{(n)}) = -K \sum_i q_i^{(n)} \ln q_i^{(n)},\ \forall n \geq 1.$$

Trecând la limită în această relaţie, vom avea:

$$\lim_{n\to\infty} \psi(q_1^{(n)}, \ldots, q_k^{(n)}) = -K \sum_i \lim_{n\to\infty} q_i^{(n)} \ln q_i^{(n)} = -K \sum_i p_i \ln p_i. \tag{16}$$

Datorită proprietăţii A1, funcţia ψ este continuă în fiecare dintre argumentele ei, astfel că

$$\lim_{n\to\infty} \psi(q_1^{(n)}, \ldots, q_k^{(n)}) = \psi(\lim_{n\to\infty} q_1^{(n)}, \ldots, \lim_{n\to\infty} q_k^{(n)}) = \psi(p_1, \ldots, p_k). \tag{17}$$

Din relaţiile (16) şi (17) putem concluziona că

$$\psi(p_1, ..., p_k) = -K \sum_i p_i \ln p_i,\ \textbf{pentru}\ \forall p_i \in [0, 1],\ \textbf{a.î.}\ \sum_i p_i = 1.$$

15. (Entropia relativă: definiţie şi proprietăţi elementare; exprimarea câştigului de informaţie cu ajutorul entropiei relative)

***Entropia relativă* sau *divergenţa Kullback-Leibler* (KL) a unei distribuţii p în raport cu o altă distribuţie q — ambele distribuţii fiind discrete — se defineşte astfel:**

$$KL(p||q) \stackrel{def.}{=} -\sum_{x\in X} p(x) \log \frac{q(x)}{p(x)}$$

Din perspectiva teoriei informaţiei, divergenţa KL specifică numărul de *biţi adiţionali* care sunt necesari în medie pentru a transmite valorile variabilei X atunci când presupunem că aceste valori sunt distribuite conform distribuţiei („model“) q, dar în realitate ele urmează o altă distribuţie, p.[17]

[17] *Atenţie*: Divergenţa KL nu este o măsură de *distanţă* între două distribuţii probabiliste, fiindcă în general ea nu este simetrică ($KL(p||q) \neq KL(q||p)$) şi nici nu satisface inegalitatea triunghiului. Pentru „simetrizare“, se consideră $M(p, q) = \frac{1}{2}(p+q)$, apoi se defineşte funcţia $JSD(p||q) = \frac{1}{2}KL(p||M) + \frac{1}{2}KL(q||M)$, care se numeşte *divergenţa Jensen-Shannon*. În sfârşit, se poate arăta că $\sqrt{JSD(p||q)}$ defineşte o măsură de distanţă (metrică), adică este nengativă, simetrică, implică identitatea indiscernabililor şi satisface inegalitatea triunghiului; ea este numită *distanţa Jensen-Shannon*.

Variaţia informaţiei, definită prin

$$VI(X, Y) \stackrel{def.}{=} H(X, Y) - IG(X, Y) = H(X) + H(Y) - 2IG(X, Y) = H(X \mid Y) + H(Y \mid X),$$

este de asemenea o măsură de distanţă.

a. Demonstraţi inegalitatea $KL(p||q) \geq 0$ şi apoi arătaţi că egalitatea are loc dacă şi numai dacă $p = q$.[18]

***Indicaţie*:**
Pentru a demonstra punctul acesta puteţi folosi *inegalitatea lui Jensen*:[19]
Dacă $f : \mathbb{R} \to \mathbb{R}$ este o *funcţie convexă*, atunci pentru orice $a_i \geq 0,\ i = 1, \ldots, n$ cu $\sum_i a_i = 1$ şi orice $x_i \in \mathbb{R},\ i = 1, \ldots, n$, avem $f(\sum_i a_i x_i) \leq \sum_i a_i f(x_i)$. Dacă f este strict convexă, atunci egalitatea are loc doar dacă $x_1 = \ldots = x_n$. Pentru funcţii concave, semnul inegalităţii este $\geq$.

b. Câştigul de informaţie poate fi (re)definit ca fiind entropia relativă dintre distribuţia comună observată a lui X şi Y pe de o parte, şi produsul distribuţiilor marginale p_X şi p_Y de cealaltă parte:

$$\begin{aligned} IG(X,Y) &\overset{def.}{=} KL(p_{X,Y} \,||\, (p_X\, p_Y)) = -\sum_x \sum_y p_{X,Y}(x,y) \log\left(\frac{p_X(x)p_Y(y)}{p_{X,Y}(x,y)}\right) \\ &\overset{not.}{=} -\sum_x \sum_y p(x,y) \log\left(\frac{p(x)p(y)}{p(x,y)}\right) \end{aligned}$$

Arătaţi că această nouă definiţie a câştigului de informaţie este echivalentă cu definiţia dată anterior (vedeţi problema 6). Cu alte cuvinte, demonstraţi egalitatea

$$KL(p_{X,Y} \,||\, (p_X\, p_Y)) = H[X] - H[X \mid Y].$$

***Observaţie*: Din noua definiţie introdusă mai sus pentru câştigul de informaţie, rezultă imediat că**

$$\begin{aligned} IG(X,Y) &= \sum_y p(y) \sum_x p(x \mid y) \log \frac{p(x \mid y)}{p(x)} = \sum_y p(y) KL(p_{X|Y} \,||\, p_X) \\ &= E_Y[KL(p_{X|Y} \,||\, p_X)] \end{aligned}$$

ceea ce înseamnă că $IG(X,Y)$ poate fi văzută ca o medie (în raport cu distribuţia lui Y) a divergenţei KL dintre distribuţia condiţională a lui X în raport cu Y pe de o parte, şi distribuţia lui X pe de altă parte.

c. O consecinţă imediată a punctelor a şi b este faptul că $IG(X,Y) \geq 0$ (deci $H(X) \geq H(X|Y)$ şi $H(Y) \geq H(Y|X)$) pentru orice variabile aleatoare discrete X şi Y. Folosind din nou rezultatele de la punctele a şi b, arătaţi că $IG(X,Y) = 0$ dacă şi numai dacă X şi Y sunt independente.

16. (Cross-entropie: definiţie, o proprietate (nenegativitatea) şi un exemplu simplu de calculare a valorii cross-entropiei)

***Cross-entropia* a două distribuţii p şi q, desemnată prin *CH(p,q)*, reprezintă numărul mediu de biţi necesari pentru a codifica un eveniment dintr-o mulţime oarecare de posibilităţi, atunci când schema de cod[ific]are folosită se bazează pe o distribuţie de probabilitate dată q, în loc să se bazeze pe distribuţia „adevărată“ p. În cazul în care distribuţiile p şi q sunt discrete, această noţiune se defineşte formal astfel:**

$$CH(p,q) = -\sum_x p(x)\, \log q(x).$$

[18] Mai general, $KL(p||q)$ este cu atât mai mică cu cât „asemănarea“ dintre distribuţiile p şi q este mai mare.
[19] Vedeţi problema 1.

În cazul distribuţiilor continue, definiţia se obţine / construieşte prin analogie:

$$CH(p,q) = -\int_X p(x)\, \log q(x)\, dx.$$

***Observaţie*: Ţinând cont de definiţia entropiei relative (cunoscută şi sub numele de divergenţa Kullback-Leibler), vedeţi pr. 15, putem scrie:**

$$KL\,(p||q) = CH(p,q) - H(p).$$

Cross-entropia — ca şi entropia relativă; vedeţi problema 15 —, spre deosebire de entropia comună, nu este simetrică în raport cu cele două distribuţii / argumente: în general, $CH(p,q) \neq CH(q,p)$.

a. Poate oare cross-entropia să ia valori negative? Faceţi o demonstraţie sau daţi un contraexamplu.

b. În multe experimente, pentru a stabili calitatea diferitelor ipoteze / modele, se procedează la evaluarea / compararea lor pe un set de date. Să presupunem că, urmărind să faci predicţia *funcţiei de probabilitate* asociate unei anumite *variabile aleatoare* care are 7 valori posibile, ai obţinut (printr-un procedeu oarecare) două *modele* diferite, iar *distribuţiile de probabilitate* prezise de către aceste două modele sunt respectiv:

$$q_1 = \left(\frac{1}{10}, \frac{1}{10}, \frac{1}{5}, \frac{3}{10}, \frac{1}{5}, \frac{1}{20}, \frac{1}{20}\right) \text{ şi } q_2 = \left(\frac{1}{20}, \frac{1}{10}, \frac{3}{20}, \frac{7}{20}, \frac{1}{5}, \frac{1}{10}, \frac{1}{20}\right).$$

Să zicem că pentru evaluare foloseşti un set de date caracterizat de următoarea distribuţie *empirică*:

$$p_{empiric} = \left(\frac{1}{20}, \frac{1}{10}, \frac{1}{5}, \frac{3}{10}, \frac{1}{5}, \frac{1}{10}, \frac{1}{20}\right).$$

Calculează cross-entropiile $CH(p_{empiric}\,, q_1)$ şi $CH(p_{empiric}\,, q_2)$.
Care dintre aceste două modele va conduce la o cross-entropie mai mică? Putem oare garanta că acest model este într-adevăr [cel] mai bun? Explică / justifică răspunsul [pe care l-ai] dat.

17. **(Inegalitatea lui Gibbs: un caz particular; comparaţie între valorile entropiei şi ale cross-entropiei)**

Fie $P = \{p_1, \ldots p_n\}$ o distribuţie de probabilitate discretă.

a. Arătaţi că pentru orice distribuţie de probabilitate $Q = \{q_1, \ldots q_n\}$ are loc inegalitatea:

$$-\sum_{i=1}^{n} p_i \log_2 p_i \leq -\sum_{i=1}^{n} p_i \log_2 q_i$$

Altfel spus, $H(P) \leq CH(P,Q)$, unde $H(P)$ este entropia distribuţiei P, iar $CH(P,Q)$ este *cross-entropia* lui P în raport cu Q.

b. Arătaţi că în formula de mai sus egalitatea are loc dacă şi numai dacă $p_i = q_i$ pentru $i = 1, \ldots, n$.

***Observaţie*: În formula din enunţ, în locul bazei 2 pentru logaritm poate fi folosită orice bază supraunitară.**

***Indicaţii*:**
1. Dacă în inegalitatea dată se trece termenul din partea stângă în partea dreaptă, obţinem $0 \leq \sum_{i=1}^{n} p_i \log_2 p_i - \sum_{i=1}^{n} p_i \log_2 q_i \Leftrightarrow 0 \leq -\sum_{i=1}^{n} p_i \log_2 \frac{q_i}{p_i}$. Puteţi face legătura dintre expresia din partea dreaptă a acestei ultime inegalităţi şi definiţia *entropiei relative* (numită de asemenea *divergenţa Kullback-Leibler*; vedeţi problema 15) şi apoi să folosiţi proprietăţile entropiei relative.
2. Pentru a demonstra într-o manieră mai directă inegalitatea lui Gibbs, puteţi folosi inegalitatea lui Jensen (vedeţi problema 1).

18. (O „regulă de înlănţuire“ pentru entropia relativă / divergenţa KL)

Vă readucem aminte — vedeţi problema 15 — că entropia relativă sau divergenţa Kullback-Leibler (KL) dintre două distribuţii discrete $P(X)$ şi $Q(X)$ este definită astfel:[20]

$$KL(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}.$$

Pentru convenienţă, vom presupune că $P(x) > 0$ şi $Q(x) > 0$ pentru orice x. Uneori în cele ce urmează vom scrie $KL(P||Q)$ sub forma $KL(P(X)||Q(X))$.

Entropia relativă (sau, divergenţa KL) dintre două distribuţii condiţionale $P(X|Y)$ şi $Q(X|Y)$ se defineşte în felul următor:

$$KL(P(X|Y)||Q(X|Y)) = \sum_y P(y) \Big(\sum_x P(x|y) \log \frac{P(x|y)}{Q(x|y)} \Big).$$

Aceasta poate fi văzută ca fiind media [probabilistă a] divergenţei KL dintre distribuţiile condiţionale corespunzătoare pentru x (adică, dintre $P(X|Y=y)$ şi $Q(X|Y=y)$), media fiind calculată în raport cu valorile y ale variabilei aleatoare Y.

Demonstraţi că are loc următoarea proprietate (de tip „regulă de înlănţuire“) pentru divergenţa KL:

$$KL(P(X,Y)||Q(X,Y)) = KL(P(X)||Q(X)) + KL(P(Y|X)||Q(Y|X)). \tag{18}$$

19. (Echivalenţa dintre minimizarea entropiei relative (divergenţa KL) şi maximizarea funcţiei de log-verosmilitate)

Presupunem că ni se dă o problemă de estimare a parametrilor unei distribuţii probabiliste şi că dispunem de setul de date de antrenament $\{x_i; i = 1, \ldots, m\}$. Considerăm distribuţia empirică $\hat{P}(x) = \frac{1}{m}$, adică distribuţia uniformă peste setul de date de antrenament. Aşadar, a face eşantionare (engl., sampling) cu distribuţia empirică înseamnă a alege în mod aleatoriu un exemplu din setul de antrenament.

Presupunem că avem o anumită familie de distribuţii P_θ, indexată după parametrul θ. (Dacă doriţi, puteţi considera $P_\theta(x)$ ca notaţie alternativă pentru $P(x;\theta)$.)

Demonstraţi că a găsi valoarea de verosimilitate maximă (engl., maximum likelihood estimate) pentru parametrul θ este echivalent cu a găsi acea distribuţie P_θ pentru

[20] Atunci când P şi Q sunt p.d.f.-uri (funcţii de densitate de probabilitate) pentru variabile aleatoare continue, dacă în formula de mai sus înlocuim simbolul de sumare cu integrala, proprietatea enunţată în acest exerciţiu rămâne valabilă. Totuşi, din motive de simplitate, aici vom lucra doar cu funcţii masă de probabilitate / distribuţii discrete.

care entropia relativă (divergenţa KL) faţă de distribuţia empirică $\hat{P}$ este minimă. Aşadar, vă cerem să demonstraţi relaţia următoare:

$$\arg\min_{\theta} KL(\hat{P}||P_\theta) = \arg\max_{\theta} \sum_{i=1}^{m} \ln P_\theta(x_i).$$

***Comentariu*:** Ne vom referi acum la relaţia dintre proprietatea enunţată în acest exerciţiu pe de o parte şi estimarea parametrilor făcută de algoritmul de clasificare binară Bayes Naiv cu atribute de tip Bernoulli de cealaltă parte (vedeţi capitolul de *Clasificare bayesiană*). În modelul de clasificare Bayes Naiv se presupune că P_θ este de forma următoare: $P_\theta(x, y) = p(y) \prod_{i=1}^{n} p(x_i|y)$. Conform regulei de înlănţuire pentru divergenţa KL (vedeţi relaţia (18) de la ex. 18), avem:

$$KL(\hat{P}||P_\theta) = KL(\hat{P}(y)||p(y)) + \sum_{i=1}^{n} KL(\hat{P}(x_i|y)||p(x_i|y)).$$

Această relaţie arată că problema găsirii maximului verosimilităţii / minimumului divergenţei KL (pentru a estima valorile parametrilor) se descompune în $2n + 1$ probleme de optimizare independente: una pentru distribuţia a priori a clasei, $p(y)$, şi câte una pentru fiecare distribuţie condiţională $p(x_i|y)$, corespunzătoare feacărei trăsături x_i în raport cu fiecare din cele două valori posibile pentru eticheta y. Concret, găsirea estimării de verosimilitate maximă pentru fiecare dintre aceste probleme în mod individual rezultă în maximizarea verosimilităţii distribuţiei comune. (O observaţie similară se poate formula şi în legătură cu estimarea parametrilor reţelelor bayesiene (engl., Bayesian networks.)

20. (Proprietăţi ale entropiei: Adevărat sau Fals?)

Stabiliţi dacă următoarele propoziţii sunt adevărate sau false.

a. Entropia nu este negativă.

b. $H(X, Y) \geq H(X) + H(Y)$ pentru orice două variabile aleatoare X şi Y.

c. Dacă X şi Y sunt variabile aleatoare independente, atunci $H(X, Y) = H(X) + H(Y)$.

21. (Cât de multe date de antrenament necesită algoritmul Bayes Naiv vs. algoritmul Bayes Optimal? [LC: complexitatea la eşantionare])

Unul dintre motivele pentru care folosim clasificatorul Bayes Naiv este faptul că el necesită mult mai puţine date de antrenament (în vederea estimării parametrilor) decât clasificatorul Bayes Optimal.

Acest exerciţiu te va ajuta să înţelegi cât de importantă este această diferenţă dintre cei doi algoritmi.

Presupunem că o *observaţie / instanţă* este o valoare generată în mod aleatoriu de către vectorul de variabile aleatoare $\bar{X} = (X_1, \ldots, X_{d-1}, X_d)$, unde fiecare X_i este o variabilă aleatoare urmând distribuţia probabilistă Bernoulli de parametru $p = 0.5$. Considerăm $X_1, \ldots, X_{d-1}$ variabilele de intrare, iar $X_d = Y$ variabila de ieşire.

Pentru a estima în sensul verosimilităţii maxime (MLE) parametrii clasificatorului Bayes Optimal, avem nevoie să *observăm* / întâlnim fiecare valoare a lui $\bar{X}$ de un număr rezonabil de ori. Similar, pentru a antrena clasificatorul Bayes Naiv avem

nevoie să întâlnim fiecare valoare a fiecărei variabile X_i ($i = \overline{1,d}$) de un număr rezonabil de ori.

Ne întrebăm cât de multe observaţii sunt necesare (a fi generate) pentru ca fiecare valoare a variabilei comune $\bar{X}$ în cazul algoritmului Bayes Optimal, şi respectiv fiecare valoare a variabilelor X_i ($i = \overline{1,d}$) în cazul Bayes Naiv să fie întâlnită cel puţin o dată. (În practică este nevoie de mult mai multe observaţii, dar în acest exerciţiu ne limităm la câte o singură observaţie pentru fiecare valoare în parte.)

***Indicaţie*: La rezolvarea punctelor de mai jos vă sugerăm să folosiţi următoarele două inegalităţi:**

- **pentru orice evenimente $E_1, \ldots, E_n$, avem $P(E_1 \cup \ldots \cup E_n) \leq \sum_{i=1}^{n} P(E_i)$.**[21]
- **$(1 - \frac{1}{k})^k \leq \frac{1}{e}$ pentru orice $k \geq 1$, unde $e \approx 2.71828$ este baza logaritmului natural.**

a. Începem cu algoritmul Bayes Naiv. Fie $i \in \{1, \ldots, d\}$ fixat. Arătaţi că dacă s-au făcut N observaţii (având forma $\bar{x}_j = (x_1^j, \ldots, x_{d-1}^j, x_d^j)$ cu $j = 1, \ldots, N$), atunci probabilitatea să nu fi întâlnit ambele valori ale variabilei X_i este $\frac{1}{2^{N-1}}$. (Observaţi că această fracţie reprezintă un număr foarte mic atunci când N este suficient de mare.)

b. Fie $\varepsilon > 0$ fixat. Folosind prima inegalitate din *indicaţia* de mai sus, arătaţi că dacă au fost făcute câte $N_{NB} = 1 + \log_2 \frac{d}{\varepsilon}$ observaţii de aceeaşi formă ca mai sus, atunci probabilitatea să nu fi întâlnit ambele valori pentru fiecare dintre variabilele X_i ($i = \overline{1,d}$) este mai mică sau egală cu ε.

c. Acum trecem la algoritmul Bayes Optimal. Fie $\bar{x}$ o instanţă (fixată) a variabilei comune $\bar{X}$. Folosind a doua inegalitate din *indicaţia* de mai sus, arătaţi că dacă s-au făcut N observaţii (fiecare observaţie implicând simultan toate variabilele X_i cu $i = \overline{1,d}$), atunci probabilitatea ca să nu se fi întâlnit niciodată $\bar{x}$ este mai mică sau egală cu $e^{-\frac{N}{2^d}}$.

d. Arătaţi că dacă au fost făcute cel puţin $N_{JB} = 2^d \ln \frac{2^d}{\varepsilon}$ observaţii, atunci probabilitatea ca să nu se fi întâlnit toate instanţele variabilei comune $\bar{X}$ este mai mică sau egală cu ε.

e. Dacă se fixează $\varepsilon = 0.1$, calculaţi valorile N_{NB} şi N_{JB} pentru $d = 2$, $d = 5$ şi $d = 10$. Ce concluzie puteţi trage? (Altfel spus, cum interpretaţi rezultatele?)

***Observaţii*:**

1. Ştim că pentru clasificatorul Bayes Naiv, este necesar să estimăm din date probabilităţile $P(Y = y)$ — adică, $P(X_d = x_d)$ — şi $P(X_i = x_i | Y = y)$, pentru $i = 1, \ldots, d-1$. Putem considera că este suficient să întâlnim cu o probabilitate de cel puţin $1 - \varepsilon$ toate valorile y, precum şi perechile de forma (x_i, y), cu $i = 1, \ldots, d-1$. În mod implicit, problema noastră simplifică şi mai mult cerinţele, considerând că este suficient să găsim cu probabilitate de cel puţin $1-\varepsilon$ toate valorile x_i, pentru $i = 1, \ldots, d$ (subînţelegând că atunci vor apărea în date, cu probabilităţi semnificative, atât valorile y cât şi fiecare dintre combinaţiile (x_i, y), cu $i = 1, \ldots, d-1$).
2. Pentru clasificatorul Bayes Optimal, în mod similar, este necesar să estimăm probabilităţile $P(Y = y)$ şi $P(X_1 = x_1, \ldots, X_{d-1} = x_{d-1} | Y = y)$, ceea ce, evident, va permite calcularea probabilităţilor de forma $P(x_1, \ldots, x_{d-1}, y)$. Problema noastră

[21] Aceasta se numeşte *proprietatea de subaditivitate* a probabilităţilor.

consideră în mod implicit că este suficient să găsim cu probabilitate de cel puţin $1-\varepsilon$ toate combinaţiile $(x_1, \ldots, x_{d-1}, y)$.

22. **(Algoritmul Bayes Naiv: raportul cu regresia logistică şi natura separatorului decizional; cazul când variabilele de intrare sunt de tip boolean)**

a. [Bayes Naiv şi Regresia Logistică: relaţia dintre regulile de decizie]
Fie Y o variabilă aleatoare Bernoulli, iar $X = (X_1, \ldots, X_d)$ un vector de variabile booleene. Demonstraţi că distribuţia condiţională $P(Y|X)$ are forma funcţiei logistice de argument $z = -(w_0 + w_1X_1 \ldots + w_dX_d)$, cu parametrii $w_0, w_1, \ldots, w_d \in \mathbb{R}$,[22] adică

$$P(Y = 1|X) = \frac{1}{1 + \exp(w_0 + \sum_{i=1}^{d} w_iX_i)}$$

şi, prin urmare

$$P(Y = 0|X) = \frac{\exp(w_0 + \sum_{i=1}^{d} w_iX_i)}{1 + \exp(w_0 + \sum_{i=1}^{d} w_iX_i)}.$$

Vă reamintim că *funcţia logistică* (sau *sigmoidală*) este definită prin expresia $\sigma(z) = 1/(1 + e^{-z})$ pentru orice $z \in \mathbb{R}$.

***Comentariu*:**
Regresia logistică (şi, mai general, *clasificatorii „discriminativi“*) învaţă [în mod] direct parametrii distribuţiei $P(Y|X)$,[23] pe când algoritmul Bayes Naiv (şi, mai general, *clasificatorii „generativi“*) învaţă [parametrii pentru] distribuţiile $P(X|Y)$ şi $P(Y)$, cu ajutorul cărora va calcula apoi $P(Y|X)$ şi cea mai probabilă valoare pentru Y (atunci când X are o valoare fixată / dată). Vom spune că regresia logistică este corespondentul „discriminativ“ al clasificatorului „generativ“ Bayes Naiv.

***Indicaţii*:**
1. Vom introduce o *notaţie* simplă, care ne va fi de folos în continuare. Întrucât variabilele X_i sunt sunt booleene, odată fixată o valoare y_k pentru variabila Y, vom avea nevoie de un singur parametru pentru a defini distribuţia condiţională $P(X_i|Y = y_k)$, pentru fiecare $i = 1, \ldots, d$. Aşadar, vom desemna cu θ_{i1} probabilitatea $P(X_i = 1|Y = 1)$ şi, prin urmare $P(X_i = 0|Y = 1) = 1 - \theta_{i1}$. În mod similar, vom desemna cu θ_{i0} probabilitatea $P(X_i = 1|Y = 0)$.
2. Remarcaţi că odată ce am introdus notaţiile de mai sus, vom putea scrie $P(X_i|Y = 1)$ după cum urmează:

$$P(X_i|Y = 1) = \theta_{i1}^{X_i}(1 - \theta_{i1})^{(1-X_i)}, \tag{19}$$

bineînţeles, cu excepţia cazurilor când $\theta_{i1} = 0$ şi $X_i = 0$, respectiv $\theta_{i1} = 1$ şi $X_i = 1$. Observaţi că atunci când X_i are valoarea 1, cel de-al doilea factor din partea dreaptă a egalităţii (19) este 1, pentru că exponentul lui este zero. Deci $P(X_i|Y = 1) = \theta_{i1}^{X_i} = \theta_{i1}$ pentru $X_i = 1$. În mod similar, atunci când $X_i = 0$ primul factor este egal cu 1, pentru că exponentul lui este zero. Deci $P(X_i|Y = 1) = (1 - \theta_{i1})^{1-X_i} = 1 - \theta_{i1}$ pentru $X_i = 0$.

[22] LC: Prin urmare, *separatorul decizional* (sau, *graniţa de decizie*) pentru algoritmul Bayes Naiv este — într-o astfel de situaţie — liniar (în funcţie de argumentele $X_1, \ldots, X_d$). Ecuaţia separatorului decizional va fi $w_0 + w_1X_1 \ldots + w_dX_d = 0$.

[23] LC: Parametrii distribuţiei $P(Y|X)$ sunt în acest caz $w_i \in \mathbb{R}$, cu $i = 0, 1, \ldots, d$, iar învăţarea lor se face prin maximizarea funcţiei de verosimilitate $\mathcal{L}(w) \overset{not.}{=} P(D|w)$, unde D este setul de date de antrenament. La rândul ei, maximizarea aceasta se realizează prin aplicarea unei metode de optimizare, de exemplu metoda gradientului ascendent sau metoda lui Newton.

b. [Relaxarea presupoziţiei de independenţă condiţională]

Pentru a putea exprima interacţiunile dintre trăsături, modelul regresiei logistice poate fi extins cu nişte termeni suplimentari. De exemplu, putem adăuga un termen care să exprime dependenţa dintre trăsăturile X_1 şi X_2:

$$P(Y = 1|X) = \frac{1}{1 + \exp(w_0 + w_{1,2}X_1X_2 + \sum_{i=1}^{d} w_iX_i)}.$$

În mod similar, presupoziţia de independenţă condiţională asumată de către algoritmul Bayes Naiv poate fi relaxată astfel încât trăsăturile X_1 şi X_2 să nu mai trebuiască să satisfacă independenţa condiţională. Aşadar, vom putea scrie:

$$P(Y|X) = \frac{P(Y)\,P(X_1, X_2|Y)\,\prod_{i=3}^{d} P(X_i|Y)}{P(X)}.$$

Demonstraţi că în acest caz distribuţia $P(Y|X)$ are aceeaşi formă ca şi modelul de regresie logistică augmentat cu un termen suplimentar, care exprimă dependenţa dintre X_1 şi X_2 (şi, în acest fel, modelul extins al regresiei logistice rămâne corespondentul discriminativ al clasificatorului nostru generativ).

***Indicaţii*:**
3. De data aceasta o altă notaţie simplă ne va ajuta. Vom avea nevoie de mai mulţi parametri decât la punctul a pentru a defini distribuţia comună $P(X_1, X_2|Y)$. Aşa că vom nota $\beta_{ijk} = P(X_1 = i, X_2 = j|Y = k)$, pentru fiecare combinaţie posibilă de valori pentru indicii i, j şi k.
4. Această nouă notaţie poate fi folosită acum pentru a exprima probabilitatea $P(X_1, X_2|Y = k)$ după cum urmează:

$$P(X_1, X_2|Y = k) = (\beta_{11k})^{X_1X_2}\,(\beta_{10k})^{X_1(1-X_2)}\,(\beta_{01k})^{(1-X_1)X_2}\,(\beta_{00k})^{(1-X_1)(1-X_2)} \tag{20}$$

pentru $k \in \{0, 1\}$, cu excepţia următoarelor cazuri: *i.* $\beta_{11k} = 0$ şi $X_1X_2 = 0$, *ii.* $\beta_{10k} = 0$ şi $X_1(1 - X_2) = 0$, *iii.* $\beta_{01k} = 0$ şi $(1 - X_1)X_2 = 0$ şi *iv.* $\beta_{00k} = 0$ şi $(1 - X_1)(1 - X_2) = 0$.

23. (Asupra folosirii algoritmului k-NN în spaţii ($\mathbb{R}^p$) de dimensiune (p) mare: un avertisment: „blestemul marilor dimensiuni“)

Considerăm punctele $x_1, x_2, \ldots, x_n$ distribuite în mod independent şi uniform într-o sferă (notată cu B) care are raza egală cu unitatea[24] şi centrul în O, originea spaţiului $\mathbb{R}^p$. Aşadar, $B = \{x : ||x||^2 \leq 1\} \subset \mathbb{R}^p$, unde $||x|| = \sqrt{x \cdot x}$, iar operatorul $\cdot$ desemnează produsul scalar din $\mathbb{R}^p$.

În această problemă veţi studia „mărimea“ vecinătăţii de tip 1-NN pentru originea O şi cum anume variază ea în raport cu dimensiunea p. În acest fel, veţi putea vedea care sunt dezavantajele folosirii algoritmului k-NN într-un spaţiu de dimensiune mare.

Din punct de vedere formal, „mărimea“ menţionată mai sus va fi identificată cu d^*, distanţa de la O la cel mai apropiat vecin din mulţimea $\{x_1, x_2, \ldots, x_n\}$:

$$d^* \overset{not.}{=} \min_{1 \leq i \leq n} ||\,x_i\,|| \,.$$

***Observaţie*: Din moment ce eşantionul $\{x_1, x_2, \ldots, x_n\}$ este generat în mod aleatoriu, distanţa d^* poate fi văzută ca fiind [produsă de către] o variabilă aleatoare.**

[24] Termenul folosit în limba engleză pentru o astel de sferă este *unit ball*.

a. În cazul particular $p = 1$, calculaţi expresia *funcţiei de distribuţie cumulativă*[25] a lui d^* (văzută ca variabilă aleatoare), şi anume $P(d^* \leq t)$ pentru $t \in [0, 1]$.

b. Determinaţi expresia *funcţiei de distribuţie cumulativă* (c.d.f.) a lui d^* în cazul general, adică pentru $p \in \{1, 2, 3, \ldots\}$.

***Sugestie*: Puteţi folosi următoarea formulă pentru volumul unei sfere de rază r din $\mathbb{R}^p$:**

$$V_p(r) = \frac{(r\sqrt{\pi})^p}{\Gamma\left(\frac{p}{2} + 1\right)},$$

unde Γ reprezintă funcţia Gamma a lui Euler, care are proprietăţile:

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi},\ \Gamma(1) = 1, \text{ iar } \Gamma(x+1) = x\Gamma(x) \text{ pentru } x > 0.^{26}$$

c. Care este *mediana* variabilei aleatoare d^* (adică, valoarea lui t pentru care $P(d^* \leq t) = 1/2$)? Va trebui ca răspunsul să fie formulat în funcţie de n şi p (dimensiunea eşantionului şi, respectiv, dimensiunea spaţiului din care se face extragerea instanţelor, $\mathbb{R}^p$).
Pentru $n = 100$, alcătuiţi un grafic cu valorile [funcţiei] mediane pentru $p = 1, 2, 3, \ldots, 100$. Valorile lui p vor fi plasate pe axa Ox, iar valorile medianei pe axa Oy. Ce observaţi?

d. Folosind funcţia de distribuţie cumulativă (c.d.f.) de la punctul b, determinaţi cât de mare ar trebui să fie n (mărimea eşantionului) astfel încât

$$P(d^* \leq 0.5) \geq 0.9,$$

adică, cu probabilitate de cel puţin $9/10$, distanţa d^* de la originea O la cel mai apropiat vecin să fie mai mică decât $1/2$ (adică, jumătate din distanţa de la O la marginea sferei). Va trebui să formulaţi răspunsul ca expresie a unei funcţii în raport cu variabila p.
Reprezentaţi grafic valorile acestei funcţii pentru $p = 1, 2, \ldots, 20$, plasând valorile lui p pe axa Ox şi valorile funcţiei pe axa Oy. Ce observaţi?

***Sugestie*: Pentru $\ln(1 - x)$, puteţi face apel la dezvoltarea sa sub formă de *serie Taylor*:**

$$\ln(1 - x) = -\sum_{i=1}^{\infty} \frac{x^i}{i} \text{ pentru } -1 \leq x < 1.$$

e. În urma rezolvării punctelor de mai sus, ce puteţi spune despre dezavantajele algoritmului k-NN în raport cu [diferitele valori posibile pentru] n şi p?

[25] Engl., cummulative distribution function, c.d.f.

[26] Se verifică uşor că pentru $p = 3$ se obţine volumul sferei: $V_3(r) = \frac{(r\sqrt{\pi})^3}{\frac{3}{4}\sqrt{\pi}} = \frac{4\pi r^3}{3}$.