Învăţare automată

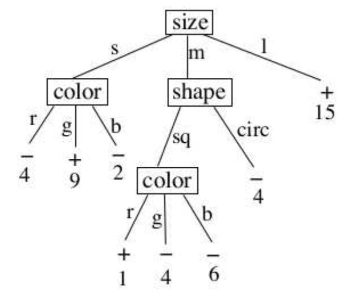
— Licenţă, anul III, 2021-2022, examenul parţial I —

Nume student: **Grupa:**

1. **(Algoritmul ID3: 2 exemple simple de aplicare (antrenare + testare))**

a. Considerăm arborele de decizie alăturat. Numerele situate lângă etichetele +/− reprezintă cât de multe exemple de antrenament au fost asignate la nodul de decizie respectiv în cursul antrenării.

Folosind acest arbore de decizie, clasificaţi următoarele instanţe de test:



Size	Color	Shape	Smell	Classification
small	green	square	pine	
large	blue	square	pine	
medium	green	circle	rotten egg	
medium	red	square	lemon	

***Observaţie*: Pentru convenienţă, pe arborele din imaginea de mai sus, valorile atributelor au fost scrise în mod prescurtat.**

---

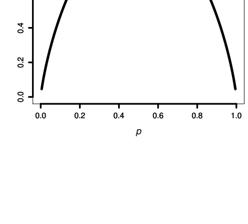
b. Acum vom presupune că aplicaţi algoritmul ID3 pe un set de date de antrenament (neprecizat) şi că folosiţi criteriul câştigului de informaţie maxim (sau, un criteriu *echivalent*(!) cu acesta). Presupunem că

- **setul de date de antrenament este constituit din 128 de exemple, dintre care 64 sunt etichetate pozitiv, iar 64 sunt etichetate negativ;**
- **instanţele de antrenament sunt caracterizate de trei atribute booleene:** *Home-Owner*, *In-Debt* **şi** *Rich*;
- **pentru 64 de exemple de antrenament, atributul** *Home-Owner* **este** *true*, **iar dintre aceste exemple 1/4 sunt negative şi restul sunt pozitive;**
- **pentru 96 de exemple de antrenament, atributul** *In-Debt* **este** *true*, **iar dintre aceste exemple 1/2 sunt pozitive şi restul sunt negative;**
- **pentru 32 de exemple de antrenament, atributul** *Rich* **este** *true*, **iar dintre aceste exemple 3/4 sunt pozitive şi restul sunt negative.**

i. **Ce atribut trebuie pus în nodul rădăcină al arborelui ID3?**
(În prealabil, veţi desena compaşii de decizie corespunzători celor trei atribute de intrare care au fost indicate mai sus.)

***Observaţie*: Atunci când veţi calcula entropii şi / sau câştiguri de informaţie, le veţi exprima în biţi.**

ii. **Puteţi determina atributul care trebuie pus în nodul rădăcină al arborelui ID3 fără să faceţi calcule efectiv [pentru entropii], ci doar bazându-vă pe monotonia funcţiei entropie ($H(p)$) pentru distribuţia aleatoare Bernoulli de parametru p (vedeţi graficul alăturat)?**



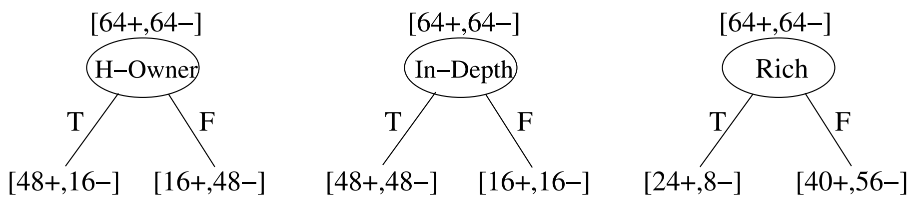
Răspuns:

a.

Size	Color	Shape	Smell	Classification
small	green	square	pine	+
large	blue	square	pine	+
medium	green	circle	rotten egg	−
medium	red	square	lemon	+

b.

i.



Pentru uşurinţa calculelor, vom folosi criteriul entropiei condiţionale medii minime (care este echivalent cu criteriul specificat în enunţ). Vom nota cu Y variabila de ieşire (clasa), iar cu $H(p)$ entropia unei distribuţii Bernoulli (oarecare) de parametru p.

$H(Y|\textit{Home-Owner}) = 2 \cdot \dfrac{64}{128} \cdot H(16/64) = H(1/4) = \ldots = 0.8112$ **biţi;**

$H(Y|\textit{In-Depth}) = 1$ **bit (este imediat);**

$H(Y|\textit{Rich}) = \dfrac{32}{128} \cdot H(8/32) + \dfrac{96}{128} \cdot H(40/96) = \dfrac{1}{4} \cdot H(1/4) + \dfrac{3}{4} \cdot H(5/12) = \ldots = 0.9377$ **biţi.**

Se observă că *Home-Owner* **este atributul cel mai bun, deci va fi pus în nodul rădăcină.**

ii. **Răspunsul este pozitiv.**

De fapt, în raţionamentul de mai sus nu mai era nevoie să facem calculele (acolo unde am pus ...). Justificarea este simplă: a determina cel mai bun atribut revine la a vedea care dintre atributele *Home-Owner* **şi** *Rich* **este mai bun (fiindcă** *In-Depth* **are entropie condiţională medie maximă (1), deci câştig de informaţie minim (0)). Aşadar, trebuie să comparăm**

$$H(1/4) \quad \textbf{şi} \quad \frac{1}{4} \cdot H(1/4) + \frac{3}{4} \cdot H(5/12),$$

ceea ce este echivalent cu a compara $\frac{3}{4} \cdot H(1/4)$ **cu** $\frac{3}{4} \cdot H(5/12)$, **adică** $H(1/4)$ **cu** $H(5/12)$.

Întrucât $\frac{1}{4} < \frac{5}{12} < \frac{1}{2}$, **rezultă — datorită monotoniei lui** $H(p)$, **entropia distribuţiei Bernoulli de parametru** p **— că**

$$H(1/4) < H(5/12).$$

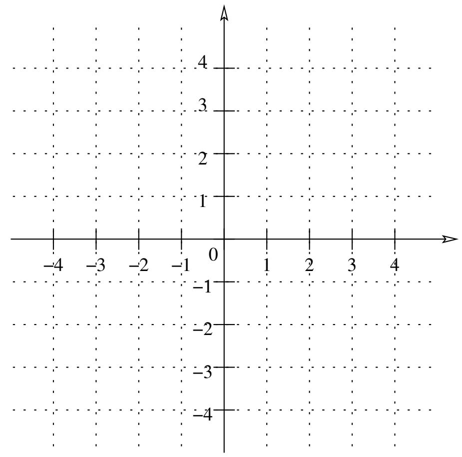
Prin urmare, $H(Y|\textit{Home-Owner}) < H(Y|\textit{Rich})$, **deci în concluzie atributul** *Home-Owner* **trebuie pus în nodul rădăcină.**

2. **(Algoritmul k-NN, cu aplicare pe date din $\mathbb{R}^2$, respectiv $\mathbb{R}$: desenarea de diagrame Voronoi / identificarea zonelor de decizie; calcularea erorii la CVLOO)**

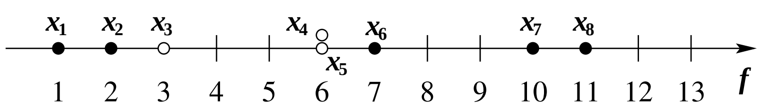
A. Se dă un set de date de antrenament din planul euclidian ($\mathbb{R}^2$), format din instanţele negative $(-1, 0)$, $(2, 1)$ şi $(2, -2)$ şi instanţele pozitive $(0, 0)$ şi $(1, 0)$.

a. Pe reperul de axe de coordonate ortogonale alăturat marcaţi aceste exemple de antrenament. Veţi folosi *convenţia* noastră de notare: simbolul $\bullet$ desemnează instanţe pozitive $(+)$, iar simbolul $\circ$ instanţe negative $(-)$. Desenaţi apoi graniţele de decizie determinate de algoritmul 1-NN. Veţi haşura zona / zonele de decizie corespunzătoare instanţelor pozitive.

b. Cum va clasifica instanţa de test $(1, -1.01)$ de către algoritmul 1-NN? Dar de către algoritmul 3-NN? Justificaţi, în ambele cazuri.

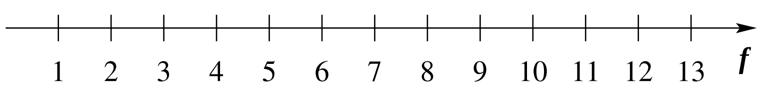


B. În desenul de mai jos este reprezentat un set de antrenament care conţine 8 instanţe, fiecare dintre ele având doar o trăsătură (engl., feature), notată cu f.



Remarcaţi faptul că sunt două instanţe pentru care valoarea trăsăturii f este aceeaşi, şi anume 6. Aceste două instanţe sunt reprezentate prin două simboluri $\circ$, situate unul deasupra celuilalt, dar de fapt ele ar fi trebuit să fie reprezentate ca două simboluri $\circ$ suprapuse (unul peste celălalt), întrucât aceste instanţe au exact aceeaşi valoare pentru trăsătura f.

c. La acest punct al problemei veţi folosi algoritmul 1-NN. Vă cerem ca pe linia de mai jos să haşuraţi — şi, bineînţeles, să delimitaţi prin separatori decizionali — zonele în care algoritmul 1-NN va prezice semnul $+$, dat fiind setul de date de antrenament din figura de mai sus.

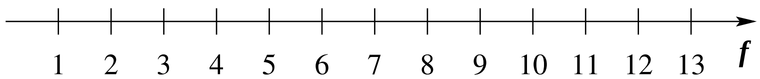


Cu ce etichetă vor fi clasificate instanţele de test $f = 2.5$ şi $f = 6.5$? Justificaţi.

***Convenţii / reguli*:**
1. În cazul în care există două sau mai multe instanţe situate exact pe „marginea“ [adică, pe conturul circular al] k-NN-vecinătăţii asociate instanţei de clasificat, se va considera că toate aceste instanţe aparţin respectivei vecinătăţi, iar fiecare dintre ele dispune de un vot întreg.
2. În caz de *paritate de voturi*, veţi alege eticheta instanţei de antrenament care este situată la *distanţă minimă* către stânga faţă de instanţa de test respectivă. (Distanţă minimă este 0 atunci când instanţa de test coincide cu o instanţă de antrenament!)
3. Aceste reguli vor fi aplicate şi la punctele următoare.

d. **Dacă faceţi cross-validare folosind metoda "leave-one-out" pe acest set de date, în conjuncţie cu algoritmul 1-NN, cât va fi eroarea produsă? Veţi arăta în mod clar — alcătuind un *tabel* care să conţină 1-NN-vecinătăţile respective — cum anume aţi ajuns la rezultatul pe care l-aţi indicat.**

e. **Similar cu punctul c, însă aici veţi folosi algoritmul 5-NN.**[1]



***Indicaţie*: Justificaţi în mod riguros rezultatul, referindu-vă la diverse intervale de valori (sau valori particulare) ale lui f:**

Cazul 1: $f \in (-\infty, 4)$**:** ...

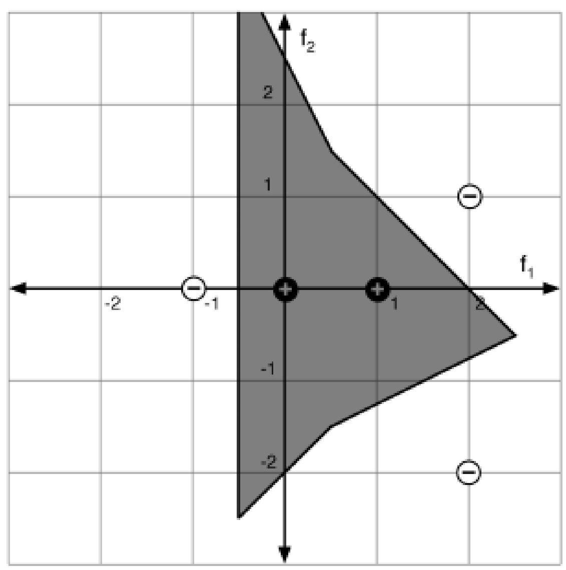
Cazul 2: $f = 4$**:** ...

...

Cu ce etichetă vor fi clasificate instanţele de test $f = 4$, $f = 6$ şi $f = 7$? Justificaţi.

Răspuns:

a.



b.
$(1, -1.01)$**: Positive (+) since this is the class of the closest data point** $(1, 0)$**.**
$(1, -1.01)$**: Positive (+) since it is the majority class of the three closest data points** $(0, 0)$**,** $(1, 0)$ **and** $(2, -2)$**.**

c.

***Justificare*:**

[1]Fără să aplicaţi 5-NN pentru instanţele de test $f = 2.5$ şi $f = 6.5$.

Cazul 1.1: pentru orice $f \in (-\infty, 1.5) \cup (1.5, 2.5) \cup (6.5, 8.5) \cup (8.5, 10.5) \cup (10.5, +\infty)$, există un unic $x_i \in \{x_1, \ldots, x_8\}$ care este cel mai apropiat vecin al lui f, iar x_i este clasificat pozitiv, deci şi f va fi clasificat pozitiv;

Cazul 1.2: pentru $f \in \{1.5, 8.5, 10.5\}$, 1-NN-vecinătatea lui f este respectiv $\{x_1, x_2\}$, $\{x_6, x_7\}$, $\{x_7, x_8\}$, deci, conform *regulii* 1, f va fi clasificat pozitiv (fiindcă x_1, x_2, x_6, x_7 şi x_8 au semnul $+$);

Cazul 2.1: $f = 2.5$ este clasificat pozitiv, fiindcă 1-NN-vecinătatea sa este $\{x_1, x_2\}$ şi se aplică *regula* (2) *pentru paritate de voturi* (x_1 are etichetă pozitivă);

Cazul 2.2: $f = 6.5$ este clasificat negativ, fiindcă 1-NN-vecinătatea sa este $\{x_4, x_5, x_6\}$, iar x_4 şi x_5 au etichete negative (se aplică regula 1);

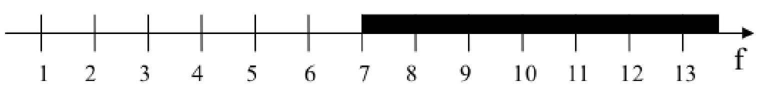
Cazul 3.1: pentru orice $f \in (2.5, 4.5)$, 1-NN-vecinătatea lui f este $\{x_3\}$, deci f este clasificat ca şi x_3, adică negativ.

Cazul 3.2: pentru orice $f \in (4.5, 6.5)$, 1-NN-vecinătatea lui f este $\{x_4, x_5\}$, deci f este clasificat negativ (ca şi x_4 şi x_5), conform regulii 1.

***Concluzie*:** orice instanţă având $f \in (-\infty, 2.5] \cup (6.5, +\infty)$ este clasificată pozitiv de către algoritmul 1-NN; altfel, este clasifictă negativ.

d. The point at 1 would be correct, nearest neighbor is at 2.
The point at 2 would be correct, nearest neighbor is at 1 (tie).
The point at 3 would be incorrect, nearest neighbor is 2.
Both points at 6 would be correct, nearest neighbor is the other point at 6.
The point at 7 would be incorrect, nearest neighbor is 6 (tie).
The point at 10 would be correct, nearest neighbor is 11.
The point at 11 would be correct, nearest neighbor is 10.
So, 6 correct, 2 incorrect ⇒ 75% would be predicted performance.

e.



***Justificare*:**

Cazul 1: $\forall f \in (-\infty, 4)$ va fi clasificat cu $-$, fiindcă 5-NN-vecinătatea oricărui astfel de punct este $\{x_1, \ldots, x_5\}$, în care semnul majoritar este $-$;

Cazul 2: $f = 4$: 5-NN-vecinătatea sa este $\{x_1, \ldots, x_6\}$; se aplică *regula pentru paritate de voturi* şi, deci, $f = 4$ primeşte semnul lui x_3: $-$;

Cazul 3: $\forall f \in (4, 6)$ va fi clasificat cu $-$, fiindcă 5-NN-vecinătatea oricărui astfel de punct este $\{x_2, \ldots, x_6\}$, în care semnul majoritar este $-$;

Cazul 4: $f = 6$: 5-NN-vecinătatea sa este $\{x_2, \ldots, x_7\}$; se aplică *regula pentru paritate de voturi* şi, deci, $f = 6$ primeşte semnul lui x_4 (sau x_5): $-$;

Cazul 5: $\forall f \in (6, 7)$ va fi clasificat cu $-$, fiindcă 5-NN-vecinătatea oricărui astfel de punct este $\{x_3, \ldots, x_7\}$, în care semnul majoritar este $-$;

Cazul 6: $f = 7$: 5-NN-vecinătatea sa este $\{x_3, \ldots, x_8\}$; se aplică *regula pentru paritate de voturi* şi, deci, $f = 7$ primeşte semnul lui x_6: $+$;

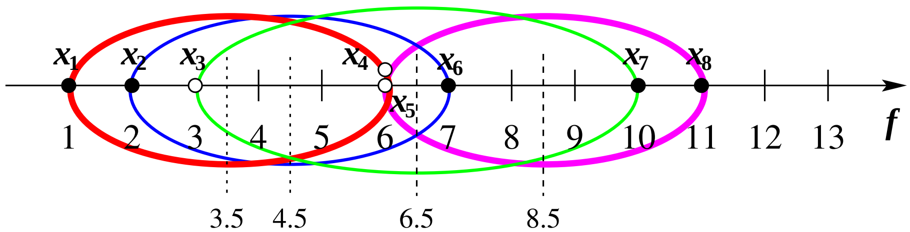
Cazul 7: $\forall f \in (7, +\infty)$ va fi clasificat cu $+$, fiindcă 5-NN-vecinătatea oricărui astfel de punct este $\{x_4, \ldots, x_8\}$, în care semnul majoritar este $+$.

***Concluzie*:** orice instanţă având $f \in [7, +\infty)$ este clasificată pozitiv de către algoritmul 5-NN; altfel, este clasifictă negativ.

***Explicaţie*:** Cum au fost alese / determinate aceste praguri, respectiv intervale?

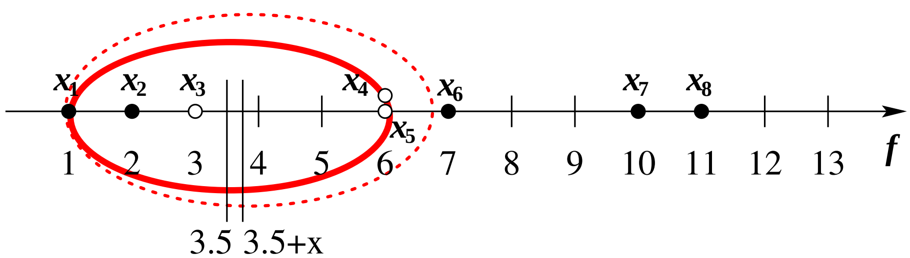
Se poate demonstra faptul că în cazul seturilor de date de antrenament $x_1, x_2, \ldots, x_n \subset \mathbb{R}$ (adică, atunci când se lucrează pe axa reală) k-NN-vecinătăţile nu sunt discontinue, adică: dat fiind un punct de test oarecare $x_q \in \mathbb{R}$, dacă instanţele de antrenament x_i şi x_j, cu $x_i \leq x_j$, se află în k-NN-vecinătatea lui x_q, atunci orice instanţă de antrenament x_l care satisface proprietatea $x_i \leq x_l \leq x_j$ aparţine şi ea respectivei k-NN-vecinătăţi a lui x_q.

În figura următoare punem în evidenţă cele patru (singurele!) 5-NN-vecinătăţi având în componenţă câte 5 exemple de antrenament din setul $\{x_1, \ldots, x_8\}$ dat în enunţul problemei noastre. Cazurile 1, 3, 5 şi 7 de mai sus corespund acestor patru 5-NN-vecinătăţi.



Pragurile $f = 4$, $f = 6$ şi $f = 7$ (cazurile 2, 4 şi 6 de mai sus) corespund acelor 5-NN-vecinătăţi care au în componenţă câte 6 exemple de antrenament din setul $\{x_1, \ldots, x_8\}$ dat în enunţul problemei noastre. Cum au fost determinate aceste praguri?

Justificare intuitivă, pentru $f = 4$:



Justificare analitică:

Cazul 1: **Mijlocul intervalului $[x_1, x_6]$ este 3.5. Este imediat că orice punct situat la stânga lui 3.5 are aceeaşi vecinătate ca şi 3.5 (adică, $\{x_1, \ldots, x_5\}$). Vom identifica acum punctele la dreapta lui 3.5 care au aceeaşi vecinătate, $\{x_1, \ldots, x_5\}$:**
Aceasta revine la a găsi $x > 0$ a.î. $1 + 2(3.5 + x - 1) < 7 = x_6$.
$1 + 2(3.5 + x - 1) < 7 \Leftrightarrow 3.5 + x - 1 < 3 \Leftrightarrow x \in (0, 0.5)$.
Aşadar, intervalul corespunzător acestui caz (adică, acestei 5-NN vecinătăţi!) este $(-\infty, 4)$.

Cazul 3: **Mijlocul intervalului $[x_2, x_6]$ este 4.5. Este imediat că orice punct situat în intervalul $(4, 3.5)$ are aceeaşi vecinătate ca şi 4.5 (adică, $\{x_2, \ldots, x_6\}$). Vom identifica acum punctele la dreapta lui 4.5 care au aceeaşi vecinătate, $\{x_2, \ldots, x_6\}$:**
Aceasta revine la a găsi $y > 0$ a.î. $2 + 2(4.5 + y - 2) < 10 = x_7$.
$2 + 2(4.5 + y - 2) < 10 \Leftrightarrow 4.5 + y - 2 < 4 \Leftrightarrow y \in (0, 1.5)$.
Aşadar, intervalul corespunzător acestui caz (adică, acestei 5-NN vecinătăţi!) este $(4, 6)$.

Cazul 5: **Mijlocul intervalului $[x_3, x_7]$ este 6.5. Este imediat că orice punct situat în intervalul $(6, 6.5)$ are aceeaşi vecinătate ca şi 6.5 (adică, $\{x_3, \ldots, x_7\}$). Vom identifica acum punctele la dreapta lui 6.5 care au aceeaşi vecinătate, $\{x_3, \ldots, x_7\}$:**

Aceasta revine la a găsi $z > 0$ **a.î.** $3 + 2(6.5 + z - 3) < 11 = x_8$**.**
$3 + 2(6.5 + y - 3) < 11 \Leftrightarrow 6.5 + z - 3 < 4 \Leftrightarrow z \in (0, 0.5)$**.**
Aşadar, intervalul corespunzător acestui caz (adică, acestei 5-NN vecinătăţi!) este $(6, 7)$**.**

Cazul 5: Prin complementaritate (sau, în mod direct, printr-un raţionament realizat ca „în oglindă" faţă de cele de mai sus), rezultă că intervalul corespunzător 5-NN vecinătăţii $\{x_4, \ldots, x_8\}$ **este** $(7, +\infty)$**.**

3. **(Un exemplu de clasificator de tip bayesian care combină avantajele algoritmilor Bayes Naiv şi Bayes Optimal)**

Fie A, B şi C variabile aleatoare binare independente, fiecare dintre ele având posibilitatea să ia valoarea 0 cu probabilitate de 50%. Considerăm funcţia

$$Y = ((\neg B) \wedge (\neg C)) \vee (A \wedge B).$$

a. Scrieţi *tabela de adevăr* a funcţiei Y.

b. i. Folosind tabela de adevăr a funcţiei Y — văzută acum ca un set de date de antrenament pentru un clasificator de tip bayesian — *estimaţi* valorile următoarelor probabilităţi condiţionate:[2]

$P(A=0\|B=0,Y=0) = \ldots$	$P(A=0\|C=0,B=0,Y=0) = \ldots$ $P(A=0\|C=1,B=0,Y=0) = \ldots$
$P(A=0\|B=0,Y=1) = \ldots$	$P(A=0\|C=0,B=0,Y=1) = \ldots$ $P(A=0\|C=1,B=0,Y=1) = \ldots$
$P(A=0\|B=1,Y=0) = \ldots$	$P(A=0\|C=0,B=1,Y=0) = \ldots$ $P(A=0\|C=1,B=1,Y=0) = \ldots$
$P(A=0\|B=1,Y=1) = \ldots$	$P(A=0\|C=0,B=1,Y=1) = \ldots$ $P(A=0\|C=1,B=1,Y=1) = \ldots$

Atenţie! **Unele dintre aceste probabilităţi condiţionate s-ar putea să nu fie definite. De *exemplu*, $P(A=0|C=0,B=0,Y=0)$ este nedfinită, fiindcă $P(C=0,B=0,Y=0)$, adică (veţi vedea) nu există în tabelul de date nicio combinaţie de forma $C=0, B=0, Y=0$. În dreptul unor astfel de probabilităţi condiţionate nedfinite veţi pune semnul $*$ *(nedefinit)*.**

ii. Analizând rezultatele estimărilor obţinute mai sus, se poate deduce o relaţie de tip *independenţă condiţională* între variabilele A, B, C şi Y. Care este această relaţie?

c. i. Pe baza rezultatului de la punctul $b.ii$, scrieţi *regula de decizie* pe care o poate folosi un clasificator de tip bayesian — diferit de Bayes Naiv şi Bayes Optimal / Comun; să-i spunem *New-Bayes* — pentru învăţarea conceptului / funcţiei Y, astfel încât
– să producă eroare 0 pe setul de date [de antrenament] de la punctul a, dar
– să necesite un număr de parametri „liberi“ (engl., free) care trebuie estimaţi mai mic decât cel al algoritmului Bayes Optimal.

ii. *Justificaţi* în mod riguros regula de decizie pe care aţi scris-o şi că eroarea la antrenare produsă de la punctul a este într-adevăr 0.

iii. Care sunt aceşti parametri „liberi“?

iv. *Comparaţi* numărul acestor parametri „liberi“ cu numărul minimal de parametri necesari de estimat de către clasificatorii Bayes Naiv şi respectiv Bayes Optimal pe acelaşi set de date.

[2]Estimarea va fi făcută în mod clasic, adică în sensul verosimilităţii maxime (engl., Maximum Likelihood Estimation, MLE). Nu veţi aplica nicio regulă de netezire a acestor probabilităţi, cum ar fi regula "add-one" a lui Laplace.

<u>**Răspuns:**</u>

a.

A	B	C	$A \wedge B$	$B \vee C$	$\neg(B \vee C)$	Y
0	0	0	0	0	1	1
0	0	1	0	1	0	0
0	1	0	0	1	0	0
0	1	1	0	1	0	0
1	0	0	0	0	1	1
1	0	1	0	1	0	0
1	1	0	1	1	0	1
1	1	1	1	1	0	1

b. Probabilităţile estimate sunt:

$P(A=0 \mid B=0, Y=0) = 1/2$	$P(A=0 \mid C=0, B=0, Y=0) = *$*(nedef.)* $P(A=0 \mid C=1, B=0, Y=0) = 1/2$
$P(A=0 \mid B=0, Y=1) = 1/2$	$P(A=0 \mid C=0, B=0, Y=1) = 1/2$ $P(A=0 \mid C=1, B=0, Y=1) = *$*(nedef.)*
$P(A=0 \mid B=1, Y=0) = 1$	$P(A=0 \mid C=0, B=1, Y=0) = 1$ $P(A=0 \mid C=1, B=1, Y=0) = 1$
$P(A=0 \mid B=1, Y=1) = 0$	$P(A=0 \mid C=0, B=1, Y=1) = 0$ $P(A=0 \mid C=1, B=1, Y=1) = 0.$

Se observă că

$$P(A=a \mid B=b, Y=y) = P(A=a \mid C=c, B=b, Y=y) \text{ **pentru orice** } a,b,c,y \in \{0,1\},$$

ori de câte ori probabilităţile condiţionate respective sunt definite.[3]
Prin urmare, variabilele A şi C sunt independente condiţional, în raport cu variabilele B şi Y (luate împreună).

c. [Justificare] Datorită independenţei condţionale demonstrate la punctul a, avem

$$\begin{aligned} P(A,B,C,Y) &= P(A,B,C \mid Y) \cdot P(Y) = P(A,C \mid B,Y) \cdot P(B \mid Y) \cdot P(Y) \\ &= P(A \mid B,Y) \cdot P(C \mid B,Y) \cdot P(B \mid Y) \cdot P(Y). \end{aligned} \tag{1}$$

La scrierea acestor egalităţi, am folosit în ordine: regula de înmulţire / multiplicare a probabilităţilor, varianta condiţională a regulei de multiplicare — sau, direct, pentru primele două egalităţi, regula de înlănţuire a probabilităţilor — şi, în sfârşit, independenţa condiţională dedusă la punctul b.

[3] Se poate verifica uşor că definiţia mai generală a independenţei condiţionale — $P(A=a, C=c \mid B=b, Y=y) = P(A=a \mid B=b, Y=y) \cdot P(C=c \mid B=b, Y=y)$ pentru orice $a,b,c,y \in \{0,1\}$ — este satisfăcută în ambele cazuri de „nedefinire“ indicate mai sus.

În consecinţă,

$$
\begin{aligned}
v_{New\text{-}Bayes}(A=a,B=b,C=c) &\stackrel{def.}{=} v_{JB}(A=a,B=b,C=c)\\
&\stackrel{def.}{=} \arg\max_{y\in\{0,1\}} P(Y=y|A=a,B=b,C=c)\\
&= \arg\max_{y\in\{0,1\}} \frac{P(A=a,B=b,C=c|Y=y)\cdot P(Y=y)}{P(A=a,B=b,C=c)}\\
&= \arg\max_{y\in\{0,1\}} P(A=a,B=b,C=c|Y=y)\cdot P(Y=y)\\
&= \arg\max_{y\in\{0,1\}} P(A=a,B=b,C=c,Y=y)\\
&\stackrel{(1)}{=} \arg\max_{y\in\{0,1\}} P(A=a|B=b,Y=y)\cdot P(C=c|B=b,Y=y)\cdot P(B=b|Y=y)\cdot P(Y=y).
\end{aligned}
$$

De la curs ştim că algoritmul Bayes Optimal produce eroare la antrenare 0 la învăţarea oricărei funcţii boolene. Prin urmare, şi noul clasificator bayesian produce eroare la antrenare 0 la învăţarea funcţiei Y.

Parametrii de estimat sunt:

$P(Y=0)$,
$P(B=0|Y=0)$, $P(B=0|Y=1)$,
$P(A=0|B=0,Y=0)$, $P(A=0|B=0,Y=1)$, $P(A=0|B=1,Y=0)$, $P(A=0|B=1,Y=1)$,
$P(C=0|B=0,Y=0)$, $P(C=0|B=0,Y=1)$, $P(C=0|B=1,Y=0)$, $P(C=0|B=1,Y=1)$.

Sunt, deci, în total 11 parametri de estimat, ceea ce reprezintă mai mult decât numărul parametrilor care sunt necesari pentru Bayes Naiv ($7=2\cdot 3+1$), dar mai puţin decât numărul celor necesari pentru Bayes Optimal ($15=2^{3+1}-1$).