0.1 Fundamente — Probleme rezolvate

0.1.1 Evenimente aleatoare şi formula lui Bayes

1. **(Proprietăţi derivate din definiţia funcţiei de probabilitate)**

• ***CMU, 2009 spring, Tom Mitchell, HW2, pr. 1.1***

Fie două evenimente A şi B.

a. Folosind doar proprietăţile din definiţia funcţiei de probabilitate, arătaţi că $P(A \cap \bar{B}) = P(A) - P(A \cap B)$.

b. Demonstraţi *inegalitatea lui Bonferroni*: $P(A \cap B) \geq P(A) + P(B) - 1$.

c. Spunem că evenimentele A şi B sunt *incompatibile* dacă $P(A \cap B) = 0$.

În ipoteza în care $P(A) = 1/3$ şi $P(B) = 5/6$, este posibil ca evenimentele A şi B să fie incompatibile? Justificaţi răspunsul.

Răspuns:

a. $A = A \cap \Omega = A \cap (B \cup \bar{B}) = (A \cap B) \cup (A \cap \bar{B})$.

Cum evenimentele $(A \cap B)$ şi $(A \cap \bar{B})$, văzute ca mulţimi, sunt disjuncte, conform proprietăţii de „aditivitate numărabilă“ din definiţia funcţiei de probabilitate putem scrie $P(A) = P(A \cap B) + P(A \cap \bar{B})$, deci $P(A \cap \bar{B}) = P(A) - P(A \cap B)$.

b. Întrucât $A \cup B = (A \cap \bar{B}) \cup (A \cap B) \cup (\bar{A} \cap B)$, este imediat că $P(A \cup B) = P(A \cap \bar{B}) + P(A \cap B) + P(\bar{A} \cap B)$. Deci $P(A \cap B) = P(A \cup B) - P(A \cap \bar{B}) - P(\bar{A} \cap B) = [P(A \cup B) - P(A \cap \bar{B})] + [P(A \cup B) - P(\bar{A} \cap B)] - P(A \cup B) = P(A) + P(B) - P(A \cup B)$. Cum $P(A \cup B) \leq 1$ (fiind o probabilitate) putem scrie că: $P(A \cap B) = P(A) + P(B) - P(A \cup B) \geq P(A) + P(B) - 1$, deci $P(A \cap B) \geq P(A) + P(B) - 1$.

c. Pentru a studia posibilitatea ca evenimentele A şi B să fie incompatibile vom aplica inegalitatea lui Bonferroni:

$$P(A) + P(B) = \frac{1}{3} + \frac{5}{6} = \frac{7}{6} \Rightarrow P(A \cap B) \geq \frac{1}{6} > 0$$

Deci $P(A \cap B)$ nu poate fi 0 şi, în consecinţă, evenimentele A şi B nu sunt incompatibile.

2. **(Calcularea de probabilităţi elementare şi probabilităţi condiţionate)**

• ***CMU, 2009 spring, Tom Mitchell, HW2, pr. 1.4***

În acest exerciţiu vom arăta că — într-un anumit sens — probabilitatea unui eveniment se poate schimba (într-un anumit sens) dacă ştim probabilitatea unui alt eveniment, legat de cel dintâi.

Se aruncă simultan două zaruri. Notăm cu S variabila aleatoare care desemnează suma numerelor rezultate din aruncarea celor două zaruri.

a. Calculaţi $P(S = 11)$.

b. Dacă ştim că S este număr prim, cât devine probabilitatea de mai sus?

Răspuns:

a. Probabilitatea unui eveniment precum cel din enunţ este dată de raportul dintre numărul de cazuri favorabile şi numărul cazurilor posibile.

La aruncarea simultană a două zaruri, fiecare zar poate cădea pe una dintre cele 6 feţe, independent de celălalt zar. Prin urmare, există $6 \cdot 6 = 36$ posibilităţi (cazuri posibile). Pentru a se obţine $S = 11$ cele două zaruri trebuie să aibă fie valorile $(5, 6)$, fie $(6, 5)$, deci există 2 posibilităţi (cazuri favorabile). Prin urmare,

$$P(S = 11) = \frac{2}{36} = \frac{1}{18}$$

b. Probabilitatea căutată este:

$$P(S = 11 \mid S = prim) = \frac{P(S = 11 \cap S = prim)}{P(S = prim)} = \frac{P(S = 11)}{P(S = prim)}$$

Trebuie calculată probabilitatea ca S să fie număr prim. Pentru aceasta este necesar numărul de cazuri favorabile, adică numărul cazurilor pentru care $S \in \{2, 3, 5, 7, 11\}$. Există următoarele 15 posibilităţi:

- $S = 2 : (1, 1)$
- $S = 3 : (1, 2), (2, 1)$
- $S = 5 : (1, 4), (4, 1), (2, 3), (3, 2)$
- $S = 7 : (1, 6), (6, 1), (2, 5), (5, 2), (3, 4), (4, 3)$
- $S = 11 : (5, 6), (6, 5)$

Deci $P(S = prim) = 15/36$. Prin urmare,

$$P(S = 11 \mid S = prim) = \frac{2/36}{15/36} = \frac{2}{15}$$

Întrucât $2/15 > 1/18$, putem spune că probabilitatea evenimentului $S = 11$ a crescut după ce am aflat un fapt suplimentar, şi anume că suma rezultată la aruncarea celor două zaruri este un număr prim ($S = prim$).

Ca o *observaţie* cu caracter general: dacă $A \subseteq B$ (aşa cum a fost cazul în această problemă), atunci rezultă imediat că $P(A) \leq P(A|B)$.

3. **(Spaţiu de eşantionare – exemplificare; calcul de probabilităţi condiţionate)**

• ***CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW1, pr. 1.4***

O cutie conţine trei carduri. Un card este roşu pe ambele părţi, un alt card este verde pe ambele părţi iar cel care a rămas este roşu pe o parte şi verde

pe partea cealaltă. Selectăm în mod aleatoriu un card din această cutie; presupunem că nu-i vedem decât culoarea de pe faţa superioară. Dacă această faţă este verde, care este probabilitatea ca şi cealaltă faţă să fie verde?

Răspuns:

Pentru a rezolva această problemă este foarte util să stabilim mai întâi care sunt elementele ce compun *spaţiul de eşantionare* (engl., sample space), Ω.[14] Contrar intuiţiei primare, Ω nu va fi constituit din cele trei carduri, ci din feţele lor, fiindcă ceea ce observăm după o extragere este doar o faţă a unui card, nu ambele feţe ale cardului extras.

Din punct de vedere formal, vom folosi următoarea *notaţie* pentru carduri:

$$C_1 = (R1, R2),\ C_2 = (R3, V4),\ C_3 = (V5, V6),$$

unde $R1, R2, R3, V4, V5$ şi $V6$ desemnează cele 6 feţe ale cardurilor. Aşadar, $\Omega = \{R1, R2, R3, V4, V5, V6\}$.

***Observaţie*: Am fi putut nota feţele cardurilor folosind pur şi simplu numerele $1, \ldots, 6$, însă am preferat să însoţim fiecare dintre aceste numere cu litera R sau V care desemnează culoarea feţei respective.**

După ce am făcut această pregătire, probabilitatea cerută în enunţul problemei se calculează simplu, folosind regula clasică: $p = m/n$, unde m este numărul de cazuri favorabile, iar n este numărul de cazuri posibile.

Cazurile posibile sunt $V4, V5, V6$, iar cazurile favorabile sunt $V5$ şi $V6$ deoarece doar pentru ele faţa cealaltă a cardului este verde (este vorba de $V6$ şi respectiv $V5$). Aşadar, probabilitatea cerută este $\frac{2}{3}$.

4. **(Evenimente aleatoare independente)**

- ***CMU, 2009 spring, Tom Mitchell, HW2, pr. 1.2.1***
  CMU, 2009 spring, Ziv Bar-Joseph, HW1, pr. 1.1

Două evenimente A şi B sunt independente statistic dacă $P(A \cap B) = P(A) \cdot P(B)$.

a. Arătaţi că dacă A şi B sunt evenimente independente, atunci:

– A şi $\bar{B}$ sunt independente;

– $\bar{A}$ şi $\bar{B}$ sunt independente.

b. Dacă evenimentul A este independent în raport cu el însuşi, ce puteţi spune despre $P(A)$?

Răspuns:

a. $P(A \cap \bar{B}) = P(A) - P(A \cap B) = P(A) - P(A) \cdot P(B) = P(A) \cdot (1 - P(B)) = P(A) \cdot P(\bar{B})$, **deci A şi $\bar{B}$ sunt independente. (S-a folosit independenţa evenimentelor A şi B.)**

Pentru independenţa evenimentelor $\bar{A}$ şi $\bar{B}$ se procedează analog:
$P(\bar{A} \cap \bar{B}) = P(\bar{B}) - P(A \cap \bar{B}) = P(\bar{B}) - P(A) \cdot P(\bar{B}) = P(\bar{B}) \cdot (1 - P(A)) = P(\bar{A}) \cdot P(\bar{B})$, **deci**

[14] A se vedea noţiunea prezentată la curs.

$\bar{A}$ şi $\bar{B}$ sunt independente. (La cea de-a doua egalitate s-a folosit independenţa evenimentelor A şi $\bar{B}$ demonstrată mai sus.)

b. Condiţia de independenţă a evenimentului A faţă de el însuşi se scrie astfel:

$$P(A \cap A) = P(A) \cdot P(A) \Rightarrow P(A) = [P(A)]^2 \Rightarrow P(A)[P(A) - 1] = 0$$

Ţinând cont şi de restricţia $P(A) \in [0, 1]$, rezultă că $P(A) = 0$ sau $P(A) = 1$. (Atenţie: nu rezultă neapărat că $A = \emptyset$ respectiv $A = \Omega$!)

5. **(Evenimente aleatoare independente; aplicarea proprietăţilor din definiţia funcţiei de probabilitate)**

• ***CMU, 2009 spring, Tom Mitchell, HW2, pr. 1.2.3***

Robert şi Alina dau cu banul alternativ. Cel dintâi dintre ei care va obţine stema (în engleză: head) câştigă jocul. Alina este prima care va da cu banul.

a. Dacă $P(stemă) = 1/2$, care este probabilitatea ca Alina să câştige jocul?

***Indicaţie* (1): Încercaţi să enumeraţi toate situaţiile în care Alina poate câştiga.**

***Indicaţie* (2): Pentru orice $a \in [0, 1)$, avem $\sum_{i=0}^{i=\infty} a^i = 1 + a + a^2 + \ldots + a^n + \ldots = \lim_{n \to +\infty} \frac{1 - a^n}{1 - a} = \frac{1}{1 - a}$.**

b. Dacă $P(stemă) = p \in (0, 1]$, care este probabilitatea ca Alina să câştige jocul?

c. Ţinând cont de expresia obţinută la punctul b, dacă ar fi ca tu să joci acest joc, cum ai decide să intri în joc: primul ori al doilea (presupunând, bineînţeles, că ai avea posibilitatea să alegi)?

Răspuns:

a. Alina aruncă moneda în „paşii" impari. Ea câştigă jocul dacă la pasul $2n+1$ obţine stema *şi* până la pasul respectiv nimeni nu a obţinut stema.

Notăm cu A evenimentul ca Alina să câştige jocul şi cu A_i evenimentul ca Alina să câştige jocul la a $i-$a aruncare a banului. Întrucât evenimentele A_i, văzute ca mulţimi, sunt mutual disjuncte ($A_i \cap A_j = \emptyset$ pentru orice $i \neq j$) şi $A = A_1 \cup A_3 \cup \ldots$, conform proprietăţii de aditivitate numărabilă din definiţia funcţiei de probabilitate rezultă că

$$P(A) = P(A_1) + P(A_3) + P(A_5) + \ldots$$

Întrucât $p = \frac{1}{2}$, ţinând cont [şi] de faptul că toate aruncările sunt independente, probabilităţile corespunzătoare evenimentelor A_i se calculează astfel:

$$P(A_1) = P(stemă) = \frac{1}{2}$$

$$P(A_3) = (1 - P(stemă)) \cdot (1 - P(stemă)) \cdot P(stemă) = \left(\frac{1}{2}\right)^3$$

$$P(A_5) = (1 - \frac{1}{2}) \cdot (1 - \frac{1}{2}) \cdot (1 - \frac{1}{2}) \cdot (1 - \frac{1}{2}) \cdot P(stemă) = \left(\frac{1}{2}\right)^5$$

$$\ldots$$

$$P(A_{2i+1}) = (1 - P(stemă))^{2i} \cdot P(stemă) = \left(\frac{1}{2}\right)^{2i+1}$$

Prin urmare,

$$P(A) = \sum_{i=0}^{\infty} P(A_{2i+1}) = \sum_{i=0}^{\infty} \left(\frac{1}{2}\right)^{2i+1} = \sum_{i=0}^{\infty} \left(\frac{1}{2}\right)^{2i} \cdot \frac{1}{2} = \frac{1}{2} \cdot \sum_{i=0}^{\infty} \left(\frac{1}{4}\right)^{i} = \frac{1}{2} \cdot \frac{1}{1-1/4}$$
$$= \frac{1}{2} \cdot \frac{4}{3} = \frac{2}{3}.$$

Aşadar, pentru $p = 1/2$ probabilitatea ca Alina să câştige jocul este $2/3$ (deci de două ori mai mare decât probabilitatea ca Robert să câştige jocul), pur şi simplu datorită faptului că ea este prima care dă cu banul. (Avantajul primului jucător!)

b. Dacă $P(stemă) = p \in (0, 1]$, se urmează acelaşi raţionament ca mai sus, cu modificarea valorilor $P(A_i)$ astfel:

$$P(A_1) = P(stemă) = p$$
$$P(A_3) = (1 - P(stemă)) \cdot (1 - P(stemă)) \cdot P(stemă) = (1-p)^2 \cdot p$$
$$P(A_5) = (1-p) \cdot (1-p) \cdot (1-p) \cdot (1-p) \cdot P(stemă) = (1-p)^4 \cdot p$$
$$\ldots$$
$$P(A_{2i+1}) = (1 - P(stemă))^{2i} \cdot P(stemă) = (1-p)^{2i} \cdot p$$

Prin urmare, probabilitatea ca Alina să câştige devine:

$$P(A) = \sum_{i=0}^{\infty} A_{2i+1} = \sum_{i=0}^{\infty} (1-p)^{2i} \cdot p = p \cdot \sum_{i=0}^{\infty} (1-p)^{2i} \overset{p \neq 0}{=} p \cdot \frac{1}{1-(1-p)^2} = \frac{p}{2p-p^2} = \frac{1}{2-p}$$

c. Jucătorul care aruncă primul banul câştigă jocul cu o probabilitate de $1/(2-p)$ (calculată mai sus). Cum $p \geq 0$, rezultă că $\frac{1}{2-p} \geq \frac{1}{2}$. Aşadar, în cazul în care există posibilitatea de a alege, este de preferat să arunci primul.

6. **(Formula lui Bayes)**

• ***CMU, 2001 fall, Andrew Moore, midterm exam, pr. 5***

Se consideră două variabile aleatoare A şi B despre care ştim următoarele informaţii:

a. $P(A \mid B) = 2/3$
b. $P(A \mid B) = 2/3$ **şi** $P(A \mid \bar{B}) = 1/3$
c. $P(A \mid B) = 2/3$, $P(A \mid \bar{B}) = 1/3$ **şi** $P(B) = 1/3$
d. $P(A \mid B) = 2/3$, $P(A \mid \bar{B}) = 1/3$, $P(B) = 1/3$ **şi** $P(A) = 4/9$.

În care din cele patru cazuri informaţiile date sunt suficiente pentru a calcula $P(B \mid A)$? Există vreun caz în care apar informaţii superflue (i.e, informaţii care pot fi deduse din celelalte informaţii furnizate în cazul respectiv)?

Răspuns:

Conform teoremei lui Bayes, vom avea:

$$P(B \mid A) = \frac{P(A \mid B) \cdot P(B)}{P(A)} = \frac{P(A \mid B) \cdot P(B)}{P(A \mid B) \cdot P(B) + P(A \mid \bar{B}) \cdot P(\bar{B})}$$
$$= \frac{P(A \mid B) \cdot P(B)}{P(A \mid B) \cdot P(B) + P(A \mid \bar{B}) \cdot (1 - P(B))}$$

Devine astfel evident că în cazurile c şi d informaţiile deţinute sunt suficiente, pe când în celelalte două cazuri nu. În cazul d, informaţia $P(A) = 4/9$ este superfluă.

7. **(Formula lui Bayes)**

• ○ ***CMU, 2008 spring, T. Mitchell, W. Cohen, midterm, pr. 1.5***

Presupunem că, răspunzând la o întrebare cu răspuns de genul adevărat / fals, un student fie cunoaşte răspunsul, fie ghiceşte răspunsul. Probabilitatea de a cunoaşte răspunsul este p, iar probabilitatea de a ghici răspunsul este $1 - p$.

Presupunem că probabilitatea de a răspunde corect la întrebare este

1 în cazul în care studentul cunoaşte răspunsul

şi 0.5 dacă studentul ghiceşte răspunsul.

Exprimaţi în funcţie de p care este probabilitatea ca studentul examinat să cunoască răspunsul la întrebare, în ipoteza că el a răspuns corect (notaţie: $P(knew \mid correct)$).

Răspuns:

Evenimentele de interes în problema dată sunt:
***correct* = studentul a răspuns corect la întrebare,**
***knew* = studentul cunoştea răspunsul corect**
şi complementarul acestuia din urmă:
guess $\stackrel{not.}{=} \neg$ ***knew* = studentul ghiceşte răspunsul.**

Aplicând formula lui Bayes obţinem:

$$P(knew \mid correct) = \frac{P(correct \mid knew) \cdot P(knew)}{P(correct \mid knew) \cdot P(knew) + P(correct \mid guess) \cdot P(guess)}$$
$$= \frac{1 \cdot p}{1 \cdot p + 0.5 \cdot (1 - p)} = \frac{p}{0.5p + 0.5} = \frac{p}{0.5(p+1)} = \frac{2p}{p+1}$$

8. **(Probabilităţi, chestiuni elementare: Adevărat sau Fals?)**

□ • ○ ***CMU, 2016 fall, N. Balcan, M. Gormley, HW1, pr. 6.1.1-4***

În cele de mai jos vom nota cu Ω spaţiul de eşantionare, iar cu $\bar{A}$ complementul evenimentului A.
Marcaţi cu *adevărat* sau *fals* fiecare dintre afirmaţiile următoare:

a. Pentru orice $A, B \subseteq \Omega$ astfel încât $P(A) > 0$ şi $P(B) > 0$, are loc egalitatea $P(A|B)P(B) = P(B|A)P(A)$.

b. Pentru orice $A, B \subseteq \Omega$ astfel încât $P(B) > 0$, are loc egalitatea $P(A \cup B) = P(A) + P(B) - P(A|B)$.

c. Pentru orice $A, B, C \subseteq \Omega$ astfel încât $P(B \cup C) > 0$, urmează că $\frac{P(A \cup B \cup C)}{P(B \cup C)} \geq P(A|B \cup C)P(B \cup C)$

d. Pentru orice $A, B \subseteq \Omega$ astfel încât $P(A) > 0$ şi $P(\bar{A}) > 0$, urmează că $P(B|A) + P(B|\bar{A}) = 1$.

***Indicaţie*:**
Pentru fiecare afirmaţie adevărată, faceţi demonstraţia proprietăţii respective. Pentru fiecare afirmaţie falsă, daţi fie un contraexemplu fie o justificare riguroasă.

Răspuns:

a. Adevărat. Demonstraţia este imediată.

b. Fals.
Ştim că pentru orice $A, B \subseteq \Omega$, are loc egalitatea $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Aşadar, egalitatea din enunţ ar fi adevărată dacă pentru orice $A, B \subseteq \Omega$ am avea $P(A \cap B) = P(A|B)$. Însă,

$$P(A \cap B) = P(A|B) \Leftrightarrow P(A \cap B) = \frac{P(A \cap B)}{P(B)} \Leftrightarrow P(B) = 1 \text{ sau } P(A \cap B) = 0.$$

Deci egalitatea nu este adevărată pentru orice A şi B.

c. Adevărat.
Observăm că dacă se notează $D = B \cup C$, atunci inegalitatea din enunţ devine mai uşor de manipulat: $\frac{P(A \cup D)}{P(D)} \geq P(A|D) \cdot P(D)$. Este imediat că ea este echivalentă cu $\frac{P(A \cup D)}{P(D)} \geq P(A \cap D)$, ceea ce implică $P(A \cup D) \geq P(A \cap D) \cdot P(D)$. Ultima inegalitate este adevărată fiindcă pe de o parte $P(A \cup D) \geq P(A \cap D)$ şi pe de altă parte $P(D) \in [0,\ 1]$.

d. Fals.
Ştim că pentru orice $A, B \subseteq \Omega$ cu $P(A) > 0$, are loc egalitatea $P(B|A) + P(\bar{B}|A) = 1$ (o puteţi demonstra imediat). Aşadar, egalitatea din enunţ ar fi adevărată dacă pentru orice $A, B \subseteq \Omega$ astfel încât $P(A) > 0$ şi $P(\bar{A}) > 0$ am avea $P(B|\bar{A}) = P(\bar{B}|A)$.

Vom construi următorul contraexemplu: considerăm că la aruncarea unui zar perfect, evenimentul A este apariţia unei feţe pare, iar evenimentul B este apariţia feţei 1. Urmează că

$$P(B|\bar{A}) = \frac{1}{3} \text{ şi } P(\bar{B}|A) = 1.$$

Deci egalitatea nu este adevărată pentru orice evenimente A (cu $P(A) > 0$ şi $P(\bar{A}) > 0$) şi B.

0.1.2 Variabile aleatoare

9. **(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. ■ ***CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh, HW1, pr. 1.1***

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

$$E[aX] = aE[X]. \tag{1}$$

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. ***CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh, HW1, pr. 1.3***

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

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

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. ***CMU, 2009 spring, Tom Mitchell, HW2, pr. 1.3.1***

***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:

$$Cov\,(X,Y) = E[XY] - E[X] \cdot E[Y]. \tag{3}$$

***Consecinţă imediată* (din relaţiile (1) şi (3)):**

$$Cov\,(aX, bY) = ab\,Cov\,(X,Y)\ \forall a, b \in \mathbb{R}. \tag{4}$$

***Observaţii*:**

1. Este imediat faptul că proprietatea (4) generalizează proprietatea (2).

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, la problema 19 se va demonstra că atunci când $Var\,(X) \neq 0$ şi $Var\,(Y) \neq 0$, avem următoarele margini (una inferioară şi cealaltă superioară) pentru $Cov\,(X,Y)$:

$$-Var(X) \cdot Var(Y) \leq Cov\,(X,Y) \leq +Var(X) \cdot Var(Y).$$

Răspuns:

a. Pentru cazul discret vom folosi o formă echivalentă a formulei pentru media unei variabile aleatoare şi anume $E[X] = \sum_{\omega \in \Omega} X(\omega) \cdot P(\omega)$. Prin urmare, putem scrie:

$$\begin{aligned}
E[W+Z] &= \sum_{u \in Val(W+Z)} u \cdot P(W+Z=u) \\
&= \sum_{w \in Val(W), z \in Val(Z)} (w+z) \cdot P(W+Z=w+z) \\
&= \sum_{w \in Val(W)} \sum_{z \in Val(Z)} (w+z) \cdot P(\{\omega \in \Omega \mid (W+Z)(\omega) = w+z\}) \\
&= \sum_{w \in Val(W)} \sum_{z \in Val(Z)} w \cdot P(\{\omega \in \Omega \mid (W+Z)(\omega) = w+z\}) + \\
&\quad \sum_{w \in Val(W)} \sum_{z \in Val(Z)} z \cdot P(\{\omega \in \Omega \mid (W+Z)(\omega) = w+z\}) \\
&= \sum_{w \in Val(W)} w \sum_{z \in Val(Z)} P(\{\omega \in \Omega \mid W(\omega) = w, Z(\omega) = z\}) + \\
&\quad \sum_{z \in Val(Z)} z \sum_{w \in Val(W)} P(\{\omega \in \Omega \mid W(\omega) = w, Z(\omega) = z\}) \\
&= \sum_{w \in Val(W)} wP(\{\omega \in \Omega \mid W(\omega) = w\}) + \sum_{z \in Val(Z)} zP(\{\omega \in \Omega \mid Z(\omega) = z\})
\end{aligned}$$

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

Pentru cazul continuu:

$$\begin{aligned} E[W+Z] &= \int_w \int_z (w+z)\, p_{WZ}(w,z) dzdw \\ &= \int_w \int_z w\, p_{WZ}(w,z) dzdw + \int_w \int_z z\, p_{WZ}(w,z) dzdw \\ &= \int_w w \int_z p_{WZ}(w,z) dzdw + \int_z z \int_w p_{WZ}(w,z) dwdz \\ &= \int_w w p_W(w) dw + \int_z z p_Z(z) dz \\ &= E[W] + E[Z]. \end{aligned}$$

În cazul în care variabila aleatoare X este discretă, egalitatea $E[aX] = aE[X]$ se poate demonstra imediat, aplicând definiţia mediei. Vom considera mai întâi subcazul $a \neq 0$:

$$\begin{aligned} E[aX] &= \sum_{u \in Val(aX)} u\, P(aX = u) = \sum_{x \in Val(X)} ax\, P(X = x), \quad \textbf{unde } x = \frac{1}{a}u \\ &= a \sum_{x \in Val(X)} x\, P(X = x) = aE[X]. \end{aligned}$$

Pentru subcazul $a = 0$, egalitatea $E[aX] = aE[X]$ este trivială.

Similar se face demonstraţia egalităţii $E[aX] = aE[X]$ în cazul în care variabila aleatoare X este continuă.

b. Pentru a demonstra această proprietate vom ţine cont de liniaritatea mediei (vedeţi *Observaţia* de la punctul a):

$$\begin{aligned} Var\,(X) &= E[(X - \bar{X})^2] = E[X^2 - 2X\bar{X} + \bar{X}^2] \\ &= E[X^2] - E[2X\bar{X}] + E[\bar{X}^2] = E[X^2] - 2\bar{X}E[X] + \bar{X}^2 \\ &= E[X^2] - 2\bar{X}^2 + \bar{X}^2 = E[X^2] - \bar{X}^2 = E[X^2] - (E[X])^2. \end{aligned}$$

Egalitatea $Var\,(aX) = a^2\, Var\,(X)$ rezultă imediat, aplicând definiţia varianţei şi ţinând cont de proprietatea $E[aX] = aE[X]$ pe care am demonstrat-o la punctul a.

$$\begin{aligned} Var\,(aX) &\overset{def.}{=} E[(aX - \underbrace{E[aX]}_{aE[X]})^2] = E[(a(X - E[X]))^2] \\ &= E[a^2(X - E[X])^2] = a^2 E[(X - E[X])^2] \overset{def.}{=} a^2\, Var\,(X). \end{aligned}$$

Alternativ, putem folosi proprietatea pe care am demonstrat-o mai sus:

$$\begin{aligned} Var\,(aX) &= E[(aX)^2] - (E[aX])^2 = E[a^2X^2] - (aE[X])^2 = a^2E[X^2] - a^2(E[X])^2 \\ &= a^2(E[X^2] - (E[X])^2) = a^2\, Var\,(X). \end{aligned}$$

c. Se foloseşte acelaşi gen de raţionament ca la punctul precedent (prima parte):

$$\begin{aligned} Cov\,(X,Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY - XE[Y] - E[X]Y + E[X]E[Y]] \\ &= E[XY] - E[XE[Y]] - E[E[X]Y] + E[E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] = E[XY] - E[X]E[Y]. \end{aligned}$$

10. **(Rezultat teoretic: covarianţa oricăror 2 variabile aleatoare independente este 0)**

■ ***CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh, HW1, pr. 1.2***

În mod intuitiv, două variabile aleatoare X şi Y sunt *independente* atunci când cunoaşterea valorii uneia dintre ele (de exemplu X) nu furnizează niciun indiciu despre valoarea celeilalte variabile (Y, în acest caz).

Formal, dacă X şi Y sunt variabile aleatoare discrete, independenţa lor revine la egalitatea $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$ pentru orice $x \in Val(X)$ şi orice $y \in Val(Y)$.

Similar, dacă X şi Y sunt variabile aleatoare continue, atunci $p(X = x, Y = y) = p(X = x) \cdot p(Y = y)$ pentru orice valori x şi y posibile.

Arătaţi că dacă X şi Y sunt variabile aleatoare independente de acelaşi tip (adică fie discret fie continuu), atunci

$$E[XY] = E[X] \cdot E[Y].$$

Echivalent: X, Y independente $\Rightarrow Cov(X, Y) = 0$. (A se vedea problema 9 punctul c.)

***Observaţie*: Reciproca implicaţiei de mai sus nu este în general adevărată. A se vedea problema 11.**

Răspuns:

Pentru cazul în care variabilele aleatoare independente X şi Y sunt discrete, putem scrie:

$$\begin{aligned}
E[XY] &= \sum_{x \in Val(X)} \sum_{y \in Val(Y)} xyP(X = x, Y = y) \\
&\overset{indep.}{=} \sum_{x \in Val(X)} \sum_{y \in Val(Y)} xyP(X = x) \cdot P(Y = y) \\
&= \sum_{x \in Val(X)} \Big(xP(X = x) \sum_{y \in Val(Y)} yP(Y = y) \Big) \\
&= \sum_{x \in Val(X)} xP(X = x)E[Y] = \Big(\sum_{x \in Val(X)} xP(X = x) \Big) E[Y] \\
&= E[X] \cdot E[Y]
\end{aligned}$$

Pentru cazul continuu demonstraţia este similară:

$$\begin{aligned}
E[XY] &= \int_x \int_y xy \; p(X = x, Y = y) dy dx \\
&\overset{indep.}{=} \int_x \int_y xy \; p(X = x) \cdot p(Y = y) dy dx \\
&= \int_x x \; p(X = x) \int_y y \; p(Y = y) dy \; dx \\
&= \int_x x \; p(X = x) E[Y] dx = E[Y] \cdot \int_x x \; p(X = x) \; dx \\
&= E[Y] \cdot E[X] = E[X] \cdot E[Y]
\end{aligned}$$

11. **(Covarianţa nulă nu implică în mod neapărat independenţa variabilelor aleatoare)**

CMU, 2009 spring, Tom Mitchell, HW2, pr. 1.3.2
CMU, 2009 fall, Geoff Gordon, HW1, pr. 3.1

X	Y	$P(X,Y)$
0	0	1/3
1	0	0
2	0	1/3
0	1	0
1	1	1/3
2	1	0

a. Reciproca afirmaţiei din problema 10 nu este în general adevărată, deci $Cov(X,Y) = 0 \not\Rightarrow X$ şi Y sunt independente. Arătaţi aceasta folosind ca exemplu variabilele aleatoare ale căror distribuţii sunt date în tabelul alăturat.

b. Totuşi, dacă X şi Y sunt variabile aleatoare binare luând valori în mulţimea $\{0,1\}$, iar $E[XY] = E[X] \cdot E[Y]$, atunci X şi Y sunt independente. Justificaţi.

(Aşadar, două variabile aleatoare binare cu valori în mulţimea $\{0,1\}$ sunt independente atunci şi numai atunci când au covarianţa nulă.)

Răspuns:

a. Din datele furnizate în exemplul a rezultă următoarele probabilităţi:

$$\begin{array}{lll} P(X=0)=1/3 & P(Y=0)=2/3 & P(XY=0)=2/3 \\ P(X=1)=1/3 & P(Y=1)=1/3 & P(XY=1)=1/3 \\ P(X=2)=1/3 & & P(XY=2)=0 \end{array}$$

Putem calcula mediile acestor variabile aleatoare folosind formula de definiţie pentru variabile discrete $E[X] = \sum_x x \cdot P(X=x)$:

$$\begin{aligned} E[X] &= 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + 2 \cdot \frac{1}{3} = 1 \\ E[Y] &= 0 \cdot \frac{2}{3} + 1 \cdot \frac{1}{3} = \frac{1}{3} \\ E[XY] &= 0 \cdot \frac{2}{3} + 1 \cdot \frac{1}{3} + 2 \cdot 0 = \frac{1}{3}. \end{aligned}$$

Conform formulei care a fost demonstrată la problema 10, covarianţa variabilelor X şi Y este: $Cov(X,Y) = E[XY] - E[X] \cdot E[Y] = \frac{1}{3} - 1 \cdot \frac{1}{3} = 0$. Cu toate acestea, variabilele X şi Y nu sunt independente. Într-adevăr, pentru $X = 0$ şi $Y = 0$ se observă că $P(X=0, Y=0) = 1/3$ dar $P(X=0)P(Y=0) = 1/3 \cdot 2/3 = 2/9 \neq \frac{1}{3}$.

Prin acest contraexemplu am arătat că implicaţia „$Cov(X,Y) = 0 \Rightarrow X$ şi Y sunt independente“ nu este în general adevărată.

b. În continuare vom demonstra că în cazul în care X şi Y iau valori în mulţimea $\{0,1\}$ implicaţia de mai sus este adevărată.

Dacă X şi Y sunt variabile aleatoare binare, atunci:

$$\begin{aligned} E[X] &= 0 \cdot P(X=0) + 1 \cdot P(X=1) = P(X=1) \\ E[Y] &= P(Y=1) \\ E[XY] &= P(X=1, Y=1) \end{aligned}$$

Covarianţa nulă înseamnă — conform problemelor 9 şi 10 — că $E[XY] = E[X] \cdot E[Y]$, adică:

$$P(X=1, Y=1) = P(X=1)P(Y=1)$$

Însă, a demonstra independenţa variabilelor aleatoare X şi Y revine la a arăta că $P(X,Y) = P(X)P(Y)$ pentru toate combinaţiile posibile de valori ale variabilelor. Unul din cazuri ($X = 1, Y = 1$) este deja demonstrat, deci mai există încă 3 cazuri. Pentru acestea vom utiliza formulele: $P(A \cap B) = P(A) - P(A \cap \bar{B})$ şi $P(\bar{A}) = 1 - P(A)$.

Cazul $X = 1, Y = 0$:

$$\begin{aligned} P(X=1,Y=0) &= P(X=1) - P(X=1,Y=1) \\ &= P(X=1) - P(X=1)P(Y=1) \\ &= P(X=1)(1 - P(Y=1)) \\ &= P(X=1)P(Y=0) \end{aligned}$$

Cazul $X = 0, Y = 1$ se tratează similar cu cazul anterior:

$$\begin{aligned} P(X=0,Y=1) &= P(Y=1) - P(X=1,Y=1) \\ &= P(Y=1) - P(X=1)P(Y=1) \\ &= P(Y=1)(1 - P(X=1)) \\ &= P(Y=1)P(X=0) \end{aligned}$$

Cazul $X = 0, Y = 0$:

$$\begin{aligned} P(X=0,Y=0) &= P(X=0) - P(X=0,Y=1) \\ &= P(X=0) - P(X=0)P(Y=1) \\ &= P(X=0)(1 - P(Y=1)) \\ &= P(X=0)P(Y=0) \end{aligned}$$

Prin urmare, egalitatea $P(X,Y) = P(X)P(Y)$ este adevărată pentru toate cazurile, deci variabilele X şi Y sunt independente.

12. **(Variabile aleatoare: calcul de medii)**

Fie X o variabilă aleatoare pentru care $E(X) = \mu$ şi $Var(X) = \sigma^2$.

a. ***CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW1, pr. 2.4***

Cât este $E[X(X-1)]$ în funcţie de μ şi σ?

b. ***CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, midterm, pr. 1.c***

Fie $c \in R$. Care din următoarele variante sunt adevărate?

A. $E[(X-c)^2] = (\mu - c)^2 + \sigma^2$
B. $E[(X-c)^2] = (\mu - c)^2$
C. $E[(X-c)^2] = (\mu - c)^2 - \sigma^2$
D. $E[(X-c)^2] = (\mu - c)^2 + 2\sigma^2$
E. $E[(X-c)^2] = \mu^2 + c^2 + 2\sigma^2$
F. $E[(X-c)^2] = \mu^2 + c^2 - 2\sigma^2$.

Răspuns:

a. De la problema 9.a ştim că media sumei a două variabile aleatoare este suma mediilor variabilelor respective. De asemenea, este imediat demonstrabil că

$E[c \cdot X] = c \cdot E[X]$, unde X, Y sunt variabile aleatoare iar c este o constantă. Ţinând cont că $Var(X) = E[X^2] - (E[X])^2$ (vedeţi problema 9.b), putem scrie:

$$\begin{aligned} E[X(X-1)] &= E[X^2 - X] = E[X^2] - E[X] \\ &= E[X^2] - (E[X])^2 + (E[X])^2 - E[X] \\ &= Var(X) + (E[X])^2 - E[X] \\ &= \sigma^2 + \mu^2 - \mu \end{aligned}$$

b. Pentru a găsi varianta adevărată vom calcula $E[(X-c)^2]$:

$$\begin{aligned} E[(X-c)^2] &= E[X^2 - 2cX + c^2] = E[X^2] - 2cE[X] + c^2 \\ &= E[X^2] - (E[X])^2 + (E[X])^2 - 2cE[X] + c^2 \\ &= \sigma^2 + \mu^2 - 2c\mu + c^2 \\ &= \sigma^2 + (\mu - c)^2 \end{aligned}$$

Deci varianta A este adevărată. (Toate celelalte variante sunt, în general, false.)

13. **(Variabile aleatoare discrete: distribuţii de probabilitate comune, marginale, condiţionate; regula de înlănţuire)**

CMU, 2002 fall, Andrew Moore, final exam, pr. 4.a

Considerăm un set de date definit cu ajutorul a 3 variabile aleatoare cu valori booleene X, Y şi Z. Care dintre seturile de informaţii de mai jos sunt suficiente pentru a specifica distribuţia comună $P(x, y, z)$?

A.	**B.**	**C.**	**D.**
$P(\neg X \mid Z)$	$P(\neg X \mid \neg Z)$	$P(X \mid Z)$	$P(X \mid Z)$
$P(\neg X \mid \neg Z)$	$P(X \mid \neg Z)$	$P(X \mid \neg Z)$	$P(X \mid \neg Z)$
$P(\neg Y \mid X, Z)$	$P(Y \mid X, Z)$	$P(Y \mid X, Z)$	$P(Y \mid X, Z)$
$P(\neg Y \mid X, \neg Z)$	$P(Y \mid X, \neg Z)$	$P(Y \mid X, \neg Z)$	$P(Y \mid X, \neg Z)$
$P(\neg Y \mid \neg X, Z)$	$P(Y \mid \neg X, Z)$	$P(Y \mid \neg X, Z)$	$P(\neg Y \mid \neg X, \neg Z)$
$P(\neg Y \mid \neg X, \neg Z)$	$P(Y \mid \neg X, \neg Z)$	$P(\neg Y \mid \neg X, \neg Z)$	$P(Y \mid \neg X, \neg Z)$
$P(Z)$	$P(Z)$	$P(\neg Z)$	$P(Z)$

Răspuns:

Pentru a calcula distribuţia comună a mai multor variabile aleatoare se poate aplica regula de înlănţuire.[15] În cazul nostru, putem scrie:

$$P(X, Y, Z) = P(Z) \cdot P(X \mid Z) \cdot P(Y \mid X, Z)$$

Deoarece variabilele aleatoare X, Y şi Z au valori booleene, pentru a specifica distribuţia comună $P(X, Y, Z)$ este nevoie să se calculeze valoarea acesteia în fiecare din cele 8 cazuri posibile:

[15] Pentru variabile aleatoare, regula de înlănţuire

$$P(A_1, A_2, \ldots, A_n) = P(A_1) \cdot P(A_2 \mid A_1) \cdot P(A_n \mid A_1, A_2, \ldots, A_{n-1})$$

se demonstrează imediat pornind de la regula de înlănţuire pentru (probabilităţi de) evenimente aleatoare. A se vedea problema 14.

$$\begin{array}{cccc} P(X,Y,Z) & P(X,Y,\neg Z) & P(X,\neg Y,Z) & P(X,\neg Y,\neg Z) \\ P(\neg X,Y,Z) & P(\neg X,Y,\neg Z) & P(\neg X,\neg Y,Z) & P(\neg X,\neg Y,\neg Z) \end{array}$$

Cunoaştem de asemenea relaţii de calcul de forma:

$$\begin{aligned} P(\neg X) &= 1 - P(X) \\ P(\neg X \mid Y) &= 1 - P(X \mid Y) \end{aligned}$$

Aşadar, pentru a aplica regula de înlănţuire de mai sus este nevoie să cunoaştem

$P(Z)$ **sau** $P(\neg Z)$**;**

$P(X \mid Z)$ **sau** $P(\neg X \mid Z)$**;**
$P(X \mid \neg Z)$ **sau** $P(\neg X \mid \neg Z)$**;**

$P(Y \mid X, Z)$ **sau** $P(\neg Y \mid X, Z)$**;**
$P(Y \mid \neg X, Z)$ **sau** $P(\neg Y \mid \neg X, Z)$**;**
$P(Y \mid X, \neg Z)$ **sau** $P(\neg Y \mid X, \neg Z)$**;**
$P(Y \mid \neg X, \neg Z)$ **sau** $P(\neg Y \mid \neg X, \neg Z)$**.**

Cu aceste precizări, putem specifica pentru fiecare dintre seturile de informaţii din enunţ dacă sunt suficiente pentru a calcula distribuţia comună $P(X, Y, Z)$**.**

Cazul A. Da. Se observă că putem calcula $P(X \mid Z)$ **şi** $P(X \mid \neg Z)$ **din primele două probabilităţi din enunţ. De asemenea, utilizând următoarele 4 probabilităţi se pot deduce:** $P(Y \mid X, Z)$**,** $P(Y \mid X, \neg Z)$**,** $P(Y \mid \neg X, Z)$ **şi** $P(Y \mid \neg X, \neg Z)$**. Iar din** $P(Z)$ **se obţine** $P(\neg Z)$**. Prin urmare, există toate informaţiile necesare distribuţiei comune** $P(X, Y, Z)$**.**

Cazul B. Nu, informaţiile din enunţ nu sunt suficiente. Nu putem deduce valoarea pentru $P(X \mid Z)$**.**

Cazul C. Da, informaţiile din enunţ sunt suficiente. Din $P(X \mid Z)$ **se obţine** $P(\neg X \mid Z)$**, iar din** $P(X \mid \neg Z)$ **se obţine** $P(\neg X \mid \neg Z)$**. Din următoarele 4 probabilităti se obţin celelalte 4 necesare pentru** $P(Y \mid X, Z)$**, şi anume:** $P(\neg Y \mid X, Z)$**,** $P(\neg Y \mid X, \neg Z)$**,** $P(\neg Y \mid \neg X, Z)$ **şi respectiv** $P(Y \mid \neg X, \neg Z)$**.**

Cazul D. Nu, informaţiile din enunţ nu sunt suficiente. Nu putem deduce valoarea pentru $P(Y \mid \neg X, Z)$**.**

14. **(Variabile aleatoare discrete: regula de multiplicare, varianta condiţională)**

CMU, 2009 fall, Geoff Gordon, HW2, pr. 2.1

Arătaţi că pentru orice valori x, y **şi** z **ale variabilelor aleatoare** X, Y **şi** Z **respectiv, avem:**

$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Y = y, Z = z) \cdot P(Y = y \mid Z = z).$$

În notaţie simplificată: $P(X, Y \mid Z) = P(X \mid Y, Z) \cdot P(Y \mid Z)$**.**

***Indicaţie*: Folosiţi regula de înlănţuire pentru evenimente aleatoare.**

Răspuns:

Folosind definiţia probabilităţii condiţionate şi regula de înlănţuire (cu termenii ordonaţi în mod convenabil), egalitatea cerută se obţine astfel:

$$P(X,Y \mid Z) \overset{def.}{=} \frac{P(X,Y,Z)}{P(Z)} \overset{not.}{=} \frac{P(Z \cap Y \cap X)}{P(Z)}$$
$$= \frac{P(Z)P(Y \mid Z)P(X \mid Y,Z)}{P(Z)} = P(Y \mid Z) \cdot P(X \mid Y,Z)$$

***Observaţie*: O altă metodă de rezolvare constă în a aplica pentru fiecare membru al egalităţii din enunţ definiţia probabilităţii condiţionate, după simplificări obţinându-se pentru ambii membri aceeaşi valoare:**

$$P(X,Y \mid Z) = \frac{P(X,Y,Z)}{P(Z)}$$
$$P(X \mid Y,Z) \cdot P(Y \mid Z) = \frac{P(X,Y,Z)}{P(Y,Z)} \cdot \frac{P(Y,Z)}{P(Z)} = \frac{P(X,Y,Z)}{P(Z)}$$

15. **(Variabile aleatoare discrete: independenţă condiţională)**

CMU, 2005 fall, T. Mitchell, A. Moore, midterm, pr. 4

Fie variabilele aleatoare discrete A, B şi C având distribuţia comună conform tabelului de mai jos.

a. Este variabila A independentă condiţional de B în raport cu variabila C?

b. Dacă aţi răspuns afirmativ la întrebarea a, faceţi o schimbare în primele două linii ale tabelului de mai sus pentru a obţine o distribuţie pentru care răspunsul la aceeaşi întrebare să devină negativ. Invers, dacă aţi răspuns negativ la întrebarea a, faceţi o schimbare în primele două linii ale tabelului încât răspunsul să devină afirmativ.

A	B	C	$P(A,B,C)$
0	0	0	1/8
0	0	1	1/8
0	1	0	1/8
0	1	1	1/8
1	0	0	1/8
1	0	1	1/8
1	1	0	1/8
1	1	1	1/8

Răspuns:

a. Faptul că variabila A este independentă condiţional de B în raport cu variabila C se mai notează prin $A \perp B \mid C$ şi poate fi demonstrat prin una din următoarele două relaţii:

$$P(A=a, B=b \mid C=c) = P(A=a \mid C=c) \cdot P(B=b \mid C=c) \text{ sau}$$
$$P(A=a \mid B=b, C=c) = P(A=a \mid C=c), \text{ dacă } P(B=b, C=c) \neq 0.$$

pentru orice $a \in Val(A), b \in Val(B), c \in Val(C)$. Deşi în general se foloseşte prima relaţie, pentru acest exerciţiu este mai uşor să utilizăm cea de-a doua relaţie. Conform tabelului dat în enunţ, rezultă imediat că $P(B=b, C=c) \neq 0$ pentru orice $b, c \in \{0,1\}$. Vom demonstra că pentru orice $a, b, c \in \{0,1\}$ este adevărat că $P(A=a \mid B=b, C=c) = P(A=a \mid C=c)$. Cele două probabilităţi condiţionate vor fi calculate folosind datele din tabel:

Cazul $(0,0,0)$: $P(A=0 \mid B=0, C=0) = \dfrac{1 \cdot \frac{1}{8}}{2 \cdot \frac{1}{8}} = \dfrac{1}{2}$ **şi** $P(A=0 \mid C=0) = \dfrac{2 \cdot \frac{1}{8}}{4 \cdot \frac{1}{8}} = \dfrac{1}{2}$**.**

Cazul $(0,0,1)$**:** $P(A=0 \mid B=0, C=1) = \dfrac{1 \cdot \frac{1}{8}}{2 \cdot \frac{1}{8}} = \dfrac{1}{2}$ **şi** $P(A=0 \mid C=1) = \dfrac{2 \cdot \frac{1}{8}}{4 \cdot \frac{1}{8}} = \dfrac{1}{2}$**.**

Se observă că pentru toate celelalte cazuri se obţin aceleaşi valori, deci

$$P(A=a \mid B=b, C=c) = P(A=a \mid C=c),\ \forall a,b,c \in \{0,1\}.$$

Aşadar, variabila A **este independentă condiţional de** B **în raport cu variabila** C**.**

b. Schimbarea trebuie făcută în aşa fel încât să se păstreze relaţia $\sum P(A,B,C) = 1$**. O variantă posibilă este:**

A	B	C	$P(A,B,C)$
0	0	0	1/4
0	0	1	0
	...		...

Pentru aceste noi valori se observă că: $P(A=0 \mid B=0, C=0) = \dfrac{1 \cdot 1/4}{1 \cdot 1/4 + 1 \cdot 1/8} = \dfrac{2}{3}$ **şi** $P(A=0 \mid C=0) = \dfrac{1 \cdot 1/4 + 1 \cdot 1/8}{1 \cdot 1/4 + 3 \cdot 1/8} = \dfrac{3}{5}$**. Este suficient un singur caz în care probabilităţile respective nu sunt egale, prin urmare variabila** A **nu este independentă condiţional de variabila** B **în raport cu a treia variabilă,** C**.**

16. **(Variabile aleatoare discrete: distribuţii comune, distribuţii marginale, distribuţii condiţionale; independenţă, independenţă condiţională)**

• ***CMU, 2016 fall, N. Balcan, M. Gormley, HW2, pr. 1.4***

Fie trei variabile aleatoare X**,** Y **şi** Z **care iau valori în mulţimea** $\{0,1\}$**. În tabelul de mai jos este dată distribuţia probabilistă comună a acestor trei variabile,** $P(X,Y,Z)$**.**

	$Z=0$		$Z=1$	
	$X=0$	$X=1$	$X=0$	$X=1$
$Y=0$	1/24	1/12	1/12	5/24
$Y=1$	1/12	p	q	7/24

a. Considerând că X **şi** Y **sunt independente, găsiţi valorile lui** p **şi** q**.**

***Indicaţie*: Este util (deşi nu obligatoriu) să calculaţi mai întâi** $P(X,Y)$**, distribuţia comună a variabilelor** X **şi** Y**, completând tabelul alăturat, după care veţi calcula şi distribuţiile (marginale) pentru** X **şi** Y**, de preferinţă ca o linie şi respectiv o coloană suplimentară la acest tabel.**

	$X=0$	$X=1$
$Y=0$		
$Y=1$		

b. Considerând valorile lui p **şi** q **determinate la punctul** a**, sunt** X **şi** Y **independente condiţional în raport cu** Z**? De ce?**

Răspuns:

a. Conform definiţiei, variabilele aleatoare X şi Y sunt independente dacă şi numai dacă

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y), \forall x \in Val\,(X), \forall y \in Val\,(Y). \tag{5}$$

***Observaţie*: Evident, pentru a determina p şi q, ar fi de dorit să scriem un sistem de două ecuaţii cu aceste două necunoscute. Cele două ecuaţii pot fi obţinute prin particularizarea relaţiei (5) pentru două perechi distincte de valori $(x, y) \in Val(X) \times Val(Y)$. Alternativ, va fi suficient să reţinem doar o astfel de ecuaţie, pentru că în locul celeilalte putem folosi una din proprietăţile din definiţia funcţiei de probabilitate, şi anume**

$$\sum_{x,y,z} P(X = x, Y = y, Z = z) = 1. \tag{6}$$

Aceasta revine, practic la a însuma elementele din tabelul dat în enunţ pentru distribuţia $P(X, Y, Z)$.[16]

Pentru a calcula probabilităţile marginale $P(X = x, Y = y)$, vom proceda conform definiţiei: $P(X = x, Y = y) = \sum_{z \in Val\,(Z)} P(X = x, Y = y, Z = z)$. Şi anume:

$$\begin{aligned}
P(X = 0, Y = 0) &= P(X = 0, Y = 0, Z = 0) + P(X = 0, Y = 0, Z = 1) = \frac{1}{24} + \frac{1}{12} = \frac{1}{8} \\
P(X = 0, Y = 1) &= P(X = 0, Y = 1, Z = 0) + P(X = 0, Y = 1, Z = 1) = \frac{1}{12} + q \\
P(X = 1, Y = 0) &= P(X = 1, Y = 0, Z = 0) + P(X = 1, Y = 0, Z = 1) = \frac{1}{12} + \frac{5}{24} = \frac{7}{24} \\
P(X = 1, Y = 1) &= P(X = 1, Y = 1, Z = 0) + P(X = 1, Y = 1, Z = 1) = \frac{7}{24} + p.
\end{aligned}$$

În mod similar,

$$\begin{aligned}
P(X = 0) &= \sum_{y \in Val\,(Y)} P(X = x, Y = y) = \frac{1}{8} + \frac{1}{12} + q = \frac{5}{24} + q \\
P(X = 1) &= \frac{7}{24} + p + \frac{7}{24} = \frac{7}{12} + p
\end{aligned}$$

şi

$$\begin{aligned}
P(Y = 0) &= \sum_{x \in Val\,(X)} P(X = x, Y = y) = \frac{1}{8} + \frac{7}{24} = \frac{5}{12} \\
P(Y = 1) &= 1 - P(Y = 0) = \frac{7}{12}.
\end{aligned}$$

Punând acum toate aceste rezultate împreună, obţinem:

	$X = 0$	$X = 1$	
$Y = 0$	$1/8$	$7/24$	$5/12$
$Y = 1$	$1/12 + q$	$7/24 + p$	$7/12$
	$5/24 + q$	$7/12 + p$	

[16] Remarcăm că, dacă vom urma *Indicaţia* din enunţ, după ce vom scrie probabilităţile marginale $P(X, Y)$, $P(X)$ şi $P(Y)$, va fi uşor să scriem încă alte trei proprietăţi similare cu (6). Deci, o vom putea alege atunci pe cea mai „directă"/simplă dintre ele.

Acum este simplu de văzut că $5/24 + q + p + 14/24 = 1 \Leftrightarrow p + q = 5/24$**. (Aceasta constituie prima noastră ecuaţie în** p **şi** q**.) În consecinţă, am putea chiar să eliminăm** q **din tabelul de mai sus, scriind** $P(X = 0, Y = 1) = 1/12 + q = 1/12 + 5/24 - p = 7/24 - p$ **şi** $P(X = 0) = 5/24 + q = 5/24 + 5/24 - p = 5/12 - p$**.**

Aplicând definiţia independenţei variabilelor X **şi** Y **pentru** $x = 0$ **şi** $y = 0$**, vom avea:**

$$P(X = 0, Y = 0) = P(X = 0) \cdot P(Y = 0) \Leftrightarrow \frac{1}{8} = \left(\frac{5}{12} - p\right) \cdot \frac{5}{12} \Leftrightarrow \frac{5}{12} - p = \frac{3}{2} \cdot \frac{1}{5}$$
$$\Leftrightarrow p = \frac{5}{12} - \frac{3}{10} = \frac{25 - 18}{60} = \frac{7}{60},$$

de unde rezultă imediat

$$q = \frac{5}{24} - p = \frac{5}{24} - \frac{7}{60} = \frac{25 - 14}{120} = \frac{11}{120}.$$

De asemenea, vom avea $P(X = 0) = 5/12 - p = 5/12 - 7/60 = 3/10$ **şi** $P(X = 1) = 7/10$**.**

În final, va trebui să verificăm şi celelalte trei egalităţi din definiţia independenţei lui X **şi** Y**, pentru că, în general, este posibil ca sistemul de ecuaţii corespunzător relaţiei (5) să fie supra-resctricţionat (deci, incompatibil):**[17]

$P(X = 0, Y = 1) = P(X = 0) \cdot P(Y = 1)$ **:**

$$\frac{7}{24} - p = \frac{3}{10} \cdot \frac{7}{12} \Leftrightarrow \frac{7}{24} - \frac{7}{60} = \frac{7}{40} \Leftrightarrow 7 \cdot \left(\frac{1}{24} - \frac{1}{60}\right) = \frac{7}{40} \Leftrightarrow 7 \cdot \frac{5-2}{120} = \frac{7}{40} \ (A)$$

$P(X = 1, Y = 0) = P(X = 1) \cdot P(Y = 0)$ **:**

$$\frac{7}{24} = \frac{7}{10} \cdot \frac{5}{12} \ (A)$$

$P(X = 1, Y = 1) = P(X = 1) \cdot P(Y = 1)$ **:**

$$\frac{7}{24} + p = \frac{7}{10} \cdot \frac{7}{12} \Leftrightarrow \frac{7}{24} + \frac{7}{60} = 7^2 \cdot \frac{1}{10} \cdot \frac{1}{12} \Leftrightarrow \frac{1}{24} + \frac{1}{60} = \frac{7}{120} \Leftrightarrow \frac{5+2}{120} = \frac{7}{120} \ (A)$$

Aşadar, date fiind cele două valori ale lui p **şi** q**, variabilele** X **şi** Y **sunt independente.**

b. Conform definiţiei, variabilele aleatoare X **şi** Y **sunt independente condiţional în raport cu variabila** Z **dacă şi numai dacă**

$$P(X = x, Y = y | Z = z) = P(X = x | Z = z) \cdot P(Y = y | Z = z), \quad (7)$$
$$\forall x \in Val(X), \forall y \in Val(Y), \forall z \in Val(Z).$$

Dacă X **şi** Y ***nu*** **sunt independente condiţional în raport cu** Z**, atunci** $\exists x \in Val(X), \exists y \in Val(Y), \exists z \in Val(Z)$ **astfel încât** $P(X = x, Y = y | Z = z) \neq P(X = x | Z = z) \cdot P(Y = y | Z = z)$**.**

Pentru a putea exprima probabilităţile condiţionate care apar în formulele de mai sus, vom calcula în prealabil distribuţia marginală a lui Z **pornind de la tabelul din enunţ:**

$$\begin{aligned} P(Z = 0) &= \sum_{x \in Val(X)} \sum_{y \in Val(Y)} P(X = x, Y = y, Z = z) \\ &= 5/24 + p = 5/24 + 7/60 = 39/120 = 13/40 \\ P(Z = 1) &= 1 - P(Z = 0) = 27/40. \end{aligned}$$

[17] Ar fi suficient chiar să verificăm doar două dintre cele trei egalităţi, fiindcă ultima decurge automat din celelalte, ţinând cont de proprietăţile $\sum_{x,y} P(X = x, Y = y) = 1$, $\sum_x P(X = x) = 1$ şi $\sum_y P(Y = y) = 1$.

Vom verifica relaţia (7) pentru $x = y = z = 0$:

$$\begin{aligned}
&P(X=0,Y=0|Z=0) = P(X=0|Z=0)\cdot P(Y=0|Z=0)\\
\Leftrightarrow\ &\frac{P(X=0,Y=0,Z=0)}{P(Z=0)} = \frac{P(X=0,Z=0)}{P(Z=0)}\cdot\frac{P(Y=0,Z=0)}{P(Z=0)}\\
\Leftrightarrow\ &P(X=0,Y=0,Z=0)\cdot P(Z=0) = P(X=0,Z=0)\cdot P(Y=0,Z=0)\\
\Leftrightarrow\ &\frac{1}{24}\cdot\frac{13}{40} = \left(\frac{1}{24}+\frac{1}{12}\right)\cdot\left(\frac{1}{24}+\frac{1}{12}\right)\\
\Leftrightarrow\ &\frac{13}{24\cdot 40} = \frac{1}{8}\cdot\frac{1}{8} \Leftrightarrow \frac{13}{3\cdot 5} = 1\ (F)
\end{aligned}$$

Prin urmare, variabilele X şi Y nu sunt independente condiţional în raport cu variabila Z.

***Observaţie*: Această problemă pune în evidenţă următoarea *proprietate*, care merită a fi reţinută: independenţa a două variabile aleatoare nu implică (în general) independenţa lor condiţională în raport cu o a treia variabilă aleatoare.**

17. **(Variabile aleatoare continue: funcţia densitate de probabilitate)**

CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW1, pr. 1.5

Fie variabila aleatoare continuă X a cărei funcţie densitate de probabilitate (în limba engleză: "probability density functions"; scris, sub formă abreviată, p.d.f.) este:

$$p(x) = \begin{cases} cx^2 & \text{pentru } 1 \le x \le 2\\ 0 & \text{în caz contrar.}\end{cases}$$

a. Ţinând cont de proprietăţile funcţiei densitate de probabilitate (a se vedea notiţele de la curs), cât trebuie să fie valoarea constantei c?

b. Desenaţi graficul funcţiei de mai sus.

c. Calculaţi $P(X > 3/2)$.

Răspuns:

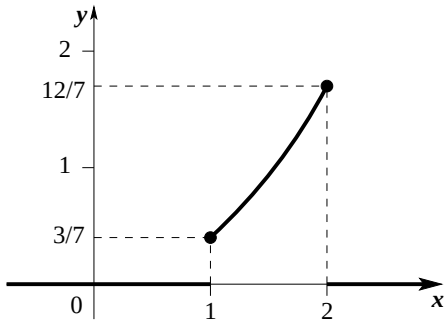
a. Faptul că $p(x)$ este funcţie densitate de probabilitate pentru variabila aleatoare continuă X înseamnă că $p(x) \ge 0, \forall x$ şi că $\int_{-\infty}^{+\infty} p(x)dx = 1$. Pentru a afla constanta c vom calcula valoarea integralei:

$$\begin{aligned}
\int_{-\infty}^{+\infty} p(x)dx &= \underbrace{\int_{-\infty}^{1} p(x)dx}_{=0} + \int_1^2 p(x)dx + \underbrace{\int_2^{+\infty} p(x)dx}_{=0}\\
&= \int_1^2 cx^2dx = c\cdot\int_1^2 x^2dx = c\cdot\left.\frac{x^3}{3}\right|_1^2 = c\left(\frac{2^3}{3}-\frac{1^3}{3}\right) = c\cdot\frac{7}{3}
\end{aligned}$$

Prin urmare, $c\cdot\frac{7}{3} = 1 \Rightarrow c = \frac{3}{7}$.

b. Trebuie să reprezentăm grafic funcţia:

$$p(x) = \begin{cases} \frac{3}{7}x^2 & \textbf{pentru } 1 \le x \le 2 \\ 0 & \textbf{în caz contrar}. \end{cases}$$



În intervalul $[1,2]$, această funcţie este un fragment din parabola corespunzătoare funcţiei de gradul al doilea: $\frac{3}{7}x^2$, parabolă care are vârful în punctul $(0,0)$. Putem calcula $p(1) = 3/7 \approx 0.42$ şi $p(2) = 12/7 \approx 1.71$.

c. Valoarea probabilităţii cerute se poate calcula astfel:

$$\begin{aligned} P(X > 3/2) &\overset{not.}{=} P(\{\omega \mid X(\omega) > \frac{3}{2}\}) \overset{def.}{=} \int_{X(\omega)=x>3/2} p(x)dx = \int_{3/2}^{+\infty} p(x)dx \\ &= \int_{3/2}^{2} p(x)dx + \underbrace{\int_{2}^{+\infty} p(x)dx}_{=0} = \int_{3/2}^{2} \frac{3}{7}x^2 dx = \frac{3}{7} \cdot \int_{3/2}^{2} x^2 dx \\ &= \frac{3}{7} \cdot \frac{x^3}{3}\Big|_{3/2}^{2} = \frac{1}{7} \cdot x^3\big|_{3/2}^{2} = \frac{1}{7} \cdot \left(8 - \frac{27}{8}\right) = \frac{1}{7} \cdot \frac{64-27}{8} = \frac{37}{56}. \end{aligned}$$

O altă variantă de rezolvare este bazată pe folosirea *funcţiei cumulative de distribuţie*, care, după cum ştim, se defineşte prin relaţia $F(x) \overset{def.}{=} P(X \le x)$, pentru orice $x \in \mathbb{R}$:

$$\begin{aligned} P(X > 3/2) &= 1 - P(X \le 3/2) = 1 - F\left(\frac{3}{2}\right) = 1 - \int_{-\infty}^{3/2} p(x)dx \\ &= 1 - \left(\int_{-\infty}^{1} \underbrace{p(x)}_{0} dx + \int_{1}^{3/2} p(x)dx\right) = 1 - \left(0 + \int_{1}^{3/2} p(x)dx\right) \\ &= 1 - \int_{1}^{3/2} p(x)dx = 1 - \int_{1}^{3/2} \frac{3}{7}x^2 dx = 1 - \frac{1}{7} \cdot x^3\big|_{1}^{3/2} \\ &= 1 - \frac{1}{7} \cdot \left(\frac{27}{8} - 1\right) = 1 - \frac{1}{7} \cdot \frac{19}{8} \\ &= 1 - \frac{19}{56} = \frac{37}{56}. \end{aligned}$$

18. **(Variabile aleatoare continue: funcţia densitate de probabilitate)**

CMU, 2008 spring, Eric Xing, HW1, pr. 1.1.b

Fie funcţia

$$p(x) = \begin{cases} cx^{-d} & \textbf{pentru } x > 1 \\ 0 & \textbf{în caz contrar}. \end{cases}$$

Care sunt valorile posibile pentru c şi d în aşa fel ca p să poată reprezenta o funcţie densitate de probabilitate?

Răspuns:

O funcţie p poate reprezenta o funcţie densitate de probabilitate dacă $p(x) \geq 0$ pentru $\forall x$, şi $\int_{-\infty}^{\infty} p(x)dx = 1$. Vom folosi aceste două condiţii pentru a calcula valorile posibile pentru c şi d.

Prima condiţie $p(x) \geq 0$ implică faptul că $c \geq 0$, asupra lui d neimpunând nicio restricţie. Mai mult, ţinând cont de forma lui p şi de cea de-a doua condiţie, vom avea chiar $c > 0$, fiindcă $c = 0$ ar implica $\int_{-\infty}^{\infty} p(x)dx = 0 \neq 1$.

Pentru a aplica cea de-a doua condiţie, calculăm integrala:

$$\int_{-\infty}^{\infty} p(x)dx = \int_{1}^{\infty} cx^{-d}dx = c\int_{1}^{\infty} x^{-d}dx.$$

Vom avea de tratat două cazuri:

$$\textbf{Cazul 1:} \quad d = 1 \quad \Rightarrow c\int_{1}^{\infty} x^{-d}dx = c\int_{1}^{\infty} \frac{1}{x}dx = c \cdot \ln x\big|_1^{\infty} = \infty$$

$$\textbf{Cazul 2:} \quad d \neq 1 \quad \Rightarrow c\int_{1}^{\infty} x^{-d}dx = c \cdot \frac{x^{-d+1}}{-d+1}\bigg|_1^{\infty} = \frac{c}{1-d} \cdot x^{1-d}\big|_1^{\infty}$$

Subcazul $d > 1$: $x^{1-d}\big|_1^{\infty} = 0 - 1 = -1$. **Deci în acest caz** $c\int_1^{\infty} x^{-d}dx = \frac{c}{d-1}$.

Subcazul $d < 1$: $x^{1-d}\big|_1^{\infty} = +\infty$. **Deci în acest caz** $c\int_1^{\infty} x^{-d}dx = +\infty$.

Prin urmare, $\int_{-\infty}^{\infty} p(x)dx = 1 \Rightarrow d > 1$ **şi** $\frac{c}{d-1} = 1$.

În concluzie, p poate reprezenta o funcţie densitate de probabilitate dacă

$$c > 0,\ d > 1 \text{ şi } c = d - 1.$$

Referitor la aceste trei restricţii, se observă că ultimele două o implică pe cea dintâi, deci o fac superfluă.

19. **(Coeficientul de corelaţie pentru două variabile aleatoare: două proprietăţi)**

Liviu Ciortuz, 2019, după
Sheldon Ross, A First Course in Probability, 5th ed., Prentice Hall, 1997, pag. 332

Pentru două variabile aleatoare oarecare X şi Y având $Var(X) \neq 0$ şi $Var(Y) \neq 0$, *coeficientul de corelaţie* se defineşte astfel:

$$\rho(X,Y) \stackrel{def.}{=} \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}}.$$

a. Să se demonstreze că $-1 \leq \rho(X,Y) \leq 1$.

***Consecinţă*: $Cov(X,Y) \in [-\sqrt{Var(X) \cdot Var(Y)}, +\sqrt{Var(X)\ Var(Y)}]$, dacă $Var(X) \neq 0$ şi $Var(Y) \neq 0$.**

b. Să se arate că dacă $\rho(X,Y)=1$, atunci $Y=aX+b$, cu $a=Var\,(Y)/Var\,(X)>0$. Similar, dacă $\rho(X,Y)=-1$, atunci $Y=aX+b$, cu $a=-Var\,(Y)/Var\,(X)<0$.[18]

***Indicaţii*:**

1. La punctul a, notând $Var(X)=\sigma_X$ şi $Var(Y)=\sigma_Y$, pentru a demonstra inegalitatea $\rho(X,Y)\geq -1$ vă sugerăm să dezvoltaţi expresia $Var\Big(\frac{X}{\sigma_X}+\frac{Y}{\sigma_Y}\Big)$ folosind următoarele două proprietăţi, valabile pentru orice variabile aleatoare X şi Y: $Var(X+Y)=Var(X)+Var(Y)+2Cov(X,Y)$,[19] şi $Cov(aX,bY)=abCov(X,Y)$, pentru orice $a,b\in\mathbb{R}$.[20] Apoi veţi proceda similar, dezvoltând expresia $Var\Big(\frac{X}{\sigma_X}-\frac{Y}{\sigma_Y}\Big)$, ca să demonstraţi inegalitatea $\rho(X,Y)\leq 1$.

2. La punctul b, veţi ţine cont de faptul că pentru o variabilă aleatoare oarecare X, avem $Var(X)=0$ dacă şi numai dacă variabila X este constantă.[21]

Răspuns:

a. Pentru a demonstra inegalitatea $\rho(X,Y)\geq -1$, procedăm conform *Indicaţiei* 1:

$$\begin{aligned} Var\Big(\frac{X}{\sigma_X}+\frac{Y}{\sigma_Y}\Big) &= Var\Big(\frac{X}{\sigma_X}\Big)+Var\Big(\frac{Y}{\sigma_Y}\Big)+2\,Cov\Big(\frac{X}{\sigma_X},\frac{Y}{\sigma_Y}\Big)\\ &= \frac{1}{\sigma_X^2}Var(X)+\frac{1}{\sigma_Y^2}Var(Y)+2\frac{1}{\sigma_X}\frac{1}{\sigma_Y}Cov(X,Y)\\ &= 1+1+2\rho(X,Y)=2[1+\rho(X,Y)]. \end{aligned}$$

Întrucât $Var(X)\geq 0$ pentru orice variabilă aleatoare X, rezultă că $Var\Big(\frac{X}{\sigma_X}+\frac{Y}{\sigma_Y}\Big)\geq 0$, deci $1+\rho(X,Y)\geq 0$, adică $\rho(X,Y)\geq -1$.

În mod similar, putem să arătăm că

$$Var\Big(\frac{X}{\sigma_X}-\frac{Y}{\sigma_Y}\Big)=2[1-\rho(X,Y)]\geq 0,$$

deci $\rho(X,Y)\leq 1$.

b. Dacă $\rho(X,Y)=-1$, atunci din primul calcul de la punctul a va rezulta că $Var\Big(\frac{X}{\sigma_X}+\frac{Y}{\sigma_Y}\Big)=0$. Ştim că aceasta se întâmplă dacă $\frac{X}{\sigma_X}+\frac{Y}{\sigma_Y}$ este o variabilă aleatoare constantă, adică, mai precis, există $a'\in\mathbb{R}$ astfel încât $\frac{X}{\sigma_X}+\frac{Y}{\sigma_Y}=a'$ cu probabilitate 1. Prin urmare, putem scrie $Y=a'\sigma_Y-\frac{\sigma_Y}{\sigma_X}X$. Rezultă că $Y=aX+b$, unde $a=-\frac{\sigma_Y}{\sigma_X}<0$ şi $b=a'\sigma_Y$.

[18] Aşadar, coeficientul de corelaţie reprezintă o „măsură“ a gradului de „dependenţă liniară“ dintre X şi Y. LC: Coeficientul a din relaţia $Y=aX+b$ nu poate lua valori în afara intervalului $[-1,1]$ — aşa cum ne-am aştepta dacă facem legătura cu ecuaţia unei drepte oarecare din planul euclidian — din cauza simetriei $Cov\,(X,Y)=Cov\,(Y,X)$.

[19] Vedeţi rezolvarea problemei 21.c.

[20] Vedeţi *Consecinţa* de la problema 9.c.

[21] Mai precis, există $c\in\mathbb{R}$ astfel încât $P(X=c)=1$, unde P este distribuţia de probabilitate considerată la definirea variabilelor din enunţul problemei.

În mod similar, dacă $\rho(X,Y)=1$, atunci din al doilea calcul de la punctul a va rezulta că $Var\Big(\frac{X}{\sigma_X}-\frac{Y}{\sigma_Y}\Big)=0$, deci există $a''\in\mathbb{R}$ astfel încât $\frac{X}{\sigma_X}-\frac{Y}{\sigma_Y}=a''$ cu probabilitate 1. Aşadar, $Y=-a''\sigma_Y+\frac{\sigma_Y}{\sigma_X}X$. Renotând, obţinem $Y=aX+b$, cu $a=\frac{\sigma_Y}{\sigma_X}>0$ şi $b=-a''\sigma_Y$.

20. **(O proprietate: matricea de covarianţă a oricărui vector de variabile aleatoare este simetrică şi pozitiv semidefinită)**

■ □ ***prelucrare de Liviu Ciortuz, după "The Multivariate Gaussian Distribution", Chuong B. Do, 2008***

Fie variabilele aleatoare $X_1,\ldots,X_n$, cu $X_i:\Omega\to\mathbb{R}$ pentru $i=1,\ldots,n$. *Matricea de covarianţă* a *vectorului de variabile aleatoare* $X=(X_1,\ldots,X_n)$ este o matrice pătratică de dimensiune $n\times n$, ale cărei elemente se definesc astfel: $[Cov(X)]_{ij}\overset{def.}{=}Cov(X_i,X_j)$, pentru orice $i,j\in\{1,\ldots,n\}$.

Arătaţi că $\Sigma\overset{not.}{=}Cov(X)$ este matrice simetrică şi pozitiv semidefinită, cea de-a doua proprietate însemnând că pentru orice vector $z\in\mathbb{R}^n$ are loc inegalitatea $z^\top\Sigma z\geq 0$. (Vectorii $z\in\mathbb{R}^n$ sunt consideraţi vectori-coloană, iar simbolul $\top$ reprezintă operaţia de transpunere a matricelor.)

Răspuns:

Faptul că matricea Σ este simetrică decurge imediat din definiţia ei: dacă $X=(X_1,\ldots,X_n)$, atunci $[Cov(X)]_{i,j}\overset{def.}{=}Cov(X_i,X_j)=E[(X_i-E[X_i])(X_j-E[X_j])]=E[(X_j-E[X_j])(X_i-E[X_i])]=Cov(X_j,X_i)=[Cov(X)]_{j,i}$, pentru orice $i,j\in\{1,\ldots,n\}$.

Apoi, pentru orice vector $z\in\mathbb{R}^n$ de forma $z=(z_1,\ldots,z_n)^\top$, avem:

$$\begin{aligned}
z^\top\Sigma z &= \sum_{i=1}^n z_i(\sum_{j=1}^n \Sigma_{ij}z_j)=\sum_{i=1}^n\sum_{j=1}^n(z_i\Sigma_{ij}z_j)=\sum_{i=1}^n\sum_{j=1}^n(z_i\,Cov[X_i,X_j]\,z_j)\\
&= \sum_{i=1}^n\sum_{j=1}^n(z_i\,E[(X_i-E[X_i])(X_j-E[X_j])]\,z_j)\\
&= E\left[\sum_{i=1}^n\sum_{j=1}^n z_i\,(X_i-E[X_i])(X_j-E[X_j])\,z_j\right]
\end{aligned}$$

Ultima dintre egalităţile de mai sus derivă din proprietatea de liniaritate a mediilor. Mai departe,

$$\begin{aligned}
z^\top\Sigma z &= E\Big[\Big(\sum_{i=1}^n z_i\,(X_i-E[X_i])\Big)\Big(\sum_{j=1}^n(X_j-E[X_j])\,z_j\Big)\Big]\\
&= E\Big[\Big(\sum_{i=1}^n(X_i-E[X_i])\,z_i\Big)\Big(\sum_{j=1}^n(X_j-E[X_j])\,z_j\Big)\Big]
\end{aligned}$$

Pentru a finaliza demonstraţia, să mai observăm că ultima expresie obţinută mai sus se scrie sub formă vectorială astfel:

$$E[((X - E[X])^\top \cdot z)^2],$$

ceea ce evident, reprezintă o cantitate nenegativă. Aşadar, $z^\top \Sigma z \geq 0$**.**

21. **(Variabile aleatoare: Adevărat sau Fals?)**

CMU, 2006 fall, E. Xing, T. Mitchell, final exam, pr. 1.b
CMU, 2008 fall, Eric Xing, midterm exam, pr. 1.1, 1.2, 1.3

a. Dacă o variabilă aleatoare continuă X **are funcţia densitate de probabilitate** p **diferită de zero pe tot domeniul de definiţie, atunci probabilitatea ca** X **să ia o valoare oarecare** x **(notaţie:** $P(X = x)$**) este egală cu** $p(x)$**.**

b. $E[X + Y] = E[X] + E[Y]$ **pentru orice două variabile aleatoare** X **şi** Y**.**

c. $Var[X + Y] = Var[X] + Var[Y]$ **pentru orice două variabile aleatoare** X **şi** Y**.**

d. $E[XY] = E[X] \cdot E[Y]$ **pentru orice două variabile aleatoare** X **şi** Y**.**

Răspuns:

a. Fals (în general). Dacă variabila aleatoare continuă X **are funcţia densitate de probabilitate** p**, atunci** $P(a \leq X \leq b) = \int_a^b p(x)dx$**. Prin urmare,** $P(X = x) = 0$ **pentru orice** $x \in \mathbb{R}$**. Enunţul afirmă pe de o parte că** $p(x) = P(X = x) = 0$ **pentru un anume** $x \in \mathbb{R}$**, iar pe de altă parte că** p **este diferită de zero pe tot domeniul de definiţie, ceea ce este absurd.**

b. Adevărat. Demonstraţia este făcută în problema 9 punctul a**.**

c. Fals (în general). Se poate considera situaţia $Y = -X$**, caz în care** $Var[X + Y] = 0$**, dar** $Var[X] + Var[Y] = E[X^2] - (E[X])^2 + E[(-X)^2] - (E[-X])^2 = 2\,Var[X]$**. Aşadar, pentru orice variabilă aleatoare** X **cu** $Var[X] \neq 0$**, luând** $Y = -X$**, rezultă că** $Var[X + Y] = 0$ **iar** $Var[X] + Var[Y] \neq 0$**. Un exemplu de variabilă aleatoare cu varianţa nenulă este distribuţia gaussiană.**

Mai *general*, se poate demonstra că $Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y]$ **astfel:**

$$\begin{aligned}
Var[X + Y] &= E[(X + Y)^2] - (E[X + Y])^2 \\
&= E[X^2 + 2XY + Y^2] - (E[X] + E[Y])^2 \\
&= E[X^2] + 2E[XY] + E[Y^2] - (E[X])^2 - 2E[X] \cdot E[Y] - (E[Y])^2 \\
&= (E[X^2] - (E[X])^2) + (E[Y^2] - (E[Y])^2) + (2E[XY] - 2E[X] \cdot E[Y]) \\
&= Var[X] + Var[Y] + 2Cov[X, Y]
\end{aligned}$$

Prin urmare, egalitatea din enunţ este adevărată pentru orice două variabile aleatoare X **şi** Y **pentru care** $Cov[X, Y] = 0$ **(vedeţi problema 9.c), dar este falsă în rest. Egalitatea** $Cov[X, Y] = 0$ **este adevărată de exemplu atunci când** X **şi** Y **sunt variabile independente (vedeţi problema 10).**

d. Fals, în general. Afirmaţia din enunţ este echivalentă cu $Cov[X, Y] = 0$**. Aşa cum am menţionat la punctul** c**, ea este adevărată dacă, de exemplu** X

şi Y sunt variabile aleatoare independente. Dacă, în schimb, vom considera de pildă cazul $Y = X$, cu X variabilă aleatoare binară care ia valoarea 1 cu probabilitatea p şi valoarea 0 cu probabilitatea $1 - p$, se poate deduce imediat că $E[X^2] \neq (E[X])^2$, fiindcă $p(1-p) \neq 0$ pentru orice $p \in (0, 1)$.

0.1.3 Distribuţii probabiliste uzuale

22. **(Variabile aleatoare discrete – distribuţia Bernoulli – şi evenimente aleatoare identic distribuite: calcul de medii)**

• ***CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW1, pr. 2.2***

Un iepuraş se joacă „de-a săritelea". Poziţia lui iniţială coincide cu originea axei reale. După aceea, iepuraşul face câte un salt de-a lungul axei, fie la stânga fie la dreapta. Pentru a determina în ce direcţie să sară, iepuraşul dă cu banul. Dacă obţine stema, va sări spre dreapta, iar dacă obţine banul, va sări spre stânga. Probabilitatea de a obţine stema este p. Se presupune că toate salturile iepuraşului au aceeaşi lungime, şi anume 1.

Care va fi poziţia (medie) la care ne aşteptăm să fie iepuraşul după ce face n salturi?

Răspuns:

Fiecare săritură a iepuraşului este modelată de o variabilă aleatoare. Un salt spre dreapta înseamnă o deplasare cu $+1$ pe axă, iar un salt spre stânga -1. Să notăm cu X_i variabila aleatoare corespunzătoare săriturii i. Aceasta este:

$$X_i : \begin{pmatrix} -1 & 1 \\ 1-p & p \end{pmatrix}$$

Media acestei variabile aleatoare este $E[X_i] = -1 \cdot (1-p) + 1 \cdot p = 2p - 1$.

Ţinând cont de proprietatea de liniaritate a mediilor (vedeţi pr. 9.a), poziţia iepuraşului după n salturi este de aşteptat să fie:

$$E[X_1 + X_2 + \ldots + X_n] = E[X_1] + E[X_2] + \ldots + E[X_n] = n(2p - 1).$$

De exemplu, pentru $p = 1/2$, se va obţine poziţia 0 pe axa reală (aşa cum este de aşteptat dacă n este număr par), în vreme ce pentru $p = 2/3$ va rezulta poziţia $n/3$ (dacă $n/3 \in \mathbb{N}$), iar pentru $p = 1/3$ poziţia $-n/3$ (similar).

23. **(Distribuţia binomială: verificarea condiţiilor de definiţie pentru p.m.f.; calculul mediei şi al varianţei)**

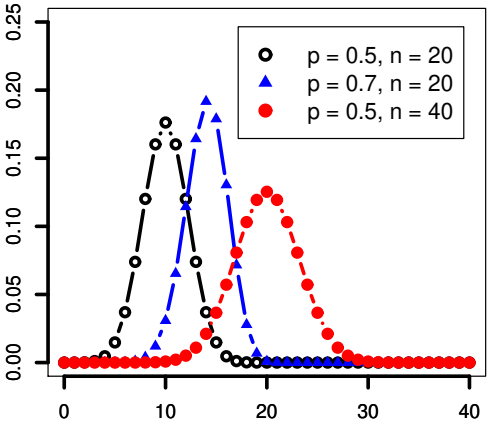
■ □ ***Liviu Ciortuz, 2015***

Distribuţia binomială de parametri n şi p are funcţia masă de probabilitate (engl., probability mass function, p.m.f.) definită astfel:

$$b(r;n,p) = C_n^r p^r (1-p)^{n-r} \ \forall r \in \{0,\ldots,n\}.$$

Distribuţia binomială: p.m.f.

Vă reamintim că $b(r;n,p)$ este probabilitatea care corespunde numărului (r) de apariţii ale feţei cu stema (corespondent engl., *head*) obţinute la efectuarea a n aruncări independente ale unei monede, atunci când se presupune că probabilitatea de apariţie a stemei la o aruncare oarecare a acestei monede este p.



a. Verificaţi că funcţia $b(r;n,p)$, aşa cum a fost definită mai sus, reprezintă într-adevăr o funcţie masă de probabilitate. Aceasta revine la a arăta că $b(r;n,p) \geq 0$ pentru orice $p \in [0,1]$, $n \in \mathbb{N}$ şi $r \in \{0,1,\ldots,n\}$, iar $\sum_{r=0}^{n} b(r;n,p) = 1$ pentru orice astfel de n şi p, fixaţi.

b. Calculaţi media şi varianţa distribuţiei binomiale.

Răspuns:

a. Evident, $b(r;n,p) \stackrel{def.}{=} C_n^r p^r (1-p)^{n-r} \geq 0$ **pentru orice** $p \in [0,1]$, $n \in \mathbb{N}$ **şi** $r \in \{0,1,\ldots,n\}$**, iar**

$$\begin{aligned}\sum_{r=0}^{n} b(r;n,p) &= (1-p)^n + C_n^1 p(1-p)^{n-1} + \cdots + C_n^{n-1} p^{n-1}(1-p) + p^n \\ &= [p+(1-p)]^n = 1\end{aligned}$$

b. Calculul mediei se poate face pornind de la definiţie:

$$\begin{aligned}E[b(r;n,p)] &\stackrel{def.}{=} \sum_{r=0}^{n} r \cdot b(r;n,p) = \\ &= 1 \cdot C_n^1 p(1-p)^{n-1} + 2 \cdot C_n^2 p^2 (1-p)^{n-2} + \cdots + (n-1) \cdot C_n^{n-1} p^{n-1}(1-p) + n \cdot p^n \\ &= p\left[C_n^1 (1-p)^{n-1} + 2 \cdot C_n^2 p(1-p)^{n-2} + \cdots + (n-1) \cdot C_n^{n-1} p^{n-2}(1-p) + n \cdot p^{n-1}\right] \\ &= np\left[(1-p)^{n-1} + C_{n-1}^1 p(1-p)^{n-2} + \cdots + C_{n-1}^{n-2} p^{n-2}(1-p) + C_{n-1}^{n-1} p^{n-1}\right] \quad (8) \\ &= np[p+(1-p)]^{n-1} = np. \quad (9)\end{aligned}$$

Pentru egalitatea (8) am folosit faptul că

$$\begin{aligned}k\, C_n^k &= k\frac{n!}{k!\,(n-k)!} = \frac{n!}{(k-1)!\,(n-k)!} = \frac{n\,(n-1)!}{(k-1)!\,(n-1-(k-1))!} \\ &= n\, C_{n-1}^{k-1}, \forall k = 1,\ldots,n.\end{aligned}$$

Pentru calculul varianţei, vom folosi formula $Var[X] = E[X^2] - E^2[X]$**, care a fost demonstrată la problema 9.b. Întrucât am calculat deja** $E[b(r;n,p)]$**, rămâne să calculăm** $E[b^2(r;n,p)]$**.**[22] **Notând** $q = 1-p$**, vom avea:**

$$\begin{aligned} E[b^2(r;n,p)] &\overset{def.}{=} \sum_{r=0}^{n} r^2 C_n^r p^r q^{n-r} = \sum_{r=0}^{n} r^2 \frac{n(n-1)\ldots(n-r+1)}{r!} p^r q^{n-r} \\ &= \sum_{r=1}^{n} rn \frac{(n-1)\ldots(n-r+1)}{(r-1)!} p^r q^{n-r} = \sum_{r=1}^{n} rn\, C_{n-1}^{r-1}\, p^r q^{n-r} \\ &= np \sum_{r=1}^{n} r\, C_{n-1}^{r-1}\, p^{r-1} q^{(n-1)-(r-1)}. \end{aligned}$$

Mai departe, notând pentru convenienţă $r-1$ **cu** j**, urmează:**

$$\begin{aligned} E[b^2(r;n,p)] &= np \sum_{j=0}^{n-1} (j+1)\, C_{n-1}^{j}\, p^j q^{(n-1)-j} \\ &= np\Big[\sum_{j=0}^{n-1} j\, C_{n-1}^{j}\, p^j q^{(n-1)-j} + \sum_{j=0}^{n-1} C_{n-1}^{j}\, p^j q^{(n-1)-j} \Big]. \end{aligned}$$

Prima sumă din paranteza pătrată de mai sus este chiar $E[b^2(r;n-1,p)]$**, conform relaţiei** (9)**, iar cea de-a doua sumă este egală cu** 1**, conform unui calcul absolut similar cu cel de la punctul** a**. Prin urmare,**

$$E[b^2(r;n,p)] = np[(n-1)p+1] = n^2p^2 - np^2 + np.$$

Aşadar, $Var[X] = E[b^2(r;n,p)] - (E[b(r;n,p)])^2 = n^2p^2 - np^2 + np - n^2p^2 = np(1-p)$**.**

***Observaţie*: O altă cale de a calcula varianţa distribuţiei binomiale este următoarea:**

– se demonstrează relativ uşor că orice variabilă aleatoare urmând distribuţia binomială $b(r;n,p)$ **poate fi văzută ca o sumă de** n **variabile independente care urmează distribuţia Bernoulli de parametru** p**;**[23]

– se ştie (sau, se poate dovedi imediat) că varianţa distribuţiei Bernoulli de parametru p **este** $p(1-p)$**;**

– ţinând cont de proprietatea $Var[X_1+X_2\ldots+X_n] = Var[X_1]+Var[X_2]\ldots+Var[X_n]$ **atunci când** $X_1, X_2, \ldots, X_n$ **sunt variabile independente, conform demonstraţiei de la problema 21.c, rezultă că** $Var[X] = np(1-p)$**.**

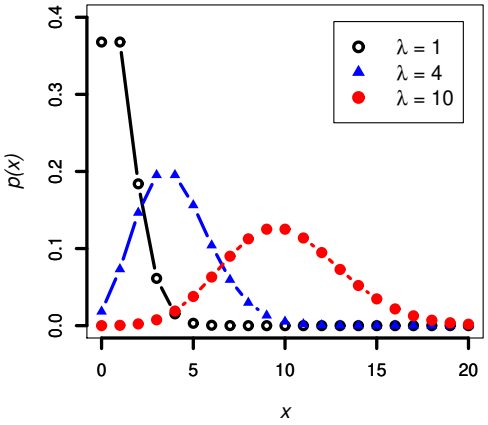
[22] Rezolvarea de mai jos urmează îndeaproape linia demonstraţiei găsite pe site-ul www.proofwiki.org/wiki/Variance_of_Binomial_Distribution, accesat la data de 5 octombrie 2015. La rândul său, acest site menţionează ca sursă "Probability: An Introduction" de Geoffrey Grimmett şi Dominic Welsh, Oxford Science Publications, 1986.

[23] Vedeţi www.proofwiki.org/wiki/Bernoulli_Process_as_Binomial_Distribution, care citează ca sursă "Probability: An Introduction" de Geoffrey Grimmett şi Dominic Welsh, Oxford Science Publications, 1986.

24. **(Distribuţia Poisson: verificarea condiţiilor de definiţie pentru p.m.f.; calculul mediei şi al varianţei)**

Liviu Ciortuz, 2017

Distribuţia Poisson: pmf



Distribuţia *Poisson* este o distribuţie discretă de parametru $\lambda > 0$, a cărei funcţie masă de probabilitate este dată de expresia

$$p(x \mid \lambda) = \frac{1}{e^\lambda} \cdot \frac{\lambda^x}{x!}, \text{ pentru orice } x \in \mathbb{N}.$$

Prin convenţie, se consideră că $0! = 1$. Factorul $\frac{1}{e^\lambda}$, care nu depinde de x, este aşa-numitul *factor de normalizare*.

Demonstraţi mai întâi că funcţia $p(\ )$ este într-adevăr funcţie masă de probabilitate (engl., probability mass function, p.m.f.) şi apoi că media acestei distribuţii este λ, iar varianţa ei este tot λ.

***Sugestie*: Ţineţi cont că $\sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^\lambda$ (limită fundamentală).**

Răspuns:

Vom arăta mai întâi că $p(\ )$ este într-adevăr o funcţie masă de probabilitate (p.m.f). Evident, $p(x|\lambda) > 0$ pentru orice $x \in \mathbb{N}$, fiindcă $\lambda > 0$. Apoi,

$$\sum_{x \in \mathbb{N}} \frac{1}{e^\lambda} \frac{\lambda^x}{x!} = \frac{1}{e^\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = \frac{1}{e^\lambda} e^\lambda = 1,$$

ţinând cont de limita fundamentală care apare în *Sugestia* din enunţ.

Pentru calcularea mediei acestei distribuţii, aplicăm definiţia:

$$\sum_{x=0}^{\infty} x \frac{1}{e^\lambda} \frac{\lambda^x}{x!} = \frac{1}{e^\lambda} \sum_{x=0}^{\infty} x \frac{\lambda^x}{x!} = \frac{1}{e^\lambda} \sum_{x=1}^{\infty} \frac{\lambda^x}{(x-1)!} = \frac{\lambda}{e^\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} = \frac{\lambda}{e^\lambda} \underbrace{\sum_{x=0}^{\infty} \frac{\lambda^x}{x!}}_{e^\lambda} = \frac{\lambda}{e^\lambda} e^\lambda = \lambda.$$

În ce priveşte calculul varianţei pentru distribuţia Poisson, ştim că $Var[X] = E[X^2] - E^2[X]$ (proprietate demonstrată la problema 9.b) pentru orice distribuţie aleatoare X şi, prin urmare, pentru că am calculat deja media acestei distribuţii, trebuie să mai calculăm $\sum_{x=0}^{\infty} x^2 p(x|\lambda)$.

$$\begin{aligned}
\sum_{x=0}^{\infty} x^2 p(x|\lambda) &= \sum_{x=0}^{\infty} x^2 \frac{1}{e^\lambda} \frac{\lambda^x}{x!} = \frac{1}{e^\lambda} \sum_{x=1}^{\infty} x^2 \frac{\lambda^x}{x!} = \frac{1}{e^\lambda} \left[\sum_{x=1}^{\infty} [x(x-1) + x] \frac{\lambda^x}{x!} \right] \\
&= \frac{1}{e^\lambda} \left[\sum_{x=1}^{\infty} x(x-1) \frac{\lambda^x}{x!} + \sum_{x=1}^{\infty} x \frac{\lambda^x}{x!} \right] = \frac{1}{e^\lambda} \left[\sum_{x=2}^{\infty} \frac{\lambda^x}{(x-2)!} + \sum_{x=1}^{\infty} \frac{\lambda^x}{(x-1)!} \right] \\
&= \frac{1}{e^\lambda} \left[\lambda^2 \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} + \lambda \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \right] = \frac{1}{e^\lambda} \left[\lambda^2 \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} + \lambda \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \right]
\end{aligned}$$

$$= \frac{1}{e^\lambda}\left[\lambda^2 e^\lambda + \lambda e^\lambda\right] = \frac{1}{e^\lambda} e^\lambda(\lambda^2+\lambda) = \lambda^2 + \lambda.$$

Prin urmare, varianţa distribuţiei Poisson este $\lambda^2 + \lambda - \lambda^2 = \lambda$.

25. **(Distribuţia geometrică: numărul „aşteptat“ / mediu de „observaţii“ necesare pentru ca un anumit eveniment să se producă)**

□ • ○ ***CMU, 2012 spring, Ziv Bar-Joseph, HW1, pr. 1.4***

În cazul unui zar perfect cu şase feţe, probabilităţile de apariţie pentru fiecare dintre feţele zarului sunt egale.

Mickey se duce la un cazinou şi doreşte ca, folosindu-şi cunoştinţele din domeniul probabilităţilor, să-şi evalueze şansa pe care o are de a obţine la aruncarea unui astfel de zar faţa 6.

Mai precis, Mickey se întreabă care este numărul mediu (sau, numărul „aşteptat“; engl., expected number) de aruncări ale zarului pe care ar trebui să le efectueze până să obţină faţa 6.

Justificaţi răspunsul în detaliu.

Răspuns:

Experimentul lui Mickey poate fi modelat cu ajutorul *distribuţiei geometrice*.[24] **Ştim că distribuţia geometrică poate fi reprezentată de o tabelă de „repartiţie“ de forma**

$$\begin{pmatrix} 1 & 2 & 3 & \dots & n & \dots \\ q & pq & p^2q & \dots & p^{n-1}q & \dots \end{pmatrix},$$

unde $p, q \geq 0$ şi $p + q = 1$. Se verifică imediat că $\sum_{i=1}^{\infty} p^{i-1}q = 1$ atunci când $p \in [0, 1)$. În particular, pentru $p = \frac{5}{6}$ şi $q = \frac{1}{6}$, avem

$$\sum_{i=1}^{\infty} \frac{1}{6} \cdot \left(\frac{5}{6}\right)^{i-1} = 1. \tag{10}$$

Vom nota cu E numărul mediu de aruncări ale zarului pe care ar trebui să le efectueze Mickey până să obţină faţa 6.

Conform definiţiei mediei, E se poate exprima astfel:

$$E = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} \cdot \frac{5}{6} + 3 \cdot \frac{1}{6} \cdot \left(\frac{5}{6}\right)^2 \dots + n \cdot \frac{1}{6} \cdot \left(\frac{5}{6}\right)^{n-1} \dots + \dots.$$

Punând termenul generic $n \cdot \frac{1}{6} \cdot \left(\frac{5}{6}\right)^{n-1}$ sub forma $(1 + (n-1)) \cdot \frac{1}{6} \cdot \left(\frac{5}{6}\right)^{n-1}$ şi aplicând apoi distributivitatea înmulţirii faţă de adunare, putem rescrie E în mod echivalent astfel:

$$E = \frac{1}{6} + \frac{5}{6} \cdot \left[\frac{1}{6} + \frac{1}{6} \cdot \frac{5}{6} + \dots + \frac{1}{6} \cdot \left(\frac{5}{6}\right)^{n-1} + \dots\right]$$

[24] Simplu spus, distribuţia gemetrică poate fi gândită ca modelând următorul experiment aleatoriu: Fie o monedă a cărei probabilitate de apariţie a feţei-stemă este p. Aruncăm moneda o dată sau de mai multe ori, până când apare stema. Notăm numărul de aruncări care au precedat apariţia stemei cu k. Acest număr $k \in \{0, 1, \dots\}$ va fi [asociat cu] valoarea unei variabile aleatoare X, despre care spunem că urmează distribuţia geometrică. Evident, $P(X = k) = (1-p)^k p$.

$$+\frac{5}{6}\cdot\left[1\cdot\frac{1}{6}+2\cdot\frac{1}{6}\cdot\frac{5}{6}+\ldots+(n-1)\cdot\frac{1}{6}\cdot\left(\frac{5}{6}\right)^{n-1}+\ldots\right].$$

Ţinând cont de relaţia (10)**, egalitatea precedentă se rescrie astfel:**

$$E=\frac{1}{6}+\frac{5}{6}\cdot 1+\frac{5}{6}\cdot E.$$

Rezultă imediat că $\frac{1}{6}E=\frac{1}{6}+\frac{5}{6}$**, deci** $E=6$**.**

26. **(O mixtură de distribuţii categoriale: calculul mediei şi al varianţei)**

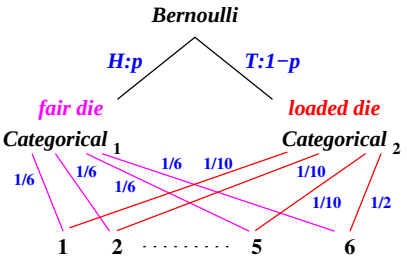
■ □ • ○ ***CMU, 2010 fall, Aarti Singh, HW1, pr. 2.2.1-2***

Presupunem că avem două zaruri cu şase feţe. Un zar este perfect — deci vom considera funcţia sa masă de probabilitate (p.m.f.) definită prin $P_1(x)=1/6$ **pentru** $x=1,\ldots,6$ **—, iar celălalt zar este măsluit şi are următoarea funcţie masă de probabilitate:**

$$P_2(x)=\begin{cases}\frac{1}{2} & \text{pentru } x=6;\\ \frac{1}{10} & \text{pentru } x\in\{1,2,3,4,5\}.\end{cases}$$

Pentru a decide ce zar să aruncăm, vom folosi o monedă; considerăm că probabilitatea să obţinem stema la aruncarea monedei este $p\in(0,1)$**. Dacă obţinem stema (engl., head), vom arunca apoi zarul perfect; în caz contrar vom arunca zarul măsluit.**

Putem reprezenta grafic mixtura formată din cele două distribuţii categoriale ca în figura alăturată.



a. Calculaţi în funcţie de p **media variabilei aleatoare (notată cu** X**) care reprezintă „mixtura“ de distribuţii [categoriale] descrisă mai sus.**

b. Calculaţi în funcţie de p **varianţa variabilei aleatoare** X**.**

Răspuns:

a. Ţinând cont de modul în care a fost definită în enunţ mixtura celor două distribuţii categoriale, putem calcula media variabilei X**, care reprezintă această mixtură, în felul următor:**

$$\begin{aligned}E[X] &= \sum_{i=1}^{6} i\cdot P(i)=\sum_{i=1}^{6} i\cdot[P(i|fair)\cdot p+P(i|loaded)\cdot(1-p)]\\ &= \sum_{i=1}^{6} i\cdot[P_1(i)\cdot p+P_2(i)\cdot(1-p)]=\Big[\sum_{i=1}^{6} i\cdot P_1(i)\Big]p+\Big[\sum_{i=1}^{6} i\cdot P_2(i)\Big](1-p)\end{aligned}$$

b. Considerăm familia mixturilor de două distribuţii gaussiene unidimensionale:

$$f(x|\theta,\mu_1,\sigma_1,\mu_2,\sigma_2) = \theta\,\mathcal{N}(x|\mu_1,\sigma_1) + (1-\theta)\,\mathcal{N}(x|\mu_2,\sigma_2),$$

unde $0<\theta<1,\ 0<\sigma_1,\ 0<\sigma_2$, iar μ_1,μ_2 sunt numere reale oarecare, distincte. Este imediat că pentru orice estimări bine-definite $\hat\theta, \hat\mu_1, \hat\mu_2, \hat\sigma_1$ şi $\hat\sigma_2$ obţinute pe un set oarecare de instanţe $x_1,\ldots,x_n$, următorul set de estimări

$$\hat\theta' = 1-\hat\theta,\ \hat\mu_1' = \hat\mu_2,\ \hat\mu_2' = \hat\mu_1,\ \hat\sigma_1' = \hat\sigma_2,\ \hat\sigma_2' = \hat\sigma_1$$

vor produce întotdeauna aceeaşi verosimilitate ca şi estimările iniţiale. Şi totuşi $(\hat\theta', \hat\mu_1', \hat\mu_2', \hat\sigma_1', \hat\sigma_2') \neq (\hat\theta, \hat\mu_1, \hat\mu_2, \hat\sigma_1, \hat\sigma_2)$. Aşadar, estimările în sensul MLE nu sunt în mod necesar unice.

c. În exemplul de la punctul a, domeniul de de existenţă pentru parametrul distribuţiei pe care am ales-o este, ca şi în multe alte cazuri, un interval deschis. Însă valoarea acestui parametru pentru care se atinge maximul funcţiei de log-verosimilitate este situată la „marginea" [inferioară a] intervalului, nu în interiorul intervalului respectiv. Această problemă apare adeseori atunci când mărimea eşantionului (adică numărul instanţelor x_i) este mică în comparaţie cu numărul parametrilor care trebuie estimaţi.

În exemplul de la punctul b, funcţia de log-verosimilitate nu este concavă — pentru exemplificare, vedeţi problema 17 de la capitolul de *Clusterizare*, în special graficele care reprezintă curbele de izocontur ale funcţiei de log-verosimilitate ale mixturilor de gaussiene, la pag. 777 şi pag. 778 —, ceea ce implică faptul că pot exista mai multe puncte de maxim local.

0.1.5 Elemente de teoria informaţiei[92]

50. **(Exerciţiu cu caracter teoretic: proprietăţi dezirabile ale entropiei)**

■ □ • ***CMU, 2005 fall, T. Mitchell, A. Moore, HW1, pr. 2.2***

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*:[93]

[92] Observaţie importantă: În toate problemele care urmează, referitor la entropie / teoria informaţiei se va considera în mod implicit că notaţia 'log' desemnează logaritmul în baza 2. De asemenea, prin convenţie, se va considera $p \cdot \log p = 0$ pentru $p = 0$.

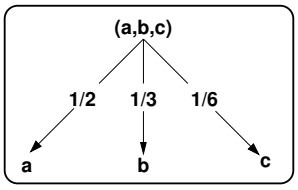
[93] LC: Deşi nu se specifică în enunţul original al problemei, este necesar / natural să considerăm şi proprietatea următoare: [A0.] $\psi_n(p_1,\ldots,p_n) \geq 0$ pentru orice $n \in \mathbb{N}^*$ şi orice $p_1,\ldots,p_n \in [0,1]$ astfel încât $\sum_i p_i = 1$, pentru că ψ_n este văzută că măsură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

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.

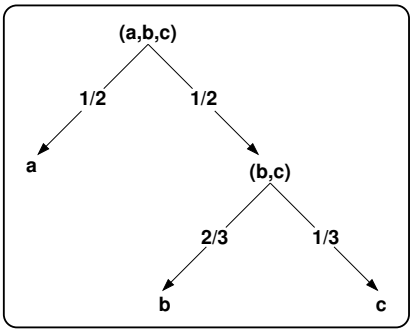
A2. Funcţia $\psi_n(1/n, \ldots, 1/n)$ este monoton *crescătoare* în raport cu n.
Altfel spus, dacă toate evenimentele sunt echiprobabile, atunci entropia creşte odată cu numărul de evenimente posibile.

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 $\psi_2(1/2, 1/2) + 1/2 \cdot \psi_2(2/3, 1/3)$.



Encoding 1



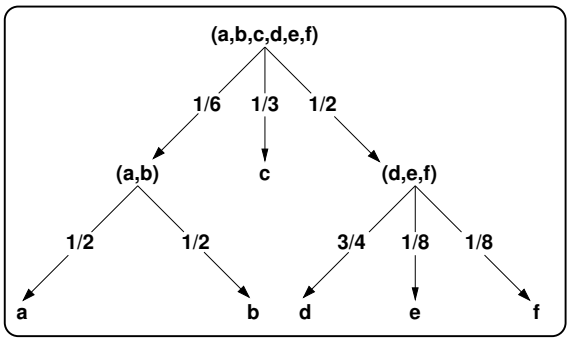
Encoding 2

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, vom vedea, K depinde de $\psi_s\left(\frac{1}{s}, \ldots, \frac{1}{s}\right)$ pentru un anumit $s \in \mathbb{S}^*$.

***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 $A(n) \stackrel{not.}{=} \psi(1/n, 1/n, \ldots, 1/n)$**.**

c. Arătaţi că

$$A(s^m) = m\,A(s) \text{ pentru orice } s, m \in \mathbb{N}^*. \tag{49}$$

Mai departe, pentru $s, m \in \mathbb{N}^*$ **fixaţi, vom considera** $t, n \in \mathbb{N}^*$ **astfel încât**

$$s^m \leq t^n \leq s^{m+1}. \tag{50}$$

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**

$$|\, \frac{m}{n} - \frac{\log t}{\log s} \,| \leq \frac{1}{n}. \tag{51}$$

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

f. Combinând ultima inegalitate de mai sus cu inegalitatea (49)**, 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**

$$|\, \frac{m}{n} - \frac{A(t)}{A(s)} \,| \leq \frac{1}{n} \text{ pentru } s \neq 1. \tag{52}$$

g. Combinând inegalităţile (51) **şi** (52)**, arătaţi că**

$$|\, \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \,| \leq \frac{2}{n} \text{ pentru } s \neq 1 \tag{53}$$

şi, în consecinţă

$$A(t) = K \log t \text{ cu } K > 0 \text{ (din cauza proprietăţii A2)}. \tag{54}$$

***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** $\psi(p_1, \ldots, p_n)$ **cu** $p_i \in \mathbb{Q}$ **pentru** $i = 1, \ldots, n$ **(cazul** $p_i \notin \mathbb{Q}$ **nu este tratat aici):**

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

Conform egalităţii (54), 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,

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

Răspuns:

a. Facem calculele:

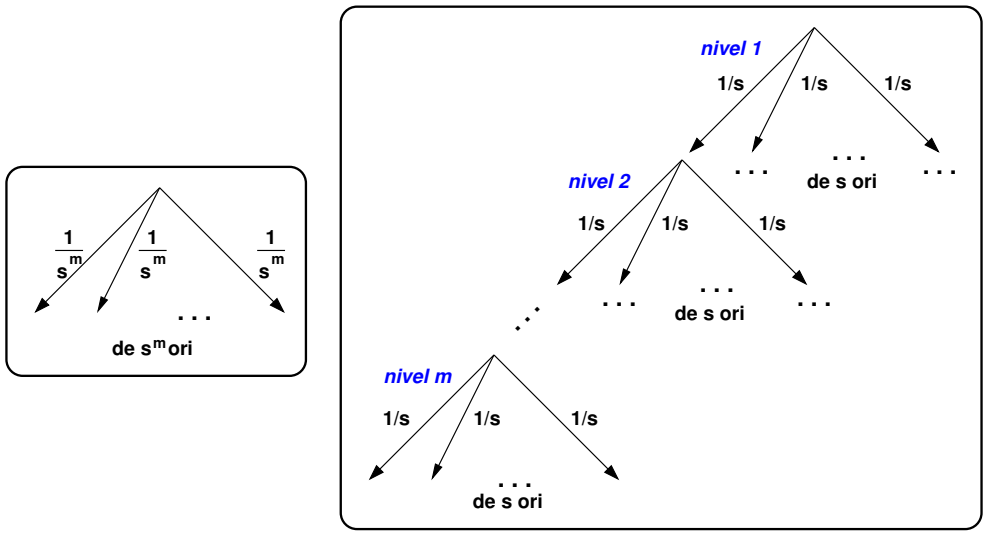
$$\begin{aligned}H\left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right) &= \frac{1}{2}\log 2 + \frac{1}{3}\log 3 + \frac{1}{6}\log 6 = \left(\frac{1}{2} + \frac{1}{6}\right)\log 2 + \left(\frac{1}{3} + \frac{1}{6}\right)\log 3 \\ &= \frac{2}{3} + \frac{1}{2}\log 3 \\ H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2}H\left(\frac{2}{3}, \frac{1}{3}\right) &= 1 + \frac{1}{2}\left(\frac{2}{3}\log\frac{3}{2} + \frac{1}{3}\log 3\right) = 1 + \frac{1}{2}\left(\log 3 - \frac{2}{3}\right) \\ &= \frac{2}{3} + \frac{1}{2}\log 3\end{aligned}$$

şi rezultă că $H\left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right) = H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2}H\left(\frac{2}{3}, \frac{1}{3}\right)$.

b. Folosind proprietatea A3, entropia „codificării“ din figura dată în enunţ este:

$$\begin{aligned}H\left(\frac{1}{6}, \frac{1}{3}, \frac{1}{2}\right) &+ \frac{1}{6}H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2}H\left(\frac{3}{4}, \frac{1}{8}, \frac{1}{8}\right) \\ &= \frac{1}{6}\log 6 + \frac{1}{3}\log 3 + \frac{1}{2}\log 2 + \frac{1}{6} + \frac{1}{2}\left(\frac{3}{4}\log\frac{4}{3} + \frac{2}{8}\log 8\right) \\ &= \frac{1}{6}\log 2 + \frac{1}{6}\log 3 + \frac{1}{3}\log 3 + \frac{1}{2} + \frac{1}{6} + \frac{3}{8}(2 - \log 3) + \frac{3}{8} \\ &= \frac{1}{6} + \frac{1}{2} + \frac{1}{6} + \frac{3}{4} + \frac{3}{8} + \left(\frac{1}{6} + \frac{1}{3} - \frac{3}{8}\right)\log 3 \\ &= \frac{47}{24} + \frac{1}{8}\log 3 = 1.958 + 0.125 \log 3 = 2.156\end{aligned}$$

c. Pentru calculul lui $A(s^m)$ se poate folosi atât o „codificare“ imediată cât şi una (des)compusă, ca în figura următoare:



Aplicând proprietatea A3 pe „codificarea" din figura de mai sus, partea dreaptă, avem:

$$\begin{aligned} A(s^m) &= A(s) + s \cdot \frac{1}{s} A(s) + s^2 \cdot \frac{1}{s^2} A(s) + \ldots + s^{m-1} \cdot \frac{1}{s^{m-1}} A(s) \\ &= \underbrace{A(s) + A(s) + A(s) + \ldots + A(s)}_{\text{de } m \text{ ori}} = mA(s) \end{aligned}$$

d. Aplicând funcţia $\log$ **fiecărui termen al inegalităţii** $s^m \leq t^n \leq s^{m+1}$ **obţinem** $m \log s \leq n \log t \leq (m+1) \log s$**. Apoi, pentru** $s \neq 1$**, împărţind prin** $n \log s$**, rezultă:**

$$\frac{m}{n} \leq \frac{\log t}{\log s} \leq \frac{m}{n} + \frac{1}{n} \Rightarrow 0 \leq \frac{\log t}{\log s} - \frac{m}{n} \leq \frac{1}{n} \Rightarrow \left| \frac{\log t}{\log s} - \frac{m}{n} \right| \leq \frac{1}{n}$$

e. Datorită proprietăţii A2 din enunţ, inegalitatea $s^m \leq t^n \leq s^{m+1}$ **implică**

$$\psi_{s^m}\left(\frac{1}{s^m}, \ldots, \frac{1}{s^m}\right) \leq \psi_{t^n}\left(\frac{1}{t^n}, \ldots, \frac{1}{t^n}\right) \leq \psi_{s^{m+1}}\left(\frac{1}{s^{m+1}}, \ldots, \frac{1}{s^{m+1}}\right)$$

ceea ce reprezintă exact $A(s^m) \leq A(t^n) \leq A(s^{m+1})$**.**

f. Datorită proprietăţii (54)**, dubla inegalitate** $A(s^m) \leq A(t^n) \leq A(s^{m+1})$ **devine** $m\,A(s) \leq n\,A(t) \leq (m+1)\,A(s)$**. Împărţind această inegalitate prin** $nA(s)$**, despre care se poate spune că este nenul pentru orice** $s \neq 1$**,**[94] **rezultă:**

$$\frac{m}{n} \leq \frac{A(t)}{A(s)} \leq \frac{m}{n} + \frac{1}{n} \Rightarrow 0 \leq \frac{A(t)}{A(s)} - \frac{m}{n} \leq \frac{1}{n} \Rightarrow \left| \frac{A(t)}{A(s)} - \frac{m}{n} \right| \leq \frac{1}{n}$$

g. Inegalităţile duble de mai jos rescriu convenabil proprietăţile (51) **şi** (52)**:**

$$-\frac{1}{n} \leq \frac{m}{n} - \frac{\log t}{\log s} \leq \frac{1}{n} \quad \text{şi} \quad -\frac{1}{n} \leq \frac{A(t)}{A(s)} - \frac{m}{n} \leq \frac{1}{n}$$

[94] Putem considera în mod natural $A(1) = 0$. Conform proprietăţii A2, urmează că $A(s) > A(1) = 0$ pentru orice $s > 1$.

Însumându-le membru cu membru, rezultă

$$-\frac{2}{n} \leq \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \leq \frac{2}{n} \Rightarrow \left|\frac{A(t)}{A(s)} - \frac{\log t}{\log s}\right| \leq \frac{2}{n}$$

Din inegalitatea dublă $s^m \leq t^n \leq s^{m+1}$ rezultă că odată cu m creşte şi n (acesta din urmă depinzând de valorile lui s, t şi m).[95] **Aşadar, atunci când m tinde la infinit vom avea şi $n \to \infty$. Dacă trecem la limită inegalitatea $\left| \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \right| \leq \frac{2}{n}$ pentru $n \to \infty$, rezultă $\left| \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \right| \to 0$, de unde avem $\left| \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \right| = 0$ şi deci $\frac{A(t)}{A(s)} = \frac{\log t}{\log s}$. Rezultă că $A(t) = \frac{A(s)}{\log s} \log t = K \log t$. Evident, constanta $K = \frac{A(s)}{\log s} = \frac{1}{\log s}\psi\left(\frac{1}{s}, \ldots, \frac{1}{s}\right)$ nu depinde de t. Variind valorile lui t, rezultă că $A(t) = K \log t$ pentru orice $t \in \mathbb{N}^*$ astfel încât relaţia (50) are loc.**

O *observaţie* finală: din inegalitatea $s^m \leq t^n \leq s^{m+1}$, dacă $s \neq 1$ va rezulta că de fapt şi $t \neq 1$. Atunci când pentru A se foloseşte formula clasică a entropiei, egalitatea $A(t) = K \log t$ se verifică şi în cazul $t = 1$, fiindcă $A(1) = \sum_p p \log p = 1 \log 1 = 0$, iar $K \log 1 = 0$.

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

■ • ***Liviu Ciortuz, pornind de la CMU, 2005 fall, T. Mitchell, A. Moore, HW1, pr. 2***

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

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

[95] Mai precis, este util să observaţi o succesiune de forma $s^m \leq t^{n_1} \leq s^{m+1} \leq t^{n_2} \leq s^{m+2} \leq \ldots$.

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).**

Răspuns:

a. Este uşor să arătăm că $H(X) = -\sum_i P(X = x_i) \log P(X = x_i) \geq 0$.

Ştim că $\log x \leq 0$ **pentru** $\forall x \leq 1$ **şi** $\log x \geq 0$ **pentru** $\forall x \geq 1$. **De asemenea ştim că** $P(X = x_i) \in [0, 1]$ **(fiind o probabilitate). Aşadar,**

$$H(X) = -\sum_i P(X = x_i) \log P(X = x_i) = \sum_i \underbrace{P(X = x_i)}_{\geq 0} \underbrace{\log \frac{1}{P(X = x_i)}}_{\geq 0} \geq 0$$

Pentru a arăta că $H(X) = 0$ **dacă şi numai dacă** X **este constantă vom demonstra că ambele implicaţii au loc:**

„⇒“ Presupunem că $H(X) = 0$, **adică** $\sum_i P(X = x_i) \log \frac{1}{P(X = x_i)} = 0$. **Datorită faptului că fiecare termen din această sumă este mai mare sau egal cu** 0, **rezultă că** $H(X) = 0$ **doar dacă pentru** $\forall i$, $P(X = x_i) = 0$ **sau** $\log \frac{1}{P(X = x_i)} = 0$, **adică dacă pentru** $\forall i$, $P(X = x_i) = 0$ **sau** $P(X = x_i) = 1$. **Cum însă** $\sum_i P(X = x_i) = 1$ **rezultă că există o singură valoare** x_1 **pentru** X **astfel încât** $P(X = x_1) = 1$, **iar** $P(X = x) = 0$ **pentru orice** $x \neq x_1$. **Altfel spus, variabila aleatoare discretă** X **este constantă.**[96]

„⇐“ Presupunem că variabila X **este constantă, ceea ce înseamnă că** X **ia o singură valoare** x_1, **cu probabilitatea** $P(X = x_1) = 1$. **Prin urmare,** $H(X) = -1 \cdot \log 1 = 0$.

b. Pentru a demonstra egalitatea cerută vom porni de la definiţia lui $H(Y \mid X)$ **şi apoi vom efectua câteva transformări elementare:**

$$\begin{aligned}
H(Y \mid X) &= \sum_i P(X = x_i) H(Y \mid X = x_i) \\
&= \sum_i P(X = x_i) \Big[-\sum_j P(Y = y_j \mid X = x_i) \log P(Y = y_j \mid X = x_i) \Big] \\
&= -\sum_i \sum_j \underbrace{P(X = x_i) P(Y = y_j \mid X = x_i)}_{= P(X = x_i, Y = y_j)} \log P(Y = y_j \mid X = x_i) \\
&= -\sum_i \sum_j P(X = x_i, Y = y_j) \log P(Y = y_j \mid X = x_i)
\end{aligned}$$

[96] Mai corect spus, X este constantă pe tot domeniul de definiţie, evenual cu excepţia unei mulţimi de probabilitate 0.

c. În primul rând, trebuie să demonstrăm că

$$H(X,Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y)$$

Din definiţia entropiei comune ştim că $H(X,Y) = -\sum_i \sum_j P(X = x_i, Y = y_j) \log P(X = x_i, Y = y_j)$**. Vom aplica mai întâi regula de multiplicare,** $P(X,Y) = P(X) \cdot P(Y \mid X)$**, după care vom transforma logaritmul produsului în sumă de logaritmi. Pentru claritatea demonstraţiei vom nota prescurtat** $p(x_i) = P(X = x_i)$**,** $p(x_i, y_j) = P(X = x_i, Y = y_j)$ **etc.**

$$\begin{aligned}
H(X,Y) &= -\sum_i \sum_j p(x_i, y_j) \log p(x_i, y_j) \\
&= -\sum_i \sum_j p(x_i) \cdot p(y_j \mid x_i) \log[p(x_i) \cdot p(y_j \mid x_i)] \\
&= -\sum_i \sum_j p(x_i) \cdot p(y_j \mid x_i)[\log p(x_i) + \log p(y_j \mid x_i)] \\
&= -\sum_i \sum_j p(x_i) \cdot p(y_j \mid x_i) \log p(x_i) - \sum_i \sum_j p(x_i) \cdot p(y_j \mid x_i) \log p(y_j \mid x_i) \\
&= -\sum_i p(x_i) \log p(x_i) \cdot \underbrace{\sum_j p(y_j \mid x_i)}_{=1} - \sum_i p(x_i) \sum_j p(y_j \mid x_i) \log p(y_j \mid x_i) \\
&= H(X) + \sum_i p(x_i) H(Y \mid X = x_i) = H(X) + H(Y \mid X)
\end{aligned}$$

Egalitatea $\sum_j p(y_j \mid x_i) = 1$ **se justifică uşor ţinând cont de *proprietatea de aditivitate numărabilă* din definiţia funcţiei / distribuţiei de probabilitate.**

Pentru a demonstra egalitatea $H(X,Y) = H(Y) + H(X \mid Y)$**, se procedează analog, înlocuind** $p(x_i, y_j)$ **nu cu** $p(x_i) \cdot p(y_j \mid x_i)$**, ci cu** $p(y_i) \cdot p(x_j \mid y_i)$**.**

Pentru cazul general

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2 \mid X_1) + \ldots + H(X_n \mid X_1, \ldots, X_{n-1}),$$

vom folosi regula de înlănţuire de la variabile aleatoare

$$P(X_1, \ldots, X_n) = P(X_1) \cdot P(X_2 \mid X_1) \cdot P(X_3 \mid X_1, X_2) \cdot \ldots \cdot P(X_n \mid X_1, \ldots, X_{n-1}),$$

precum şi scrierea entropiei sub formă de medie, $H(X) = E\left[\log \frac{1}{P(X)}\right]$**:**

$$\begin{aligned}
H(X_1, \ldots, X_n) &= E\left[\log \frac{1}{p(x_1, \ldots, x_n)}\right] \\
&= -E_{p(x_1,\ldots,x_n)}[\log \underbrace{p(x_1, \ldots, x_n)}_{p(x_1) \cdot p(x_2|x_1) \cdot \ldots \cdot p(x_n|x_1,\ldots,x_{n-1})}] \\
&= -E_{p(x_1,\ldots,x_n)}[\log p(x_1) + \log p(x_2 \mid x_1) + \ldots + \log p(x_n \mid x_1, \ldots, x_{n-1})] \\
&= -E_{p(x_1)}[\log p(x_1)] - E_{p(x_1,x_2)}[\log p(x_2 \mid x_1)] - \ldots \\
&\quad - E_{p(x_1,\ldots,x_n)}[\log p(x_n \mid x_1, \ldots, x_{n-1})] \\
&\overset{(b)}{=} H(X_1) + H(X_2 \mid X_1) + \ldots + H(X_n \mid X_1, \ldots, X_{n-1})
\end{aligned}$$

La penultima egalitate am ţinut cont de definiţia distribuţiei marginale pornind de la distribuţia comună, iar la ultima egalitate am folosit rezultatul de la punctul b**.**

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

■ □ • ○ ***CMU, 2012 spring, Roni Rosenfeld, HW2, pr. 2***

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*

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

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.

Răspuns:

a. Redăm distribuţia lui S (uşor de calculat) în următorul tabel:

S	2	3	4	5	6	7	8	9	10	11	12
$P(S)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

b. Conform definiţiei date, vom avea:

$$\begin{aligned}
Information(S = 2) &= -\log_2(1/36) = \log_2 36 = 2\log_2 6 = 2(1 + \log_2 3) \\
&= 5.169925001 \text{ biţi} \\
Information(S = 11) &= -\log_2 2/36 = \log_2 18 = 1 + 2\log_2 3 = 4.169925001 \text{ biţi} \\
Information(S = 5) &= -\log_2 4/36 = \log_2 9 = 2\log_2 3 = 3.169925001 \text{ biţi} \\
Information(S = 7) &= -\log_2 6/36 = \log_2 6 = 1 + \log_2 3 = 2.584962501 \text{ biţi}
\end{aligned}$$

c. Conform definiţiei pentru entropie (vedeţi problema 51), $H(S)$ este media ponderată (cu ajutorul probabilităţilor) a „surprizelor“ / cantităţilor de informaţie produse la „observarea“ tuturor valorilor variabilei S. Făcând calculele, vom obţine:

$$\begin{aligned}
H(S) &= -\sum_{i=1}^{n} p_i \log_2 p_i \\
&= -\left(2 \cdot \frac{1}{36} \log_2 \frac{1}{36} + 2 \cdot \frac{2}{36} \log_2 \frac{2}{36} + 2 \cdot \frac{3}{36} \log_2 \frac{3}{36} + 2 \cdot \frac{4}{36} \log_2 \frac{4}{36} + \right. \\
&\quad \left. 2 \cdot \frac{5}{36} \log_2 \frac{5}{36} + \frac{6}{36} \log_2 \frac{6}{36}\right) \\
&= \frac{1}{36}\left(2 \log_2 36 + 4 \log_2 18 + 6 \log_2 12 + 8 \log_2 9 + 10 \log_2 \frac{36}{5} + 6 \log_2 6\right) \\
&= \frac{1}{36}\left(2 \log_2 6^2 + 4 \log_2 6 \cdot 3 + 6 \log_2 6 \cdot 2 + 8 \log_2 3^2 + 10 \log_2 \frac{6^2}{5} + 6 \log_2 6\right) \\
&= \frac{1}{36}(40 \log_2 6 + 20 \log_2 3 + 6 - 10 \log_2 5) \\
&= \frac{1}{36}(60 \log_2 3 + 46 - 10 \log_2 5) = 3.274401919 \text{ biţi.}
\end{aligned}$$

d. Distribuţia variabilei S condiţionată de observarea feţei 4 la prima aruncare este:

S	2	3	4	5	6	7	8	9	10	11	12
$P(S\|...)$	0	0	0	1/6	1/6	1/6	1/6	1/6	1/6	0	0

În consecinţă, folosind definiţia entropiei condiţionale specifice (vedeţi de asemenea problema 51), vom avea:

$$H(S|\textit{First-die-shows-4}) = -6 \cdot \frac{1}{6} \log_2 \frac{1}{6} = \log_2 6 = 2.58 \text{ biţi,}$$

ceea ce înseamnă că se obţine următorul *câştig* de informaţie:

$$IG(S; \textit{First-die-shows-4}) = H(S) - H(S|\textit{First-die-shows-4}) = 3.27 - 2.58 = 0.69 \text{ biţi.}$$

Altfel spus, atunci când ni se comunică faptul că la aruncarea celor două zaruri primul dintre ele produce faţa 4, această informaţie va reduce ulterior entropia variabilei S (sau, am putea spune, „surpriza“ medie provocată de valorile ei) cu 0.69 biţi.

53. **(Probabilităţi marginale, entropii, entropii condiţionale medii)**

■ □ • ***CMU, 2012 spring, Roni Rosenfeld, HW2, pr. 3***

Un doctor trebuie să pună un diagnostic unui pacient care are simptome de răceală (C, de la engl. cold). Factorul principal pe care doctorul îl ia în considerare pentru a elabora diagnosticul este timpul, adică starea vremii de afară (T). Variabila aleatoare C ia două valori, *yes* şi *no*, iar variabila aleatoare T ia 3 valori: *sunny* (însorit), *rainy* (ploios) şi *snowy* (foarte rece, să zicem). Distribuţia comună a celor două variabile este dată în tabelul următor:

	$T = sunny$	$T = rainy$	$T = snowy$
$C = no$	0.30	0.20	0.10
$C = yes$	0.05	0.15	0.20

a. Calculaţi probabilităţile marginale $P(C)$ şi $P(T)$.

***Sugestie*: Folosiţi formula $P(X = x) = \sum_y P(X = x; Y = y)$. De exemplu,**

$P(C = no) = P(C = no, T = sunny) + P(C = no, T = rainy) + P(C = no, T = snowy)$.

b. Calculaţi entropiile $H(C)$ şi $H(T)$.

c. Calculaţi entropiile condiţionale medii $H(C|T)$ şi $H(T|C)$.

Răspuns:

a. Folosind formula dată, vom obţine: $P_C = (0.6, 0.4)$ şi $P_T = (0.35, 0.35, 0.30)$.

b. Aplicând definiţia pentru entropie (vedeţi problema 51), rezultă:

$$\begin{aligned} H(C) &= 0.6 \log_2 \frac{5}{3} + 0.4 \log_2 \frac{5}{2} = \log_2 5 - 0.6 \log_2 3 - 0.4 = 0.971 \text{ biţi} \\ H(T) &= 2 \cdot 0.35 \log_2 \frac{20}{7} + 0.3 \log_2 \frac{10}{3} \\ &= 0.7(2 + \log_2 5 - \log_2 7) + 0.3(1 + \log_2 5 - \log_2 3) \\ &= 1.7 + \log_2 5 - 0.7 \log_2 7 - 0.3 \log_2 3 = 1.581 \text{ biţi.} \end{aligned}$$

c. Aplicând definiţia pentru entropie condiţională medie (vedeţi de asemenea problema 51), vom avea:

$$\begin{aligned} H(C|T) &\stackrel{def.}{=} \sum_{t \in Val(T)} P(T = t) \cdot H(C|T = t) \\ &= P(T = sunny) \cdot H(C|T = sunny) + P(T = rainy) \cdot H(C|T = rainy) + \\ &\quad P(T = snowy) \cdot H(C|T = snowy) \\ &= 0.35 \cdot H\left(\frac{0.30}{0.30 + 0.05}, \frac{0.05}{0.30 + 0.05}\right) + 0.35 \cdot H\left(\frac{0.20}{0.20 + 0.15}, \frac{0.15}{0.20 + 0.15}\right) + \\ &\quad 0.30 \cdot H\left(\frac{0.10}{0.10 + 0.20}, \frac{0.20}{0.20 + 0.10}\right) \\ &= \frac{7}{20} \cdot H\left(\frac{6}{7}, \frac{1}{7}\right) + \frac{7}{20} \cdot H\left(\frac{4}{7}, \frac{3}{7}\right) + \frac{3}{10} \cdot H\left(\frac{1}{3}, \frac{2}{3}\right) \\ &= \frac{7}{20} \cdot \left(\frac{6}{7} \log_2 \frac{7}{6} + \frac{1}{7} \log_2 7\right) + \frac{7}{20} \cdot \left(\frac{4}{7} \log_2 \frac{7}{4} + \frac{3}{7} \log_2 \frac{7}{3}\right) + \frac{3}{10} \cdot \left(\frac{1}{3} \log_2 3 + \frac{2}{3} \log_2 \frac{3}{2}\right) \\ &= \frac{7}{20} \cdot \left(\log_2 7 - \frac{6}{7} - \frac{6}{7} \log_2 3\right) + \frac{7}{20} \cdot \left(\log_2 7 - \frac{8}{7} - \frac{3}{7} \log_2 3\right) + \frac{3}{10} \cdot \left(\log_2 3 - \frac{2}{3}\right) \\ &= \frac{7}{10} \log_2 7 - \left(\frac{3}{10} + \frac{4}{10} + \frac{2}{10}\right) - \left(\frac{6}{20} + \frac{3}{20} - \frac{3}{10}\right) \cdot \log_2 3 \\ &= \frac{7}{10} \log_2 7 - \frac{3}{20} \log_2 3 - \frac{9}{10} = 0.82715 \text{ biţi.} \end{aligned}$$

Similar,

$$\begin{aligned} H(T|C) &\stackrel{def.}{=} \sum_{c \in Val(C)} P(C = c) \cdot H(T|C = c) \\ &= P(C = no) \cdot H(T|C = no) + P(C = yes) \cdot H(T|C = yes) \\ &= 0.60 \cdot H\left(\frac{0.30}{0.30 + 0.20 + 0.10}, \frac{0.20}{0.30 + 0.20 + 0.10}, \frac{0.10}{0.30 + 0.20 + 0.10}\right) \end{aligned}$$

$$
\begin{aligned}
& +0.40 \cdot H\left(\frac{0.05}{0.05+0.15+0.20}, \frac{0.15}{0.05+0.15+0.20}, \frac{0.20}{0.05+0.15+0.20}\right) \\
= \; & \frac{3}{5} \cdot H\left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right) + \frac{2}{5} \cdot H\left(\frac{1}{8}, \frac{3}{8}, \frac{1}{2}\right) \\
= \; & \frac{3}{5}\left(\frac{1}{2} + \frac{1}{3}\log_2 3 + \frac{1}{6}(1+\log_2 3)\right) + \frac{2}{5}\left(\frac{1}{8}\cdot 3 + \frac{3}{8}(3-\log_2 3) + \frac{1}{2}\right) \\
= \; & \frac{3}{5}\left(\frac{2}{3} + \frac{1}{2}\log_2 3\right) + \frac{2}{5}\left(2 - \frac{3}{8}\log_2 3\right) = \frac{6}{5} + \frac{3}{20}\log_2 3 = 1.43774 \text{ biţi.}
\end{aligned}
$$

54. **(Calcularea entropiei unei variabile aleatoare continue: cazul distribuţiei exponenţiale)**

■ □ ***CMU, 2011 spring, Roni Rosenfeld, HW2, pr. 2.c***

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

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

Calculaţi entropia *distribuţiei* continue *exponenţiale* de parametru $\lambda > 0$. Vă reamintim că definiţia p.d.f.-ului acestei distribuţii este următoarea:[97]

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

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

Răspuns:

Dat fiind faptul că funcţia p se anulează pe intervalul $(-\infty, 0)$, este natural ca mai întâi să „rupem" intervalul de integrare pentru $\int_{-\infty}^{\infty} p(x) \log_2 \frac{1}{p(x)}\, dx$ în două: $(-\infty, 0)$ şi $[0, \infty)$. Aşadar,

$$
\begin{aligned}
H(X) &= \int_{-\infty}^{0} p(x) \log_2 \frac{1}{p(x)}\, dx + \int_{0}^{\infty} p(x) \log_2 \frac{1}{p(x)}\, dx \\
&\overset{\text{def. } p}{=} \int_{-\infty}^{0} 0 \log_2 0\, dx + \int_{0}^{\infty} \lambda e^{-\lambda x} \log_2 \frac{1}{\lambda e^{-\lambda x}}\, dx
\end{aligned}
$$

Prima dintre aceste două ultime integrale este 0, conform *indicaţiei* din enunţ. Pentru a putea calcula mai uşor cea de-a doua integrală (în expresia căreia apare numărul e), vom schimba baza logaritmului, şi anume vom trece din baza 2 în baza e (baza logaritmului natural, $\ln$).[98]

[97] La ex. 28.a puteţi vedea graficul acestei funcţii de densitate pentru câteva valori ale parametrului (λ).

[98] Pentru aceasta, vom folosi formula $\log_a b = \frac{\log_c b}{\log_c a}$, valabilă pentru orice $a > 0$, $b > 0$ şi $c > 0$, cu $a \neq 1$ şi $c \neq 1$. În calculele care urmează vom folosi şi alte formule de la logaritmi.

Prin urmare,

$$\begin{aligned} H(X) &= \frac{1}{\ln 2}\int_0^\infty \lambda e^{-\lambda x} \ln \frac{1}{\lambda e^{-\lambda x}}\, dx \\ &= \frac{1}{\ln 2}\int_0^\infty \lambda e^{-\lambda x} \left(\ln \frac{1}{\lambda} + \ln \frac{1}{e^{-\lambda x}}\right) dx \\ &= \frac{1}{\ln 2}\int_0^\infty \lambda e^{-\lambda x} \left(-\ln \lambda + \ln e^{\lambda x}\right) dx \\ &= \frac{1}{\ln 2}\int_0^\infty \lambda e^{-\lambda x} \left(-\ln \lambda + \lambda x\right) dx \\ &= \frac{1}{\ln 2}\int_0^\infty \lambda e^{-\lambda x}(-\ln \lambda)\, dx + \frac{1}{\ln 2}\int_0^\infty \lambda e^{-\lambda x}\lambda x\, dx \\ &= \frac{-\ln \lambda}{\ln 2}\int_0^\infty \lambda e^{-\lambda x}\, dx + \frac{\lambda}{\ln 2}\int_0^\infty \lambda e^{-\lambda x} x\, dx \end{aligned}$$

Prima integrală are valoarea 1, întrucât p este p.d.f.-ul distribuţiei exponenţiale (vedeţi ex.28.a) Cea de-a doua integrală are valoarea $1/lambda$ (vedeţi rezolvarea aceluiaşi ex.28.a). Prin urmare,

$$H(X) = \frac{\ln \lambda}{\ln 2}(-1) - \frac{\lambda}{\ln 2}\left(-\frac{1}{\lambda}\right) = -\frac{\ln \lambda}{\ln 2} + \frac{1}{\ln 2} = \frac{1 - \ln \lambda}{\ln 2}.$$

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

■ □ • ***prelucrare de Liviu Ciortuz, după CMU, 2007 fall, Carlos Guestrin, HW1, pr. 1.2***

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:**[99]

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

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.[100]

[99] În învăţarea automată, divergenţa KL este folosită de exemplu la fundamentarea schemei algoritmice EM. Vedeţi problemele 1.b şi 2 de la capitolul *Algoritmul EM*. Mai general, maximizarea verosimilităţii datelor — în vederea estimării parametrilor distribuţiilor probabiliste — poate fi văzută ca fiind echivalentă cu minimizarea divergenţei KL (vedeţi problema 135).

[100] *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 egalitatea 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 $p = q$.[101]

***Indicaţie*:**
Pentru a demonstra punctul acesta puteţi folosi *inegalitatea lui Jensen*:[102]
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 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 pe de altă parte:

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

Arătaţi că această nouă definiţie a câştigului de informaţie este echivalentă cu definiţia dată anterior (vedeţi problema 51). 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. Arătaţi că $IG(X,Y) = 0$ dacă şi numai dacă X şi Y sunt independente.

Răspuns:

a. Vom dovedi inegalitatea $KL(p||q) \geq 0$ folosind inegalitatea lui Jensen, în expresia căreia vom înlocui f cu funcţia convexă $-\log_2$, pe a_i cu $p(x_i)$ şi pe x_i cu $\frac{q(x_i)}{p(x_i)}$. (Pentru convenienţă, în cele ce urmează vor renunţa la indicele variabilei x.) Vom avea:

$$KL\,(p \,||\, q) \stackrel{def.}{=} -\sum_x p(x) \log \frac{q(x)}{p(x)}$$
$$\stackrel{Jensen}{\geq} -\log \Big(\sum_x p(x) \frac{q(x)}{p(x)} \Big) = -\log \big(\underbrace{\sum_x q(x)}_{1} \big) = -\log 1 = 0.$$

[101] Mai general, $KL(p||q)$ este cu atât mai mică cu cât „asemănarea“ dintre distribuţiile p şi q este mai mare.
[102] Vedeţi problema 71.

Aşadar, $KL\,(p \,||\, q) \geq 0$, oricare ar fi distribuţiile (discrete) p şi q.

Vom demonstra acum că $KL(p||q) = 0 \Leftrightarrow p = q$.
Egalitatea $p(x) = q(x)$ implică $\frac{q(x)}{p(x)} = 1$, deci $\log \frac{q(x)}{p(x)} = 0$ pentru orice x, de unde rezultă imediat $KL(p||q) = 0$.
Pentru a demonstra implicaţia inversă, se ţine cont că în inegalitatea lui Jensen, în cazul funcţiilor strict convexe (cum este $-\log_2$) are loc egalitatea doar în cazul în care $x_i = x_j$ pentru orice i şi j. În cazul de faţă, această condiţie se traduce prin faptul că raportul $\frac{q(x)}{p(x)}$ este acelaşi (α) pentru orice valoare a lui x. Ţinând cont că $\sum_x p(x) = 1$ şi $\sum_x q(x) = \sum_x p(x)\frac{q(x)}{p(x)} = 1$, rezultă că $\alpha = \frac{q(x)}{p(x)} = 1$, deci $p(x) = q(x)$ pentru orice x, ceea ce înseamnă că distribuţiile p şi q sunt identice.

b. Vom folosi regula de multiplicare, şi anume $p(x, y) = p(x \mid y)p(y)$:

$$
\begin{aligned}
KL\,(p_{X,Y} \,||\, (p_X\, p_Y)) &\overset{def.\ KL}{=} -\sum_x \sum_y p(x,y) \log \left(\frac{p(x)p(y)}{p(x,y)} \right) \\
&= -\sum_x \sum_y p(x,y) \log \left(\frac{p(x)}{p(x \mid y)} \right) = -\sum_x \sum_y p(x,y)[\log p(x) - \log p(x \mid y)] \\
&= -\sum_x \sum_y p(x,y) \log p(x) - \Big(-\sum_x \sum_y p(x,y) \log p(x \mid y) \Big) \\
&\overset{pr.\ 51.b}{=} -\sum_x \log p(x) \underbrace{\sum_y p(x,y)}_{=p(x)} - H[X \mid Y] = \sum_x p(x) \log p(x) - H[X \mid Y] \\
&= H[X] - H[X \mid Y] = IG(X,Y)
\end{aligned}
$$

c. Conform punctului b, egalitatea $IG(X, Y) = 0$ este echivalentă cu egalitatea $KL\,(p_{X,Y} \,||\, p_X\, p_Y) = 0$. Conform punctului a, această a doua relaţie este adevărată dacă şi numai dacă distribuţiile $p_{X,Y}$ şi $p_X\, p_Y$ sunt identice, or aceasta este exact definiţia independenţei variabilelor X şi Y.

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

■ □ • ○ ***CMU, 2009 spring, Ziv Bar-Joseph, HW5, pr. 6***

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 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

Răspuns:

a. Pentru calculul informaţiei mutuale putem folosi formula din problema 55.b:

$$MI(X,Y) = \sum_x \sum_y p_{X,Y}(x,y) \log \left(\frac{p_{X,Y}(x,y)}{p_X(x) p_Y(y)} \right)$$

Probabilităţile marginale, estimate în sensul verosimilităţii maxime (MLE), sunt:

	P_{X_1}	P_{X_2}	P_{X_3}	P_{X_4}	P_{X_5}	P_Y
0	1/2	3/8	1/2	1/2	0.5	1/4
1	1/2	5/8	1/2	1/2	0.5	3/4

iar probabilităţile comune sunt:

X_i	Y	$P_{X_1,Y}$	$P_{X_2,Y}$	$P_{X_3,Y}$	$P_{X_4,Y}$	$P_{X_5,Y}$
0	0	1/8	1/8	1/8	1/4	0
0	1	3/8	1/4	3/8	1/4	1/2
1	0	1/8	1/8	1/8	0	1/4
1	1	3/8	1/2	3/8	1/2	1/4

Se poate observa că X_1 şi Y sunt independente, deci $MI(X_1, Y) = 0$, conform proprietăţii care a fost demonstrată la problema 55.c. Similar, $MI(X_3, Y) = 0$. În rest, efectuând calculele obţinem $MI(X_2, Y) = 0.01571$, $MI(X_4, Y) = 0.3113$ şi $MI(X_5, Y) = 0.3113$.

b. La selecţia de trăsături vom alege acele trăsături X_i care au informaţie mutuală nenulă în raport cu Y. Acestea sunt X_2, X_4 şi X_5. Celelalte două trăsături, X_1 şi X_3 sunt independente în raport cu Y.

Totuşi, inspectând datele, observăm că dacă vom selecta doar trăsăturile X_2, X_4 şi X_5 vom avea două instanţe (vedeţi prima şi ultima linie din tabel) care au aceleaşi trăsături ($X_2 = 1, X_4 = 0, X_5 = 1$) dar au etichete / ieşiri diferite: $Y = 0$, respectiv $Y = 1$. Aşadar, vom adăuga la setul de trăsături selectate anterior şi variabila X_1, care va permite dezambiguizarea în cazul acestor două instanţe, menţinând astfel „consistenţa“ setului de date.

***Observaţie*: Deşi $MI(X_1, Y) = 0$ — sau, echivalent spus, X_1 este independent de Y —, nu rezultă că variabila X_1 combinată cu una sau mai multe variabile X_j, cu $j \in \{2, 4, 5\}$, şi formând astfel o nouă variabilă aleatoare, rămâne independentă de Y. Nouă variabilă poate avea câştig de informaţie nenul (în unele cazuri chiar maxim!) în raport cu Y.**

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

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y). \tag{55}$$

Demonstraţi că dacă X şi Y sunt variabile aleatoare independente discrete, atunci $H(X,Y) = H(X) + H(Y)$.

Este adevărată şi reciproca acestei afirmaţii? Adică, atunci când are loc egalitatea $H(X,Y) = H(X) + H(Y)$ rezultă că variabilele X şi Y sunt independente?

Răspuns:

Conform definiţiei câştigului de informaţie (vedeţi problema 51),

$$IG(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{56}$$

De asemenea, conform problemei 55.c,

$$IG(X,Y) = 0 \Leftrightarrow X \text{ şi } Y \text{ sunt independente.} \tag{57}$$

Din relaţiile (56) şi (57) rezultă că

$$H(Y) = H(Y|X) \Leftrightarrow X \text{ şi } Y \text{ sunt independente.} \tag{58}$$

Aşadar, dacă X şi Y sunt independente, coroborând relaţiile (58) şi (55) vom avea $H(X,Y) = H(Y) + H(X)$.

Invers, dacă $H(X,Y) = H(X)+H(Y)$, din relaţia (55) rezultă că $H(Y) = H(Y|X)$, ceea ce implică faptul că X şi Y sunt independente, conform relaţiei (58).

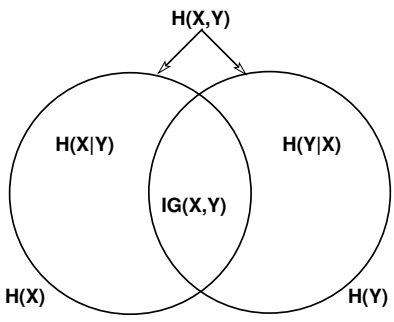
***Observaţie*: Din egalităţile (55) şi (56), rezultă**

$$H(X,Y) = H(X) + H(Y) - IG(X,Y).$$

Conform proprietăţii de nenegativitate a câştigului de informaţie ($IG(X,Y) \geq 0$; vedeţi problema 55.c), rezultă

$$H(X,Y) \leq H(X) + H(Y).$$

Această ultimă relaţie, precum şi relaţiile (55) şi (56) sunt ilustrate în figura alăturată.



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

■ □ • ○ ***CMU, 2011 spring, Roni Rosenfeld, HW2, pr. 3.c***

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:**[103]

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

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. 55, putem scrie:**

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

Cross-entropia — ca şi entropia relativă; vedeţi problema 55 —, 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.

[103] Pentru exemple de folosire a cross-entropiei în învăţarea automată, vedeţi problema 15 de la capitolul *Reţele neuronale artificale*, precum şi problema 2 de la capitolul *Algoritmul EM*.

Răspuns:

a. Nu, cross-entropia nu poate lua valori negative. Iată cum demonstrăm: Ştim că pentru orice funcţii de probabilitate p şi q şi pentru orice x (care aparţine domeniului de valori al unei variabile aleatoare care are o astfel de distribuţie de probabilitate), valorile $p(x)$ şi $q(x)$ satisfac inegalităţile $0 \leq p(x) \leq 1$ and $0 \leq q(x) \leq 1$. Inegalitatea $q(x) \leq 1$ implică faptul că $\log q(x) \leq 0$. Din $0 \leq p(x)$ şi $-\log q(x) \geq 0$, rezultă că $0 \leq -p(x) \log q(x)$. În consecinţă, suma tuturor acestor termeni va fi de asemenea mai mare sau egală cu 0, deci cross-entropia nu poate fi niciodată negativă.

***Observaţie importantă*: Spre deosebire de entropie (vedeţi problema 129), cross-entropia nu este mărginită superior. Ea poate creşte la infinit; vedeţi cazul când pentru o anumită valoare x sunt adevărate simultan relaţiile $p(x) \neq 0$ şi $q(x) = 0$.[104] Prin urmare, nici entropia relativă (adică divergenţa Kullback-Leibler) nu este mărginită superior (vedeţi *Observaţia* din enunţ).**

b. Facem calculele, folosind formula cross-entropiei:

$$\begin{aligned}
&CH(p_{empiric}, q_1) = \\
&\quad -\left(\frac{1}{20}\log_2\frac{1}{10} + \frac{1}{10}\log_2\frac{1}{10} + \frac{1}{5}\log_2\frac{1}{5} + \frac{3}{10}\log_2\frac{3}{10} + \frac{1}{5}\log_2\frac{1}{5} + \frac{1}{10}\log_2\frac{1}{20}\right. \\
&\quad \left. + \frac{1}{20}\log_2\frac{1}{20}\right) = \frac{3}{20}\log_2 10 + \frac{2}{5}\log_2 5 + \frac{3}{10}\log_2\frac{10}{3} + \frac{3}{20}\log_2 20 = \\
&= \frac{3}{20}\log_2 2\cdot 5 + \frac{2}{5}\log_2 5 + \frac{3}{10}\log_2\frac{2\cdot 5}{3} + \frac{3}{20}\log_2 2^2\cdot 5 \\
&= \left(\frac{3}{20} + \frac{3}{10} + 2\cdot\frac{3}{20}\right) + \left(\frac{3}{20} + \frac{2}{5} + \frac{3}{10} + \frac{3}{20}\right)\log_2 5 - \frac{3}{10}\log_2 3 \\
&= \frac{3}{4} + \log_2 5 - \frac{3}{10}\log_2 3 = 2.596439345 \text{ biţi}
\end{aligned}$$

$$\begin{aligned}
&CH(p_{empiric}, q_2) = \\
&\quad -\left(\frac{1}{20}\log_2\frac{1}{20} + \frac{1}{10}\log_2\frac{1}{10} + \frac{1}{5}\log_2\frac{3}{20} + \frac{3}{10}\log_2\frac{7}{20} + \frac{1}{5}\log_2\frac{1}{5} + \frac{1}{10}\log_2\frac{1}{10}\right. \\
&\quad \left. + \frac{1}{20}\log_2\frac{1}{20}\right) = \\
&= \frac{1}{10}\log_2 20 + \frac{1}{5}\log_2 10 + \frac{1}{5}\log_2\frac{20}{3} + \frac{3}{10}\log_2\frac{20}{7} + \frac{1}{5}\log_2 5 \\
&= \frac{1}{10}\log_2 2^2\cdot 5 + \frac{1}{5}\log_2 2\cdot 5 + \frac{1}{5}\log_2\frac{2^2\cdot 5}{3} + \frac{3}{10}\log_2\frac{2^2\cdot 5}{7} + \frac{1}{5}\log_2 5 \\
&= \left(2\cdot\frac{1}{10} + \frac{1}{5} + 2\cdot\frac{1}{5} + 2\cdot\frac{3}{10}\right) + \left(\frac{1}{10} + 3\cdot\frac{1}{5} + \frac{3}{10}\right)\log_2 5 - \frac{1}{5}\log_2 3 - \frac{3}{10}\log_2 7 \\
&= \frac{7}{5} + \log_2 5 - \frac{1}{5}\log_2 3 - \frac{3}{10}\log_2 7 = 2.562729118 \text{ biţi.}
\end{aligned}$$

Aşadar, distribuţia $p_{empiric}$ are o cross-entropie mai mică în [raport cu] modelul q_2. Este deci rezonabil să afirmăm că alegerea modelului q_2 este mai bună.

Totuşi, nu putem garanta că acest model este întotdeauna cel mai bun, fiindcă aici lucrăm cu o distribuţie „empirică", iar distribuţia „adevărată" nu neapărat se reflectă în mod complet / perfect în această distribuţie empirică.

[104] Mai precis, $\lim_{q(x)\to+0}(-p(x)\cdot\log_2 q(x)) = -p(x)(-\infty) = +\infty$.

De obicei, *bias-ul de eşantionare* (engl., sampling bias), precum şi *insuficienţa datelor de antrenament* vor contribui la lărgirea „spaţiului“ care diferenţiază distribuţia adevărată de distribuţia empirică. Prin urmare, în practică, atunci când concepem un [astfel de] experiment de evaluare a mai multor distribuţii probabiliste, trebuie să avem permanent în minte faptul acesta şi, dacă este posibil, să folosim tehnici care reduc / minimizează aceste riscuri.

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

□ ***Liviu Ciortuz, 2012, după www.en.wikipedia.org***

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

a. Arătaţi că pentru orice distribuţie de probabilitate $Q = \{q_1, \dots q_n\}$ 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 $i = 1, \dots, n$.

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

***Indicaţie*: 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 55) şi apoi să folosiţi proprietăţile entropiei relative.**

Răspuns:

Expresia $-\sum_{i=1}^{n} p_i \log_2 \frac{q_i}{p_i}$ la care s-a ajuns în *Indicaţie* este exact divergenţa Kullback-Leibler dintre distribuţiile P şi Q. Formal, scriem acest lucru astfel: $KL(P||Q) = -\sum_{i=1}^{n} p_i \log_2 \frac{q_i}{p_i} = CH(P,Q) - H(P)$.

a. La problema 55.a am demonstrat inegalitatea $KL(P||Q) \geq 0$, care are loc pentru orice distribuţii probabiliste discrete P şi Q. Aceasta este exact proprietatea de care avem nevoie pentru a justifica inegalitatea dată în enunţ la acest punct ($H(P) \leq CH(P,Q)$).[105]

b. Tot la problema 55.a s-a demonstrat că $KL(P||Q)$ are valoarea 0 dacă şi numai dacă distribuţiile P şi Q sunt identice. În contextul nostru, această proprietate se transpune imediat sub forma $H(P) = CH(P,Q) \Leftrightarrow p_i = q_i$ pentru $i = 1, \dots, n$.

[105] Pentru o demonstraţie mai directă a inegalităţii lui Gibbs, de această dată folosind inegalitatea lui Jensen,[106] vedeţi nota de subsol 749 (pagina 850) de la problema 2 de la capitolul *Algoritmul EM*.

4 Arbori de decizie

Sumar

Noţiuni preliminare

– **partiţie a unei mulţimi: ex. 82 de la cap. *Fundamente*;**

– **proprietăţi elementare ale funcţiei logaritm; formule uzuale pentru calcule cu logaritmi;**

• **Elemente de *teoria informaţiei* (vedeţi secţiunea corespunzătoare din cap. *Fundamente*):**

  – **entropie, definiţie: T. Mitchell, *Machine Learning*, 1997 (desemnată în continuare simplu prin *cartea ML*), pag. 57; ex. 2.a, ex. 39.a, ex. 35.a;**
  – **entropie condiţională specifică: ex. 14.a;**
  – **entropie condiţională medie: ex. 2.cd, ex. 35.c;**
  – **câştig de informaţie (definiţie: cartea ML, pag. 58): ex. 2.cd, ex. 5.a, ex. 33, ex. 39.b, ex. 35.e;**

– ***arbori de decizie*, văzuţi *ca structură de date*: ex. 1, ex. 31**
şi, respectiv, ca program în logica propoziţiilor: ex. 2.e, ex. 38.bc;
• (P0) *expresivitatea arborilor de decizie* cu privire la *funcţii boolene*: ex. 32;

– ***spaţiu de versiuni* pentru un concept (de învăţat): ex. 1, ex. 31, ex. 37.**

Algoritmul ID3

– **pseudo-cod: cartea ML, pag. 56;**

– ***bias-ul inductiv*: *ibidem*, pag. 63-64;**

– **exemple simple de aplicare: ex. 2, ex. 3, ex. 5, ex. 36, ex. 37, ex. 39, ex. 40;**

– **ID3 ca algoritm *per se*:**

  • **este un *algoritm de căutare*;**
  ***spaţiul de căutare* — mulţimea tuturor arborilor de decizie care se pot construi cu atributele de intrare în *nodurile de test* şi cu valorile atributului de ieşire în *nodurile de decizie* — este de dimensiune exponenţială în raport cu numărul de atribute: ex. 1, ex. 3, ex. 31, ex. 37;**
  ID3 are ca *obiectiv* căutarea unui arbore / *model* care *i.* să explice cât mai bine datele (în particular, atunci când datele sunt *consistente*, modelul trebuie să fie *consistent* cu acestea), *ii.* să fie cât mai *compact*, din motive de *eficienţă* la generalizare / testare şi *iii.* în final să aibă o [cât mai] bună putere de *generalizare*;[371]
  ◦ **ID3 ar putea fi văzut şi ca algoritm de *optimizare*;**[372]

[371] LC: Alternativ, putem spune că algoritmul ID3 produce o *structură* de tip *ierarhie* (arbore) între diferite *partiţionări* ale setului de instanţe de antrenament, această ierarhie fiind generată pe baza *corespondenţei* dintre atributul de *ieşire* şi atributele de *intrare*, care sunt adăugate la model câte unul pe rând.

[372] LC: Am putea să-l interpretăm pe ID3 ca fiind un algoritm care caută între diferitele *distribuţii probabiliste* discrete care pot fi definite pe setul de date de antrenament una care să satisfacă cerinţa de *ierarhizare*, şi pentru care entropia să fie *minimală* (vedeţi proprietatea de structuralitate de la ex. 50 de la cap. *Fundamente*. Cerinţa ca arborele ID3 să fie *minimal* (ca număr de niveluri / noduri) este însă mai importantă, mai practică şi mai uşor de înţeles.

- ***greedy*** ⇒ **nu garantează obţinerea soluţiei optime d.p.v. al numărului de niveluri / noduri:**
  ex. 4, ex. 21.a, ex. 38 (vs. ex. 37.b, ex. 3.b), ex. 46;
- **de tip *divide-et-impera* (⇒ "*Iterative Dichotomizer*"), recursiv;**
  - ***1-step look-ahead*;**
- **complexitate de timp, cf. *Weka* book:**[373]
  la antrenare, în anumite condiţii: $\mathcal{O}(d\,m\log m)$**; la testare** $\mathcal{O}(d)$**,**
  unde d **este numărul de atribute, iar** m **este numărul de exemple;**

– **ID3 ca algoritm de învăţare automată:**

- ***bias*-ul inductiv al algoritmului ID3:**
  [dorim ca modelul să aibă structură ierarhică, să fie compatibil / consistent cu datele dacă acestea sunt consistente (adică, necontradictorii), iar]
  ***arborele* produs de *ID3* trebuie să aibă un număr cât mai mic de niveluri / noduri;**
- **algoritm de învăţare de tip *"eager"*;**
- ***analiza erorilor*:**
  la antrenare: ex. 7.a, ex. 10.a, ex. 42;[374] **[acurateţe la antrenare: ex. 6;]**
  la validare: CMU, 2003 fall, T. Mitchell, A. Moore, midterm, pr. 1;
  la n**-fold cross-validare**
  la cross-validare leave-one-out (CVLOO): ex. 10.b, ex. 45.bc;[375]
- ***robusteţe la „zgomote" şi overfitting*: ex. 10, ex. 21.bc, ex. 45, ex. 67.b;**[376]
- ***zone de decizie* şi *graniţe de separare / decizie* pentru arbori de decizie cu variabile continue: ex. 10, ex. 44, ex. 45, ex. 46.**
  ***Observaţie*: Zonele de decizie produse de algoritmul ID3 nu sunt în mod neapărat unice, fiindcă arborele de decizie creat de ID3 nu este determinat în mod unic (vedeţi proprietatea (P2), mai jos).**

Extensii / variante ale algoritmului ID3

– ***atribute cu valori continue*: ex. 10-12, ex. 14.c, ex. 43-47; cap. *Învăţare bazată pe memorare*, ex. 11.b;**

– **atribute discrete cu multe valori: ex. 13;**

– **atribute cu valori nespecificate / absente pentru unele instanţe;**

– **atribute cu diferite costuri asociate: ex. 14;**

– **reducerea caracterului "eager" al învăţării: ex. 16;**

– **reducerea caracterului "greedy" al învăţării:**
IG cu "2-step look-ahead": ex. 17, ex. 18;
variante de tip "look-ahead" specifice atributelor continue: ex. 48;

– **folosirea altor *măsuri de „impuritate"* în locul câştigului de informaţie:**
Gini Impurity, Misclassification Impurity: ex. 15;

[373] "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations", Ian Witten, Eibe Frank (3rd ed.), Morgan Kaufmann Publishers, 2011.
[374] De asemenea, ex. 4.ab de la capitolul *Clasificare bayesiană*.
[375] De asemenea, ex. 4.ab de la capitolul *Clasificare bayesiană*.
[376] De asemenea, ex. 4.ab de la capitolul *Clasificare bayesiană*.

– **reducerea *overfitting*-ului:**
reduced-error pruning (folosind un set de date de validare):
cartea ML, pag. 69-71; A. Cornuéjols, L. Miclet, 2nd ed., pag. 418-421;
rule post-pruning: cartea ML, pag. 71-72; ex. 50
top-down vs. bottom-up pruning, folosind un criteriu bazat pe câştigul de informaţie: ex. 19, ex. 49;
pruning folosind testul statistic χ^2: ex. 20, ex. 51.

Proprietăţi ale arborilor ID3

- **(P1) arborele produs de algoritmul ID3 este *consistent* (adică, în concordanţă) cu datele de antrenament, dacă acestea sunt *consistente* (adică, necontradictorii). Altfel spus, *eroarea la antrenare* produsă de algoritmul ID3 pe orice set de *date consistente* este 0: ex. 2-4, ex. 37-38;**
- **(P2) arborele produs de algoritmul ID3 *nu* este în mod neapărat *unic*: ex. 3, ex. 37;**
- **(P3) arborele ID3 nu este neapărat *optimal* (ca nr. de noduri / niveluri): ex. 4, ex. 21, ex. 38;**
- **(P4) influenţa *atributelor identice* şi, respectiv, a *instanţelor multiple* asupra arborelui ID3: ex. 8;**
  - **(P5) *o margine superioară* pentru *eroarea la antrenare* a algoritmului ID3, în funcţie de numărul de valori ale variabilei de ieşire): ex. 7.b;**
  - **(P6) o aproximare simplă a numărului de *instanţe greşit clasificate* din totalul de M instanţe care au fost asignate la un nod frunză al unui arbore ID3, cu ajutorul entropiei (H) nodului respectiv: ex. 41;**
- **(P7) *graniţele de separare / decizie* pentru arborii ID3 cu *atribute de intrare continue* sunt întotdeauna paralele cu axele de coordonate: ex. 10, ex. 12, ex. 44, ex. 45, ex. 46 şi cap. *Învăţare bazată pe memorare*, ex. 11.b;**

  ***Observaţie*: Următoarele trei proprietăţi se referă la arbori de decizie în general, nu doar la arbori ID3.**
- **(P8) *adâncimea maximă* a unui arbore de deczie, când atributele de intrare sunt categoriale: numărul de atribute: ex. 52.c;**
  - **(P9) o *margine superioară* pentru *adâncimea* unui arbore de deczie când atributele de intrare sunt continue, iar datele de antrenament sunt (ne)separabile liniar: ex. 11;**
  - **(P10) o *margine superioară* pentru numărul de *noduri-frunză* dintr-un arbore de decizie, în funcţie de numărul de exemple şi de numărul de atribute de intrare, atunci când acestea (atributele de intrare) sunt binare: ex. 9.**

Învăţare automată de tip ansamblist folosind arbori de decizie: Algoritmul AdaBoost

- **Noţiuni preliminare:**
distribuţie de probabilitate discretă, factor de normalizare pentru o distribuţie de probabilitate, ipoteze „slabe“ (engl., weak hypotsesis), compas de decizie (engl., decision stump), prag de separare (engl., threshold split) pentru un

compas de decizie, prag exterior de separare (engl., outside threshold split), eroare ponderată la antrenare (engl., weighted training error), vot majoritar ponderat (engl., weighted majority vote), overfitting, ansambluri de clasificatori (vedeţi ex. 64), funcţii de cost / pierdere (engl., loss function) (vedeţi ex. 29 şi ex. 23);

- **pseudo-codul algoritmului AdaBoost + proprietăţi de bază + convergenţa erorii la antrenare: ex. 22 şi 23;**
- **exemple de aplicare: ex. 24, 56, 54, 55, 57, 58.**

- **AdaBoost ca algoritm *per se*:**
  algoritm iterativ, algoritm de căutare (spaţiul de căutare este mulţimea combinaţiilor liniare care se pot construi peste clasa de ipoteze „slabe“ considerate), algoritm de opimizare secvenţială (minimizează o margine superioară pentru eroarea la antrenare), algoritm greedy (dacă la fiecare iteraţie se alege cea mai bună ipoteză „slabă“).
  - ***învăţabilitate empirică* γ*-slabă*:**
    definiţie: ex. 23.e
    exemplificarea unor cazuri când nu există *garanţie* pentru învăţabilitate γ-slabă: ex. 25, 60;
  - **AdaBoost ca algoritm de *optimizare secvenţială* în raport cu funcţia de cost / „pierdere“ negativ-exponenţială: ex. 26;**
  - **marginea de votare: ex. 27, 28 şi 65;**
  - ***selectarea trăsăturilor* folosind AdaBoost; aplicare la clasificarea de documente: ex. 62;**
  - **o variantă generalizată a algoritmului AdaBoost: ex. 29 şi ex. 65;**
  - **recapitulare (întrebări cu răspuns *adevărat* / *fals*): ex. 30 şi 66.**

- **Proprietăţi ale algoritmului AdaBoost:**

  (P0) AdaBoost poate produce rezultate diferite atunci când are posibilitatea să aleagă între două sau mai multe [cele mai bune] ipoteze „slabe“: ex. 24, 54;

  (P1) $err_{D_{t+1}}(h_t) = \frac{1}{2}$ **(ex. 22.*vii*);**
  ca o *consecinţă*, rezultă că ipoteza h_t nu poate fi reselectată şi la iteraţia $t+1$; ea poate fi reselectată la o iteraţie ulterioară;

  (P2) Din relaţia de definiţie pentru distribuţia D_{t+1} rezultă
  $Z_t = e^{-\alpha_t} \cdot (1-\varepsilon_t) + e^{\alpha_t} \cdot \varepsilon_t = 2\sqrt{\varepsilon_t(1-\varepsilon_t)}$ **şi**
  $\varepsilon_t \in (0, 1/2) \Rightarrow Z_t \in (0,1)$ **(ex. 22).**

  (P3) $D_{t+1}(i) = \frac{1}{m \prod_{t'=1}^{t} Z_{t'}} e^{-y_i f_t(x_i)}$, **unde** $f_t(x_i) \stackrel{def.}{=} -\sum_{t'=1}^{t} \alpha_{t'} h_{t'}(x_i)$ **(ex. 23.a).**
  Produsul $y_i f_t(x_i)$ se numeşte *margine algebrică*;

  (P4) $err_S(H_t) \leq \prod_{t'=1}^{t} Z_t$, **adică eroarea la antrenare comisă de ipoteza combinată produsă de AdaBoost este majorată de produsul factorilor de normalizare (ex. 23.b);**

  (P5) AdaBoost nu optimizează în mod direct $err_S(H_t)$, ci marginea sa superioară, $\prod_{t'=1}^{t} Z_t$; optimizarea se face în mod secvenţial (greedy): la iteraţia t se minimizează valoarea lui Z_t ca funcţie de α_t, ceea ce conduce la $\alpha_t = \ln \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}}$

(ex. 23.c);

(P5′) $\varepsilon_i > \varepsilon_j \Leftrightarrow \alpha_i < \alpha_j$ **(consecinţă imediată din relaţia** $\alpha_t = \ln\sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}}$**): ex. 22.**$v$**;**

(P6) O consecinţă din relaţia (175) **şi (P5):** $D_{t+1}(i) = \begin{cases} \frac{1}{2\varepsilon_t} D_t(i), & i \in M \\ \frac{1}{2(1-\varepsilon_t)} D_t(i), & i \in C \end{cases}$

(ex. 22.iv**);**

(P7) $err_S(H_t)$ **nu neapărat descreşte de la o iteraţie la alta; în schimb, descresc marginile sale superioare:** $\prod_{t'=1}^{t} Z_{t'}$ **şi** $\exp(-\sum_{t'=1}^{t} \gamma_{t'}^2)$ **(ex. 23.d);**

(P8) O condiţie suficientă pentru învăţabilitate γ**-slabă, bazată pe marginea de votare: marginea de votare a oricărei instanţe de antrenament să fie de cel puţin** 2γ**, la orice iteraţie a algoritmului AdaBoost (ex. 28);**

(P9) Orice mulţime formată din m **instanţe din** $\mathbb{R}$ **care sunt etichetate în mod consistent poate fi corect clasificată de către o combinaţie liniară formată din cel mult** $2m$ **compaşi de decizie (ex. 64.a);**

(P10) Orice mulţime de instanţe distincte [şi etichetate] din $\mathbb{R}$ **este** γ**-slab învăţabilă cu ajutorul compaşilor de decizie (ex. 63).**

- **Alte metode de învăţare ansamblistă bazate pe arbori de decizie: Bagging, Random Forests;**
- **Alte metode de învăţare automată bazate pe arbori: arbori de regresie (CART).**

4.1 Arbori de decizie — Probleme rezolvate

4.1.1 Algoritmul ID3

1. **(Arbori de decizie; optimalitate, relativ la numărul de noduri)**

Reprezentaţi arborele / arborii de decizie care are / au numărul minim de noduri posibile şi corespunde / corespund funcţiei booleene $(\neg A \vee B) \wedge \neg(C \wedge A)$ definită peste atributele booleene A, B şi C.

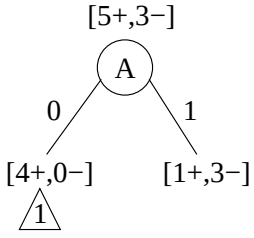
Răspuns:

A	B	C	X
0	0	0	1
0	0	1	1
0	1	0	1
0	1	1	1
1	0	0	0
1	0	1	0
1	1	0	1
1	1	1	0

Vom determina arborele de decizie optimal (ca număr de noduri) parcurgând în mod *exhaustiv spaţiul de versiuni*, adică mulţimea tuturor arborilor de decizie (construiţi cu variabilele A, B şi C) care sunt *consistenţi* cu funcţia dată. Aşadar, vom examina ce se întâmplă când în nodul rădăcină se pun pe rând atributele A, B şi respectiv C.

Notăm cu X funcţia $(\neg A \vee B) \wedge \neg(C \wedge A)$, ale cărei valori sunt date în tabelul alăturat.

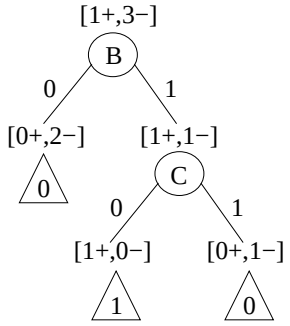
• ***Cazul 1*: Dacă în nodul rădăcină se plasează atributul A, mulţimea de exemple va fi re-partiţionată conform reprezentării alăturate.**

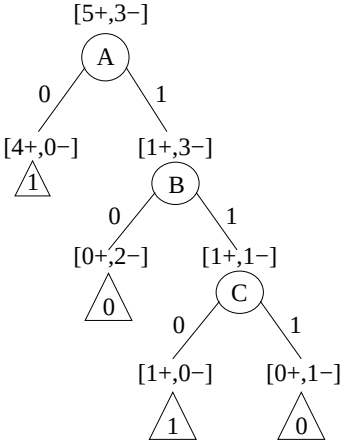


Subarborele drept al acestui arbore va trebui să reprezinte arborele de decizie pentru funcţia

$$X_1 = X[A/1] = (\neg 1 \vee B) \wedge \neg(C \wedge 1) = B \wedge \neg C,$$

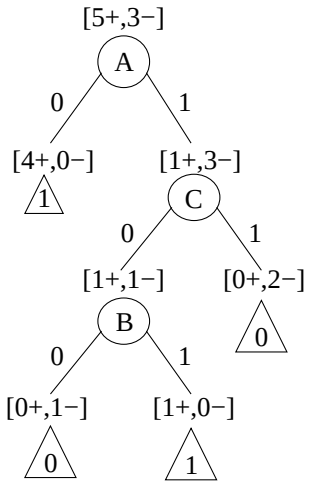
pentru care o reprezentare optimă este redată mai jos, în partea stângă:



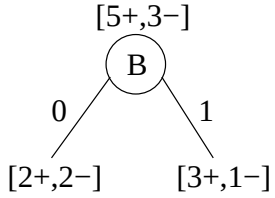


Prin urmare, un arbore optim (ca număr de noduri) care are variabila A în nodul rădăcină este cel reprezentat mai sus, în partea dreaptă.

***Observaţie*: Evident, există încă un arbore optim care are variabila A în nodul rădăcină (el corespunde unei alte reprezentări optimale a conjuncţiei $B \wedge \neg C$ faţă de cea de mai sus). Vedeţi desenul alăturat.**



- ***Cazul 2*: Dacă în nodul rădăcină se alege atributul B, se obţine partiţia**



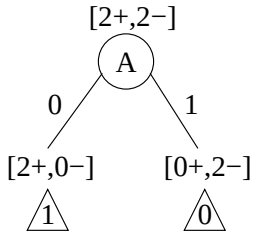
Subarborele stâng şi subarborele drept trebuie să reprezinte arborele de decizie pentru funcţiile

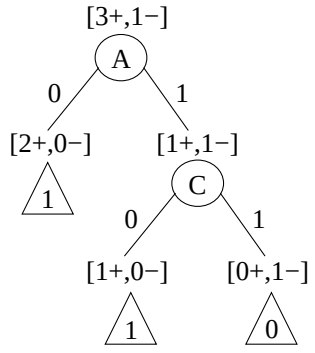
$$X_2 = X[B/0] = (\neg A \vee 0) \wedge \neg(C \wedge A) = \neg A \wedge (\neg C \vee \neg A) = (\neg A \wedge \neg C) \vee \neg A = \neg A,$$

şi respectiv

$$X_3 = X[B/1] = (\neg A \vee 1) \wedge \neg(C \wedge A) = 1 \wedge (\neg C \vee \neg A) = \neg C \vee \neg A,$$

care au ca reprezentări optime arborii de mai jos:

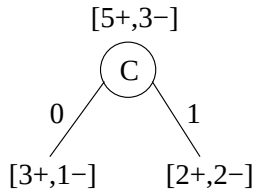




Pentru arborele din dreapta există un arbore de decizie echivalent, obţinut prin interschimbarea lui A cu C.

Prin urmare, orice arbore optim având variabila B în rădăcină are 3 niveluri şi 4 noduri, aşadar cu un nod (de test) mai mult decât cel determinat în primul caz.

• ***Cazul 3*: În sfârşit, când în nodul rădăcină se alege atributul C, obţinem partiţia**



Subarborele stâng şi subarborele drept trebuie să reprezinte arborele de decizie pentru funcţiile

$$X_4 = X[C/0] = (\neg A \vee B) \wedge \neg(0 \wedge A) = (\neg A \vee B) \wedge \neg 0 = \neg A \vee B$$

şi respectiv

$$X_5 = X[C/1] = (\neg A \vee B) \wedge \neg(1 \wedge A) = (\neg A \vee B) \wedge \neg A = \neg A.$$

Urmând un raţionament similar cu cel de la cazul anterior, putem spune că orice arbore optim cu atributul C în rădăcină are 3 niveluri şi 4 noduri, cu un nod (de test) mai mult decât cel determinat în primul caz.

Aşadar, putem concluziona că arborii de decizie optimi corespunzători funcţiei date sunt cei determinaţi în primul caz.

2. **(Algoritmul ID3: aplicare)**

■ • ***CMU, 2002 spring, A. Moore, midterm example questions, pr. 2***

Ai naufragiat pe o insulă pustie, unde nu găseşti niciun alt fel de hrană decât ciuperci. Despre unele dintre aceste ciuperci se ştie că sunt otrăvitoare, despre altele se ştie că sunt comestibile, iar despre restul nu se ştie ce fel sunt. Ai rămas singur pe insulă — foştii tăi camarazi, fiind epuizaţi de foame, au folosit metoda 'trial and error'... — şi ai la dispoziţie următoarele date:

Exemplu	*Uşoară*	*Mirositoare*	*ArePete*	*Netedă*	*Comestibilă*
A	1	0	0	0	1
B	1	0	1	0	1
C	0	1	0	1	1
D	0	0	0	1	0
E	1	1	1	0	0
F	1	0	1	1	0
G	1	0	0	1	0
H	0	1	0	0	0
U	0	1	1	1	?
V	1	1	0	1	?
W	1	1	0	0	?

Atunci când nu vei mai avea la dispoziţie pentru a supravieţui decât ciuperci U, V, sau W, ai putea estima care dintre ele sunt comestibile, folosind arbori de decizie.

În primele trei întrebări care urmează, ne vom referi la ciupercile $A - H$:

a. Care este entropia atributului *Comestibilă*?

b. Doar privind datele — adică fără a face explicit calculul câştigului de informaţie (engl., information gain) pentru cele patru atribute — poţi determina ce atribut vei alege ca rădăcină a arborelui de decizie?

c. Calculează câştigul de informaţie pentru atributul pe care l-ai ales la întrebarea precedentă.

d. Elaborează întregul arbore de decizie ID3 bazat pe datele din tabel şi apoi clasifică ciupercile U, V, W.

e. Exprimă cu ajutorul calculului propoziţional (logica predicatelor de ordinul 0) clasificarea produsă de arborele de decizie obţinut. ($Comestibilă \leftrightarrow \ldots$)

f. Există vreun risc dacă vei consuma ciuperci care au fost clasificate de arborele de decizie ca fiind comestibile? De ce da? sau, de ce nu?

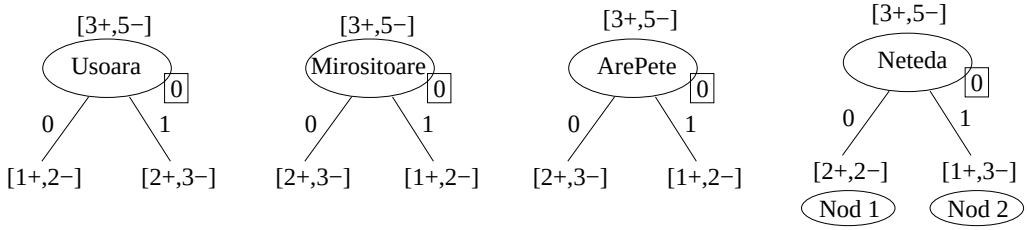
Răspuns:

a. Entropia atributului *Comestibilă* este:

$$\begin{aligned} H_{Comestibilă} &\overset{not.}{=} H[3+,5-] \overset{def.}{=} -\frac{3}{8}\log_2\frac{3}{8} - \frac{5}{8}\log_2\frac{5}{8} = \frac{3}{8}\log_2\frac{8}{3} + \frac{5}{8}\log_2\frac{8}{5} = \\ &= \frac{3}{8}3 - \frac{3}{8}\log_2 3 + \frac{5}{8}3 - \frac{5}{8}\log_2 5 = 3 - \frac{3}{8}\log_2 3 - \frac{5}{8}\log_2 5 \approx \\ &\approx 0.9544 \end{aligned}$$

***Observaţie* (1): Notaţia $[3+,5-]$ simbolizează o mulţime partiţionată în 3 exemple pozitive şi 5 exemple negative. Vom folosi acest gen de notaţie peste tot în continuare, cu mici variaţii determinate de valorile pe care le poate lua atributul de ieşire. De exemplu, dacă vorbim despre o mulţime cu 5 obiecte roşii, 3 albastre şi 4 verzi, am putea nota: $[5R, 3A, 4V]$.**

b. În rădăcina arborelui de decizie se alege atributul care aduce cel mai mare câştig de informaţie. Adică, atributul care, intuitiv vorbind, partiţionează cel mai bine datele de antrenament în raport cu atributul de ieşire. În cazul nostru, variantele pe care le avem la dispoziţie pentru rădăcina arborelui (nodul 0) sunt:



Este uşor de observat că atributele *Uşoară*, *Mirositoare* şi *ArePete* împart mulţimea exemplelor în mod similar: o submulţime cu 3 elemente, dintre care unul este pozitiv iar două sunt negative, şi o submulţime cu 5 elemente, dintre care două sunt pozitive, iar trei sunt negative.

Dacă am considera un arbore de decizie cu un singur nod de test în care plasăm atributul *Netedă*, atunci numărul minim de erori la antrenare pe care îl putem obţine este 3, utilizând următoarea clasificare:

- *Netedă* $= \mathbf{0}$ **:** *Comestibilă* $= \mathbf{1} \Rightarrow$ **ciupercile** E **şi** H **sunt clasificate greşit**
- *Netedă* $= \mathbf{1}$ **:** *Comestibilă* $= \mathbf{0} \Rightarrow$ **ciuperca** C **este clasificată greşit**

Dacă, în schimb, vom pune în rădăcina arborelui de decizie unul dintre celelalte trei atribute, spre exemplu atributul *Uşoară*, şi dacă vom lua votul majoritar în fiecare nod descendent din nodul rădăcină, eroarea rezultată la antrenare va fi aceeaşi ca mai sus (3/8), însă toate instanţele vor fi clasificate la fel (şi anume, negativ). Dacă nu lucrăm cu vot majoritar pentru ambii descendenţi, ci doar pentru cel cu entropie mai mică (în vreme ce pentru celălalt nod descendent luăm decizia contrară), se observă că pentru atributul *Uşoară* vom obţine 4 erori pe setul de antrenament, iar pentru atributul *Netedă* vom obţine 3 erori.

Sumarizând, suntem înclinaţi să credem că ar fi o alegere sensibil mai bună să punem în rădăcină atributul *Netedă*. Pentru o justificare numerică riguroasă a acestei alegeri folosind criteriul maximizării câştigului de informaţie, vedeţi punctul d.

c. Pentru a obţine câştigul de informaţie pentru atributul *Netedă*, se fac calculele:

$$\begin{aligned} H_{0/Netedă} &\overset{def.}{=} \frac{4}{8}H[2+,2-] + \frac{4}{8}H[1+,3-] = \frac{1}{2}\cdot 1 + \frac{1}{2}\left(\frac{1}{4}\log_2\frac{4}{1} + \frac{3}{4}\log_2\frac{4}{3}\right) \\ &= \frac{1}{2} + \frac{1}{2}\left(\frac{1}{4}\cdot 2 + \frac{3}{4}\cdot 2 - \frac{3}{4}\log_2 3\right) = \frac{1}{2} + \frac{1}{2}\left(2 - \frac{3}{4}\log_2 3\right) \\ &= \frac{1}{2} + 1 - \frac{3}{8}\log_2 3 = \frac{3}{2} - \frac{3}{8}\log_2 3 \approx 0.9056 \end{aligned}$$

$$IG_{0/Netedă} \overset{def.}{=} H_{Comestibilă} - H_{0/Netedă} = 0.9544 - 0.9056 = 0.0488$$

***Observaţie* (2): În cele de mai sus am notat cu $H_{0/Netedă}$ entropia *partiţiei* [mulţimii de exemple de antrenament] determinate de alegerea atributului *Netedă* în nodul 0,[377] iar cu $IG_{0/Netedă}$ câştigul de informaţie corespunzător acestei alegeri. În general, prin notaţia $H_{n/A}$ vom înţelege entropia partiţiei determinate de alegerea atributului A în nodul n.**

d. Arborele de decizie ID3 se construieşte pornind din rădăcină şi alegând atributul pentru fiecare nod de test în modul următor:

***Nodul 0* (rădăcina):**
Să verificăm dacă alegerea făcută la punctul b este cea corectă:

$$\begin{aligned} H_{0/Uşoară} &\overset{def.}{=} \frac{3}{8}H[1+,2-] + \frac{5}{8}H[2+,3-] \\ &= \frac{3}{8}\left(\frac{1}{3}\log_2\frac{3}{1} + \frac{2}{3}\log_2\frac{3}{2}\right) + \frac{5}{8}\left(\frac{2}{5}\log_2\frac{5}{2} + \frac{3}{5}\log_2\frac{5}{3}\right) \\ &= \frac{3}{8}\left(\frac{1}{3}\log_2 3 + \frac{2}{3}\log_2 3 - \frac{2}{3}\cdot 1\right) + \frac{5}{8}\left(\frac{2}{5}\log_2 5 - \frac{2}{5}\cdot 1 + \frac{3}{5}\log_2 5 - \frac{3}{5}\log_2 3\right) \\ &= \frac{3}{8}\left(\log_2 3 - \frac{2}{3}\right) + \frac{5}{8}\left(\log_2 5 - \frac{3}{5}\log_2 3 - \frac{2}{5}\right) \\ &= \frac{3}{8}\log_2 3 - \frac{2}{8} + \frac{5}{8}\log_2 5 - \frac{3}{8}\log_2 3 - \frac{2}{8} = \frac{5}{8}\log_2 5 - \frac{4}{8} \approx 0.9512 \end{aligned}$$

[377] Mai riguros, folosind terminologia din *Teoria informaţiei*, vom spune că notaţia $H_{0/Netedă}$ se referă la *entropia condiţională medie* a atributului de ieşire *Comestibilă* în raport cu atributul de intrare *Netedă*.

Urmează că

$$IG_{0/\,Uşoară} \overset{def.}{=} H_{Comestibilă} - H_{0/\,Uşoară} = 0.9544 - 0.9512 = 0.0032,$$

deci

$$IG_{0/\,Uşoară} = IG_{0/Mirositoare} = IG_{0/ArePete} = 0.0032 < IG_{0/Netedă} = 0.0488$$

Am avut deci dreptate să alegem atributul *Netedă* **la punctul** b**.**

***Observaţie importantă*:**
În loc să fi calculat efectiv aceste câştiguri de informaţie, pentru a determina atributul cel mai „bun", ar fi fost *suficient* să elaborăm un *raţionament* de tip *relaţional*, bazat pe comparaţia dintre valorile entropiilor condiţionale medii $H_{0/Netedă}$ **şi** $H_{0/\,Uşoară}$**:**

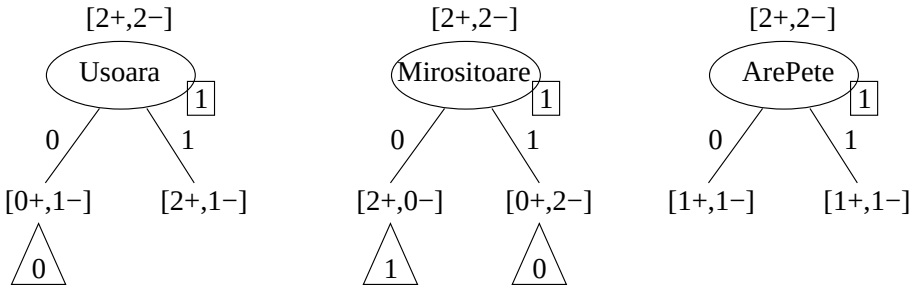
$$\begin{aligned} & IG_{0/Netedă} > IG_{0/\,Uşoară} \Leftrightarrow H_{0/Netedă} < H_{0/\,Uşoară} \\ \Leftrightarrow \quad & \frac{3}{2} - \frac{3}{8}\log_2 3 < \frac{5}{8}\log_2 5 - \frac{1}{2} \Leftrightarrow 12 - 3\log_2 3 < 5\log_2 5 - 4 \\ \Leftrightarrow \quad & 16 < 5\log_2 5 + 3\log_2 3 \Leftrightarrow 16 < 11.6096 + 4.7548 \ \textbf{(adev.)} \end{aligned}$$

În mod *alternativ*, ţinând cont de relaţia (197) **de la problema 35, putem proceda chiar *mai simplu* relativ la calcule (nu doar aici, ci ori de câte ori *nu* avem de-a face cu un *număr mare de instanţe*):**

$$H_{0/Netedă} < H_{0/\,Uşoară} \Leftrightarrow \frac{4^4}{\cancel{2^2}\cdot\cancel{2^2}} \cdot \frac{4^4}{3^3} < \frac{5^5}{\cancel{2^2}\cdot\cancel{3^3}} \cdot \frac{\cancel{3^3}}{\cancel{2^2}} \Leftrightarrow \frac{4^8}{3^3} < 5^5 \Leftrightarrow 4^8 < 3^3 \cdot 5^5$$
$$\Leftrightarrow 2^{16} < 3^3 \cdot 5^5 \Leftrightarrow 64 \cdot 2^{10} < 27 \cdot 25 \cdot 125 \ \textbf{(adev.)}$$

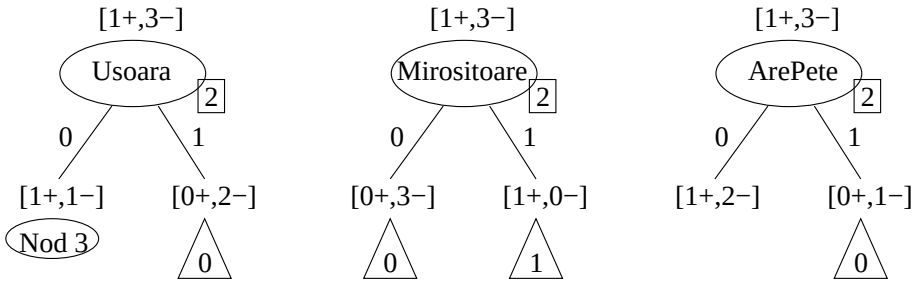
Vă sfătuim să *procedaţi aşa* la rezolvarea problemelor propuse din acest capitol, acolo unde este cazul.

***Nodul 1*: Trebuie să clasificăm acele exemple care au** *Netedă* $= 0$**; avem de ales între 3 atribute -** *Uşoară*, *Mirositoare* **şi** *ArePete*.



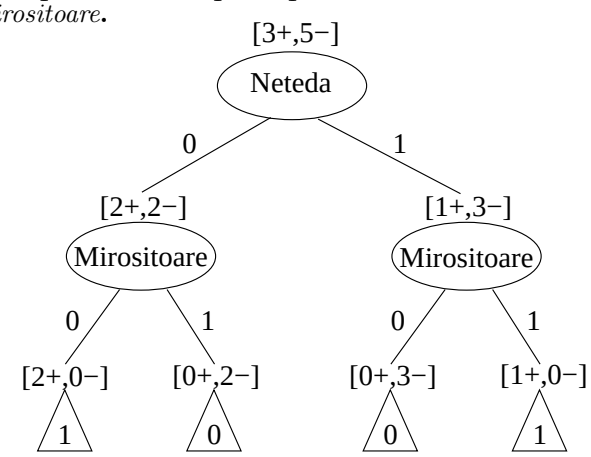
Avem $H_{1/Mirositoare} = \frac{2}{4}H[2+,0-] + \frac{2}{4}H[0+,2-] = 0$**. Oricare ar fi valorile pentru** $H_{1/ArePete}$ **şi** $H_{1/\,Uşoară}$**, întrucât ştim că entropia are întotdeauna valori nenegative, rezultă că atributul** *Mirositoare* **maximizează în nodul 1 câştigul de informaţie. În imaginea de mai sus valorile din triunghi reprezintă decizia luată de subarborele construit în nodul-frunză respectiv.**

***Nodul 2*: Avem de clasificat exemplele pentru care** *Netedă* $= 1$**. Atributele disponibile sunt:** *Uşoară*, *Mirositoare* **şi** *ArePete*.



Evident, $H_{2/Mirositoare} = \frac{3}{4}H[0+,3-] + \frac{1}{4}H[1+,0-] = \frac{3}{4}\cdot 0 + \frac{1}{4}\cdot 0 = 0$ **Aşadar, pentru nodul 2 putem alege atributul** *Mirositoare*.

Arborele complet arată astfel:



Parcurgând arborele construit, ciupercile U, V **şi** W **vor fi clasificate astfel:**

U	$\boldsymbol{Netedă = 1,\ Mirositoare = 1 \Rightarrow Comestibilă = 1}$
V	$\boldsymbol{Netedă = 1,\ Mirositoare = 1 \Rightarrow Comestibilă = 1}$
W	$\boldsymbol{Netedă = 0,\ Mirositoare = 1 \Rightarrow Comestibilă = 0}$

e. $Comestibilă \leftrightarrow (\neg Netedă \wedge \neg Mirositoare) \vee (Netedă \wedge Mirositoare)$

Acelaşi lucru poate fi exprimat şi sub forma unui pseudo-cod ***if*** $\cdots$ ***then Comestibilă else*** $\neg$ ***Comestibilă*****:**

IF (***Netedă*** $=$ **0 AND** ***Mirositoare*** $=$ **0) OR**
(***Netedă*** $=$ **1 AND** ***Mirositoare*** $=$ **1)**
THEN ***Comestibilă*****;**
ELSE $\neg$***Comestibilă*****;**

f. Arborele de decizie produs de către algoritmul ID3 elaborat mai sus este consistent cu datele de antrenament pe care le-am avut la dispoziţie (fiindcă aceste date sunt necontradictorii). Întrucât în realitate clasificarea poate depinde şi de alte trăsături / informaţii decât cele de care dispunem noi, nu avem garanţia că arborele ID3 face identificarea corectă a etichetei / clasei pentru toate instanţele din setul de test. Aşadar, nu putem fi siguri că nu ne vom îmbolnăvi dacă vom consuma ciupercile U **şi** V**, sau că ne vom îmbolnăvi dacă vom consuma ciuperca** W**. În multe aplicaţii practice, calitatea unui model de învăţare automată (în cazul de faţă, un arbore de decizie) se verifică pe un set de *date de validare*.**

3. **(Algoritmul ID3, aplicat pe expresii booleene; exploatarea simetriilor operaţiilor $\vee, \wedge$ în alegerea nodurilor; analiza „optimalităţii“ arborelui ID3, ca număr de noduri de test)**

prelucrare de Liviu Ciortuz, după Tom Mitchell, "Machine Learning", 1997, ex. 3.1.b

Considerăm următoarea funcţie booleană: $A \vee (B \wedge C)$. Presupunem că această funcţie este deja definită — adică valoarea ei este cea cunoscută din logica propoziţiilor —, însă dorim să o reprezentăm ca arbore de decizie.

a. Aplicaţi algoritmul ID3 [tabelei de adevăr corespunzătoare] acestei funcţii. *Observaţie*: Dacă exploataţi simetriile, veţi avea nevoie doar de puţine calcule, altfel vă veţi complica inutil.

b. Arborele ID3 obţinut la punctul precedent este optimal?
Alfel spus, puteţi găsi un alt arbore de decizie, de adâncime mai mică sau cu număr mai mic de noduri (comparativ cu arborele obţinut la punctul a), care să reprezinte această funcţie? (Ţineţi cont că în fiecare nod al unui arbore de decizie se poate testa un sigur atribut.)

Răspuns:

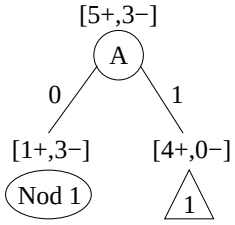
A	B	C	$Y = A \vee (B \wedge C)$
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	1
1	0	1	1
1	1	0	1
1	1	1	1

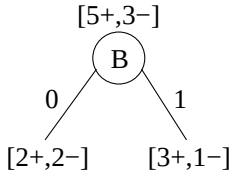
a. Observăm că funcţia dată este simetrică în B şi C, datorită comutativităţii operatorului logic $\wedge$. O consecinţă a acestui fapt este că dacă, pe parcursul algoritmului ID3, avem de ales (şi) între cele două atribute este nevoie să-l studiem doar pe unul dintre ele, celălalt comportându-se identic.

***Nodul 0* (rădăcină):**

$$H_{0/A} = \frac{4}{8}H[1+,3-] + \frac{4}{8}H[4+,0-] =$$
$$= \frac{1}{2}H[1+,3-] + \frac{1}{2}\cdot 0 = \frac{1}{2}H[1+,3-]$$

$$H_{0/B} = \frac{4}{8}H[2+,2-] + \frac{4}{8}H[3+,1-] =$$
$$= \frac{1}{2}\cdot 1 + \frac{1}{2}H[3+,1-] = \frac{1}{2} + \frac{1}{2}H[1+,3-]$$





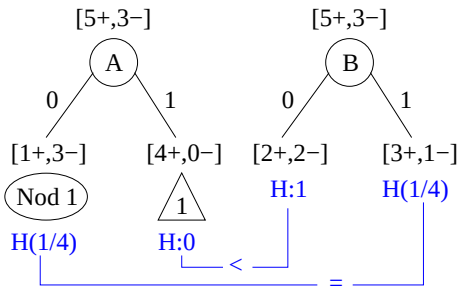
Este evident că $H_{0/A} < H_{0/B}$, deci vom alege atributul A în rădăcină.

***Observaţii importante*:**

1. La aceeaşi concluzie se putea ajunge *imediat* pe baza unui *raţionament* de tip *calitativ*, şi anume, comparând atent cei doi arbori („compaşi de de decizie“) de mai sus. Mai precis, vom compara (două câte două) entropiile

condiţionale specifice din nodurile descendente, precum şi ponderile cu care se combină aceste entropii în scrierea entropiilor condiţionale medii corespunzătoare atributelor A şi B.

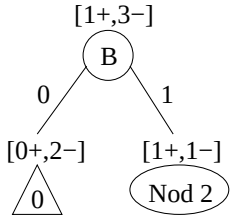
Putem pune în evidenţă acest fapt în figura alăturată, în care simbolul H, scris uneori însoţit de un argument (aşadar, ca $H(p)$), se referă la entropia unei variabile Bernoulli de parametru p.



Mai facem *precizarea* că semnele $<$ şi $=$ din figura de mai sus se referă de fapt nu [doar] la entropiile condiţionale specifice, ci [şi] la produsul acestora cu ponderile asociate în mod corespunzător: $\frac{4}{8}H[1+,3-] = \frac{4}{8}H[3+,1-]$ şi respectiv $\frac{4}{8}H[4+,0-] < \frac{4}{8}H[2+,2-]$.

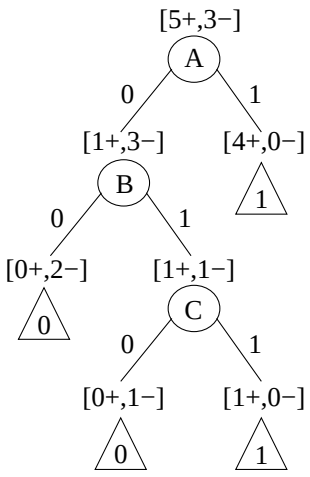
2. Pentru câteva formule de calcul convenabile pentru entropii şi câştiguri de informaţie relative la compaşi de decizie, atunci când se foloseşte calculatorul de buzunar, dar numărul de instanţe de antrenament asociate nu este prea mare, vedeţi problema 35.

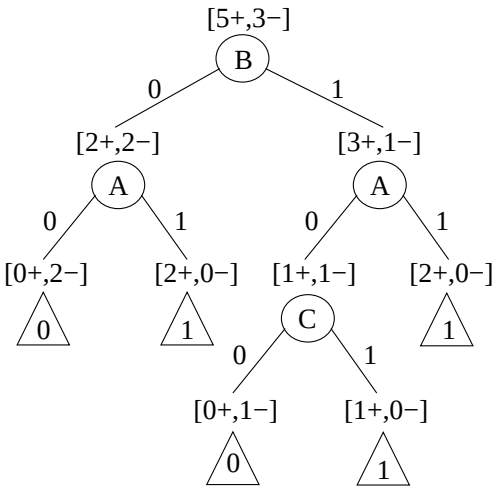
***Nodul 1*: Avem de clasificat instanţele care au $A = 0$ şi putem alege între atributele B şi C. Datorită simetriei, îl putem alege pe oricare dintre ele. Pentru fixare, îl alegem pe B.**



***Nodul 2*: La acest punct a mai rămas disponibil doar atributul C.**

Arborele construit de ID3 este cel reprezentat mai jos, în partea stângă:[378]





[378] Un alt arbore ID3 este cel obţinut din acesta interschimbând atributele B şi C. (Vedeţi *Observaţia* din enunţ.)

b. Pentru a vedea dacă arborele construit de algoritmul ID3 este cel optimal, trebuie să reconsiderăm toate deciziile pe care le-am luat în construirea acestuia:

– La nodul 1 al arborelui avem de clasificat exemplele pentru care $A = 0$, deci funcţia care trebuie reprezentată de subarborele în cauză este $f' = f[A/0] = 0 \vee (B \wedge C) = B \wedge C$, funcţie care este reprezentată în mod optimal de subarborele construit de ID3. (Notaţia $A/0$ semnifică faptul că variabila logică A este instanţiată la valoarea 0.) Prin urmare, nu există un arbore mai bun care să reprezinte funcţia dată şi să aibă în rădăcină atributul A.

– În rădăcină am ales atributul A în detrimentul celorlalte două atribute deoarece am demonstrat că aduce cel mai mare câştig de informaţie. Să vedem ce se întâmplă dacă alegem unul dintre atributele B sau C. După cum am discutat mai sus, datorită simetriei, pe oricare dintre cele două l-am alege, arborele rezultat ar avea aceeaşi formă. Pentru fixare, să-l alegem pe B.

Subarborele stâng şi drept vor trebui să reprezinte funcţiile:

$$f'' = f[B/0] = A \vee (0 \wedge C) = A \vee 0 = A$$

şi respectiv

$$f''' = f[B/1] = A \vee (1 \wedge C) = A \vee C.$$

Arborele minimal care poate fi construit în aceste circumstanţe este cel reprezentat mai sus în partea dreaptă. Acest arbore are 3 niveluri şi 4 noduri, cu un nod în plus faţă de cel construit de algoritmul ID3.

Putem deci conchide că arborele construit respectând specificaţiile algoritmului ID3 este cel optimal.

***Observaţie*:**

Această problemă pune în evidenţă două modalităţi de parcurgere a spaţiului de versiuni pentru un concept, în particular unul din logica propoziţiilor, care este reprezentat cu ajutorul arborilor de decizie. Pe de o parte avem explorarea (incompletă) făcută de algoritmul ID3 care este de tip "greedy", iar pe de altă parte avem explorarea exhaustivă. Prima strategie de explorare procedează la o căutare „orientată" a soluţiei (şi din această cauză este mai eficientă, dar se va vedea, ca revers, că nu asigură întotdeauna găsirea optimului), iar cea de-a doua strategie de explorare, deşi asigură găsirea optimului, nu este utilizabilă în cazurile (frecvente!) în care spaţiul de versiuni este foarte mare.

4. **(ID3, ca algoritm "greedy";
un exemplu când arborele ID3 nu este optimal
ca număr de noduri şi de niveluri)**

prelucrare de Liviu Ciortuz, după
■ ***CMU, 2003 fall, T. Mitchell, A. Moore, midterm exam, pr. 9.a***

Fie atributele binare de intrare A, B, C, atributul de ieşire Y şi următoarele exemple de antrenament:

A	B	C	Y
1	1	0	0
1	0	1	1
0	1	1	1
0	0	1	0

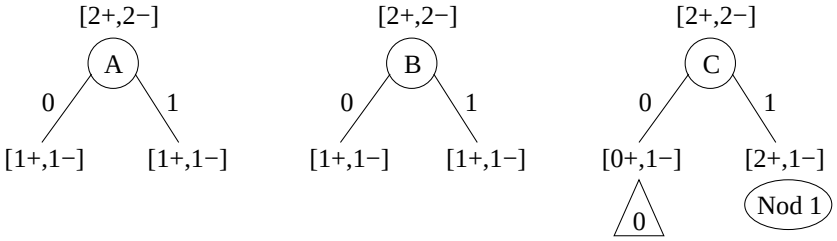
a. Determinaţi arborele de decizie calculat de algoritmul ID3. Este acest arbore de decizie *consistent* cu datele de antrenament?

b. Există un arbore de decizie de adâncime mai mică (decât cea a arborelui ID3) consistent cu datele de mai sus? Dacă da, ce concept (logic) reprezintă acest arbore?

Răspuns:

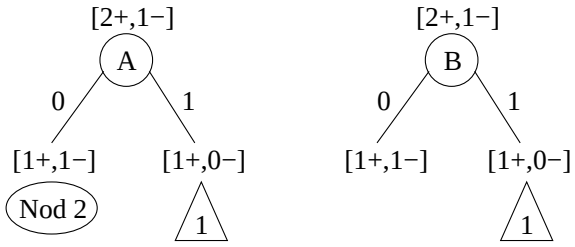
a. Se construieşte arborele de decizie cu algoritmul ID3 astfel:

***Nodul 0* (rădăcina):**



Observăm că $H_{0/A} = H_{0/B} = \frac{2}{4}H[1+,1-] + \frac{2}{4}H[1+,1-] = H[1+,1-] = 1$, care este valoarea maximă a entropiei [condiţionale medii a] unei variabile booleene. Prin urmare, $H_{0/C}$ nu poate fi decât mai mică sau egală cu $H_{0/A}$ şi $H_{0/B}$. Deci vom alege în nodul rădăcină atributul C.

***Nodul 1*: Avem de clasificat instanţele cu $C = 1$, deci alegerea se face între atributele A şi B.**

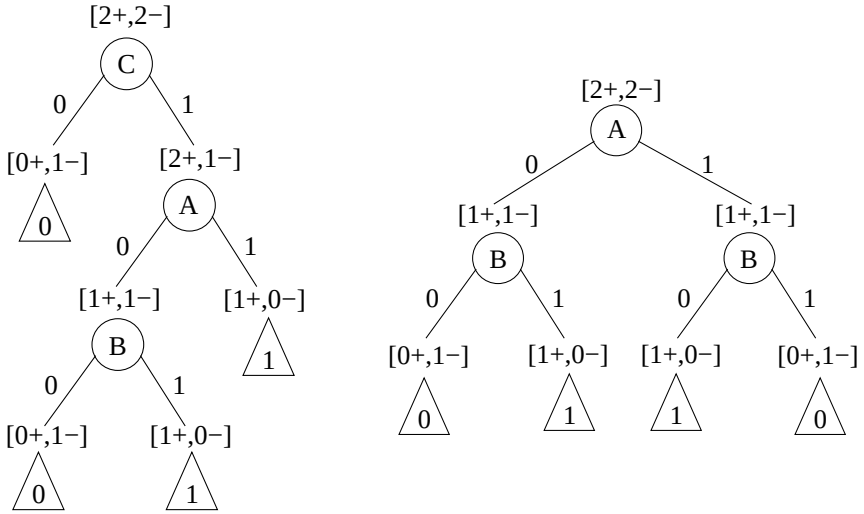


Cele două entropii condiţionale medii sunt egale:

$$H_{1/A} = H_{1/B} = \frac{2}{3}H[1+,1-] + \frac{1}{3}H[1+,0-]$$

Aşadar, putem alege oricare dintre cele două atribute. Pentru fixare, îl alegem pe A.

***Nodul 2*: La acest nod nu mai avem decât atributul B, deci îl vom pune pe acesta. Arborele complet este reprezentat în partea stângă:**



Prin construcţie, arborele ID3 este consistent cu datele de antrenament dacă acestea sunt consistente (i.e., necontradictorii). În cazul nostru, se verifică imediat că datele de antrenament sunt consistente.

b. Se observă că atributul de ieşire Y reprezintă de fapt funcţia logică A XOR B. Reprezentând această funcţie ca arbore de decizie, vom obţine arborele desenat mai sus în partea dreaptă. Acest arbore are cu un nivel mai puţin decât arborele construit cu algoritmul ID3. Prin urmare, arborele obţinut de algoritmul ID3 pe datele din enunţ *nu* este *optim* din punctul de vedere al numărului de niveluri. Aceasta este o *consecinţă* a caracterului "greedy" al algoritmului ID3, datorat faptului că la fiecare iteraţie alegem „cel mai bun" atribut în raport cu criteriul câştigului de informaţie. Se ştie că algoritmii de tip "greedy" nu grantează obţinerea optimului global.

5. (Clasificare ternară: "decision stump" produs de ID3, pe date care conţin duplicări şi „zgomote")

Presupunem că se dau şase date de antrenament (precizate în tabel) pentru o problemă de clasificare cu două atribute binare şi trei clase $Y \in \{1, 2, 3\}$. Se va crea un arbore ID3, bazat pe câştigul de informaţie.

X_1	X_2	Y
1	1	1
1	1	1
1	1	2
1	0	3
0	0	2
0	0	3

a. Calculaţi câştigul de informaţie atât pentru X_1 cât şi pentru X_2. Se va folosi aproximarea $\log_2 3 = 19/12$ şi se va scrie câştigul de informaţie sub formă de fracţii.

b. Pe baza rezultatelor anterioare, ce atribut va fi folosit pentru primul nod al arborelui ID3? Desenaţi arborele de decizie care rezultă folosind doar acest singur nod. Etichetaţi corespunzător nodul, ramurile şi eticheta prevăzută în fiecare frunză.

c. Cum va clasifica acest arbore instanţa determinată de $X_1 = 0$ şi $X_2 = 1$?

Răspuns:

a. Redăm formula pentru calculul câştigului de informaţie în varianta folosită de Tom Mitchell:

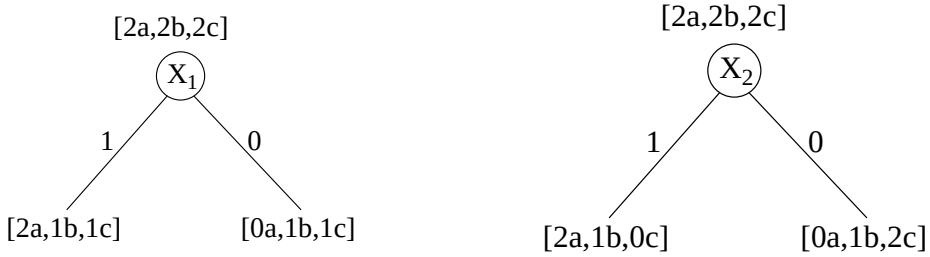
$$Gain(X_i) = Entropy(S) - \sum_{v \in Values(X_i)} \frac{|S_v|}{|S|} Entropy(S_v)$$

unde S este mulţimea celor 6 date de antrenament, iar

$$Entropy(S) = -\sum_{y \in Y} p_y \log_2 p_y$$

Notăm cu a clasa instanţelor având eticheta $Y = 1$, cu b clasa instanţelor cu $Y = 2$ şi cu c clasa instanţelor cu $Y = 3$. În mulţimea S există câte 2 elemente din fiecare clasă şi atunci putem scrie că $S = [2a, 2b, 2c]$.

Pentru nodul rădăcină, putem alege fie atributul X_1, fie atributul X_2, ceea ce determină următoarele împărţiri ale mulţimii S:



Putem calcula câştigurile de informaţie pentru cele două atribute:

$$Gain(X_1) = H[2a, 2b, 2c] - \left(\frac{4}{6} H[2a, 1b, 1c] + \frac{2}{6} H[0a, 1b, 1c] \right)$$

unde $H[2a, 2b, 2c]$ este o altă notaţie pentru entropia mulţimii compuse din două exemple de clasă a, două de clasă b şi două de clasă c.

Calculăm entropiile care intervin în formulă:

$$\begin{aligned}
H[2a, 2b, 2c] &= -\frac{2}{6}\log_2\frac{2}{6} - \frac{2}{6}\log_2\frac{2}{6} - \frac{2}{6}\log_2\frac{2}{6} \\
&= -3 \cdot \frac{2}{6}\log_2\frac{2}{6} = -\log_2\frac{1}{3} = \log_2 3 = \frac{19}{12} \\
H[2a, 1b, 1c] &= -\frac{2}{4}\log_2\frac{2}{4} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{4}\log_2\frac{1}{4} = +\frac{1}{2}\cdot\log_2 2 + \frac{1}{4}\log_2 4 + \frac{1}{4}\log_2 4 \\
&= \frac{1}{2} + \frac{1}{4}\cdot 2 + \frac{1}{4}\cdot 2 = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = \frac{3}{2} \\
H[0a, 1b, 1c] &= 0 - \frac{1}{2}\cdot\log_2 2 - \frac{1}{2}\cdot\log_2 2 = \log_2 2 = 1
\end{aligned}$$

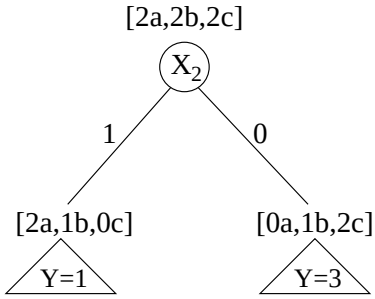
Înlocuind aceste valori numerice în formula câştigului de informaţie, obţinem:

$$Gain(X_1) = \frac{19}{12} - \left(\frac{2}{3}\cdot\frac{3}{2} + \frac{1}{3}\cdot 1 \right) = \frac{19}{12} - \left(1 + \frac{1}{3} \right) = \frac{19}{12} - \frac{4}{3} = \frac{3}{12} = \frac{1}{4}$$

Se aplică aceleaşi formule şi pentru atributul X_2:

$$\begin{aligned} Gain(X_2) &= H[2a,2b,2c] - \left(\frac{3}{6}H[2a,1b,0c] + \frac{3}{6}H[0a,1b,2c]\right) \\ H[2a,1b,0c] &= H[0a,1b,2c] \\ &= -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = -\frac{2}{3}(\log_2 2 - \log_2 3) - \frac{1}{3}(-1)\log_2 3 \\ &= -\frac{2}{3}\cdot 1 + \frac{2}{3}\cdot\frac{19}{12} + \frac{1}{3}\cdot\frac{19}{12} = \frac{19}{12} - \frac{2}{3} = \frac{11}{12} \\ \Rightarrow Gain(X_2) &= \frac{19}{12} - \left(\frac{1}{2}\cdot\frac{11}{12} + \frac{1}{2}\cdot\frac{11}{12}\right) = \frac{19}{12} - \frac{11}{12} = \frac{8}{12} = \frac{2}{3}. \end{aligned}$$

b. Deoarece $Gain(X_1) = \frac{3}{12}$, $Gain(X_2) = \frac{8}{12}$, şi deci $Gain(X_1) < Gain(X_2)$, se va alege atributul X_2 ca rădăcină a arborelui. Arborele de decizie construit din acest singur nod este cel din figura alăturată.



c. O instanţă care are $X_1 = 0$ şi $X_2 = 1$ va fi clasificată de acest arbore cu $Y = 1$ (cu probabilitate $2/3$).

6. **(Algoritmul ID3: aplicare pe date inconsistente, "decision stumps", calculul acurateţii)**

• ○ ***CMU, 2012 fall, T. Mitchell, Z. Bar-Joseph, HW1, pr. 2.ab***

Tabelul de mai jos sumarizează situaţia celor 2201 de pasageri şi membri ai echipajului de la bordul vasului Titanic, în urma naufragiului din data de 15 Aprilie 1912. Pentru fiecare combinaţie de valori ale celor 3 variabile (Clasă, Sex, Vârstă) am indicat în tabel câţi oameni au supravieţuit şi câţi nu au supravieţuit. (*Observaţie*: Datele originale au patru valori pentru atributul Clasă; am comasat valorile II, III, şi Echipaj într-o singură valoare, denumită „Inferioară".)

			Supravieţuitori		
Clasa	Sexul	Vârsta	Nu	Da	Total
I	Masculin	Copil	0	5	5
I	Masculin	Adult	118	57	175
I	Feminin	Copil	0	1	1
I	Feminin	Adult	4	140	144
Inferioară	Masculin	Copil	35	24	59
Inferioară	Masculin	Adult	1211	281	1492
Inferioară	Feminin	Copil	17	27	44
Inferioară	Feminin	Adult	105	176	281
		Total	1490	711	2201

Pentru a vă uşura calculele pe care va trebui să le faceţi, am făcut noi totalurile pentru fiecare variabilă:

	Supravieţuitori		
Clasa	**Nu**	**Da**	**Total**
I	**122**	**203**	**325**
Inferioară	**1368**	**508**	**1876**

	Supravieţuitori		
Sexul	**Nu**	**Da**	**Total**
Masculin	**1364**	**367**	**1731**
Feminin	**126**	**344**	**470**

	Supravieţuitori		
Vârsta	**Nu**	**Da**	**Total**
Copil	**52**	**57**	**109**
Adult	**1438**	**654**	**2092**

a. Folosind un arbore de decizie, dorim să prezicem variabila de ieşire Y (Supravieţuitor), pornind de la atributele de intrare C (Clasa), S (Sexul), V (Vârsta). Utilizaţi criteriul câştigului de informaţie pentru a alege care dintre aceste trei atribute C, S sau V trebuie să fie folosit în nodul-rădăcină al arborelui de decizie.

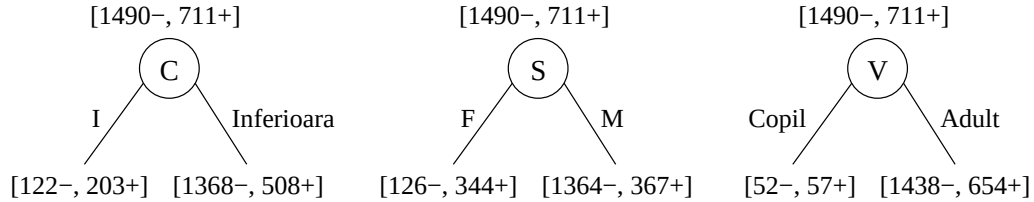
De fapt, ce vi se cere este să învăţaţi un arbore de decizie de adâncime 1 care foloseşte doar atributul din rădăcină pentru a clasifica datele. (Astfel de arbori de decizie de adâncime 1 sunt adesea numiţi în terminologia de limbă engleză "decision stumps".) Parcurgeţi toate etapele rezolvării, redând inclusiv calculele pentru câştigul de informaţie al fiecărui atribut.

b. Care este acurateţea [medie] obţinută pe datele de antrenament de către arborele de decizie cu adâncime 1 de la punctul precedent?

c. Dacă aţi crea un arbore de decizie care foloseşte toate cele trei variabile, care ar fi acurateţea lui [medie] pe datele de antrenament? (*Observaţie*: Nu trebuie neapărat să creaţi arborele de decizie pentru a afla răspunsul!)

Răspuns:

a. Totalurile care au fost furnizate în enunţ pentru fiecare dintre variabilele C, S şi V ne servesc foarte bine pentru a crea rapid cei trei "decision stumps":



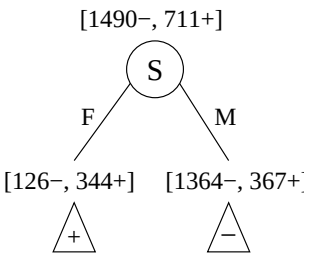
Analizând datele conform figurii de mai sus, se poate „intui" că atributul S va avea un câştig de informaţie (în raport cu atributul de ieşire Y − *Supravieţuitor*) mai bun decât al celorlalte două atribute de intrare (C şi V). Intuiţia se verifică făcând calculele:

$$\begin{aligned} IG(Y,C) &= H[1490-,711+] - \left(\frac{325}{2201}H[122-,203+] + \frac{1876}{2201}H[1368-,508+]\right) \\ &= 0.048501 \\ IG(Y,S) &= H[1490-,711+] - \left(\frac{470}{2201}H[126-,344+] + \frac{1731}{2201}H[1364-,367+]\right) \\ &= 0.142391 \end{aligned}$$

$$IG(Y,V) = H[1490-,711+] - \left(\frac{109}{2201}H[52-,57+] + \frac{2092}{2201}H[1438-,654+]\right)$$
$$= 0.006411.$$

Deci, într-adevăr, câştigul maxim de informaţie se obţine pentru atributul S.

b. Arborele de decizie de adâncime 1 care are în nodul rădăcină atributul S este cel din figura alăturată. Acurateţea [medie a] acestui arbore de decizie este:



$$\frac{470}{2201}\cdot\frac{344}{470} + \frac{1731}{2201}\cdot\frac{1364}{1731} = \frac{344+1364}{2201} = \frac{1708}{2201} = 0.776.$$

c. Se poate constata imediat că arborele ID3 produs pe datele din această problemă va avea 8 noduri-frunză, iar în fiecare dintre aceste noduri-frunză se va asigna câte una dintre mulţimile descrise (pe linie) în coloanele 4 şi 5 ale tabelului principal din enunţ: $[0-,5+]$, $[118-,57+]$, $\ldots$, $[17-,27+]$, $[105-,176+]$. Decizia care va fi luată în fiecare nod-frunză este dictată de votul majoritar, şi anume: $+, -, \ldots, +$ şi respectiv $+$.

Putem calcula acurateţea [medie] astfel:

$$\frac{5+118+1+140+35+1211+27+176}{2201} = \frac{1713}{2201} = 0.778.$$

Se observă că se produce (din păcate) o creştere foarte mică în raport cu acurateţea celui mai bun "decision stump": doar 0.002.

7. **(Algoritmul ID3: cazul când există repetiţii şi inconsistenţe în datele de antrenament; o margine superioară pentru eroarea la antrenare în funcţie de numărul de valori ale variabilei de ieşire)**

CMU, 2002 fall, Andrew Moore, midterm exam, pr. 1.fg

A	B	C	Z
0	0	0	0
0	0	1	0
0	0	1	0
0	1	0	0
0	1	1	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	1
1	1	0	1
1	1	1	0
1	1	1	1

Presupunem că învăţăm un arbore de decizie care să prezică atributul de ieşire Z pornind de la atributele de intrare A, B, C. Se folosesc datele de antrenament din tabelul alăturat.

a. Care va fi eroarea la antrenare pe acest set de date? Exprimaţi răspunsul sub forma fracţiei de înregistrări care vor fi clasificate eronat ($n/12$).

b. Considerăm un arbore de decizie construit pe un set arbitrar de date. Dacă atributul de ieşire este cu valori discrete şi poate lua k valori distincte, care este eroarea de antrenare maximă (exprimată ca fracţie)?

Răspuns:

a. Este uşor de observat că datele de antrenament conţin „inconsistenţe" (contradicţii, relativ la etichetare), şi anume la exemplele $(0,1,1)$ şi $(1,1,1)$. Fiecare dintre aceste exemple sunt etichetate o dată cu 0 şi altă dată cu 1. Prin urmare, jumătate din aceste exemple vor fi clasificate eronat de arborele învăţat de către algoritmul ID3. Eroarea la antrenare va fi deci $\frac{2}{12}$.

b. Vom analiza pe rând mai multe cazuri, care sunt din ce în ce mai generale.

Cazul i: Mai întâi vom calcula eroarea la antrenare pentru cazul în care setul de date de antrenament este compus din k instanţe care sunt identice ca tupluri de valori pentru atributele ce le caracterizează, dar sunt clasificate pe rând cu fiecare din cele k valori posibile ale atributului de ieşire. Este evident că arborele de decizie învăţat va clasifica eronat $k-1$ instanţe. Aşadar, în acest caz, eroarea la antrenare va fi

$$E = \frac{k-1}{k}$$

Cazul *ii*: Aceeaşi eroare [la antrenare] ca mai sus se va înregistra dacă în locul fiecărei instanţe dintre cele considerate la cazul precedent vom avea l instanţe identice, inclusiv în ce priveşte eticheta. (În total sunt kl instanţe de antrenament.)

$$E = \frac{(k-1)\cdot l}{k\cdot l} = \frac{k-1}{k}$$

Cazul *iii*: Dacă relaxăm condiţia de mai sus considerând $l_1, l_2, \ldots, l_k$ instanţe identice, iar $l = \max_{i=1}^{k} l_i$, este imediat că eroarea maximă se va atinge în cazul $l_1 = l_2 = \ldots = l_k = l$, şi va avea aceeaşi valoare ca mai sus. Aşadar, în continuare vom putea renunţa la a considera factorul de multiplicare l, fără ca prin aceasta să restrângem generalitatea raţionamentului.

Cazul *iv*: Fie n exemple de antrenament (instanţe etichetate) dintre care d sunt distincte (ca tupluri de valori ale atributelor de intrare). Fie k_1, k_2 $\cdots$ k_d numărul de instanţe etichetate pentru fiecare caz distinct în parte din cele d. Atunci vom avea:

$$k_1 \le k,\, k_2 \le k, \cdots, k_d \le k \Rightarrow n = k_1 + k_2 + \cdots + k_d \le k \cdot d \textbf{ deci } n \le k \cdot d$$

Eroarea maximă la antrenare va fi dată de formula

$$E = \frac{(k_1 - 1) + (k_2 - 1) + \cdots + (k_d - 1)}{n} = \frac{n-d}{n}$$

Avem:

$$E \le \frac{k-1}{k} \Leftrightarrow \frac{n-d}{n} \le \frac{k-1}{k} \Leftrightarrow k\cdot n - k\cdot d \le k\cdot n - n \Leftrightarrow k\cdot d \ge n \textbf{ (adev.)}$$

Prin urmare, eroarea maximă la antrenare care poate fi atinsă atunci când atributul de ieşire poate lua k valori distincte este $\frac{k-1}{k}$.

8. **(ID3, aspecte computaţionale: influenţa atributelor duplicate, respectiv a instanţelor de antrenament duplicate asupra arborelui ID3 rezultat)**

CMU, 2009 spring, Ziv Bar-Joseph, final exam, pr. 3.1-2

Se doreşte construirea unui arbore de decizie pentru n vectori, cu m atribute.

a. Să presupunem că există i şi j astfel încât pentru TOŢI vectorii X din datele de antrenament, aceste atribute au valori egale (adică, $x_i = x_j$ pentru toţi vectorii, unde x_i este valoarea atributului i în vectorul X). Să presupunem de asemenea că în cazul în care ambele atribute duc la acelaşi câştig de informaţie vom folosi atributul i. Ştergerea atributului j din datele de antrenament poate schimba arborele de decizie obţinut? Explicaţi pe scurt.

b. Să presupunem că există în mulţimea de antrenament doi vectori egali X şi Z (adică, toate atributele lui X şi Z sunt exact la fel, inclusiv etichetele). Ştergerea vectorului Z din datele de antrenament poate schimba arborele de decizie obţinut? Explicaţi pe scurt.

Răspuns:

a. Nu, îndepărtarea atributului j nu schimbă arborele de decizie, deoarece atributele i şi j conduc la valori egale pentru câştigul de informaţie în fiecare nod al arborelui.

b. Da, în acest caz arborele de decizie se poate schimba, fiindcă entropia condiţională — care se calculează în fiecare nod pentru a determina atributul cu câştigul de informaţie cel mai mare — depinde de numărul de instanţe de antrenament luate în considerare.

9. **(Arbori de decizie: o margine superioară pentru numărul de noduri frunză, în funcţie de numărul atributelor şi numărul de exemple)**

CMU, 2005 fall, T. Mitchell, A. Moore, midterm exam, pr. 2.d

Presupunem că învăţăm un arbore de decizie folosind un set de R instanţe de antrenament descrise de M atribute de intrare având valori binare.

Care este numărul maxim posibil de noduri frunză din arborele de decizie, presupunând că fiecărui nod frunză îi este asociat măcar un exemplu de antrenament? Încercuiţi unul din răspunsurile de mai jos; justificaţi alegerea făcută.

R, $\log_2(R)$, R^2, 2^R, M, $\log_2(M)$, M^2, 2^M,
$\min(R, M)$, $\min(R, \log_2(M))$, $\min(R, M^2)$, $\min(R, 2^M)$,
$\min(\log_2(R), M)$, $\min(\log_2(R), \log_2(M))$, $\min(\log_2(R), M^2)$, $\min(\log_2(R), 2^M)$,
$\min(R^2, M)$, $\min(R^2, \log_2(M))$, $\min(R^2, M^2)$, $\min(R^2, 2^M)$,
$\min(2^R, M)$, $\min(2^R, \log_2(M))$, $\min(2^R, M^2)$, $\min(2^R, 2^M)$,
$\max(R, M)$, $\max(R, \log_2(M))$, $\max(R, M^2)$, $\max(R, 2^M)$,
$\max(\log_2(R), M)$, $\max(\log_2(R), \log_2(M))$, $\max(\log_2(R), M^2)$, $\max(\log_2(R), 2^M)$,
$\max(R^2, M)$, $\max(R^2, \log_2(M))$, $\max(R^2, M^2)$, $\max(R^2, 2^M)$,
$\max(2^R, M)$, $\max(2^R, \log_2(M))$, $\max(2^R, M^2)$, $\max(2^R, 2^M)$.

Răspuns:

Notăm cu $\max_{frunze}$ valoarea căutată. Trebuie luate în considerare două aspecte:

(a) fiecare nod frunză trebuie să clasifice măcar un exemplu de antrenament $\Rightarrow$ nu putem să avem mai multe frunze decât exemple de antrenament $\Rightarrow \max_{frunze} \leq R$

(b) fiecare atribut poate fi testat o singură dată pe un drum de la rădăcină la o frunză oarecare $\Rightarrow$ arborele obţinut va avea adâncimea cel mult M $\Rightarrow \max_{frunze} \leq 2^M$

$$\left.\begin{array}{l} \max_{frunze} \leq R \\ \max_{frunze} \leq 2^M \end{array}\right\} \Rightarrow \max_{frunze} \leq \min(R, 2^M)$$

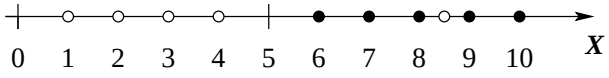
Această valoare poate fi atinsă: putem să luăm, de exemplu, cazul trivial când avem o singură instanţă de antrenament. Prin urmare, avem $\max_{frunze} = \min(R, 2^M)$.

10. **(Extensii ale algoritmului ID3: variabile de intrare continue; "decision stumps"; eroarea la antrenare, eroarea la CVLOO; overfitting)**

■ • ***CMU, 2002 fall, Andrew Moore, midterm exam, pr. 3***

Fie setul de date de mai jos. X este atribut de intrare şi ia valori reale, iar Y este variabilă de ieşire cu valori booleene. (*Observaţie*: Am reprezentat acest set de date sub formă grafică, marcând valoarea / eticheta $Y = 1$ prin bulină neagră, iar valoarea / eticheta $Y = 0$ prin cerculeţ alb.) Pe acest set de date se foloseşte algoritmul ID3 pentru învăţare de arbori de decizie.

X	Y
1	0
2	0
3	0
4	0
6	1
7	1
8	1
8.5	0
9	1
10	1



Algoritmul ID3 (extins) va trebui să decidă cum divide (engl., split) intervale de valori asociate variabilei reale X. Separarea în intervale diferite va fi stabilită cu ajutorul unor *praguri* (engl., split thresholds), determinate în felul următor:

– **Mai întâi se ordonează crescător acele valori ale variabilei X care apar în datele de antrenament.**

– **Se stabilesc apoi perechi de valori consecutive pentru care există instanţe de antrenament care sunt etichetate în mod diferit pentru o valore, comparativ cu cealaltă valoare. Pentru fiecare pereche de valori de acest fel, va fi plasat un prag de separare la jumătatea distanţei dintre cele două valori.**

Iniţial, se alege pragul de separare care conduce la un câştig de informaţie maxim. Apoi, la fiecare nouă execuţie a buclei principale a algoritmului ID3 — vă readucem aminte că acest algoritm este recursiv — se va selecta câte un alt prag dintre cele rămase disponibile, aplicând acelaşi criteriu: maximizarea câştigului de informaţie.

De *exemplu*, pentru $X = 4$ avem o instanţă de antrenament negativă, iar pentru $X = 6$ avem o instanţă de antrenament pozitivă. Se poate arăta că algoritmul ID3 va decide să spliteze mai întâi la valoarea $X = 5$ (care reprezintă jumătatea distanţei dintre $X = 4$ şi $X = 6$) şi apoi la valoarea $X = 8.25$ (care reprezintă jumătatea distanţei dintre $X = 8$ şi $X = 8.5$).

Fie DT* arborele de decizie complet, obţinut de algoritmul ID3 fără a face pruning, iar DT2 arborele de decizie produs de ID3 urmat de pruning, care are doar două noduri frunză (deci DT2 face o singură divizare de interval).

a. Care este *eroarea la antrenare* produsă de DT2 respectiv DT* (exprimată ca număr de exemple clasificate eronat din totalul de 10 exemple)?

b. Care este *eroarea* produsă de DT2 respectiv DT* la *cross-validare* folosind metoda *Leave-One-Out* (CVLOO)?

Răspuns:

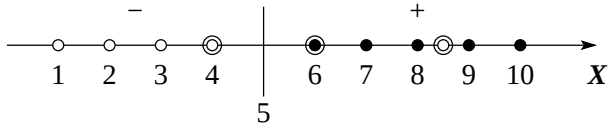
a. Deoarece DT2 reţine doar un singur split, şi anume la $X = 5$, regula de decizie pe care o reprezintă este:

IF $\mathbf{X \leq 5}$
THEN $\mathbf{Y = 0}$
ELSE $\mathbf{Y = 1}$

Este evident că această regulă produce o clasificare eronată doar pentru una dintre instanţele de antrenament, şi anume $X = 8.5$. Avem deci $E_{antren.}(DT2) = 1/10$.

Întrucât exemplele nu conţin inconsistenţe, arborele de decizie ID3 clasifică corect toate datele de antrenament. Avem deci $E_{antren.}(DT*) = 0/10 = 0$.

b. Figura de mai jos reprezintă împărţirea axei reale în *intervale / zone de decizie* conform arborelui (şi *pragului de decizie*) învăţat de către algoritmul DT2, folosind întregul set de exemple date. (Am încercuit acele puncte care, după cum se va vedea mai jos, vor constitui cazuri aparte la calcularea erorii de tip CVLOO cu algoritmul DT2.)



În ce priveşte cross-validarea prin metoda "Leave-One-Out", se poate demonstra — calculele nu sunt arătate aici[379] — următorul fapt: pentru fiecare din cele 10 exemple ($X = 1$, $X = 2$, ..., $X = 10$) considerate pe rând, pragul de decizie identificat de algoritmul DT2 va fi $X = 5$, cu excepţia următoarelor două cazuri:

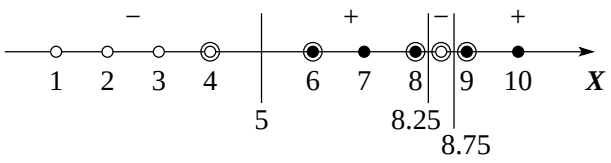
- **$X = 4$: în acest caz, splitarea se va face la mijlocul intervalului $[3, 6]$, deci la 4.5. Cum $4 \leq 4.5$, rezultă că exemplul $X = 4$ va fi clasificat corect;**
- **$X = 6$: splitarea se va face la mijlocul intervalului $[4, 7]$, deci la 5.5. Cum $6 > 5.5$, rezultă că exemplul $X = 6$ va fi clasificat corect.**

Atunci când punctul $X = 8.5$ este lăsat deoparte, pragul de decizie selectat fiind $X = 5$, va rezulta că punctul $X = 8.5$ este clasificat pozitiv (deci eronat), întrucât $8.5 > 5$.

Este imediat că pentru restul de 7 cazuri ($X = 1, 2, 3, 7, 8, 9, 10$), arborele determinat de algoritmul DT2 va clasifica corect punctul X.

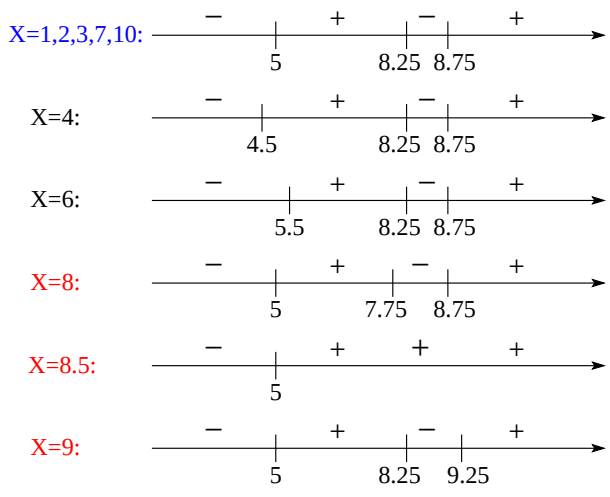
Aşadar, pentru DT2 rezultă $E_{CVLOO} = 1/10$.

Figura de mai jos reprezintă împărţirea axei reale în *intervale / zone de decizie* conform *pragurilor de decizie* (şi arborelui DT*) determinate de către algoritmul ID3 pe întregul set de date de antrenament.[380]



În această figură am încercuit exemplele care ar putea conduce la eroare la cross-validarea de tip "Leave-One-Out":

- **$X = 4$: corect clasificat, explicaţia este aceeaşi ca în cazul DT2;**
- **$X = 6$: idem;**
- **$X = 8$: split-ul trebuie făcut la mijlocul intervalului $[7, 8.5]$, adică 7.75. Cum $8 > 7.75$, rezultă că punctul $X = 8$ va fi clasificat negativ, ceea ce este eronat;**
- **$X = 8.5$: nu este nevoie decât de un singur split, arborele DT* învăţat în acest caz fiind identic cu cel construit de algoritmul DT2. Cum $8 > 5$, rezultă că punctul $X = 8.5$ va fi clasificat pozitiv, deci eronat;**



[379] Aceste calcule nu sunt dificile, mai ales dacă se utilizează raţionamente de tip „calitativ", aşa cum am arătat la problema 3.

[380] *Observaţie*: Atunci când ne raportăm doar la zonele de decizie în ansamblu, ordinea în care s-au stabilit testele în arborele ID3 este irelevantă! Acest fapt este valabil şi mai jos, unde discutăm despre eroarea la CVLOO cu algoritmul DT*.

• $X = 9$**: intervalul de split devine** $[8.5, 10]$**, split-ul făcându-se la** 9.25**. Cum** $9 \leq 9.25$**, rezultă că punctul** $X = 9$ **va fi clasificat negativ, ceea ce este eronat.**

Aşadar, pentru DT* avem $E_{CVLOO} = 3/10$**.**

În *concluzie*, se observă că $E_{antren.}(DT2) = 1/10 > 0 = E_{antren.}(DT*)$**, în vreme ce** $E_{CVLOO}(DT2) = 1/10 < 3/10 = E_{CVLOO}(DT*)$**. Acesta este un caz tipic de manifestare a fenomenului de *overfitting* (*supra-specializare*).**

11. **(Arbori de decizie cu variabile de intrare continue; o margine superioară pentru adâncimea arborilor în cazul (ne)separabilităţii liniare în** $\mathbb{R}^2$**)**

CMU, 2009 spring, Ziv Bar-Joseph, midterm exam, pr. 5.cd

Se consideră n **vectori bidimensionali** $(x = \{x_1, x_2\})$ **care pot fi clasificaţi folosind o funcţie [de regresie] liniară, adică există** $w \in \mathbb{R}^2$ **şi** $b \in \mathbb{R}$ **astfel încât:**

$$y = \begin{cases} +1 & \text{if } w \cdot x + b > 0 \\ -1 & \text{if } w \cdot x + b \leq 0 \end{cases}$$

a. Poate un arbore binar de decizie — atenţie!, nu neapărat arborele ID3 — să clasifice corect aceşti vectori? Dacă nu, justificaţi. Dacă da, determinaţi adâncimea maximă (adică numărul maxim de niveluri de test) ale unui astfel de arbore de decizie, care este optim (ca număr de niveluri de test).

b. Acum să presupunem că aceste n **date nu sunt separabile liniar (adică nu există** $w \in \mathbb{R}^2$ **şi** $b \in \mathbb{R}$ **cu proprietatea de mai sus). Poate un arbore de decizie (binar) să clasifice corect aceşti vectori? Dacă nu, justificaţi. Dacă da, care este adâncimea maximă a unui arbore de decizie corespunzător, optim ca număr de niveluri?**

Răspuns:

a. Da. O strategie posibilă pentru a construi un arbore de decizie binar care să clasifice corect aceste puncte este următoarea:

Mai întâi vom construi un arbore de decizie considerând doar atributul x_1**. În particular, făcând „înjumătăţiri" succesive ale mulţimii de valori ale acestui atribut, arborele rezultat va avea o adâncime maximă de** $\lceil \log_2 n \rceil$ **(adică partea întreagă superioară din** $\log_2 n$**) niveluri. În fiecare nod frunză al acestui arbore se va găsi o mulţime de puncte, însă nu neapărat cu aceeaşi clasificare. Totuşi, deoarece datele sunt liniar separabile, pentru fiecare dintre aceste mulţimi de puncte se poate găsi o valoare a atributului** x_2 **care să o împartă în două submulţimi clasificate corect.**

Prin urmare, arborele construit iniţial considerând doar valoarea x_1 **are nevoie să i se adauge doar cel mult câte un nod în fiecare frunză, nod care să clasifice corect datele luând în considerare valoarea** x_2**. Adâncimea totală a arborelui astfel construit este** $1 + \lceil \log_2 n \rceil$**, adică de ordinul** $O(\log n)$**.**

b. Da. Similar punctului anterior, se construieşte un arbore de decizie considerând doar atributul x_1**, ceea ce înseamnă o adâncime de** $\lceil \log_2 n \rceil$**. În fiecare**

nod frunză al acestui arbore se va găsi o mulţime de puncte, posibil cu clasificări diferite. Datele de antrenament nu mai sunt liniar separabile, deci nu pot fi clasificate corect printr-un singur nod.

Pentru fiecare astfel de nod frunză în care punctele nu au aceeaşi clasificare, se aplică algoritmul de determinare a arborelui de decizie, de această dată luând în considerare doar valoarea lui x_2. Aceasta presupune din nou o adâncime maximă a subarborelui de $\lceil \log_2 n \rceil$. În total, arborele obţinut are adâncimea maximă $\lceil \log_2 n \rceil + \lceil \log_2 n \rceil$, deci tot de ordinul $O(\log n)$.

12. (Algoritmul ID3 cu atribute discrete, respectiv atribute discrete şi un atribut continuu: aplicare; predicţie)

■ • ***CMU, 2012 fall, E. Xing, A. Singh, HW1, pr. 1.1***

Până în luna septembrie a anului 2012, 800 de planete extrasolare (numite în continuare exoplanete) au fost identificate în galaxia noastră. Nişte navete spaţiale super-secrete au fost trimise pentru a survola toate aceste exoplanete, cu scopul de a stabili dacă ele sunt locuibile sau nu de către oameni. Evident, a trimite câte o navetă spaţială la fiecare dintre aceste exoplanete este extrem de costisitor. De aceea, în această problemă vă propunem să elaboraţi un arbore de decizie pentru a prezice dacă o exoplanetă este locuibilă sau nu, folosind doar trăsături / caracteristici (engl., features) observabile cu ajutorul telescoapelor terestre.

a. În tabelul de mai jos vi se dau anumite date în legatură cu toate cele 800 de planete survolate până acum. Trăsăturile observate cu ajutorul telescoapelor sunt *Size* ("Big" sau "Small") şi *Orbit* ("Near" sau "Far").

Fiecare linie din tabel indică valori ale acestor două trăsături, caracterul habitabil ("Yes" sau "No"), precum şi de câte ori a fost identificată fiecare combinaţie de valori [pentru cele trei trăsături]. De exemplu, au fost identificate 20 de planete mari ("Big"), care sunt situate pe orbite apropiate ("Near") de soarele / steaua lor şi sunt locuibile.

Size	Orbit	Habitable	Count
Big	Near	Yes	20
Big	Far	Yes	170
Small	Near	Yes	139
Small	Far	Yes	45
Big	Near	No	130
Big	Far	No	30
Small	Near	No	11
Small	Far	No	255

Elaboraţi şi desenaţi arborele de decizie învăţat de către algoritmul ID3 pe aceste date. (Folosiţi criteriul câştigului de informaţie; nu aplicaţi pruningul.) La fiecare nod din arbore, scrieţi numărul de planete locuibile şi respectiv nelocuibile din datele de antrenament care sunt asociate cu nodul respectiv.

b. Pentru doar 9 dintre aceste exoplanete, a fost măsurată o a treia trăsătură, *Temperature* (exprimată în grade Kelvin), după cum se arată în tabelul de mai jos.

Refaceţi toţi paşii de la punctul a, de data aceasta folosind toate cele trei trăsături de intrare. Pentru trăsătura *Temperature* (văzută ca atribut numeric cu valori continue), la fiecare iteraţie va trebui să faceţi maximizarea în raport cu toate pragurile adecvate pentru separare binară (engl., binary thresholding splits). Iată un *exemplu* de test pentru o astfel de separare binară: $T \leq 250$ vs. $T > 250$.

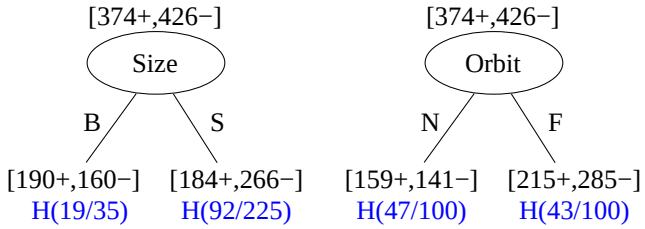
Size	Orbit	Temperature	Habitable
Big	Far	205	No
Big	Near	205	No
Big	Near	260	Yes
Big	Near	380	Yes
Small	Far	205	No
Small	Far	260	Yes
Small	Near	260	Yes
Small	Near	380	No
Small	Near	380	No

c. Conform arborelui de decizie pe care l-aţi obţinut la punctul b, cum va fi clasificată o planetă având trăsăturile (Big, Near, 280), locuibilă sau nelocuibilă?

***Indicaţie*: Este posibil să aveţi nevoie de următoarele valori pentru entropia ($H(p)$) unei variabile aleatoare Bernoulli de parametru p: $H(1/3) = 0.9182$, $H(2/5) = 0.9709$, $H(92/225) = 0.9759$, $H(43/100) = 0.9858$, $H(16/35) = 0.9946$, $H(47/100) = 0.9974$.**

Răspuns:

a. „Compaşii de decizie“ corespunzători nodului rădăcină sunt înfăţişaţi în desenul următor:



Sub fiecare nod descendent am notat entropia nodului respectiv, făcând referire la distribuţia Bernoulli şi dând (de fiecare dată) parametrului acestei distribuţii valoarea corespunzătoare. Pentru nodul descendent corespunzător lui Orbit = Near am ţinut cont şi de faptul că entropia distribuţiei Bernoulli, ca funcţie de parametru p, este simetrică faţă de valoarea $1/2$.[381]

Entropiile condiţionale medii corespunzătoare acestor doi „compaşi de decizie“ sunt:

$$\begin{aligned} H(Habitable|Size) &= \frac{35}{80} \cdot H\left(\frac{19}{35}\right) + \frac{45}{80} \cdot H\left(\frac{92}{225}\right) \\ &= \frac{35}{80} \cdot 0.9946 + \frac{45}{80} \cdot 0.9759 = 0.9841 \end{aligned}$$

[381] Este util să revedeţi explicaţiile date în *observaţia importantă* de la pagina 420.

$$H(Habitable|Orbit) = \frac{3}{8} \cdot H\left(\frac{47}{100}\right) + \frac{5}{8} \cdot H\left(\frac{43}{100}\right)$$
$$= \frac{3}{8} \cdot 0.9974 + \frac{5}{8} \cdot 0.9858 = 0.9901$$

Suntem acum în măsură să desemnăm atributul care va fi plasat în nodul rădăcină al arborelui care va fi construit de algoritmul ID3 pe datele din enunţ: este atributul *Size*, întrucât $H(Habitable|Size) < H(Habitable|Orbit)$**.**

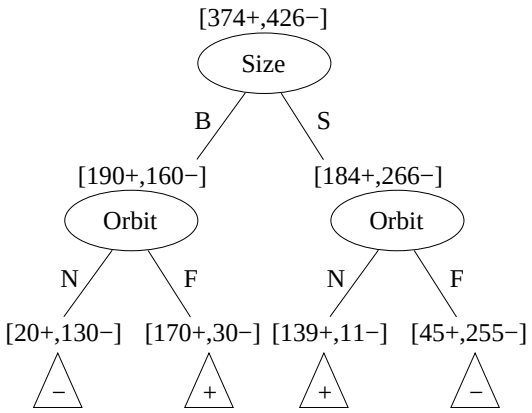
Doar cu titlu de informare, precizăm şi câştigurile de informaţie realizate de cele două atribute de intrare în raport cu atributul de ieşire:

$$IG(Habitable; Size) = H(Habitable) - H(Habitable|Size) = 0.9969 - 0.9841 = 0.0128$$

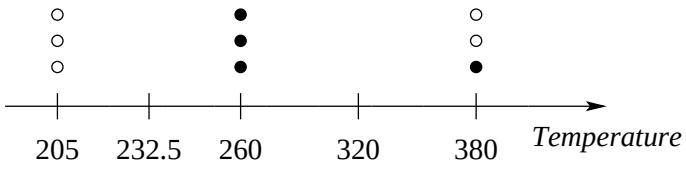
şi, similar,

$$IG(Habitable; Orbit) = 0.0067.$$

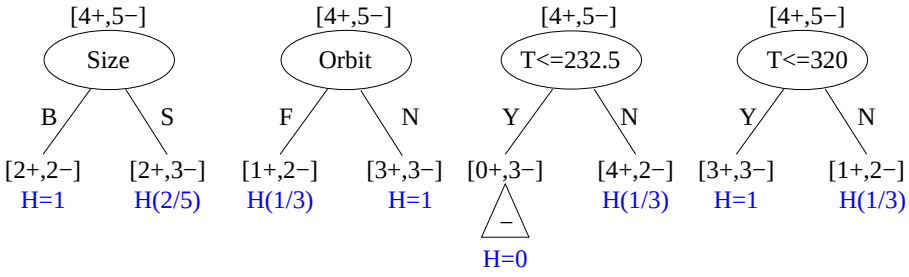
Întrucât nu avem decât două atribute de intrare, arborele de decizie va fi:



b. Pragurile de separare corespunzătoare atributului *Temperature* **sunt determinate conform imaginii următoare:**



Nivelul 1 **(rădăcina):**
„Compaşii de decizie“ corespunzători acestui nivel sunt:



Observând cu atenţie partiţiile formate,[382] vom constata că entropia condiţională medie pentru testul $Temperature \leq 232.5$ **(în raport cu atributul de ieşire** $Habitable$**) are o valoare mai mică decât fiecare dintre entropiile condiţionale medii** $H(Habitable|Orbit)$ **şi** $H(Habitable|\, Temperature \leq 320)$ **(care, de fapt, sunt egale între ele).[383]**

Relaţia dintre $H(Habitable|\, Temperature \leq 232.5)$ **şi** $H(Habitable|Size)$ **se determină cu ajutorul calculelor:**

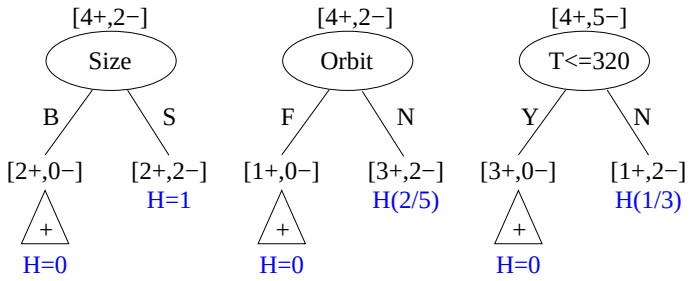
$$\begin{aligned} H(Habitable|Size) &= \frac{4}{9} + \frac{5}{9} \cdot H\left(\frac{2}{5}\right) = \frac{4}{9} + \frac{5}{9} \cdot 0.9709 = 0.9838 \\ H(Habitable|\, Temp \leq 232.5) &= \frac{2}{3} \cdot H\left(\frac{1}{3}\right) = \frac{2}{3} \cdot 0.9182 = 0.6121. \end{aligned}$$

Deşi nu mai este necesar, indicăm şi valorile câştigurilor de informaţie:

$$\begin{aligned} IG(Habitable; Size) &= 0.0072 \\ IG(Habitable; Orbit) &= 0.0183 \\ IG(Habitable; Temp \leq 232.5) &= 0.3788 \\ IG(Habitable; Temp \leq 320) &= 0.0183. \end{aligned}$$

Prin urmare, vom reţine pentru nivelul 0 testul $Temperature \leq 232.5$**).**

Nivelul 2 **(mai exact, aici ne limităm la datele cu** $Temperature > 232.5$**: „Compaşii de decizie“ corespunzători [completării] acestui nivel sunt:**



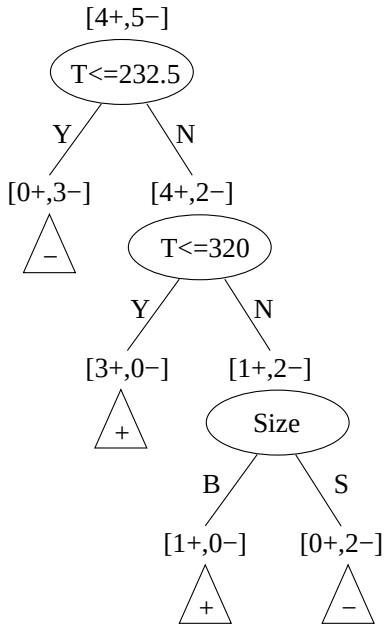
Este imediat că, pe aceste date, $H(Habitable|\, Temp \leq 320) < H(Habitable|Size)$ **şi, de asemenea,** $H(Habitable|\, Temp \leq 320) < H(Habitable|Orbit)$**.[384] Prin urmare, aici va fi ales testul** $Temp \leq 320$**.**

[382] Se compară două câte două entropiile condiţionale specifice, precum şi ponderile datelor respective în [raport cu] numărul total de instanţe asociate nodului părinte.

[383] Observaţi de exemplu că $H(Habitable|Orbit = F) > H(Habitable|\, Temperature \leq 232.5)$ şi $H(Habitable|Orbit = N > H(Habitable|\, Temperature > 232.5)$, iar ponderile corespunzătoare acestor entropii condiţionale specifice în calculul entropiilor condiţionale medii $H(Habitable|Orbit = N)$ şi $Temperature \leq 232.5$ sunt egale două câte două (şi anume, cu 3/9 şi respectiv 6/9).

[384] Mai exact, ar fi trebuit să scriem: $H(Habitable|\, Temp > 232.5, Temp \leq 320) < H(Habitable|\, Temp > 232.5, Size)$ şi respectiv $H(Habitable|\, Temp > 232.5, Temp \leq 320) < H(Habitable|\, Temp > 232.5, Orbit)$.

Nivelul 3 **(mai exact, aici ne limităm la datele cu** $Temperature > 320$**):**
Se poate observa că, pentru aceste date, atributul *Size* are putere discriminativă maximă. Aşadar, arborele de decizie final va fi cel din figura alăturată.



c. Conform arborelui de decizie obţinut la punctul anterior, o exoplanetă având trăsăturile (Big, Near, 280) va fi clasificată ca fiind locuibilă.

13. **(Extensii ale algoritmului ID3: cazul atributelor discrete cu număr mare de valori)**

CMU, 2008(?) spring, HW2, pr. 1

Se doreşte antrenarea unui arbore de decizie care să clasifice exemple cu două atribute de intrare X_1, X_2 **şi un atribut de ieşire** Y **care are valorile** 1 **şi** 2**. Primul atribut,** X_1**, este binar, pe când al doilea atribut,** X_2**, are 6 valori posibile** A, B, C, D, E, F**. Se dau următoarele 12 exemple de antrenament, câte 6 din fiecare clasă:**

$Y = 1$	$(1, A)$	$(0, E)$	$(1, B)$	$(1, B)$	$(1, F)$	$(0, D)$
$Y = 2$	$(0, A)$	$(0, C)$	$(1, E)$	$(0, F)$	$(0, B)$	$(1, D)$

Vă reamintim formula câştigului de informaţie la partajarea mulţimii de exemple S **în funcţie de valorile atributului** A**:**

$$Gain(S, A) = H(S) - H(S \mid A), \text{ cu } H(S \mid A) = \sum_{v \in valori(A)} \frac{|S_v|}{|S|} \cdot H(S_v),$$

unde $H(S)$ **este entropia mulţimii de exemple** S**, iar** $H(S \mid A)$ **este entropia condiţională medie a mulţimii** S **în raport cu atributul** A**, calculată aşa cum se vede mai sus, ca sumă ponderată a entropiilor submulţimilor lui** S **determinate de valorile atributului** A**.**

a. Determinaţi atributul ales în rădăcină folosind câştigul de informaţie. Folosiţi aproximarea $\log_2 3 = 1.585$**.**

b. Determinaţi atributul ales în nodul rădăcină folosind o măsură numită *gain ratio impurity*, adică alegeţi acel atribut care maximizează raportul

$$\frac{Gain(S, A)}{-\sum_v P(A = v) \cdot \log_2 P(A = v)} = \frac{H(S) - H(S \mid A)}{H(A)}.$$

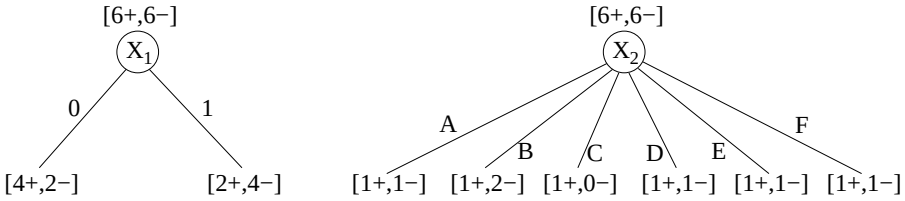
c. Având în vedere rezultatele de la punctele precedente, analizaţi utilitatea folosirii măsurii *gain ratio impurity* în cazurile în care atributele au numere diferite de valori posibile.

Răspuns:

a. Calculăm în primul rând entropia nodului rădăcină (sau a atributului de ieşire), care are 6 exemple pozitive şi 6 negative:

$$H(Y) = H[6+, 6-] = -\frac{6}{12}\log_2\frac{6}{12} - \frac{6}{12}\log_2\frac{6}{12} = 2\cdot\frac{1}{2}\log_2 2 = 1$$

În rădăcină poate fi ales fie atributul X_1, fie atributul X_2, obţinând următoarele împărţiri:



Dacă vom pune atributul X_1 în nodul rădăcină, câştigul de informaţie va fi:

$$\begin{aligned}
Gain(S, X_1) &= H(Y) - \frac{6}{12}H[4+, 2-] - \frac{6}{12}H[2+, 4-] = \\
&= 1 - \frac{6}{12}\left(-\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}\right) - \frac{6}{12}\left(-\frac{2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6}\right) \\
&= 1 - 2\cdot\frac{1}{2}\left(\frac{1}{3}\log_2 3 + \frac{2}{3}\log_2\frac{3}{2}\right) = 1 - \left(\log_2 3 - \frac{2}{3}\right) \\
&= \frac{5}{3} - \log_2 3 \approx 0.0817
\end{aligned}$$

Altminteri, punând atributul X_2 în nodul rădăcină, câştigul de informaţie este:

$$\begin{aligned}
Gain(S, X_2) &= H(Y) - 4\cdot\frac{2}{12}H[1+, 1-] - \frac{3}{12}H[1+, 2-] - \frac{1}{12}H[1+, 0-] \\
&= 1 - \frac{2}{3}\cdot 1 - \frac{3}{12}\left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) - \frac{1}{12}\cdot 0 \\
&= 1 - \frac{2}{3} - \frac{1}{4}\left(\frac{1}{3}\log_2 3 + \frac{2}{3}\log_2\frac{3}{2}\right) = 1 - \frac{2}{3} - \frac{1}{4}\left(\log_2 3 - \frac{2}{3}\right) \\
&= \frac{1}{2} - \frac{1}{4}\log_2 3 \approx 0.1037
\end{aligned}$$

Folosind drept criteriu de optimizat în fiecare nod câştigul de informaţie, în rădăcină vom pune atributul X_2, întrucât $Gain(S, X_1) < Gain(S, X_2)$.

b. Dacă vom pune atributul X_1 în nodul rădăcină, *gain ratio impurity* va fi:

$$\frac{Gain(S, X_1)}{-\sum_{i\in\{0,1\}} P(X_1 = i)\cdot\log_2 P(X_1 = i)} = \frac{Gain(S, X_1)}{-\frac{6}{12}\log_2\frac{6}{12} - \frac{6}{12}\log_2\frac{6}{12}} = \frac{Gain(S, X_1)}{1}$$
$$\approx 0.0817$$

În schimb, plasând atributul X_2 în nodul rădăcină, *gain ratio impurity* va fi:

$$\frac{Gain(S,X_2)}{-\sum_{j\in\{A,\ldots,F\}} P(X_2=j)\cdot\log_2 P(X_2=j)} = \frac{Gain(S,X_2)}{-4\cdot\frac{2}{12}\log_2\frac{2}{12}-\frac{3}{12}\log_2\frac{3}{12}-\frac{1}{12}\log_2\frac{1}{12}}$$

$$= \frac{Gain(S,X_2)}{\frac{2}{3}\log_2 6+\frac{1}{4}\log_2 4+\frac{1}{12}\log_2 12} = \frac{Gain(S,X_2)}{\frac{4}{3}+\frac{3}{4}\log_2 3}\approx 0.0411$$

Prin urmare, în nodul rădăcină vom pune atributul X_1, pentru care măsura *gain ratio impurity* are valoarea cea mai mare.

c. Câştigul de informaţie favorizează alegerea atributelor care au un număr mare de valori, indiferent dacă ele determină sau nu partiţionarea în mod semnificativ a datelor de antrenament. În schimb, măsura *gain ratio impurity* ia în considerare, prin cantitatea de la numitor (vedeţi definiţia), numărul de valori ale atributului respectiv, mai exact mărimea mulţimilor de instanţe asignate nodurilor-fii, care au fost generate ca urmare a alegerii respectivului atribut. Valoarea de la numitor va creşte odată cu numărul de noduri-fii, şi totodată cu numărul de noduri-fii care au asignate puţine exemple. Prin urmare, această măsură penalizează atributele cu multe valori, evitând favorizarea de care se face vinovat câştigul de informaţie într-o atare situaţie.

14. **(Extensii ale algoritmului ID3: cazul când se ia în considerare costul calculării atributelor; atribute continue; "decision stumps")**

CMU, 2008 spring, T. Mitchell, W. Cohen, HW1, pr. 1

Se dă următorul set de date. Acesta reprezintă fişele a 12 pacienţi ipotetici, ţinând cont de sex, vârstă (peste 60 ani sau nu), dacă suferă sau nu de diabet, dacă au pulsul mărit (sau nu) şi EKG-ul anormal (sau nu). Pacienţii sunt clasificaţi în final după cum prezintă (sau nu) aritmie.

Pacient	Sex	Peste60	Diabetic	Puls	EKG	AreAritmie
1	M	1	1	0	0	0
2	M	0	0	1	1	1
3	M	0	1	1	0	0
4	M	1	0	0	1	1
5	M	1	1	1	0	1
6	M	0	1	1	0	1
7	F	0	0	1	0	0
8	F	1	1	1	1	1
9	F	0	1	0	1	1
10	F	1	0	0	0	0
11	F	1	1	0	0	0
12	F	1	0	1	1	1

a. Calculaţi entropia condiţională specifică $H(\text{AreAritmie} \mid \text{Sex} = F)$.

b. Dacă pentru selectarea atributelor se foloseşte măsura $\frac{Gain^2(S,A)}{Cost(A)}$ în schimbul câştigului informational, care va fi atributul pus în nodul rădăcină? Se

consideră că $Cost(\text{Sex}) = Cost(\text{Peste60}) = \mathbf{1}$, $Cost(\text{Diabetic}) = \mathbf{3}$, $Cost(\text{Puls}) = \mathbf{2}$ **şi** $Cost(\text{EKG}) = \mathbf{5}$.

***Observaţie*: În calcule se vor utiliza următoarele aproximări:** $\log_2 3 = 1.585$, $\log_2 5 = 2.322$ **şi** $\log_2 7 = 2.807$.

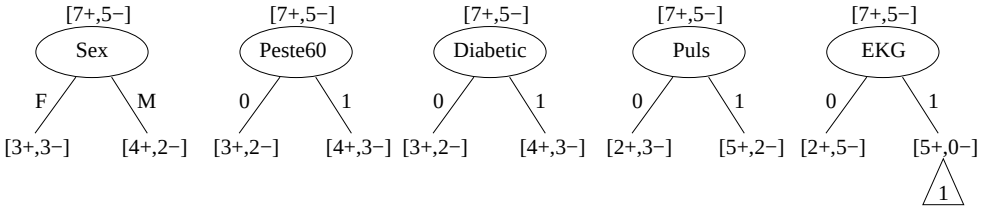
c. Să presupunem că, pentru un alt set de pacienţi, se cunoaşte vârsta lor exactă. Pentru exemplele pozitive, vârstele sunt: {40, 60, 62, 64, 70, 74, 75, 82}, iar pentru exemplele negative sunt: {33, 35, 42, 45, 49, 52, 58, 59, 80}. Să presupunem că toate celelalte atribute sunt predictori „slabi", prin urmare dorim ca arborele să aibă un singur nod, rădăcina, care să împartă exemplele cu valori continue ale atributului vârstă în două: $\boldsymbol{vârstă} < k$ **şi** $\boldsymbol{vârstă} \geq k$. **Care va fi valoarea aleasă pentru** k**, bazat pe câştigul de informaţie?**

Răspuns:

a. Entropia condiţională cerută este:

$$H(\text{AreAritmie} \mid \text{Sex} = F) = H[3+, 3-] = 1$$

b. În rădăcina arborelui de decizie se alege atributul A **pentru care raportul** $\frac{Gain^2(S,A)}{Cost(A)}$ **este maxim. În cazul nostru, variantele pe care le avem sunt:**



Se observă direct că $Gain(\text{S,EKG})$ **este mai mare decât** $Gain(S,A)$ **pentru orice atribut** $\boldsymbol{A} \neq \text{EKG}$. **Însă şi** $Cost(\text{EKG})$ **este mai mare decât** $Cost(A)$ **pentru orice** $A \neq \text{EKG}$. **Aşadar, trebuie să facem calculele în detaliu.**

Entropia atributului de ieşire, AreAritmie**, este:**

$$\begin{aligned} H(\text{AreAritmie}) = H[7+, 5-] &= \frac{7}{12} \cdot \log_2 \frac{12}{7} + \frac{5}{12} \cdot \log_2 \frac{12}{5} \\ &= \frac{7}{12} \cdot \log_2 12 - \frac{7}{12} \cdot \log_2 7 + \frac{5}{12} \cdot \log_2 12 - \frac{5}{12} \cdot \log_2 5 \\ &= \log_2(3 \cdot 4) - \frac{7}{12} \cdot \log_2 7 - \frac{5}{12} \cdot \log_2 5 \\ &= \log_2 3 + \underbrace{\log_2 4}_{=2} - \frac{7}{12} \cdot \log_2 7 - \frac{5}{12} \cdot \log_2 5 \approx 0.98 \end{aligned}$$

Vom calcula câştigul de informaţie pentru fiecare din cele 5 atribute — se observă că pentru atributele Peste60 **şi** Diabetic**, câştigurile de informaţie sunt egale —, şi apoi vom face raportul** $\frac{Gain^2}{Cost}$**:**

Pentru atributul Sex**:**

$$\begin{aligned} Gain(S, \text{Sex}) &= H(\text{AreAritmie}) - \frac{6}{12} H[3+, 3-] - \frac{6}{12} H[4+, 2-] \\ &= H(\text{AreAritmie}) - \frac{1}{2} \cdot 1 - \frac{1}{2} \left(-\frac{4}{6} \log_2 \frac{4}{6} - \frac{2}{6} \log_2 \frac{2}{6} \right) \end{aligned}$$

$$= 0.98 - \frac{1}{2} - \frac{1}{2}\left(\log_2 3 - \frac{2}{3}\right) = 0.98 - \frac{1}{2} - \frac{1}{2}\log_2 3 + \frac{1}{3} \approx 0.02$$

Prin urmare, $\frac{Gain^2(S, \text{Sex})}{Cost(\text{Sex})} \approx \frac{0.02^2}{1} = 0.0004$**.**

Pentru atributele Peste60 **şi** Diabetic**:**

$$\begin{aligned} Gain(S, \text{Peste60}) &= Gain(S, \text{Diabetic}) \\ &= H(\text{AreAritmie}) - \frac{5}{12}H[3+, 2-] - \frac{7}{12}H[4+, 3-] \\ &= 0.98 - \frac{5}{12}\left(\frac{3}{5}\log_2\frac{5}{3} + \frac{2}{5}\log_2\frac{5}{2}\right) - \frac{7}{12}\left(\frac{4}{7}\log_2\frac{7}{4} + \frac{3}{7}\log_2\frac{7}{3}\right) \\ &= 0.98 - \frac{5}{12}\left(\log_2 5 - \frac{3}{5}\log_2 3 - \frac{2}{5}\cdot 1\right) - \frac{7}{12}\left(\log_2 7 - \frac{4}{7}\cdot 2 - \frac{3}{7}\log_2 3\right) \\ &= 0.98 - \frac{5}{12}\log_2 5 - \frac{7}{12}\log_2 7 + \frac{1}{2}\log_2 3 + \frac{5}{6} \approx 0.0009 \end{aligned}$$

Deci $\frac{Gain^2(S, \text{Peste60})}{Cost(\text{Peste60})} \approx \frac{0.0009^2}{1} = 81 \cdot 10^{-8}$ **şi** $\frac{Gain^2(S, \text{Diabetic})}{Cost(\text{Diabetic})} \approx \frac{0.0009^2}{3} = 27 \cdot 10^{-8}$**, ambele fiind nişte valori foarte mici.**

Pentru atributul Puls**:**

$$\begin{aligned} Gain(S, \text{Puls}) &= H(\text{AreAritmie}) - \frac{5}{12}H[2+, 3-] - \frac{7}{12}H[5+, 2-] \\ &= 0.98 - \frac{5}{12}\left(\frac{2}{5}\log_2\frac{5}{2} + \frac{3}{5}\log_2\frac{5}{3}\right) - \frac{7}{12}\left(\frac{5}{7}\log_2\frac{7}{5} + \frac{2}{7}\log_2\frac{7}{2}\right) \\ &= 0.98 - \frac{7}{12}\log_2 7 + \frac{1}{4}\log_2 3 + \frac{1}{3} \approx 0.072 \end{aligned}$$

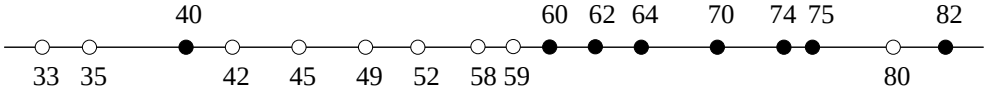
Prin urmare, $\frac{Gain^2(S, \text{Puls})}{Cost(\text{Puls})} \approx \frac{0.072^2}{2} = 0.002592$**.**

Pentru atributul EKG**:**

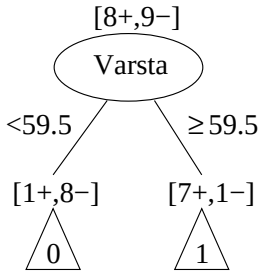
$$\begin{aligned} Gain(S, \text{EKG}) &= H(\text{AreAritmie}) - \frac{7}{12}H[2+, 5-] - \frac{5}{12}H[5+, 0-] \\ &= H(\text{AreAritmie}) - \frac{7}{12}\left(\frac{2}{7}\log_2\frac{7}{2} + \frac{5}{7}\log_2\frac{7}{5}\right) - \frac{5}{12}\cdot 0 \\ &= 0.98 + \frac{5}{12}\log_2 5 - \frac{7}{12}\log_2 7 + \frac{1}{6} \approx 0.476 \end{aligned}$$

Deci $\frac{Gain^2(S, \text{EKG})}{Cost(\text{EKG})} \approx \frac{0.476^2}{5} = 0.0453152$**. Este evident că aceasta este cea mai mare valoare, de aceea în nodul rădăcină va fi ales atributul *EKG*, deşi costul acestuia este cel mai mare.**

c. Putem reprezenta exemplele astfel:



Se observă că alegerea cea mai bună este $k = 59.5$, arborele de decizie fiind cel din figura alăturată.



15. **(Alte criterii posibile pentru selecţia atributelor în ID3: Gini impurity şi Misclassification impurity)**

prelucrare de Liviu Ciortuz, după
■ □ • ***CMU, 2003 fall, T. Mitchell, A. Moore, HW1, pr. 4***

Entropia **este o mărime care cuantifică gradul de neomogenitate (engl., impurity) al unui set de instanţe în raport cu etichetele asignate. Algoritmul ID3 foloseşte entropia drept criteriu de partiţionare (engl., splitting criterion), calculând *câştigul de informaţie* pentru a decide care este atributul care trebuie testat în nodul curent. Există, însă, şi alte măsuri de neomogenitate care pot fi folosite, de asemenea, drept criterii de partiţionare. În această problemă, vom investiga două astfel de măsuri.**

Presupunem că nodul curent (n) din arborele de decizie aflat în curs de elaborare are asignate instanţe din k clase: $c_1, c_2, \ldots, c_k$. Definim

Gini Impurity: $i(n) = 1 - \sum_{i=1}^{k} P^2(c_i)$

şi

Misclassification Impurity: $i(n) = 1 - \max_{i=1}^{k} P(c_i)$,

unde am notat cu $P(c_i)$ probabilitatea [sau: frecvenţa de apariţie a instanţelor aparţinând] clasei c_i în ansamblul instanţelor asignate la nodul curent.

a. Presupunem $k = 2$. Aşadar, în acest caz nodul n are două clase: c_1 şi c_2. Desenaţi un grafic în care cele trei măsuri de neomogenitate — *Entropia*, *Gini Impurity* şi *Misclassification Impurity* — sunt reprezentate în funcţie de $P(c_1)$.

b. Acum putem da definiţia unui nou criteriu de partiţionare, bazat pe măsurile de neomogenitate *Gini* şi *Misclassification*. În literatura de specialitate, acest nou criteriu este denumit uneori *Drop-of-Impurity* (pentru care propunem ca traducere în limba română termenul de *diminuarea neomogenităţii*). El reprezintă diferenţa dintre neomogenitatea nodului curent pe de o parte şi suma ponderată a neomogenităţilor fiilor pe de altă parte. În cazul partiţionării atributelor binare, definim *Drop-of-Impurity* ca fiind:

$$\Delta i(n) = i(n) - P(n_l)\, i(n_l) - P(n_r)\, i(n_r),$$

unde n_l şi n_r reprezintă fiul-stânga şi, respectiv, fiul-dreapta, care au fost derivaţi din nodul n după partiţionare.

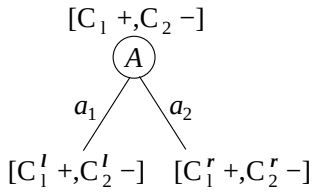
Folosind mai întâi *Gini Impurity* şi apoi *Misclassification Impurity*, calculaţi *Drop-of-Impurity* pentru următorul set de instanţe asignate nodului pentru care se

testează atributul A luând valorile a_1 şi a_2. Am notat cu C variabila de ieşire (desemnând clasa) şi cu c_1 şi c_2 cele două valori ale ei.

A	a_1	a_1	a_1	a_2	a_2	a_2
C	c_1	c_1	c_2	c_2	c_2	c_2

c. Se poate crea un set de date de antrenament (sau: se poate modifica setul de date de mai sus) astfel încât, pe noul set, *Drop-of-Impurity* **bazat pe** *Misclassification* **să fie 0 dar bazat pe** *Entropy* **şi, respectiv,** *Gini* **să fie diferit de 0?**

***Sugestie*: Puteţi folosi următoarea *proprietate*, care este uşor de demonstrat: Dacă într-un set de date avem C_1 instanţe din clasa (sau: cu eticheta) c_1 şi C_2 instanţe din clasa c_2, cu $C_1 < C_2$, iar după partiţionarea în funcţie de valorile atributului A această relaţie se păstrează, adică $C_1^l < C_2^l$ şi $C_1^r < C_2^r$ (evident, cu $C_1 = C_1^l + C_1^r$ şi $C_2 = C_2^l + C_2^r$), unde l şi r desemnează nodul-fiu stâng şi respectiv nodul-fiu drept, atunci *Drop-of-Impurity* pentru *Misclassification* va fi 0. Însă, în aceleaşi condiţii, *Drop-of-Impurity* bazat pe *Gini* sau pe *Entropy* va avea, în general, valori nenule.**



(Evident, *proprietatea* de mai sus se menţine dacă în locul relaţiei $<$ vom considera peste tot relaţia $>$.)

Răspuns:

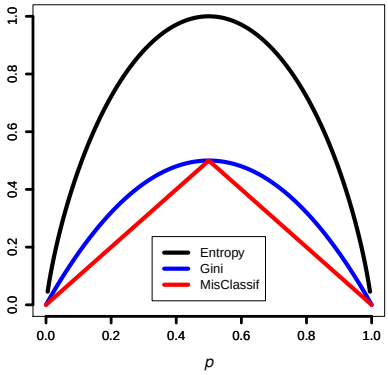
a. Vom scrie mai întâi expresiile celor trei funcţii, iar apoi vom trasa graficele lor:

$$\begin{aligned} \mathit{Entropy}(p) &= -p\log_2 p - (1-p)\log_2(1-p) \\ \mathit{Gini}(p) &= 1 - p^2 - (1-p)^2 = 2p(1-p) \end{aligned}$$

$$\mathit{MisClassif}(p) =$$

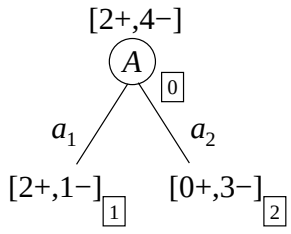
$$= \begin{cases} 1-(1-p), & \textbf{pentru } p \in [0;1/2) \\ 1-p, & \textbf{pentru } p \in [1/2;1] \end{cases}$$

$$= \begin{cases} p, & \textbf{pentru } p \in [0;1/2) \\ 1-p, & \textbf{pentru } p \in [1/2;1] \end{cases}$$



Se observă că toate cele trei funcţii iau valori în intervalul $[0;1]$, sunt simetrice în raport cu punctul $p = 1/2$, sunt strict crescătoare pe intervalul $[0;1/2]$ şi strict descrescătoare pe intervalul $[1/2;1]$, maximul fiecăreia dintre ele fiind obţinut pentru $p = 1/2$.

b. Partiţionarea datelor de antrenament în funcţie de valorile atributului A se face aşa cum se arată în figura alăturată. (În această figură şi, de asemenea, în calculele de mai jos, pentru convenienţă / simplitate, am asociat semnul $+$ etichetei c_1 şi semnul $-$ etichetei c_2.)



Aplicând formula $\Delta i(n) = i(n) - P(n_l)\, i(n_l) - P(n_r)\, i(n_r)$, vom obţine pentru *Drop-of-Impurity* următoarele valori:

Gini: $p = 2/6 = 1/3 \Rightarrow$

$$\left.\begin{array}{rcl} i(0) & = & 2 \cdot \frac{1}{3}(1 - \frac{1}{3}) = \frac{2}{3} \cdot \frac{2}{3} = \frac{4}{9} \\ i(1) & = & 2 \cdot \frac{2}{3}(1 - \frac{2}{3}) = \frac{4}{3} \cdot \frac{1}{3} = \frac{4}{9} \\ i(2) & = & 0 \end{array}\right\} \Rightarrow \Delta\, i(0) = \frac{4}{9} - \frac{3}{6} \cdot \frac{4}{9} = \frac{4}{9} - \frac{2}{9} = \frac{2}{9}.$$

Misclassification: $p = 1/3 < 1/2 \Rightarrow$

$$\left.\begin{array}{rcl} i(0) & = & p = \frac{1}{3} \\ i(1) & = & 1 - \frac{2}{3} = \frac{1}{3} \\ i(2) & = & 0 \end{array}\right\} \Rightarrow \Delta\, i(0) = \frac{1}{3} - \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}.$$

c. Într-adevăr, putem justifica uşor *proprietatea* din enunţ:

$$\begin{aligned} \Delta i(n) &= \frac{C_1}{C_1 + C_2} - \left(\frac{C_1^l + C_2^l}{C_1 + C_2} \cdot \frac{C_1^l}{C_1^l + C_2^l} + \frac{C_1^r + C_2^r}{C_1 + C_2} \cdot \frac{C_1^r}{C_1^r + C_2^r} \right) \\ &= \frac{C_1}{C_1 + C_2} - \frac{C_1^l + C_1^r}{C_1 + C_2} = \frac{C_1}{C_1 + C_2} - \frac{C_1}{C_1 + C_2} = 0. \end{aligned}$$

Dacă în setul de date din enunţ una dintre instanţele (a_1, c_1) se modifică în (a_1, c_2) şi se adaugă o instanţă (a_2, c_1), atunci *Drop-of-Impurity* va avea următoarele valori:

Entropy: $\Delta\, i(0) = H[5+, 2-] - \left(\frac{3}{7} H[2+, 1-] + \frac{4}{7} H[3+, 1-] \right) = 0.006 \neq 0$;

Gini: $2 \left\{ \frac{2}{7} \left(1 - \frac{2}{7} \right) - \left[\frac{3}{7} \cdot \frac{1}{3} \left(1 - \frac{1}{3} \right) + \frac{4}{7} \cdot \frac{1}{4} \left(1 - \frac{1}{4} \right) \right] \right\} = 2 \left\{ \frac{10}{49} - \left[\frac{2}{21} + \frac{3}{28} \right] \right\}$

$= 2 \left(\frac{10}{49} - \frac{17}{84} \right) \neq 0$;

Misclassification: $\Delta\, i(0) = \frac{2}{7} - \left(\frac{3}{7} \cdot \frac{1}{3} + \frac{4}{7} \cdot \frac{1}{4} \right) = 0$.

16. (Algoritmul ID3 ca metodă de învăţare de tip "eager": posibilitatea suplimentării datelor de antrenament)

CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh, HW1, pr. 3.4

Presupunem că, pornind de la un set de date de antrenament D, obţinem un arbore de decizie ID3, notat cu T. Ulterior, cineva ne mai dă un set suplimentar de date de antrenament, D'. Putem proceda într-unul din următoarele două moduri:

– Putem rula din nou ID3, de această dată pe datele de antrenament $D \cup D'$, obţinând arborele T_1. (Dezavantaj: dacă $|D|$ este foarte mare, această procedură poate fi costisitoare ca timp.)

– Putem extinde T, arborele ID3 obţinut pe mulţimea de antrenament D, ţinând cont de datele D'. Arborele nou, T_2, ar putea să nu fie la fel de bun ca T_1 (vedeţi cazul de mai sus), dar el este consistent cu datele din $D \cup D'$ în cazul în care aceste mulţimi de antrenament nu conţin zgomote. În special dacă $|D'|$ este mic, această metodă este acceptabilă din punct de vedere practic.

Propuneţi o procedură pentru obţinerea efectivă a arborelui T_2.

Răspuns:

Instanţele din D' se ataşează nodurilor frunză ale arborelui T, conform procedurii de clasificare de la arbori de decizie. Pentru fiecare dintre aceste noduri frunză, dacă instanţele ataşate nodului respectiv nu sunt toate etichetate identic, se aplică algoritmul ID3 (folosind mulţimea de atribute neutilizate pe drumul care uneşte nodul rădăcină al arborelui cu acest nod frunză). În acest mod se obţine un alt arbore de decizie T_2 care este consistent cu datele din $D \cup D'$.

17. **(Reducerea caracterului "greedy" al algoritmului ID3 prin calcularea câştigului de informaţie în maniera "2-step look-ahead")**

CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW4, pr. 1.2-6

Învăţarea automată a arborilor de decizie depinde mult de utilizarea unui mecanism "greedy" de selecţie a atributelor.

a. Dacă un set de date are A atribute booleene, calculaţi (ca expresie în funcţie de A) numărul total de apeluri la funcţia care calculează câştigul de informaţie pentru elaborarea întregului arbore de decizie.

Pentru *simplitate*, se va presupune că toate atributele sunt necesare pentru clasificarea unei instanţe, iar setul de date de antrenament conţine toate instanţele posibile (adică toate combinaţiile posibile de perechi atribut-valoare, inclusiv pentru atributul de ieşire).

b. Este posibil să îmbunătăţim algoritmul ID3, făcându-l să se comporte mai puţin "greedy", prin explorarea / prospectarea în avans (engl., look-ahead) a spaţiului de căutare. La o explorare cu 2 paşi înainte (engl., 2-step look-ahead), calculul câştigului de informaţie pentru un atribut a_i pe mulţimea de instanţe D va fi făcut cu ajutorul formulei

$$IG_{2-\text{step}}(D, a_i) = \max_{a_l, a_r} \left\{ \frac{n_l}{n_l + n_r} IG(D_l, a_l) + \frac{n_r}{n_l + n_r} IG(D_r, a_r) \right\},$$

unde

– a_l şi a_r sunt atributele din nodurile descendente din nodul marcat cu atributul a_i,

– D_l şi D_r sunt seturile de instanţe asignate nodurilor descendente din a_i, iar n_l şi n_r reprezintă cardinalul mulţimii D_l şi respectiv D_r;

– $IG(D_l, a_l)$ este câştigul de informaţie calculat (în sens clasic) pentru atributul a_l pe setul D_l; similar, $IG(D_r, a_r)$ este câştigul de informaţie pentru atributul a_r pe setul D_r.

b1. Explicaţi pe scurt de ce sunt necesari factorii $\frac{n_l}{n_l+n_r}$ şi $\frac{n_r}{n_l+n_r}$ în formula de mai sus.

b2. Dacă se folosesc A atribute booleene, câte apeluri la funcţia $IG(\ldots,\ldots)$ sunt necesare pentru a stabili atributul din nodul rădăcină?

b3. Vom evalua acum cât de costisitoare este această explorare în avans a spaţiului de căutare.
Dacă $A = 10$ şi se face aceeaşi presupoziţie ca la punctul a, calculaţi câte niveluri complete din arborele ID3 standard se pot calcula cu acelaşi efort de calcul — exprimat ca număr de apeluri la funcţia IG — ca la stabilirea atributului rădăcină în varianta *2-step look-ahead*.

b4. Metoda de învăţare a arborilor de decizie folosind *2-step look-ahead* crează o clasă de modele / ipoteze mai largă decât algoritmul ID3 simplu? Altfel spus, putem să calculăm în acest mod funcţii de clasificare pe care nu le putem reprezenta cu arborii de decizie standard?

Răspuns:

a. Datorită presupunerii făcute pentru *simplitate*, arborele de decizie final va fi un arbore binar complet cu $A+1$ niveluri (incluzând şi nodurile de decizie), notate în mod convenţional de la 0 la A. Pe fiecare nivel $i = \overline{0, A-1}$ al arborelui binar se găsesc 2^i noduri. În fiecare dintre aceste noduri poate fi ales un atribut din cele $A - i$ rămase disponibile. Deci pentru determinarea nivelului i se fac $2^i \cdot (A - i)$ apeluri ale funcţiei IG.

Desigur, pe penultimul nivel, $A-1$, nu mai există decât un singur atribut rămas disponibil, prin urmare nu este necesar să se calculeze câştigul de informaţie. Aşadar, numărul total de apeluri ale funcţiei IG pentru elaborarea întregului arbore de decizie este:

$$N_{IG} = A + 2(A-1) + 4(A-2) + \ldots + 2^{A-2}(A-(A-2)) = \sum_{i=0}^{A-2} 2^i(A-i)$$

b1. n_l şi n_r reprezintă numărul de instanţe asignate nodurilor descendente din a_i, deci factorii $\frac{n_l}{n_l+n_r}$ şi $\frac{n_r}{n_l+n_r}$ reprezintă ponderile acestor sub-mulţimi în raport cu reuniunea lor. Cei doi factori vor pondera corespunzător câştigurile de informaţie de pe cele două ramuri din arborele de decizie. Dacă nu facem astfel de ponderări, este posibil să avem un câştig de informaţie mare pe o mulţime mică sau invers. În consecinţă, simpla însumare a celor două câştiguri de informaţie nu ar emula în mod veridic câştigul de informaţie pe întreg ansamblul lui D.

b2. Pentru stabilirea atributului din nodul rădăcină făcând o explorare cu 2 paşi înainte, se calculează pentru fiecare dintre cele A atribute câştigurile de informaţie corespunzătoare celor 2 descendenţi, adică:

$$N_{IG_{2-\text{step}}}(\text{rădăcină}) = A \cdot 2(A-1) = 2A^2 - 2A$$

b3. Dacă $A = 10$, atunci pentru stabilirea atributului rădăcină în varianta *2-step look-ahead* se apelează funcţia IG de $N_{IG_{2-\text{step}}}(\text{rădăcină}) = 200 - 20 = 180$ de ori. Trebuie determinat numărul x de niveluri complete din arborele ID3

standard care se pot calcula folosind maxim 180 **de apeluri ale funcţiei** IG**, adică** argmax_x **astfel încât** $N_{IG}(x) = \sum_{i=0}^{x-1} 2^i(10-i) \leq 180$**.**

$$\begin{array}{lllll} x=1 & \Rightarrow & N_{IG}(1) = & 2^0(10-0) = & 10 \\ x=2 & \Rightarrow & N_{IG}(2) = & 10+2^1(10-1) = & 28 \\ x=3 & \Rightarrow & N_{IG}(3) = & 28+2^2(10-2) = & 60 \\ x=4 & \Rightarrow & N_{IG}(4) = & 60+2^3(10-3) = & 116 \\ x=5 & \Rightarrow & N_{IG}(5) = & 116+2^4(10-4) = & 212 \quad > 180 \end{array}$$

Aşadar, se pot calcula 4 niveluri complete din arborele ID3 standard cu acelaşi efort de calcul ca la stabilirea atributului din rădăcină în varianta (mai puţin "greedy") a algoritmului ID3 cu *2-step look-ahead*.

b4. Nu. Clasa de modele / ipoteze pe care lucrează algoritmul ID3 cu *2-step look-ahead* este aceeaşi ca la algoritmul ID3 standard. Însă este foarte posibil ca arborele de decizie construit să fie mai bun dacă se face cătarea în maniera *2-step look-ahead*.

18. **(Îmbunătăţirea algoritmului ID3, folosind IG cu "2-step look-ahead")**

Liviu Ciortuz, folosind date de la CMU, 2008 fall, Eric Xing, HW2, pr. 3

Se consideră variabilele booleene X_1**,** X_2 **şi** X_3**, precum şi clasificarea** $Y = \{0, 1\}$**. Fie setul de date de antrenament din tabelul de mai jos.**

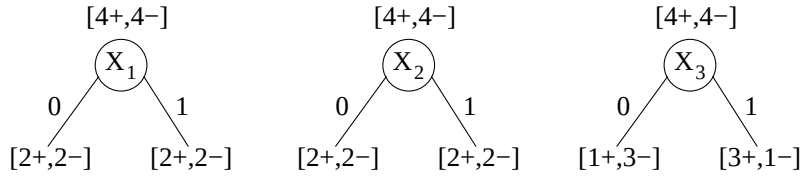
	X_1	X_2	X_3	Y
1	0	0	0	0
2	0	0	1	0
3	0	1	0	1
4	0	1	1	1
5	1	0	1	1
6	1	0	1	1
7	1	1	0	0
8	1	1	0	0

a. Elaboraţi arborele de decizie cu algoritmul ID3 standard.

b. Elaboraţi arborele de decizie cu algoritmul ID3 folosind câştigul de informaţie cu *2-step look-ahead*, aşa cum este acesta definit în problema 17.

Răspuns:

a. Pentru nodul rădăcină vom alege unul dintre atributele X_1**,** X_2 **şi** X_3**. Corespunzător, vom obţine următoarele partajări ale setului de date de antrenament:**



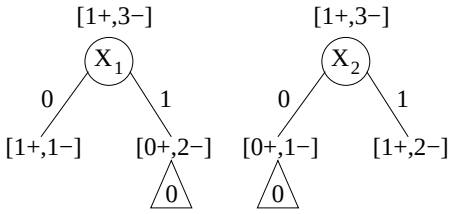
În cele ce urmează vom scrie câştigul de informaţie (IG**) omiţând primul argument, întrucât se consideră cunoscut din context. Vom calcula câştigul de informaţie corespunzător fiecărui atribut:**

$$IG(X_1) = IG(X_2) \quad = \quad H[4+,4-] - \left(\frac{4}{8}H[2+,2-] + \frac{4}{8}H[2+,2-]\right) = 1 - \left(\frac{1}{2}+\frac{1}{2}\right) = 0$$

$$IG(X_3) = H[4+,4-] - \left(\frac{4}{8}H[1+,3-] + \frac{4}{8}H[3+,1-]\right)$$

Însă $H[1+,3-] = H[3+,1-]$, iar valoarea corespunzătoare este sub-unitară, deci $IG(X_3) > 0$. Aşadar, în nodul rădăcină se alege atributul X_3.

Pentru nodul corespunzător ramurii $X_3 = 0$ se poate alege dintre atributele care au mai rămas, adică X_1 sau X_2.
Se calculează câştigurile de informaţie corespunzătoare:

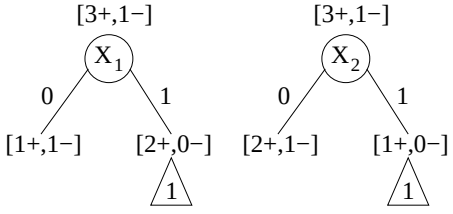


$$\begin{aligned}
IG(X_1 \mid X_3 = 0) &= H[1+,3-] - \left(\frac{2}{4}H[1+,1-] + \frac{2}{4}H[0+,2-]\right) \\
&= 2 - \frac{3}{4}\log_2 3 - \left(\frac{1}{2} + 0\right) = \frac{3}{2} - \frac{3}{4}\log_2 3 = 0.311 \\
IG(X_2 \mid X_3 = 0) &= H[1+,3-] - \left(\frac{1}{4}H[0+,1-] + \frac{3}{4}H[1+,2-]\right) \\
&= 2 - \frac{3}{4}\log_2 3 - \left(0 + \frac{3}{4}H[1+,2-]\right) \\
&= 2 - \frac{3}{4}\log_2 3 - \frac{3}{4}\left(\frac{1}{3}\log_2 3 + \frac{2}{3}\log_2 \frac{3}{2}\right) \\
&= 2 - \frac{3}{4}\log_2 3 - \frac{3}{4}\left(\log_2 3 - \frac{2}{3}\right) = \frac{5}{2} - \frac{3}{2}\log_2 3 = 0.122
\end{aligned}$$

Cum $IG(X_1 \mid X_3 = 0) > IG(X_2 \mid X_3 = 0)$, se alege atributul X_1.

În cele de mai sus, s-au folosit notaţiile în maniera condiţională $IG(X_1 \mid X_3 = 0)$ şi $IG(X_2 \mid X_3 = 0)$ doar pentru a identifica în mod neambiguu care este nodul din arbore pentru care se calculează câştigul de informaţie respectiv.

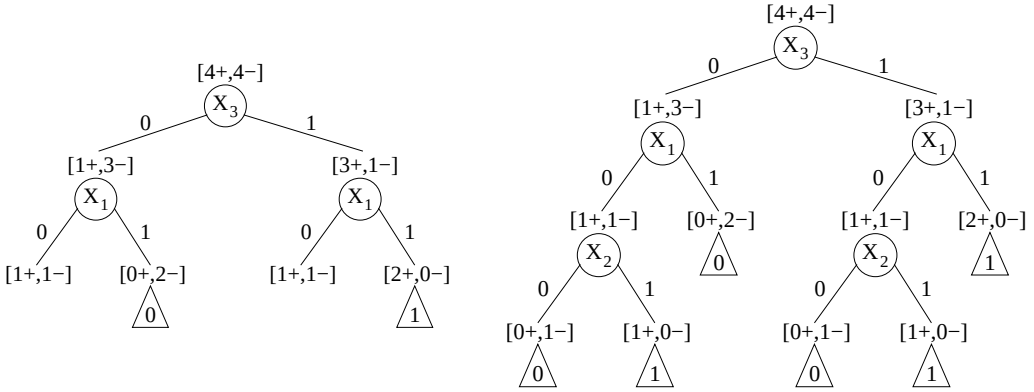
Pentru nodul corespunzător ramurii $X_3 = 1$ ce porneşte din nodul rădăcină se poate alege din nou unul dintre atributele X_1 şi X_2.
Pentru aceasta, se compară câştigurile de informaţie corespunzătoare. Se observă că acestea sunt egale cu cele din cazul precedent. Aşadar,



$$IG(X_1 \mid X_3 = 1) = IG(X_1 \mid X_3 = 0) = \frac{3}{2} - \frac{3}{4}\log_2 3 = 0.311$$

$$IG(X_2 \mid X_3 = 1) = IG(X_2 \mid X_3 = 0) = \frac{5}{2} - \frac{3}{2}\log_2 3 = 0.122$$

Prin urmare, se alege tot atributul X_1. Arborele de decizie construit până în acest moment este cel reprezentat mai jos în partea stângă:



În cele două noduri care au rămas, nu poate fi ales decât atributul X_2. Deci arborele de decizie complet construit de algoritmul ID3 este cel reprezentat mai sus în partea dreaptă.

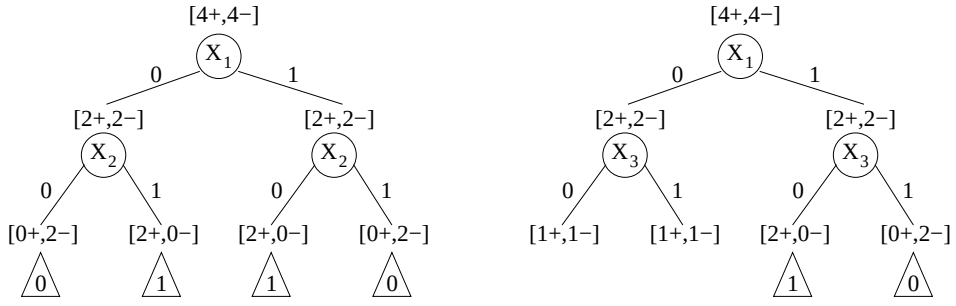
b. Dacă se foloseşte câştigul de informaţie cu *2-step look-ahead* pentru algoritmul ID3, atunci la alegerea fiecărui nod se iau în considerare [şi] toate combinaţiile posibile de noduri de pe nivelul următor, utilizându-se formula:

$$IG_{2-\text{step}}(D, a_i) = \max_{a_l, a_r} \left\{ \frac{n_l}{n_l + n_r} IG(D_l, a_l) + \frac{n_r}{n_l + n_r} IG(D_r, a_r) \right\}$$

Pentru nodul rădăcină se poate alege, la fel ca la punctul precedent, unul dintre atributele X_1, X_2 şi X_3. Aşadar, câştigul de informaţie corespunzător atributului X_1 va fi calculat astfel:

$$IG_{2-\text{step}}(X_1) = \max \begin{cases} \frac{4}{8} IG(X_2 \mid X_1 = 0) + \frac{4}{8} IG(X_2 \mid X_1 = 1) \\ \frac{4}{8} IG(X_2 \mid X_1 = 0) + \frac{4}{8} IG(X_3 \mid X_1 = 1) \\ \frac{4}{8} IG(X_3 \mid X_1 = 0) + \frac{4}{8} IG(X_2 \mid X_1 = 1) \\ \frac{4}{8} IG(X_3 \mid X_1 = 0) + \frac{4}{8} IG(X_3 \mid X_1 = 1) \end{cases}$$

Pentru a determina această valoare, va trebui să calculăm cele 4 câştiguri de informaţie (în sens clasic) implicate în formulă, şi este util să reprezentăm două din cele 4 situaţii posibile:



$$\begin{aligned} IG(X_2 \mid X_1 = 0) &= IG(X_2 \mid X_1 = 1) = IG(X_3 \mid X_1 = 1) \\ &= H[2+, 2-] - \left(\frac{1}{2}H[0+, 2-] + \frac{1}{2}H[2+, 0-]\right) = 1 - 0 = 1 \\ IG(X_3 \mid X_1 = 0) &= H[2+, 2-] - \left(\frac{1}{2}H[1+, 1-] + \frac{1}{2}H[1+, 1-]\right) = 1 - \left(\frac{1}{2} + \frac{1}{2}\right) = 0 \end{aligned}$$

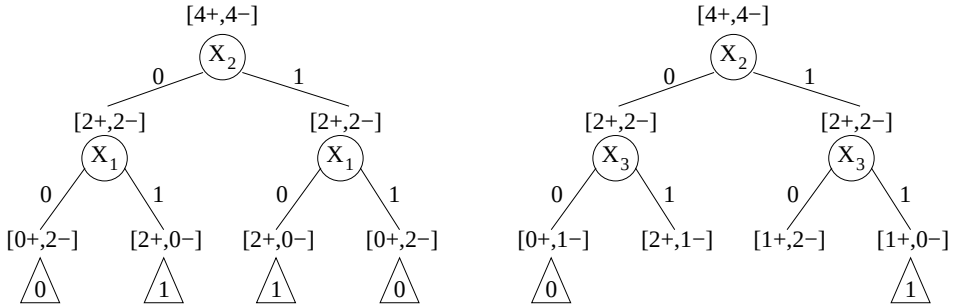
Prin urmare,

$$IG_{2-\text{step}}(X_1) = \max \left\{ \begin{array}{l} \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 1 \\ \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 1 \\ \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 1 \\ \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 1 \end{array} \right. = \max \left\{ 1, 1, \frac{1}{2}, \frac{1}{2} \right\} = 1.$$

Câştigul de informaţie corespunzător atributului X_2 plasat în nodul rădăcină este:

$$IG_{2-\text{step}}(X_2) = \max \left\{ \begin{array}{l} \frac{4}{8}IG(X_1 \mid X_2 = 0) + \frac{4}{8}IG(X_1 \mid X_2 = 1) \\ \frac{4}{8}IG(X_1 \mid X_2 = 0) + \frac{4}{8}IG(X_3 \mid X_2 = 1) \\ \frac{4}{8}IG(X_3 \mid X_2 = 0) + \frac{4}{8}IG(X_1 \mid X_2 = 1) \\ \frac{4}{8}IG(X_3 \mid X_2 = 0) + \frac{4}{8}IG(X_3 \mid X_2 = 1) \end{array} \right.$$

Vom reprezenta din nou două dintre situaţiile posibile:



$$\begin{aligned} IG(X_1 \mid X_2 = 0) &= IG(X_1 \mid X_2 = 1) = H[2+, 2-] - 0 = 1 \\ IG(X_3 \mid X_2 = 0) &= IG(X_3 \mid X_2 = 1) = H[2+, 2-] - \left(\frac{1}{4}H[0+, 1-] + \frac{3}{4}H[2+, 1-]\right) = \\ &= 1 - \frac{3}{4}\left(\log_2 3 - \frac{2}{3}\right) = \frac{3}{2} - \frac{3}{4}\log_2 3 = 0.311 \end{aligned}$$

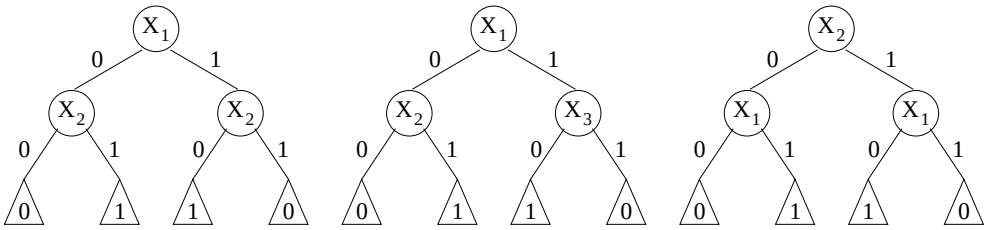
Prin urmare,

$$IG_{2-\text{step}}(X_2) = \max \left\{ \begin{array}{l} \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 1 \\ \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 0.311 \\ \frac{1}{2} \cdot 0.311 + \frac{1}{2} \cdot 1 \\ \frac{1}{2} \cdot 0.311 + \frac{1}{2} \cdot 0.311 \end{array} \right. = \max\{1, 0.655, 0.655, 0.311\} = 1$$

Pentru calculul câştigului de informaţie corespunzător atributului X_3 plasat în nodul rădăcină, au fost calculate la punctul a cele 4 câştiguri de informaţie în sens clasic implicate în formulă, deci rezultă:

$$
\begin{aligned}
IG_{2-\text{step}}(X_3) &= \max \begin{cases} \frac{4}{8}IG(X_1 \mid X_3=0)+\frac{4}{8}IG(X_1 \mid X_3=1) \\ \frac{4}{8}IG(X_1 \mid X_3=0)+\frac{4}{8}IG(X_2 \mid X_3=1) \\ \frac{4}{8}IG(X_2 \mid X_3=0)+\frac{4}{8}IG(X_1 \mid X_3=1) \\ \frac{4}{8}IG(X_2 \mid X_3=0)+\frac{4}{8}IG(X_2 \mid X_3=1) \end{cases} \\
&= \max \begin{cases} \frac{1}{2}\left(\frac{3}{2}-\frac{3}{4}\log_2 3\right)+\frac{1}{2}\left(\frac{3}{2}-\frac{3}{4}\log_2 3\right) \\ \frac{1}{2}\left(\frac{3}{2}-\frac{3}{4}\log_2 3\right)+\frac{1}{2}\left(\frac{5}{2}-\frac{3}{2}\log_2 3\right) \\ \frac{1}{2}\left(\frac{5}{2}-\frac{3}{2}\log_2 3\right)+\frac{1}{2}\left(\frac{3}{2}-\frac{3}{4}\log_2 3\right) \\ \frac{1}{2}\left(\frac{5}{2}-\frac{3}{2}\log_2 3\right)+\frac{1}{2}\left(\frac{5}{2}-\frac{3}{2}\log_2 3\right) \end{cases} = \max \begin{cases} \frac{3}{2}-\frac{3}{4}\log_2 3 \\ 2-\frac{9}{8}\log_2 3 \\ 2-\frac{9}{8}\log_2 3 \\ \frac{5}{2}-\frac{3}{2}\log_2 3 \end{cases} \\
&= \max\{0.311, 0.216, 0.216, 0.122\} = 0.311
\end{aligned}
$$

Comparând $IG_{2-\text{step}}(X_1)$, $IG_{2-\text{step}}(X_2)$ şi $IG_{2-\text{step}}(X_3)$, obţinem valoarea maximă 1 fie pentru X_1, fie pentru X_2 în nodul rădăcină. Având în vedere calculele realizate pentru a obţine aceste valori, nu mai sunt necesare operaţii suplimentare pentru determinarea arborelui de decizie construit de ID3 cu metoda *2-step look-ahead*. Există de fapt 3 arbori de decizie optimi (ca număr de niveluri şi / sau noduri) pentru aceste date de antrenament, şi anume:



Este important de remarcat faptul că algoritmul ID3 cu *2-step look-ahead* identifică toate aceste trei soluţii optimale, în vreme ce algoritmul ID3 standard nu identifică niciuna dintre ele.

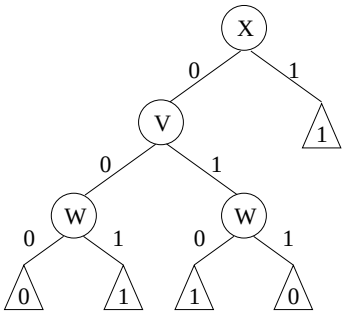
19. **(O strategie de pruning pentru arborele ID3: eliminarea nodurilor cu $IG < \varepsilon$; explorare top-down vs. bottom-up)**

prelucrare de Liviu Ciortuz, după

■ □ • ○ ***CMU, 2006 spring, Carlos Guestrin, midterm exam, pr. 4***

Puteţi constata uşor că, aplicând algoritmul ID3 pe datele din tabelul de mai jos, se va obţine arborele de decizie alăturat.

V	W	X	Y
0	0	0	0
0	1	0	1
1	0	0	1
1	1	0	0
1	1	1	1



a. Pentru un astfel de arbore de decizie, o strategie simplă de trunchiere (engl., pruning) în vederea contracarării fenomenului de "overfitting" constă în a parcurge arborele de sus în jos, începând deci cu nodul-rădăcină şi identificând fiecare nod de test pentru care câştigul de informaţie (sau un alt criteriu fixat în avans) are o valoare mai mică decât o valoare pozitivă, mică, fixată de la început, ε. Orice astfel de nod de test este imediat înlocuit — împreună cu subarborele corespunzător lui — cu un nod de decizie, conform etichetei majoritare a instanţelor asignate nodului de test. Această strategie se numeşte "top-down pruning".

Care este arborele de decizie obţinut aplicând această strategie pe arborele de mai sus, dacă se consideră $\varepsilon = 0.0001$? Care este eroarea la antrenare pentru noul arbore?

b. O altă posibilitate de a face pruning este să parcurgem arborele de decizie începând cu părinţii nodurilor-frunză şi să eliminăm în mod recursiv acele noduri de test pentru care câştigul de informaţie (sau un alt criteriu ales) este mai mic decât ε. Aceasta este strategia de "bottom-up pruning".

***Observaţie*: Spre deosebire de strategia top-down, în varianta de pruning de tip bottom-up nu vor fi eliminate noduri (cu $IG < \varepsilon$) pentru care există descendenţi al căror câştig de informaţie este mai mare sau egal cu ε.**

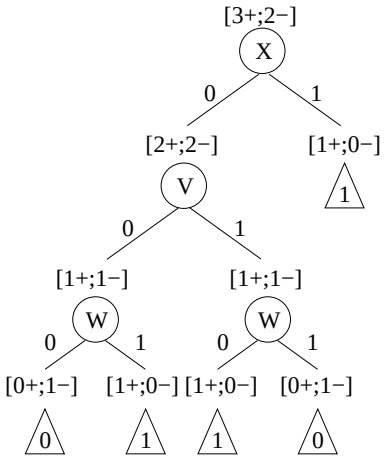
Ce arbore se obţine făcând "bottom-up pruning" pe arborele dat mai sus, dacă se consideră $\varepsilon = 0.0001$? Care este eroarea la antrenare pentru arborele rezultat?

c. Stabiliţi în ce situaţii ar fi indicat să alegem strategia "bottom-up pruning" în loc de "top-down pruning" şi viceversa. Comparaţi acurateţea la antrenare şi complexitatea computaţională a celor două strategii de pruning.

d. Cât este înălţimea — adică, numărul de niveluri de test — pentru arborele returnat de ID3 urmat de "bottom-up pruning"? Puteţi găsi un arbore de decizie având o înălţime mai mică, dar care clasifică perfect setul de antrenament? Ce concluzie putem trage despre calitatea [outputului] algoritmului ID3?

Răspuns:

Înainte de a rezolva efectiv punctele a şi b, vom augmenta arborele de decizie dat cu informaţiile referitoare la numărul de instanţe (pozitive şi, respectiv, negative) asignate fiecărui nod de test. Obţinem astfel figura alăturată.



a. Câştigul de informaţie al atributului X plasat în nodul-rădăcină este:

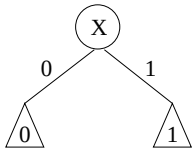
$$H[3+;2-] - 1/5 \cdot 0 - 4/5 \cdot 1 = 0.971 - 0.8 = 0.171 > \varepsilon.$$

Prin urmare, acest nod nu va fi eliminat din arbore.

Câştigul de informaţie al atributului V este:

$$H[2+;2-] - 1/2 \cdot 1 - 1/2 \cdot 1 = 1 - 1 = 0 < \varepsilon.$$

Aşadar, nodul reprezentat de atributul V va fi eliminat şi vom obţine arborele de decizie [trunchiat], reprezentat în figura alăturată. Menţionăm că am fi putut la fel de bine (din punctul de vedere al numărului de erori la antrenare) să alegem decizia $Y = 1$ în nodul-fiu stâng, însă în acel caz arborele s-ar fi redus de fapt la un singur nod de decizie (cu outputul $Y = 1$).
Eroarea la antrenare produsă de acest arbore este $2/5$.



b. Câştigul de informaţie al celor două noduri marcate cu atributul W în arborele dat în enunţ este acelaşi, şi anume 1 (se poate verifica imediat). Prin urmare, aplicând strategia de "bottom-up pruning", arborele de decizie rămâne identic cu cel iniţial. Evident, eroarea la antrenare pentru acest arbore este 0, întrucât datele de antrenament nu conţin inconsistenţe.

c. Din cauza faptului că la top-down pruning, odată cu un nod de test pentru care câştigul de informaţie (IG) este mai mic decât valoarea ε se elimină întregul subarbore care are ca rădăcină acel nod de test, această strategie este mai rapidă decât (sau, în cel mai rău caz, la fel de rapidă / lentă ca şi) pruning-ul de tip bottom-up.

Aşa cum s-a menţionat în *observaţia* din enunţ şi s-a exemplificat apoi la punctele a şi b, dezavantajul pruning-ului de tip top-down este că odată ce este eliminat un nod cu IG mai mic decât ε, este posibil ca între descendenţii săi (eliminaţi) să fie şi noduri care au câştigul de informaţie mai mare sau egal cu ε. Pruning-ul bottom-up nu are acest dezavantaj; el este deci mai „conservativ".

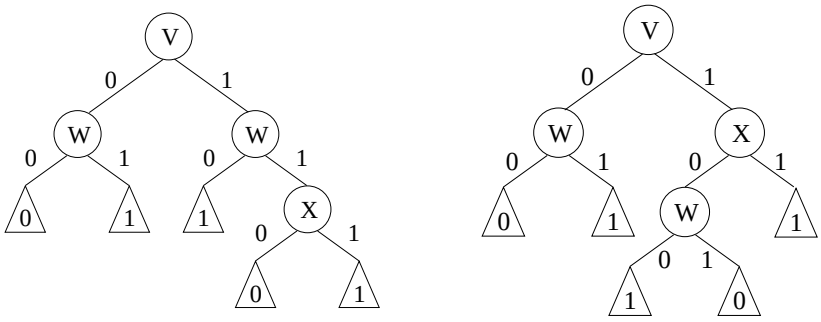
O *problemă* care poate însă să apară şi în cazul pruning-ului de tip bottom-up este faptul că în nodurile apropiate de nodurile de decizie (acestea din urmă

fiind nodurile-frunză), câştigul de informaţie se calculează uneori pe mulţimi mici de exemple, deci testul $IG \geq \varepsilon$ nu este neapărat semnificativ din punct de vedere statistic. Aşadar, este *recomandabil* ca în astfel de situaţii să se folosească un test statistic, de exemplu *testul* χ^2. (A se vedea problemele 20 şi 51.)

În ce priveşte comparaţia dintre acurateţile arborilor obţinuţi prin aplicarea celor două variante de pruning: se observă că arborele mai simplu (cel de la punctul a) are o eroare la antrenare mai mare decât arborele mai complex (cel de la punctul b), însă este mai probabil ca acesta din urmă să producă overfitting.

d. Din rezolvarea dată la punctul b, rezultă imediat că înălţimea arborelui obţinut prin aplicarea algoritmului ID3 urmat de pruning de tip bottom-up este 3. (Nu se iau în considerare nodurile frunză.) Vom demonstra — exact ca la problema 1 — că nu există un arbore de decizie consistent cu datele de antrenament, care să aibă adâncimea strict mai mică decât 3:

Se observă uşor din tabelul dat în enunţ că $Y = (V \text{ XOR } W) \vee X$, aşadar variabilele V şi W au rol simetric în definirea funcţiei reprezentate de variabila Y. Punând atributul X în nodul rădăcină, arborele de decizie minimal care se poate obţine este cel dat în enunţ (sau, echivalent, arborele care obţine din acesta interschimbând V şi W). Dacă, în schimb, punem atributul V în nodul rădăcină, se pot obţine doi arbori de decizie minimali, aşa cum se arată grafic mai jos. (Şi similar, dacă în nodul rădăcină punem atributul W.)



20. **(ID3 cu post-pruning: folosirea testului statistic χ^2 pentru limitarea overfitting-ului)**

prelucrare de Liviu Ciortuz, după
■ • ***CMU, 2010 fall, Ziv Bar-Joseph, HW2, pr. 2.1***

În acest exerciţiu veţi face pruning asupra unui arbore ID3 (după ce s-a făcut antrenarea pe toate datele disponibile), folosind o metodă statistică de testare / verificare a ipotezelor.

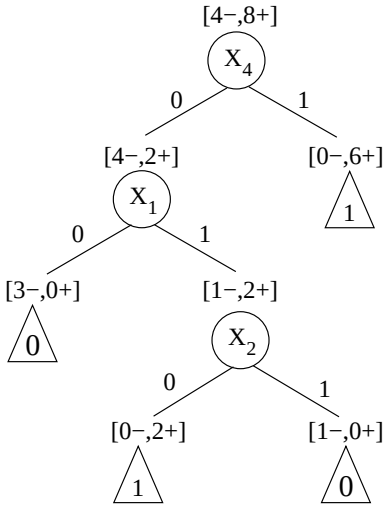
După ce a fost învăţat arborele de decizie, vizităm fiecare nod intern (inclusiv nodul rădăcină) şi testăm dacă atributul care a fost pus în nodul respectiv nu este cumva necorelat cu eticheta / clasa specificată de către atributul de ieşire.

Pentru aceasta, mai întâi presupunem că atributul din nodul respectiv este independent de atributul de ieşire (aceasta este aşa-numita „ipoteză nulă"), iar apoi folosim testul χ^2 al lui Pearson pentru a genera o „statistică", care

poate constitui temeiul pentru respingerea „ipotezei nule“. Dacă ipoteza nulă nu poate fi respinsă, eliminăm sub-arborele din nodul respectiv (de fapt, îl înlocuim cu un nod de decizie).

Pentru a ilustra aceste chestiuni, considerăm arborele de decizie de mai jos (partea dreaptă); el a fost construit pornind de la datele din tabelul alăturat lui, folosind algoritmul ID3.

X_1	X_2	X_3	X_4	*Class*
1	1	0	0	0
1	0	1	0	1
0	1	0	0	0
1	0	1	1	1
0	1	1	1	1
0	0	1	0	0
1	0	0	0	1
0	1	0	1	1
1	0	0	1	1
1	1	0	1	1
1	1	1	1	1
0	0	0	0	0



a. Pentru fiecare nod intern din arborele de decizie vom crea o *tabelă de contingenţă* (engl., contingency table) pentru exemplele de antrenare care sunt asignate nodului respectiv. Tabela aceasta va avea coloanele etichetate cu cele c clase / valori ale variabilei de ieşire. Similar, valorile atributului testat în nodul respectiv (având în total r valori) vor fi asignate liniilor tabelei. Dacă acceptăm o uşoară simplificare în forma de exprimare, vom putea spune că un element oarecare $O_{i,j}$ din tabela de contingenţă reprezintă numărul de „observaţii“ (adică, instanţe de antrenament asignate nodului respectiv) pentru care valoarea atributului testat este i, iar eticheta / clasa este j.

Calculaţi tabelele de contingenţă pentru cele trei noduri de test ale arborelui de decizie dat mai sus. Apoi, pornind de la datele conţinute în fiecare dintre aceste matrice de contingenţă, estimaţi (în sensul verosimilităţii maxime) probabilităţile pentru valorile variabilei de ieşire (*Class*), precum şi pentru valorile atributului din nodul corespunzător.

b. Pentru a aplica testul statistic χ^2, avem nevoie să calculăm pentru fiecare nod intern al arborelui de decizie încă o tabelă (pe care o vom nota cu E), în care să consemnăm *numărul aşteptat* de apariţii (engl., expected counts) ale instanţelor de antrenament la nodul respectiv, pentru fiecare pereche de indici i, j având semnificaţia de mai sus. Acest număr aşteptat este numărul (mediu) de instanţe de antrenament pe care le-am „observa“ în nodul respectiv dacă atributul selectat şi clasa (variabila de ieşire) ar fi independente.

Derivaţi o formulă pentru calculul fiecărui element (notat $E_{i,j}$) din această a doua tabelă.

***Întrebări ajutătoare*: Care este probabilitatea ca exemplele de antrenament asignate nodului respectiv să aibă o anumită etichetă (j)? Ţinând cont de această probabilitate, precum şi de presupoziţia de independenţă formulată**

prin ipoteza „nulă“, care este numărul de exemple cu o anumită valoare (i) pentru atributul selectat în nodul respectiv, care ar trebui (i.e., „ne aşteptăm“) să aibă acea etichetă / clasă (j)?

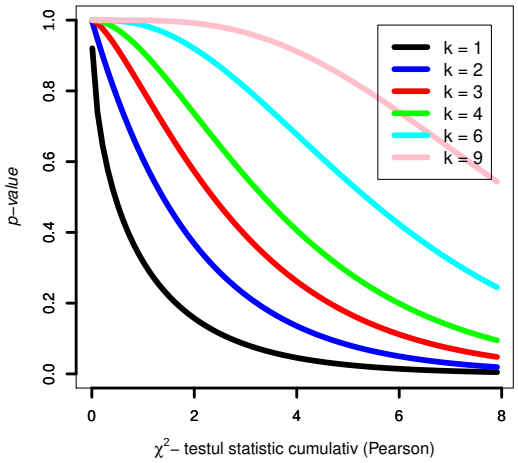
Folosind formula pe care tocmai aţi derivat-o, calculaţi matricea E pentru fiecare dintre cele trei noduri de test ale arborelui de decizie dat mai sus.

c. Date fiind cele două tabele pentru nodul considerat, puteţi calcula acum testul statistic χ-pătrat:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

Puteţi introduce valoarea calculată χ^2 precum şi numărul de grade de libertate $(r-1)(c-1)$ într-un program[385] sau într-un calculator on-line[386] pentru a calcula o aşa-numită p-valoare (engl., p-value).[387]

În general, dacă $p < 0.05$, se va considera că nu avem suficientă *evidenţă* în favoarea ipotezei „nule“, care afirma că atributul selectat şi clasa (variabila de ieşire) sunt independente, şi o vom respinge. Într-o astfel de situaţie, spunem că testul [din nodul respectiv] este *semnificativ* din punct de vedere statistic.



Pentru fiecare dintre cele trei noduri interne din arborele de decizie dat mai sus, găsiţi p-valoarea corespunzătoare şi precizaţi dacă testul din nodul respectiv este sau nu semnificativ d.p.v. statistic. Cât de multe noduri interne vor fi eliminate din arbore dacă la pruning impunem condiţia $p \geq 0.05$ [pentru a elimina un nod de test din arbore şi a-l înlocui cu un nod de decizie]?

Răspuns:

a. Pentru fiecare nod din arborele ID3 vom alcătui matricea de contingenţă asociată, pornind de la partiţionările mulţimilor de exemple care au fost asignate (de către algoritmul ID3) descendenţilor nodului respectiv. Apoi, din fiecare matrice de contingenţă vom estima (în sensul verosimilităţii maxime)

[385] Folosiţi 1-chi2cdf(x,df) în MATLAB sau CHIDIST(x,df) în Excel.

[386] `http//faculty.vassar.edu/lowry/tabs.html#csq` este un astfel de calculator.

[387] *Statistica* χ^2 este aproximată de *distribuţia* χ^2 cu numărul corespunzător (k) de grade de libertate. (Distribuţia χ^2 reprezintă suma pătratelor a k variabile gaussiene standard independente.)

p-valoarea despre care este vorba mai sus reprezintă probabilitatea ca distribuţia χ^2 să ia valori mai mari sau egale cu valoarea considerată (i.e., valoarea calculată pentru statistica χ^2). Aşadar, p-valoarea pentru testul χ^2 se calculează făcând diferenţa dintre 1 şi valoarea funcţiei de distribuţie cumulative (c.d.f.) pentru distribuţia χ^2 cu k de grade de libertate. Vedeţi site-ul https://en.m.wikipedia.org/wiki/Chi-square_distribution#Table_of_.CF.872_value_vs_p-value (accesat la 5.09.2015).

probabilităţile pentru valorile variabilei de ieşire (*Class*), precum şi probabilităţile pentru valorile atributului din acel nod. (Atenţie la condiţionarea probabilităţilor!)

O_{X_4}	$Class=0$	$Class=1$
$X_4=0$	4	2
$X_4=1$	0	6

$$\overset{N=12}{\Rightarrow} \begin{cases} P(X_4=0)=\frac{6}{12}=\frac{1}{2},\ P(X_4=1)=\frac{1}{2} \\ P(Class=0)=\frac{4}{12}=\frac{1}{3},\ P(Class=1)=\frac{2}{3} \end{cases}$$

$O_{X_1\mid X_4=0}$	$Class=0$	$Class=1$
$X_1=0$	3	0
$X_1=1$	1	2

$$\overset{N=6}{\Rightarrow} \begin{cases} P(X_1=0\mid X_4=0)=\frac{3}{6}=\frac{1}{2} \\ P(X_1=1\mid X_4=0)=\frac{1}{2} \\ P(Class=0\mid X_4=0)=\frac{4}{6}=\frac{2}{3} \\ P(Class=1\mid X_4=0)=\frac{1}{3} \end{cases}$$

$O_{X_2\mid X_4=0,X_1=1}$	$Class=0$	$Class=1$
$X_2=0$	0	2
$X_2=1$	1	0

$$\overset{N=3}{\Rightarrow} \begin{cases} P(X_2=0\mid X_4=0,X_1=1)=\frac{2}{3} \\ P(X_2=1\mid X_4=0,X_1=1)=\frac{1}{3} \\ P(Class=0\mid X_4=0,X_1=1)=\frac{1}{3} \\ P(Class=1\mid X_4=0,X_1=1)=\frac{2}{3} \end{cases}$$

b. Considerăm i o valoare arbitrar aleasă (dar fixată) pentru atributul de intrare A care este testat în nodul curent, iar j o valoare arbitrar aleasă (de asemenea, fixată) pentru atributul de ieşire *Class* (renotat cu C). Ţinând cont de presupoziţia de independenţă stipulată de ipoteza „nulă", putem scrie:

$$P(A=i,C=j)=P(A=i)\cdot P(C=j)$$

Probabilităţile $P(A=i)$ şi $P(C=j)$ pot fi estimate — în sensul verosimilităţii maxime (MLE) —, cu ajutorul celor N instanţe de antrenament asignate nodului respectiv. Instanţele pentru care atributul A are valoarea i sunt tocmai cele din linia i a matricei. În mod similar, instanţele care au eticheta / clasa j sunt cele din coloana j a matricei de count-uri observate. Aşadar,

$$P(A=i)=\frac{\sum_{k=1}^{c} O_{i,k}}{N} \text{ şi } P(C=j)=\frac{\sum_{k=1}^{r} O_{k,j}}{N}$$

În consecinţă,

$$P(A=i,C=j)=\frac{\left(\sum_{k=1}^{c} O_{i,k}\right)\left(\sum_{k=1}^{r} O_{k,j}\right)}{N^2},$$

iar valoarea aşteptată — repetăm, în condiţiile presupoziţiei de independenţă — pentru numărul de instanţe având atributul $A=i$ şi clasa $C=j$ va fi dată de formula

$$E_{i,j}=N\cdot P(A=i,C=j)=\frac{\left(\sum_{k=1}^{c} O_{i,k}\right)\left(\sum_{k=1}^{r} O_{k,j}\right)}{N}$$

c. Folosind probabilităţile calculate la punctul precedent şi ţinând cont de presupoziţia de indepenedenţă, calculăm numărul de observaţii aşteptate în fiecare nod, pentru a completa matricele E:

E_{X_4}	$Class=0$	$Class=1$
$X_4=0$	2	4
$X_4=1$	2	4

$E_{X_1\|X_4}$	$Class=0$	$Class=1$
$X_1=0$	2	1
$X_1=1$	2	1

$E_{X_2\|X_4,X_1=1}$	$Class=0$	$Class=1$
$X_2=0$	$\frac{2}{3}$	$\frac{4}{3}$
$X_2=1$	$\frac{1}{3}$	$\frac{2}{3}$

Ca să *exemplificăm* cum am procedat, detaliem mai jos calculul pentru primul element din matricea E_{X_4}:

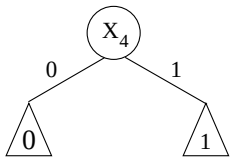
$$N=12,\ P(X_4=0)=\frac{1}{2}\ \textbf{şi}\ P(Class=0)=\frac{1}{3}\Rightarrow$$
$$N\cdot P(X_4=0,Class=0)=N\cdot P(X_4=0)\cdot P(Class=0)=12\cdot\frac{1}{2}\cdot\frac{1}{3}=2$$

Acum putem aplica pentru fiecare nod din arborele de decizie formula de calcul a valorilor / statisticilor χ^2 care ne-a fost dată în enunţ:

$$\chi^2{}_{X_4}=\frac{(4-2)^2}{2}+\frac{(2-4)^2}{4}+\frac{(0-2)^2}{2}+\frac{(6-4)^2}{4}=2+2+1+1=6$$
$$\chi^2{}_{X_1|X_4=0}=\frac{(3-2)^2}{2}+\frac{(0-1)^2}{1}+\frac{(1-2)^2}{2}+\frac{(2-1)^2}{1}=3$$
$$\chi^2{}_{X_2|X_4=0,X_1=1}=\frac{\left(0-\frac{2}{3}\right)^2}{\frac{2}{3}}+\frac{\left(2-\frac{4}{3}\right)^2}{\frac{4}{3}}+\frac{\left(1-\frac{1}{3}\right)^2}{\frac{1}{3}}+\frac{\left(0-\frac{2}{3}\right)^2}{\frac{2}{3}}=\frac{4}{9}\cdot\frac{27}{4}=3$$

Accesând pagina web indicată în enunţ, am obţinut p-valorile următoare: 0.0143, 0.0833 şi 0.0833.

În consecinţă, cu un grad de încredere de cel puţin 95%, nodurile situate pe nivelurile 1 şi 2 din arborele ID3 pot fi eliminate. Pentru nodul rădăcină, ipoteza „nulă“ nu se verifică, adică variabilele X_4 şi $Class$ nu sunt independente. Arborele obţinut în urma pruning-ului este cel din figura alăturată.



***Observaţie*: Este de remarcat faptul că arborele ID3 furnizat în enunţ are (întâmplător) atât pentru atributul X_4 (din nodul rădăcină) cât şi pentru nodul X_1 (de pe primul nivel) acelaşi câştig de informaţie, 0.4591. În urma testului χ^2, se va elimina însă doar nodul care-l conţine pe X_1. Valoarea statisticii χ^2 asociate nodului X_1 este mult mai mică (3, faţă de 6, cât este pentru nodul care-l conţine pe X_4), deci vom avea suficientă „evidenţă“ pentru a trage concluzia, cu un grad de încredere de 95%, că variabila de ieşire, *Class*, şi $X_1|X_4=0$ sunt independente.**[388] **Dacă am fi aplicat o metodă de pruning**

[388] Observaţi şi faptul că nodul care conţine atributul X_1 are mai puţine instanţe asociate (şi anume, jumătate) faţă de cele asociate nodului rădăcină.

bazată pe câştigul de informaţie, aceasta n-ar fi putut să trateze în mod diferit cele două noduri, adică să-l păstreze pe unul şi să-l elimine pe celălalt.

21. **(Adevărat sau Fals?)**

a. ***Liviu Ciortuz***

Algoritmul ID3 garantează obţinerea arborelui de decizie optimal (ca număr de niveluri sau de noduri).

CMU, 2002 spring, A. Moore, midterm example questions, pr. 1.c

b. Întrucât arborii de decizie pot învăţa să clasifice instanţe într-un număr discret de clase (deci nu învaţă funcţii cu valori reale), este imposibil ca ei să manifeste fenomenul de overfitting.

CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh, midterm, pr. 1.4

c. Fie A şi B doi algoritmi de clasificare automată. Algoritmul A este mai bun decât algoritmul B dacă eroarea la antrenare a algoritmului A este mai mică decât eroarea la antrenare a algoritmului B. Justificaţi.

CMU, 2005 spring, C. Guestrin, T. Mitchell, midterm, pr. 2.c

d. Presupunem că avem m instanţe şi că vom folosi jumătate dintre ele pentru antrenarea unui clasificator oarecare (nu neapărat ID3) şi jumătate pentru testare. Diferenţa dintre eroarea la antrenare şi eroarea la testare descreşte pe măsură ce numărul m creşte.

Răspuns:

a. Fals. ID3 nu garantează obţinerea arborelui optim (relativ la numărul de niveluri şi / sau noduri), ci încearcă să găsească o soluţie convenabilă, însă fără să caute în mod exhaustiv în tot spaţiul soluţiilor. Mai exact, căutarea soluţiei se face în manieră "greedy", maximizând un anumit criteriu (e.g., câştigul de informaţie) la fiecare iteraţie. O astfel de căutare nu garantează obţinerea optimului.

b. Fals. Arborii de decizie manifestă fenomenul de overfitting. Prima parte a afirmaţiei din enunţ este adevarată — arborii de decizie pot învăţa să clasifice instanţe într-un număr discret de clase —, însă a doua parte este falsă, deci avem o implicaţie de forma: $T \rightarrow F \equiv \neg T \vee F \equiv F$.

c. Fals, fiindcă la testare algoritmul B poate să aibă o eroare mai mică decât algoritmul A. Într-un astfel de caz, se spune că algoritmul A este "overfit" (rom., supra-antrenat).

d. Adevărat. Pe măsură ce dispunem de tot mai multe date de antrenament, dacă datele de antrenament sunt inconsistente, eroarea la antrenare va creşte, fiindcă va fi din ce în ce mai greu ca modelul învăţat să se adapteze la „zgomotele" din date. Similar, eroarea la testare va descreşte, fiindcă producem un clasificator care este din ce în ce mai puţin afectat de "overfitting" pe datele de antrenament. Cele două erori vor converge la aşa-numita *eroare adevărată* (engl., true error) fiindcă diferenţele statistice dintre datele de antrenament şi datele de testare vor dispărea.

2 Clasificare bayesiană

Sumar

Noţiuni preliminare

– **probabilităţi şi probabilităţi condiţionate;**

– **formula lui Bayes: ex. 5.b;**
cap. *Fundamente*, ex. 6, ex. 7, ex. 83, ex. 84;

– **independenţa [condiţională a] evenimentelor aleatoare:**
cap. *Fundamente*, ex. 4, ex. 80, ex. 81;

– **independenţa [condiţională a] variabilelor aleatoare: ex. 9, ex. 10, ex. 12, ex. 31-38; vedeţi şi cap. *Fundamente*, ex. 15, ex. 27, ex. 88.b, ex. 97, ex. 95;**

– **distribuţii probabiliste comune, marginale şi condiţionale: ex. 8, ex. 10, ex. 12, ex. 31; vedeţi şi cap. *Fundamente*, ex. 13, ex. 14;**

– **distribuţia gaussiană: de la cap. *Fundamente*, ex. 29, ex. 30 (pentru cazul unidimensional), ex. 32 (pentru cazul bidimensional), ex. 20, ex. 31, ex. 33, ex. 34 (pentru cazul multidimensional);**

– **estimarea parametrilor pentru distribuţii de tip Bernoulli, categorial şi gaussian (ultimul doar pentru cazul clasificării bayesiene de tip gaussian);**[305]

– **ipoteze MAP vs. ipoteze ML:**
formulare [ca soluţii la] probleme de optimizare:[306] **ex. 25;**
exemplificare: ex. 1, ex. 2, ex. 3, ex. 24, ex. 37;
exemplificare în cazul arborilor de decizie: ex. 4;

– **regresia logistică, chestiuni introductive:**[307] **de la cap. *Metode de regresie*, ex. 12.**

Algoritmi de clasificare bayesiană

– **Algoritmul Bayes Naiv şi algoritmul Bayes Optimal:**[308]
formulare ca probleme de optimizare / estimare în sens MAP: cartea ML, pag. 167;
pseudo-cod: cartea ML, pag. 177; vedeţi şi slide-urile lui Tom Mitchell;
exemple de aplicare: ex. 5, ex. 7, ex. 8, ex. 9, ex. 26, ex. 27, ex. 28;

[305] De la cap. *Fundamente*, pentru estimarea parametrului unei distribuţii Bernoulli vedeţi ex. 40 şi ex. 113.a, pentru estimarea parametrilor unei distribuţii categoriale vedeţi ex. 115, iar pentru estimarea parametrilor unei distribuţii gaussiene vedeţi ex. 45, ex. 46, ex. 122 (pentru cazul unidimensional) şi ex. 48 (pentru cazul multidimensional).

[306] Vedeţi cartea ML, pag. 156-157.

[307] Vedeţi draftul capitolului suplimentar pentru cartea ML a lui T. Mitchell, *Generative and discriminative classifiers: Naive Bayes and logistic regression* (în special secţiunea 3).

[308] La secţiunea aceasta, precum şi la următoarea secţiune, considerăm (implicit) că toate variabilele de intrare sunt de tip Bernoulli sau, mai general, de tip categorial. După aceea vom considera şi variabile de intrare de tip continuu, în genere de tip gaussian. Variabila de ieşire se consideră întotdeauna de tip Bernoulli / categorial.

– **aplicarea / adaptarea algoritmului Bayes Naiv pentru clasificare de texte:**[309] **ex. 6, ex. 29;**
folosirea regulii "add-one" [a lui Laplace] pentru „netezirea" parametrilor: ex. 6, ex. 30;

– **calculul ratei medii a erorilor pentru algoritmii Bayes Naiv şi Bayes Optimal: ex. 10, ex. 11, ex. 31, ex. 32, ex. 33, ex. 34, ex. 38;**

– **evidenţierea grafică a neconcordanţei predicţiilor făcute de clasificatorii Bayes Naiv şi Bayes Optimal: ex. 12.**

Proprietăţi ale algoritmilor Bayes Naiv şi Bayes Optimal

– **(P0) dacă proprietatea de independenţă condiţională a atributelor de intrare în raport cu variabila de ieşire se verifică, atunci rezultatele produse de către cei doi algoritmi (Bayes Naiv şi Bayes Optimal) în faza de testare coincid;**

– **(P1) numărul de parametri necesari de estimat din date: liniar pentru Bayes Naiv** $(2d + 1)$ **şi exponenţial pentru Bayes Optimal** $(2^{d+1} - 1)$**:**[310] **ex. 7.e, ex. 28.ab, ex. 33.ac;**

– **(P2) complexitatea algoritmului Bayes Naiv:**

complexitatea de spaţiu: $\mathcal{O}(dn)$
complexitatea de timp:
la antrenare: $\mathcal{O}(dn)$
la testare: $\mathcal{O}(d')$,
unde n **este numărul de exemple, iar** d **este numărul de atribute de intrare [LC:** d' **este numărul de atribute de intrare din instanţa de test];**

– **(P3) algoritmul Bayes Optimal poate produce eroare [la clasificare] din cauza faptului că ia decizia în sensul unui vot majoritar. Algoritmul Bayes Naiv are şi el această „sursă" de eroare; în plus el poate produce eroare şi din cauza faptului că lucrează cu presupoziţia de independenţă condiţională (care nu este satisfăcută în mod neapărat);**

– **(P4) acurateţea [la clasificare a] algoritmului Bayes Naiv scade atunci când unul sau mai multe atribute de intrare sunt duplicate: ex. 10.d, ex. 31.def;**

– **(P5) în cazul „învăţării" unei funcţii booleene (oarecare), rata medie a erorii produse la antrenare de către algoritmul Bayes Optimal (spre deosebire de Bayes Naiv!) este** 0**: ex. 33.d;**

– **(P6) complexitatea de eşantionare: de ordin logaritmic pentru Bayes Naiv şi de ordin exponenţial pentru Bayes Optimal: ex. 13;**

– **(P7) corespondenţa dintre regula de decizie a algoritmului Bayes Naiv (când toate variabilele de intrare sunt de tip Bernoulli) şi regula de decizie a *regresiei logistice* şi, în consecinţă, liniaritatea graniţelor de decizie: ex. 14.**

– ***comparaţii* între algoritmul Bayes Naiv şi alţi algoritmi de clasificare automată: ex. 36, ex. 38.**

[309] Atenţie: Noi am folosit aici versiunea de bază a algoritmului Bayes Naiv; varianta "bag of words" (vedeţi cartea Machine Learning a lui Tom Mitchell, pag. 183) diferă uşor de aceasta.
[310] Numărul de parametri indicaţi în paranteze se referă la cazul când atât atributele de intrare cât şi atributul de ieşire sunt de tip Bernoulli.

Algoritmii Bayes Naiv şi Bayes Optimal cu variabile de intrare *de tip gaussian*

• Aplicare: G[N]B: ex. 15, ex. 40 şi ex. 48; GJB: ex. 45, ex. 46 şi ex. 47; GNB vs. GJB: ex. 21.
• Numărul de parametri necesari de estimat din date: ex. 42.

• Proprietăţi:

- **(P0′) presupunem că variabila de ieşire este booleană, i.e. ia valorile 0 sau 1; dacă pentru orice atribut de intrare, variabilele condiţionale $X_i|Y = 0$ şi $X_i|Y = 1$ au distribuţii gaussiene de varianţe egale ($\sigma_{i0} = \sigma_{i1}$), atunci regula de decizie GNB (Gaussian Naive Bayes) este echivalentă (ca formă) cu cea a regresiei logistice, deci separarea realizată de către algoritmul GNB este de formă liniară: demonstraţie: ex. 17; exemplificare în $\mathbb{R}$: ex. 40.a; exemplificare în $\mathbb{R}^2$: ex. 41.c;**
- **(P1′) similar, presupunem că variabila de ieşire este booleană; dacă variabilele de intrare (notaţie: $X = (X_1, \ldots, X_d)$) au distribuţiile [comune] condiţionale $X|Y = 0$ şi $X|Y = 1$ de tip gaussian [multidimensional], cu matricele de covarianţă egale ($\Sigma_0 = \Sigma_1$), atunci regula de decizie a algoritmului "full" / Joint Gaussian Bayes este şi ea echivalentă (ca formă) cu cea a regresiei logistice, deci separarea realizată este tot de formă liniară: ex. 18, ex. 20.a.$i - ii$;**
- **(P2′) când variabilele de intrare satisfac condiţii mixte de tip (P0′) sau (P7), atunci concluzia – separare liniară – se menţine: ex. 44.b;**
- **(P3′) dacă în condiţiile de la propoziţiile (P0′)-(P2′) presupoziţia de independenţă condiţională este satisfăcută, iar numărul de instanţe de antrenament tinde la infinit, atunci rezultatul de clasificare obţinut de către algoritmul Bayes Naiv gaussian este identic cu cel al regresiei logistice: ex. 22.a.**

  Atunci când presupoziţia de independenţă condiţională nu este satisfăcută, iar numărul de instanţe de antrenament tinde la infinit, regresia logistică se comportă mai bine decât algoritmul Bayes Naiv [gaussian]: ex. 22.b;
- **(P4′) nu există o corespondenţă 1-la-1 între parametrii calculaţi de regresia logistică şi între parametrii calculaţi de algoritmul Bayes Naiv [gaussian]: ex. 23.a;**
- **(P5′) atunci când varianţele distribuţiilor gaussiene care corespund probabilităţilor condiţionale $P(X_i|Y = k)$ depind şi de eticheta k, separatorul decizional determinat de algoritmul Bayes Naiv gaussian nu mai are (în mod necesar) forma regresiei logistice: ex. 40.bc, ex. 43; similar, pentru algoritmul Bayes Optimal gaussian, atunci când $\Sigma_0 \neq \Sigma_1$, ecuaţia separatorului decizional este de ordin pătratic: ex. 19, ex. 20.a.$iii - vi$, ex. 42.e (separatorul decizional este un cerc), ex. 42.f (o hiperbolă), ex 47 (o reuniune de două drepte);**
- **(P6′) parametrii algoritmilor Bayes Naiv gaussian şi Bayes Optimal gaussian se pot estima în timp liniar în raport cu numărul de instanţe din setul de date de antrenament: ex. 41.bd, ex. 23.b.**

2.1 Clasificare bayesiană — Probleme rezolvate

2.1.1 Ipoteze de probabilitate maximă a posteriori (MAP)

1. **(Formula lui Bayes; medii ale unor variabile aleatoare discrete; [ipoteze MAP;] măsuri statistice folosite în clasificare)**

■ • ***CMU, 2009 fall, Geoff Gordon, HW1, pr. 2***

O anumită boală afectează una din 500 de persoane în medie. Identificarea persoanelor care au această boală se poate face cu ajutorul unei analize a sângelui, care costă 100 de dolari de persoană. Acestă analiză indică în cazul unui rezultat *pozitiv* faptul că se *poate* ca persoana respectivă să sufere de acea boală.

Testul / analiza are o *sensibilitate* (engl., *sensitivity* sau *recall*) perfectă — adică raportul dintre numărul instanţelor pozitive identificate ca atare de acel test şi numărul total de instanţe pozitive este 1 —, ceea ce înseamnă că pentru orice persoană care are boala respectivă, rezultatul testului este pozitiv cu probabilitate de 100%. Pe de altă parte, testul are o *specificitate* — raportul dintre numărul instanţelor negative identificate ca atare de acel test şi numărul total de instanţe negative — de 99%, adică o persoană care nu suferă de acea boală va avea cu probabilitate de 1% rezultatul testului pozitiv.

a. Se testează o persoană selectată în mod aleatoriu, iar rezultatul este pozitiv. Care este probabilitatea ca persoana respectivă să sufere de acea boală?

b. Există şi un al doilea test, care costă 10.000 de dolari şi are atât sensibilitatea cât şi specificitatea de 100%. Dacă am cere ca toate persoanele detectate pozitiv la testul precedent să fie supuse acestui test mult mai scump, care ar fi costul mediu pentru testarea / analiza unui individ?

c. O companie farmaceutică încearcă să reducă preţul celui de-al doilea test (care este perfect), adică are atât *sensibilitatea* cât şi *specificitatea* de 100%. Cât ar trebui să fie preţul acesta pentru ca primul test să nu mai fie necesar? (Adică, la ce preţ va rezulta că este mai ieftin să se utilizeze doar testul al doilea, decât să se facă ambele teste, ca la punctul b?)

Răspuns:

Definim următoarele variabile aleatoare:

B: ia valoarea 1 / adevărat pentru persoanele care suferă de această boală şi 0 / fals în caz contrar

T_1: rezultatul primului test, care poate fi $+$ (în caz de boală) sau $-$

T_2: rezultatul celui de-al doilea test, care poate fi tot $+$ sau $-$.

Folosind aceste variabile aleatoare, datele problemei se pot rescrie astfel:

$$
\begin{aligned}
&P(B) = \frac{1}{500} \\
&P(T_1 = + \mid B) = 1 \\
&P(T_1 = + \mid \bar{B}) = \frac{1}{100} \\
&P(T_2 = + \mid B) = 1 \\
&P(T_2 = + \mid \bar{B}) = 0
\end{aligned}
$$

a. Probabilitatea ca o persoană oarecare să sufere de boala respectivă, ştiind că rezultatul primului test este pozitiv, este $P(B \mid T_1 = +)$ şi se calculează cu ajutorul formulei lui Bayes:

$$
\begin{aligned}
P(B \mid T_1 = +) &= \frac{P(T_1 = + \mid B) \cdot P(B)}{P(T_1 = + \mid B) \cdot P(B) + P(T_1 = + \mid \bar{B}) \cdot P(\bar{B})} \\
&= \frac{1 \cdot \frac{1}{500}}{1 \cdot \frac{1}{500} + \frac{1}{100} \cdot \frac{499}{500}} = \frac{100}{599} \approx 0.1669
\end{aligned}
$$

***Observaţie*: Remarcaţi faptul că $P(\bar{B} \mid T_1 = +) = 0.8331$. Este imediat care dintre cele două probabilităţi, $P(B \mid T_1 = +)$ şi $P(\bar{B}$, este mai mare. În mod echivalent, pentru a stabili acest fapt era suficient să comparăm $P(B \mid T_1 = +)$ cu $1/2$, sau să stabilim care dintre produsele $P(T_1 = + \mid B) \cdot P(B)$ şi $P(T_1 = + \mid \bar{B}) \cdot P(\bar{B})$ este mai mare. Aceste observaţii sunt utile pentru că ele fac legătura cu noţiunea de *ipoteză de probabilitate maximă a posteriori* (vedeţi pr. 3).**

b. Vom calcula costul mediu al testării unui individ folosind o nouă variabilă aleatoare, notată cu C, care reprezintă costul total de testare al unei persoane. Notând cu c_1 şi c_2 costurile celor două teste, putem scrie:

$$
C = \begin{cases} c_1 & \text{dacă persoana este testată doar cu primul test} \\ c_1 + c_2 & \text{dacă persoana este testată cu ambele teste} \end{cases}
$$

O persoană este testată cu al doilea test doar dacă are rezultatul pozitiv la primul test, deci probabilităţile pentru variabila aleatoare C sunt:

$$
P(C = c_1) = P(T_1 = -) \text{ şi } P(C = c_1 + c_2) = P(T_1 = +)
$$

Costul mediu cerut de problemă este media variabilei aleatoare C, deci se poate calcula astfel:

$$
\begin{aligned}
E[C] &= c_1 \cdot P(C = c_1) + (c_1 + c_2) \cdot P(C = c_1 + c_2) \\
&= c_1 \cdot P(T_1 = -) + (c_1 + c_2) \cdot P(T_1 = +)
\end{aligned}
$$

Ştim că $P(T_1 = -) = 1 - P(T_1 = +)$, iar din formula probabilităţii totale avem:

$$
\begin{aligned}
P(T_1 = +) &= P(T_1 = + \mid B) \cdot P(B) + P(T_1 = + \mid \bar{B}) \cdot P(\bar{B}) \\
&= 1 \cdot \frac{1}{500} + \frac{1}{100} \cdot \frac{499}{500} = \frac{599}{50000} = 0.01198
\end{aligned}
$$

Aşadar, vom obţine:

$$\begin{aligned} E[C] &= c_1 \cdot (1 - P(T_1 = +)) + (c_1 + c_2) \cdot P(T_1 = +) \\ &= c_1 - c_1 \cdot P(T_1 = +) + c_1 \cdot P(T_1 = +) + c_2 \cdot P(T_1 = +) \\ &= c_1 + c_2 \cdot P(T_1 = +) \\ &= 100 + 10000 \cdot \frac{599}{50000} \\ &= 219.8 \approx 220\$ \end{aligned}$$

c. Notăm cu c_n noul preţ pentru al doilea test (T_2'). Acest preţ trebuie să fie mai mic sau egal cu costul mediu de aplicare al ambelor teste (T_1 şi T_2'), deci:

$$\begin{aligned} c_n \leq E[C'] &= c_1 \cdot P(C = c_1) + (c_1 + c_n) \cdot P(C = c_1 + c_n) \\ &= c_1 + c_n \cdot P(T_1 = +) = 100 + c_n \cdot \frac{599}{50000} \end{aligned}$$

Rezolvând ecuaţia $c_n = 100 + c_n \cdot 0.01198$, obţinem $c_n \approx 101.2125$.

Aşadar, dacă al doilea test ar costa cel mult 101.21 dolari, atunci primul test n-ar mai fi necesar.

2. **(Formula lui Bayes; [ipoteze MAP;] inferenţe statistice)**

prelucrare de Liviu Ciortuz, după
CMU, 2009 fall, Geoff Gordon, HW1, pr. 1
("Monty's haunted house" problem)

Fără să ştii cum s-a întâmplat, ai nimerit într-o casă plină de fantome. Acum eşti blocat în faţa unui perete care are 3 uşi (pentru convenienţă, le vom nota cu numerele 1, 2, 3). Apare o fantomă care îţi spune: „Scăparea ta este să ieşi din casă printr-una din aceste uşi. Însă doar una dintre ele dă în afară; celelalte două sunt păzite de câte un monstru care te va ucide imediat dacă încerci să ieşi pe acolo. Trebuie să alegi o uşă!"

Decizi să alegi la întâmplare una din cele trei uşi, să zicem uşa 1.

***Observaţie*: Probabilitatea *a priori* ca uşa aceasta să dea în afară este 1/3. (La fel este şi în cazul celorlalte două uşi.) Prin adăgarea altor informaţii — vedeţi continuarea problemei — probabilitatea *a posteriori* a aceluiaşi eveniment se poate modifica, deci într-un caz fericit ea poate creşte.**

Într-adevăr, tocmai când ai pus mâna pe clanţă ca să o deschizi, fantoma îţi spune: „Aşteaptă puţin! Îţi voi mai da o informaţie." Zicând aceasta, fantoma întredeschide o altă uşă (să zicem uşa 2) şi îţi arată că în spatele ei se află un monstru groaznic. Apoi fantoma te întreabă: „Vrei acum să alegi ultima uşă (adică uşa 3) sau consideri că este mai bine să rămâi la alegerea pe care ai făcut-o iniţial?"

Fantoma te mai ajută spunându-ţi că în alegerea ei, ea a urmat o *strategie* bazată pe două *principii*:

P1. După ce tu ai făcut alegerea iniţială, fantoma a ales una din celelalte două uşi, mai precis o uşă în spatele căruia se află un monstru. (Este evident că întotdeauna există o astfel de uşă, indiferent de alegerea ta.)

P2. În cazul în care care ambele uşi între care are de ales fantoma au în spate câte un monstru, ea procedează după *una* din următoarele trei *variante* (pe care o alege a priori şi ţi-o aduce la cunoştinţă):

a. **Fantoma alege una din cele două uşi cu probabilitate egală (1/2).**
b. **Dacă, aşa cum a fost cazul mai sus, tu ai ales uşa 1, fantoma alege uşa 2 (cu probabilitate 1).**
c. **Dacă, tot aşa, tu ai ales uşa 1, fantoma alege uşa 3 (cu probabilitate 1).**
(*Notă*: Alte variante nu interesează pentru rezolvarea care se cere mai jos.)

Se cere ca, pentru fiecare din aceste 3 variante în parte, să determini probabilităţile ca ieşirea să se afle în spatele uşii 1, respectiv în spatele uşii 3, dacă fantoma a deschis uşa 2.

***Indicaţie*: Pentru rezolvare, vom folosi două *variabile aleatoare*:**

- O **(de la engl. outside), cu valori în mulţimea $\{1, 2, 3\}$, indicând unde este ieşirea cea bună;**
- G **(de la engl. ghost), pentru a desemna uşa aleasă de fantomă.**

Pentru fiecare din variantele de strategie ale fantomei (a, b, c), folosind formula lui Bayes se calculează $P(O = 1 \mid G = 2)$ şi $P(O = 3 \mid G = 2)$.

Răspuns:

Pentru variabila aleatoare O avem probabilităţi *a priori* egale pentru toate ieşirile:

$$P(O = 1) = P(O = 2) = P(O = 3) = \frac{1}{3}. \tag{161}$$

Folosind formula lui Bayes combinată cu formula probabilităţii totale, probabilităţile (*a posteriori*) cerute vor putea fi calculate astfel:

$$P(O = 1 \mid G = 2) = \frac{P(G = 2 \mid O = 1) \cdot P(O = 1)}{P(G = 2 \mid O = 1) \cdot P(O = 1) + P(G = 2 \mid O = 3) \cdot P(O = 3)}$$
$$P(O = 3 \mid G = 2) = \frac{P(G = 2 \mid O = 3) \cdot P(O = 3)}{P(G = 2 \mid O = 3) \cdot P(O = 3) + P(G = 2 \mid O = 1) \cdot P(O = 1)}.$$

Remarcăm că la fiecare din numitorii celor două fracţii de mai sus ar fi trebuit să mai scriem $P(G = 2 \mid O = 2) \cdot P(O = 2)$, însă acesta este 0 fiindcă $P(G = 2 \mid O = 2) = 0$, conform principiului P1. Deci $P(O = 3 \mid G = 2) = 1 - P(O = 1 \mid G = 2)$.

Pentru a exprima succint probabilităţile condiţionate necesare pentru calculul probabilităţilor $P(O = 1 \mid G = 2)$ şi $P(O = 3 \mid G = 2)$ folosind formulele de mai sus, completăm următorul tabel:[311]

G	O	$P(G \mid O)$		
		varianta a	**varianta** b	**varianta** c
2	1	1/2	1	0
3	1	1/2	0	1
2	2	0	0	0
3	2	1	1	1
2	3	1	1	1
3	3	0	0	0

[311] Se observă că pe fiecare dintre cele trei coloane din tabel, ultimele patru linii sunt identice, ceea ce este natural, conform principiilor P1 şi P2.

În aceste condiţii putem calcula uşor probabilităţile cerute în fiecare din cele trei variante:

***Varianta a*:**

$$P(O=1 \mid G=2)=\frac{\frac{1}{2}\cdot\frac{1}{3}}{\frac{1}{2}\cdot\frac{1}{3}+1\cdot\frac{1}{3}}=\frac{1}{3} \text{ şi } P(O=3 \mid G=2)=1-\frac{1}{3}=\frac{2}{3},$$

deci vei alege uşa a treia, fiindcă ea are probabilitatea mai mare de a te duce afară.

***Varianta b*:**

$$P(O=1 \mid G=2)=\frac{1\cdot\frac{1}{3}}{1\cdot\frac{1}{3}+1\cdot\frac{1}{3}}=\frac{1}{2} \text{ şi } P(O=3 \mid G=2)=1-\frac{1}{2}=\frac{1}{2},$$

deci vei alege la întâmplare oricare din cele două uşi rămase (uşa 1 şi uşa 3), ele având aceeaşi probabilitate de salvare.

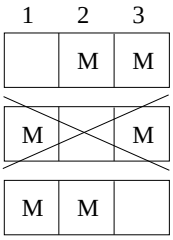
***Varianta c*:**

$$P(O=1 \mid G=2)=0 \text{ şi } P(O=3 \mid G=2)=1-0=1,$$

deci vei alege uşa a treia, care este cu siguranţă ieşirea cea bună.

***Observaţii*:**

1. Echivalent, pentru a determina maximul dintre $P(O=1 \mid G=2)$ şi $P(O=3 \mid G=2)$ ar fi fost suficient, conform formulei lui Bayes, să comparăm $P(G=2 \mid O=1)\cdot P(O=1)$ şi $P(G=2 \mid O=3)\cdot P(O=3)$. Mai mult, ţinând cont de relaţia (161), aceasta revine la a compara $P(G=2 \mid O=1)$ şi $P(G=2 \mid O=3)$. Răspunsul poate fi citit imediat din tabelul de mai sus (vedeţi prima linie şi penultima linie): în cazul variantei a, $O=3$ este varianta pentru care se obţine probabilitatea [a posteriori] maximă. Altfel spus, pentru variantei a, $O=3$ este *ipoteza de probabilitate maximă a posteriori* (engl., maximum a posteriori probability (MAP) hypothesis). Absolut similar se poate proceda şi pentru variantele b şi c.[312]



2. Alternativ, pentru variantele b şi c putem răspunde la întrebare şi făcând un raţionament care nu foloseşte formula lui Bayes. Ieşirea cea bună se poate găsi în spatele uneia dintre cele trei uşi (vedeţi figura alăturată). Cum fantoma a deschis deja uşa cu numărul 2, una dintre aceste situaţii (şi anume, a doua din figură) este eliminată, fiindcă în spatele ei este un monstru. În continuare, putem raţiona în felul următor:

***Varianta b*: Întrucât fantoma alege uşa 2 cu probabilitate 1, vom putea afirma că ambele variante – 1 şi 3 – au probabilităţi egale, şi anume $\frac{1}{2}$. Într-adevăr,**

[312] *Observaţie*: În cazuri precum cel de mai sus ($P(O=1)=P(O=2)=P(O=3)=1/3$), ipoteza MAP coincide cu *ipoteza de verosimilitate maximă* (engl., maximul likelihood (ML) hypothesis).

– **fie uşa 1, cea aleasă de mine, dă înspre afară, iar atunci fantoma trebuie, conform principiului P2, să aleagă uşa 2;**
– **fie uşa 3 dă înspre afară, iar atunci, din nou conform principiului P2, fantoma trebuie să aleagă uşa 2;**
– **conform principiului P1, nu există o a treia posibilitate;**
– **nu dispun de alte informaţii pentru a decide între cele două situaţii de mai sus.**

***Varianta c*: Ştiind că fantoma nu a deschis uşa 3 (care ar fi opţiunea corespunzătoare principiului P2), ci a ales uşa 2, înseamnă că nu a putut face altfel, deci uşa 3 reprezintă ieşirea.**

3. **(Formula lui Bayes; inferenţe statistice; exemplificarea noţiunii de ipoteză / ipoteze MAP ("Maximum A posteriori Probability"))**

■ ● ○ ***CMU, 2012 spring, Ziv Bar-Joseph, HW1, pr. 1.5***

Mickey dă cu zarul de mai multe ori, sperând să obţină un 6. Secvenţa celor 10 rezultate obţinute de el în urma acestor aruncări este următoarea: $1, 3, 4, 2, 3, 3, 2, 5, 1, 6$**. Mickey se întreabă dacă nu cumva zarul este măsluit (având tendinţa să producă de mai multe ori faţa 3 decât ar fi normal dacă zarul ar fi perfect).**

Concepeţi o analiză simplă bazată pe teorema lui Bayes care să-i furnizeze lui Mickey informaţia care-l interesează: [în ce măsură putem spune că] zarul este măsluit [sau nu]?

Explicaţi raţionamentul dumneavoastră.

Veţi presupune că în general fiecare set de 100 de zaruri conţine 5 zaruri măsluite (engl., unfair) în aşa fel încât este favorizată apariţia feţei 3, rezultând următoarea distribuţie de probabilitate a celor şase feţe, $(1, 2, 3, 4, 5, 6)$**:** $P = [0.1,\ 0.1,\ 0.5,\ 0.1,\ 0.1,\ 0.1]$**.**

Răspuns:

Acesta este un exerciţiu [tipic] de punere în evidenţă a noţiunii de ipoteză de *probabilitate maximă a posteriori* (engl., Maximum A posteriori Probability, MAP).

Vă reamintim definiţia:

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}}\, P(h|D),$$

unde D **este setul de date cu care se lucrează, iar** H **este mulţimea de ipoteze considerate.**

În cazul nostru, $D = \{1, 3, 4, 2, 3, 3, 2, 5, 1, 6\}$**, iar** $H = \{FD, LD\}$**, unde am notat cu** FD **zarul corect / cinstit (engl., fair dice) şi cu** LD **zarul măsluit (engl., loaded dice).**

Folosind formula lui Bayes, definiţia de mai sus, poate fi „rafinată" astfel:

$$h_{MAP} \overset{def.}{=} \underset{h \in H}{\operatorname{argmax}}\, P(h|D) \overset{F.B.}{=} \underset{h \in H}{\operatorname{argmax}}\, \frac{P(D|h) \cdot P(h)}{P(D)} = \underset{h \in H}{\operatorname{argmax}}\, P(D|h) \cdot P(h),$$

ultima egalitate având loc datorită faptului că $P(D)$ este o cantitate pozitivă care nu depinde de h.[313]

Aşadar, în cazul de faţă, a determina ipoteza de probablitate maximă a posteriori (h_{MAP}) revine la a determina maximul dintre două produse: $P(D|FD) \cdot P(FD)$ şi $P(D|LD) \cdot P(LD)$.

Facem observaţia că pentru a calcula $P(D|FD)$ şi $P(D|LD)$, vom ţine cont de faptul că aruncările zarului au fost independente unele de altele. Prin urmare, notând $D = \{x_1, x_2, \ldots, x_{10}\}$, vom putea scrie:

$$\begin{aligned} P(D|FD) \cdot P(FD) &= P(x_1, x_2, \ldots, x_{10}|FD) \cdot P(FD) \overset{i.i.d.}{=} \Big(\prod_{i=1}^{10} P(x_i|FD)\Big) \cdot P(FD) \\ &= \left(\frac{1}{6}\right)^{10} \cdot \frac{95}{100} = \frac{1}{2^{10} \cdot 3^{10}} \cdot \frac{19}{20}. \end{aligned}$$

Similar,

$$\begin{aligned} P(D|LD) \cdot P(LD) &= P(x_1, x_2, \ldots, x_{10}|LD) \cdot P(LD) \overset{i.i.d.}{=} \Big(\prod_{i=1}^{10} P(x_i|LD)\Big) \cdot P(LD) \\ &= \left(\frac{1}{10} \cdot \frac{1}{2} \cdot \frac{1}{10} \cdot \frac{1}{10} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{10} \cdot \frac{1}{10} \cdot \frac{1}{10} \cdot \frac{1}{10} \cdot\right) \cdot \frac{5}{100} \\ &= \frac{1}{10^7 \cdot 2^3} \cdot \frac{1}{20} = \frac{1}{2^{10} \cdot 5^7} \cdot \frac{1}{20}. \end{aligned}$$

Aşadar, a vedea care dintre produsele $P(D|FD) \cdot P(FD)$ şi $P(D|LD) \cdot P(LD)$ este mai mare revine la a compara fracţiile $\frac{19}{3^{10}}$ şi $\frac{1}{5^7}$. În loc să facem ridicările la putere (3^{10} şi 5^7), este mai convenabil să logaritmăm, folosind ca bază un număr supra-unitar:

$$\ln \frac{19}{3^{10}} = \ln 19 - \ln 3^{10} = \ln 19 - 10 \ln 3 = 2.9444 - 10.9861 = -8.0417$$

şi

$$\ln \frac{1}{5^7} = -\ln 5^7 = -7 \ln 5 = -11.2661.$$

Conchidem că ipoteza de probabilitate maximă a posteriori este FD, deci că zarul lui Mickey *nu* este măsluit.

***Observaţie*: Se obişnuieşte ca, în loc să se lucreze cu cele două produse ($P(D|FD) \cdot P(FD)$ şi $P(D|LD) \cdot P(LD)$) în mod separat, aşa cum am procedat noi mai sus, să se facă raportul lor,**

$$\frac{P(D|LD) \cdot P(LD)}{P(D|FD) \cdot P(FD)} = \frac{P(LD|D)}{P(FD|D)}.$$

Acest raport se numeşte *raportul de şanse* (engl., odds ratio). (Dacă acest raport este supra-unitar, înseamnă că ipoteza LD este mai plauzibilă decât ipoteza FD.) Mai departe, aplicând logaritmul — fiindcă la calcule probabilitatea $P(D|\ldots)$ se exprimă ca produs de n factori —, se obţine ceea ce în limba

[313] Folosind formula probabilităţii totale, atunci când H este o mulţime discretă, putem exprima $P(D)$ ca $\sum_{h' \in H} P(D|h') \cdot P(h')$. Ar fi util să calculaţi această probabilitate *a priori* în cazul datelor din acest exerciţiu.

engleză se numeşte *log-odds ratio*. (Evident, dacă *log-odds ratio* are valoare pozitivă, ipoteza LD este mai plauzibilă.) Pe datele noastre,

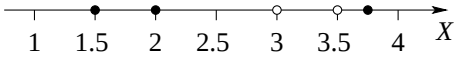
$$\ln \frac{P(LD|D)}{P(FD|D)} = -11.2661 - (-8.0417) = -3.2244 < 0.$$

În concluzie, folosind un astfel de raport (de „potrivire"), vom putea spune nu doar care ipoteză este mai plauzibilă, ci şi *în ce măsură* acea ipoteză (în cazul nostru, FD) este mai plauzibilă decât cealaltă ipoteză (LD).

4. **(Arbori ID3 cu decizii probabiliste, ca ipoteze ML şi respectiv MAP)**

■ ***CMU, 2009 spring, T. Mitchell, midterm, pr. 2***

Se consideră următorul set de date de antrenament din spaţiul real unidimensional:



Este vorba de 5 date caracterizate de atributul real X, împărţite în două clase: clasa 0 constituită din mulţimea $\{3, 3.5\}$, şi clasa 1 – mulţimea $\{1.5, 2, 3.75\}$.

Pe acest set de date se va aplica algoritmul ID3 pentru construirea unor arbori de decizie. Deoarece atributul are valori reale, testele vor fi de forma $X > t$, unde t reprezintă o valoare-prag.

Se notează cu DT^* algoritmul care construieşte arborele de decizie cu număr minim de noduri necesare pentru clasificarea perfectă a datelor de antrenament, şi cu $DT1$ algoritmul care construieşte un arbore de decizie cu un singur nod de test.

***Indicaţie*: Veţi presupune că atunci când algoritmul de învăţare de arbori de decizie găseşte două praguri astfel încât testele $x < t_1$ şi $x < t_2$ produc acelaşi câştig de informaţie, el (adică, algoritmul) va alege pragul cel mai din stânga. (Aşadar, dacă $t_1 < t_2$, atunci el va alege testul $(x < t_1)$).**

a. Care este eroarea la antrenare a lui DT^* pe datele specificate? Dar eroarea la cross-validare folosind metoda "Leave-One-Out" (CVLOO)?

b. Care este eroarea la antrenare a lui $DT1$ pe datele specificate? Dar eroarea la cross-validare folosind metoda "Leave-One-Out"?

În continuare se consideră o nouă clasă de arbori de decizie, care au *etichete probabiliste*. Fiecare nod frunză specifică probabilitatea fiecărei etichete posibile, probabilitate scrisă sub forma raportului dintre datele cu acea etichetă din nodul respectiv şi toate datele din acel nod.

De *exemplu*, un arbore de decizie neavând niciun nod de test, construit pe datele specificate mai sus, clasifică astfel: $P(Y = 1) = 3/5$ şi $P(Y = 0) = 2/5$. Un arbore de decizie cu un singur nod de test (în raport cu valoarea / pragul 2.5) conţine probabilităţile: $P(Y = 1) = 1$ dacă $X \leq 2.5$, şi $P(Y = 1) = 1/3$ dacă $X > 2.5$.

c. Pentru setul de date de mai sus, determinaţi arborele de decizie de tip ML (engl., maximum likelihood), adică acel arbore cu decizii probabiliste care maximizează *verosimilitatea* datelor de antrenament:

$$T_{ML} = \arg\max_T P_T(D), \text{ unde}$$

$$P_T(D) \stackrel{def.}{=} P(D|T) \stackrel{indep.}{=} \prod_{i=1}^{5} P(Y = y_i | X = x_i, T),$$

cu y_i eticheta / clasa instanţei $x_i \in \{1.5, 2, 3, 3.5, 3.75\}$.

d. Se consideră o distribuţie a priori $P(T)$ care penalizează numărul de teste / split-uri din arborele de decizie T, şi anume:

$$P(T) \propto \left(\frac{1}{4}\right)^{splits(T)^2}$$

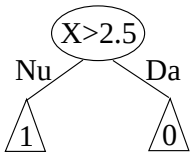
unde $splits(T)$ reprezintă numărul nodurilor de test din arborele T, iar simbolul $\propto$ înseamnă „este proporţional cu".

Pentru acelaşi set de date, folosind această distribuţie a priori $P(T)$, găsiţi arborele de decizie de tip MAP (engl., Maximum A posteriori Probability):

$$T_{MAP} = \arg\max_T P_T(T|D)$$

Răspuns:

a. Dacă se aplică algoritmul $DT1$, întrucât câştigurile de informaţie calculate pentru testele $X > 2.5$ şi $X > 3.625$ sunt 0.419 şi respectiv 0.171, rezultatul este:



Eroarea la antrenare este $1/5$, deoarece punctul 3.75 este clasificat greşit.

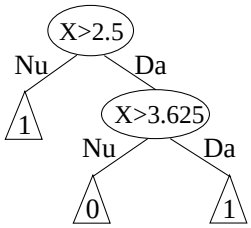
Eroarea la cross-validare cu metoda "Leave-One-Out" se determină astfel:

- $x_1 = 1.5$. **Testul se face relativ la pragul 2.5 — ţinând cont de *Indicaţia* din enunţ —, deci punctul 1.5 este clasificat corect.**
- $x_2 = 2$. **Testul se face relativ la pragul 2.25 — ţinând din nou cont de *Indicaţia* din enunţ —, deci punctul $x_2 = 2$ este clasificat corect.**
- $x_3 = 3$. **Testul se face relativ la pragul 2.75, fiindcă se verifică imediat că IG-ul acestui prag este $1/2$, deci mai mare decât IG-ul pragului 3.625, şi anume $0,311$. Prin urmare, punctul $x_3 = 3$ va fi clasificat fie corect (în cazul în care se consideră că decizia arborelui $DT1$ este 0 pentru $X > 2.75$), fie eronat (în cazul în care se consideră că decizia arborelui $DT1$ este 1 pentru $X > 2.75$). Însă în al doilea caz arborele DT1 s-ar reduce la un singur nod (care este nod frunză), ceea ce este contrar definiţiei sale, aşa că vom reţine doar primul caz.**

- $x_4 = 3.5$. **Testul se face din nou relativ la pragul** 2.5 **— justificarea este similară cu cea de la punctul precedent —, iar punctul** $x_4 = 3.5$ **este clasificat fie corect (în cazul în care se consideră că decizia arborelui** $DT1$ **este** 0 **pentru** $X > 2.5$**), fie eronat (în cazul în care se consideră că decizia arborelui** $DT1$ **este** 1 **pentru** $X > 2.5$**). Însă (din nou!) în al doilea caz arborele DT1 s-ar reduce la un singur nod (frunză), aşa că vom reţine doar primul caz.**
- $x_5 = 3.75$. **Testul se face tot relativ la pragul** 2.5**, iar punctul** $x_5 = 3.75$ **este clasificat greşit.**

Deci eroarea la cross-validare cu metoda "Leave-One-Out" este de $1/5$**, punctul** 3.75 **fiind clasificat eronat.**

b. Dacă se aplică algoritmul DT^***, rezultatul este:**



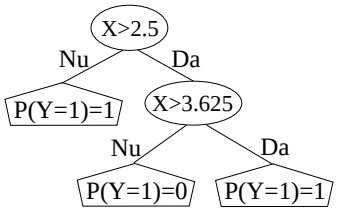
Eroarea la antrenare este bineînţeles 0, deoarece datele de antrenament nu conţin inconsistenţe.

La cross-validare cu metoda "Leave-One-Out", punctele care ar putea genera probleme sunt:

- $x_2 = 2$. **Cele două teste se fac la pragurile** 2.25 **şi la** 3.625 **(ordinea nu contează), deci punctul** $x_2 = 2$ **este corect clasificat.**
- $x_3 = 3$. **Primul test se face la pragul** 2.75**, deci punctul** $x_3 = 3$ **este corect clasificat.**
- $x_4 = 3.5$. **Al doilea test se face la pragul** 3.375**, deci punctul** $x_4 = 3.5$ **este clasificat greşit.**
- $x_5 = 3.75$. **Se face un singur test (la pragul** 2.5**), iar punctul** $x_5 = 3.75$ **este clasificat greşit.**

Deci eroarea la cross-validare pentru arborele DT^* **folosind metoda "Leave-One-Out" este** $2/5$**.**

c. Un arbore de decizie care maximizează probabilităţile datelor de antrenament va fi unul care clasifică în mod perfect aceste date. Deci un arbore de decizie ML poate fi obţinut din arborele construit de DT^***, extins cu etichete probabiliste în frunze, conform figurii alăturate.**



d. Pentru a determina arborele de decizie MAP folosind distribuţia a priori $P(T)$**, va trebui să comparăm probabilităţile a posteriori ale celor 3 arbori de decizie posibili pe setul de date de antrenament. Vom scrie aceste probabilităţi a posteriori folosind formula lui Bayes:**

$$P(T_j \mid D) = \frac{P(D \mid T_j) \cdot P(T_j)}{P(D)} = \frac{\prod_{i=1}^{5} P(Y = y_i \mid T_j, X = x_i) \cdot P(T_j)}{P(D)}$$

Evident, la compararea propriu-zisă a probabilităţilor $P(T_j \mid D)$ **cu** $j = 0, 1, 2$**, nu vom avea nevoie de numitorul** $P(D)$**. Aşadar,**

- **pentru 0 noduri de test:**
  După cum s-a precizat şi în enunţ, acest arbore va avea $P(Y = 1) = 3/5$ şi $P(Y = 0) = 2/5$. Deci

$$P(T_0 \mid D) \propto \left(\frac{3}{5}\right)^3 \cdot \left(\frac{2}{5}\right)^2 \cdot \left(\frac{1}{4}\right)^0 = \frac{3^3 \cdot 2^2}{5^5} = \frac{108}{3125} = 0.0336.$$

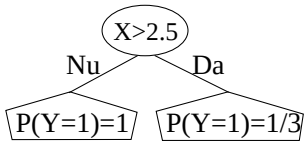
- **pentru un nod de test:**
  Acest arbore va avea probabilităţile: $P(Y = 1) = 1$ dacă $X < 2.5$, şi $P(Y = 1) = 1/3$ dacă $X \geq 2.5$. Prin urmare,

$$P(T_1 \mid D) \propto 1^2 \cdot \left(\frac{2}{3}\right)^2 \cdot \frac{1}{3} \cdot \left(\frac{1}{4}\right)^1 = \frac{1}{27} = 0.037$$

- **pentru două noduri de test:**
  Este vorba de arborele construit la punctul c. Acesta clasifică perfect datele, dar

$$P(T_2) \propto \left(\frac{1}{4}\right)^4 \Rightarrow P(T_2 \mid D) \propto 1 \cdot \left(\frac{1}{4}\right)^4 = \frac{1}{256} = 0.0039$$

Probabilitatea a posteriori maximă o are arborele T_1, deci acesta va fi arborele MAP. Reprezentarea grafică a acestui arbore este:



2.1.2 Algoritmii Bayes Naiv şi Bayes Optimal

5. **(Algoritmul Bayes Naiv: aplicare; comparaţie cu estimarea MLE)**

prelucrare de Liviu Ciortuz, după CMU, 2004 fall, T. Mitchell Z. Bar-Joseph, final exam, pr. 2

a	b	c	K
1	0	1	1
1	1	1	1
0	1	1	0
1	1	0	0
1	0	1	0
0	0	0	1
0	0	0	1
0	0	1	0

Fie setul de date alăturat, cu trei variabile booleene de intrare a, b, c şi o variabilă booleană de ieşire K.

a. Estimaţi probabilităţile $P(K = 1 \mid a = 1, b = 1)$ şi $P(K = 1 \mid a = 1, b = 1, c = 0)$ în sensul verosimilităţii maxime (engl., Maximum Likelihood Estimation, MLE).

b. Cum clasifică algoritmul Bayes Naiv instanţa $(a = 1, b = 1)$?
Dar instanţa $(a = 1, b = 1, c = 0)$?

Răspuns:

a. $P(K = 1 \mid a = 1, b = 1) = \frac{1}{2}$**, fiindcă avem două instanţe în care ambele atribute** ***a, b*** **sunt adevărate, dintre care una este clasificată** $K = 1$**, iar cealaltă** $K = 0$**.**

$P(K = 1 \mid a = 1, b = 1, c = 0) = 0$**, fiindcă există o singură instanţă cu** $a = 1, b = 1, c = 0$ **în setul de antrenament şi ea este clasificată** $K = 0$**.**

b. Cazul $(a = 1, b = 1)$**:**

$$\begin{aligned}
\hat{k}_{MAP} &= \underset{k\in\{0,1\}}{\operatorname{argmax}} P(K = k|a = 1, b = 1) = \\
&= \underset{k\in\{0,1\}}{\operatorname{argmax}} \frac{P(a = 1, b = 1|K = k) \cdot P(K = k)}{P(a = 1, b = 1)} = \\
&= \underset{k\in\{0,1\}}{\operatorname{argmax}} P(a = 1, b = 1|K = k) \cdot P(K = k) = \\
&= \underset{k\in\{0,1\}}{\operatorname{argmax}} P(a = 1|K = k) \cdot P(b = 1|K = k) \cdot P(K = k)
\end{aligned}$$

Avem:

$$\begin{aligned}
p_0 &= P(a = 1|K = 0) \cdot P(b = 1|K = 0) \cdot P(K = 0) = \frac{2}{4} \cdot \frac{2}{4} \cdot \frac{4}{8} = \frac{1}{8} \\
p_1 &= P(a = 1|K = 1) \cdot P(b = 1|K = 1) \cdot P(K = 1) = \frac{2}{4} \cdot \frac{1}{4} \cdot \frac{4}{8} = \frac{1}{16}
\end{aligned}$$

Prin urmare, $p_0 > p_1$**. Aşadar, clasificatorul Bayes Naiv va prezice** $K = 0$ **pentru instanţa** $(a = 1, b = 1)$ **cu probabilitatea**

$$\begin{aligned}
&P(K = 0 \mid a = 1, b = 1) \\
&= \frac{P(a = 1, b = 1 \mid K = 0) \cdot P(K = 0)}{P(a = 1, b = 1 \mid K = 0) \cdot P(K = 0) + P(a = 1, b = 1 \mid K = 1) \cdot P(K = 1)} \\
&= \frac{p_0}{p_0 + p_1} = \frac{\frac{1}{8}}{\frac{1}{8} + \frac{1}{16}} = \frac{2}{3}
\end{aligned}$$

Observaţie**: Se constată că probabilitatea** $P(K = 1|a = 1, b = 1)$ **estimată în sensul MLE (adică,** $1/2$**; vedeţi punctul** a**) diferă de probabilitatea (*a posteriori*) cu care algoritmul Bayes Naiv clasifică instanţa** $(a = 1, b = 1)$ **ca aparţinând clasei** $K = 1$ **(adică,** $2/3$**). Explicaţia rezidă în faptul că presupoziţia de independenţă condiţională asumată de către algoritmul Bayes Naiv nu este validă. Într-adevăr,** $P(a = 0|K = 1) = 2/4$ **şi** $P(a = 0|b = 1, K = 1) = 0$**, deci** $P(a = 0|K = 1) \neq P(a = 0|b = 1, K = 1)$**.**

Cazul $(a = 1, b = 1, c = 0)$**:**

$$\begin{aligned}
\hat{k}_{MAP} &= \underset{k\in\{0,1\}}{\operatorname{argmax}} P(K = k|a = 1, b = 1, c = 0) = \\
&= \underset{k\in\{0,1\}}{\operatorname{argmax}} \frac{P(a = 1, b = 1, c = 0|K = k) \cdot P(K = k)}{P(a = 1, b = 1, c = 0)} = \\
&= \underset{k\in\{0,1\}}{\operatorname{argmax}} P(a = 1, b = 1, c = 0|K = k) \cdot P(K = k) = \\
&= \underset{k\in\{0,1\}}{\operatorname{argmax}} P(a = 1|K = k) \cdot P(b = 1|K = k) \cdot P(c = 0|K = k) \cdot P(K = k)
\end{aligned}$$

Avem:

$$\begin{aligned} p_0 &= P(a=1|K=0)\cdot P(b=1|K=0)\cdot P(c=0|K=0)\cdot P(K=0) = \\ &= \frac{2}{4}\cdot\frac{2}{4}\cdot\frac{1}{4}\cdot\frac{4}{8} = \frac{1}{32} \\ p_1 &= P(a=1|K=1)\cdot P(b=1|K=1)\cdot P(c=0|K=1)\cdot P(K=1) = \\ &= \frac{2}{4}\cdot\frac{1}{4}\cdot\frac{2}{4}\cdot\frac{4}{8} = \frac{1}{32} \end{aligned}$$

Întrucât $p_0 = p_1$, clasificatorul Bayes Naiv va prezice $K = 0$ sau $K = 1$ cu aceeaşi probabilitate $(1/2)$.

***Notă*: *Observaţia* de mai sus este valabilă şi pentru cazul $(a = 1, b = 1, c = 0)$.**

6. **(Algoritmul Bayes Naiv: aplicaţie la filtrarea emailurilor spam)**

■ • ***CMU, 2009 spring, Ziv Bar-Joseph, midterm, pr. 2***

Circa $2/3$ dintre emailurile tale sunt spam, aşadar te-ai decis să descarci de pe internet un filtru spam open-source care utilizează un clasificator Bayes Naiv.

Presupunem că ai strâns următoarele emailuri spam şi non-spam (engl., regular), şi de asemenea că doar trei cuvinte sunt discriminative pentru această clasificare, deci fiecare email este reprezentat ca un vector de 3 componente binare, fiecare dintre ele indicând dacă respectivul cuvânt este conţinut (sau nu) în email.

'study'	'free'	'money'	Category	count
1	0	0	**Regular**	1
0	0	1	**Regular**	1
1	0	0	**Regular**	1
1	1	0	**Regular**	1
0	1	0	**Spam**	4
0	1	1	**Spam**	4

a. Descoperi că filtrul spam open-source foloseşte o probabilitate a priori $P(\text{spam}) = 0.1$. Explică în mod succint de ce crezi că această alegere este rezonabilă.

b. Calculează următorii parametri ai modelului prin metoda estimării de verosimilitate maximă (MLE), folosind netezire (engl., smoothing) de tip "add-one" (regula lui Laplace).

$P(\text{study}|\text{spam}) =$

$P(\text{study}|\text{regular}) =$

$P(\text{free}|\text{spam}) =$

$P(\text{free}|\text{regular}) =$

$P(\text{money}|\text{spam}) =$

$P(\text{money}|\text{regular}) =$

c. Folosind probabilitatea a priori şi probabilităţile condiţionate de mai sus, calculează probabilitatea ca mesajul $s =$ "money for psychology study" să fie spam, adică $P(\text{spam}|\ s)$.

d. Care ar trebui să fie valoarea probabilităţii a priori $P(\text{spam})$ în cazul în care dorim ca mesajul de mai sus să aibă aceeaşi probabilitate de fi spam respectiv non-spam (i.e., el va fi clasificat ca spam cu probabilitatea 0.5)?

Răspuns:

a. Diferenţa dintre $P_{MLE}(\text{Category}|\text{Spam}) = 2/3$ şi probabilitatea a priori indicată în enunţ (0.1) se explică prin faptul că se preferă trecerea prin filtru a unor emailuri spam, decât să fie marcate ca spam unele emailuri non-spam şi astfel să nu ajungă în Inbox.

b. Dacă nu am folosi regula de tip "add-one" a lui Laplace (pentru „netezirea" probabilităţilor), parametrii modelului ar avea valorile (obţinute prin metoda estimării de verosimilitate maximă – MLE) care apar mai jos în partea stângă. Folosind regula lui Laplace, parametrii primesc valorile indicate în partea dreaptă. Observaţi că apariţiile lui 2 de la numitorul fracţiilor corespund numărului de valori pentru fiecare dintre atributele / variabilele de intrare.

$$P(\textbf{study}|\textbf{spam}) = \frac{0}{8} = 0 \qquad P(\textbf{study}|\textbf{spam}) \overset{Laplace}{=} \frac{0+1}{8+2} = \frac{1}{10}$$

$$P(\textbf{study}|\textbf{regular}) = \frac{3}{4} \qquad P(\textbf{study}|\textbf{regular}) \overset{Laplace}{=} \frac{3+1}{4+2} = \frac{2}{3}$$

$$P(\textbf{free}|\textbf{spam}) = \frac{8}{8} = 1 \qquad P(\textbf{free}|\textbf{spam}) \overset{Laplace}{=} \frac{8+1}{8+2} = \frac{9}{10}$$

$$P(\textbf{free}|\textbf{regular}) = \frac{1}{4} \qquad P(\textbf{free}|\textbf{regular}) \overset{Laplace}{=} \frac{1+1}{4+2} = \frac{1}{3}$$

$$P(\textbf{money}|\textbf{spam}) = \frac{4}{8} = \frac{1}{2} \qquad P(\textbf{money}|\textbf{spam}) \overset{Laplace}{=} \frac{4+1}{8+2} = \frac{1}{2}$$

$$P(\textbf{money}|\textbf{regular}) = \frac{1}{4} \qquad P(\textbf{money}|\textbf{regular}) \overset{Laplace}{=} \frac{1+1}{4+2} = \frac{1}{3}$$

c. Avem mesajul $s =$ "money for psychology study", deci trebuie să calculăm $P(\textbf{spam} \mid s) = P(\textbf{spam} \mid \textbf{study}, \neg\textbf{free}, \textbf{money})$.

$$P(\textbf{spam} \mid s) \overset{F.\ Bayes}{=}$$

$$= \frac{P(\textbf{study, } \neg\textbf{free, money} \mid \textbf{spam}) \cdot P(\textbf{spam})}{P(\textbf{study,}\neg\textbf{free,money} \mid \textbf{spam})P(\textbf{spam}) + P(\textbf{study,}\neg\textbf{free,money} \mid \textbf{reg})P(\textbf{reg})}$$

Calculăm probabilităţile folosind ipoteza de independenţă condiţională:

$$P(\textbf{study, } \neg\textbf{free, money}|\textbf{spam}) = P(\textbf{study}|\textbf{spam}) \cdot P(\neg\textbf{free}|\textbf{spam}) \cdot P(\textbf{money}|\textbf{spam})$$
$$= \frac{1}{10} \cdot \frac{1}{10} \cdot \frac{1}{2} = \frac{1}{200}$$
$$P(\textbf{study, } \neg\textbf{free, money}|\textbf{reg}) = P(\textbf{study}|\textbf{reg}) \cdot P(\neg\textbf{free}|\textbf{reg}) \cdot P(\textbf{money}|\textbf{reg})$$
$$= \frac{2}{3} \cdot \frac{2}{3} \cdot \frac{1}{3} = \frac{4}{27}$$

Înlocuind valorile în formulă, obţinem: $P(\textbf{spam}|\ s) = \dfrac{\frac{1}{200} \cdot \frac{1}{10}}{\frac{1}{200} \cdot \frac{1}{10} + \frac{4}{27} \cdot \frac{9}{10}} \approx 0.0037$

Aceasta este o probabilitate mică. Se observă însă că dacă nu am fi folosit regula lui Laplace, probabilitatea ca emailul s să fie spam ar fi fost 0. Aceasta se datorează faptului că niciunul dintre emailurile spam din datele de antrenament nu conţine cuvântul *study*, care apare însă în emailul s.

d. Dacă notăm cu p probabilitatea a priori cerută, P(spam), ştiind că P(spam$|s) = 0.5$, putem scrie:

$$0.5 = \frac{\frac{1}{200} \cdot p}{\frac{1}{200} \cdot p + \frac{4}{27} \cdot (1-p)} \Leftrightarrow \frac{1}{2} = \frac{\frac{p}{200}}{\frac{p}{200} + \frac{4}{27} - \frac{4p}{27}} \Leftrightarrow \frac{2p}{200} = \frac{p}{200} + \frac{4}{27} - \frac{4p}{27}$$

$$\Leftrightarrow 54p = 27p + 800 - 800p \Leftrightarrow p = \frac{800}{827} \approx 0.9673$$

7. (Algoritmul Bayes Naiv şi algoritmul Bayes Optimal; comparaţie relativ la numărului de parametri)

■ • ***CMU, 2008 fall, Eric Xing, HW1, pr. 2***

Fie următoarea problemă de clasificare: X_1 şi X_2 sunt variabile aleatoare observabile, Y este eticheta clasei asignate fiecărei instanţe observate, conform tabelului alăturat.
În acest exerciţiu veţi compara rezultatele care se obţin în urma antrenării pe acest set de date de către doi algoritmi de clasificare: Bayes Naiv şi Bayes Comun (engl., Joint Bayes), care este numit adeseori şi *Bayes Optimal* (engl., Optimal Bayes).

X_1	X_2	Y	**Nr. apariţii**
0	0	0	2
0	0	1	18
1	0	0	4
1	0	1	1
0	1	0	4
0	1	1	1
1	1	0	2
1	1	1	18

***Comentariu*:** Pentru cel de-al doilea algoritm, denumirea *Bayes Comun* se datorează faptului că acest clasificator lucrează cu distribuţia comună a variabilelor / atributelor de intrare, în vreme ce denumirea *Bayes Optimal* este justificată de faptul că nicio altă distribuţie probabilistă asupra datelor de intrare nu poate conduce la o eroare medie la clasificare mai mică decât cea obţinută de către acest clasificator. (În cele ce urmează, vom opta pentru cea de-a doua denumire.)

a. Clasificaţi instanţa $X_1 = 0, X_2 = 0$ folosind clasificatorul Bayes Naiv.

b. Clasificaţi instanţa $X_1 = 0, X_2 = 0$ folosind clasificatorul Bayes Optimal.

c. Notăm cu P_{NB} şi respectiv P_{JB} valoarea probabilităţii $P(Y = 1 \mid X_1 = 0, X_2 = 0)$ calculate pentru clasificatorul Bayes Naiv, respectiv pentru clasificatorul Bayes Optimal. De ce diferă cele două valori? *Sugestie*: Calculaţi $P(X_1, X_2 \mid Y)$.

d. Care ar fi situaţia pentru P_{NB} şi P_{JB} de la întrebarea precedentă în situaţia în care datele observate ar proveni din tabelul alăturat?

X_1	X_2	Y	**Nr. apariţii**
0	0	0	3
0	0	1	9
1	0	0	3
1	0	1	9
0	1	0	3
0	1	1	9
1	1	0	3
1	1	1	9

e. De câţi parametri independenţi (i.e., probabilităţi estimate) este nevoie în total pentru a construi clasificatorul Bayes Naiv? Dar în cazul clasificatorului Bayes Optimal?

Răspundeţi la aceste întrebări şi în cazul general, cînd se folosesc n variabile binare observate. Comentaţi rezultatul.

Răspuns:

a. Clasificatorul Bayes Naiv face predicţia pentru valoarea lui Y după formula de mai jos:

$$\hat{y}_{NB} = \underset{y \in \{0,1\}}{\operatorname{argmax}} P(X_1 = 0|Y = y) \cdot P(X_2 = 0|Y = y) \cdot P(Y = y)$$

Avem:

$$\begin{aligned} p_0 &\overset{not.}{=} P(X_1 = 0|Y = 0) \cdot P(X_2 = 0|Y = 0) \cdot P(Y = 0) \\ &\overset{MLE}{=} \frac{6}{12} \cdot \frac{6}{12} \cdot \frac{12}{50} = \frac{3}{50} = \frac{6}{100} \end{aligned}$$

$$\begin{aligned} p_1 &\overset{not.}{=} P(X_1 = 0|Y = 1) \cdot P(X_2 = 0|Y = 1) \cdot P(Y = 1) \\ &\overset{MLE}{=} \frac{19}{38} \cdot \frac{19}{38} \cdot \frac{38}{50} = \frac{19}{100} \end{aligned}$$

Întrucât $p_0 < p_1$, clasificatorul Bayes Naiv va prezice $Y = 1$ pentru instanţa $(X_1 = 0, X_2 = 0)$.

b. Deoarece clasificatorul Bayes Optimal nu lucrează cu presupunerea de independenţă condiţională a atributelor de intrare în raport cu atributul de ieşire, el va face predicţia folosind formula de mai jos:

$$\hat{y}_{JB} = \underset{y \in \{0,1\}}{\operatorname{argmax}} P(X_1 = 0, X_2 = 0|Y = y) \cdot P(Y = y)$$

Probabilităţile din formulă sunt, ca şi în cazul clasificatorului Bayes Naiv, cele estimate din distribuţia datelor de antrenament.

Aşadar, avem:

$$p'_0 \overset{not.}{=} P(X_1 = 0, X_2 = 0 \mid Y = 0) \cdot P(Y = 0) \overset{MLE}{=} \frac{2}{12} \cdot \frac{12}{50} = \frac{2}{50}$$

$$p'_1 \overset{not.}{=} P(X_1 = 0, X_2 = 0 \mid Y = 1) \cdot P(Y = 1) \overset{MLE}{=} \frac{18}{38} \cdot \frac{38}{50} = \frac{18}{50}$$

Fiindcă $p'_0 < p'_1$, clasificatorul Bayes Optimal va prezice tot $Y = 1$.

c. Vom calcula cele două valori pentru $P(Y = 1 \mid X_1 = 0, X_2 = 0)$:

$$\begin{aligned} P_{NB} &\overset{not.}{=} P(Y = 1 \mid X_1 = 0, X_2 = 0) \\ &\overset{F.\ Bayes}{=} \frac{P(X_1 = 0, X_2 = 0 \mid Y = 1) \cdot P(Y = 1)}{P(X_1 = 0, X_2 = 0 \mid Y = 1)P(Y = 1) + P(X_1 = 0, X_2 = 0 \mid Y = 0)P(Y = 0)} \\ &\overset{indep.\ cdt.}{=} \frac{P(X_1 = 0|Y = 1) \cdot P(X_2 = 0|Y = 1) \cdot P(Y = 1)}{\begin{array}{r} P(X_1 = 0|Y = 0) \cdot P(X_2 = 0|Y = 0) \cdot P(Y = 0) + \\ P(X_1 = 0|Y = 1) \cdot P(X_2 = 0|Y = 1) \cdot P(Y = 1) \end{array}} \\ &= \frac{p_1}{p_0 + p_1} = \frac{\frac{19}{100}}{\frac{6}{100} + \frac{19}{100}} = \frac{19}{25} \end{aligned}$$

$$P_{JB} \overset{not.}{=} P(Y=1 \mid X_1=0, X_2=0)$$
$$\overset{F.\ Bayes}{=} \frac{P(X_1=0, X_2=0 \mid Y=1)\cdot P(Y=1)}{P(X_1=0, X_2=0 \mid Y=1)P(Y=1) + P(X_1=0, X_2=0 \mid Y=0)P(Y=0)}$$
$$= \frac{p_1'}{p_0' + p_1'} = \frac{\frac{18}{50}}{\frac{2}{50} + \frac{18}{50}} = \frac{18}{20}$$

Cele două valori diferă deoarece presupunerea de independenţă condiţională a variabilelor X_1 şi X_2 în raport cu Y făcută de clasificatorul Bayes Naiv este falsă. Acest lucru se poate vedea uşor din valorile estimate pentru $P(X_1 = 0, X_2 = 0 \mid Y = 0)$, $P(X_1 = 0 \mid Y = 0)$ şi $P(X_2 = 0 \mid Y = 0)$:

$$\left.\begin{aligned} P(X_1=0, X_2=0 \mid Y=0) &\overset{MLE}{=} \frac{2}{12} \\ P(X_1=0 \mid Y=0)\cdot P(X_2=0 \mid Y=0) &\overset{MLE}{=} \frac{6}{12}\cdot\frac{6}{12} = \frac{1}{4} \end{aligned}\right\} \Rightarrow$$
$$\Rightarrow P(X_1=0, X_2=0 \mid Y=0) \neq P(X_1=0 \mid Y=0)\cdot P(X_2=0 \mid Y=0) \Rightarrow$$
$$\Rightarrow P(X_1, X_2 \mid Y) \neq P(X_1 \mid Y)\cdot P(X_2 \mid Y)$$

Aşadar, variabilele X_1 şi X_2 nu sunt independente condiţional în raport cu variabila de ieşire Y.

d. Vom calcula cele două valori pentru $P(Y = 1 \mid X_1 = 0, X_2 = 0)$ în cazul noilor date:

$$P_{NB} = P(Y=1 \mid X_1=0, X_2=0)$$
$$= \frac{P(X_1=0, X_2=0 \mid Y=1)\cdot P(Y=1)}{P(X_1=0, X_2=0 \mid Y=1)P(Y=1) + P(X_1=0, X_2=0 \mid Y=0)P(Y=0)}$$
$$= \frac{P(X_1=0|Y=1)\cdot P(X_2=0|Y=1)\cdot P(Y=1)}{\begin{aligned}&P(X_1=0|Y=0)\cdot P(X_2=0|Y=0)\cdot P(Y=0) + \\ &\qquad P(X_1=0|Y=1)\cdot P(X_2=0|Y=1)\cdot P(Y=1)\end{aligned}}$$
$$= \frac{\frac{18}{36}\cdot\frac{18}{36}\cdot\frac{36}{48}}{\frac{6}{12}\cdot\frac{6}{12}\cdot\frac{12}{48} + \frac{18}{36}\cdot\frac{18}{36}\cdot\frac{36}{48}} = \frac{\frac{9}{48}}{\frac{3}{48} + \frac{9}{48}} = \frac{9}{12} = \frac{3}{4}$$

$$P_{JB} = P(Y=1 \mid X_1=0, X_2=0)$$
$$= \frac{P(X_1=0, X_2=0 \mid Y=1)\cdot P(Y=1)}{P(X_1=0, X_2=0 \mid Y=1)P(Y=1) + P(X_1=0, X_2=0 \mid Y=0)P(Y=0)}$$
$$= \frac{\frac{9}{36}\cdot\frac{36}{48}}{\frac{3}{12}\cdot\frac{12}{48} + \frac{9}{36}\cdot\frac{36}{48}} = \frac{\frac{9}{48}}{\frac{3}{48} + \frac{9}{48}} = \frac{9}{12} = \frac{3}{4}$$

Aşadar, în acest caz avem $P_{NB} = P_{JB}$.

De fapt, se poate constata uşor că în cazul distribuţiei probabiliste date la acest punct al problemei se verifică independenţa condiţională a variabilelor X_1 şi X_2 în raport cu Y. Prin urmare, predicţiile făcute de cei doi clasificatori, Bayes Naiv şi Bayes Optimal, vor coincide întotdeauna.

e. În contextul problemei noastre, clasificatorul Bayes Naiv are nevoie de estimările următoarelor probabilităţi:

$$
\begin{array}{lcl}
P(Y=0) & \Rightarrow & P(Y=1) = 1 - P(Y=0) \\
P(X_1=0 \mid Y=0) & \Rightarrow & P(X_1=1 \mid Y=0) = 1 - P(X_1=0 \mid Y=0) \\
P(X_1=0 \mid Y=1) & \Rightarrow & P(X_1=1 \mid Y=1) = 1 - P(X_1=0 \mid Y=1) \\
P(X_2=0 \mid Y=0) & \Rightarrow & P(X_2=1 \mid Y=0) = 1 - P(X_2=0 \mid Y=0) \\
P(X_2=0 \mid Y=1) & \Rightarrow & P(X_2=1 \mid Y=1) = 1 - P(X_2=0 \mid Y=1)
\end{array}
$$

Avem nevoie, prin urmare, doar de 5 valori pentru a construi complet clasificatorul Bayes Naiv.

În cazul general, dacă avem n variabile de intrare, avem nevoie de estimări pentru probabilităţile $P(Y)$, $P(X_i \mid Y)$ şi $P(X_i \mid \neg Y)$ pentru $i = \overline{1,n}$, deci $2n+1$ valori.

Pentru clasificatorul Bayes Optimal avem nevoie de:

$P(Y=0)$ — $P(Y=1) = 1 - P(Y=0)$

$P(X_1=0, X_2=0 \mid Y=0)$
$P(X_1=0, X_2=1 \mid Y=0)$
$P(X_1=1, X_2=0 \mid Y=0)$
— **$P(X_1=1, X_2=1 \mid Y=0)$ se poate determina din celelalte 3 valori, aşa cum vom arăta mai jos.**

$P(X_1=0, X_2=0 \mid Y=1)$
$P(X_1=0, X_2=1 \mid Y=1)$
$P(X_1=1, X_2=0 \mid Y=1)$
— **Similar, $P(X_1=1, X_2=1 \mid Y=1)$ se poate determina din celelalte 3 valori.**

Notăm evenimentul $X_1 = 0$ cu $\boldsymbol{A}$, $X_2 = 0$ cu $\boldsymbol{B}$ şi $Y = 0$ cu $\boldsymbol{C}$. Ştim că:

$$
\begin{array}{l}
\Omega = (A \wedge B) \vee (\neg A \wedge B) \vee (A \wedge \neg B) \vee (\neg A \wedge \neg B) \Rightarrow \\
\Omega \wedge C = ((A \wedge B) \vee (\neg A \wedge B) \vee (A \wedge \neg B) \vee (\neg A \wedge \neg B)) \wedge C \Rightarrow \\
C = ((A \wedge B) \wedge C) \vee ((\neg A \wedge B) \wedge C) \vee ((A \wedge \neg B) \wedge C) \vee ((\neg A \wedge \neg B) \wedge C)
\end{array}
$$

De asemenea, deoarece toate evenimentele din partea dreaptă a egalităţii sunt disjuncte două câte două, putem scrie egalitatea de mai sus şi cu probabilităţi:

$$
\begin{array}{l}
P(C) = P((A \wedge B) \wedge C) + P((\neg A \wedge B) \wedge C) + P((A \wedge \neg B) \wedge C) + P((\neg A \wedge \neg B) \wedge C) \\
\Rightarrow 1 = \dfrac{P((A \wedge B) \wedge C)}{P(C)} + \dfrac{P((\neg A \wedge B) \wedge C)}{P(C)} + \dfrac{P((A \wedge \neg B) \wedge C)}{P(C)} + \dfrac{P((\neg A \wedge \neg B) \wedge C)}{P(C)} \\
\Rightarrow 1 = P(A, B \mid C) + P(\neg A, B \mid C) + P(A, \neg B \mid C) + P(\neg A, \neg B \mid C) \\
\Rightarrow P(\neg A, \neg B \mid C) = 1 - (P(A, B \mid C) + P(\neg A, B \mid C) + P(A, \neg B \mid C))
\end{array}
$$

Prin urmare, ştiind 3 valori o putem afla şi pe a patra. La fel şi pentru $\neg C$.

Pentru a avea un clasificator Bayes Optimal complet avem deci nevoie de 7 valori diferite.

În cazul general, pentru n variabile de intrare, avem nevoie de probabilităţile $P(Y)$, $P(\tilde{X_1}, \cdots, \tilde{X_n} \mid Y)$ şi $P(\tilde{X_1}, \cdots, \tilde{X_n} \mid \neg Y)$, unde

$$\tilde{X_i} \in \{X_i, \neg X_i\}\ \forall i \in \overline{1,n} \qquad \text{şi} \qquad (\tilde{X_1}, \cdots, \tilde{X_n}) \neq (\neg X_1, \cdots, \neg X_n).$$

Avem, deci, $2(2^n - 1) + 1 = 2^{n+1} - 1$ valori.

Se observă că algoritmul Bayes Naiv foloseşte un număr liniar de parametri (în raport cu n, numărul de atribute de intrare), în vreme ce algoritmul Bayes Optimal foloşte un număr exponenţial de parametri (în raport cu acelaşi n).

8. **(Calculul parametrilor pentru clasificatorul Bayes Naiv pornind de la distribuţia comună a variabilelor; comparaţie între algoritmii Bayes Naiv şi Bayes Optimal)**

prelucrare de L. Ciortuz, după

■ • ***CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh, midterm, pr. 2.1***

Fie P o distribuţie de *probabilitate comună* peste variabilele aleatoare booleene x_1, x_2 şi y.

a. Exprimaţi $P(y = 0 \mid x_1, x_2)$ în funcţie de $P(x_1, x_2, y = 0)$ şi $P(x_1, x_2, y = 1)$.

În cele ce urmează vom considera distribuţia P, definită conform tabelului alăturat.

x_1	x_2	y	$P(x_1, x_2, y)$
0	0	0	0.15
0	0	1	0.25
0	1	0	0.05
0	1	1	0.08
1	0	0	0.10
1	0	1	0.02
1	1	0	0.20
1	1	1	0.15

b. Pornind de la această distribuţie comună, calculaţi probabilităţile necesare pentru clasificare bayesiană naivă. Completaţi tabelele de mai jos:

y	$P(y)$
$y = 0$	
$y = 1$	

$P(x_1 \mid y)$	$x_1 = 0$	$x_1 = 1$
$y = 0$		
$y = 1$		

$P(x_2 \mid y)$	$x_2 = 0$	$x_2 = 1$
$y = 0$		
$y = 1$		

c. Cât este probabilitatea $P(y = 1 \mid x_1 = 1, x_2 = 0)$ calculată de către clasificatorul Bayes Naiv?

d. Cât este probabilitatea $P(y = 1 \mid x_1 = 1, x_2 = 0)$ calculată de către clasificatorul Bayes Optimal? (Vă readucem aminte că acest algoritm este similar cu algoritmul Bayes Naiv, însă nu foloseşte presupoziţia de independenţă condiţională a atributelor.)

e. Răspunsurile la precedentele două întrebări ar trebui să fie diferite. Care este motivul? Justificaţi.

Răspuns:

a. Aplicând definiţia probabilităţii condiţionate şi apoi proprietatea de aditivitate numărabilă din definiţia funcţiei de probabilitate, obţinem:

$$P(y = 0 \mid x_1, x_2) = \frac{P(x_1, x_2, y = 0)}{P(x_1, x_2)} = \frac{P(x_1, x_2, y = 0)}{P(x_1, x_2, y = 0) + P(x_1, x_2, y = 1)}$$

b. Probabilităţile cerute în enunţ sunt $P(x_1 \mid y)$, $P(x_2 \mid y)$ şi $P(y)$.

$P(y)$ **este o *probabilitate marginală* a distribuţiei comune, deci se calculează astfel:**

$$\begin{aligned} P(y = 0) &= P(x_1 = 0, x_2 = 0, y = 0) + P(x_1 = 0, x_2 = 1, y = 0) + \\ &\quad + P(x_1 = 1, x_2 = 0, y = 0) + P(x_1 = 1, x_2 = 1, y = 0) \\ &= 0.15 + 0.05 + 0.1 + 0.2 = 0.5 \\ P(y = 1) &= 1 - P(y = 0) = 0.5 \end{aligned}$$

$P(x_1 \mid y)$ se calculează folosind din nou definiţia probabilităţii condiţionate şi formula probabilităţii totale:

$$P(x_1 = 0 \mid y = 0) = \frac{P(x_1 = 0, y = 0)}{P(y = 0)} = \frac{P(x_1 = 0, y = 0)}{P(x_1 = 0, y = 0) + P(x_1 = 1, y = 0)}$$

Probabilităţile implicate în formulă sunt probabilităţi marginale ale distribuţiei comune şi se calculează astfel:

$$P(x_1 = 0, y = 0) = P(x_1 = 0, x_2 = 0, y = 0) + P(x_1 = 0, x_2 = 1, y = 0) = 0.15 + 0.05 = 0.2$$

$$P(x_1 = 1, y = 0) = P(x_1 = 1, x_2 = 0, y = 0) + P(x_1 = 1, x_2 = 1, y = 0) = 0.1 + 0.2 = 0.3$$

Prin urmare,

$$P(x_1 = 0 \mid y = 0) = \frac{0.2}{0.2 + 0.3} = 0.4,$$

iar

$$P(x_1 = 1 \mid y = 0) = 1 - P(x_1 = 0 \mid y = 0) = 0.6$$

Analog se calculează şi celelalte probabilităţi corespunzătoare lui $P(x_1 \mid y)$:

$$P(x_1 = 0 \mid y = 1) = \frac{P(x_1 = 0, y = 1)}{P(y = 1)} = \frac{0.25 + 0.08}{0.5} = 0.66$$
$$P(x_1 = 1 \mid y = 1) = 1 - P(x_1 = 0 \mid y = 1) = 0.34$$

$P(x_2 \mid y)$ se calculează în acelaşi mod.

Putem completa tabelele următoare cu valorile numerice ale probabilităţilor calculate la acest punct:

y	$P(y)$
$y = 0$	0.5
$y = 1$	0.5

$P(x_1 \mid y)$	$x_1 = 0$	$x_1 = 1$
$y = 0$	0.40	0.60
$y = 1$	0.66	0.34

$P(x_2 \mid y)$	$x_2 = 0$	$x_2 = 1$
$y = 0$	0.50	0.50
$y = 1$	0.54	0.46

c. Clasificatorul Bayes Naiv face presupunerea de *independenţă condiţională* a variabilelor, deci:

$$\begin{aligned}
P(y = 1 \mid x_1 = 1, x_2 = 0) &\overset{Bayes}{=} \frac{P(x_1 = 1, x_2 = 0 \mid y = 1) \cdot P(y = 1)}{P(x_1 = 1, x_2 = 0)} \\
&= \frac{P(x_1 = 1, x_2 = 0 \mid y = 1) \cdot P(y = 1)}{P(x_1 = 1, x_2 = 0 \mid y = 1)P(y = 1) + P(x_1 = 1, x_2 = 0 \mid y = 0)P(y = 0)} \overset{indep.\ cdt.}{=} \\
&= \frac{P(x_1 = 1 \mid y = 1) \cdot P(x_2 = 0 \mid y = 1) \cdot P(y = 1)}{P(x_1 = 1|y = 1)P(x_2 = 0|y = 1)P(y = 1) + P(x_1 = 1|y = 0)P(x_2 = 0|y = 0)P(y = 0)} \\
&= \frac{0.34 \cdot 0.54 \cdot 0.5}{0.34 \cdot 0.54 \cdot 0.5 + 0.6 \cdot 0.5 \cdot 0.5} \approx 0.3796
\end{aligned}$$

d. Clasificatorul Bayes Optimal nu face niciun fel de presupunere, deci folosind o formulă similară cu cea obţinută la punctul a, vom avea:

$$\begin{aligned}
P(y = 1 \mid x_1 = 1, x_2 = 0) &= \frac{P(x_1 = 1, x_2 = 0, y = 1)}{P(x_1 = 1, x_2 = 0, y = 1) + P(x_1 = 1, x_2 = 0, y = 0)} \\
&= \frac{0.02}{0.02 + 0.1} = 0.1(6).
\end{aligned}$$

e. Valorile calculate de clasificatorul Bayes Naiv şi de clasificatorul Bayes Optimal pentru $P(y = 1 \mid x_1 = 1, x_2 = 0)$ sunt diferite deoarece presupunerea de independenţă condiţională făcută de clasificatorul Bayes Naiv nu este adevărată. Într-adevăr, se observă că variabilele x_1 şi x_2 nu sunt independente condiţional în raport cu variabila y:

$$P(x_1 = 1, x_2 = 0 \mid y = 1) = \frac{P(x_1 = 1, x_2 = 0, y = 1)}{P(y = 1)} = \frac{0.02}{0.5} = 0.04$$
$$P(x_1 = 1 \mid y = 1) \cdot P(x_2 = 0 \mid y = 1) = 0.34 \cdot 0.54 = 0.1836 \neq 0.04$$

9. (Clasificare bayesiană: un caz particular)

CMU, 2010 spring, E. Xing, A. Singh, T. Mitchell, midterm, pr. 2.2

Se consideră variabilele aleatoare $X_1,\ X_2,\ X_3$ şi X_4. Aceste variabile sunt independente condiţional două câte două în raport cu variabila Y, cu excepţia perechii $X_3,\ X_4$. (Aşadar, dacă am aplica algoritmul Bayes Naiv, acesta ar produce erori de clasificare.)

Cum am putea modifica regula de decizie a algoritmului Bayes Naiv pentru a ţine cont de această particularitate a datelor?

Răspuns:

Întrucât are loc egalitatea

$$\operatorname*{argmax}_y P(Y = y \mid X_1, X_2, X_3, X_4) =$$
$$\operatorname*{argmax}_y P(X_1 \mid Y = y) \cdot P(X_2 \mid Y = y) \cdot P(X_3, X_4 \mid Y = y) \cdot P(Y = y),$$

urmează că regula de decizie a algoritmului Bayes Naiv în cazul dat este:

$$\operatorname*{argmax}_y P_{MLE}(X_1 \mid Y = y) \cdot P_{MLE}(X_2 \mid Y = y) \cdot P_{MLE}(X_3, X_4 \mid Y = y) \cdot P_{MLE}(Y = y).$$

Pentru a justifica egalitatea de mai sus, se foloseşte regula de înlănţuire condiţională:

$$P(A_1, A_2, A_3 \mid B) = P(A_3 \mid B) \cdot P(A_2 \mid A_3, B) \cdot P(A_1 \mid A_2, A_3, B),$$

care se demonstrează uşor. Varianta necondiţională a regulii de înlănţuire este:

$$P(A_1, A_2, A_3) = P(A_3) \cdot P(A_2 \mid A_3) \cdot P(A_1 \mid A_2, A_3).$$

Apoi, regula de înlănţuire condiţională se particularizează pentru evenimentele $A_1 = (X_1 = x_1)$, $A_2 = (X_2 = x_2)$, $A_3 = (X_3 = x_3, X_4 = x_4)$ şi $B = (Y = y)$. În fine, se va ţine cont că $P(X_1 = x_1 \mid X_2 = x_2, X_3 = x_3, X_4 = x_4, Y = y) = P(X_1 = x_1 \mid Y = y)$ şi $P(X_2 = x_2 \mid X_3 = x_3, X_4 = x_4, Y = y) = P(X_2 = x_2 \mid Y = y)$ datorită proprietăţii de independenţă condiţională din enunţ.

10. **(Algoritmul Bayes Naiv: calculul ratei medii a erorii la antrenare)**

CMU, 2006 fall, T. Mitchell, E. Xing, midterm, pr. 6

Considerăm o problemă de clasificare binară în care fiecare exemplu X are două atribute binare $X_1, X_2 \in \{0,1\}$ şi eticheta $Y \in \{0,1\}$. Vom presupune că X_1 şi X_2 sunt independente condiţional în raport cu Y,[314] şi că $P(Y=0) = P(Y=1) = 0,5$. De asemenea, probabilităţile condiţionate sunt date în tabelele următoare:

$P(X_1 \mid Y)$	$Y=0$	$Y=1$
$X_1=0$	$0,7$	$0,2$
$X_1=1$	$0,3$	$0,8$

$P(X_2 \mid Y)$	$Y=0$	$Y=1$
$X_2=0$	$0,9$	$0,5$
$X_2=1$	$0,1$	$0,5$

a. Calculaţi predicţia $\hat{Y}$ făcută de clasificatorul Bayes Naiv pentru fiecare din cele patru combinaţii posibile de valori ale variabilelor X_1 şi X_2. Completaţi următorul tabel:

X_1	X_2	$P(X_1, X_2, Y=0)$	$P(X_1, X_2, Y=1)$	$\hat{Y}(X_1, X_2)$
0	0	$0,7 \cdot 0,9 \cdot 0,5$	$0,2 \cdot 0,5 \cdot 0,5$	0
0	1			
1	0			
1	1			

b. Presupunând că se folosesc o infinitate de exemple, calculaţi *rata medie a erorii* (engl., the expected error rate) făcute de acest clasificator *la antrenare*, folosind formula:

$$P(Y \neq \hat{Y}(X_1, X_2)) = \Sigma_{X_1=0}^{1} \Sigma_{X_2=0}^{1} P(X_1, X_2, Y = 1 - \hat{Y}(X_1, X_2))$$

c. Care din următorii doi clasificatori are rata medie a erorii la antrenare mai mică:

- **clasificatorul Bayes Naiv care prezice Y având ca input doar X_1;**
- **clasificatorul Bayes Naiv care prezice Y având ca input doar X_2.**

d. Presupunem că definim un nou atribut X_3, care este o copie a lui X_2. Care este rata medie a erorii la antrenare a clasificatorului Bayes Naiv care prezice Y folosind toate atributele X_1, X_2, X_3? (Se presupune că datele de antrenament sunt în număr infinit.)

e. Explicaţi de ce rata erorii de la punctul d diferă faţă de cea de la punctul a.

Răspuns:

a. Predicţia $\hat{Y}$ făcută de clasificatorul Bayes Naiv pentru fiecare din cele patru combinaţii posibile de valori ale variabilelor X_1 şi X_2 este înregistrată în ultima coloană a tabelului de mai jos. Calculele necesare pentru justificare

[314] Aşadar, în această situaţie rezultatele algoritmilor Bayes Naiv şi Bayes Optimal vor coincide.

— **folosind formula** $P(X_1, X_2, Y) = P(X_1, X_2|Y) \cdot P(Y) = P(X_1|Y) \cdot P(X_2|Y) \cdot P(Y)$
— **sunt conţinute în coloanele a treia şi a patra.**

X_1	X_2	$P(X_1, X_2, Y = 0)$	$P(X_1, X_2, Y = 1)$	$\hat{Y}(X_1, X_2)$
0	0	$0,7 \cdot 0,9 \cdot 0,5 = 0,315$	$0,2 \cdot 0,5 \cdot 0,5 = \mathbf{0,05}$	0
0	1	$0,7 \cdot 0,1 \cdot 0,5 = \mathbf{0,035}$	$0,2 \cdot 0,5 \cdot 0,5 = 0,05$	1
1	0	$0,3 \cdot 0,9 \cdot 0,5 = \mathbf{0,135}$	$0,8 \cdot 0,5 \cdot 0,5 = 0,2$	1
1	1	$0,3 \cdot 0,1 \cdot 0,5 = \mathbf{0,015}$	$0,8 \cdot 0,5 \cdot 0,5 = 0,2$	1

***Observaţie* (1):**
În acest tabel putem vedea valorile distribuţiei comune $P(X_1, X_2, Y)$. Spre deosebire de problema 8, unde distribuţia comună era dată iar distribuţiile marginale condiţionale erau calculate pornind de la aceasta, aici se procedează invers, ţinând cont [şi] de presupoziţia de independenţă condiţională.

b. Rata medie a erorii este:

$$\begin{aligned}
&P(Y \neq \hat{Y}(X_1, X_2)) = \\
&\quad = \sum_{X_1=0}^{1} \sum_{X_2=0}^{1} P(X_1, X_2, Y = 1 - \hat{Y}(X_1, X_2)) \\
&\quad = P(X_1 = 0, X_2 = 0, Y = 1 - 0) + P(X_1 = 0, X_2 = 1, Y = 1 - 1) \\
&\qquad\quad + P(X_1 = 1, X_2 = 0, Y = 1 - 1) + P(X_1 = 1, X_2 = 1, Y = 1 - 1) \\
&\quad = 0,05 + 0,035 + 0,135 + 0,015 = 0,235
\end{aligned}$$

Pentru justificarea penultimei egalităţi, vedeţi tabelul de la punctul precedent.

c. Făcând prezicerea doar cu X_1 ca atribut de intrare — folosind formula de multiplicare $P(X_1, Y) = P(X_1|Y) \cdot P(Y)$ —, obţinem:

X_1	$P(X_1, Y = 0)$	$P(X_1, Y = 1)$	$\hat{Y}_1(X_1, X_2)$
0	$0,7 \cdot 0,5 = 0,35$	$0,2 \cdot 0,5 = \mathbf{0,1}$	0
1	$0,3 \cdot 0,5 = \mathbf{0,15}$	$0,8 \cdot 0,5 = 0,4$	1

Rata medie a erorii în acest caz va fi:

$$\begin{aligned}
&P(Y \neq \hat{Y}_1(X_1, X_2)) = \\
&\quad = \sum_{X_1=0}^{1} \sum_{X_2=0}^{1} P(X_1, X_2, Y = 1 - \hat{Y}_1(X_1, X_2)) \\
&\quad = P(X_1 = 0, X_2 = 0, Y = 1 - 0) + P(X_1 = 0, X_2 = 1, Y = 1 - 0) \\
&\qquad\quad + P(X_1 = 1, X_2 = 0, Y = 1 - 1) + P(X_1 = 1, X_2 = 1, Y = 1 - 1) \\
&\quad = 0,05 + 0,05 + 0,135 + 0,015 = 0,1 + 0,15 = 0,25
\end{aligned}$$

Similar, dacă luăm în considerare doar variabila X_2, avem:

X_2	$P(X_2, Y = 0)$	$P(X_2, Y = 1)$	$\hat{Y}_2(X_1, X_2)$
0	$0,9 \cdot 0,5 = 0,45$	$0,5 \cdot 0,5 = \mathbf{0,25}$	0
1	$0,1 \cdot 0,5 = \mathbf{0,05}$	$0,5 \cdot 0,5 = 0,25$	1

Acum, rata medie a erorii va fi:

$$\begin{aligned}
&P(Y \neq \hat{Y}_2(X_1, X_2)) = \\
&\quad = \sum_{X_1=0}^{1} \sum_{X_2=0}^{1} P(X_1, X_2, Y = 1 - \hat{Y}_2(X_1, X_2)) \\
&\quad = P(X_1 = 0, X_2 = 0, Y = 1 - 0) + P(X_1 = 0, X_2 = 1, Y = 1 - 1) \\
&\qquad\quad + P(X_1 = 1, X_2 = 0, Y = 1 - 0) + P(X_1 = 1, X_2 = 1, Y = 1 - 1) \\
&\quad = 0,05 + 0,035 + 0,2 + 0,015 = 0,25 + 0,05 = 0,3
\end{aligned}$$

Prin urmare, rata medie a erorii la antrenare este mai mică pentru clasificatorul Bayes Naiv care prezice Y având ca input doar X_1 (decât pornind doar de la X_2).[315]

***Observaţie* (2):**
Atât în cazul lui X_1 cât şi în cazul lui X_2, rata medie a erorii (pentru algoritmul Bayes Naiv) putea fi calculată folosind direct probabilităţile din cele două tabele de mai sus. Justificarea ţine de faptul că distribuţiile calculate de Bayes Naiv în aceste două tabele sunt distribuţii marginale în raport cu distribuţia comună (reală!) din tabelul de la punctul a. La punctul d veţi vedea că acolo trebuie procedat altfel, fiindcă în cazul respectiv distribuţia calculată de către Bayes Naiv *nu* mai coincide cu distribuţia reală a datelor!

d. Clasificatorul Bayes Naiv care prezice valoarea lui Y în funcţie de toate cele trei variabilele $X_1, X_2, X_3 = X_2$ va lua deciziile conform tabelului următor:[316]

X_1	X_2	X_3	$P(X_1, X_2, X_3, Y = 0)$	$P(X_1, X_2, X_3, Y = 1)$	$\hat{Y}_3(X_1, X_2)$
0	0	0	$0,7 \cdot 0,9 \cdot 0,9 \cdot 0,5 = 0,2835$	$0,2 \cdot 0,5 \cdot 0,5 \cdot 0,5 = \mathbf{0,025}$	0
0	1	1	$0,7 \cdot 0,1 \cdot 0,1 \cdot 0,5 = \mathbf{0,0035}$	$0,2 \cdot 0,5 \cdot 0,5 \cdot 0,5 = 0,025$	1
1	0	0	$0,3 \cdot 0,9 \cdot 0,9 \cdot 0,5 = 0,1215$	$0,8 \cdot 0,5 \cdot 0,5 \cdot 0,5 = \mathbf{0,1}$	0
1	1	1	$0,3 \cdot 0,1 \cdot 0,1 \cdot 0,5 = \mathbf{0,0015}$	$0,8 \cdot 0,5 \cdot 0,5 \cdot 0,5 = 0,1$	1

***Observaţie* (3):**
Este util să observaţi că distribuţia de probabilitate comună calculată aici de către algoritmul Bayes Naiv nu mai coincide cu distribuţia „reală“ (vedeţi tabelul de la punctul a). Rata erorii se calculează în raport cu distribuţia „reală“ a datelor, nu cu cea calculată de către Bayes Naiv (deşi pentru a identifica situaţiile în care $\hat{Y} \neq Y$ se foloseşte ultimul tabel)!

Rata medie a erorii produse la antrenare va fi:

$$\begin{aligned}
P(Y \neq \hat{Y}_3(X_1, X_2)) &= \\
&= \sum_{X_1=0}^{1} \sum_{X_2=0}^{1} P(X_1, X_2, Y = 1 - \hat{Y}_3(X_1, X_2)) \\
&= P(X_1 = 0, X_2 = 0, Y = 1 - 0) + P(X_1 = 0, X_2 = 1, Y = 1 - 1) \\
&\qquad + P(X_1 = 1, X_2 = 0, Y = 1 - 0) + P(X_1 = 1, X_2 = 1, Y = 1 - 1) \\
&= 0,05 + 0,035 + 0,2 + 0,015 = 0,3
\end{aligned}$$

e. Diferenţa dintre cele două rate medii ale erorilor care au fost calculate la punctele a şi d se datorează faptului că presupunerea de independenţă condiţională a variabilelor nu este adevărată. Într-adevăr, X_2 nu este independent condiţional faţă de X_3 deoarece cele două variabile au tot timpul valori identice.

***Observaţii importante*:**
1. Este imediat că atunci când datele satisfac presupoziţia de independenţă condiţională, algoritmii Bayes Naiv şi Bayes Optimal produc aceleaşi rezultate

[315] Însă ambii clasificatori au rata medie a erorii mai mare decât clasificatorul Bayes Naiv care foloseşte atât X_1 cât şi X_2, ceea ce era de aşteptat întrucât în condiţiile date Bayes Naiv are acelaşi comportament ca şi Bayes Optimal.

[316] Este bine de observat că în această situaţie presupoziţia de independenţă condiţională este încălcată. Aşadar, Bayes Naiv nu va mai furniza aceleaşi rezultate ca Bayes Optimal.

şi au aceeaşi rată medie a erorilor.
2. Se poate arăta uşor că algoritmul Bayes Optimal nu produce în mod neapărat o rată medie a erorilor nulă, chiar dacă lucrează cu distribuţia reală a datelor. De *exemplu*, în condiţiile definite iniţial de problema noastră, algoritmul Bayes Optimal produce rata medie 0.235, ca şi algoritmul Bayes Naiv (vedeţi punctul b). Erorile produse de către algoritmul Bayes Optimal se datorează faptului că el aplică operatorul $\arg\max$, echivalent cu luarea unui vot majoritar (impus deci minorităţii).

11. **(Cât de naiv / prost este algoritmul Bayes Naiv?)**

■ • ***CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh, HW2, pr. 1.2***

În mod evident, clasificatorul Bayes Naiv lucrează cu o presupoziţie foarte restrictivă (engl., strong presupposition). Însă ne putem întreba dacă acest clasificator nu este totuşi destul de folositor chiar şi în cazul în care respectiva presupoziţie nu este satisfăcută.

În consecinţă, în acest exerciţiu ne propunem să folosim un exemplu simplu pentru a explora limitările algoritmului Bayes Naiv.

Fie X_1 şi X_2 variabile aleatoare binare de tip Bernoulli de parametru $p = 0.5$, iar Y o funcţie deterministă în raport cu valorile lui X_1 şi X_2, luând valori în mulţimea $\{1, 2\}$.

a. Definiţi Y astfel încât (pe setul de date respectiv) algoritmul Bayes Naiv să aibă rata medie a erorii de 50%.
Pe acest caz, observaţi cum se corelează valorile lui X_1 şi X_2 când valorea lui Y este fixată. (Altfel spus, observaţi cât de (in)dependente sunt în acest context valorile lui X_1 şi X_2, dată fiind valoarea lui Y.)

X_1	X_2	Y
0	0	
0	1	
1	0	
1	1	

b. Există în total $2^4 = 16$ moduri în care poate fi definită funcţia Y. Însă, datorită simetriei (relativ la valorile lui Y), problema se reduce la doar 4 cazuri,

X_1	X_2	Y
0	0	1
0	1	1
1	0	1
1	1	1

X_1	X_2	Y
0	0	1
0	1	1
1	0	1
1	1	2

X_1	X_2	Y
0	0	1
0	1	2
1	0	1
1	1	2

X_1	X_2	Y
0	0	1
0	1	2
1	0	2
1	1	1

dintre care un caz corespunde punctului a de mai sus. În fiecare din acele trei cazuri rămase după rezolvarea de la punctul a, arătaţi că rata erorii înregistrate de algoritmul Bayes Naiv este 0.

Răspuns:

a. Considerăm Y definit conform tabelului de mai jos.

***Observaţie*: Dacă se consideră valoarea lui Y fixată (fie 1, fie 2), atunci putem să stabilim o regulă astfel încât dacă îl cunoaştem pe X_1 să-l determinăm pe X_2 (şi invers).[317] Altfel spus, X_1 este unic determinat de X_2 (şi invers), dată fiind o valoare fixată a lui Y. Deci condiţia de independenţă condiţională este încălcată. Mai mult, în acest caz avem maximul posibil de „dependenţă" între cele două variabile (în raport cu Y).**

X_1	X_2	Y
0	0	1
0	1	2
1	0	2
1	1	1

Dorim să calculăm rata erorii înregistrate de algoritmul Bayes Naiv pe datele din tabelul de mai sus. Bayes Naiv estimează valoarea lui Y astfel:

$$\hat{y} = \underset{y \in \{1,2\}}{\operatorname{argmax}} P(X_1 \mid Y = y) \cdot P(X_2 \mid Y = y) \cdot P(Y = y)$$

Pentru $X_1 = 0, X_2 = 0$, algoritmul compară următoarele două valori:

$$\begin{aligned} p_1 &= P(X_1 = 0 \mid Y = 1) \cdot P(X_2 = 0 \mid Y = 1) \cdot P(Y = 1) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8} \\ p_2 &= P(X_1 = 0 \mid Y = 2) \cdot P(X_2 = 0 \mid Y = 2) \cdot P(Y = 2) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8} \end{aligned}$$

Cum $p_1 = p_2$, algoritmul va alege una dintre ele cu o probabilitate de 0.5. Deoarece valoarea lui Y din tabel este 1, înseamnă că algoritmul va alege greşit în 50% din cazuri.

Pentru celelalte 3 cazuri, $(X_1 = 0, X_2 = 1)$, $(X_1 = 1, X_2 = 0)$ şi $(X_1 = 1, X_2 = 1)$, se observă uşor că se obţin de asemenea valori egale, iar algoritmul va alege pentru Y una dintre valorile 1 sau 2 cu o probabilitate de 0.5.

Deci pentru această definiţie a lui Y rata erorii este de 50%.

b. Vom calcula rata erorii pentru fiecare dintre cele 3 moduri de definire a lui Y care nu a fost studiat.

***Cazul 1*:**

X_1	X_2	Y
0	0	1
0	1	1
1	0	1
1	1	1

Este similar cu cazul:

Y
2
2
2
2

- **Pentru $X_1 = 0, X_2 = 0$, algoritmul compară:**

$$\begin{aligned} p_1 &= P(X_1 = 0 \mid Y = 1) \cdot P(X_2 = 0 \mid Y = 1) \cdot P(Y = 1) = \frac{2}{4} \cdot \frac{2}{4} \cdot 1 = \frac{1}{4} \\ p_2 &= P(X_1 = 0 \mid Y = 2) \cdot P(X_2 = 0 \mid Y = 2) \cdot P(Y = 2) = 0 \cdot 0 \cdot 0 = 0 \end{aligned}$$

  Cum $p_1 > p_2$ algoritmul alege pentru Y valoarea 1, ceea ce este corect.

- **Pentru celelalte 3 cazuri, $(X_1 = 0, X_2 = 1)$, $(X_1 = 1, X_2 = 0)$ şi $(X_1 = 1, X_2 = 1)$, se observă că se obţin aceleaşi valori pentru p_1 şi p_2 ca mai sus, deci algoritmul alege (în mod corect) pentru Y valoarea 1.**

Aşadar, am obţinut că rata erorii este în acest caz 0.

[317] Pentru $Y = 1$, regula este: X_2 are aceeaşi valoare ca şi X_1. Pentru $Y = 2$, regula este: X_1 şi X_2 au valori complementare.

***Cazul 2*:**

X_1	X_2	Y
0	0	1
0	1	1
1	0	1
1	1	2

Cazuri similare:

Y	Y	Y	Y	Y	Y	Y
1	1	2	2	2	2	1
1	2	1	2	2	1	2
2	1	1	2	1	2	2
1	1	1	1	2	2	2

- **Pentru** $X_1 = 0, X_2 = 0$**:**

$$\left.\begin{aligned} p_1 &= P(X_1 = 0 \mid Y = 1) \cdot P(X_2 = 0 \mid Y = 1) \cdot P(Y = 1) = \frac{2}{3} \cdot \frac{2}{3} \cdot \frac{3}{4} = \frac{1}{3} \\ p_2 &= P(X_1 = 0 \mid Y = 2) \cdot P(X_2 = 0 \mid Y = 2) \cdot P(Y = 2) = 0 \cdot 0 \cdot \frac{1}{4} = 0 \end{aligned}\right\} \Rightarrow \hat{y} = 1$$

- **Pentru** $X_1 = 0, X_2 = 1$**:**

$$\left.\begin{aligned} p_1 &= P(X_1 = 0 \mid Y = 1) \cdot P(X_2 = 1 \mid Y = 1) \cdot P(Y = 1) = \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{3}{4} = \frac{1}{6} \\ p_2 &= P(X_1 = 0 \mid Y = 2) \cdot P(X_2 = 1 \mid Y = 2) \cdot P(Y = 2) = 0 \cdot 1 \cdot \frac{1}{4} = 0 \end{aligned}\right\} \Rightarrow \hat{y} = 1$$

- **Pentru** $X_1 = 1, X_2 = 0$**:**

$$\left.\begin{aligned} p_1 &= P(X_1 = 1 \mid Y = 1) \cdot P(X_2 = 0 \mid Y = 1) \cdot P(Y = 1) = \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{3}{4} = \frac{1}{6} \\ p_2 &= P(X_1 = 1 \mid Y = 2) \cdot P(X_2 = 0 \mid Y = 2) \cdot P(Y = 2) = 1 \cdot 0 \cdot \frac{1}{4} = 0 \end{aligned}\right\} \Rightarrow \hat{y} = 1$$

- **Pentru** $X_1 = 1, X_2 = 1$**:**

$$\left.\begin{aligned} p_1 &= P(X_1 = 1 \mid Y = 1) \cdot P(X_2 = 1 \mid Y = 1) \cdot P(Y = 1) = \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{3}{4} = \frac{1}{12} \\ p_2 &= P(X_1 = 1 \mid Y = 2) \cdot P(X_2 = 1 \mid Y = 2) \cdot P(Y = 2) = 1 \cdot 1 \cdot \frac{1}{4} = \frac{1}{4} \end{aligned}\right\} \Rightarrow \hat{y} = 2$$

Deci rata erorii este 0 **pentru acestă definiţie a lui** Y**.**

***Cazul 3*:**

X_1	X_2	Y
0	0	1
0	1	2
1	0	1
1	1	2

Cazuri similare:

Y	Y	Y
2	1	2
1	1	2
2	2	1
1	2	1

- **Pentru** $X_1 = 0, X_2 = 0$**:**

$$\left.\begin{aligned} p_1 &= \frac{1}{2} \cdot 1 \cdot \frac{1}{2} = \frac{1}{4} \\ p_2 &= \frac{1}{2} \cdot 0 \cdot \frac{1}{2} = 0 \end{aligned}\right\} \Rightarrow p_1 > p_2 \Rightarrow \hat{y} = 1 \textbf{ (corect)}$$

- **Pentru** $X_1 = 0, X_2 = 1$**:**

$$\left.\begin{aligned} p_1 &= \frac{1}{2} \cdot 0 \cdot \frac{1}{2} = 0 \\ p_2 &= \frac{1}{2} \cdot 1 \cdot \frac{1}{2} = \frac{1}{4} \end{aligned}\right\} \Rightarrow p_1 < p_2 \Rightarrow \hat{y} = 2 \textbf{ (corect)}$$

- **Pentru** $X_1 = 1, X_2 = 0$**:**

$$\left.\begin{aligned} p_1 &= \frac{1}{2} \cdot 1 \cdot \frac{1}{2} = \frac{1}{4} \\ p_2 &= \frac{1}{2} \cdot 0 \cdot \frac{1}{2} = 0 \end{aligned}\right\} \Rightarrow p_1 > p_2 \Rightarrow \hat{y} = 1 \textbf{ (corect)}$$

- **Pentru** $X_1 = 1, X_2 = 1$**:**

$$\left.\begin{aligned} p_1 &= \frac{1}{2} \cdot 0 \cdot \frac{1}{2} = 0 \\ p_2 &= \frac{1}{2} \cdot 1 \cdot \frac{1}{2} = \frac{1}{4} \end{aligned}\right\} \Rightarrow p_1 < p_2 \Rightarrow \hat{y} = 2 \textbf{ (corect)}$$

Prin urmare, rata erorii este 0 **şi în acest caz.**

***Cazul 4*: (cel de la punctul** a**)**

X_1	X_2	Y
0	0	1
0	1	2
1	0	2
1	1	1

Este similar cu cazul:

Y
2
1
1
2

În concluzie, doar pentru 2 moduri (cazul 4) de definire a lui Y **rata erorii este de** 50%**; pentru celelalte 14 moduri (cazurile 1, 2, 3) rata erorii este** 0**.**

12. **(O reprezentare grafică a neconcordanţei deciziilor luate de algoritmii Bayes Naiv şi Bayes Optimal)**

CMU, 2009 fall, Geoff Gordon, HW4, pr. 1
CMU, 2009 fall, Carlos Guestrin, HW1, pr. 4.1.5

Pentru un task de clasificare se consideră atributele X_1, X_2 **şi** X_3 **şi eticheta** Y**. Toate acestea sunt variabile aleatoare binare.** X_1 **şi** X_2 **sunt independente condiţional în raport cu** Y**, iar** X_3 **este o copie a lui** X_2 **(aşadar, întotdeauna** $X_2 = X_3$**).**

Sunt date următoarele probabilităţi condiţionate:

$$\begin{aligned} &P(X_1 = T \mid Y = T) = p, \quad P(X_1 = T \mid Y = F) = 1 - p, \\ &P(X_2 = F \mid Y = T) = q, \quad P(X_2 = F \mid Y = F) = 1 - q, \\ &P(Y = T) = 0.5. \end{aligned}$$

Se dă instanţa de test $X_1 = T,\ X_2 = X_3 = F$**. Vrem să clasificăm această instanţă, adică să prezicem valoarea lui** Y **pentru ea.**

a. Arătaţi că dacă se foloseşte clasificatorul Bayes Naiv, atunci eticheta pentru instanţa aceasta de test este T **— ceea ce revine la** $P(Y = T \mid X_1 = T, X_2 = F, X_3 = F) \geq 0.5$ **— dacă** $p \geq \frac{(1-q)^2}{q^2+(1-q)^2}$**.**

b. Ce devine inegalitatea de la punctul precedent dacă în schimbul clasificatorului Bayes Naiv se utilizează clasificatorul Bayes Optimal?

c. Desenaţi cele două curbe de decizie obţinute la punctele a **şi** b**. Pe axa** ***Ox*** **marcaţi valorile lui** q**, iar pe axa** ***Oy*** **marcaţi valorile lui** p**. Atât** p **cât şi** q **variază în intervalul** $[0, 1]$**. Indicaţi pe grafic zona în care clasificatorul Bayes Naiv produce un output (**Y**) diferit de cel al algoritmului Bayes Optimal.**

***Atenţie*! Acest exerciţiu *nu* studiază în ce condiţii cei doi algoritmi clasifică corect (ori dimpotrivă, eronat) instanţa** $X_1 = T, X_2 = X_3 = F$**, ci doar când anume (în funcţie de** p **şi** q**) produc ei clasificări diferite pentru această instanţă. Se va vedea (grafic!) că algoritmul Bayes Naiv nu se comportă deloc rău în comparaţie cu algoritmul Bayes Optimal.**

Răspuns:

Se lucrează cu variabile aleatoare binare, iar pentru claritatea calculelor vom nota prin X faptul că valoarea variabilei aleatoare X este T, iar prin $\neg X$ faptul că $X = F$.

Folosind aceste notaţii, putem transcrie datele din enunţ astfel:

$$\begin{array}{ll} P(X_1 \mid Y) = p, & P(X_1 \mid \neg Y) = 1 - p, \\ P(\neg X_2 \mid Y) = q, & P(\neg X_2 \mid \neg Y) = 1 - q, \\ P(Y) = 0.5. & \end{array}$$

a. Eticheta pentru instanţa aceasta de test este T dacă $P(Y \mid X_1, \neg X_2, \neg X_3) \geq 0.5$. Vom calcula această probabilitate utilizând regula lui Bayes precum şi ipoteza de independenţă condiţională făcută de algoritmul Bayes Naiv:

$$\begin{aligned} P(Y \mid X_1, \neg X_2, \neg X_3) &\overset{form.\ Bayes}{=} \\ &= \frac{P(X_1, \neg X_2, \neg X_3 \mid Y)P(Y)}{P(X_1, \neg X_2, \neg X_3 \mid Y)P(Y) + P(X_1, \neg X_2, \neg X_3 \mid \neg Y)P(\neg Y)} \overset{indep.\ cdt.}{=} \\ &= \frac{P(X_1 \mid Y)P(\neg X_2 \mid Y)P(\neg X_3 \mid Y)P(Y)}{P(X_1|Y)P(\neg X_2|Y)P(\neg X_3|Y)P(Y) + P(X_1|\neg Y)P(\neg X_2|\neg Y)P(\neg X_3|\neg Y)P(\neg Y)} \\ &= \frac{p \cdot q \cdot q \cdot 0.5}{p \cdot q \cdot q \cdot 0.5 + (1-p) \cdot (1-q) \cdot (1-q) \cdot 0.5} = \frac{pq^2}{pq^2 + (1-p)(1-q)^2} \end{aligned}$$

Deci eticheta este T dacă $\frac{pq^2}{pq^2 + (1-p)(1-q)^2} \geq 0.5$, ceea ce înseamnă că

$$pq^2 \geq 0.5(pq^2 + (1-p)(1-q)^2) \Leftrightarrow pq^2 - 0.5pq^2 \geq 0.5(1-p)(1-q)^2$$
$$\Leftrightarrow pq^2 \geq (1-q)^2 - p(1-q)^2 \Leftrightarrow p(q^2 + (1-q)^2) \geq (1-q)^2 \Leftrightarrow p \geq \frac{(1-q)^2}{q^2 + (1-q)^2}$$

b. Dacă în locul clasificatorului Bayes Naiv se foloseşte clasificatorul Bayes Optimal, nu se mai foloseşte presupunerea de independenţă condiţională. În locul acesteia se folosesc informaţiile furnizate în enunţ, şi anume că X_1 şi X_2 sunt independente condiţional în raport cu Y, iar X_3 este o copie a lui X_2.

$$\begin{aligned} P(Y \mid X_1, \neg X_2, \neg X_3) &\overset{form.\ Bayes}{=} \\ &= \frac{P(X_1, \neg X_2, \neg X_3 \mid Y)P(Y)}{P(X_1, \neg X_2, \neg X_3 \mid Y)P(Y) + P(X_1, \neg X_2, \neg X_3 \mid \neg Y)P(\neg Y)} \\ &= \frac{P(X_1 \mid Y)P(\neg X_2, \neg X_3 \mid Y)P(Y)}{P(X_1|Y)P(\neg X_2, \neg X_3|Y)P(Y) + P(X_1|\neg Y)P(\neg X_2, \neg X_3|\neg Y)P(\neg Y)} \\ &= \frac{P(X_1 \mid Y)P(\neg X_2 \mid Y)P(Y)}{P(X_1|Y)P(\neg X_2|Y)P(Y) + P(X_1|\neg Y)P(\neg X_2|\neg Y)P(\neg Y)} \\ &= \frac{p \cdot q \cdot 0.5}{p \cdot q \cdot 0.5 + (1-p) \cdot (1-q) \cdot 0.5} = \frac{pq}{pq + (1-p)(1-q)} \end{aligned}$$

S-a folosit egalitatea

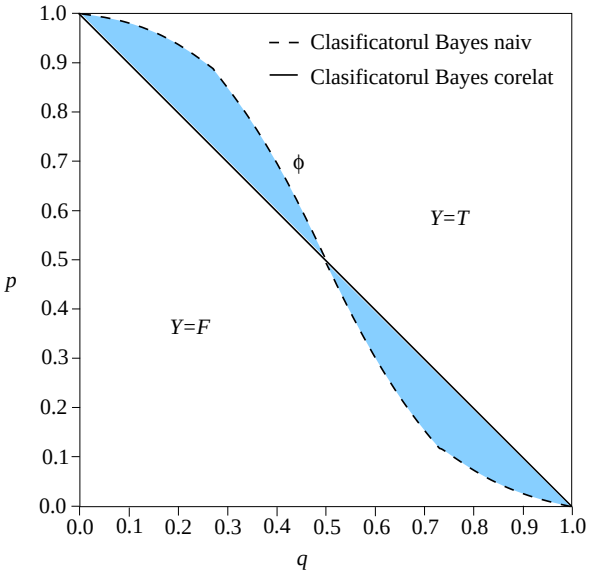
$$P(X_1, \neg X_2, \neg X_3 | Y) = P(X_1 | \neg X_2, \neg X_3, Y) \cdot P(\neg X_2, \neg X_3 | Y) = P(X_1 | Y) \cdot P(\neg X_2 | Y).$$

În acest caz eticheta este T dacă $\frac{pq}{pq+(1-p)(1-q)} \geq 0.5$, adică

$$pq \geq 0.5(pq+(1-p)(1-q)) \Leftrightarrow pq - 0.5pq \geq 0.5(1-p-q+pq)$$
$$\Leftrightarrow pq \geq 1-p-q+pq \Leftrightarrow p \geq 1-q.$$

c. Dreapta de ecuaţie $p = 1-q$ este uşor de reprezentat.

Notând $\phi(q) = \frac{(1-q)^2}{q^2+(1-q)^2}$, se observă imediat că $\phi(q) + \phi(1-q) = 1 \Leftrightarrow \phi(1-q) = 1-\phi(q)$. De aici, cu ajutorul unui raţionament geometric simplu, se ajunge imediat la concluzia că graficul funcţiei ϕ este simetric faţă de punctul de coordonate $(1/2, 1/2)$. Aşadar, va fi suficient să studiem graficul lui ϕ pe intervalul $[0, 1/2]$. Proprietăţile funcţiei ϕ necesare elaborării graficului sunt uşor de studiat. Adiţional, se poate arăta imediat că $\phi(q) \geq 1-q$ pentru orice $q \in [0, 1/2]$ şi $\phi(q) \leq 1-q$ pentru orice $q \in [1/2, 1]$.



În figura de mai sus am reprezentat curbele $p = 1-q$ şi $p = \phi(q)$, precum şi zonele de decizie pentru cei doi clasificatori obţinuţi la punctele a şi b. Se observă uşor zona în care rezultatul produs de clasificatorul Bayes Naiv (Y) este diferit / „eronat“ în raport cu algoritmul Bayes Optimal.

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

■ • ***CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh, HW2, pr. 1.1***

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 variabila aleatoare comună $\bar{X} = (X_1, \ldots, X_{d-1}, X_d)$, 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 $P(E_1 \cup \ldots \cup E_n) \leq \sum_{i=1}^{n} P(E_i)$.**[318]
- **$(1 - \frac{1}{k})^k \leq \frac{1}{e}$ 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 pentru fiecare din variabilele X_i ($i = \overline{1,d}$), atunci probabilitatea să nu fi întâlnit ambele valori pentru fiecare dintre aceste variabile 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

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

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

Răspuns:

a. Dacă s-au facut N observaţii şi nu s-au întâlnit ambele valori ale variabilei X_i, înseamnă că ea are aceeaşi valoare în toate aceste observaţii (adică ea este fie 0 în toate observaţiile, fie 1 în toate observaţiile). Pentru fiecare dintre aceste două cazuri probabilitatea este $1/2^N$, deci:

$$P(\text{doar una dintre valorile variabilei } X_i \text{ a apărut în } N \text{ observaţii}) = \left(\frac{1}{2}\right)^N + \left(\frac{1}{2}\right)^N = \frac{2}{2^N} = \frac{1}{2^{N-1}}$$

b. Se cere să se calculeze probabilitatea să nu fi întâlnit ambele valori pentru fiecare dintre variabilele $X_1, \ldots, X_{d-1}, X_d$. Pentru acesta vom folosi prima inegalitate din indicaţia dată în enunţ:

$$\begin{aligned} &P(\text{nu toate valorile variabilelor } X_i,\ i = \overline{1,d},\ \text{au apărut în } N_{NB} \text{ observaţii}) \\ &\leq \sum_{i=1}^{d} P(\text{numai una dintre valorile variabilei } X_i \text{ a apărut în } N_{NB} \text{ observaţii}) \\ &= \sum_{i=1}^{d} \frac{1}{2^{N_{NB}-1}} = d \cdot \frac{1}{2^{N_{NB}-1}} = d \cdot \frac{1}{2^{1+\log_2 \frac{d}{\varepsilon}-1}} = d \cdot \frac{1}{2^{\log_2 \frac{d}{\varepsilon}}} = d \cdot \frac{1}{\frac{d}{\varepsilon}} = d \cdot \frac{\varepsilon}{d} = \varepsilon. \end{aligned}$$

c. Pentru algoritmul Bayes Optimal s-au făcut N observaţii. Trebuie să calculăm probabilitatea ca să nu se fi întâlnit niciodată instanţa $\bar{x}$.
Cum există în total 2^d posibile observaţii, probabilitatea ca instanţa $\bar{x}$ să nu fie obţinută la *una* dintre observaţii este $1 - \frac{1}{2^d}$. Observaţiile fiind independente, probabilitatea ca $\bar{x}$ să nu fie obţinută *după* N observaţii este $\left(1 - \frac{1}{2^d}\right)^N$. Aşadar,

$$\begin{aligned} P(\text{instanţa } \bar{x} \text{ n-a fost întâlnită în } N \text{ observaţii}) &= \left(1 - \frac{1}{2^d}\right)^N \\ &= \left[\left(1 - \frac{1}{2^d}\right)^{2^d}\right]^{N/2^d} \leq \left(\frac{1}{e}\right)^{N/2^d} = e^{-N/2^d} \end{aligned}$$

d. Vom calcula probabilitatea ca să nu se fi întâlnit toate instanţele variabilei $\bar{X}$ utilizând din nou prima inegalitate din *indicaţie*:

$$\begin{aligned} &P(\text{nu toate instanţele variabilei } \bar{X} \text{ au fost întâlnite în } N_{JB} \text{ observaţii}) \\ &\leq \sum_{\bar{x}} P(\text{instanţa } \bar{x} \text{ n-a fost întâlnită în } N_{JB} \text{ observaţii}) \end{aligned}$$

$$\leq \sum_{\bar{x}} e^{-N_{JB}/2^d} = 2^d \cdot e^{-N_{JB}/2^d} = 2^d \cdot e^{-\ln \frac{2^d}{\varepsilon}} = 2^d \cdot \frac{1}{e^{\ln \frac{2^d}{\varepsilon}}} = \frac{2^d}{\frac{2^d}{\varepsilon}} = \varepsilon.$$

e. Vom înlocui datele numerice în formulele $N_{NB} = 1 + \log_2 \frac{d}{\varepsilon}$, $N_{JB} = 2^d \ln \frac{2^d}{\varepsilon}$.

$$\varepsilon = 0.1,\ d = 2 \Rightarrow \begin{cases} N_{NB} = 1 + \log_2 \frac{2}{0.1} = 1 + \log_2 20 \approx 5.32 \\ N_{JB} = 2^2 \cdot \ln \frac{2^2}{0.1} = 4 \cdot \ln 40 \approx 14.75 \end{cases}$$

$$\varepsilon = 0.1,\ d = 5 \Rightarrow \begin{cases} N_{NB} = 1 + \log_2 \frac{5}{0.1} = 1 + \log_2 50 \approx 6.64 \\ N_{JB} = 2^5 \cdot \ln \frac{2^5}{0.1} = 32 \cdot \ln 320 \approx 184.58 \end{cases}$$

$$\varepsilon = 0.1,\ d = 10 \Rightarrow \begin{cases} N_{NB} = 1 + \log_2 \frac{10}{0.1} = 1 + \log_2 100 \approx 7.64 \\ N_{JB} = 2^{10} \cdot \ln \frac{2^{10}}{0.1} = 1024 \cdot \ln 10240 \approx 9455.67 \end{cases}$$

Acum se observă uşor [diferenţa dintre] numărul de date de antrenament necesare pentru cei doi algoritmi: de ordin logaritmic pentru Bayes Naiv, respectiv de ordin exponenţial pentru Bayes Optimal.

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

■ □ ● ○ ***CMU, 2005 fall, T. Mitchell, A. Moore, HW2, pr. 2***
CMU, 2009 fall, Carlos Guestrin, HW1, pr. 4.1.2
CMU, 2009 fall, Geoff Gordon, HW4, pr. 1.2-3
CMU, 2012 fall, T. Mitchell, Z. Bar-Joseph, HW2, pr. 3.a

a. [Bayes Naiv şi Regresia Logistică: relaţia dintre regulile de decizie][319]

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 $z = -(w_0 + w_1 X_1 \ldots + w_d X_d)$, cu parametrii $w_0, w_1, \ldots, w_d \in \mathbb{R}$,[320] adică

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

şi, prin urmare

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

[319] Pentru o introducere în chestiunea regresiei logistice, vedeţi Tom Mitchell, *Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression*, draft pentru un capitol suplimentar pentru o nouă ediţie a cărţii *Machine Learning*, 2016. Puteţi vedea de asemenea problema 12 de la capitolul *Metode de regresie* din prezenta culegere.

[320] 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 $w_0 + w_1 X_1 \ldots + w_d X_d = 0$.

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 $P(Y|X)$,[321] 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:

$$P(X_i|Y = 1) = \theta_{i1}^{X_i}(1 - \theta_{i1})^{(1-X_i)}, \tag{162}$$

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 (162) 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 $X_i = 0$.

b. [Relaxarea presupoziţiei de independenţă condiţională][322]

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).

[321] 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. Vedeţi de exemplu problemele 12 şi 13 de la capitolul *Metode de regresie*.
[322] Vedeţi de exemplu problema 9.

***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:

$$P(X_1, X_2|Y = k) = (\beta_{11k})^{X_1X_2} (\beta_{10k})^{X_1(1-X_2)} (\beta_{01k})^{(1-X_1)X_2} (\beta_{00k})^{(1-X_1)(1-X_2)} \quad (163)$$

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 $(1 - X_1)(1 - X_2) = 0$.

Răspuns:

a. Mai întâi vom rescrie probabilitatea $P(Y = 1|X = x)$ ca o fracţie, folosind formula lui Bayes, compusă cu formula probabilităţii totale, apoi vom împărţi atât numărătorul cât şi numitorul fracţiei astfel obţinute cu expresia de la numărător:[323]

$$\begin{aligned} P(Y = 1|X = x) &\overset{FB}{=} \frac{P(X = x|Y = 1)\, P(Y = 1)}{\sum_{y' \in \{0,1\}} P(X = x|Y = y')\, P(Y = y')} \\ &= \frac{1}{1 + \dfrac{P(X = x|Y = 0)P(Y = 0)}{P(X = x|Y = 1)P(Y = 1)}}. \end{aligned}$$

După aceea, folosind formula $e^{\ln a} = a$ (valabilă pentru orice $a > 0$), vom forţa punerea fracţiei de mai sus într-o formă apropiată de cea a funcţiei sigmoidale ($\sigma(x) = 1/(1 + e^{-x})$):[324]

$$\begin{aligned} P(Y = 1|X = x) &= \frac{1}{1 + \exp\left(\ln \dfrac{P(X = x|Y = 0)P(Y = 0)}{P(X = x|Y = 1)P(Y = 1)}\right)} \\ &= \frac{1}{1 + \exp\left(\ln \dfrac{P(X_1 = x_1, \ldots, X_d = x_d|Y = 0)P(Y = 0)}{P(X_1 = x_1, \ldots, X_d = x_d|Y = 1)P(Y = 1)}\right)}. \end{aligned}$$

Mai departe, ţinând cont de presupoziţia de independenţă condiţională şi de proprietăţile funcţiei logaritm, obţinem:[325]

$$P(Y = 1|X = x) = \frac{1}{1 + \exp\left(\ln \dfrac{P(Y = 0)}{P(Y = 1)} + \sum_{i=1}^{d} \ln \dfrac{P(X_i = x_i|Y = 0)}{P(X_i = x_i|Y = 1)}\right)}.$$

Vom nota probabilităţile a priori $P(Y = 1)$ şi $P(Y = 0)$ cu π şi respectiv $1 - \pi$. Apoi, conform *indicaţiei* 2, vom scrie $P(X_i|Y = 1)$ ca $\theta_{i1}^{X_i}(1 - \theta_{i1})^{(1-X_i)}$ şi $P(X_i|Y = 0)$ ca $\theta_{i0}^{X_i}(1 - \theta_{i0})^{(1-X_i)}$. În consecinţă,

[323] Cu excepţia cazului când $P(X = x|Y = 1)\, P(Y = 1) = 0$.
[324] Cu excepţia cazului când $P(X = x|Y = 0)\, P(Y = 0) = 0$.
[325] Cu excepţia cazurilor când $P(X = x_i|Y = 0) = 0$ sau $P(X = x_i|Y = 1) = 0$ pentru $i = 1, \ldots, d$.

$$P(Y=1|X=x) = \frac{1}{1+\exp\left(\ln\frac{1-\pi}{\pi} + \sum_{i=1}^{d}\ln\frac{\theta_{i0}^{X_i}(1-\theta_{i0})^{(1-X_i)}}{\theta_{i1}^{X_i}(1-\theta_{i1})^{(1-X_i)}}\right)}$$

$$= \frac{1}{1+\exp\left(\ln\frac{1-\pi}{\pi} + \sum_{i=1}^{d}\left(X_i\ln\frac{\theta_{i0}}{\theta_{i1}} + (1-X_i)\ln\frac{1-\theta_{i0}}{1-\theta_{i1}}\right)\right)}$$

$$= \frac{1}{1+\exp\left(\ln\frac{1-\pi}{\pi} + \sum_{i=1}^{d}\ln\frac{1-\theta_{i0}}{1-\theta_{i1}} + \sum_{i=1}^{d}X_i\left(\ln\frac{\theta_{i0}}{\theta_{i1}} - \ln\frac{1-\theta_{i0}}{1-\theta_{i1}}\right)\right)}.$$

Pentru a pune această ultimă expresie sub forma dorită, adică $P(Y=1|X=x) = 1/(1+\exp(w_0+\sum_{i=1}^{d}w_iX_i)$**, vom alege valorile parametrilor** w_i **în mod natural:**

$$w_0 = \ln\frac{1-\pi}{\pi} + \sum_{i=1}^{d}\ln\frac{1-\theta_{i0}}{1-\theta_{i1}} \quad \text{şi} \quad w_i = \ln\frac{\theta_{i0}}{\theta_{i1}} - \ln\frac{1-\theta_{i0}}{1-\theta_{i1}} \text{ pentru } i=1,\ldots,d.$$

b. Vom începe ca şi la punctul precedent, prin a pune probabilitatea $P(Y=1|X=x)$ **sub o formă apropiată de cea a funcţiei sigmoidale:**[326]

$$P(Y=1|X) \overset{FB}{=} \frac{P(X|Y=1)P(Y=1)}{P(X|Y=1)P(Y=1)+P(X|Y=0)P(Y=0)}$$

$$= \frac{1}{1+\frac{P(X|Y=0)P(Y=0)}{P(X|Y=1)P(Y=1)}} = \frac{1}{1+\exp\left(\ln\frac{P(X|Y=0)P(Y=0)}{P(X|Y=1)P(Y=1)}\right)}.$$

Până aici nu avem încă nicio diferenţă în raport cu calculul de la punctul a**. Însă acum vom ţine cont că toate variabilele** X_i **ci** $i=1,\ldots,d$ **sunt independente condiţional două câte două în raport cu** Y**, cu excepţia perechii** X_1, X_2**:**[327]

$$P(Y=1|X) = \frac{1}{1+\exp\left(\ln\frac{P(X_1,X_2|Y=0)\ \prod_{i=3}^{d}P(X_i|Y=0)P(Y=0)}{P(X_1,X_2|Y=1)\ \prod_{i=3}^{d}P(X_i|Y=1)P(Y=1)}\right)}$$

$$= \frac{1}{1+\exp\left(\ln\frac{1-\pi}{\pi} + \sum_{i=3}^{d}\ln\frac{P(X_i|Y=0)}{P(X_i|Y=1)} + \ln\frac{P(X_1,X_2|Y=0)}{P(X_1,X_2|Y=1)}\right)}.$$

Ca şi la punctul a**, vom ţine cont că**

$$\ln\frac{P(X_i|Y=0)}{P(X_i|Y=1)} = \ln\frac{\theta_{i0}^{X_i}(1-\theta_{i0})^{(1-X_i)}}{\theta_{i1}^{X_i}(1-\theta_{i1})^{(1-X_i)}}$$

atunci când condiţiile asociate cu relaţia (162) sunt satisfăcute. Mai departe, folosind *indicaţia* 4, vom putea înlocui $P(X_1,X_2|Y=0)$ **şi** $P(X_1,X_2|Y=1)$ **în**

[326] Cu excepţia cazurilor când $P(X|Y=1)P(Y=1)=0$ sau $P(X|Y=0)P(Y=0)=0$.

[327] Cu excepţia cazurilor când $P(Y=0)=0$ sau $P(Y=1)=0$, respectiv $P(X_1X_2|Y=0)=0$ sau $P(X_1X_2|Y=1)=0$ şi încă $P(X_i|Y=0)=0$ sau $P(X_i|Y=1)=0$ pentru $i=3,\ldots,d$.

funcţie de β_{ijk}, obţinând (atunci când condiţiile asociate cu relaţia (163) sunt satisfăcute):

$$\ln \frac{P(X_1, X_2|Y=0)}{P(X_1, X_2|Y=1)} = \ln \frac{(\beta_{110})^{X_1X_2}\,(\beta_{100})^{X_1(1-X_2)}\,(\beta_{010})^{(1-X_1)X_2}\,(\beta_{000})^{(1-X_1)(1-X_2)}}{(\beta_{111})^{X_1X_2}\,(\beta_{101})^{X_1(1-X_2)}\,(\beta_{011})^{(1-X_1)X_2}\,(\beta_{001})^{(1-X_1)(1-X_2)}}.$$

Aşadar, va rezulta:

$$P(Y=1|X) = \frac{1}{1+\exp\left(w_0 + \sum_{i=3}^{d} w_i X_i + w_1 X_1 + w_2 X_2 + w_{1,2} X_1 X_2\right)},$$

unde

$$\begin{aligned}
w_0 &= \ln\frac{1-\pi}{\pi} + \sum_{i=3}^{d} \ln\frac{1-\theta_{i1}}{1-\theta_{i0}} + \ln\frac{\beta_{000}}{\beta_{001}} \\
w_1 &= \ln\frac{\beta_{100}}{\beta_{101}} + \ln\frac{\beta_{001}}{\beta_{000}} \\
w_2 &= \ln\frac{\beta_{010}}{\beta_{011}} + \ln\frac{\beta_{001}}{\beta_{000}} \\
w_{1,2} &= \ln\frac{\beta_{110}}{\beta_{111}} + \ln\frac{\beta_{101}}{\beta_{100}} + \ln\frac{\beta_{011}}{\beta_{010}} + \ln\frac{\beta_{000}}{\beta_{001}} \\
w_i &= \ln\frac{\theta_{i0}}{\theta_{i1}} + \ln\frac{1-\theta_{i1}}{1-\theta_{i0}} \text{ pentru } i = 3, \ldots, d.
\end{aligned}$$

2.1.3 Clasificare bayesiană [cu atribute de intrare] de tip gaussian

15. **(Algoritmul Bayes [Naiv] gaussian: aplicare pe date din $\mathbb{R}$)**

prelucrare de Liviu Ciortuz, după

■ • ***CMU, 2001 fall, Andrew Moore, midterm, pr. 3.a***

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, va trebui mai întâi să învăţaţi *parametrii* clasificatorului Bayes gaussian, prin metoda estimării de verosimilitate maximă (MLE).[328] Centralizaţi rezultatele, completând tabelul următor:

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

[328] Vedeţi secţiunea corespunzătoare din capitolul de *Fundamente*, în speţă (pentru acest caz) problemele 45.a şi 46.a.

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. Cum va clasifica algoritmul Bayes [Naiv] gaussian punctul $X = 2$? Puteţi exprima răspunsul fie în funcţie de α şi β, fie — mai bine! — calculând în prealabil valorile lui α şi β în funcţie de parametrii calculaţi / estimaţi la punctul precedent.

Răspuns:

a. Pentru a estima mediile μ_A şi μ_B, vom folosi formula demonstrată la problema 45.a de la capitolul de *Fundamente*:

$$\mu_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n},$$

unde n este numărul instanţelor de antrenament. Aşadar, $\mu_A = \frac{\sum_{i=1}^{2} X_i}{2} = \frac{0+2}{2} = 1$, iar $\mu_B = \frac{\sum_{i=3}^{7} X_i}{5} = \frac{3+4+5+6+7}{5} = 5$.

Similar, pentru calculul varianţelor σ_A şi σ_B, vom folosi formula care a fost demonstrată la problema 46.a de la capitolul de *Fundamente*:

$$\sigma^2_{MLE} = \frac{\sum_{i=1}^{n}(x_i - \mu_{MLE})^2}{n}.$$

Aşadar, $\sigma^2_A = \frac{1}{2}[(0-1)^2 + (2-1)^2] = 1$, iar $\sigma^2_B = \frac{1}{5}[(3-5)^2 + (4-5)^2 + 0^2 + (6-5)^2 + (7-5)^2] = \frac{1}{5} \cdot 2 \cdot [4+1] = 2$.

Pentru calculul probabilităţilor $P(Y = A)$ şi $P(Y = B)$ se foloseşte formula clasică (numărul de cazuri favorabile împărţit la numărul de cazuri posibile; vedeţi problema 40 sau problema 113 de la capitolul de *Fundamente*), fiindcă Y este variabilă de tip Bernoulli. Aşadar, $P(Y = A) = 2/7$ şi $P(Y = B) = 5/7$.

Centralizând aceste estimări, obţinem:

$\mu_A = 1$	$\sigma^2_A = 1$	$P(Y = A) = 2/7$
$\mu_B = 5$	$\sigma^2_B = 2$	$P(Y = B) = 5/7$

b. Folosind formula de multiplicare [a probabilităţilor], calculăm $p(X = 2, Y = A) = p(X = 2|Y = A) \cdot P(Y = A) = \frac{2\alpha}{7}$ şi $p(X = 2, Y = B) = p(X = 2|Y = B) \cdot P(Y = B) = \frac{5\beta}{7}$.

Probabilitatea $p(X = 2)$ se poate obţine aplicând formula probabilităţii totale: $p(X = 2) = p(X = 2|Y = A) \cdot P(Y = A) + p(X = 2|Y = B) \cdot P(Y = B) = \frac{1}{7}(2\alpha + 5\beta)$.

Probabilitatea condiţionată $p(Y = A|X = 2)$ se calculează folosind definiţia: $p(Y = A|X = 2) = \dfrac{p(Y = A, X = 2)}{p(X = 2)} = \dfrac{2\alpha}{2\alpha + 5\beta}$.

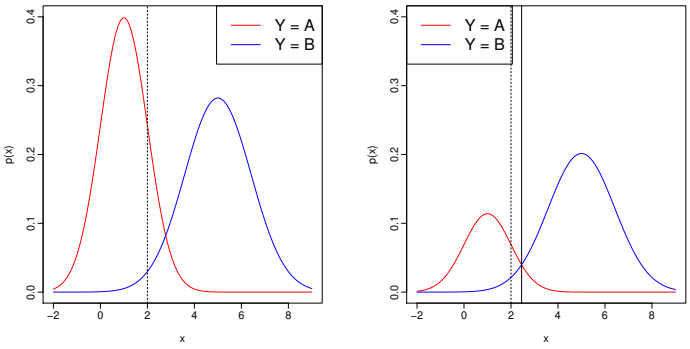
c. Algoritmul Bayes [Naiv] gaussian va asocia punctului $X = 2$ eticheta $Y = A$ dacă $p(Y = A|X = 2) \geq p(Y = B|X = 2) \Leftrightarrow \dfrac{2}{7}\alpha \geq \dfrac{5}{7}\beta$.

Folosind valorile estimate pentru parametrii μ_A, μ_B, σ_A şi σ_B la punctul a, vom putea scrie: $\alpha = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{(2-1)^2}{2}} = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}} = 0.24197$ **şi** $\beta = \dfrac{1}{\sqrt{2\pi} \cdot \sqrt{2}} e^{-\frac{(2-5)^2}{2 \cdot 2}} = \dfrac{1}{2\sqrt{\pi}} e^{-\frac{9}{4}} = 0.029732$. **Deci,**

$$\frac{2}{7}\alpha \geq \frac{5}{7}\beta \Leftrightarrow \frac{2}{7} \cdot 0.24197 \geq \frac{5}{7} \cdot 0.029732 \Leftrightarrow 0.06913 \geq 0.02123 \text{ (adevărat).}$$

Prin urmare, algoritmul Bayes [Naiv] gaussian va asocia punctului $X = 2$ eticheta $Y = A$, cu probabilitatea $\dfrac{0.06913}{0.06913 + 0.02123} = 0.76499$.

***Observaţie*: În cele două grafice alăturate am reprezentat $\mathcal{N}(\mu_A, \sigma_A^2)$ şi $\mathcal{N}(\mu_B, \sigma_B^2)$, p.d.f.-urile gaussienelor asociate claselor A şi respectiv B (în partea stângă), precum şi funcţiile obţinute din aceste două p.d.f.-uri prin înmulţirea cu *factorii de selecţie* $2/7$ şi respectiv $5/7$ (în partea dreaptă).**



Se poate constata relativ uşor că există două puncte de intersecţie ($x_1 = -8.451$ şi $x_2 = 2.451$) pentru graficele funcţiilor $\dfrac{2}{7}\mathcal{N}(\mu_A, \sigma_A^2)$ şi $\dfrac{5}{7}\mathcal{N}(\mu_B, \sigma_B^2)$. Toate instanţele de test x situate între aceste puncte de intersecţie ($x_1 < x < x_2$) vor aparţine clasei A (acolo curba roşie este situată deasupra celei albastre). Instanţele situate fie la stânga lui x_1 fie la dreapta lui x_2 vor aparţine clasei B (acolo curba albastră este situată deasupra celei roşii). *Separatorul decizional* este de tip pătratic, fiind constituit din punctele x_1 şi x_2.[329]

[329] LC: Mulţumesc studentului MSc Dinu Sergiu pentru această observaţie.

16. **(Clasificatorul Bayes [Naiv] gaussian, cazul când se foloseşte un singur atribut de intrare: zone de decizie şi graniţe de decizie; analiza diferitelor *cazuri specifice*)**

□ ***Sergiu Dinu, Liviu Ciortuz, 2020***

În acest exerciţiu ne propunem să identificăm *zonele de decizie* şi *graniţele de decizie* determinate de clasificatorul Bayes [Naiv] gaussian (abreviat *GB*), atunci când folosim un singur atribut de intrare X, iar atributul de ieşire, pe care îl vom nota cu Y, este binar, deci poate lua două valori, desemnate în continuare cu A şi B. Vom desemna prin $p_A = p$ *probabilitatea de selecţie* pentru clasa A. Prin urmare, probabilitatea de selecţie pentru clasa B este $p_B = 1 - p$. Vom presupune că $p \in (0, 1)$ şi $X|(Y = A) \sim \mathcal{N}(x|\mu_A, \sigma_A^2)$, iar $X|(Y = B) \sim \mathcal{N}(x|\mu_B, \sigma_B^2)$.

a. Arătaţi că *regula de decizie* a acestui clasificator Bayes [Naiv] gaussian

$$\hat{Y}_{GB}(X = x) = A \Leftrightarrow p \cdot \mathcal{N}(x|\mu_A, \sigma_A^2) \geq (1 - p) \cdot \mathcal{N}(x|\mu_B, \sigma_B^2) \tag{164}$$

devine echivalentă cu

$$\left(\sigma_A^2 - \sigma_B^2\right)(x - x_1)(x - x_2) \geq 0,$$

unde

$$x_1 = \frac{\sigma_A^2\mu_B - \sigma_B^2\mu_A - \sqrt{\Delta'}}{\sigma_A^2 - \sigma_B^2} \quad \textbf{şi} \quad x_2 = \frac{\sigma_A^2\mu_B - \sigma_B^2\mu_A + \sqrt{\Delta'}}{\sigma_A^2 - \sigma_B^2},$$

cu

$$\Delta' \overset{not.}{=} \sigma_A^2\sigma_B^2 \left[(\mu_A - \mu_B)^2 + (\sigma_A^2 - \sigma_B^2) \ln\left(\frac{1-p}{p} \cdot \frac{\sigma_A}{\sigma_B}\right)^2\right],$$

în *condiţiile* în care $\sigma_A^2 \neq \sigma_B^2$ şi $\Delta' \geq 0$.

b. Arătaţi că atunci când $\sigma_A = \sigma_B \overset{not.}{=} \sigma$ şi $\mu_A \neq \mu_B$, regula de decizie (164) este echivalentă cu

$$x \geq x_0 \text{ dacă } \mu_A > \mu_B$$

şi, respectiv

$$x \leq x_0 \text{ dacă } \mu_A < \mu_B,$$

unde

$$x_0 = \frac{\mu_A + \mu_B}{2} + \frac{\sigma^2}{\mu_A - \mu_B} \cdot \ln\frac{1-p}{p}.$$

c. Arătaţi că este posibil ca în anumite cazuri să avem $\Delta' < 0$, adică inegalitatea (164) să fie ori adevărată pentru orice $x \in \mathbb{R}$, ori falsă pentru orice $x \in \mathbb{R}$. Altfel spus, arătaţi că pot exista combinaţii de valori pentru parametrii σ_A şi σ_B (ambii strict pozitivi), μ_A şi μ_B din $\mathbb{R}$, precum şi $p \in (0, 1)$, astfel încât

$$(\mu_A - \mu_B)^2 + (\sigma_A^2 - \sigma_B^2) \ln\left(\frac{1-p}{p} \cdot \frac{\sigma_A}{\sigma_B}\right)^2 < 0. \tag{165}$$

Răspuns:

a. Putem scrie următoarea succesiune de echivalenţe:

$$p \cdot \mathcal{N}(x|\mu_A, \sigma_A^2) \geq (1-p) \cdot \mathcal{N}(\mu_B, \sigma_B^2) \Leftrightarrow$$
$$p \cdot \frac{1}{\sqrt{2\pi}\sigma_A} \exp\left(-\frac{(x-\mu_A)^2}{2\sigma_A^2}\right) \geq (1-p) \cdot \frac{1}{\sqrt{2\pi}\sigma_B} \exp\left(-\frac{(x-\mu_B)^2}{2\sigma_B^2}\right) \Leftrightarrow$$
$$\exp\left(\frac{1}{2}\left[\left(\frac{x-\mu_B}{\sigma_B}\right)^2 - \left(\frac{x-\mu_A}{\sigma_A}\right)^2\right]\right) \geq \frac{1-p}{p} \cdot \frac{\sigma_A}{\sigma_B} \Leftrightarrow$$
$$\left(\frac{x-\mu_B}{\sigma_B}\right)^2 - \left(\frac{x-\mu_A}{\sigma_A}\right)^2 \geq \ln\left(\frac{1-p}{p} \cdot \frac{\sigma_A}{\sigma_B}\right)^2 \Leftrightarrow$$
$$(\sigma_A^2 - \sigma_B^2)x^2 + 2(\sigma_B^2\mu_A - \sigma_A^2\mu_B)x + \sigma_A^2\mu_B^2 - \sigma_B^2\mu_A^2 - \sigma_A^2\sigma_B^2 \ln\left(\frac{1-p}{p} \cdot \frac{\sigma_A}{\sigma_B}\right)^2 \geq 0. \tag{166}$$

Membrul stâng al acestei ultime inegalităţi este o funcţie polinomială de gradul al doilea. Prin urmare, discriminantul ei „pe jumătate" (adică, $\Delta' \overset{not.}{=} \Delta/2$) este:

$$\Delta' = (\sigma_B^2\mu_A - \sigma_A^2\mu_B)^2 - (\sigma_A^2 - \sigma_B^2)\left[\sigma_A^2\mu_B^2 - \sigma_B^2\mu_A^2 - \sigma_A^2\sigma_B^2 \ln\left(\frac{1-p}{p}\frac{\sigma_A}{\sigma_B}\right)^2\right]$$

Însă,

$$\begin{aligned}
&(\sigma_B^2\mu_A - \sigma_A^2\mu_B)^2 - (\sigma_A^2 - \sigma_B^2)(\sigma_A^2\mu_B^2 - \sigma_B^2\mu_A^2) \\
&= \cancel{\sigma_B^4\mu_A^2} + \cancel{\sigma_A^4\mu_B^2} - 2\sigma_A^2\sigma_B^2\mu_A\mu_B - \cancel{\sigma_A^4\mu_B^2} + \sigma_A^2\sigma_B^2\mu_A^2 + \sigma_A^2\sigma_B^2\mu_B^2 - \cancel{\sigma_B^4\mu_A^2} \\
&= \sigma_A^2\sigma_B^2(\mu_A - \mu_B)^2.
\end{aligned}$$

Aşadar,

$$\Delta' = \sigma_A^2\sigma_B^2\left[(\mu_A - \mu_B)^2 + (\sigma_A^2 - \sigma_B^2)\ln\left(\frac{1-p}{p} \cdot \frac{\sigma_A}{\sigma_B}\right)^2\right],$$

ceea ce ne conduce imediat la concluzia din enunţ.

***Consecinţă*: În acest caz (a), separatorul decizional este format din punctele x_1 şi x_2. Dacă $\sigma_A^2 > \sigma_B^2$, atunci zona de decizie corespunzătoare clasei A este $(-\infty, x_1] \cup [x_2, +\infty)$, iar zona de decizie corespunzătoare clasei B este intervalul (x_1, x_2). Similar, dacă $\sigma_A^2 < \sigma_B^2$, atunci zona de decizie corespunzătoare clasei A este (x_1, x_2), iar zona de decizie corespunzătoare clasei B este intervalul $(-\infty, x_1] \cup [x_2, +\infty)$.**

b. Dacă $\sigma_A = \sigma_B = \sigma$, iar $\sigma > 0$, atunci este uşor de observat că inegalitatea (166) devine

$$(\mu_A - \mu_B)x + \mu_B^2 - \mu_A^2 - \sigma^2 \ln\frac{1-p}{p} \geq 0,$$

ceea ce fundamentează concluzia din enunţ.

***Consecinţă*: În acest caz (b), separatorul decizional este punctul x_0. Zonele de decizie se stabilesc imediat conform relaţiilor din enunţ.**

c. Observăm că dacă vom considera cazul particular $\mu_A = \mu_B$, relaţia (165) va căpăta o formă mai simplă:

$$(\sigma_A^2 - \sigma_B^2)\ln\left(\frac{1-p}{p} \cdot \frac{\sigma_A}{\sigma_B}\right) < 0. \tag{167}$$

Această inegalitate este satisfăcută, de pildă atunci când σ_A şi σ_B (care se consideră, prin convenţie, strict pozitive) sunt în relaţia $\sigma_A > \sigma_B$, iar $\frac{1-p}{p} < \frac{\sigma_B}{\sigma_A}$. Se poate constata imediat că a doua inegalitate este echivalentă cu $p > \frac{\sigma_A}{\sigma_A + \sigma_B}$. Se observă că această ultimă fracţie are valori în intervalul $(1/2, 1)$, fiindcă $\sigma_A > \sigma_B$.

Similar, dacă $\sigma_A < \sigma_B$, atunci pentru a satisface inegalitatea (167) se impune condiţia $\frac{1-p}{p} > \frac{\sigma_B}{\sigma_A}$, care revine la $p < \frac{\sigma_A}{\sigma_A + \sigma_B}$. Această ultimă fracţie ia valori în intervalul $(0, 1/2)$, întrucât $\sigma_A < \sigma_B$.

***Consecinţă*: În acest caz (c), nu există separator decizional. Una din cele două zone de decizie este vidă, iar cealaltă zonă coincide cu mulţimea tuturor numerelor reale, $\mathbb{R}$.**

17. **(Algoritmul Bayes Naiv gaussian: deducerea [formei liniare a] regulei de decizie în cazul matricelor de covarianţă diagonale şi identice, i.e., $\sigma_{i0} = \sigma_{i1}$, pentru $i = 1, \ldots, d$)**

■ • ○ ***CMU, 2009 spring, Ziv Bar-Joseph, HW2, pr. 2***

Considerăm un model de tip Bayes Naiv cu două clase ($Y \in \{0, 1\}$), definit peste spaţiul real $\mathbb{R}^d$ al atributelor de intrare $X_1, \ldots, X_d$. Presupunem că în acest model distribuţia [comună] condiţionată $X|Y = 0$, unde $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$, poate fi definită ca un vector de distribuţii gaussiene unidimensionale independente şi-l vom desemna prin notaţia

$$Gaussian(\mu_0 =< \mu_{10}, \ldots, \mu_{d0} >,\ \sigma =< \sigma_1, \ldots, \sigma_d >)$$

şi analog pentru $X|Y = 1$:

$$Gaussian(\mu_1 =< \mu_{11}, \ldots, \mu_{d1} >,\ \sigma =< \sigma_1, \ldots, \sigma_d >).$$

Observaţi că intrările $X_1, \ldots, X_d$ au — la condiţionare în raport cu clasa — medii diferite dar varianţe (de fapt, matrice de covarianţă, diagonale) identice pentru ambele clase.

În acest exerciţiu vă vom arăta că, în modelul specificat mai sus, probabilitatea condiţionată $P(Y = 1|X = x)$, unde $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, se poate rescrie ca valoare a unei funcţii sigmoidale / „logistice", $f(x) = \frac{1}{1 + e^{-(w_0 + w \cdot x)}}$, cu parametrii $w_0 \in \mathbb{R}$ şi $w = (w_1, \ldots, w_d) \in \mathbb{R}^d$ aleşi în mod convenabil.[330]

a. Folosiţi regula lui Bayes (compusă cu formula probabilităţii totale) pentru a rescrie $P(Y = 1|X = x)$ sub forma unei fracţii. Împărţiţi atât numărătorul cât şi numitorul fracţiei astfel obţinute cu expresia de la numărător.

b. La punctul a ar fi trebuit să ajungeţi la un rezultat de forma

$$\frac{1}{1 + f(x, X, Y)},$$

[330] În consecinţă, [se poate arăta imediat că] regula de decizie a clasificatorului Bayes Naiv pentru acest model este de tip *liniar*.

unde f este o anumită funcţie având argumentele x, X şi Y. Folosind formula $e^{\ln a} = a$ (valabilă pentru orice $a > 0$), putem forţa punerea acestei fracţii într-o formă apropiată de funcţia sigmoidală:

$$\frac{1}{1 + e^{\ln f(x,X,Y)}}.$$

Vă cerem să rescrieţi sub o astfel de formă rezultatul de la punctul a.

c. Explicitaţi presupoziţia Bayes „naivă“ în cadrul modelului probabilist dat. Apoi folosiţi-o pentru a converti exponentul care apare în fracţia rezultată la punctul b la o sumă de forma

$$\ln g(Y) + \sum_{i=1}^{d} \ln h(x_i, X_i, Y).$$

Precizaţi expresiile funcţiilor g şi h.

d. Acum folosiţi specificul modelului dat — şi anume, de tip gaussian, având pentru componentele condiţionate (şi anume, variabilele aleatoare condiţionate $X_i|Y = 1$, respectiv $X_i|Y = 0$, pentru $i = 1, \ldots, n$) medii diferite dar varianţe egale —, pentru a aduce expresia de la punctul c la o formă mai convenabilă.

e. Rescriind $P(Y = 1|X = x)$ conform expresiilor obţinute la punctele b şi d, rezultatul ar trebui să semene cu un alt model pe care l-am întâlnit la curs. Care anume? Exprimaţi *parametrii* acelui model în raport cu $P(Y = 1)$, μ_{i0}, μ_{i1} şi σ_i, cu $i = 1, \ldots, d$.

Răspuns:

a. Procedând conform cerinţelor, vom scrie:

$$\begin{aligned} P(Y = 1|X = x) &= \frac{P(X = x|Y = 1)\, P(Y = 1)}{\sum_{y \in \{0,1\}} P(X = x|Y = y)\, P(Y = y)} \\ &= \frac{1}{1 + \dfrac{P(X = x|Y = 0)P(Y = 0)}{P(X = x|Y = 1)P(Y = 1)}}. \end{aligned}$$

Considerând $f(x, X, Y) = \dfrac{P(X = x|Y = 0)P(Y = 0)}{P(X = x|Y = 1)P(Y = 1)}$, se observă că ultima expresie obţinută are într-adevăr forma $\dfrac{1}{1 + f(x, X, Y)}$.

b. Folosind formula $a = e^{\ln a}$, obţinem:

$$P(Y = 1|X = x) = \frac{1}{1 + \exp\left(\ln \dfrac{P(X = x|Y = 0)P(Y = 0)}{P(X = x|Y = 1)P(Y = 1)}\right)}.$$

c. Conform presupoziţiei specifice algoritmului Naive Bayes, vom considera că atributele X_i sunt independente condiţional două câte două în raport cu variabila de ieşire. Aşadar, vom avea egalităţile următoare:

$$P(X=x|Y=1) = \prod_{i=1}^{d} P(X_i=x_i|Y=1)$$
$$P(X=x|Y=0) = \prod_{i=1}^{d} P(X_i=x_i|Y=0)$$

Prin urmare, vom scrie argumentul funcţiei exp( ) **din fracţia de la punctul** b **astfel:**

$$\ln\frac{P(X=x|Y=0)P(Y=0)}{P(X=x|Y=1)P(Y=1)} = \ln\frac{P(Y=0)}{P(Y=1)} + \sum_{i=1}^{d}\ln\frac{P(X_i=x_i|Y=0)}{P(X_i=x_i|Y=1)}.$$

În consecinţă, funcţiile g **şi** h **care au fost cerute în enunţ vor fi:**

$$g(Y)=\frac{P(Y=0)}{P(Y=1)} \qquad h(x_i,X_i,Y)=\frac{P(X_i=x_i|Y=0)}{P(X_i=x_i|Y=1)} \text{ pentru } i=1,\ldots,d.$$

d. Ţinând cont de specificul gaussian al modelului din enunţ, vom putea scrie rezultatul de la punctul b **astfel:**

$$\begin{aligned}
&\ln\frac{P(X=x|Y=0)P(Y=0)}{P(X=x|Y=1)P(Y=1)}\\
&= \ln\frac{P(Y=0)}{P(Y=1)} + \sum_{i=1}^{d}\ln\left(\frac{\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{(x_i-\mu_{i0})^2}{2\sigma_i^2}\right)}{\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{(x_i-\mu_{i1})^2}{2\sigma_i^2}\right)}\right)\\
&= \ln\frac{P(Y=0)}{P(Y=1)} + \sum_{i=1}^{d}\left(\frac{(x_i-\mu_{i1})^2}{2\sigma_i^2}-\frac{(x_i-\mu_{i0})^2}{2\sigma_i^2}\right)\\
&= \ln\frac{P(Y=0)}{P(Y=1)} + \sum_{i=1}^{d}\frac{2x_i(\mu_{i0}-\mu_{i1})+(\mu_{i1}^2-\mu_{i0}^2)}{2\sigma_i^2}\\
&= \ln\frac{P(Y=0)}{P(Y=1)} + \sum_{i=1}^{d}\left(\frac{x_i(\mu_{i0}-\mu_{i1})}{\sigma_i^2}+\frac{(\mu_{i1}^2-\mu_{i0}^2)}{2\sigma_i^2}\right)\\
&= \ln\frac{P(Y=0)}{P(Y=1)} + \sum_{i=1}^{d}\frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2} + \sum_{i=1}^{d}\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}x_i.
\end{aligned}$$

e. Notând în expresia obţinută la punctul d

$$w_0 = \ln\frac{P(Y=0)}{P(Y=1)} + \sum_{i=1}^{d}\frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2} \quad \textbf{şi} \quad w_i=\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2} \text{ pentru } i=1,\ldots,d$$

şi apoi revenind la expresia de la punctul b**, vom putea scrie:**

$$P(Y=1|X=x)=\frac{1}{1+\exp(w_o+\sum_{i=1}^{d}w_ix_i)}.$$

Aşadar, expresia probabilităţii a posteriori este exact cea a funcţiei sigmoidale, $\frac{1}{1+e^{-x}}$**. Rezultatul seamănă foarte bine cu modelul de *regresie logistică*.**[331]

[331] Pentru o introducere la modelul regresiei logistice, vedeţi problema 12 de la capitolul *Metode de regresie*.

18. **(Algoritmul Bayes Optimal gaussian: raportul faţă de regresia logistică în cazul $\Sigma_0 = \Sigma_1$)**

■ □ • ○ ***CMU, 2011 spring, Tom Mitchell, HW2, pr. 2.2***

În cele ce urmează, vom considera:
1. Y**, o variabilă booleană care urmează o distribuţie de tip Bernoulli, cu parametrul** $\pi = P(Y = 1)$**, ceea ce implică** $P(Y = 0) = 1 - \pi$**;**
2. $X = (X_1, X_2, \ldots, X_d)^\top$**, un vector de variabile aleatoare care *nu* sunt independente condiţional în raport cu variabila** Y**, probabilitatea [comună şi] condiţionată** $P(X|Y = k)$ **urmând o *distribuţie gaussiană multidimensională*,** $\mathcal{N}(\mu_k, \Sigma)$**, unde** $k \in \{0, 1\}$**.**

Reţineţi faptul că μ_k**, media acestei distribuţii multidimensionale, este un vector-coloană (altfel spus, o matrice** $d \times 1$**) şi există două astfel de medii, câte una pentru fiecare dintre cele două valori ale variabilei** Y**. De asemenea,** Σ**, matricea de covarianţă are dimensiunea** $d \times d$**, iar ea nu depinde de valorile variabilei** Y**.**[332]

În rezolvarea problemei, veţi folosi funcţia de densitate [de probabilitate] a distribuţiei gaussiene multidimensionale în notaţie matriceală:[333]

$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right),$$

unde simbolul $\top$ **desemnează operaţia de transpunere a matricelor.**

Răspundeţi la următoarea *întrebare*:
Este oare distribuţia $P(Y|X)$ **corespunzătoare acestui clasificator de tip Bayes Optimal gaussian (nu naiv!) de aceeaşi formă cu cea a *regresiei logistice*?**

***Sugestie*: Rafinaţi expresia distribuţiei probabiliste** $P(Y|X)$**.**

***Observaţie*: La problema 17 am arătat că în cazul (particular!) în care intrările** X_1**,** X_2**,** $\ldots$**,** X_d **sunt independente condiţional în raport cu ieşirea** Y **— ceea ce, conform problemei 31 de la capitolul de *Fundamente*, este echivalent cu a spune că matricea** Σ **este diagonală —, răspunsul la întrebarea pusă în enunţ este pozitiv.**

Răspuns:

Vom demara calculele în maniera standard (adică, similar cu prima parte a rezolvării problemelor 14 şi 17):

$$\begin{aligned} P(Y=1|X) &= \frac{P(X|Y=1)\,P(Y=1)}{P(X|Y=1)\,P(Y=1) + P(X|Y=0)\,P(Y=0)} \\ &= \frac{1}{1 + \dfrac{P(Y=0)\,P(X|Y=0)}{P(Y=1)\,P(X|Y=1)}} = \frac{1}{1 + \exp\left(\ln \dfrac{P(Y=0)\,P(X|Y=0)}{P(Y=1)\,P(X|Y=1)}\right)} \\ &= \frac{1}{1 + \exp\left(\ln \dfrac{P(Y=0)}{P(Y=1)} + \ln \dfrac{P(X|Y=0)}{P(X|Y=1)}\right)}. \end{aligned}$$

[332] Altfel spus, presupunând că Σ_k, pentru $k \in \{0, 1\}$, desemnează matricea de convarianţă a distribuţiei gaussiene multidimensionale $\mathcal{N}(\mu_k, \Sigma_k)$, atunci considerăm că $\Sigma_0 = \Sigma_1$.
[333] Vedeţi problema 34 de la capitolul de *Fundamente*.

Acum ne vom concentra atenţia asupra termenului $\ln \dfrac{P(X|Y=0)}{P(X|Y=1)}$**, ţinând cont de faptul că** $X|Y=0 \sim \mathcal{N}(\mu_0, \Sigma)$ **şi** $X|Y=1 \sim \mathcal{N}(\mu_1, \Sigma)$**:**

$$\begin{aligned}
&\ln \frac{P(X|Y=0)}{P(X|Y=1)} \\
&= \ln \frac{\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}}{\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}} + \ln \exp\Big(\frac{1}{2}\left[(X-\mu_1)^\top \Sigma^{-1}(X-\mu_1) - (X-\mu_0)^\top \Sigma^{-1}(X-\mu_0)\right]\Big) \\
&= \frac{1}{2}\left[(X-\mu_1)^\top \Sigma^{-1}(X-\mu_1) - (X-\mu_0)^\top \Sigma^{-1}(X-\mu_0)\right] \\
&= \frac{1}{2}\left[-X^\top \Sigma^{-1}\mu_1 - \mu_1^\top \Sigma^{-1} X + \mu_1^\top \Sigma^{-1}\mu_1 + X^\top \Sigma^{-1}\mu_0 + \mu_0^\top \Sigma^{-1} X - \mu_0^\top \Sigma^{-1}\mu_0\right] \\
&= \frac{1}{2}\left[\mu_1^\top \Sigma^{-1}\mu_1 - \mu_0^\top \Sigma^{-1}\mu_0 + X^\top \Sigma^{-1}(\mu_0-\mu_1) + (\mu_0^\top - \mu_1^\top)\Sigma^{-1} X\right] \\
&= \frac{1}{2}\mu_1^\top \Sigma^{-1}\mu_1 - \frac{1}{2}\mu_0^\top \Sigma^{-1}\mu_0 + (\mu_0-\mu_1)^\top \Sigma^{-1} X.
\end{aligned}$$

Remarcaţi faptul că $((\mu_0^\top - \mu_1^\top)\Sigma^{-1}X)^\top = ((\mu_0-\mu_1)^\top \Sigma^{-1}X)^\top = X^\top (\Sigma^{-1})^\top (\mu_0 - \mu_1) = X^\top (\Sigma^\top)^{-1}(\mu_0-\mu_1) = X^\top \Sigma^{-1}(\mu_0-\mu_1)$**, întrucât matricea** Σ^{-1} **este simetrică (ca urmare a faptului că** Σ **însăşi, ca matrice de covarianţă, este simetrică; vedeţi problema 20 de la capitolul de *Fundamente*).**

Prin urmare,

$$\begin{aligned}
P(Y=1|X) &= \frac{1}{1+\exp\Big(\ln\frac{1-\pi}{\pi} + \frac{1}{2}\mu_1^\top \Sigma^{-1}\mu_1 - \frac{1}{2}\mu_0^\top \Sigma^{-1}\mu_0 + (\mu_0-\mu_1)^\top \Sigma^{-1} X\Big)} \\
&= \frac{1}{1+\exp(w_0 + w^\top X)},
\end{aligned}$$

cu $w_0 = \ln\dfrac{1-\pi}{\pi} + \dfrac{1}{2}\mu_1^\top \Sigma^{-1}\mu_1 - \dfrac{1}{2}\mu_0^\top \Sigma^{-1}\mu_0$ **şi** $w = \Sigma^{-1}(\mu_0-\mu_1)$**. Evident,** w_0 **este un număr real (constant), iar** w **un vector-coloană (mai precis, o matrice de dimensiune** $d \times 1$**).**

În concluzie, distribuţia probabilistă $P(Y|X)$ **are (şi în acest caz!) aceeaşi formă cu cea din modelul regresiei logistice.**

19. **(Clasificatorul Bayes Optimal Gaussian: natura separatorului decizional în cazul în care** $\Sigma_0 \neq \Sigma_1$**)**

■ ● ○ ***Stanford, 2014 fall, Andrew Ng, midterm, pr. 2.b***

Fie setul de date de antrenament $\{(x_1, y_1), \ldots, (x_n, y_n)\}$**, cu** x_i **vectori-coloană din** $\mathbb{R}^d$ **şi** $y_i \in \{0, 1\}$ **pentru orice** i**. Algoritmul Bayes Optimal gaussian estimează următorii parametri:** $\phi \in (0,1)$**, vectorii-coloană** μ_0 **şi** μ_1 **din** $\mathbb{R}^d$**, precum şi matricele de covarianţă** Σ_0 **şi** Σ_1 **de dimensiune** $d \times d$**, corespunzători următoarelor distribuţii:**

$$\begin{aligned}
p(y) &= \phi^y (1-\phi)^{1-y}, \text{ unde } \phi = p(y=1) \\
p(x|y=0) &= \frac{1}{(2\pi)^{d/2}|\Sigma_0|^{1/2}} \exp\Big(-\frac{1}{2}(x-\mu_0)^\top \Sigma_0^{-1}(x-\mu_0)\Big)
\end{aligned}$$

$$p(x|y=1) \quad = \quad \frac{1}{(2\pi)^{d/2}|\Sigma_1|^{1/2}} \exp\Big(-\frac{1}{2}(x-\mu_1)^\top \Sigma_1^{-1}(x-\mu_1)\Big)$$

Vă readucem aminte *regula de decizie* a acestui clasificator:

$$y=1 \textbf{ dacă } p(y=1|x) \geq p(y=0|x) \textbf{ şi } y=0 \textbf{ în caz contrar.} \tag{168}$$

Arătaţi că atunci când $\Sigma_0 \neq \Sigma_1$, *separatorul decizional* determinat de algoritmul Bayes Optimal gaussian este de ordin pătratic, adică inegalitatea dintre probabilităţile condiţionate a posteriori, $p(y=1|x) \geq p(y=0|x)$, este echivalentă cu o inegalitate numerică de forma

$$x^\top A x + B^\top x + C \geq 0, \tag{169}$$

unde A este o anumită matrice de dimensiune $d \times d$, cu $A \neq 0$ (matricea nulă de dimensiune $d \times d$), $B \in \mathbb{R}^d$ (un anumit vector-coloană) şi $C \in \mathbb{R}$ (o anumită constantă). Veţi specifica în mod clar valorile pentru A, B şi C.

Răspuns:

Elaborând relaţia dintre probabilităţile care determină decizia luată de către clasificatorul Bayes Optimal gaussian pentru o instanţă oarecare de test x (vedeţi relaţia (168)), putem obţine următoarele echivalenţe:

$$\begin{aligned}
&p(y=1|x) \geq p(y=0|x) \Leftrightarrow \ln p(y=1|x) \geq \ln p(y=0|x) \\
&\quad \Leftrightarrow \quad \ln p(y=1|x) - \ln p(y=0|x) \geq 0 \Leftrightarrow \ln \frac{p(y=1|x)}{p(y=0|x)} \geq 0 \\
&\overset{\text{F. Bayes}}{\Leftrightarrow} \ln \frac{p(x|y=1)\,p(y=1)}{p(x|y=0)\,p(y=0)} \geq 0 \Leftrightarrow \ln \frac{p(y=1)}{p(y=0)} + \ln \frac{p(x|y=1)}{p(x|y=0)} \geq 0 \\
&\quad \Leftrightarrow \quad \ln \frac{\phi}{1-\phi} - \ln \frac{|\Sigma_1|^{1/2}}{|\Sigma_0|^{1/2}} - \frac{1}{2}\Big((x-\mu_1)^\top \Sigma_1^{-1}(x-\mu_1) - (x-\mu_0)^\top \Sigma_0^{-1}(x-\mu_0)\Big) \geq 0 \\
&\quad \Leftrightarrow \quad -\frac{1}{2}\Big(x^\top(\Sigma_1^{-1} - \Sigma_0^{-1})x - 2(\mu_1^\top \Sigma_1^{-1} - \mu_0^\top \Sigma_0^{-1})x + \mu_1^\top \Sigma_1^{-1}\mu_1 - \mu_0^\top \Sigma_0^{-1}\mu_0\Big) \\
&\qquad + \ln \frac{\phi}{1-\phi} - \ln \frac{|\Sigma_1|^{1/2}}{|\Sigma_0|^{1/2}} \geq 0 \\
&\quad \Leftrightarrow \quad x^\top\Big(\frac{1}{2}(\Sigma_0^{-1} - \Sigma_1^{-1})\Big)x + \Big(\mu_1^\top \Sigma_1^{-1} - \mu_0^\top \Sigma_0^{-1}\Big)x \\
&\qquad + \ln \frac{\phi}{1-\phi} + \ln \frac{|\Sigma_0|^{1/2}}{|\Sigma_1|^{1/2}} + \frac{1}{2}\Big(\mu_0^\top \Sigma_0^{-1}\mu_0 - \mu_1^\top \Sigma_1^{-1}\mu_1\Big) \geq 0.
\end{aligned} \tag{170}$$

Coroborând rezultatul pe care tocmai l-am obţinut cu relaţia (168) din enunţ, observăm că putem considera

$$\begin{aligned}
A &= \frac{1}{2}(\Sigma_0^{-1} - \Sigma_1^{-1}) \\
B^\top &= \mu_1^\top \Sigma_1^{-1} - \mu_0^\top \Sigma_0^{-1} \Leftrightarrow \\
B &= (\mu_1^\top \Sigma_1^{-1} - \mu_0^\top \Sigma_0^{-1})^\top = (\Sigma_1^{-1})^\top \mu_1 - (\Sigma_0^{-1})^\top \mu_0 = \Sigma_1^{-1}\mu_1 - \Sigma_0^{-1}\mu_0 \\
C &= \ln \frac{\phi}{1-\phi} + \ln \frac{|\Sigma_0|^{1/2}}{|\Sigma_1|^{1/2}} + \frac{1}{2}\Big(\mu_0^\top \Sigma_0^{-1}\mu_0 - \mu_1^\top \Sigma_1^{-1}\mu_1\Big).
\end{aligned} \tag{171}$$

Se poate constata imediat că ipoteza $\Sigma_0 \neq \Sigma_1$ din enunţ implică faptul că $\Sigma_0^{-1} \neq \Sigma_1^{-1}$, deci $A \neq 0$. Prin urmare, graniţa de separare (engl., decision boundary) determinată în acest caz de către algoritmul Bayes Optimal gaussian este de ordin pătratic.

20. **(Clasificatorul Bayes Optimal gaussian: aplicare pe date din $\mathbb{R}^2$; raportul cu regresia logistică)**

prelucrare de Liviu Ciortuz, după MIT, 2001 fall, Tommi Jaakkola, HW2, pr. 2.abe

a. Precizaţi separatorii decizionali determinaţi de algoritmul Bayes Optimal gaussian pentru fiecare din cazurile de mai jos. Veţi menţiona în mod explicit, în fiecare caz în parte, tipul — liniar, pătratic (cerc, elipsă, parabolă, hiperbolă etc.), etc. — şi ecuaţia acestor separatori decizionali.

$(i.)$

$P_0 = 0.5,\ P_1 = 0.5$
$\mu_0 = (1,1)^\top$
$\mu_1 = (-1,-1)^\top$
$\Sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
$\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

$(ii.)$

$P_0 = 0.995,\ P_1 = 0.005$
$\mu_0 = (1,1)^\top$
$\mu_1 = (-1,-1)^\top$
$\Sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
$\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

$(iii.)$

$P_0 = 0.5,\ P_1 = 0.5$
$\mu_0 = (1,1)^\top$
$\mu_1 = (-1,-1)^\top$
$\Sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
$\Sigma_1 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$

$(iv.)$

$P_0 = 0.5,\ P_1 = 0.5$
$\mu_0 = (0,0)^\top$
$\mu_1 = (0,0)^\top$
$\Sigma_0 = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}$
$\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$

$(v.)$

$P_0 = 0.5,\ P_1 = 0.5$
$\mu_0 = (1,1)^\top$
$\mu_1 = (1,1)^\top$
$\Sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
$\Sigma_1 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$

$(vi.)$

$P_0 = 0.5,\ P_1 = 0.5$
$\mu_0 = (-2,1)^\top$
$\mu_1 = (1,1)^\top$
$\Sigma_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1/2 \end{pmatrix}$
$\Sigma_1 = \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix}$

***Sugestie*: În cazul general, ecuaţia separatorului decizional determinat de clasificatorul Bayes Optimal gaussian este $x^\top Ax + B^\top x + C = 0$ (vedeţi realţia (169) de la problema 19), unde valorile pentru A, B şi C au fost determinate prin relaţiile (171).**

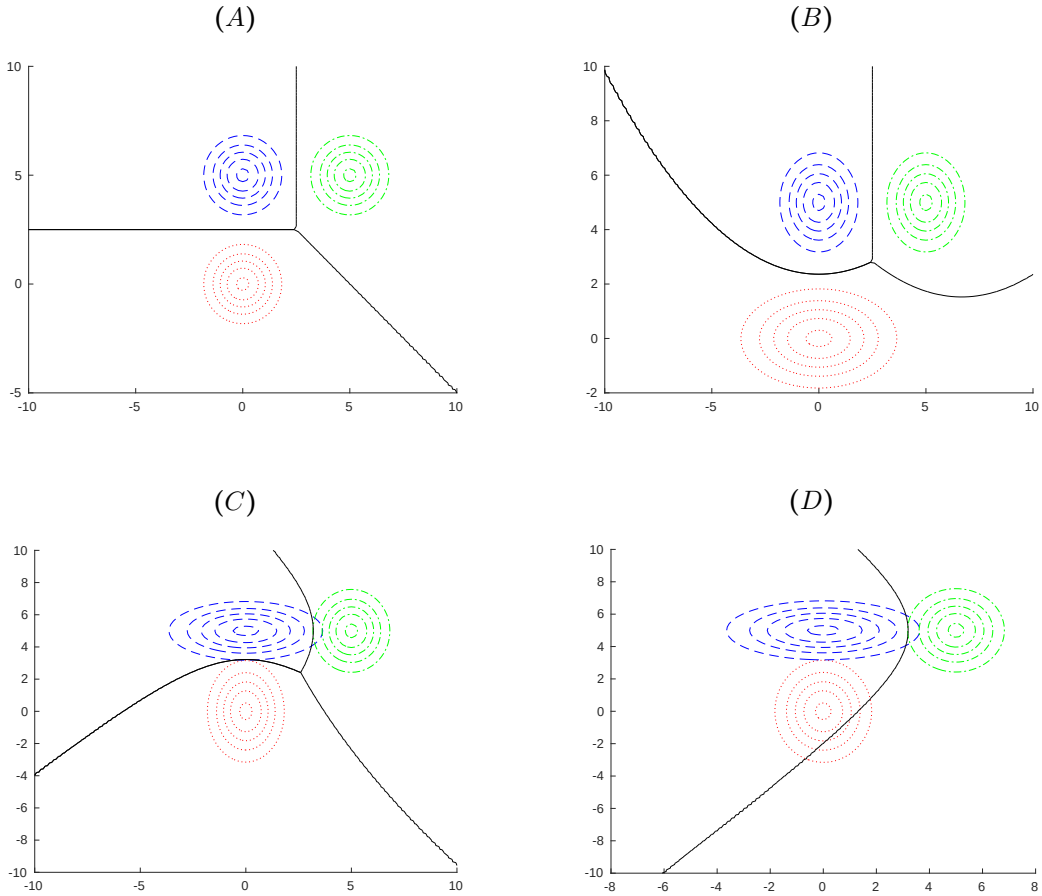
b. Arătaţi că este posibil ca întreg planul euclidian să fie alocat de către algoritmul Bayes Optimal gaussian[334] unei singure clase, situaţie în care, practic, nu avem separator decizional.

c. Care dintre *separatorii decizionali* de la punctul a poate corespunde *şi* unui model de tip *regresie logistică*? Justificaţi răspunsul dumneavoastră.

d. Care dintre *zonele de decizie* din figurile de mai jos poate să corespundă unui model de tip *regresie logistică multinomială* (*softmax*)? Justificaţi răspunsul dumneavoastră.[335]

[334] LC: De fapt, este suficient să ne gândim la cazul $\Sigma_0 = \Sigma_1$. (Menţionăm că acest caz particular de clasificare bayesiană gaussiană este numit în literatura de specialitate *analiză gaussiană discriminativă*.)

[335] Pentru o introducere în regresia logistică multinomială, vedeţi problema 17 de la capitolul *Metode de regresie*.



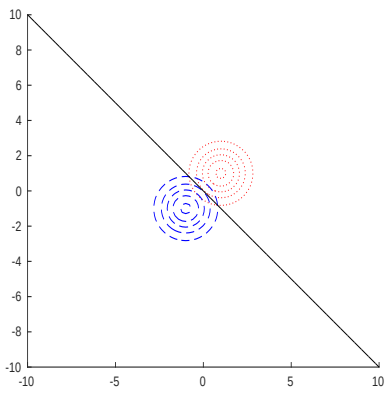
Răspuns:

a. Folosind formulele (169)**, se calculează coeficienţii** A**,** B **şi** C **din ecuaţia** $x^\top Ax + B^\top x + C = 0$**, pentru fiecare dintre cazurile date:**[336]

i. $A = 0$ **− matricea nulă de dimensiune** 2×2**,** $B^\top = (2, 2)$**,** $C = 0$**.**
Ecuaţia separatorului decizional este

$$2x_1 + 2x_2 = 0 \Leftrightarrow x_2 = -x_1.$$

Această ecuaţie reprezintă o dreaptă care trece prin originea sistemului de coordonate. Observaţi că această dreaptă este mediatoarea segmentului de dreaptă care uneşte punctele reprezentate de μ_1 **şi** μ_2 **(mediile celor două gaussiene).**

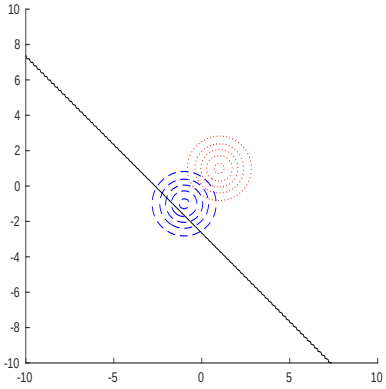


[336] Am omis calculele efective, fiindcă nu sunt dificile. Invităm studentul să verifice că este capabil să obţină el însuşi aceste rezultate.

ii. $A = 0$, $B^\top = (2, 2)$, $C = \ln 199$.
Ecuaţia separatorului decizional este

$$2x_1 + 2x_2 + \ln 199 = 0 \Leftrightarrow x_2 = -x_1 - \underbrace{\ln \sqrt{199}}_{\approx 2.6466}.$$

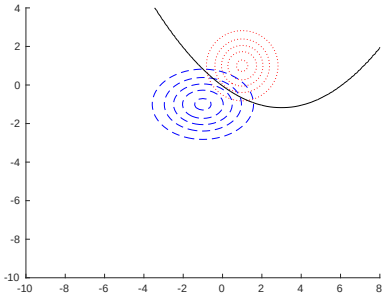
Această ecuaţie reprezintă o dreaptă paralelă cu cea de la punctul precedent (i). Ea este situată acum mult mai aproape de punctul μ_1 — datorită faptului că în acest caz probabilitatea P_0 este mult mai mare decât probabilitatea P_1.



iii. $A = \begin{pmatrix} 1/4 & 0 \\ 0 & 0 \end{pmatrix}$, $B^\top = (-3/2, -2)$, $C = 1/4$.
În acest caz, spre deosebire de cazurile precedente, ecuaţia separatorului decizional este de ordin pătratic:

$$\frac{1}{4}x_1^2 - \frac{3}{2}x_1 - 2x_2 + \frac{1}{4} = 0 \Leftrightarrow x_2 = \frac{1}{8}x_1^2 - \frac{3}{4}x_1 + \frac{1}{8}.$$

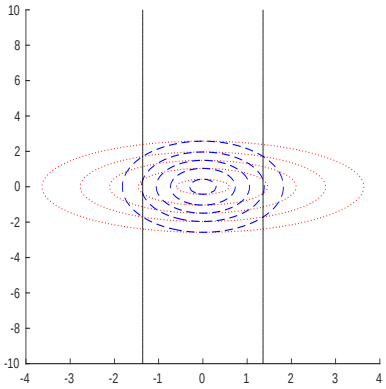
Această ecuaţie reprezintă o parabolă, al cărei vârf are coordonatele $(3, -1)$.



iv. $A = \begin{pmatrix} -3/8 & 0 \\ 0 & 0 \end{pmatrix}$, $B^\top = 0$ – **vectorul nul din** $\mathbb{R}^2$, $C = \ln 2$.
Ecuaţia separatorului decizional este

$$-\frac{3}{8}x_1^2 + \ln 2 = 0 \Leftrightarrow x_1^2 = \frac{8}{3}\ln 2.$$

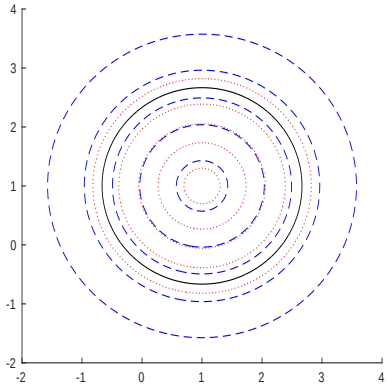
Aşadar, în acest caz, separatorul decizional este tot de ordin pătratic, dar este reprezentat de două drepte, având ecuaţiile $x_1 = -\sqrt{\frac{8}{3}\ln 2} \approx -1.3595$ **şi respectiv** $x_1 = \sqrt{\frac{8}{3}\ln 2} \approx 1.3595$.
Ţinând cont de relaţia (170) **care a fost demonstrată la problema 19, deducem că zona de decizie corespunzătoare clasei** 1 **este situată între aceste două drepte, iar zona de decizie corespunzătoare clasei** 0 **este în exteriorul lor.**



v. $A = \begin{pmatrix} 1/4 & 0 \\ 0 & 1/4 \end{pmatrix}$, $B^\top = (-1/2, -1/2)$, $C = -\ln 2 + 1/2$.

Ecuaţia separatorului decizional este

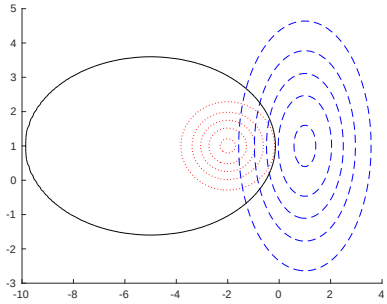
$$\frac{1}{4}(x_1^2+x_2^2)-\frac{1}{2}(x_1+x_2)-\ln 2+\frac{1}{2}=0 \Leftrightarrow$$
$$\frac{1}{4}(x_1^2-2x_1+x_2^2-2x_2)=\ln 2-\frac{1}{2} \Leftrightarrow$$
$$(x_1-1)^2+(x_2-1)^2=4\big(\ln 2-\frac{1}{2}\big)+2.$$



Această ecuaţie — tot de ordin pătratic — reprezintă un cerc cu centrul în punctul $(1,1)$ şi raza de $2\sqrt{\ln 2}\approx 1.3862$.

vi. $A=\begin{pmatrix}1/4 & 0\\ 0 & 7/8\end{pmatrix}$, $B^\top=(5/2,-7/4)$, $C=-\ln(1/4)+21/8$.
Ecuaţia separatorului decizional este

$$\frac{1}{4}x_1^2+\frac{7}{8}x_2^2+\frac{5}{2}x_1-\frac{7}{4}x_2-2\ln 2+\frac{21}{8}=0 \Leftrightarrow$$
$$\frac{1}{4}(x_1^2+10x_1)+\frac{7}{8}(x_2^2-2x_2)=2\ln 2-\frac{21}{8} \Leftrightarrow$$
$$\frac{(x_1+5)^2}{4}+\frac{(x_2-1)^2}{\frac{8}{7}}=2\ln 2-\frac{21}{8}+\frac{25}{4}+\frac{7}{8}.$$



Această ecuaţie — tot de ordin pătratic — reprezintă o elipsă care are centrul de simetrie în punctul $(-5,1)$.

b. Pornind de la regula de decizie $y_{GNB}=\operatorname{argmax}_{y\in\{0,1\}}P(Y=y|X=x)$, putem raţiona astfel:

$$P(Y=1|X=x)\geq P(Y=0|X=x) \overset{F.\ Bayes}{\Leftrightarrow}$$
$$P(X=x|Y=1)P(Y=1)\geq P(X=x|Y=0)P(Y=0).$$

Dacă alegem ca distribuţiile gaussiene corespunzătoare claselor $Y=1$ şi $Y=0$ să fie identice — adică să aibă aceiaşi parametri μ şi Σ —, în inegalitatea de mai sus factorii $P(X=x|Y=1)$ şi $P(X=x|Y=0)$ vor fi egali. Prin urmare, atunci când $P(Y=1)\neq P(Y=0)$, va rezulta că întreg planul / spaţiul va fi alocat unei singure clase, şi anume acelei clase $y\in\{0,1\}$ care are probablititatea a priori ($P(Y=y)$) mai mare.

c. Se ştie că regresia logistică determină separatori decizionali liniari.[337] Prin urmare, doar separatorii din cazurile i şi ii pot corespunde (şi) unor modele de regresie logistică.

d. Justificarea este similară cu cea de la punctul c: doar în cazul (A) separatorii decizionali pot corespunde (şi) unui model de regresie logistică multinomială / softmax.

[337] Vedeţi problema 12 de la capitolul *Metode de regresie*, în special relaţiile (123), din care rezultă:

$$P(Y=1|X=x)\geq P(Y=0|X=x)\Leftrightarrow\sigma(z)\geq 1/2\Leftrightarrow z\geq 0,$$

unde $\sigma(z)\overset{def.}{=}\frac{1}{1+e^{-z}}$, $z\overset{not.}{=}w_0+\sum_{i=1}^d w_ix_i$ şi $x\overset{not.}{=}(x_1,\ldots,x_d)$. Aşadar, în cazul regresiei logistice separatorul decizional are ecuaţia de forma $w_0+\sum_{i=1}^d w_ix_i=0$, deci este de tip liniar.

21. **(Clasificatori de tip Bayes gaussian, cazul datelor cu atribute de intrare corelate: exemplificare în $\mathbb{R}^2$; comparaţie)**

■ □ • ○ ***CMU, 2014 fall, W. Cohen, Z. Bar-Joseph, HW2, pr. 5.c***

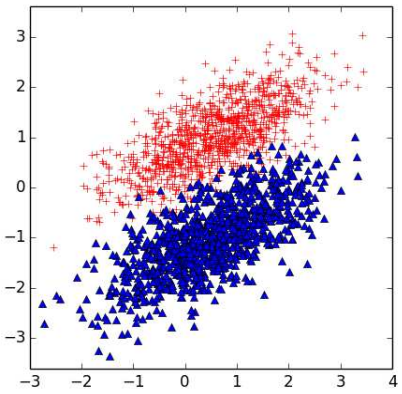
În cazul bidimensional, putem vizualiza modul în care se comportă algoritmul Bayes Naiv gaussian atunci când atributele de intrare (adică, trăsăturile; engl., features) sunt corelate (engl., correlated).[338]

Fie setul de date din figura (A) de mai jos, în care instanţele roşii sunt din clasa 0, iar instanţele albastre din clasa 1. Distribuţiile comune condiţionale ($(X_1, X_2)|Y = 0$ şi $(X_1, X_2)|Y = 1$) sunt de tip gaussian bidimensional. Elipsele din figurile (B), (C) şi (D) reprezintă curbe de izocontur pentru [diverse] distribuţii condiţionale asociate celor două clase. Centrele elipselor corespund mediilor, iar curbele de izocontur sunt situate la o distanţă de două deviaţii standard faţă de medii.[339]

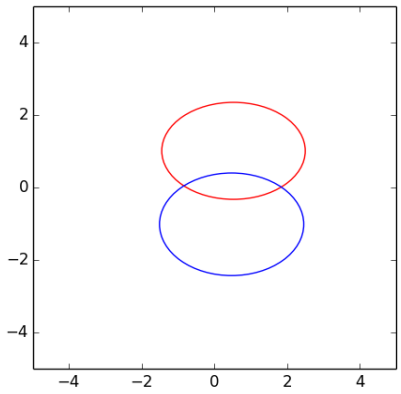
a. Care anume dintre perechile de elipse din figurile (B), (C) şi (D) corespunde cel mai probabil distribuţiilor condiţionale care au generat datele din figura (A)?

b. Care anume dintre aceste elipse corespunde [cel mai probabil] estimărilor de parametri făcute de către algoritmul Bayes Naiv gaussian?

c. Dacă presupunem că probabilităţile a priori pentru cele două clase sunt egale, care dintre modelele (B), (C) şi (D) va obţine o acurateţe mai mare pe datele de antrenament? Care va fi natura separatorului decizional în acest caz?



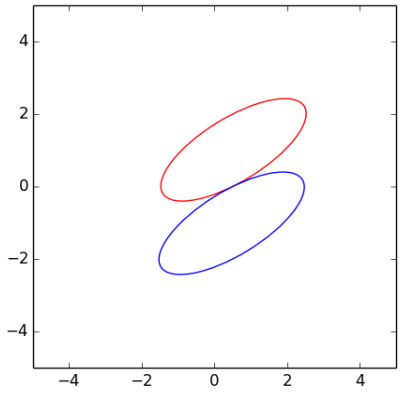
(A) Date



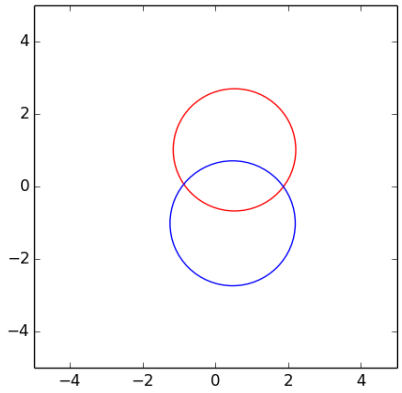
(B)

[338] Pentru definiţia noţiunii de *corelare* pentru două variabile aleatore, vedeţi problema 19 de la capitolul de *Fundamente*.

[339] Mai exact, o astfel de curbă de izocontur este constituită din punctele x din planul euclidian pentru care $\Sigma^{-1/2}(x - \mu) = 2 \Leftrightarrow (x - \mu)^\top \Sigma^{-1}(x - \mu) = 4$, unde μ este media distribuţiei gaussiene considerate. Vedeţi *Pattern Classification*, R. Duda, P. Hart, D. Stork, 2nd ed. (Wiley-Interscience, 2000), Appendix A, pag. 625.



(C)



(D)

Răspuns:

a. Se observă în figura (A) că datele au fost generate de distribuţii gaussiene bidimensionale având matrice de covarianţă nediagonale şi identice, însă având medii diferite. În plus, dacă notăm o instanţă oarecare cu $x = (x_1, x_2)$ este evident că x_1 şi x_2 nu sunt independente, ci există o anumită corelare între ele (mai precis, x_1 şi x_2 sunt într-o relaţie de dependenţă de tip liniar). În consecinţă, curbele de izocontur pentru aceste distribuţii sunt elipse identice ca mărime, având axele de simetrie neparalele cu axele sistemului de coordonate. Evident, doar desenul (C) corespunde acestei situaţii.

b. Clasificatorul Bayes Naiv gaussian presupune independenţa condiţională a celor două atribute (X_1 şi X_2) în raport cu eticheta / variabila de ieşire (Y). Această independenţă corespunde unor matrice de covarianţă diagonale, respectiv unor curbe de izocontur reprezentate de elipse ale căror axe de simetrie sunt paralele cu axele sistemului de coordonate. Atât elipsele din desenul (B) cât şi cele din desenul (D) satisfac aceste specificaţii, însă în cazul (D) elipsele sunt chiar cercuri, în vreme ce datele din figura (A) sunt dispuse în elipse având deviaţii standard diferite pe cele două axe de simetrie. Aşadar, cazul (B) corespunde estimărilor făcute de clasificatorul Bayes Naiv gaussian pe aceste date.

c. Evident, cazul (C) furnizează cea mai mică eroare la antrenare, deci cea mai mare acurateţe. În acest caz, se lucrează cu algoritmul Bayes Optimal gaussian. Întrucât matricele de covarianţă sunt egale, separatorul decizional este de tip liniar. (Pentru justificare riguroasă, vedeţi rezultatul teoretic demonstrat la problema 18.)

22. **(Algoritmul Bayes Naiv [gaussian] vs. regresia logistică — comparaţii)**

□ ● ○ ***CMU, 2012 fall, T. Mitchell, Z. Bar-Joseph, HW2, pr. 1.f, 3.b-d***

Fie un set de date caracterizate de atributele $X_1, \ldots, X_n$ (pe care le vom considera ca fiind fie toate Bernoulli, fie toate gaussiene) şi de eticheta Y.[340]

[340] Aşa este cazul problemei 14, în care variabilele condiţionate $X_i|Y$ sunt de tip Bernoulli şi sunt independente

Modelul Bayes Naiv [eventual de tip gaussian] va fi identificat în continuare cu abrevierea NB, iar modelul regresiei logistice cu LR.

a. Presupunem că datele satisfac presupoziţia de independenţă condiţională de tip Bayes Naiv. Atunci când numărul de exemple de antrenament tinde la infinit, care dintre cei doi clasificatori va produce rezultate mai bune, NB sau LR? Justificaţi.

b. Presupunem acum că datele *nu* satisfac presupoziţia de independenţă condiţională de tip Bayes Naiv. Ne punem aceeaşi întrebare ca mai sus: atunci când numărul de exemple de antrenament tinde la infinit, care dintre cei doi clasificatori va produce rezultate mai bune, NB sau LR? Justificaţi.

c. Este oare posibil să calculăm distribuţia $P(X)$ cu ajutorul parametrilor estimaţi de către algoritmul Bayes Naiv? Explicaţi în mod succint.

d. Este oare posibil să calculăm distribuţia $P(X)$ cu ajutorul parametrilor w calculaţi de către regresia logistică? Explicaţi în mod succint.

Răspuns:

a. Regresia logistică este un clasificator probabilist de tip *discriminativ*, adică aproximează / modelează $P(Y|X)$ cu ajutorul funcţiei logistice de argument $w \cdot X$ (unde w este vectorul de parametri, $w \in \mathbb{R}^d$, sau $w \in \mathbb{R}^{d+1}$ dacă extindem fiecare instanţă X cu componenta $X_0 = 1$). În contrast cu acesta, NB este un clasificator probabilist de tip *generativ*, deci calculează distribuţiile $P(X|Y = y)$ (care în cazul GNB sunt de tip gaussian multidimensional) şi de asemenea $P(Y)$, estimând parametrii acestor distribuţii.

Atunci când numărul de instanţe tinde la infinit, pe de o parte aproximaţia calculată de regresia logistică pentru distribuţia $P(Y|X)$ tinde la distribuţia reală $P(Y|X)$, iar pe de altă parte estimările făcute de clasificatorul Bayes Naiv pentru distribuţiile $P(Y)$ şi $P(X|Y)$ vor tinde la distribuţiile reale corespunzătoare, dat fiind (în cazul lui $P(X|Y)$) că datele de antrenament satisfac presupoziţia de independenţă condiţională. Corespondenţa dintre distribuţiile reale $P(Y|X)$ (pe de o parte) şi $P(Y)$ şi $P(X|Y)$ (pe de altă parte) este dată de formula lui Bayes. Prin urmare, în aceste condiţii cei doi clasificatori vor produce rezultate echivalente.

b. Regresia logistică va produce rezultate mai bune, fiindcă ea nu lucrează cu presupoziţia de independenţă condiţională. (Vedeţi pe de o parte problema 32 de la capitolul *Metode de regresie*, iar pe de altă parte problema 10 de la prezentul capitol.)

c. Da, algoritmul Bayes Naiv este un clasificator de tip *generativ* (engl., generative classifier). Putem calcula $P(X)$ prin „marginalizarea“ distribuţiei condiţionate $P(X|Y)$ în raport cu eticheta / clasa Y, şi anume: $P(X) = \sum_y P(X|Y = y) \cdot P(Y = y)$.

d. Nu, nu este posibil. Aşa cum am precizat la punctul a, regresia logistică estimează $P(Y|X)$ (nu $P(X|Y)$ şi $P(Y)$, cum calculează clasificatorii de tip Bayes Naiv).

condiţional două câte două, sau cel al problemei 17, în care toate variabilele condiţionate, $X_i|Y$ pentru $i = 1, \ldots, n$ urmează distribuţii gaussiene — având varianţele $\sigma_{i0} = \sigma_{i1}$ — şi sunt, de asemenea, independente condiţional două câte două. (Similar este şi cazul problemei 44, care constituie o combinaţie a precedentelor două tipuri.)

23. **(Clasificarea bayesiană gaussiană vs. regresia logistică: Adevărat sau Fals?)**

□ ○ ***CMU, 2010 fall, Aarti Singh, midterm, pr. 1.2***
CMU, (?) 15-781, midterm example questions, pr. 1.d

a. Corespondenţa dintre regresia logistică şi clasificatorul Bayes Naiv de tip gaussian înseamnă — în cazul în care matricele de covarianţă corepunzătoare claselor sunt toate egale cu matricea-identitate[341] — că există o corespondenţă 1-la-1 între parametrii celor doi clasificatori.

b. Presupunând că lucrăm cu un număr fix de atribute, putem învăţa un clasificator Bayes Optimal de tip gaussian în timp liniar în raport cu numărul de instanţe din setul de date de antrenament.

Răspuns:

a. Fals. Se poate preciza de la început că deşi cei doi clasificatori învaţă separatori decizionali care au aceeaşi formă (si anume, o formă liniară, vedeţi problema 17) nu rezultă în mod neapărat că pe un acelaşi set de date cei doi separatori decizionali învăţaţi coincid. Faptul că matricele de covarianţă sunt, toate, matrice identitate / diagonale înseamnă că presupoziţia de independenţă condiţională este satisfăcută.[342] Suntem, aşadar, în condiţii similare cu cele de la problema 22.a, însă aici nu avem neapărat satisfăcută ipoteza că numărul de instanţe de antrenament tinde la infinit, caz în care rezultatele de clasificare furnizate de LR şi NB ar fi echivalente. Aşadar, nu există (în general) o corespondenţă de tip 1-la-1 între parametrii w_{LR} calculaţi de regresia logistică[343] şi parametrii w_{NB} corespunzători clasificatorului Bayes Naiv gaussian.[344]

***Observaţie*: Rezultatul acesta este valabil şi în cazul algoritmului Bayes Naiv cu variabile categoriale.**

b. Adevărat. În cazul clasificatorului Bayes Optimal de tip gaussian, regula de calcul pentru clasificarea unei instanţe noi $x \in \mathbb{R}^d$ este următoarea:

$$\begin{aligned} y_{GNB} &\overset{def.}{=} \operatorname*{argmax}_{y \in Val(Y)} P(Y=y|X=x) = \operatorname*{argmax}_{y \in Val(Y)} \frac{P(X=x|Y=y)P(Y=y)}{P(X=x)} \\ &= \operatorname*{argmax}_{y \in Val(Y)} P(X=x|Y=y)P(Y=y) = \operatorname*{argmax}_{y \in Val(Y)} \mathcal{N}(x; \mu_y, \Sigma_y)P(Y=y). \end{aligned}$$

Aşadar, pentru fiecare $y \in Val(Y)$ vom avea de estimat câte o pereche de parametri, care caracterizează o distribuţie gaussiană multidimensională: vectorul de medii μ_y şi matricea de covarianţă Σ_y. Conform problemei 48 de la capitolul de *Fundamente*, estimările celor doi parametri sunt media la eşantionare şi respectiv matricea de covarianţă la eşantionare, deci se calculează în timp liniar în raport cu numărul de instanţe de antrenament.

[341] LC: Chiar mai general, putem considera că aceste matrice sunt diagonale.
[342] Vedeţi problema 31 de la capitolul de *Fundamente*.
[343] Vedeţi problema 12 de la capitolul *Metode de regresie*.
[344] În primul rând, *din punctul de vedere al terminologiei*, dacă ne referim la *parametrii* calculaţi de NB ca fiind parametrii distribuţiilor $P(X|Y=y)$ şi $P(Y)$, este imediat că nu există o corespondenţă 1-la-1 între aceştia şi parametrii w_{LR} calculaţi de LR, care servesc la estimarea / aproximarea distribuţiei $P(Y|X)$.

2.2 Clasificare bayesiană — Probleme propuse

2.2.1 Ipoteze de probabilitate maximă a posteriori (MAP)

24. **(Formula lui Bayes; inferenţe statistice; ilustrarea noţiunii de ipoteze MAP)**

* ***CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW1, pr. 5***

Imaginează-ţi că te afli în faţa a trei cutii, care sunt etichetate cu literele A, B, C. Două dintre ele sunt goale, iar una conţine un premiu. Tu nu ştii în care dintre ele se află premiul; trebuie să ghiceşti. Procedezi în felul următor:

Mai întâi alegi la întâmplare o cutie X (să zicem că $X = A$). Totuşi, chiar înainte de a deschide cutia X observi că altcineva, înaintea ta, a deschis cutia Y (să zicem că $Y = B$). Ai dreptul să priveşti în cutia Y, ca să vedeţi dacă ea conţine sau nu premiul. Dacă ea nu conţine premiul, vei avea dreptul să schimbi alegerea pe care ai făcut-o iniţial.

În vederea luării unei decizii cât mai bune, ţi se comunică *strategia* după care a fost aleasă cutia Y, şi anume, folosind una din următoarele trei *variante*:

a. **Dacă cutia pe care ai ales-o iniţial conţine premiul, atunci cutia Y este aleasă cu probabilitate de $1/2$ una din cele două cutii goale (diferite de cutia X). Dacă X este vidă, atunci Y se alege ca fiind cutia goală diferită de X.**

b. **Se alege aleatoriu cu probabilitate de $1/2$ una din cutiile diferite de cea pe care ai ales-o tu iniţial, X. (În consecinţă, cutia Y poate sau nu să conţină premiul. Dacă Y conţine premiul, ai pierdut jocul.)**

c. **Se alege în mod aleatoriu cu probabilitate de $1/2$ una din cutiile goale. (Deci este posibil ca $Y = X$. Aşadar, în cazul $Y = X$ observi că a fost deschisă anterior chiar cutia X. În continuare vei putea alege una din celelalte două cutii.)**

Considerând (pentru simplitate) că $X = A$, $Y = B$, iar cutia B este vidă, pentru fiecare din cele trei *variante* de mai sus decide ce cutie ar trebui să alegi în final pentru a-ţi maximiza şansele de a obţine premiul. Justifică-ţi decizia, elaborând calculul probabilistic aferent.

25. **(Adevărat sau Fals?)**

CMU, 2002 fall, Andrew Moore, final exam, pr. 11.a

Fie D un set de exemple (date de antrenament), iar H o mulţime de ipoteze pentru (un algoritm de) învăţare automată pe datele D. Precizaţi care este *valoarea de adevăr* a următoarelor afirmaţii:

$\text{argmax}_{h \in H} P(D|h)$ **este ipoteză de probabilitate maximă a posteriori, [iar]**
$\text{argmax}_{h \in H} P(h|D)$ **este ipoteză de verosimilitate maximă.**

2.2.2 Algoritmii Bayes Naiv şi Bayes Optimal

26. **(Algoritmii Bayes Naiv şi Bayes Optimal; aplicare; numărul minimal de parametri de estimat)**

• ○ ***Liviu Ciortuz, 2017, pornind de la setul de date din Machine Learning, Tom Mitchell, 1997, ch. Decision Trees, page 59***

Considerăm următorul set de date de antrenament, în care variabila de ieşire este *EnjoyTennis*:

Day	Outlook	Temperature	Humidity	Wind	EnjoyTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

a. Determinaţi decizia luată de către algoritmul Bayes Naiv pentru instanţa de test

$$X = \langle Outlook = sunny,\ Temp = cool,\ Humidity = high,\ Wind = strong \rangle,$$

precum şi probabilitatea cu care este luată această decizie.

b. Care este numărul *minim* de parametri pe care trebuie să-l estimeze algoritmul Bayes Naiv pe aceste date [pentru a face apoi predicţii pe un set oarecare de instanţe de test]? Dar în cazul clasificatorului Bayes Optimal?

c. Implementaţi algoritmul Bayes Naiv, iar apoi cu ajutorul acestei implementări calculaţi eroarea la antrenare şi eroarea la CVLOO pe acest set de date.

27. **(Algoritmul Bayes Naiv: aplicare)**

∗ ***CMU, 2004 fall, T. Mitchell Z. Bar-Joseph, midterm, pr. 6.a***

Se dă setul de date din tabelul alăturat, în care A, B, C sunt atribute (de intrare) binare, iar Y este atribut de ieşire. Care va fi răspunsul algoritmului de clasificare Bayes Naiv pentru intrarea $A = 0,\ B = 0,\ C = 1$?

A	B	C	Y
0	0	1	0
0	1	0	0
1	1	0	0
0	0	1	1
1	1	1	1
1	0	0	1
1	1	0	1

28. **(Algoritmul Bayes Naiv şi algoritmul Bayes Optimal: aplicare)**

prelucrare de Liviu Ciortuz, după

• ∗ *CMU, 2002 fall, Andrew Moore, final exam, pr. 4.b-e*

Se dă setul de date alăturat, cu A şi B variabile de intrare, iar C variabilă de ieşire.

A	B	C	nr. apariţii
0	0	1	3
0	1	0	1
0	1	1	4
1	0	0	5
1	1	0	2
1	1	1	1

a. Care este numărul minim de probabilităţi ce trebuie estimate pentru a putea construi după aceea (pe acest set de date) un clasificator de tip Bayes Naiv? Justificaţi.

b. Similar, pentru clasificatorul Bayes Optimal. Justificaţi.

c. Care este decizia clasificatorului Bayes Naiv pentru $A = 0$, $B = 1$? Precizaţi cu ce probabilitate este luată această decizie.

d. Care este decizia clasificatorului Bayes Optimal pentru $A = 0$, $B = 1$? Precizaţi cu ce probabilitate este luată această decizie.

e. Dacă rezultatele obţinute la punctele c şi d diferă (fie şi numai în privinţa *probabilităţilor* cu care sunt luate deciziile), care este explicaţia? Justificaţi în mod riguros.

29. **(Aplicarea algoritmului Bayes Naiv la clasificarea de texte)**

• ∗ *Edinburgh, 2009 fall, C. Williams, V. Lavrenko, tutorial 2, pr. 2*

Firma Whizzco decide să implementeze un clasificator de texte. Pentru început, ei vor să clasifice documente aparţinând fie clasei *sport* fie clasei *politică*. Ei decid să reprezinte fiecare document ca un vector de atribute descriind prezenţa ori absenţa unor cuvinte-cheie:

goal, football, golf, defence, offence, wicket, office, strategy.

Datele de antrenament sunt reprezentate folosind o matrice în care fiecare linie este un vector de valori (0 sau 1) pentru cele 8 atribute.

```
xP=[1 0 1 1 1 0 1 1; % Politica
    0 0 0 1 0 0 1 1;
    1 0 0 1 1 0 1 0;
    0 1 0 0 1 1 0 1;
    0 0 0 1 1 0 1 1;
    0 0 0 1 1 0 0 1]
```

```
xS=[1 1 0 0 0 0 0 0; % Sport
    0 0 1 0 0 0 0 0;
    1 1 0 1 0 0 0 0;
    1 1 0 1 0 0 0 1;
    1 1 0 1 1 0 0 0;
    0 0 0 1 0 1 0 0;
    1 1 1 1 1 0 1 0]
```

Folosind algoritmul Bayes Naiv, care este probabilitatea cu care documentul $x = (1, 0, 0, 1, 1, 1, 1, 0)$ va fi clasificat ca aparţinând clasei *politică*?

30. **(Aplicarea algoritmului Bayes Naiv: chestiunea valorilor lipsă (engl., missing values) în datele de antrenament)**

□ • ***CMU, 2013 fall, A. Smola, G. Gordon, midterm practice, pr. 9***

Fie setul de date de antrenament (x, y) şi datele de test z:

$$\begin{array}{ll} x_1 = (0,0,0,1,0,0,1) & y_1 = 1 \\ x_2 = (0,0,1,1,0,0,0) & y_2 = 1 \\ x_3 = (1,1,0,0,0,1,0) & y_3 = -1 \\ x_4 = (1,0,0,0,1,1,0) & y_4 = -1 \end{array}$$

$$\begin{array}{l} z_1 = (1,0,0,0,0,1,0) \\ z_2 = (0,1,1,0,0,1,1) \end{array}$$

Ce *problemă* va întâmpina clasificatorul Bayes Naiv pe aceste date? (*Indicaţie*: Pentru ca răspunsul dumneavoastră să fie cât mai bine justificat, veţi estima toţi parametrii necesari şi veţi aplica algoritmul pe cele două instanţe de test. Veţi nota atributele cu $A1, A2, \ldots$.)

La curs am prezentat un „remediu“ standard pentru o astfel de *problemă*. Precizaţi cum se numeşte „tehnica“ respectivă şi aplicaţi-o pe aceste date. După aceea, veţi aplica algoritmul Bayes Naiv pentru a clasifica instanţele de test z_1 şi z_2.

31. **(Algoritmului Bayes Naiv: calculul ratei medii a erorilor)**

■ • ○ ***CMU, 2010 fall, Aarti Singh, HW1, pr. 4.2***

Considerăm următoarea problemă de clasificare:
Fie variabila aleatoare $Y : \boldsymbol{Hike} \in \{T, F\}$ care denotă faptul că Alice şi Bob merg sau nu în drumeţie în funcţie de condiţiile vremii: $X_1 : \boldsymbol{Sunny} \in \{T, F\}$ şi $X_2 : \boldsymbol{Windy} \in \{T, F\}$.

Se presupune că au fost estimaţi următorii parametri:

$$\begin{array}{ll} P(\boldsymbol{Hike}) = 0.5 & \\ P(\boldsymbol{Sunny} \mid \boldsymbol{Hike}) = 0.8, & P(\boldsymbol{Sunny} \mid \neg\boldsymbol{Hike}) = 0.7 \\ P(\boldsymbol{Windy} \mid \boldsymbol{Hike}) = 0.4, & P(\boldsymbol{Windy} \mid \neg\boldsymbol{Hike}) = 0.5 \end{array}$$

De asemenea, se consideră că este satisfăcută presupoziţia de independenţă condiţională a algoritmului Bayes Naiv.

a. Care este probabilitatea (comună) ca Alice şi Bob să meargă în drumeţie atunci când vremea este însorită şi bate vântul, adică

$$P(\boldsymbol{Sunny} = T, \boldsymbol{Windy} = T, \boldsymbol{Hike} = T) = ?$$

Care este decizia luată de algoritmul Bayes Naiv în acest caz?

b. Completaţi tabelul următor:

X_1	X_2	Y	$P(X_1, X_2, Y)$	$P_{NB}(Y\|X_1, X_2)$	***decizia algoritmului Bayes Naiv***
F	F	F			
F	F	T			
F	T	F			
F	T	T			
T	F	F			
T	F	T			
T	T	F			
T	T	T			

***Observaţie*: Calculele de la punctul a corespund ultimei linii din tabelul de mai sus.**

c. Care este rata medie a erorilor (engl., expected error rate) produse de algoritmul Bayes Naiv? Vă reamintim că această (rată) medie este definită ca fiind suma probabilităţilor $P(X_1, X_2, Y)$ pentru acele (triplete de) valori ale variabilelor X_1, X_2, Y pentru care decizia luată de algoritmul Bayes Naiv diferă de valoarea variabilei Y.

În cele ce urmează se presupune că se obţin mai multe informaţii despre vreme. Se introduce o nouă trăsătură X_3:*Rainy* $\in\{T, F\}$. Se presupune că în fiecare zi vremea poate fi fie *Rainy* fie *Sunny*, dar nu şi *Rainy* şi *Sunny*. Similar, se presupune că vremea nu poate fi într-o zi $\neg$*Rainy* şi $\neg$*Sunny*.

d. În noile condiţii, presupoziţia de independenţă condiţională rămâne oare adevărată? Justificaţi.

e. Calculaţi $P(\textit{Sunny} = T, \textit{Windy} = T, \textit{Rainy} = F, \textit{Hike} = T)$.

f. Care este rata medie a erorilor produse de clasificatorul Bayes Naiv când se folosesc toate cele 3 atribute de intrare?

g. S-a îmbunătăţit performanţa algoritmului Bayes Naiv prin adăugarea atributului *Rainy*? Explicaţi de ce.

32. **(Algoritmul Bayes Naiv: calculul ratei medii a erorii – exemplificare; comparaţie cu regresia logistică)**

prelucrare de Liviu Ciortuz, după
• ∗ ***CMU, 2009 fall, Carlos Guestrin, HW1, pr. 4.1.4***

Considerăm o problemă de clasificare binară în care fiecare exemplu de antrenament are două atribute binare $X_1, X_2 \in \{T, F\}$ şi eticheta / clasa $Y \in \{T, F\}$. Presupunem că $P(Y = T) = 0.5$, iar $P(X_1 = T|Y = T) = 0.8$, $P(X_1 = F|Y = F) = 0.7$, $P(X_2 = T|Y = T) = 0.5$ şi $P(X_2 = F|Y = F) = 0.9$. (Se poate observa că atributul X_1 furnizează / constituie un indiciu întrucâtva mai puternic decât atributul X_2 în ce priveşte determinarea clasei unei instanţe oarecare.)

În cele ce urmează vom presupune că X_1 şi X_2 sunt independente în raport cu Y.

a. Calculaţi probabilităţile $P(X_1 = F|Y = T)$, $P(X_1 = T|Y = F)$, $P(X_2 = F|Y = T)$ şi $P(X_2 = T|Y = F)$. Asociaţi răspunsului dumneavoastră o justificare generală, sub forma unei formule din teoria probabilităţilor:

$$P(\neg A|B) = \ldots, \text{ unde } A \text{ şi } B \text{ sunt evenimente aleatoare oarecare.}$$

b. Scrieţi regula de decizie a algoritmului Bayes Naiv pentru $X_1 = x_1$ şi $X_2 = x_2$, justificând în mod succint obţinerea ei.

c. Calculaţi rata medie a erorii produse de algoritmul Bayes Naiv, atunci când se folosesc ambele atribute, X_1 şi X_2. (Veţi da în prealabil definiţia ratei medii a erorii.) Este oare această rată mai bună decât în cazul în care se foloseşte un singur atribut (X_1 sau X_2)? De ce?

d. Să presupunem acum că se crează un nou atribut, X_3, care este o copie exactă a lui X_2. Aşadar, pentru fiecare exemplu de antrenament, atributele X_2 şi X_3 au aceeaşi valoare, $X_2 = X_3$. Răspundeţi la următoarele întrebări:

– Sunt X_2 şi X_3 independente condiţional în raport cu Y?

– Cât este rata medie a erorii pentru Bayes Naiv acum? (Atenţie! Distribuţia „adevărată“ a datelor nu s-a modificat.)

– Explicaţi ce se întâmplă cu algoritmul Bayes Naiv. Oare *regresia logistică* are aceeaşi problemă? Explicaţi de ce.

33. **(Clasificare bayesiană: calculul ratei medii a erorilor pentru diverşi clasificatori bayesieni)**

prelucrare de L. Ciortuz, după
■ • ○ ***CMU, 2004 fall, T. Mitchell Z. Bar-Joseph, HW3, pr. 1.2***

Fie funcţia $Y = (A \wedge B) \vee \neg(B \vee C)$, unde A, B şi C sunt variabile aleatoare binare independente, fiecare dintre ele având posibilitatea să ia valoarea 0 cu probabilitate de 50%.

a. Câţi parametri trebuie să estimeze clasificatorul Bayes Naiv pentru a învăţa funcţia Y? Enumeraţi aceşti parametri. *Atenţie*: $P(\neg x)$ nu va fi socotit ca parametru dacă $P(x)$ a fost deja estimat ca parametru.

b. Care este rata medie a erorii la antrenare pentru clasificatorul Bayes Naiv la învăţarea conceptului Y, presupunând că avem o infinitate de date de antrenament?

***Indicaţie*: Scrieţi mai întâi tabela de adevăr a funcţiei Y, apoi estimaţi valorile parametrilor (în sensul verosimilităţii maxime, MLE). Pentru convenienţă, centralizaţi toate calculele făcute de către algoritmul Bayes Naiv într-un tabel. (Sau, altfel, puteţi adăuga nişte coloane suplimentare la tabela de adevăr a funcţiei Y.)**

***Convenţie*: În cazul în care, pentru o setare oarecare a variabilelor A, B şi C, cele două probabilităţi calculate de către algoritmul Bayes Naiv în vederea determinării valorii y_{NB} sunt egale, convenim că algoritmul va lua decizia $y_{NB} = 1$.**

c. Câţi parametri trebuie să estimeze clasificatorul Bayes Optimal pentru a „învăţa“ funcţia Y? Justificaţi în detaliu.

d. Care este rata medie a erorii la antrenare pentru clasificatorul Bayes Optimal la învăţarea conceptului Y, presupunând acelaşi lucru ca mai sus? *Atenţie*: Nu este nevoie să calculaţi efectiv această rată; este suficient să indicaţi valoarea ei şi să o justificaţi printr-un *raţionament calitativ*.[345]

e. Considerăm un alt clasificator de tip Bayes, care presupune că A este independent de C, condiţionat de B şi Y — în contrast cu clasificatorul Bayes Naiv, care presupune că variabilele A, B şi C sunt independente două câte două în raport cu Y.

Arătaţi că acest clasificator Bayes va avea nevoie să estimeze mai puţini parametri decât clasificatorul Bayes Optimal la învăţarea conceptului Y, şi totuşi va obţine aceeaşi rată medie a erorii la antrenare (considerând că este valabilă aceeaşi presupoziţie în legătură cu datele de antrenament).

***Indicaţii*:**

– Scrieţi tabela de adevăr a funcţiei Y.

– Folosind această tabelă de adevăr, dovediţi că într-adevăr proprietatea de independenţă condiţională formulată mai sus este satisfăcută. Şi anume: pentru fiecare pereche de valori b şi y ale variabilelor aleatoare B şi respectiv Y, arătaţi că

$$P(A = a, C = c|B = b, Y = y) = P(A = a|B = b, Y = y) \cdot P(C = c|B = b, Y = y),$$

pentru $\forall a \in Val(A)$ şi $\forall c \in Val(C)$.

– Scrieţi regula de decizie pentru acest nou clasificator de tip Bayes şi arătaţi că ea este echivalentă cu regula de decizie a clasificatorului Bayes Optimal.

– Calculaţi numărul minim de parametri de estimat de către noul nostru clasificator Bayes şi enumeraţi-i.

34. **(Algoritmul Bayes Naiv: independenţa [condiţională a] atributelor de intrare; calculul ratei medii a erorii)**

□ • ***CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW3, pr. 1.1***

Presupunem că A şi B sunt variabile aleatoare binare independente, fiecare dintre ele având posibilitatea de a lua valoarea 0 cu o probabilitate de 50%.

Definiţi o funcţie booleană $y = f(A, B)$ în aşa fel încât variabila A să *nu* fie independentă de variabila B în raport cu y (văzut şi el ca variabilă aleatoare), însă clasificatorul Bayes Naiv să producă o rată medie a erorii de 0% (presupunând că datele de antrenament sunt în număr infinit).

Demonstraţi că acest clasificator are într-adevăr rata erorii de 0%.

[345] LC: Totuşi, este recomandabil să procedaţi aşa după ce în prealabil aţi văzut ce valoare produce algoritmul Bayes Optimal pentru [măcar] una dintre combinaţiile de valori ale variabilelor, de exemplu, $A = B = C = 0$.

35. **(Algoritmul Bayes Naiv: comparaţie cu alţi clasificatori)**

prelucrare de L. Ciortuz, după
∗ *CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, midterm, pr. 6.b*

Considerăm un clasificator Bayes Naiv care lucrează pe un set de date descrise de atributele de intrare A şi B şi atributul de ieşire Y. (Exemplu: $Y = A \text{ XOR } B$). A şi B sunt variabile aleatoare independente între ele.

a. Este posibil ca, în situaţia aceasta, vreun alt clasificator — de exemplu, ID3, regresia logistică (eventual kernel-izată) sau SVM — să lucreze mai bine decît clasificatorul Bayes Naiv?

b. Care este motivul?

36. **(Comparaţie între clasificatorul Bayes Naiv şi algoritmul ID3)**

□ • ○ *CMU, 2010 fall, Ziv Bar-Joseph, midterm, pr. 5.b*

Care dintre afirmaţiile de mai jos sunt adevărate atât pentru clasificatorul Bayes Naiv cât şi pentru algoritmul ID3 pentru învăţarea de arbori de decizie? (Veţi putea alege nu neapărat una singură dintre aceste afirmaţii.)

1. În cazul ambilor clasificatori se presupune că orice pereche de atribute X_i şi X_j cu $i \neq j$ — văzute ca variabile aleatoare — sunt independente.
2. În cazul ambilor clasificatori se presupune că orice pereche de atribute X_i şi X_j cu $i \neq j$ sunt dependente.
3. În cazul ambilor clasificatori se presupune că orice pereche de atribute sunt independente în raport cu eticheta (adică variabila care reprezintă clasa).
4. În cazul ambilor clasificatori se presupune că orice pereche de atribute sunt dependente în raport cu eticheta.

37. **(Algoritmul Bayes Naiv – clasificator de tip MAP; o condiţie [suficientă] pentru echivalenţa cu clasificarea de tip ML)**

□ • ○ *CMU, 2012 fall, T. Mitchell, Z. Bar-Joseph, HW3, pr. 4.2*

Algoritmul Bayes Naiv asociază unui exemplu x clasa c dacă aceasta maximizează probabilitatea $P(c|x)$.
Când este această condiţie [din formularea algoritmului Bayes Naiv] echivalentă cu selecţionarea acelei clase c care maximizează probabilitatea $P(x|c)$?

38. **(Algoritmii Bayes Naiv şi Bayes Optimal: Adevărat sau Fals?)**

• ○ *CMU, 2005 spring, C. Guestrin, T. Mitchell, midterm, pr. 2.b.5*
CMU, 2011 spring, Tom Mitchell, midterm, pr. 1.1.ab

a. Clasificatorul Bayes Optimal poate să obţină rata de eroare 0 [la antrenare] pentru orice set de date. Justificaţi.

b. Dacă antrenăm un clasificator Bayes Naiv folosind un număr infinit de date de antrenament care satisfac toate presupoziţiile luate în calcul de acest tip de modelare (de exemplu, independenţa condiţională), atunci acest clasificator va produce *eroare* zero pe setul de exemple de antrenament considerat.

c. Dacă antrenăm un clasificator Bayes Naiv folosind un număr infinit de date de antrenament care satisfac toate presupoziţiile luate în calcul de acest tip de modelare (de exemplu, independenţa condiţională), atunci acest clasificator va produce *eroare „adevărată“* zero pentru exemple de test generate conform aceleiaşi distribuţii.

2.2.3 Clasificare bayesiană [cu atribute de intrare] de tip gaussian

39. **(Distribuţia gaussiană unidimensională: exemplificare pentru estimarea mediei în sens MLE şi calcularea unei probabilităţi a posteriori)**

● ○ ∗ ***U. Edinburgh, 2009 fall, C. Williams, V. Lavrenko, HW2, pr. 1***

Considerăm un set de date [de antrenament] care constă din *exemple* pentru două *clase*. Exemplele pentru clasa 1 sunt 0.5, 0.1, 0.2, 0.4, 0.3, 0.2, 0.2, 0.1, 0.35, 0.25**, iar exemplele pentru clasa 2 sunt** 0.9, 0.8, 0.75, 1.0**.**

a. Vă cerem să antrenaţi (engl., fit) câte o distribuţie gaussiană unidimensională pentru fiecare din cele două clase, folosind metoda estimării în sensul verosimilităţii maxime (engl., Maximum Likelihood estimation, MLE). Veţi presupune că varianţa distribuţiei corespunzătoare clasei 1 este 0.0149**, iar varianţa distribuţiei pentru clasa 2 este** 0.0092**.**

b. Estimaţi de asemenea în sens MLE probabilităţile *a priori* de selecţie pentru [generarea de exemple din] cele două clase.

c. Care este probabilitatea *a posteriori* ca punctul $x = 0.6$ să aparţină clasei 1?

40. **(Algoritmul Bayes [Naiv] gaussian: exemplificare pe date din $\mathbb{R}$, graniţe de decizie)**

□ ● ○ ***CMU, 2009 spring, Tom Mitchell, midterm, pr. 5***

În acest exerciţiu vom considera mai mulţi clasificatori de tip Bayes Naiv gaussian (GNB) pentru un set de date având un singur atribut x şi două clase, 0 şi 1.[346] **Ca de obicei pentru clasificatori bayesieni, vom clasifica o instanţă x ca aparţinând clasei 1 dacă**

$$P(y = 1|x) \geq P(y = 0|x) \Leftrightarrow \ln \frac{P(y = 1|x)}{P(y = 0|x)} \geq 0.$$

[346] Fiind dat un singur atribut de intrare, putem renunţa la termenul „Naiv“ din expresia care desemnează tipul clasificatorului.

3 Învăţare bazată pe memorare

Sumar

Noţiuni preliminare

– **măsuri de distanţă, măsuri de similaritate: ex. 2;**

– **normă într-un spaţiu vectorial; [măsura de] distanţă indusă de către o normă: ex. 7;**

– **k-NN vecinătate a unui punct din $\mathbb{R}^d$.**

Algoritmul k-NN

– **pseudo-cod: cartea ML, pag. 232;**

– **bias-ul inductiv: „Cine se aseamănă se adună“ (sau: „Spune-mi cu cine te împrieteneşti, ca să-ţi spun cine eşti“): ex. 15.a;**

– **exemple (simple) de aplicare: ex. 1, ex. 2;**

– **complexitate de spaţiu:** $\mathcal{O}(d\,n)$
complexitate de timp:

la antrenare: $\mathcal{O}(d\,n)$
la testare: $\mathcal{O}(d\,n\log n)$
[**LC:** $\mathcal{O}(d\,n\,k\log k)$ **pt.** $k>1$ **(worst case) şi** $\mathcal{O}(d\,n)$ **pt.** $k=1$],

unde d este numărul de atribute, iar n este numărul de exemple;

– **arbori kd (engl., kd-trees): *Statistical Pattern Recognition*, Andrew R. Webb, 3rd ed., 2011, Willey, pag. 163-173;**

– **k-NN ca algoritm ML “lazy” (vs. “eager”):**
suprafeţe de decizie şi graniţe de decizie:
diagrame Voronoi pentru 1-NN: ex. 4, ex. 11.a, ex. 18, ex. 19, ex. 20.a;

– **analiza erorilor:**

- **1-NN pe date consistente: eroarea la antrenare este 0: ex. 2, ex. 12.a;**
- **variaţia numărului de erori (la antrenare şi respectiv testare) în funcţie de valorile lui k: ex. 22, ex. 23.ab;**
  k-NN ca metodă neparametrică; alegerea lui k: CV: ex. 23.c;
- **CVLOO: ex. 3, ex. 12.b, ex. 15.bc, ex. 16, ex. 24.a, ex. 20.b;**
- **sensibilitatea / robusteţea la „zgomote“: ex. 5, ex. 15;**
- **eroarea asimptotică: ex. 10, ex. 25.**

– **efectul trăsăturilor redundante sau irelevante;**

– **alegerea valorii convenabile pentru k: ex. 21.**

Proprietăţi ale algoritmului k-NN

- **(P0) output-ul algoritmului k-NN pentru o instanţă oarecare de test x_q depinde de valoarea lui k: ex. 1;**
- **(P1) pe seturi de date de antrenament *consistente*, eroarea la antrenare produsă de algoritmul 1-NN este 0: ex. 2, ex. 12.a;**
- **(P2) output-ul algoritmului k-NN, precum şi suprafeţele de decizie şi separatorii decizionali depind de *măsura de distanţă* folosită: ex. 7;**
- **(P3) „blestemul marilor dimensiuni“ (engl., the curse of dimensionality): în anumite condiţii, numărul de instanţe de antrenament necesare pentru a avea un *cel mai apropiat vecin* situat la distanţă *rezonabilă* faţă de instanţa de test x_q creşte exponenţial în funcţie de numărul de atribute folosite: ex. 9;**
- **(P4) în anumite condiţii, rata medie a *erorii asimptotice* a algoritmului 1-NN este mărginită superior de dublul ratei medii a erorii algoritmului Bayes Optimal: ex. 10, ex. 25.**

Comparaţii cu alţi algoritmi de clasificare automată

– **ID3: ex.11.b, ex. 13.ab;**

– **SVM: ex.12, ex. 13.c, ex. 24.b;**

– **regresia logistică: ex. 24.b;**

– **1-NN cu mapare cu RBF: ex. 14.**

Variante ale algoritmului k-NN

– **k-NN folosind alte măsuri de distanţă (decât distanţa euclidiană): ex. 7;**

– **k-NN cu *ponderarea distanţelor* (engl., distance-weighted k-NN): cartea ML, pag. 236-238 (formulele 8.2, 8.3, 8.4);**[351]

– **algoritmul lui Shepard: ex.8.**

Alte metode de tip IBL

– **reţele RBF: cartea ML, pag. 238-240;**

– **raţionare bazată pe cazuri (engl., case-based reasoning): cartea ML, pag. 240-244.**

[351] Sectiunea 8.3 din cartea ML (pag. 236-238) se referă la regresia [liniară] local-ponderată ca o formă mai generală de aproximare a [valorilor] funcţiilor, în raport cu cele calculate de către algoritmul k-NN atunci când se foloseşte ponderarea distanţelor.

3.1 Învăţare bazată pe memorare — Probleme rezolvate

1. **(Algoritmul k-NN: aplicare în $\mathbb{R}^2$ pentru diferite valori ale lui k)**

■ • ***CMU, 2006 fall, E. Xing, T. Mitchell, final exam, pr. 2***

Fie setul de instanţe de antrenament din tabelul alăturat:

a. Vizualizaţi datele într-un sistem de axe din $\mathbb{R}^2$.

b. Presupunând că se foloseşte distanţa euclidiană, care va fi predicţia făcută pentru punctul (1,1) de către

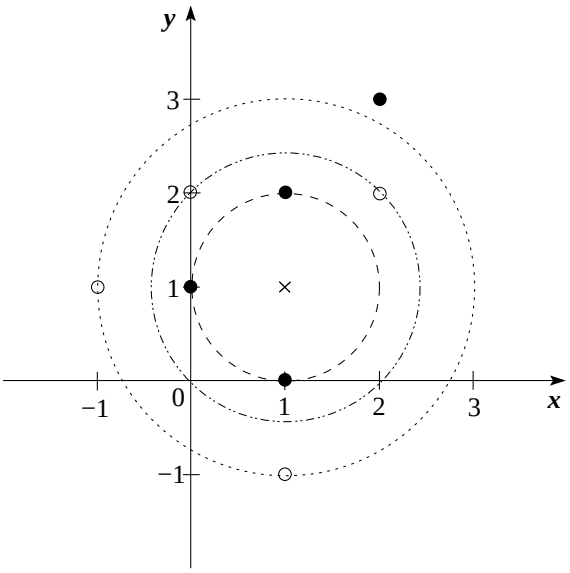
– clasificatorul 3-NN?

– clasificatorul 5-NN?

– clasificatorul 7-NN?

x	y	
−1	1	−
0	1	+
0	2	−
1	−1	−
1	0	+
1	2	+
2	2	−
2	3	+

Răspuns:



În figura alăturată am reprezentat:

– datele de antrenament, folosind cerculeţe albe pentru cele clasificate cu − şi cerculeţe negre pentru cele clasificate cu +;

– punctul care trebuie clasificat, marcat cu ×;

– vecinătăţile luate în considerare de către cei trei clasificatori precizaţi în enunţ; aceste vecinătăţi sunt reprezentate prin cele trei cercuri concentrice având centrul în punctul (1,1).

Analizând pe rând etichetele instanţelor din aceste trei vecinătăţi, ajungem la concluzia că rezultatele obţinute de cei trei clasificatori vor fi următoarele:

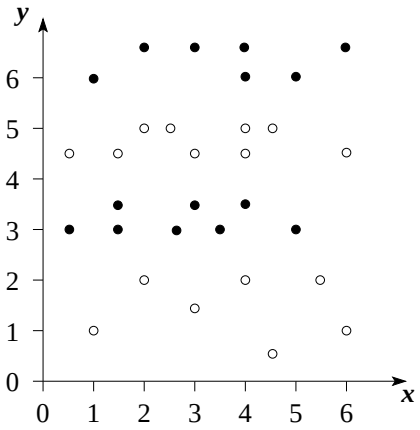
- **3-NN: +**
- **5-NN: +**
- **7-NN: −**

***Observaţie*: Acest exerciţiu pune în evidenţă faptul că la clasificarea cu algoritmul k-NN rezultatul poate diferi în funcţie de diversele valori ale lui k, adică în funcţie de cum se lărgeşte (sau se restrânge) vecinătatea punctului de test considerat.**

2. **(Algoritmul 1-NN: eroarea la antrenare)**

CMU, 2002 fall, Andrew Moore, final exam, pr. 6.e

Figura alăturată prezintă un set de date având două atribute de intrare x şi y, şi un atribut de ieşire, ale cărui valori sunt reprezentate prin culoarea punctului (alb sau negru).



Putem alege o *metrică* (adică, *măsură* sau *funcţie de distanţă*) astfel încât, folosind algoritmul de învăţare 1-NN (engl., [one] nearest neighbour), să obţinem eroare 0 la antrenare pe setul acesta de date?

Răspuns:

Fie $d : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ o metrică oarecare. Conform definiţiei matematice, d îndeplineşte următoarele *condiţii*:

$d(x, y) \geq 0, \forall x, y \in \mathbb{R}^2$	**(nenegativitatea)**
$d(x, y) = 0 \Leftrightarrow x = y$	**(identitatea indiscernabililor)**
$d(x, y) = d(y, x)$	**(simetria)**
$d(x, y) \leq d(x, z) + d(z, y), \forall x, y, z \in \mathbb{R}^2$	**(inegalitatea triunghiului)**

În particular, d poate fi metrica euclidiană, definită astfel:

$$d((x_1, y_1), (x_2, y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2},$$

pentru orice $(x_1, y_1), (x_2, y_2) \in \mathbb{R}^2$.

În general, când algoritmul 1-NN primeşte o instanţă de test (x, y), el caută în setul de antrenament punctul (x', y') cu proprietatea că distanţa $d((x, y), (x', y'))$ este minimă.

În cazul particular în care punctul (x, y) însuşi aparţine datelor de antrenament, urmează că $(x', y') = (x, y)$, datorită primelor două proprietăţi ale lui d enunţate mai sus.

Cum setul de date din enunţ este *consistent* — adică nu există nicio instanţă care să apară de două sau mai multe ori, însă cu etichete diferite —, vom avea de luat în considerare o singură valoare pentru calculul atributului de ieşire pentru punctul de test (x, y). Evident, algoritmul 1-NN o va folosi pe aceasta ca rezultat al clasificării. Concluzia acestui raţionament este că eroarea la antrenare produsă de algoritmul 1-NN pe acest set de date este 0.

Facem *observaţia* că nu doar pe acest set de date ci pe orice set de date de antrenament fără zgomote / inconsistenţe în ce priveşte etichetarea, algoritmul 1-NN va avea eroarea la antrenare 0, indiferent de metrica folosită (bineînţeles, dacă există o metrică în spaţiul instanţelor respective).

3. **(Algoritmul k-NN: calculul erorii la CVLOO pentru diferite valori ale lui k)**

CMU, 2003 fall, T. Mitchell, A. Moore, final exam, pr. 5.ab

X	Y
-0.1	$-$
0.7	$+$
1.0	$+$
1.6	$-$
2.0	$+$
2.5	$+$
3.2	$-$
3.5	$-$
4.1	$+$
4.9	$+$

Fie setul de date de antrenament din tabelul alăturat.
Vom folosi algoritmul k-NN cu distanţa euclidiană (neponderată) pentru a prezice valorile atributului de ieşire Y (de tip boolean) plecând de la atributul de intrare X, care ia valori reale.

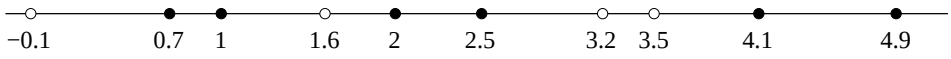
Care este eroarea produsă de algoritmul k-NN la cross-validare cu metoda "Leave-One-Out" în cazurile următoare:

a. $k = 1$.
b. $k = 3$.

Exprimaţi răspunsul sub forma numărului de instanţe clasificate eronat.

Răspuns:

Figura următoare vizualizează pe o axă datele de antrenament, precum şi clasificările acestora (○ pentru instanţe negative şi • pentru instanţe pozitive):



Comportamentul celor doi clasificatori la cross-validare prin metoda "Leave-One-Out" este cel explicat mai jos:

• Clasificatorul 1-NN:

Data	Eticheta	Vecinătate	Clasificare la CVLOO	Eroare?
-0.1	$-$	$\{0.7\}$	$+$	**da**
0.7	$+$	$\{1.0\}$	$+$	**nu**
1.0	$+$	$\{0.7\}$	$+$	**nu**
1.6	$-$	$\{2.0\}$	$+$	**da**
2.0	$+$	$\{1.6\}$	$-$	**da**
2.5	$+$	$\{2.0\}$	$+$	**nu**
3.2	$-$	$\{3.5\}$	$-$	**nu**
3.5	$-$	$\{3.2\}$	$-$	**nu**
4.1	$+$	$\{3.5\}$	$-$	**da**
4.9	$+$	$\{4.1\}$	$+$	**nu**

• Clasificatorul 3-NN:

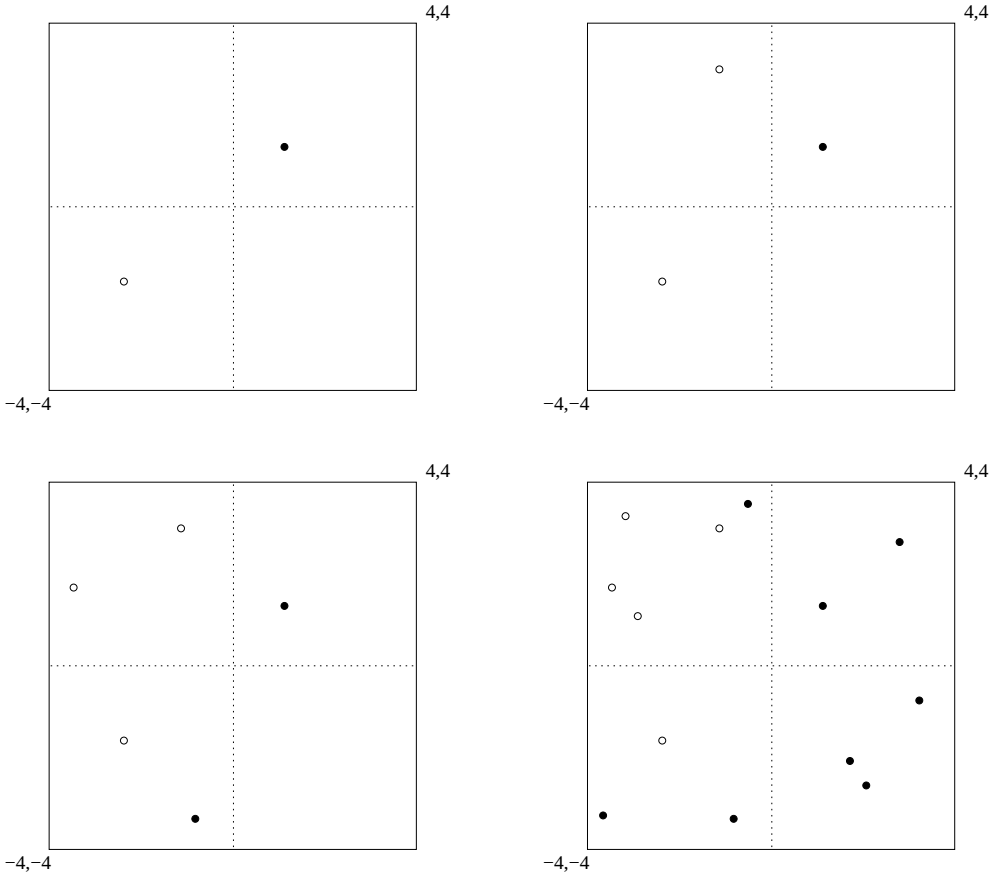
Data	Eticheta	Vecinătate	Clasificare la CVLOO	Eroare?
-0.1	$-$	$\{0.7; 1.0; 1.6\}$	$+$	**da**
0.7	$+$	$\{-0.1; 1.0; 1.6\}$	$-$	**da**
1.0	$+$	$\{0.7; 1.6; 2.0\}$	$+$	**nu**
1.6	$-$	$\{1.0; 2.0; 0.7/2.5\}$	$+$	**da**
2.0	$+$	$\{1.0; 1.6; 2.5\}$	$+$	**nu**
2.5	$+$	$\{1.6; 2.0; 3.2\}$	$-$	**da**
3.2	$-$	$\{2.5; 3.5; 4.1\}$	$+$	**da**
3.5	$-$	$\{2.5; 3.2; 4.1\}$	$+$	**da**
4.1	$+$	$\{3.2; 3.5; 4.9\}$	$-$	**da**
4.9	$+$	$\{3.2; 3.5; 4.1\}$	$-$	**da**

În concluzie, la cross-validare cu metoda "Leave-One-Out" avem 4 erori la clasificarea cu algoritmul 1-NN şi 8 erori la clasificarea cu algoritmul 3-NN. Putem concluziona că rezultatul algoritmului 3-NN este foarte afectat de către puternica „mixare" (adică, de frecventele schimbări de clasă, de la o instanţă oarecare la vecinii ei) din setul de antrenament.

4. **(Algoritmul 1-NN: graniţe / suprafeţe de decizie; diagrame Voronoi ca modalitate de învăţare rapidă / "eager")**

■ • ***CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh, HW1, pr. 3.1-2***

În fiecare din figurile următoare se dau câteva puncte în spaţiul euclidian bidimensional. Fiecare dintre aceste puncte este etichetat fie pozitiv (cerc plin) fie negativ (cerc simplu).



a. Presupunând că folosim ca metrică distanţa euclidiană, desenaţi suprafeţele de decizie corespunzătoare clasificatorului 1-NN, în fiecare din aceste patru cazuri.

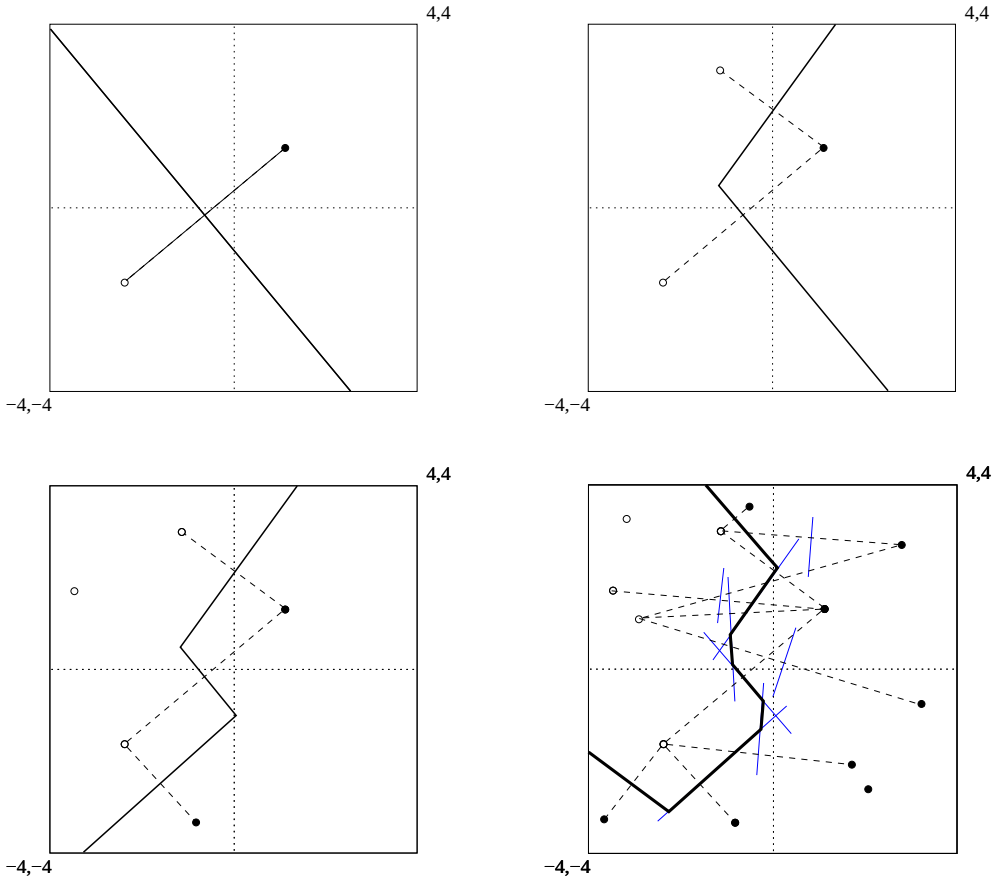
b. La curs am afirmat despre k-NN că este un clasificator lent (engl., lazy), care doar memorează toate instanţele de antrenament până ajunge la faza de test.

Totuşi, la punctul precedent am văzut că putem să trasăm suprafaţele de decizie pentru clasificatorul 1-NN, înainte de a intra în faza de test / generalizare. Atunci, în această fază, în loc să calculăm diverse distanţe şi apoi să determinăm care sunt cei mai apropiaţi vecini faţă de punctul de test dat (notat x_q), pur şi simplu i se va asigna lui x_q clasa / eticheta corespunzătoare suprafeţei de decizie în care se plasează.

Dacă am decide să memorăm toate aceste suprafeţe de decizie (ca linii poligonale) în loc să memorăm toate datele de antrenament, am obţine *întotdeauna* o îmbunătăţire în ceea ce priveşte consumul de memorie necesar pentru acest clasificator?

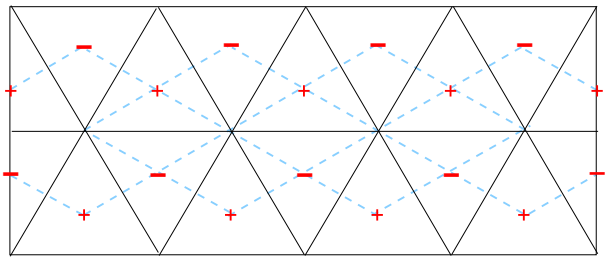
Răspuns:

a. Suprafeţele de decizie corespunzătoare clasificatorului 1-NN pentru cele patru cazuri sunt:



b. Răspunsul este negativ, adică: nu întotdeauna memorarea separatorului decizional este mai convenabilă decât memorarea datelor de antrenament. Vom justifica dând un exemplu, care reprezintă o situaţie-limită, şi anume, cazul care apare atunci când se creează tot atâtea suprafeţe de decizie câte instanţe sunt în setul de antrenament.

În figura alăturată, pentru cele n instanţe de antrenament avem nevoie să memorăm $3\frac{n}{2}$ puncte care determină triunghiurile – suprafeţele de decizie pentru clasa $+$. (Clasa $-$ va fi determinată prin excluziune / complementaritate.)

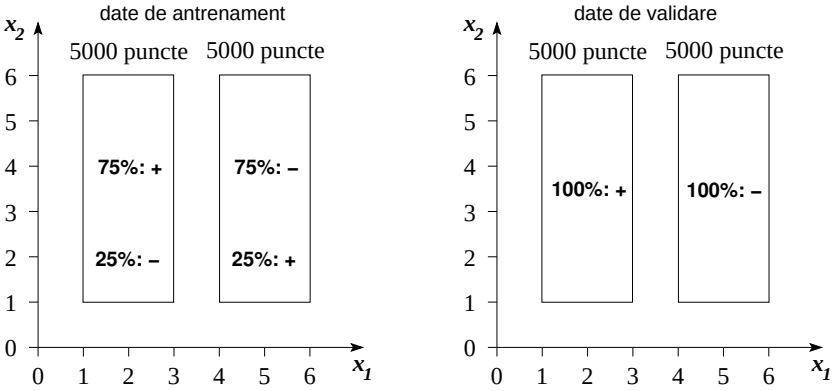


Se poate demonstra uşor următoarea proprietate: Chiar dacă am memora o singură dată cele $3\frac{n}{2}$ puncte — fiecare punct având câte 2 coordonate — care determină contururile suprafeţelor de decizie şi apoi am folosi indecşi pentru a indica ce puncte determină fiecare suprafaţă de decizie, consumul de memorie ar fi mai mare decât dacă am memora doar instanţele de antrenament.

5. **(Algoritmul k-NN: rata medie a erorii la antrenare, la CVLOO şi la testare în prezenţa unor „zgomote“ în datele de antrenament)**

• ○ ***CMU, 2002 fall, Andrew Moore, final exam, pr. 7***

Se constituie un set de date de antrenament şi un set de date de validare, alegând în mod uniform aleatoriu puncte situate în anumite regiuni dreptunghiulare, conform imaginii următoare:



Se observă că datele de antrenament sunt „bruiate“ (engl., noisy): în fiecare regiune, 25% din date provin din clasa adversă. Datele de validare nu sunt bruiate.

Pentru fiecare dintre întrebările următoare, încercuiţi fracţia care este *cea mai apropiată de* media / rata erorii în cazul respectiv. Justificaţi în mod riguros alegerea făcută.

a. Care este rata medie a erorii la antrenare pentru clasificatorul 1-NN ?

`0 1/8 1/4 3/8 1/3 1/2 5/8 2/3 3/4 7/8 1`

b. Care este rata medie a erorii la cross-validare cu metoda “Leave-One-Out” pentru clasificatorul 1-NN pe setul de antrenare?

0 1/8 1/4 3/8 1/3 1/2 5/8 2/3 3/4 7/8 1

c. Care este rata medie a erorii la testare pentru clasificatorul 1-NN pe setul de validare? (Antrenarea se face pe setul de antrenare.)

0 1/8 1/4 3/8 1/3 1/2 5/8 2/3 3/4 7/8 1

d. Care este rata medie a erorii la antrenare pentru clasificatorul 21-NN ?

0 1/8 1/4 3/8 1/3 1/2 5/8 2/3 3/4 7/8 1

e. Care este rata medie a erorii la cross-validare cu metoda "Leave-One-Out" pentru clasificatorul 21-NN pe setul de antrenare?

0 1/8 1/4 3/8 1/3 1/2 5/8 2/3 3/4 7/8 1

f. Care este rata medie a erorii la testare pentru clasificatorul 21-NN pe setul de validare? (Antrenarea se face pe setul de antrenare.)

0 1/8 1/4 3/8 1/3 1/2 5/8 2/3 3/4 7/8 1

Răspuns:

a. În acest caz, rata medie a erorii este 0**, întrucât orice punct din datele de antrenament este cel mai apropiat vecin în raport cu el însuşi, iar probabilitatea ca două puncte care au fost generate în mod aleatoriu (în oricare din cele două dreptunghiuri) să aibă exact aceleaşi coordonate şi aceeaşi etichetă este foarte mică (este practic** 0**).**

b. $3/8$**.**
Să analizăm cazul primului dreptunghi. Constituie eroare cazul când un exemplu pozitiv (care apare cu probabilitatea „aşteptată" de $3/4$**) este clasificat negativ (iar aceasta se întâmplă cu probabilitatea** $1/4$**) sau, invers, când un exemplu negativ (probabilitate:** $1/4$**) este clasificat pozitiv (probabilitate:** $3/4$**). Aşadar, probabilitatea de a clasifica greşit un exemplu din primul dreptunghi o aflăm astfel:** $3/4 \cdot 1/4 + 1/4 \cdot 3/4 = 3/8$**. Pentru dreptunghiul din dreapta raţionamentul este similar, iar selecţia exemplelor din cele două dreptunghiuri se face cu aceeaşi probabilitate (**$1/2$**), deci rezultatul final este** $3/8$**.**

c. $1/4$**.**
Probabilitatea de eroare este $1/4$ **pentru fiecare dreptunghi, fiindcă (în general) pentru un punct** (x_1, x_2) **selectat în mod aleatoriu din datele de validare din dreptunghiul respectiv aceasta (adică** $1/4$**) este probabilitatea ca vecinul cel mai apropiat de** (x_1, x_2) **în setul de date de antrenament să aibă semnul opus (faţă de semnul lui** (x_1, x_2)**). Aşadar, media erorii la testare cu clasificatorul 1-NN pe setul de validare este** $1/2 \cdot 1/4 + 1/2 \cdot 1/4 = 1/4$**.**

d. $1/4$**.**
Folosind algoritmul 21-NN, în dreptunghiul din stânga $3/4$ **din datele de antrenament sunt (în general) clasificate corect, iar restul de** $1/4$ **(şi anume, cele având semnul** $-$**) sunt clasificate eronat. Într-adevăr, la testarea unui punct oarecare** (x_1, x_2) **din acel dreptunghi, în 21-NN vecinătatea lui** (x_1, x_2)**, unul (cel mai apropiat vecin) este însuşi** (x_1, x_2)**, iar dintre ceilalţi 20 cei mai apropiaţi vecini** $3/4$ **sunt (în general) de semn** $+$**, iar** $1/4$ **de semn** $-$**. Analog se raţionează pentru dreptunghiul din dreapta. Deci rata medie a erorii la antrenare este:**

$$\frac{1}{2}\left(\frac{3}{4}\cdot 0 + \frac{1}{4}\cdot 1\right) + \frac{1}{2}\left(\frac{3}{4}\cdot 0 + \frac{1}{4}\cdot 1\right) = \frac{1}{4}.$$

e. $1/4$.
În cazul primului dreptunghi, apare eroare atunci când un exemplu pozitiv (probabilitate de selectare: $3/4$**) este clasificat negativ (probabilitate:** 0**, fiind vorba de 21-NN) sau când un exemplu negativ (probabilitate de selectare:** $1/4$**) este clasificat pozitiv (probabilitate:** 1**).**[352] **Aşadar, probabilitatea de a clasifica greşit un exemplu din primul dreptunghi este** $3/4 \cdot 0 + 1/4 \cdot 1 = 1/4$**. Cazul dreptunghiului din dreapta este similar, prin urmare rezultatul final este** $1/4$**.**

f. 0.
Datorită distribuţiei uniforme a datelor de antrenament, fiecare instanţă (x_1, x_2) **din setul de validare va avea majoritatea vecinilor — dintre cei mai apropiaţi, selectaţi de către algoritmul 21-NN din setul de date de antrenament — de acelaşi semn cu semnul lui** (x_1, x_2)**. (Şi anume, în medie de 3 ori mai mulţi vecini decât cei de semn contrar.) Prin urmare, fiecare punct din setul de validare va fi clasificat corect.**

***Observaţie* (1): Comparând rezultatele obţinute la punctele** a **şi** c **pe de o parte cu cele de la punctele** d **şi** f **(sau chiar** b **şi** e**) pe de altă parte, se observă că are loc o relaţie exact de acelaşi gen cu cea din definiţia fenomenului de *overfitting* (sau, *supra-specializare*):**[353] **erorile produse de algoritmii** 1**-NN şi** 21**-NN la antrenare sunt în relaţia** $0 < 1/4$ **dar în relaţie inversă (**$1/4 > 0$**) la testare (respectiv** $3/8 > 1/4$ **la cross-validare).**

***Observaţie* (2): Remarcaţi „sublinerea" din expresia „fracţia *cea mai apropiată de* media / rata erorii" din enunţ. Generarea aleatorie a datelor poate conduce la rezultate uşor diferite faţă de cele pe care le-am obţinut mai sus. În urma realizării unei implementări,**[354] **au fost obţinute următoarele rezultate de tip *eroare medie*, calculată în urma repetării de 100 de ori a generării datelor şi aplicării algoritmilor** k**-NN, conform cerinţelor din enunţ:**

a. 0, **b.** 0.374022, **c.** 0.250472, **d.** 0.249342, **e.** 0.253088, **f.** 0.006436.

Constatăm că se verifică „previziunile" noastre din rezolvarea de mai sus.

[352] În 21-NN vecinătatea punctului considerat (pentru testare), spre deosebire de cazul erorii la antrenare, la CVLOO nu se mai include punctul respectiv.

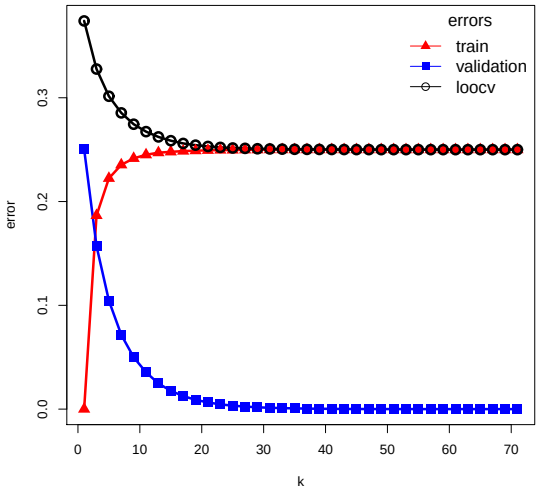
[353] Vă readucem aminte definiţia lui Tom Mitchell pentru fenomenul de *overfitting* (vedeţi cartea *Machine Learning*, pag. 67): două ipoteze h şi h' obţinute (eventual cu un acelaşi algoritm de clasificare automată) pe un set de date sunt în raport de overfitting dacă

$$error_{train}(h) < error_{train}(h'), \quad \text{dar} \quad error_{\mathcal{D}}(h) > error_{\mathcal{D}}(h'),$$

unde $\mathcal{D}$ este distribuţia reală a datelor.

[354] Implementarea a fost făcută de către studentul Sebastian Ciobanu de la Facultatea de Informatică a Universităţii „Al. I. Cuza" din Iaşi în semestrul I al anului universitar 2016-2017.

Evoluţia celor trei tipuri de eroare (la antrenare, la validare şi cross-validare cu metoda "leave-one-out"), pentru $k = 1, 3, \ldots, 69, 71$ este prezentată în figura alăturată. Se observă convergenţa la aceeaşi valoare (aprox., 0.25) pentru eroarea la antrenare şi eroarea CV-LOO (pe setul de date de antrenare) începând din jurul valorii $k = 31$, precum şi convergenţa la valori foarte apropiate de 0 pentru eroarea la testare pe setul de date de validare, începând din jurul valorii $k = 35$.



6. **(O versiune ipotetică pentru algoritmul k-NN: selectarea celor mai apropiaţi vecini nu pe baza calculării funcţiei de distanţă, ci folosind un oracol / "black box")**

CMU, 2009 spring, Ziv Bar-Joseph, final exam, pr. 2.2-3

Se încearcă clasificarea unor puncte în spaţiul euclidian bidimensional. Sunt date n instanţe P_1, P_2, ..., P_n, precum şi etichetările corespunzătoare c_1, c_2, ..., c_n, unde c_1, c_2, ..., c_n reprezintă valori dintr-o mulţime C. În schema de clasificare k-NN, fiecare element nou Q este clasificat cu eticheta majoritară obţinută în cadrul vecinătăţii formate din cei mai apropiaţi k vecini.

Să presupunem că măsura de distanţă nu este dată în mod explicit. În locul acesteia, aveţi la dispoziţie un "black box". Dacă se introduc instanţele $P_{i_1}, P_{i_2}, \ldots, P_{i_l}$ (unde l este un număr natural oarecare) şi un punct Q, black box-ul returnează cel mai apropiat vecin al lui Q, adică un element $P_{i_0} \in \{P_{i_1}, P_{i_2}, \ldots, P_{i_l}\}$, precum şi clasificarea corespunzătoare (c_{i_0}).

a. Este posibil să se construiască un algoritm de tip k-NN bazat doar pe acest black box? Dacă da, explicaţi cum anume, iar dacă nu, explicaţi de ce nu este posibil.

b. Dacă, în schimb, acel black box returnează cei mai apropiaţi j vecini (şi etichetele corespunzătoare), iar $j \neq k$, este posibil să se construiască un algoritm de tip k-NN bazat doar pe acest black box? Dacă da, explicaţi cum anume, iar dacă nu, explicaţi de ce nu este posibil.

Răspuns:

a. Da. Se foloseşte black box-ul dându-i ca intrare mai întâi mulţimea de exemple $P_1, P_2, \ldots, P_n$; se obţine cel mai apropiat vecin al lui Q şi eticheta sa. Apoi se scoate instanţa / punctul returnat din mulţimea de exemple. Se repetă acest procedeu de k ori, iar la final se va alege pentru Q eticheta corespunzătoare majorităţii din mulţimea de k vecini care au fost identificaţi de către black box.

b. Dacă $j < k$, atunci se foloseşte black box-ul de $[k/j]$ ori, obţinându-se $j*[k/j]$ cei mai apropiaţi vecini şi clasificările acestora. Notăm cu V_1 mulţimea formată din aceşti vecini. În cazul în care $k \neq j*[k/j]$, pentru a obţine restul de vecini necesari pentru k-NN, adică încă $k - j*[k/j]$ vecini, vom folosi black box-ul încă o dată. Vom nota cu V_2 mulţimea acestor noi j vecini. Dintre aceştia va trebui să alegem doar $k - j*[k/j]$ instanţe. Vom proceda astfel: considerăm o nouă mulţime de instanţe alcătuită din elementele lui V_2 şi $j - (k - j*[k/j])$ dintre toţi ceilalţi vecini obţinuţi anterior (V_1). Aplicăm black box-ul pe acest nou set de date şi vom obţine cei mai apropiaţi j vecini pentru punctul Q. Printre aceştia se vor afla cei $k - j*[k/j]$ vecini căutaţi.

Dacă $j > k$ iar black box-ul nu acceptă intrări duplicate, atunci el nu poate fi folosit pentru a determina cei mai apropiaţi k vecini ai instanţei de test.

Dar dacă $j > k$ şi black box-ul acceptă intrări duplicate, este posibil să rezolvăm problema, cel puţin în situaţia în care cei mai apropiaţi j vecini ai lui Q se află toţi la distanţe diferite faţă de acesta.[355] De exemplu, pentru $k = 1$ şi $j = 3$ vom proceda astfel:

***Pasul 1*: Aplicăm black box-ul inputului $P_1, \ldots, P_n$. Vom obţine outputul $P_{i_1}, P_{i_2}, P_{i_3}$.**

***Pasul 2*: Corespunzător fiecărui punct $P_{i_1}, P_{i_2}, P_{i_3}$, vom aplica pe rând black box-ul inputului**

- **$P_{i_1}, P_{i_1}, P_{i_2}, P_{i_3}$,**
- **$P_{i_1}, P_{i_2}, P_{i_2}, P_{i_3}$ şi respectiv**
- **$P_{i_1}, P_{i_2}, P_{i_3}, P_{i_3}$.**

Se poate constata că într-unul singur din aceste cazuri black box-ul returnează outputul $P_{i_1}, P_{i_2}, P_{i_3}$.[356] De pildă, dacă outputurile în aceste trei cazuri sunt

- **$P_{i_1}, P_{i_1}, P_{i_2}$,**
- **$P_{i_1}, P_{i_2}, P_{i_2}$,**
- **$P_{i_1}, P_{i_2}, P_{i_3}$,**

rezultă că P_{i_3} este cel mai distant dintre cei trei vecini ai lui Q.

***Pasul 3*: Considerând că lucrurile stau ca la finalul pasului precedent, pentru fiecare dintre punctele P_{i_1}, P_{i_2} vom aplica black box-ului următorul input:**

- **$P_{i_1}, P_{i_1}, P_{i_1}, P_{i_2}$ şi respectiv**
- **$P_{i_1}, P_{i_2}, P_{i_2}, P_{i_2}$.**

Dacă vom obţine outputul $P_{i_1}, P_{i_1}, P_{i_1}$, va rezulta că P_{i_1} este mai apropiat de Q decât punctul P_{i_2}. Invers, dacă obţinem outputul $P_{i_2}, P_{i_2}, P_{i_2}$, va rezulta că P_{i_2} este mai apropiat de Q decât punctul P_{i_1}.

[355] Rămâne de analizat varianta contrară.

[356] Am presupus că black box-ul returnează exact j instanţe, chiar dacă între $P_1, \ldots, P_n$ există mai multe instanţe egal depărtate faţă de punctul Q, în speţă mai multe instanţe situate la maximumul distanţelor dintre fiecare din cele j pe de o parte şi punctul Q pe de altă parte.

7. **(Algoritmul 1-NN: suprafeţele de decizie [şi separatorii decizionali] depind de măsurile de distanţă folosite)**

CMU, 2008 fall, Eric Xing, HW1, pr. 3.1.2

Se dau două puncte din spaţiul euclidian bidimensional: punctul $(-1, 0)$ clasificat negativ şi punctul $(1, 0)$ clasificat pozitiv.

Clasificatorul 1-NN care foloseşte distanţa euclidiană şi are ca set de date de antrenament cele două puncte de mai sus are următoarea *formă analitică* (uşor de dedus):

- **dat fiind un punct arbitrar (x, y),**
  în cazul în care $x > 0$, eticheta asignată punctului respectiv este $+$, iar în cazul $x < 0$ eticheta asignată este $-$;
- **dreapta $x = 0$ este *graniţa de decizie* (engl., decision boundary) corespunzătoare acestui clasificator.**

a. Care va fi forma analitică a clasificatorului 1-NN dacă în locul distanţei euclidiene (indusă de norma L_2) se foloseşte distanţa Manhattan (indusă de norma L_1)? Vă reamintim că distanţa Manhattan dintre două puncte (x_1, y_1) şi (x_2, y_2) este $|x_1 - x_2| + |y_1 - y_2|$.

b. Dar dacă se foloseşte distanţa [indusă de norma] L_∞, definită în $\mathbb{R}^2$ prin

$$d((x_1, y_1), (x_2, y_2)) = \max\{|x_1 - x_2|, |y_1 - y_2|\} \ ?$$

Răspuns:

a. Putem exprima distanţa Manhattan dintre un punct oarecare (x, y) din plan şi de fiecare dintre cele două puncte date în enunţ — punctul $(1, 0)$ clasificat pozitiv şi $punctul(-1, 0)$ clasificat negativ — astfel:

$$\begin{aligned} d_+ &\overset{not.}{=} d((x, y), (1, 0)) = |x - 1| + |y| \\ d_- &\overset{not.}{=} d((x, y), (-1, 0)) = |x + 1| + |y| \end{aligned}$$

În consecinţă, a stabili care dintre cele două distanţe (d_+ şi d_-) este mai mică revine la a compara (doar) expresiile $|x - 1|$ şi $|x + 1|$.

Avem următoarele cazuri:

- **dacă $x > 1$, atunci**

$$\left.\begin{array}{l} d_+ = x - 1 + |y| \\ d_- = x + 1 + |y| \end{array}\right\} \Rightarrow d_+ < d_- \Rightarrow \textbf{ punctul } (x, y) \textbf{ va fi clasificat } +$$

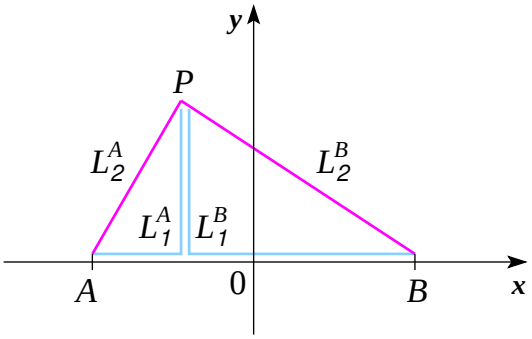
- **dacă $x < -1$:**

$$\left.\begin{array}{l} d_+ = -x + 1 + |y| \\ d_- = -x - 1 + |y| \end{array}\right\} \Rightarrow d_+ > d_- \Rightarrow \textbf{ punctul } (x, y) \textbf{ va fi clasificat } -$$

- **dacă $-1 \leq x \leq 1$:**

$$\left.\begin{array}{l} d_+ = -x + 1 + |y| \\ d_- = x + 1 + |y| \end{array}\right\} \Rightarrow \left\{\begin{array}{l} d_+ < d_- \textbf{ pentru } x > 0 \textbf{ deci } (x, y) \textbf{ va fi clasificat } + \\ d_+ > d_- \textbf{ pentru } x < 0 \textbf{ deci } (x, y) \textbf{ va fi clasificat } - \\ d_+ = d_- \textbf{ dacă şi numai dacă } x = 0. \end{array}\right.$$

Rezultă că graniţa de decizie a acestui clasificator este dreapta $x = 0$. Chiar mai mult: forma analitică a clasificatorului 1-NN care foloseşte distanţa Manhattan pe setul de date din enunţ este aceeaşi cu a clasificatorului 1-NN care foloseşte distanţa euclidiană.



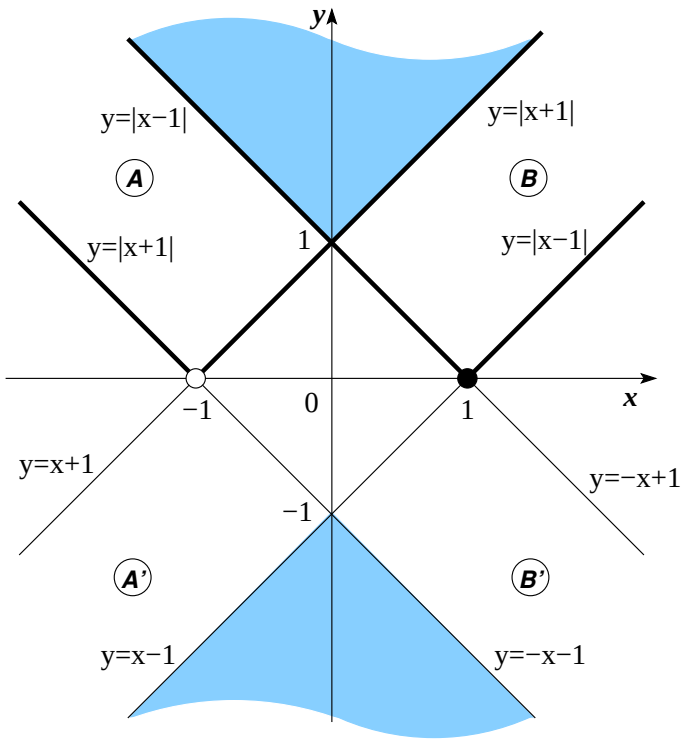
***Observaţie*: La concluzia de mai sus se putea ajunge mult mai uşor ţinând cont de următoarea *proprietate* (demonstrabilă imediat pe cale geometrică): pentru orice două puncte A şi B din $\mathbb{R}^2$ au loc următoarele relaţii de echivalenţă:**

$$L_1(P,A) < L_1(P,B) \Leftrightarrow L_2(P,A) < L_2(P,B) \text{ pentru } \forall P \in \mathbb{R}^2 \text{ şi}$$
$$L_1(P,A) = L_1(P,B) \Leftrightarrow L_2(P,A) = L_2(P,B) \text{ pentru } \forall P \in \mathbb{R}^2.$$

b. Dacă se foloseşte distanţa L_∞, atunci conform definiţiei acestei metrici vom avea:

$$\begin{aligned} d_+ &= d((x,y),(1,0)) \\ &= \max\{|x-1|,|y|\} \\ d_- &= d((x,y),(-1,0)) \\ &= \max\{|x+1|,|y|\}. \end{aligned}$$

Mai întâi vom trasa graficele funcţiilor $y = |x-1|$ şi $y = |x+1|$, şi vom obţine rezultatul din figura alăturată.



***Cazul i*: Se observă că $|y| \geq |x-1|$ şi $|y| \geq |x+1|$ pentru toate punctele (x,y) situate în zonele unghiulare haşurate. În consecinţă, d_+ va fi egal cu d_- pentru orice punct din aceste zone. Aşadar, zonele haşurate vor aparţine graniţei / suprafeţei de decizie a clasificatorului nostru.**

***Cazul ii*: Considerăm acum zonele identificate prin literele A, A', B şi B' în figura de mai sus.[357] Se observă uşor că pentru zonele A şi A' avem $d_+ =$**

[357] Analitic, zona A este definită de punctele (x,y) care satisfac inecuaţiile $|x+1| < |y| < |x-1|$ şi $y > 0$; zona A': $|x+1| < |y| < |x-1|$ şi $y < 0$; zona B: $|x-1| < |y| < |x+1|$ şi $y > 0$; zona B': $|x-1| < |y| < |x+1|$ şi $y < 0$.

$\max\{|x-1|,|y|\} = |x-1|$ **şi** $d_- = \max\{|x+1|,|y|\} = |y|$**, iar pentru zonele** B **şi** B' **avem** $d_+ = \max\{|x-1|,|y|\} = |y|$ **şi** $d_- = \max\{|x+1|,|y|\} = |x+1|$**. În consecinţă,** $d_- = |y| < |x-1| = d_+$ **pentru zonele** A **şi** A'**, iar** $d_+ = |y| < |x+1| = d_-$ **pentru zonele** B **şi** B'**. Aşadar, zonele** A **şi** A' **vor fi clasificate negativ, iar zonele** B **şi** B' **vor fi clasificate pozitiv.**

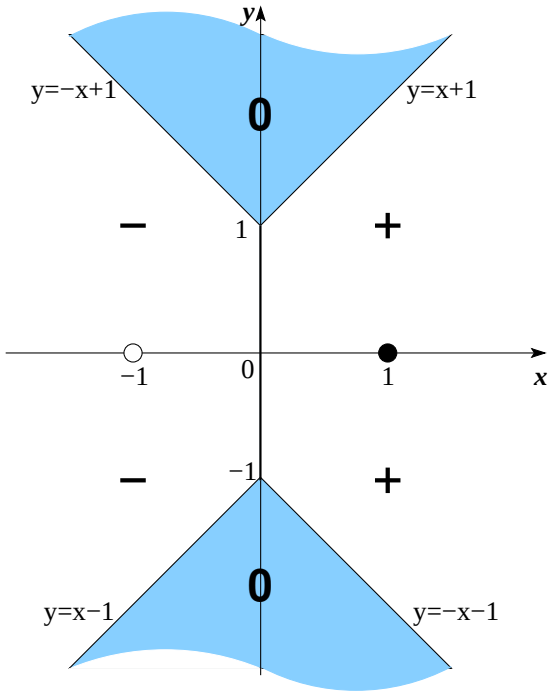
***Cazul iii*: Pentru toate celelalte zone rămase în discuţie — adică pentru orice punct** (x,y) **situat în afara celor două zone haşurate şi în afara zonelor** A, A', B, B' **—, vom avea** $d_+ = \max\{|x-1|,|y|\} = |\,x-1\,|$ **şi** $d_- = \max\{|x+1|,|y|\} = |x+1|$**. Din grafic se observă că** $|x+1| > |x-1|$ **pentru** $x > 0$ **şi** $|x-1| > |x+1|$ **pentru** $x < 0$**, iar** $|x+1| = |x-1|$ **pentru** $x = 0$**. Aşadar, sumarizând, în zone avem:** $d_+ = \min\{d_+, d_-\}$ **pentru** $x > 0$ **şi** $d_- = \min\{d_+, d_-\}$ **pentru** $x < 0$**, iar** $d_+, = d_-$ **pentru** $x = 0$**.**

Concluzionând, forma analitică a clasificatorului 1-NN care foloseşte distanţa L_∞ **este următoarea:**

- (x,y) **va fi etichetat cu** $+$ **dacă** $x > 0$ **şi** $-x-1 < y < x+1$**;**
- (x,y) **va fi etichetat cu** $-$ **dacă** $x < 0$ **şi** $x-1 < y < -x+1$**;**
- **în rest este vorba de suprafaţa de separare, adică locul geometric al punctelor** (x,y) **pentru care distanţa faţă de cele două puncte din enunţ este egală.**

Reprezentarea grafică a suprafeţelor de decizie este dată în figura alăturată.

Este de remarcat faptul că pentru acest clasificator graniţa / suprafaţa de decizie nu este formată doar din drepte, ci este reuniunea unei drepte (axa Oy**)cu două intersecţii de (câte două) semiplane.**



8. **(Algoritmul / metoda lui Shepard – aplicare)**

CMU, 2002 fall, Andrew Moore, final exam, pr. 6.a-d

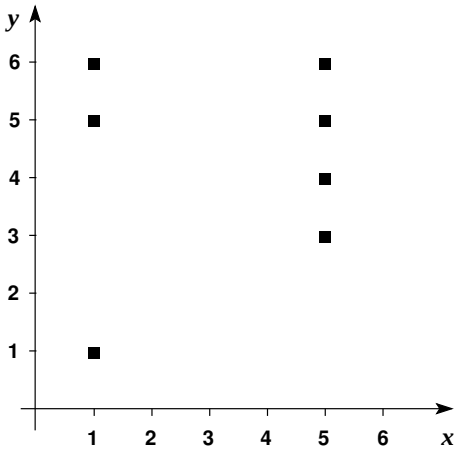
Figura de mai jos prezintă un set de date de antrenament cu un atribut de intrare $x \in \mathbb{R}$ **şi un atribut de ieşire** $y \in \mathbb{R}$**.**

Vom estima din aceste date câteva valori ale unei funcţii continue $f : \mathbb{R} \to \mathbb{R}$, folosind *metoda lui Shepard*. Aceasta este o variantă (de tip *regresie*) a algoritmului k-NN în care se iau în considerare toate punctele de antrenament, dar se aplică ponderi în funcţie de distanţă:

$$\hat{f}(x) \leftarrow \frac{\sum_i w(x, x_i) f(x_i)}{\sum_i w(x, x_i)}$$

Se va considera

$$w(x, x_i) = \begin{cases} 1, & \textbf{dacă } \mid x - x_i \mid \leq 3 \\ 0, & \textbf{în rest.} \end{cases}$$



Care va fi valoarea prezisă pentru funcţia f pentru

a. $x = 1$**?** **c.** $x = 5$**?**

b. $x = 3$**?** **d.** $x = 6$**?**

Răspuns:

Din modul cum au fost definite ponderile w ajungem la concluzia că valoarea lui f pentru un punct oarecare x va fi calculată ca medie aritmetică a valorilor / componentelor y din acele date de antrenament pentru care abscisa (x') este situată la distanţă de cel mult 3 unităţi de punctul x care ne interesează.

a. Avem $|1-1| = 0 \leq 3$ şi $|1-5| = 4 > 3$, prin urmare vor fi luate în considerare doar valorile învăţate pentru $x = 1$.

$\hat{f}(1) = \dfrac{1+5+6}{3} = \dfrac{12}{3} = 4.$

b. Avem $|3-1| = 2 \leq 3$ şi $|3-5| = 2 \leq 3$, prin urmare vor fi luate în considerare şi valorile învăţate pentru $x = 1$ şi cele pentru $x = 5$.

$\hat{f}(3) = \dfrac{1+5+6+3+4+5+6}{7} = \dfrac{30}{7}.$

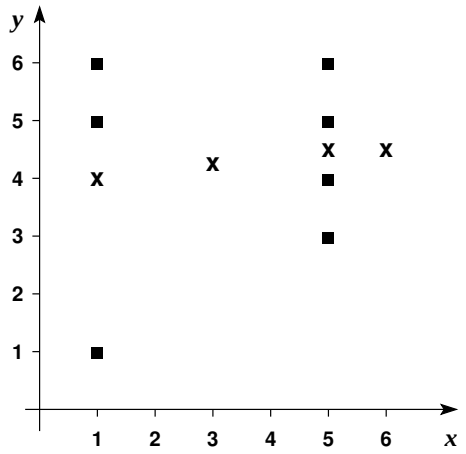
c. Avem $|5-1| = 4 > 3$ şi $|5-5| = 0 \leq 3$, prin urmare vor fi luate în considerare doar valorile învăţate pentru $x = 5$.

$\hat{f}(5) = \dfrac{3+4+5+6}{4} = \dfrac{18}{4} = \dfrac{9}{2}.$

d. Avem $|6-1| = 5 > 3$ şi $|6-5| = 1 \leq 3$, prin urmare vor fi luate în considerare doar valorile învăţate pentru $x = 5$, ca şi în cazul precedent.

$\hat{f}(6) = \hat{f}(5) = 4.5$

Dacă vom plasa rezultatele de mai sus pe grafic şi vom reprezenta punctele $(x, \hat{f}(x))$ sub forma unor cruciuliţe, vom obţine figura alăturată.



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

■ □ • ***CMU, 2010 fall, Aarti Singh, HW2, pr. 2.2***

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[358] 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.**

a. În cazul particular $p = 1$, calculaţi expresia *funcţiei de distribuţie cumulativă*[359] 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 $\mathbb{R}^p$:**

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

[358] Termenul folosit în limba engleză pentru o astel de sferă este *unit ball*.
[359] Engl., cummulative distribution function, c.d.f.

unde Γ reprezintă funcţia Gamma a lui Euler,[360] **care are proprietăţile:**

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi},\ \Gamma(1) = 1, \textbf{ iar } \Gamma(x+1) = x\Gamma(x) \textbf{ pentru } x > 0.$$[361]

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*:**

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

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?

Răspuns:

a. Pentru $p = 1$, sfera de rază 1 este intervalul $[-1, 1]$, iar funcţia de distribuţie cumulativă va avea expresia:

$$F_{n,1}(t) \overset{not.}{=} P(d^* \leq t) = 1 - P(d^* > t) = 1 - P(||x_i|| > t,\ i = 1, 2, \ldots, n)$$

Ţinem cont de presupoziţia de independenţă la generarea punctelor x_i, rezultă:

$$F_{n,1}(t) = 1 - \prod_{i=1}^{n} P(||x_i|| > t) = 1 - (1-t)^n.$$

[360] Vedeţi problema 28.b de la capitolul de *Fundamente*.

[361] Se verifică uşor că pentru $p = 3$ se obţine volumul sferei: $V_3(r) = \frac{(r\sqrt{\pi})^3}{\frac{3}{4}\sqrt{\pi}} = \frac{4\pi r^3}{3}$. Pentru demonstrarea unora dintre proprietăţile funcţiei Γ indicate mai sus, vedeţi problema 41 de la capitolul de *Fundamente*.

b. În cazul general, adică pentru p un număr natural oarecare nenul, fixat, vom exprima, mai întâi $P(d^* \leq t)$ exact ca mai înainte:

$$F_{n,p}(t) \stackrel{not.}{=} P(d^* \leq t) = 1 - P(d^* > t) = 1 - P(||x_i|| > t,\ i = 1, 2, \ldots, n)$$
$$\stackrel{indep.\ cdt.}{=} 1 - \prod_{i=1}^{n} P(||x_i|| > t).$$

Apoi, ţinând cont de presupoziţia de uniformitate la generarea punctelor x_i şi, folosind notaţia $V_p(t)$ pentru volumul sferei de rază t, obţinem:

$$F_{n,p}(t) = 1 - \left(\frac{V_p(1) - V_p(t)}{V_p(1)}\right)^n = 1 - \left(1 - \frac{V_p(t)}{V_p(1)}\right)^n.$$

În sfârşit, folosind formula *sugerată* în enunţ pentru V_p, rezultă imediat că $F_{n,p}(t) = 1 - (1 - t^p)^n$.

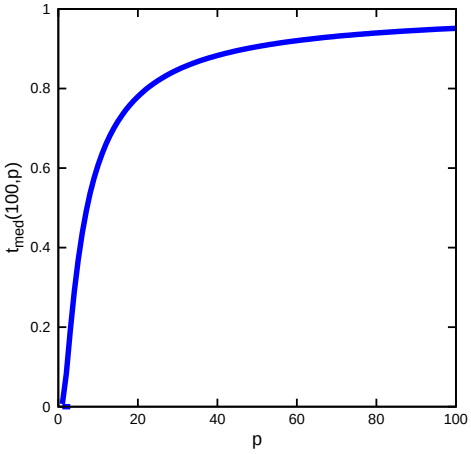
c. Pentru a afla valoarea mediană corespunzătoare variabilei aleatoare d^*, vom rezolva ecuaţia $P(d^* \leq t) = 1/2$ în funcţie de t:

$$P(d^* \leq t) = \frac{1}{2} \Leftrightarrow F_{n,p}(t) = \frac{1}{2} \stackrel{b.}{\Leftrightarrow} 1 - (1 - t^p)^n = \frac{1}{2}$$
$$\Leftrightarrow \quad (1 - t^p)^n = \frac{1}{2} \Leftrightarrow 1 - t^p = \frac{1}{2^{1/n}}$$
$$\Leftrightarrow \quad t^p = 1 - \frac{1}{2^{1/n}}$$

Prin urmare,

$$t_{med}(n, p) = \left(1 - \frac{1}{2^{1/n}}\right)^{1/p}.$$

Graficul funcţiei $t_{med}(100, p)$ pentru $p = 1, 2, \ldots, 100$ este cel din figura alăturată. Se observă că sfera minimală care conţine cel mai apropiat vecin (un x_i, cu $i \in \{1, 2, \ldots, n\}$) al originii O se lărgeşte foarte repede pe măsură ce p creşte. Pentru valori ale lui p mai mari decât 10, majoritatea dintre cele 100 de instanţe de antrenament sunt mai aproape de conturul sferei de rază 1 decât de originea O.



d. Putem scrie următorul şir de echivalenţe:

$$P(d^* \leq 0.5) \geq 0.9 \quad \Leftrightarrow \quad F_{n,p}(0.5) \geq 0.9 \Leftrightarrow$$
$$\stackrel{b.}{\Leftrightarrow} \quad 1 - \left(1 - \frac{1}{2^p}\right)^n \geq \frac{9}{10} \ \Leftrightarrow \ \left(1 - \frac{1}{2^p}\right)^n \leq \frac{1}{10}$$
$$\Leftrightarrow \quad n \cdot \ln\left(1 - \frac{1}{2^p}\right) \leq -\ln 10$$

$$\Leftrightarrow \quad n \geq \frac{\ln 10}{-\ln\left(1-\frac{1}{2^p}\right)}$$

Se poate vedea imediat că membrul din partea dreaptă a inegalităţii de mai sus tinde la $+\infty$ pentru $p \to \infty$. Este necesar să vedem *cât de repede* are loc această tindere la infinit. Pentru aceasta, vom folosi descompunerea lui $-\ln(1-1/2^p)$ sub forma unei serii Taylor (luând $x = 1/2^p$):[362]

$$P(d^* \leq 0.5) \geq 0.9 \quad \Rightarrow \quad n \geq (\ln 10)\, 2^p \frac{1}{1+\frac{1}{2}\cdot\frac{1}{2^p}+\frac{1}{3}\cdot\frac{1}{2^{2p}}+\ldots+\frac{1}{n}\frac{1}{2^{(n-1)p}}+\ldots}$$

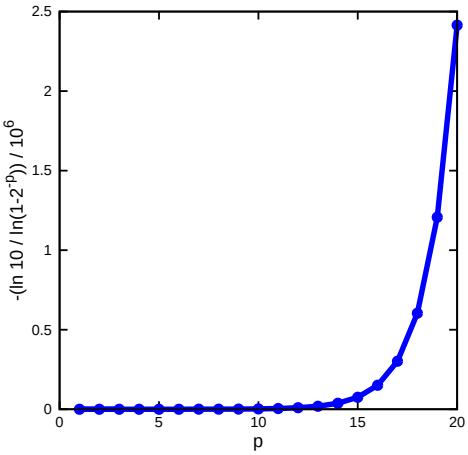
$$\Rightarrow \quad n > \frac{4}{3}2^{p-1}\ \ln 10.$$

Pentru obţinerea ultimei inegalităţi de mai sus am ţinut cont de faptul că inegalitatea $\frac{1}{n\cdot 2^{(n-1)p}} \leq \frac{1}{2^n} \Leftrightarrow 2^n \leq n\cdot 2^{(n-1)p}$ are loc pentru orice $p \geq 1$, şi $n \geq 2$,[363] deci

$$\begin{aligned} & 1+\frac{1}{2}\cdot\frac{1}{2^p}+\frac{1}{3}\cdot\frac{1}{2^{2p}}+\ldots+\frac{1}{n}\cdot\frac{1}{2^{(n-1)p}}+\ldots \\ \leq\ & 1+\frac{1}{2^2}+\frac{1}{2^3}+\ldots+\frac{1}{2^n}+\ldots \\ <\ & \left[1+\frac{1}{2}+\frac{1}{2^2}+\frac{1}{2^3}+\ldots+\frac{1}{2^n}+\ldots\right]-\frac{1}{2} \\ & \to \frac{1}{1-\frac{1}{2}}-\frac{1}{2}=2-\frac{1}{2}=\frac{3}{2}. \end{aligned}$$

Observaţi că limita de mai sus este este o limită *superioară*, de aceea inegalitatea se păstrează până la final.

Aşadar, rezultă că n creşte în mod exponenţial în raport cu p.[364] **Graficul pentru marginea inferioară dedusă mai sus ($-\ln 10/\ln(1-2^{-p})$) este cel din figura alăturată.**



Se observă că, într-adevăr, creşterea acestei margini inferioare (deci şi a lui n) este exponenţială.

[362] $-ln(1-x) = x + \frac{1}{2}x^2 + \frac{1}{3}x^3 + \ldots + \frac{1}{n}x^n + \ldots$ pentru orice $x \in (-1,+1)$.

[363] Demonstraţia se poate face prin inducţie după valorile lui p.

[364] Mai detaliat: n, numărul de instanţe de antrenament necesare pentru a ne asigura că d^* (distanţa până la cel mai apropiat vecin al originii O) este cu o probabilitate mare (şi anume, $9/10$) mai mică decât 0.5 creşte în mod exponenţial în raport cu p.

e. Conform intuiţiei, clasificatorul k-NN se comportă bine atunci când instanţa de test x_q este situată într-o vecinătate densă de instanţe de antrenament. Totuşi, analiza teoretică de mai sus ne arată că pentru a ne asigura că punctul x_q are o vecinătate densă, numărul tuturor instanţelor de antrenament trebuie să crească exponenţial în raport cu p, ceea ce nu este fezabil pentru valori mari ale lui p. (Parametrul p este dimensiunea spaţiului în care se lucrează, adică numărul de trăsături ale instanţelor de antrenament şi, respectiv, de test).

În consecinţă, pentru aplicaţiile practice în care se folosesc date cu multe atribute, este recomandat ca execuţia algoritmului k-NN să fie precedată de efectuarea unei „selecţii de trăsături“ (engl., feature selection).

10. (Algoritmul 1-NN [comparativ cu clasificatorul Bayes Optimal]: o margine superioară pentru *eroarea medie asimptotică* [la antrenare])

■ □ • ***CMU, 2005 spring, C. Guestrin, T. Mitchell, HW3, pr. 1***

Un rezultat interesant obţinut de Cover şi Hart (1967) arată că, atunci când numărul datelor de antrenament tinde la infinit, iar datele de antrenament umplu spaţiul în mod dens, rata medie a erorii produsă de către clasificatorul 1-NN este mărginită superior de dublul ratei medii a erorii pentru clasificatorul Bayes Comun (engl., Joint Bayes), (care este numit adeseori şi *Bayes Optimal* (engl., Optimal Bayes).

La acest exerciţiu vi se va arăta, pas cu pas, cum se demonstrează rezultatul lui Cover şi Hart în cazul particular al clasificării binare. Aşadar, fie $x_1, x_2, \ldots$ instanţele de antrenament, iar $y_1, y_2, \ldots$ etichetele corespunzătoare, cu $y_i \in \{0, 1\}$. Putem considera instanţele x_i ca fiind puncte într-un spaţiu euclidian d-dimensional.

Notăm $p_y(x) = P(X = x \mid Y = y)$ probabilitatea condiţionată care reprezintă distribuţia instanţelor din clasa y. Vom presupune că aceste probabilităţi condiţionate sunt continue în raport cu variabila x şi că $p_y(x) \in (0, 1)$ pentru orice x şi orice y. Notăm cu θ probabilitatea ca un exemplu de antrenament selectat în mod aleatoriu să fie din clasa 1, aşadar $\theta \stackrel{not.}{=} P(Y = 1)$. Din nou, presupunem că $\theta \in (0, 1)$.

a. Calculaţi probabilitatea ca o instanţă oarecare x să aparţină clasei 1: $q(x) \stackrel{not.}{=} P(Y = 1 \mid X = x)$. Exprimaţi $q(x)$ în funcţie de $p_0(x), p_1(x)$ şi θ.

b. Clasificatorul Bayes Optimal asignează unui punct dat x cea mai probabilă clasă, $\text{argmax}_y P(Y = y \mid X = x)$. (Aceasta implică faptul că algoritmul Bayes Optimal maximizează probabilitatea clasificării corecte a tuturor datelor.) Considerând o instanţă oarecare x, calculaţi probabilitatea ca x să fie clasificat greşit folosind clasificatorul Bayes Optimal, în funcţie de probabilitatea $q(x) \stackrel{not.}{=} P(Y = 1|X = x)$ care tocmai a fost calculată la punctul precedent. Veţi desemna această nouă probabilitate cu $Error_{Bayes}(x)$.

c. Acum considerăm clasificatorul 1-NN. Acesta îi asignează unei instanţe oarecare de test x eticheta celei mai apropiate instanţe de antrenament x'. Dată fiind o instanţă de antrenament x (aleasă în mod arbitrar, dar fixată), calculaţi *eroarea „aşteptată“* (engl., expected error) produsă de către clasificatorul 1-NN, adică probabilitatea ca instanţa x să fie clasificată greşit. Notaţi această

probabilitate cu $Error_{1\text{-}NN}(x)$ şi exprimaţi-o sub forma unei funcţii definită în raport cu probabilităţile $q(x)$ şi $q(x')$.

d. În *cazul asimptotic*, numărul de exemple de antrenament al fiecărei clase tinde la infinit, iar datele de antrenament umplu spaţiul în mod dens. Atunci $q(x') \to q(x)$, unde, ca şi mai sus, x' este cel mai apropiat vecin al lui x.[365] Făcând această substituţie în rezultatul obţinut la punctul anterior, deduceţi expresia *erorii asimptotice* pentru clasificatorul 1-NN în punctul x, adică $\lim_{x' \to x} Error_{1\text{-}NN}(x)$, în funcţie de probabilitatea $q(x)$.

e. Arătaţi că eroarea asimptotică obţinută la punctul d este mai mică decât dublul erorii clasificatorului Bayes Optimal obţinută la punctul b, adică:

$$\lim_{x' \to x} Error_{1\text{-}NN}(x) \leq 2 Error_{Bayes}(x).$$

În final, din această inegalitate deduceţi relaţia corespunzătoare între ratele medii ale erorilor:[366]

$$E[\lim_{n \to \infty} Error_{1\text{-}NN}] \leq 2E[Error_{Bayes}].$$

Răspuns:

a. Conform enunţului, $p_1(x) \stackrel{not.}{=} P(X = x|Y = 1)$, $p_0(x) \stackrel{not.}{=} P(X = x|Y = 0)$, $q(x) \stackrel{not.}{=} P(Y = 1|X = x)$ **şi** $\theta \stackrel{not.}{=} P(Y = 1)$**. Putem calcula probabilitatea $q(x)$ în funcţie de $p_1(x)$, $p_0(x)$ şi θ folosind formula lui Bayes:**

$$\begin{aligned} q(x) &\stackrel{Bayes}{=} \frac{P(X = x|Y = 1)P(Y = 1)}{P(X = x)} \\ &= \frac{P(X = x|Y = 1)P(Y = 1)}{P(X = x|Y = 1)P(Y = 1) + P(X = x|Y = 0)P(Y = 0)} \\ &= \frac{p_1(x)\ \theta}{p_1(x)\ \theta + p_0(x)(1 - \theta)} \end{aligned}$$

b. Este imediat faptul următor: probabilitatea ca algoritmul Bayes Optimal să greşească este $P(Y = 0|X = x)$ în cazul în care $P(Y = 1|X = x) \geq P(Y = 0|X = x)$, respectiv $P(Y = 1|X = x)$ atunci când $P(Y = 0|X = x) \geq P(Y = 1|X = x)$.[367] Altfel spus,

$$\begin{aligned} Error_{Bayes}(x) &= \min\{P(Y = 0|X = x), P(Y = 1|X = x)\} \\ &= \min\{1 - q(x), q(x)\} = \begin{cases} q(x) \text{ \textbf{în cazul} } q(x) \in [0\,,\, 1/2] \\ 1 - q(x) \text{ \textbf{în cazul} } q(x) \in (1/2\,,\, 1]. \end{cases} \end{aligned}$$

c. Algoritmul 1-NN greşeşte atunci când instanţa de antrenament x are eticheta 1 iar x', cel mai apropiat vecin al lui x, are eticheta 0, sau invers, adică

[365] Adică, $P(Y = 1|X = x') \to P(Y = 1|X = x)$. Aceasta se justifică ţinând cont de continuitatea lui $p_y(x) \stackrel{not.}{=} P(X = x|Y = y)$ care a fost asumată în enunţ şi, de asemenea, de rezultatul obţinut la punctul a.

[366] Cititorul atent va remarca faptul că în expresia de mai jos ($E[\lim_{n \to \infty} Error_{1\text{-}NN}]$) s-a înlocuit $\lim_{x \to x'}$ (folosită anterior) cu $\lim_{n \to \infty}$, pentru că se face trecerea la medii. Când $n \to \infty$, conform presupoziţiilor din enunţ, rezultă $x \to x'$ pentru orice x.

[367] Am ţinut cont de proprietatea (evidentă) $P(X = x, Y = y) = P(X = x|Y = y)\, P(Y = y) = P(Y = y|X = x)\, P(X = x)$, care are loc pentru orice x şi y.

atunci când x are eticheta 0 iar x' are eticheta 1. În consecinţă, folosind algoritmul 1-NN, eroarea „aşteptată“ la clasificarea lui x este:

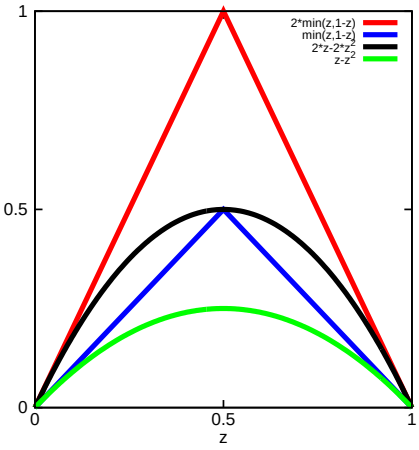
$$\begin{aligned} Error_{1\text{-}NN}(x) &= P(Y=1|X=x)P(Y=0|X=x') + \\ &\quad P(Y=0|X=x)P(Y=1|X=x') \\ &= q(x)(1-q(x')) + (1-q(x))q(x'). \end{aligned}$$

d. Este imediat că $\lim_{x'\to x} Error_{1\text{-}NN}(x) = 2q(x)(1-q(x))$.

e. Se arată imediat că $z - z^2 \leq z$ pentru $\forall z$, deci şi pentru $z \in [0\,,\,1/2]$, iar $z - z^2 \leq 1 - z$ pentru $\forall z$, deci şi pentru $\in [1/2\,,\,1]$. Aşadar, pentru orice x, vom avea:

$$q(x)(1-q(x)) \leq \begin{cases} q(x) \text{ dacă } q(x) \in [0\,,\,1/2] \\ 1-q(x) \text{ dacă } q(x) \in (1/2\,,\,1]. \end{cases}$$

Coroborând cu rezultatul de la punctul b, obţinem: $2q(x)(1-q(x)) \leq 2Error_{Bayes}(x)$ pentru orice x.



Combinând acest rezultat cu egalitatea de la punctul d, rezultă că inegalitatea

$$\lim_{n\to\infty} Error_{1\text{-}NN}(x) = \lim_{x'\to x} Error_{1\text{-}NN}(x) \leq 2Error_{Bayes}(x)$$

este adevărată pentru orice x. Înmulţind ambii membri ai acestei inegalităţi cu $P(x)$ şi însumând după toate valorile lui x — de fapt, integrând în raport cu x —, obţinem:

$$E[\lim_{n\to\infty} Error_{1\text{-}NN}] \leq 2E[Error_{Bayes}].$$

Aşadar, am demonstrat că media (sau: rata medie a) erorii asimptotice a algoritmului 1-NN este cel mult dublul mediei (sau: ratei medii a) erorii algoritmului Bayes Optimal.

***Observaţia 1*: Această margine a erorii asimptotice nu se păstrează şi în cazul neasimptotic, unde numărul de exemple de antrenament este finit.**

***Observaţia 2*: La fel, se poate arăta că $2z - 2z^2 \geq z$ pentru $\forall z \in [0\,,\,1/2]$ şi $2z - 2z^2 \geq 1 - z$ pentru $\forall z \in [1/2\,,\,1]$. Luând din nou $z = q(x)$ şi ţinând cont de rezultatul de la punctul b, obţinem că $2q(x)(1-q(x)) \geq Error_{Bayes}(x)$, pentru orice x. Combinând această inegalitate cu egalitatea de la punctul d, rezultă o nouă inegalitate:**

$$\lim_{n\to\infty} Error_{1\text{-}NN}(x) = \lim_{x'\to x} Error_{1\text{-}NN}(x) \geq Error_{Bayes}(x) \text{ pentru orice } x.$$

În final, trecând la medii, obţinem următoarea inegalitate (care era de altfel de aşteptat):

$$E[\lim_{n\to\infty} Error_{1\text{-}NN}] \geq E[Error_{Bayes}].$$

Observaţia 3 **(preluată din *An Elementary Introduction to Statistical Learning Theory*, de Sanjeev Kulkarni şi Gilbert Harman, 2011, pag. 69): În mod intuitiv, dacă mărim valoarea lui k, ar trebui ca eroarea medie a algoritmului k-NN să se reducă. Într-adevăr, în anumite condiţii (dar nu în orice condiţii!) se poate arăta că are loc următoarea inegalitate dublă:**

$$E[Error_{Bayes}] \leq E[\lim_{n\to\infty} Error_{k\text{-}NN}] \leq \left(1+\frac{1}{k}\right) E[Error_{Bayes}].$$

Este de remarcat faptul că există distribuţii probabilistice ale datelor pentru care clasificatorul 1-NN se comportă mai bine decât k-NN pentru orice $k \neq 1$.

Observaţia 4 **(preluată din aceeaşi lucrare, *An Elementary Introduction to Statistical Learning Theory*, citată mai sus): Dacă lucrăm cu k_n-NN, adică îl fixăm pe k în funcţie de n (numărul instanţelor de antrenament), se poate demonstra că în cazul în care $\frac{k_n}{n} \to 0$ pentru $n \to \infty$ (de exemplu, $k_n = \sqrt{n}$), se obţine:**

$$E[\lim_{n\to\infty} Error_{k_n\text{-}NN}] = E[Error_{Bayes}].$$

Aceasta înseamnă că, la limită, algoritmul k_n-NN se comportă la fel de bine ca algoritmul Bayes Optimal!

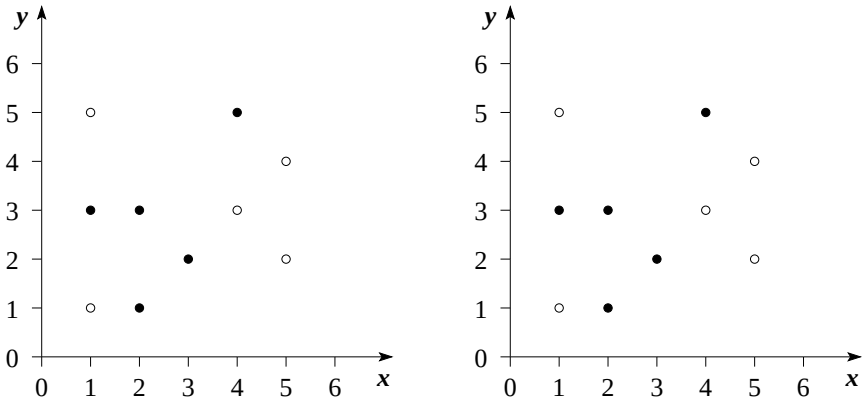
11. **(Comparaţie între algoritmii 1-NN şi ID3: zone şi suprafeţe de decizie)**

prelucrare de Liviu Ciortuz, după
■ • ***CMU, 2007 fall, Carlos Guestrin, HW2, pr 1.4***

Pe setul de date de mai jos desenaţi *graniţele de decizie* şi apoi haşuraţi *suprafeţele de decizie* produse de

a. algoritmul 1-NN (veţi obţine deci diagrama Voronoi);

b. algoritmul ID3 extins cu capacitatea de a procesa atribute cu valori continue.



Răspuns:

a. *Algoritmul 1-NN*:

Pentru a defini suprafeţele de decizie în acest caz se procedează în felul următor:

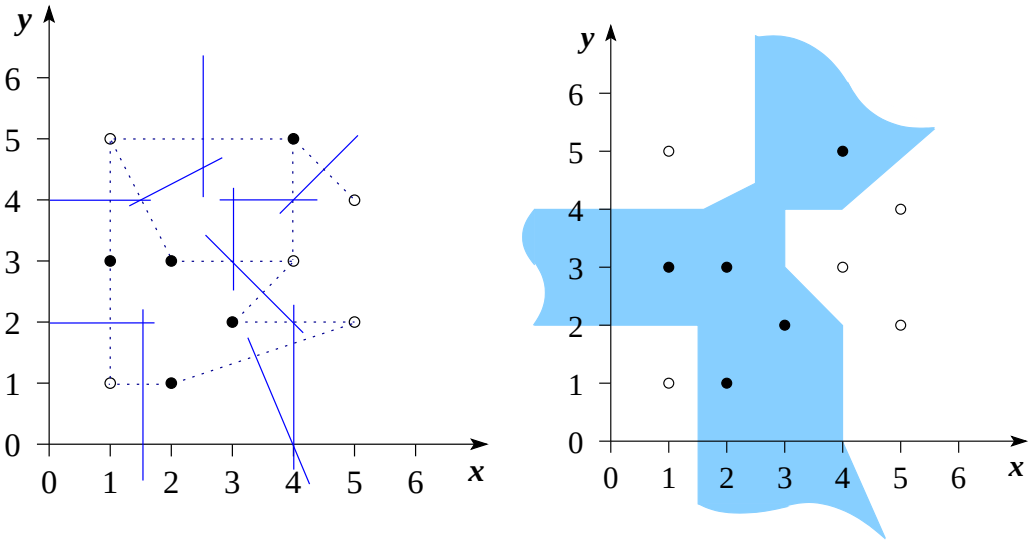
– se trasează mediatoarele segmentelor de dreaptă determinate de perechi de puncte din setul de antrenament care sunt etichetate în mod diferit;

– se stabilesc intersecţiile acestor mediatoare; acest lucru este reprezentat în figura de mai jos, partea stângă;

– apoi se marchează pe aceste mediatoare acele segmente (determinate de intersecţii) care determină zonele de decizie corespunzătoare clasificatorului 1-NN.

***Observaţii*:**
1. De fapt, întrucât nu este necesar să se lucreze cu toate perechile de instanţe cu etichete diferite sunt relevante pentru clasificarea unei instanţe / zone, în timpul „executării“ punctelor de mai sus este foarte util să se ţină cont de următoarea regulă / euristică de *ghidare*: alegerea perechilor de instanţe şi apoi a segmentelor de pe mediatoare se va face urmărind delimitarea zonelor corespunzătoare instanţelor negative (○) de zonele corespunzătoare instanţelor pozitive (•).
2. Suplimentar, în jurul fiecărui punct de antrenament A se poate identifica câte o zonă [convexă] care va constitui mulţimea punctelor mai apropiate de A decât de oricare alt punct din setul de date de antrenament. (Toate punctele din această zonă convexă vor avea aceeaşi clasificare / etichetă ca şi punctul A.) Aceasta este uşor de văzut pentru instanţele negative (○) din cazul de faţă.

În figura următoare, în partea dreaptă am haşurat zona corespunzătoare instanţelor pozitive (•).



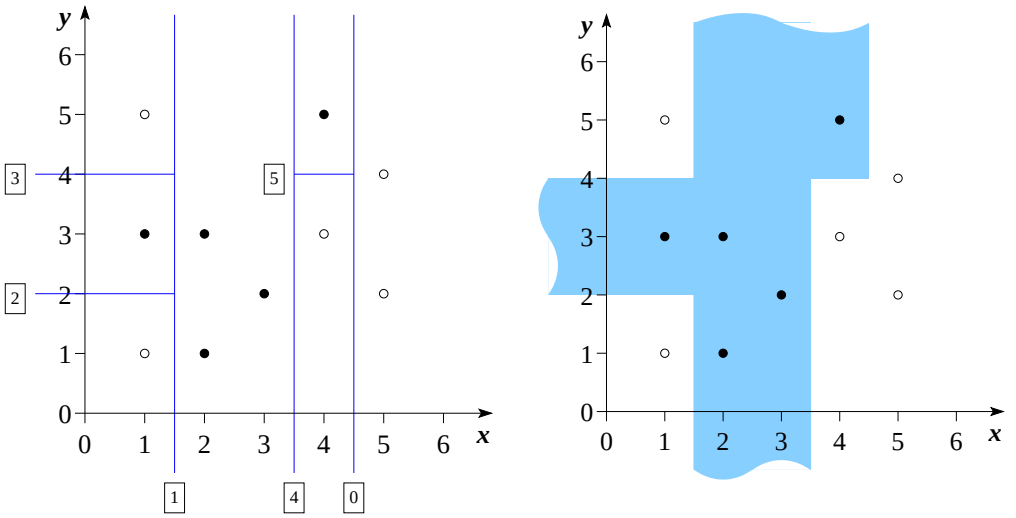
b. *Algoritmul ID3*:

Pentru construirea suprafeţelor de decizie ale algoritmului ID3,

– mai întâi se determină valorile-prag pentru teste (adică, punctele de splitare de pe fiecare axă) şi apoi se alege testul corespunzător nodului rădăcină din arborele ID3; se trasează o dreaptă prin punctul respectiv, paralelă cu cealaltă axă. În cazul nostru, testul din nodul rădăcină va fi $x < 4.5$.

– pentru fiecare test / split ulterior se trasează o semidreaptă (sau un segment de dreaptă) mărginit(ă) la un capăt de dreapta corespunzătoare noduluipărinte.

În figura de mai jos, în partea stângă am vizualizat toate testele / split-urile realizate de algoritmul ID3 pentru a învăţa complet arborele de decizie, iar în partea dreaptă, ca şi mai sus, am păstrat doar frontierele dintre zonele cu clasificări diferite.



Se observă că suprafeţele de decizie determinate de cei doi algoritmi nu sunt identice, dar sunt totuşi asemănătoare într-o anumită măsură (pentru că ambele sunt consistente cu datele de antrenament).

***Observaţii*:**
3. Este de reţinut faptul că suprafaţele de decizie produse de algoritmul ID3 nu sunt neapărat unic determinate, fiindcă sunt situaţii în care două teste diferite pot conduce la acelaşi câştig de informaţie. De exemplu, dacă în exerciţiul nostru am avea de partiţionat la un moment dat mulţimea formată din instanţele de antrenament $(1,1),(1,3),(2,1),(2,3),(3,2)$, atunci testele $x > 1.5$ şi $y > 1.5$ ar produce acelaşi câştig de informaţie, iar suprafeţele de decizie rezultate ar fi determinate (în mod diferit!) de ce anume alegem ca prim test.
4. De asemenea, trebuie să scoatem în evidenţă faptul că pragurile / spliturile care sunt calculate de algoritmul ID3 pentru un acelaşi atribut continuu pot diferi de la un nod de test la altul. De exemplu, la nodul rădăcină (nodul 0), atunci când se calculează câştigul de informaţie maxim, pentru atributul y se iau în calcul pragurile 1.5, 2.5, 3.5, şi 4.5, în vreme ce la nodurile 2 şi / sau 3 (vedeţi figura de mai sus, partea stângă) se analizează pragurile 2 şi 4, întrucât seturile / partiţiile de instanţe asignate acestor noduri sunt diferite!

12. (Comparaţie între algoritmii 1-NN, ID3 cu atribute continue şi SVM: eroarea la antrenare, eroarea la CVLOO)

prelucrare de Liviu Ciortuz, după

- ***CMU, 2003 fall, T. Mitchell, A. Moore, midterm exam, pr. 5***

Folosind metoda 1-NN cu distanţa euclidiană, învăţăm un clasificator cu două valori pentru atributul de ieşire, $Y = 0$ şi $Y = 1$, pornind de la datele de antrenament din tabelul de mai jos (X_1 şi X_2 sunt atribute de intrare).

X_1	X_2	Y
0	0	1
1	0	1
2	0	1
2.5	2	1
3	0	1
1	2	0
1.5	0	0
2	2	0
3	2	0
4	2	0

a. Care este eroarea la antrenare (exprimată ca număr de exemple clasificate eronat)?

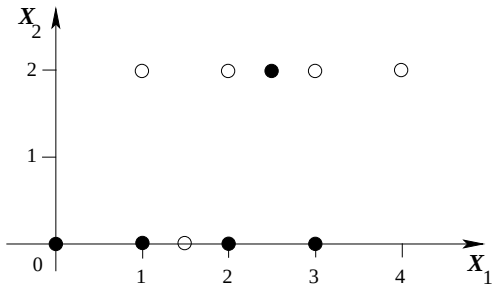
b. Care este eroarea la cross-validare folosind metoda "Leave-One-Out"?

c. Răspundeţi la întrebările de mai sus, considerând acum arbori de decizie cu atribute numerice continue în locul metodei 1-NN.

d. În sfârşit, răspundeţi la întrebările a şi b, considerând maşini cu vectori-suport în locul metodei 1-NN. (Se vor considera doar SVM-uri în cazul liniar cu margine "soft", cu un parametru C suficient de mare pentru a minimiza numărul de instanţe de antrenament clasificate eronat.)

Răspuns:

Reprezentarea datelor în planul euclidian este cea din figura alăturată.



a. După cum am explicat şi la exerciţiul 2, întrucât datele de antrenament nu conţin inconsistenţe, eroarea la antrenare produsă de algoritmul 1-NN va fi 0.

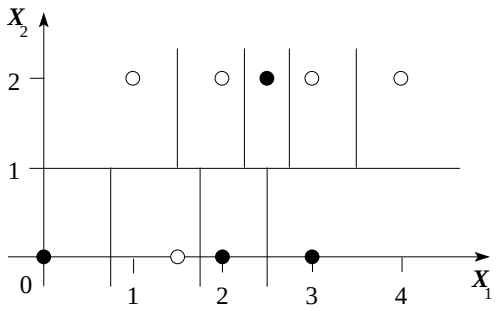
b. Comportamentul algoritmului la cross-validare cu metoda "Leave-One-Out" este cel descris în tabelul următor:

Data	Eticheta	Vecinătate	Clasificare la CVLOO	Eroare?
$(0;0)$	1	$(1;0)$	1	**nu**
$(1;0)$	1	$(1.5;0)$	0	**da**
$(2;0)$	1	$(1.5;0)$	0	**da**
$(2,5;2)$	1	$(2;2)/(3;2)$	0	**da**
$(3;0)$	1	$(2;0)$	1	**nu**
$(1;2)$	0	$(2;2)$	0	**nu**
$(1,5;0)$	0	$(1;0)/(2;0)$	1	**da**
$(2;2)$	0	$(2.5;2)$	1	**da**
$(3;2)$	0	$(2.5;2)$	1	**da**
$(4;2)$	0	$(3;2)$	0	**nu**

Deci în total avem 6 erori (din totalul de 10 instanţe), ceea ce indică faptul ca algoritmul 1-NN este foarte puţin adecvat pentru acest gen de date.

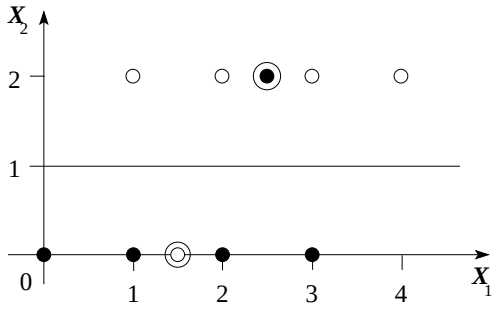
c. Eroarea la antrenare cu algoritmul ID3 pe acest set de date este 0, fiindcă instanţele sunt etichetate în mod consistent.

Eroarea la CVLOO produsă de algoritmul ID3 pe acelaşi set de date se poate calcula foarte uşor, ţinând cont că este suficient să determinăm graniţele de decizie — deci şi zonele de decizie — corespunzătoare celor două clase. De exemplu, atunci când se va elimina instanţa (1,0), se vor obţine graniţele de decizie ca în figura alăturată. Este imediat că *modelul* rezultat în acest caz va clasifica instanţa (1,0) în mod eronat.



Se constată că eroarea de tip CVLOO produsă de ID3 cu atribute continue este de 6 (din totalul de 10) instanţe. Ba chiar, — este încă o coincidenţă! — ID3 produce la CVLOO exact aceleaşi erori (mai precis, el clasifică în mod eronat exact aceleaşi instanţe) ca şi algoritmul 1-NN.

d. Se observă uşor că separatorul liniar care minimizează eroarea la antrenare este cel reprezentat în figura alăturată. Datele clasificate eronat de către acest separator sunt cele două puncte încercuite din figură; ele sunt de asemenea singurele puncte care generează eroare şi la testare cu metoda CVLOO.



Mai concret, în cazul CVLOO folosind SVM, deoarece separatorul optimal este „susţinut“ de mai mulţi vectori-suport pe fiecare parte, lipsa unuia singur dintre ei nu influenţează cu nimic construirea separatorului. În fiecare caz în parte se va învăţa ca separator dreapta paralelă cu Ox_1 care trece prin punctul $(0, 1)$.

Avem deci în ambele situaţii, adică atât la antrenare cât şi la cross-validare cu metoda "Leave-One-Out", (doar) două puncte clasificate eronat de către SVM.

Comparând cei trei clasificatori, 1-NN, ID3 cu atribute continue şi SVM, rezultă în mod clar că SVM este cel mai convenabil pe acest set de date (deşi la antrenare SVM produce două erori, iar 1-NN şi ID3 nicio eroare), fiindcă nu produce *overfitting* (aici!).

13. (Comparaţii între algoritmii 1-NN şi ID3: Da sau nu?)

CMU, 2009 spring, Ziv Bar-Joseph, final exam, pr. 2.1

a. Este posibil să se construiască un arbore de decizie (având în fiecare nod intern teste de forma $x > a$, $x \leq b$, $y > c$, sau $y \leq d$, unde a, b, c, d sunt numere reale oarecare) care să producă la clasificare aceleaşi rezultate ca şi algoritmul 1-NN folosind distanţa euclidiană? Justificaţi răspunsul.

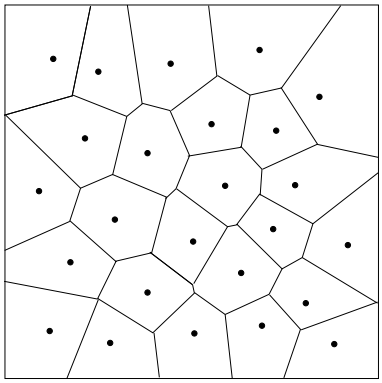
(Învăţare rapidă / "eager" vs. învăţare lentă / "lazy";
k-NN vs. ID3)

CMU, 2010 spring, HW1, pr. 3.3

b. Algoritmul ID3 este o metodă de învăţare de tip "batch", care solicită ca toate datele de antrenament să-i fie puse la dispoziţie pentru a putea elabora arborele de decizie. Aşadar, în situaţia în care date de antrenament suplimentare ne sunt puse ulterior la dispoziţie, acestea trebuie tratate cu atenţie fiindcă ele pot modifica arborele de decizie rezultat în urma învăţării.[368] Algoritmul k-NN suferă şi el de această problemă? Justificaţi.

Răspuns:

a. Nu.
Suprafeţele de decizie pentru algoritmul 1-NN corespund diagramei Voronoi şi nu sunt neapărat paralele cu axele de coordonate, după cum se observă în figura alăturată.
Suprafeţele de decizie pentru un arbore de decizie cu atribute cu valori continue sunt întotdeauna paralele cu axele de coordonate, deoarece deciziile din fiecare nod sunt de forma $x > a$, $x \leq b$, $y > c$, sau $y \leq d$, $\forall a, b, c, d \in \mathbb{R}$.



Aşadar, răspunsul este negativ pentru cazul general, deşi există situaţii în care cei doi algoritmi produc exact aceeaşi clasificare.

b. Nu.
Răspunsul decurge din modul de lucru a algoritmului k-NN, care este un algoritm de învăţare de tip "lazy": k-NN estimează valoarea locală a unei funcţii-target $\hat{f}$ pentru una sau mai multe instanţe de test x_q. Valoarea $\hat{f}(x_q)$ nu depinde decât de cei mai apropiaţi vecini ai lui x_q; celelalte instanţe de antrenament nu sunt necesare pentru calculul lui $\hat{f}(x_q)$. Evident, dacă pe măsură ce se acumulează noi date de antrenament se modifică şi vecinătatea lui x_q, atunci se prea poate să se modifice şi $\hat{f}(x_q)$. Însă în sine, algoritmul procedează exact la fel ca mai înainte. Se poate modifica doar outputul lui. Spre deosebire de algoritmul k-NN, la învăţarea arborilor de decizie adăugarea de noi instanţe modifică în general şi modelul / arborele rezultat, nu doar decizia pentru o instanţă de test particulară.

14. (1-NN cu mapare cu RBF: Adevărat sau Fals?)

■ ● ○ ***CMU, 2003 fall, T. Mitchell, A. Moore, final exam, pr. 7.f***

Algoritmul 1-NN folosind distanţa euclidiană neponderată este capabil să obţină rezultate mai bune dacă în prealabil intrările sale sunt mapate într-un „spaţiu de trăsături" folosind o funcţie-nucleu cu baza radială (RBF).

Răspuns:

[368] Vedeţi problema 16 de la capitolul *Arbori de decizie*.

Fals.
Fie $\phi : \mathbb{R}^d \to \mathbb{R}^n$ funcţia de mapare în spaţiul de trăsături, astfel încât să avem $K(x,y) \stackrel{not.}{=} e^{-\frac{||x-y||^2}{2\sigma^2}} = \phi(x) \cdot \phi(y), \forall x, y \in \mathbb{R}^d$. ($\mathbb{R}^d$ reprezintă spaţiul iniţial, $\mathbb{R}^n$ spaţiul de trăsături în care se face maparea, iar $e^{-\frac{||x-y||^2}{2\sigma^2}}$ este funcţia-nucleu cu baza radială.) Avem:

$$||\phi(x) - \phi(y)||^2 = (\phi(x) - \phi(y)) \cdot (\phi(x) - \phi(y))$$
$$= \phi(x) \cdot \phi(x) + \phi(y) \cdot \phi(y) - 2 \cdot \phi(x) \cdot \phi(y) = e^{-\frac{||x-x||^2}{2\sigma^2}} + e^{-\frac{||y-y||^2}{2\sigma^2}} - 2 \cdot e^{-\frac{||x-y||^2}{2\sigma^2}}$$
$$= e^0 + e^0 - 2 \cdot e^{-\frac{||x-y||^2}{2\sigma^2}} = 2 - 2 \cdot e^{-\frac{||x-y||^2}{2\sigma^2}} = 2 - K(x,y).$$

Prin urmare, pentru orice $x, x_i, x_j \in \mathbb{R}^d$ vom avea:

$$||\phi(x) - \phi(x_i)||^2 \leq ||\phi(x) - \phi(x_j)||^2 \Leftrightarrow 2 - K(x,x_i) \leq 2 - K(x,x_j) \Leftrightarrow K(x,x_i) \geq K(x,x_j)$$
$$\Leftrightarrow e^{-\frac{||x-x_i||^2}{2\sigma^2}} \geq e^{-\frac{||x-x_j||^2}{2\sigma^2}} \Leftrightarrow -\frac{||x-x_i||^2}{2\sigma^2} \geq -\frac{||x-x_j||^2}{2\sigma^2} \Leftrightarrow ||x-x_i||^2 \leq ||x-x_j||^2.$$

Cu alte cuvinte, dacă o instanţă de test x are drept cel mai apropiat vecin punctul x_i în spaţiul iniţial, acest lucru rămâne valabil şi în spaţiul de trăsături. Aşadar, decizia algoritmului 1-NN este identică în ambele spaţii. Aceeaşi concluzie este valabilă şi pentru k-NN.

***Observaţie*: Este însă posibil ca folosind ponderarea sau alte măsuri de distanţă (decât cea euclidiană) kernel-izarea să funcţioneze cu succes pentru k-NN.**

3.2 Învăţare bazată pe memorare — Probleme propuse

15. **(Algoritmul k-NN: acurateţe; comparaţie cu un simplu clasificator aleator)**

prelucrare de Liviu Ciortuz, după CMU, 2014 fall, W. Cohen, Z. Bar-Joseph, HW1, pr. 5.bc

a. Enunţaţi [succint] *regula de decizie* a algoritmului k-NN pentru o instanţă de test x_q.

Care este *bias*-ul inductiv al algoritmului k-NN?

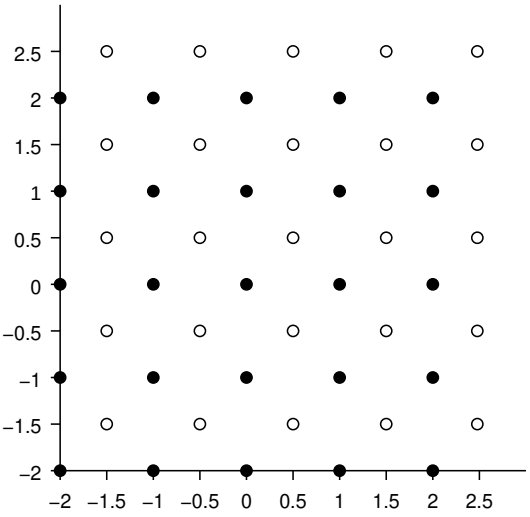
În continuare veţi lucra pe datele din figura de mai jos. Veţi aplica algoritmul k-NN, folosind distanţa euclidiană.

k-NN funcţionează bine atunci când instanţele dintr-o aceeaşi clasă sunt plasate într-una sau mai multe zone din spaţiu relativ bine delimitate, fără întrepătrunderi puternice.

***Obiectivul* nostru acum este să analizăm ce se întâmplă atunci când datele sunt puternic mixate. Rezultatele pe care le veţi obţine la calculul erorilor vor fi exprimate sub formă de numere [fracţionare] din intervalul $[0, 1]$.**

***Observaţie importantă*:**

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.



b. Pentru $k = 1$, calculaţi eroarea la antrenare şi eroarea la cross-validare cu metoda “leave-one-out” (CVLOO).

Ce puteţi spune comparând cele două rezultate? (Care este legătura între *bias*-ul inductiv al lui k-NN şi puterea de generalizare a lui 1-NN pe astfel de date? Se produce oare aici un *anumit* fenomen, specific multor situaţii din clasificarea automată?)

c. Pentru $k = 2$, calculaţi eroarea la cross-validare cu metoda “leave-one-out” (CVLOO).

d. Considerăm $k = 50$. (Remarcaţi faptul că în total în setul nostru de date sunt 50 de instanţe.) De această dată, vom impune ca algoritmul k-NN să ia decizia în mod *probabilist*. Aceasta înseamnă că dacă în vecinătatea k-NN a unei instanţe de test există n vecini pozitivi şi m vecini negativi,

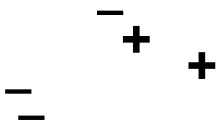
atunci algoritmul k-NN va returna (pentru instanţa respectivă) decizia $+$ cu probabilitatea $n/(n+m)$ şi decizia $-$ cu probabilitatea $m/(n+m)$. În consecinţă, pentru întreg setul de date vom putea calcula o *eroare medie*.

Calculaţi *eroarea medie* la antrenare pentru algoritmul 50-NN pe datele de mai sus. Cunoaşteţi o metodă de clasificare foarte simplă care obţine pe aceste date rezultate la fel de bune / proaste precum 50-NN?

16. **(Algoritmul 1-NN: calculul erorii la CVLOO)**

$*$ ***CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh, midterm exam, pr. 1.7***

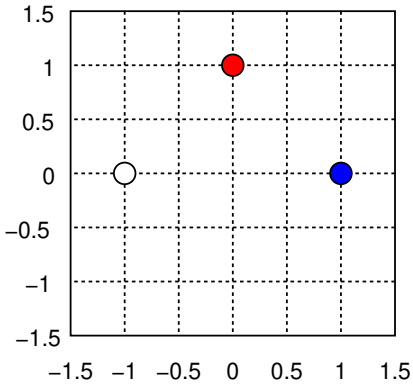
Care este eroarea clasificatorului 1-NN la cross-validare de tip "Leave-One-Out" pe setul de date alăturat?



17. **(Algoritmul 1-NN: graniţe / suprafeţe de decizie)**

• ***CMU, 2013 fall, W. Cohen, E. Xing, final exam, pr. 3.6***

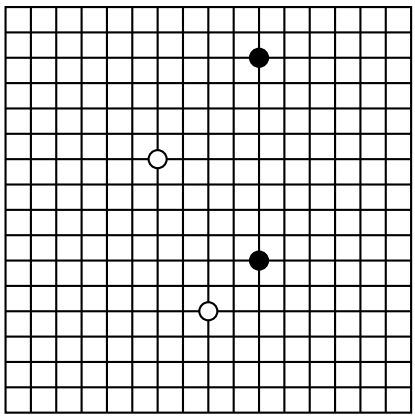
Trasaţi graniţele de decizie (engl., *decision boundaries*) produse de clasificatorul 1-NN la aplicarea pe datele din figura alăturată. Diversele forme ale punctelor reprezintă clase diferite.

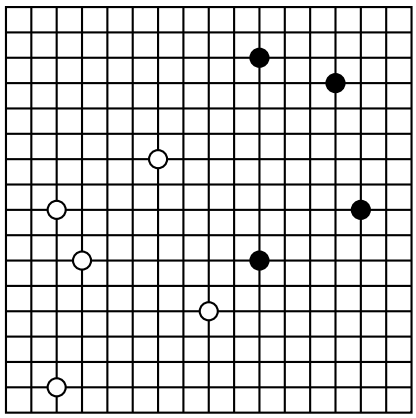


18. **(Algoritmul 1-NN: graniţe / suprafeţe de decizie)**

$*$ ***CMU, 2008 fall, Eric Xing, HW1, pr. 3***

În fiecare din figurile următoare se dau câteva puncte în spaţiul bidimensional, etichetate cu $+$ sau $-$. Indicaţi în fiecare caz graniţele / suprafeţele de decizie pentru algoritmul 1-NN presupunând că se foloseşte distanţa euclidiană.

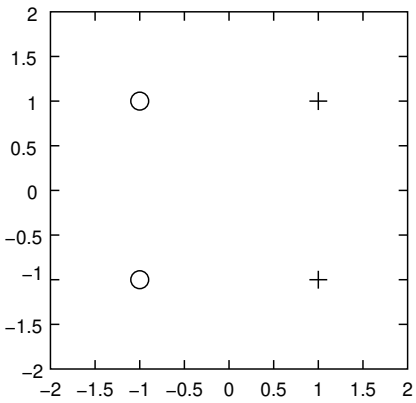


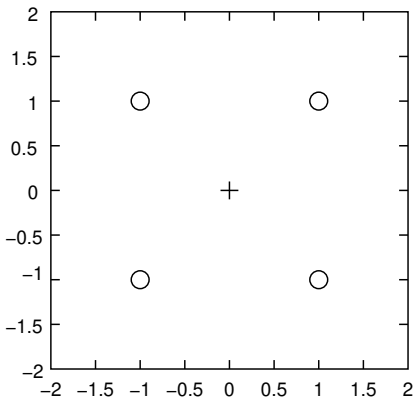


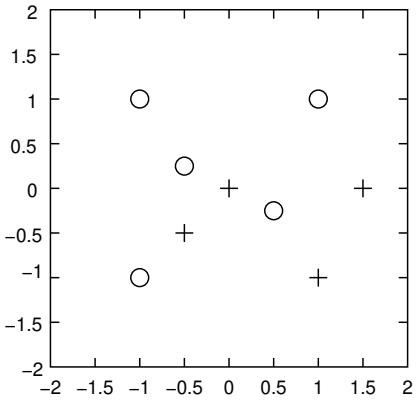
19. **(Algoritmul 1-NN: diagrame Voronoi)**

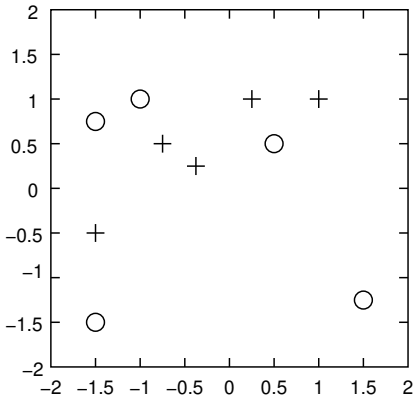
• ∗ ***CMU, 2010 fall, Ziv Bar-Joseph, HW1, pr. 3.1***

Desenaţi suprafeţele de decizie pentru clasificatorul 1-NN pentru fiecare dintre seturile de date din figurile de mai jos. Folosiţi distanţa euclidiană. Haşuraţi fin zonele corespunzătoare clasei +.





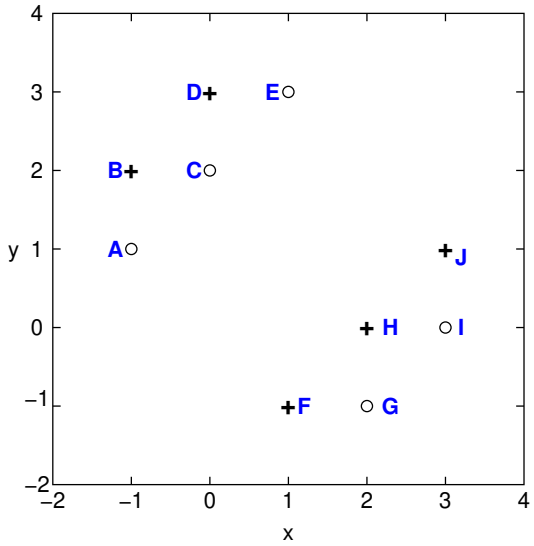




20. **(Algoritmul k-NN: diagrama Voronoi, eroarea la CVLOO; comparaţie pentru diferite valori ale lui k)**

• ***prelucrare de L. Ciortuz, după CMU, 2012 spring, Ziv Bar-Joseph, midterm exam, pr. 2***

La acest exerciţiu veţi aplica algoritmul k-NN folosind distanţa euclidiană pe setul de date din figura de mai jos. Fiecare punct aparţine la una din două clase, desemnate cu $+$ şi respectiv $\circ$.



a. Trasaţi diagrama Voronoi şi haşuraţi zona / zonele de decizie corespunzătoare etichetei $+$.

b. Care este eroarea la cross-validare cu metoda "Leave-One-Out" (CVLOO) dacă se foloseşte algoritmul 1-NN?

c. Care dintre următoarele valori ale lui k va conduce la o valoare minimă a erorii de tip CVLOO: 3, 5, 7 sau 9? Comentaţi succint rezultatul.

***Indicaţii*:**

1. k-NN-vecinătăţile vor fi construite în manieră *inclusivă*.[369] **Vă *cerem(!)* să puneţi în evidenţă toate cazurile de acest tip. Vedeţi, spre exemplu, liniile 2 şi 3 din tabelul de mai jos, coloana 1-NN vecinătăţilor.**

2. În caz de *paritate la voturi* (dar doar în acest caz!), se va considera că se aplică (în mod intuitiv) ponderarea distanţelor în sensul prezentat la curs.

3. Dacă veţi şti să exploataţi *simetriile*, veţi avea mult mai puţin de elaborat la nivel de detaliu!

21. **(Algoritmul k-NN: alegerea valorii convenabile pentru k)**

prelucrare de Liviu Ciortuz, după CMU, (?) spring, ML course 10-701, HW1, pr. 5

Pe un anumit set de date format din date de antrenament, date de validare şi date de test, după ce fost antrenat algoritmul k-NN pentru diferite valori

[369] Adică, dacă notăm cu

– $x_1, x_2, \ldots, x_n$ instanţele de antrenament,

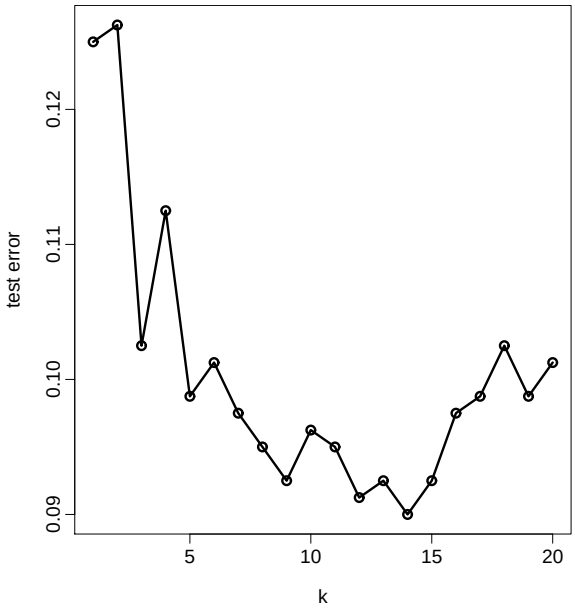
– x o instanţă oarecare căreia i se aplică la un moment dat procedura de cross-validare LOO cu algoritmul k-NN (unde k este fixat),

– $d(x, x_{i_1}) \leq d(x, x_{i_2}) \leq \ldots \leq d(x, x_{i_n})$ secvenţa ordonată a distanţelor de la x la fiecare din instanţele de antrenament,

şi există $l > 0$ astfel încât $x_{i_k} = x_{i_{k+1}} = \ldots = x_{i_{k+l}} < x_{i_{k+l+1}}$ sau $k + l = n$, atunci în *k-NN vecinătatea* lui x vor fi incluse toate instanţele $x_{i_{k+1}}, \ldots, x_{i_{k+l}}$.

ale lui k, rezultatele obţinute la validare au fost reprezentate în graficul care urmează.

Care este — în conformitate cu aceste rezultate — valoarea optimală care trebuie aleasă pentru k, în vederea folosirii ulterioare pe datele de test? Justificaţi alegerea făcută.



22. **(Algoritmul k-NN: acurateţe; comparaţii pentru diferite valori ale lui k)**

• ○ ***CMU, 2014 fall, W. Cohen, Z. Bar-Joseph, HW1, pr. 5.a***

Considerăm două clase, notate cu C_1 şi C_2, în spaţiul euclidian bidimensional. Datele din clasa C_1 sunt distribuite în mod uniform într-un cerc de rază r. Datele din clasa C_2 sunt distribuite în mod uniform într-un alt cerc de rază r. (*Observaţie*: Numărul de date din cele două clase nu este neapărat acelaşi.) Centrele celor două cercuri sunt situate la o distanţă strict mai mare decât $4r$.

Arătaţi că este posibil ca [la antrenare] acurateţea algoritmului 1-NN aplicat pe aceste date să fie *strict* mai mare decât acurateţea algoritmului k-NN, pentru un anumit număr întreg $k \geq 3$, impar, ales în mod convenabil.

23. **(Algoritmul k-NN: întrebări de ordin calitativ)**

□ • ○ ***CMU, 2012 spring, Ziv Bar-Joseph, HW1, pr. 3***

Ştim că la aplicarea algoritmului k-NN, clasificarea unei instanţe date se face pe baza votului majoritar obţinut în „vecinătatea“ instanţei respective. Presupunem că se dau două clase de instanţe, fiecare clasă având $n/2$ puncte, întrepătrunse într-o anumită măsură, într-un spaţiu bidimensional.

a. Descrieţi ce se întâmplă cu *eroarea la antrenare* (folosind toate datele disponibile) când numărul k al vecinilor consideraţi variază de la n la 1.

b. Schiţaţi grafic cum anume ar evolua *eroarea de generalizare* (de exemplu, reţinând o parte din date pentru testare) atunci când k variază. Explicaţi modul în care aţi raţionat.

c. Propuneţi o metodă de determinare a unei valori adecvate pentru k.

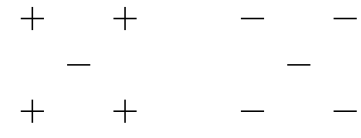
d. La folosirea algoritmului k-NN, odată ce s-a stabilit valoarea lui k, toţi cei mai apropiaţi k vecini ai punctului de clasificat au ponderi egale (adică, aceeaşi importanţă) la stabilirea etichetei respectivului punct. Sugeraţi o modificare a algoritmului k-NN care elimină această limitare.

e. Daţi două motive pentru care este de preferat să nu folosim algoritmul k-NN atunci când dimensiunea spaţiului datelor de intrare este mare.

24. **(Algoritmul k-NN: CVLOO: comparaţie pentru diferite valori ale lui k; eroarea la antrenare: comparaţie cu alţi clasificatori)**

• ○ ***CMU, 2010 fall, Aarti Singh, midterm exam, pr. 2***

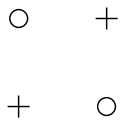
a. Care dintre clasificatorii de mai jos realizează o eroare de tip CVLOO (Leave-One-Out Cross-validation) mai mare pe setul de date alăturat?



☐ 1-NN ☐ 3-NN

***Recomandare*: Formulaţi explicit euristica pe care o veţi folosi pentru a trata cazurile în care apar tot atâtea voturi pozitive cât cele negative.**

b. Considerăm setul de date din figura alăturată. Care dintre clasificatorii de mai jos obţine / obţin eroare nulă la antrenare pe acest set de date?



☐ **arborii de decizie ID3 de adâncime 2**

☐ **clasificatorul 3-NN**

☐ **regresia logistică**

☐ **SVM (cu nucleu pătratic)**

25. **(Compararea clasificatorilor 1-NN şi Bayes Optimal: o margine superioară mai bună pentru *rata medie a erorii asimptotice* a lui 1-NN)**

∗ ■ ***Liviu Ciortuz, 2014, bazat pe un rezultat din "An Elementary Introduction to Statistical Learning Theory", S. Kulkarni, G. Harman, 2011, pag. 68-69***

La problema 10 am demonstrat că *rata medie a erorii* clasificatorului 1-NN este mărginită asimptotic[370] de dublul ratei medii a erorii clasificatorului Bayes Optimal.

[370] Adică, antunci când $n \to \infty$, unde n este numărul de instanţe de antrenament.

Arătaţi că — în aceleaşi condiţii ca la problema 10 — se poate obţine o margine chiar mai bună:

$$E[\lim_{n\to\infty} Error_{1\text{-}NN}] \leq 2E[Error_{Bayes}](1 - E[Error_{Bayes}]).$$

26. **(Adevărat sau Fals?)**

○ ***CMU, 2010 fall, Ziv Bar-Joseph, midterm, pr. 1.bc***

Care dintre următoarele afirmaţii sunt adevărate pentru clasificatorii k-NN? (Justificaţi pe scurt răspunsul, în dreptul fiecărui punct.)

a. Acurateţea la antrenare creşte pe măsură ce creşte valoarea lui k.

b. Suprafaţa de decizie este mai netedă (engl., smoother) pe măsură ce valoarea lui k scade.

c. k-NN nu necesită o procedură explicită de antrenare.

d. Suprafaţa / graniţa de decizie este liniară.

e. Este posibil ca un clasificator binar 1-NN să clasifice întotdeauna orice instanţă de test ca fiind pozitivă, chiar dacă în setul de date de antrenament există instanţe negative.

27. **(Întrebări calitative despre design-ul unor experimente din Învăţarea Automată: OK ori ...problematic?)**

● ○ ***CMU, 2009, Geoff Gordon, midterm exam, pr. 3***

Fiecare din punctele de mai jos prezintă pe scurt design-ul unui experiment practic de învăţare automată. Analizaţi fiecare din aceste cazuri, indicând apoi dacă respectivul experiment este *ok* ori *problematic* (încercuiţi varianta pe care o alegeţi). Dacă este *problematic*, identificaţi TOATE defectele [de concepţie ale] design-ului respectiv.

a. O echipă de proiectare raportează o *eroare* mică *la antrenare* şi susţine că metoda folosită este bună.

Ok
Problematic

b. O echipă de proiectare susţine că este un mare succes faptul că a obţinut 98% *acurateţe la antrenare* pentru un task de clasificare binară care are următorul specific: unul din cele două cazuri se întâlneşte foarte rar comparativ cu celălalt caz. (O astfel de problemă o constituie, de exemplu, identificarea tranzacţiile bancare frauduloase.) Datele lor au constat din 50 de exemple pozitive şi 4950 de exemple negative.

Ok
Problematic

c. O echipă de proiectare şi-a împărţit datele de care dispune în date de antrenament şi date de test. Folosind datele de antrenament, ei au construit

un *model* de clasificare caracterizat de anumiţi *parametri*. Apoi, făcând *cross-validare*, au ales cea mai bună setare a parametrilor. La final, au raportat *eroarea* obţinută pe *datele de test*.

Ok
Problematic

d. O echipă de proiectare a efectuat o procedură de *selecţie a atributelor* (engl., features) pe toate datele şi apoi a redus setul mare de atribute la un set mai mic. După aceea, membrii echipei au împărţit datele în date de test şi date de antrenament. Au construit *modelul* de clasificare pe datele de antrenament folosind mai multe setări ale parametrilor modelului, şi au raportat cea mai bună *eroare la testare* pe care au obţinut-o.

Ok
Problematic