Seminarul 4

1. Se doreşte *clasificarea naivă Bayes* a unor **r**estaurante, în *clasele*: **r**ecomandat sau **n**erecomandat, în funcţie de următoarele *atribute* cu valorile lor posibile:
- cost: ieftin, mediu, scump;
- timp de aşteptare: puţin, mediu, îndelungat;
- mâncare: fadă, acceptabilă, bună, delicioasă.

Fie $\mathbf{R}$, C, T, M variabilele aleatoare corespunzătoare şi $\mathbf{r}$, $\mathbf{n}$, i, m, s, p, m, $\hat{\imath}$, f, a, b, d valorile de mai sus, în ordinarea în care sunt menţionate.

Considerăm următorul *tabel de date* furnizat de clienţii unor restaurante:

	Cost	*Timp de aşteptare*	*Mâncare*	**Restaurant**
1	mediu	îndelungat	acceptabilă	**nerecomandat**
2	scump	puţin	bună	**recomandat**
3	ieftin	îndelungat	delicioasă	**recomandat**
4	mediu	puţin	bună	**recomandat**
5	ieftin	mediu	acceptabilă	**nerecomandat**
6	ieftin	puţin	fadă	**nerecomandat**
7	mediu	puţin	acceptabilă	**nerecomandat**
8	mediu	mediu	delicioasă	**recomandat**
9	scump	puţin	delicioasă	**recomandat**
10	ieftin	îndelungat	bună	**nerecomandat**
11	scump	puţin	acceptabilă	**nerecomandat**
12	mediu	mediu	bună	**recomandat**
13	mediu	îndelungat	fadă	**nerecomandat**
14	scump	mediu	delicioasă	**recomandat**
15	ieftin	mediu	fadă	**nerecomandat**
16	mediu	puţin	delicioasă	**recomandat**
17	ieftin	puţin	acceptabilă	**recomandat**
18	scump	îndelungat	bună	**nerecomandat**
19	ieftin	puţin	fadă	**recomandat**
20	scump	îndelungat	delicioasă	**nerecomandat**

i) Folosind datele din tabel, determinaţi probabilităţile claselor şi probabilităţile condiţionate ale atributelor, ştiind clasa.
ii) Considerăm evenimentul dat de *vectorul de atribute*: $E = (C = s) \cap (T = m) \cap (M = b)$. Alegeţi o clasă pentru E, stabilind care din următoarele probabilităţi este mai mare: $P(\mathbf{R} = \mathbf{r}|E)$ sau $P(\mathbf{R} = \mathbf{n}|E)$.
iii) Determinaţi $P(E)$.

Rezolvare:

i)

$\mathbf{R} = \mathbf{r}$	$\mathbf{R} = \mathbf{n}$	$P(\mathbf{R} = \mathbf{r})$	$P(\mathbf{R} = \mathbf{n})$
10	10	$\frac{1}{2}$	$\frac{1}{2}$

C	$\mathbf{R} = \mathbf{r}$	$\mathbf{R} = \mathbf{n}$	$P(C = ...\|\mathbf{R} = \mathbf{r})$	$P(C = ...\|\mathbf{R} = \mathbf{n})$
i	3	4	$\frac{3}{10}$	$\frac{4}{10}$
m	4	3	$\frac{4}{10}$	$\frac{3}{10}$
s	3	3	$\frac{3}{10}$	$\frac{3}{10}$

T	$\mathbf{R}=\mathbf{r}$	$\mathbf{R}=\mathbf{n}$	$P(T=...\vert\mathbf{R}=\mathbf{r})$	$P(T=...\vert\mathbf{R}=\mathbf{n})$
p	6	3	$\frac{6}{10}$	$\frac{3}{10}$
m	3	2	$\frac{3}{10}$	$\frac{2}{10}$
$\hat{\imath}$	1	5	$\frac{1}{10}$	$\frac{5}{10}$

M	$\mathbf{R}=\mathbf{r}$	$\mathbf{R}=\mathbf{n}$	$P(M=...\vert\mathbf{R}=\mathbf{r})$	$P(M=...\vert\mathbf{R}=\mathbf{n})$
f	1	3	$\frac{1}{10}$	$\frac{3}{10}$
a	1	4	$\frac{1}{10}$	$\frac{4}{10}$
b	3	2	$\frac{3}{10}$	$\frac{2}{10}$
d	5	1	$\frac{5}{10}$	$\frac{1}{10}$

ii) Pe baza formulei lui Bayes şi a ipotezei de independenţă condiţionată, deducem că:

$$P(\mathbf{R}=\mathbf{r}|E) = \frac{P(E|\mathbf{R}=\mathbf{r})P(\mathbf{R}=\mathbf{r})}{P(E)} = \frac{P(C=s,T=m,M=b|\mathbf{R}=\mathbf{r})P(\mathbf{R}=\mathbf{r})}{P(E)}$$

$$= \frac{P(C=s|\mathbf{R}=\mathbf{r})P(T=m|\mathbf{R}=\mathbf{r})P(M=b|\mathbf{R}=\mathbf{r})P(\mathbf{R}=\mathbf{r})}{P(E)} = \frac{\frac{3}{10}\cdot\frac{3}{10}\cdot\frac{3}{10}\cdot\frac{1}{2}}{P(E)} = \frac{1}{P(E)}\cdot\frac{27}{2000}$$

şi

$$P(\mathbf{R}=\mathbf{n}|E) = \frac{P(E|\mathbf{R}=\mathbf{n})P(\mathbf{R}=\mathbf{n})}{P(E)} = \frac{P(C=s,T=m,M=b|\mathbf{R}=\mathbf{n})P(\mathbf{R}=\mathbf{n})}{P(E)}$$

$$= \frac{P(C=s|\mathbf{R}=\mathbf{n})P(T=m|\mathbf{R}=\mathbf{n})P(M=b|\mathbf{R}=\mathbf{n})P(\mathbf{R}=\mathbf{n})}{P(E)} = \frac{\frac{3}{10}\cdot\frac{2}{10}\cdot\frac{2}{10}\cdot\frac{1}{2}}{P(E)} = \frac{1}{P(E)}\cdot\frac{12}{2000}.$$

Deoarece $P(\mathbf{R}=\mathbf{r}|E) > P(\mathbf{R}=\mathbf{n}|E)$, asociem vectorului de atribute E clasa $\mathbf{R}=\mathbf{r}$.

iii) Din ii) rezultă

$$1 = P(\mathbf{R}=\mathbf{r}|E) + P(\mathbf{R}=\mathbf{n}|E) = \frac{1}{P(E)}\cdot\frac{27+12}{2000},$$

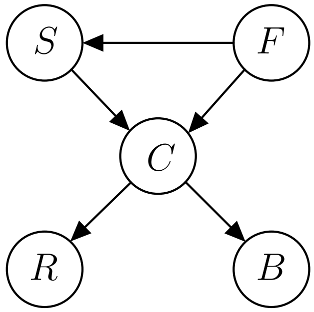
deci

$$P(E) = \frac{19,5}{1000} = 0,0195.$$

2. Considerăm următoarele variabile aleatoare care indică anumite situaţii (1=`da` şi 0=`nu`), pe care le aveţi în vedere pentru o persoană într-o seară:

- F indică dacă filmul care rulează la cinema este în premieră sau nu.
- S indică dacă biletul de intrare la film este scump sau nu.
- C indică dacă persoana vizionează filmul de la cinema sau nu.
- R indică dacă persoana ia cina la un restaurant sau nu.
- B indică dacă persoana bea un cocteil la un bar sau nu.

Variabilele aleatoare de mai sus depind unele de altele conform unei reţele Bayes cu probabilităţile condiţinate date mai jos.



$P(F=1)$	$P(F=0)$
0,8	0,2

S	$P(S=...\|F=1)$	$P(S=...\|F=0)$
1	0,9	0,6
0	0,1	0,4

C	$P(C=...\|S=1,F=1)$	$P(C=...\|S=1,F=0)$	$P(C=...\|S=0,F=1)$	$P(C=...\|S=0,F=0)$
1	0,6	0,2	0,9	0,4
0	0,4	0,8	0,1	0,6

R	$P(R=...\|C=1)$	$P(R=...\|C=0)$
1	0,3	0,5
0	0,7	0,5

B	$P(B=...\|C=1)$	$P(B=...\|C=0)$
1	0,5	0,8
0	0,5	0,2

Calculaţi probabilităţile următoarelor evenimente:
a) Persoana bea un cocteil la un bar, ştiind că nu vizionează filmul care rulează în premieră la cinema, biletul de intrare la film fiind scump.
b) Persoana vizionează un film care nu e în premieră la cinema.
c) Persoana ia cina la un restaurant.
d) Filmul care rulează la cinema este în premieră, ştiind că persoana ia cina la un restaurant.

R: a)

$$P(B=1|F=1,S=1,C=0)=P(B=1|C=0)=0,8.$$

b)

$$P(C=1,F=0)=P(C=1,F=0,S=1)+P(C=1,F=0,S=0)$$

$$=P(C=1|S=1,F=0)\cdot P(S=1|F=0)\cdot P(F=0)+P(C=1|S=0,F=0)\cdot P(S=0|F=0)\cdot P(F=0)$$

$$=0{,}2\cdot 0{,}6\cdot 0{,}2+0{,}4\cdot 0{,}4\cdot 0{,}2=0{,}056=5,6\%.$$

c)

$$P(R=1)=P(R=1|C=1)\cdot P(C=1)+P(R=1|C=0)\cdot P(C=0)$$

$$=0{,}3\cdot\sum_{i,j\in\{0,1\}}P(C=1,S=i,F=j)+0{,}5\cdot\sum_{i,j\in\{0,1\}}P(C=0,S=i,F=j)$$

$$=0{,}3\cdot\sum_{i,j\in\{0,1\}}P(C=1|S=i,F=j)\cdot P(S=i|F=j)\cdot P(F=j)$$

$$+0{,}5\cdot\sum_{i,j\in\{0,1\}}P(C=0|S=i,F=j)\cdot P(S=i|F=j)\cdot P(F=j)$$

$$=0{,}3\cdot(0{,}6\cdot 0{,}9\cdot 0{,}8+0{,}2\cdot 0{,}6\cdot 0{,}2+0{,}9\cdot 0{,}1\cdot 0{,}8+0{,}4\cdot 0{,}4\cdot 0{,}2)$$

$$+0{,}5\cdot(0{,}4\cdot 0{,}9\cdot 0{,}8+0{,}8\cdot 0{,}6\cdot 0{,}2+0{,}1\cdot 0{,}1\cdot 0{,}8+0{,}6\cdot 0{,}4\cdot 0{,}2)$$

$$=0{,}168+0{,}22=0{,}388=38{,}8\%.$$

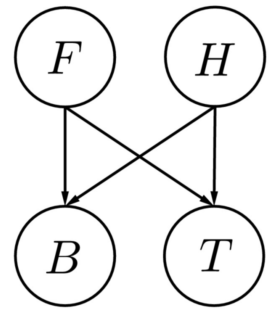
d)

$$P(F=1|R=1)=\frac{P(R=1,F=1)}{P(R=1)}=\frac{\sum_{i,j\in\{0,1\}}P(R=1,F=1,C=i,S=j)}{0{,}388}$$

$$= \frac{\sum_{i,j\in\{0,1\}} P(R=1|C=i)P(C=i|F=1,S=j)P(S=j|F=1)P(F=1)}{0{,}388}$$
$$= \frac{0{,}3\cdot 0{,}6\cdot 0{,}9\cdot 0{,}8 + 0{,}3\cdot 0{,}9\cdot 0{,}1\cdot 0{,}8 + 0{,}5\cdot 0{,}4\cdot 0{,}9\cdot 0{,}8 + 0{,}5\cdot 0{,}8\cdot 0{,}1\cdot 0{,}8}{0{,}388}$$
$$= \frac{0{,}3272}{0{,}388} \approx 84{,}3\%.$$

3. Următoarele variabile aleatoare indică decizia unui parior sportiv de a paria sau nu (`1=da`, `0=nu`) într-o zi pe anumite tipuri de meciuri: fotbal F, handbal H, baschet B şi tenis T. Pariorul ia deciziile conform reţelei Bayes alăturate, cu următoarele probabilităţi:

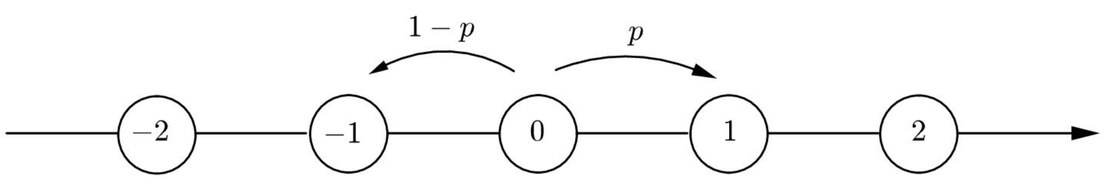
$P(F=1) = P(H=1) = 0{,}6$;
$P(B=1|F=1,H=1) = P(B=1|F=0,H=0) = 0{,}5$;
$P(T=1|F=1,H=1) = P(T=1|F=0,H=0) = 0{,}9$;
$P(B=1|F=0,H=1) = P(B=1|F=1,H=0) = 0{,}2$;
$P(T=1|F=0,H=1) = P(T=1|F=1,H=0) = 0{,}3$.



Calculaţi probabilităţile următoarelor evenimente:
a) Pariorul nu pariază pe niciun tip de meci.
b) Pariorul alege meciuri de fotbal şi baschet, ştiind că nu alege niciun meci de handbal.
c) Pariorul alege meciuri de baschet.

R: a) $P(F=0,H=0,B=0,T=0) = P(F=0)P(H=0)P(B=0|F=H=0)P(T=0|F=H=0) = 0{,}4\cdot 0{,}4\cdot 0{,}5\cdot 0{,}1 = 0{,}008$.
b) $P(F=1,B=1|H=0) = \frac{P(F=1,B=1,H=0)}{P(H=0)} = \frac{P(F=1)P(H=0)P(B=1|F=1,H=0)}{P(H=0)} = \frac{0{,}6\cdot 0{,}4\cdot 0{,}2}{0{,}4} = 0{,}12$.
c) $P(B=1) = P(B=1|F=1,H=1)P(F=1,H=1) + P(B=1|F=0,H=0)P(F=0,H=0)$
$+P(B=1|F=1,H=0)P(F=1,H=0) + P(B=1|F=0,H=1)P(F=0,H=1)$
$= 0{,}5\cdot 0{,}6\cdot 0{,}6 + 0{,}5\cdot 0{,}4\cdot 0{,}4 + 0{,}2\cdot 0{,}6\cdot 0{,}4 + 0{,}2\cdot 0{,}4\cdot 0{,}6 = 0{,}18 + 0{,}08 + 0{,}048 + 0{,}048 = 0{,}356$.

4. Un punct material se deplasează pe axa reală dintr-un nod spre un nod vecin, la fiecare pas, cu probabilitatea $p \in (0,1)$ la dreapta şi cu probabilitea $1-p$ la stânga. Nodurile sunt centrate în numerele întregi:



Fie X variabila aleatoare care indică poziţia finală a punctului material după $n \in \mathbb{N}$ paşi ai unei deplasări ce porneşte din nodul 0. Determinaţi distribuţia şi valoarea medie lui X.

R: Dacă Y_i reprezintă pasul i, atunci $Y_i \sim \begin{pmatrix} -1 & 1 \\ 1-p & p \end{pmatrix} \implies Y_i = 2X_i - 1$ cu $X_i \sim Bernoulli(p)$, $i \in \{1,\ldots,n\}$. $X = Y_1 + \ldots + Y_n = (2X_1 - 1) + \ldots + (2X_n - 1)$, $X_1 + \ldots + X_n \sim Bino(n,p) \implies X \sim \begin{pmatrix} 2k-n \\ C_n^k p^k (1-p)^{n-k} \end{pmatrix}_{k=\overline{0,n}}$ şi $E(X) = 2np - n$.

5. Trei prieteni decid cine va plăti nota la restaurant astfel: fiecare aruncă pe rând o monedă; plăteşte cel care obţine un simbol diferit de al celorlalţi doi; dacă toţi au obţinut acelaşi simbol, atunci se reia seria de trei aruncări ale monedei, ş.a.m.d. până când se decide cine plăteşte nota. Determinaţi valoarea medie a numărului de serii de aruncări până la (înainte de) seria de aruncări care va decide cine plăteşte nota.

R.: $p = 1 - P(\text{“toţi au obţinut acelaşi simbol”}) = 1 - \frac{2}{2^3} = \frac{3}{4}; \Rightarrow X \sim \text{Geom}(p) = Geom(\frac{3}{4})$,
$\Rightarrow E(X) = \frac{1-p}{p} = \frac{1}{3}$.

6. O persoană are două cutii de Tic Tac în buzunar, fiecare cu câte un număr iniţial de $n \in \mathbb{N}^*$ drajeuri. De fiecare dată când doreşte un drajeu, scoate aleatoar (şi independent de alegerile anterioare) din buzunar una dintre cutii, din care ia un drajeu. La un moment dat scoate din buzunar o cutie şi constată că e goală. Fie X numărul de drajeuri din cealaltă cutie. Determinaţi distribuţia lui X.

R.: Fixăm $k \in \{0, ..., n\}$; C_1 prima cutie; C_2 a doua cutie; notăm:
A_1 : când persoana observă că C_1 e goală, în C_2 sunt k drajeuri
A_2 : când persoana observă că C_2 e goală, în C_1 sunt k drajeuri.
$\Rightarrow A_1$ şi A_2 sunt disjuncte şi $P(A_1) = P(A_2)$, iar $P(X = k) = P(A_1 \cup A_2) = 2P(A_1)$.
Persoana observă că C_1 e goală, când alege C_1 pentru a $(n + 1)$ oară.
Definim “success”: persoana alege C_1; definim “insucces”: persoana alege C_2.

$$\Longrightarrow P(\text{“succes'}) = P(\text{“insucces”}) = 1/2$$

A_1 = obţinem al $(n + 1)$-lea “succes” după $n - k$ “insuccese” (în $n + (n - k)$ repetări, se obţin $n - k$ “insuccese”, la a $n + (n - k) + 1$ repetare se obţine “succes”)

$$\Longrightarrow P(A_1) = C_{n+(n-k)}^{n-k} \left(\frac{1}{2}\right)^n \left(\frac{1}{2}\right)^{n-k} \frac{1}{2} = \frac{1}{2} C_{2n-k}^{n} \cdot \left(\frac{1}{2}\right)^{2n-k}.$$

$$\Rightarrow P(X = k) = C_{2n-k}^{n} \left(\frac{1}{2}\right)^{2n-k}.$$

Concluzie: $X \sim \begin{pmatrix} k \\ C_{2n-k}^{n} \frac{1}{2^{2n-k}} \end{pmatrix}_{k=\overline{0,n}}$.