Seminarul 4

• **Modelul urnei cu bile de 2 culori şi bilă nereturnată:** fie $n_1, n_2, n \in \mathbb{N}$ cu $n \leq n_1 + n_2$ şi fie $k \in \mathbb{N}$ astfel încât $k \leq n_1$ şi $n - k \leq n_2$; considerând o urnă, care are iniţial n_1 bile albe şi n_2 bile negre, avem

$$\begin{aligned} p(k;n) &= \text{probabilitatea de a obţine } k \text{ bile albe din } n \text{ extrageri } \textit{fără returnarea} \text{ bilei extrase,} \\ &\quad \text{în care ordinea de extragere a bilelor nu contează} \\ &= \frac{C_{n_1}^k \cdot C_{n_2}^{n-k}}{C_{n_1+n_2}^n}. \end{aligned}$$

▷ Acest model corespunde **distribuţiei hipergeometrice**.

1. Dintr-un set de 52 de cărţi de joc se extrag aleator, pe rând, fără returnare, 13 cărţi (*bridge hand*). Calculaţi probabilităţile următoarelor evenimente:

a) A: nu s-a extras nicio treflă;

b) B: s-au obţinut 5 inimi;

c) C: s-a obţinut cel mult un as.

R: $P(A) = \frac{C_{39}^{13} \cdot C_{13}^0}{C_{52}^{13}}$; $P(B) = \frac{C_{13}^5 \cdot C_{39}^8}{C_{52}^{13}}$;

$P(C) = P(\text{nu s-a extras niciun as}) + P(\text{s-a extras exact un as}) = \frac{C_{48}^{13} \cdot C_4^0}{C_{52}^{13}} + \frac{C_{48}^{12} \cdot C_4^1}{C_{52}^{13}}$.

• **Modelul urnei cu r culori şi bilă nereturnată:** fie n_i =numărul iniţial de bile cu culoarea i din urnă, $i = \overline{1,r}$;

$$\begin{aligned} p(k_1, \ldots, k_r; n) &= \text{probabilitatea de a obţine } k_i \text{ bile cu culoarea } i,\ i = \overline{1,r}, \\ &\quad \text{din } n = k_1 + \ldots + k_r \text{ extrageri } \textit{fără returnarea} \text{ bilei extrase,} \\ &\quad \text{în care ordinea de extragere a bilelor de diverse culori nu contează} \\ &= \frac{C_{n_1}^{k_1} \cdot \ldots \cdot C_{n_r}^{k_r}}{C_{n_1+\ldots+n_r}^n}. \end{aligned}$$

▷ Cazul $r = 2$ corespunde **distribuţiei hipergeometrice**.

Observaţie: Extragerea fără returnare (engl. *sampling without replacement*) este folosită în **metoda validării încrucişate** (engl. *k-fold cross validation*): În cazul validării încrucişate eşantionul original de date este împărțit aleatoriu în k sub-eşantioane de dimensiuni egale. Din cele k sub-eşantioane, un singur sub-eşantion este folosit ca date de validare pentru testarea modelului, iar celelalte $k-1$ sub-eşantioane sunt utilizate ca date de antrenament. Procesul de validare încrucişată se repetă apoi de k ori, fiecare dintre cele k sub-eşantioane fiind utilizat exact o dată ca date de validare.

2. O echipă formată din 4 cercetători este aleasă aleator dintr-un grup de 4 matematicieni, 3 informaticieni şi 5 fizicieni. Care este probabilitatea ca echipa să fie formată din 2 matematicieni, 1 informatician şi 1 fizician?

R: $\frac{C_4^2 C_3^1 C_5^1}{C_{12}^4}$.

3. ***Clasificarea naivă Bayes***

Clasificatorii bayesieni naivi sunt o familie de clasificatori probabilistici simpli, bazaţi pe aplicarea formulei lui Bayes cu ipoteze "naive" de independenţă condiţionată între atribute (engl. *features*), cunoscând clasificarea. În aplicaţii practice pentru modelele bayesiene naive se foloseşte *metoda probabilităţii maxime*. Noţiunea folosită în acest context este condiţional independenţa între v.a.

Def. 1 *Fie U, X, Y, Z v.a. discrete, care iau valori în mulţimile $\mathcal{U}$, $\mathcal{X}$, $\mathcal{Y}, \mathcal{Z}$. V.a. U, X, Y sunt* ***condiţional independente****, cunoscând (ştiind) v.a. Z, dacă pentru fiecare $u \in \mathcal{U}, x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}$ are loc*

$$P(U = u, X = x, Y = y | Z = z) = P(U = u | Z = z) P(X = x | Z = z) P(Y = y | Z = z) .$$

Considerăm următoarea problemă de *clasificare naivă Bayes* a unor restaurante (**R**), în
- *clasele*: **r**ecomandat sau **n**erecomandat,

în funcţie de următoarele *atribute* cu valorile lor posibile:
- cost (C): ieftin, mediu, scump;
- timp de aşteptare (T): puţin, mediu, îndelungat;
- mâncare (M): fadă, acceptabilă, bună, delicioasă.

R, C, T, M sunt variabelele aleatoare (categoriale) şi **r**, **n**, i, m, s, p, m, $\hat{i}$, 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 $P(E)$.

R.:

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

M	$\mathbf{R}=\mathbf{r}$	$\mathbf{R}=\mathbf{n}$	$P(M=...\|\mathbf{R}=\mathbf{r})$	$P(M=...\|\mathbf{R}=\mathbf{n})$
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

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

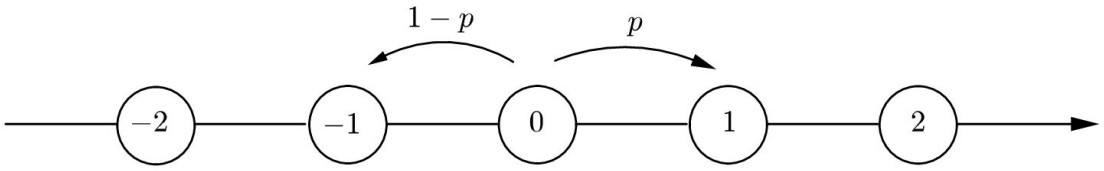
4. Un mesaj este transmis printr-un canal de comunicare cu perturbări. Probabilitatea ca mesajul să fie recepţionat este 10%. Dacă mesajul nu este recepţionat, atunci se reia transmisia mesajului, independent de transmisiile anterioare. Fie X variabila aleatoare care indică numărul de transmisii până la prima transmisie în care este recepţionat mesajul. Determinaţi valoarea medie a lui X.

R: Observăm că $X \sim \text{Geo}(p)$, $p = \frac{1}{10}$. Pe baza criteriului raportului, seria cu termeni pozitivi $\sum_{k=0}^{\infty} kp(1-p)^k$ este convergentă.

$$E(X) = \sum_{k=0}^{\infty} kp(1-p)^k = (1-p)\sum_{k=1}^{\infty} kp(1-p)^{k-1}$$
$$\overset{k=j+1}{=\!=} (1-p)\sum_{j=0}^{\infty}(j+1)p(1-p)^j = (1-p)\sum_{j=0}^{\infty} jp(1-p)^j + (1-p)\sum_{j=0}^{\infty} p(1-p)^j$$
$$= (1-p)E(X) + (1-p) \implies E(X) = \frac{1-p}{p}$$

$\implies E(X) = \frac{\frac{9}{10}}{\frac{1}{10}} = 9$, deci vor fi în medie 9 transmisii eşuate până la recepţionarea mesajului.

5. 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 $E(X) = 2np - n$.

6. Considerăm vectorul aleatoar discret (X, Y) cu distribuţia dată sub formă tabelară:

X \ Y	-2	1	2
1	0,2	0,1	0,2
2	0,1	0,1	0,3

a) Să se determine distribuţiile de probabilitate ale variabilelor aleatoare X şi Y.
b) Calculaţi probabilitatea ca $|X - Y| = 1$, ştiind că $Y > 0$.
c) Sunt evenimentele $X = 2$ şi $Y = 1$ independente?
d) Sunt variabilele aleatoare X şi Y independente?
e) Sunt evenimentele $X = 1$ şi $Y = 1$ condiţional independente, cunoscând $X + Y = 2$?
f) Este variabila aleatoare X condiţional independentă de Y, cunoscând $X + Y$?
g) Calculaţi valoarea medie a variabilei aleatoare $2X + Y^2$.

R: a) $X \sim \begin{pmatrix} 1 & 2 \\ 0{,}5 & 0{,}5 \end{pmatrix}, Y \sim \begin{pmatrix} -2 & 1 & 2 \\ 0{,}3 & 0{,}2 & 0{,}5 \end{pmatrix}$.
b) $P(|X-Y|=1 \big| Y>0) = \frac{P(|X-Y|=1, Y>0)}{P(Y>0)} = \frac{P(X=1,Y=2)+P(X=2,Y=1)}{P(Y>0)} = \frac{0{,}3}{0{,}7} = \frac{3}{7}$.
c) $P(X=2, Y=1) = 0{,}1 = 0{,}5 \cdot 0{,}2 = P(X=2) \cdot P(Y=1) \implies X=2$ şi $Y=1$ sunt independente.
d) $P(X=2, Y=2) = 0{,}3 \neq 0{,}25 = 0{,}5 \cdot 0{,}5 = P(X=2) \cdot P(Y=2) \implies X$ şi Y nu sunt independente.
e) $P(X=1, Y=1 | X+Y=2) = 1 = P(X=1|X+Y=2) \cdot P(Y=1|X+Y=2) \implies X=1$ şi $Y=1$ sunt condiţional independente, cunoscând $X+Y=2$.
f) $P(X=1, Y=2|X+Y=3) = \frac{P(X=1,Y=2)}{P(X+Y=3)} = \frac{0{,}2}{0{,}3} \neq \frac{0{,}2}{0{,}3} \cdot \frac{0{,}2}{0{,}3} = P(X=1|X+Y=3) \cdot P(Y=2|X+Y=3) \implies X$ şi Y nu sunt condiţional independente, cunoscând $X+Y$.
g) $E(2X+Y^2) = 2E(X) + E(Y^2) = 2(1 \cdot 0{,}5 + 2 \cdot 0{,}5) + (-2)^2 \cdot 0{,}3 + 1^2 \cdot 0{,}2 + 2^2 \cdot 0{,}5 = 6{,}4$.

7. O monedă este aruncată de 10 ori. Fie X variabila aleatoare care indică diferenţa dintre numărul de capete şi numărul de pajuri obţinute. Determinaţi:
i) distribuţia de probabilitate a lui X;
ii) valoarea medie a lui X.
R: i) Dacă C şi P indică numărul de capete, respectiv de pajuri, atunci $C, P \sim \text{Bino}(10, \frac{1}{2})$, $P = 10 - C$ şi $X = C - P = 2C - 10 \implies X \sim \begin{pmatrix} 2k-10 \\ C_{10}^k \frac{1}{2^{10}} \end{pmatrix}_{k=\overline{0,10}}$.
ii) $E(X) = E(C-P) = E(C) - E(P) = 0$, deoarece C şi P au aceeaşi distribuţie.