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.

• **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 împărţit 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?

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:

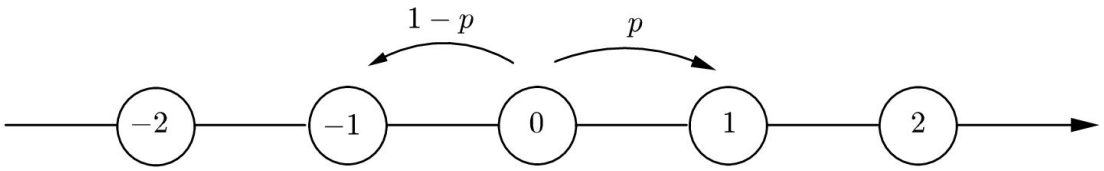
	Cost	*Timp de aşteptare*	*Mâncare*	**Restaurant**
1	mediu	îndelungat	acceptabilă	**nerecomandat**
2	scump	puţin	bună	**recomandat**
3	ieftin	îndelungat	delicioasă	**recomandat**
4	mediu	puţin	bună	**recomandat**
5	ieftin	mediu	acceptabilă	**nerecomandat**
6	ieftin	puţin	fadă	**nerecomandat**
7	mediu	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 $P(E)$.

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.

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.

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 $2X + Y^2$.

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.