

Seminarul 4

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

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

▷ Acest model corespunde **distribuției hipergeometrice**.

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

- a) A : nu s-a extras nicio treflă;
- b) B : s-au obținut 5 inimi;
- c) C : s-a obținut cel mult un as.

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

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

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

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

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

Observație: Extragerea fără returnare (engl. *sampling without replacement*) este folosită în **metoda validării încrucișate** (engl. *k-fold cross validation*): În cazul validării încrucișate eșantionul original de date este împărțit aleatoriu în k sub-eșantioane de dimensiuni egale. Din cele k sub-eșantioane, un singur sub-eșantion este folosit ca date de validare pentru testarea modelului, iar celelalte $k - 1$ sub-eșantioane sunt utilizate ca date de antrenament. Procesul de validare încrucișată se repetă de k ori, fiecare dintre cele k sub-eșantioane 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 și 5 fizicieni. Care este probabilitatea ca echipa să fie formată din 2 matematicieni, 1 informatician și 1 fizician?

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

3. Clasificarea naivă Bayes

Clasificatorii bayesieni naivi sunt o familie de clasificatori probabilistici simpli, bazați pe aplicarea formulei lui Bayes cu ipoteze “naive” de **independență condiționată între atribute** (engl. *features*), cunoscând **clasificarea**. În aplicații practice pentru modelele bayesiene naive se folosește *metoda probabilității maxime*. Noțiunea folosită în acest context este condițional independența între v.a.

Def. 1 Fie U, X, Y, Z v.a. discrete, care iau valori în mulțimile $\mathcal{U}, \mathcal{X}, \mathcal{Y}, \mathcal{Z}$. V.a. U, X, Y sunt **condițional independente**, cunoscând (știind) 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*: recomandat sau nerecomandat,
- în funcție de următoarele *attribute* cu valorile lor posibile:
- cost (**C**): ieftin, mediu, scump;
- timp de așteptare (**T**): puțin, mediu, îndelungat;
- mâncare (**M**): fadă, acceptabilă, bună, delicioasă.

R, **C**, **T**, **M** sunt variabilele aleatoare (catoriale) și **r**, **n**, *i*, *m*, *s*, *p*, *m*, *î*, *f*, *a*, *b*, *d* valorile de mai sus, în ordinea în care sunt menționate.

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

	<i>Cost</i>	<i>Timp de așteptare</i>	<i>Mâncare</i>	Restaurant
1	mediu	îndelungat	acceptabilă	nerecomandat
2	scump	puțin	bună	recomandat
3	ieftin	îndelungat	delicioasă	recomandat
4	mediu	puțin	bună	recomandat
5	ieftin	mediu	acceptabilă	nerecomandat
6	ieftin	puțin	fadă	nerecomandat
7	mediu	puțin	acceptabilă	nerecomandat
8	mediu	mediu	delicioasă	recomandat
9	scump	puțin	delicioasă	recomandat
10	ieftin	îndelungat	bună	nerecomandat
11	scump	puțin	acceptabilă	nerecomandat
12	mediu	mediu	bună	recomandat
13	mediu	îndelungat	fadă	nerecomandat
14	scump	mediu	delicioasă	recomandat
15	ieftin	mediu	fadă	nerecomandat
16	mediu	puțin	delicioasă	recomandat
17	ieftin	puțin	acceptabilă	recomandat
18	scump	îndelungat	bună	nerecomandat
19	ieftin	puțin	fadă	recomandat
20	scump	îndelungat	delicioasă	nerecomandat

i) Folosind datele din tabel, determinați probabilitățile claselor și probabilitățile condiționate ale atributelor, știind clasa.

ii) Considerăm evenimentul dat de *vectorul de attribute*: $E = (C = s) \cap (T = m) \cap (M = b)$. Alegeți o clasă pentru E , stabilind care din următoarele probabilități este mai mare: $P(\mathbf{R} = \mathbf{r} | E)$ sau $P(\mathbf{R} = \mathbf{n} | E)$.

iii) Determinați $P(E)$.

R.:

i)

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

C	$\mathbf{R} = \mathbf{r}$	$\mathbf{R} = \mathbf{n}$	$P(C = \dots \mathbf{R} = \mathbf{r})$	$P(C = \dots \mathbf{R} = \mathbf{n})$
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	$\mathbf{R} = \mathbf{r}$	$\mathbf{R} = \mathbf{n}$	$P(T = \dots \mathbf{R} = \mathbf{r})$	$P(T = \dots \mathbf{R} = \mathbf{n})$
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	$\mathbf{R} = \mathbf{r}$	$\mathbf{R} = \mathbf{n}$	$P(M = \dots \mathbf{R} = \mathbf{r})$	$P(M = \dots \mathbf{R} = \mathbf{n})$
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 și a ipotezei de independență condiționată, deducem că:

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

și

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

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 recepționat este 10%. Dacă mesajul nu este recepționat, atunci se reia transmisia mesajului, independent de transmisiile anterioare. Fie X variabila aleatoare care indică numărul de transmisi până la prima transmisie în care este recepționat mesajul. Determinați valoarea medie a lui X .

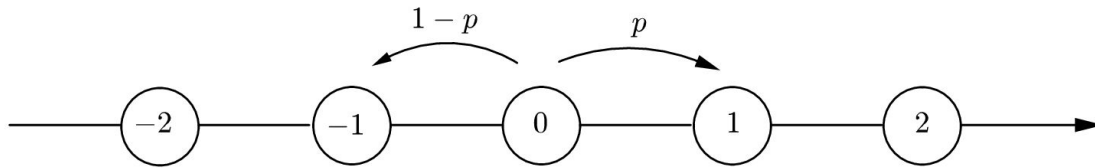
R: Observăm că $X \sim \text{Geo}(p)$, $p = \frac{1}{10}$. Pe baza criteriului raportului, seria cu termeni pozitivi

$\sum_{k=0}^{\infty} kp(1-p)^k$ este convergentă.

$$\begin{aligned} E(X) &= \sum_{k=0}^{\infty} kp(1-p)^k = (1-p) \sum_{k=1}^{\infty} kp(1-p)^{k-1} \\ &\stackrel{k=j+1}{=} (1-p) \sum_{j=0}^{\infty} (j+1)p(1-p)^j = (1-p) \sum_{j=0}^{\infty} jp(1-p)^j + (1-p) \sum_{j=0}^{\infty} p(1-p)^j \\ &= (1-p)E(X) + (1-p) \implies E(X) = \frac{1-p}{p} \end{aligned}$$

$\implies E(X) = \frac{\frac{9}{10}}{\frac{1}{10}} = 9$, 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 probabilitatea $1-p$ la stânga. Nodurile sunt centrate în numerele întregi:



Fie X variabila aleatoare care indică poziția finală a punctului material după $n \in \mathbb{N}$ pași ai unei deplasări ce pornește din nodul 0. Determinați distribuția și 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 $X_i \sim \text{Bernoulli}(p)$, $i \in \{1, \dots, n\}$. $X = Y_1 + \dots + Y_n = (2X_1 - 1) + \dots + (2X_n - 1)$, $X_1 + \dots + X_n \sim \text{Bino}(n, p) \implies X \sim \left(\begin{matrix} 2k - n \\ C_n^k p^k (1-p)^{n-k} \end{matrix} \right)_{k=0, \overline{n}}$ și $E(X) = 2np - n$.

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

$\begin{matrix} Y \\ X \end{matrix}$	-2	1	2
1	0,2	0,1	0,2
2	0,1	0,1	0,3

- Să se determine distribuțiile de probabilitate ale variabilelor aleatoare X și Y .
- Calculați probabilitatea ca $|X - Y| = 1$, știind că $Y > 0$.
- Sunt evenimentele $X = 2$ și $Y = 1$ independente?
- Sunt variabilele aleatoare X și Y independente?
- Sunt evenimentele $X = 1$ și $Y = 1$ condițional independente, cunoscând $X + Y = 2$?
- Este variabila aleatoare X condițional independentă de Y , cunoscând $X + Y$?
- Calculați 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) $P(|X - Y| = 1 | Y > 0) = \frac{P(|X - Y| = 1, Y > 0)}{P(Y > 0)} = \frac{P(X=1, Y=2) + P(X=2, Y=1)}{P(Y > 0)} = \frac{0,3}{0,7} = \frac{3}{7}.$

c) $P(X = 2, Y = 1) = 0,1 = 0,5 \cdot 0,2 = P(X = 2) \cdot P(Y = 1) \implies X = 2$ și $Y = 1$ sunt independente.

d) $P(X = 2, Y = 2) = 0,3 \neq 0,25 = 0,5 \cdot 0,5 = P(X = 2) \cdot P(Y = 2) \implies X$ și Y nu sunt independente.

e) $P(X = 1, Y = 1 | X + Y = 2) = 1 = P(X = 1 | X + Y = 2) \cdot P(Y = 1 | X + Y = 2) \implies X = 1$ și $Y = 1$ sunt condițional independente, cunoscând $X + Y = 2$.

f) $P(X = 1, Y = 2 | X + Y = 3) = \frac{P(X=1, Y=2)}{P(X+Y=3)} = \frac{0,2}{0,3} \neq \frac{0,2}{0,3} \cdot \frac{0,2}{0,3} = P(X = 1 | X + Y = 3) \cdot P(Y = 2 | X + Y = 3) \implies X$ și Y nu sunt condițional 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 diferența dintre numărul de capete și numărul de pajuri obținute. Determinați:

i) distribuția de probabilitate a lui X ;

ii) valoarea medie a lui X .

R: i) Dacă C și P indică numărul de capete, respectiv de pajuri, atunci $C, P \sim \text{Bino}(10, \frac{1}{2})$, $P = 10 - C$

și $X = C - P = 2C - 10 \implies X \sim \left(\begin{matrix} 2k - 10 \\ C_{10}^k \frac{1}{2^{10}} \end{matrix} \right)_{k=\overline{0,10}}.$

ii) $E(X) = E(C - P) = E(C) - E(P) = 0$, deoarece C și P au aceeași distribuție.