Seminarul 4

Exemplu de clasificare naivă Bayes

Se doreşte *clasificarea* traficului pe un anumit bulevard, în *clasele*: aglomerat sau relaxat, în funcţie de următoarele *atribute* cu valorile lor posibile:
• vreme: ploaie, zăpadă, senin;
• temperatură: scăzută, medie, ridicată;
• timp: dimineaţă, amiază, seară, noapte.

Fie $\mathbf{T}$, V, Te, Ti variabelele aleatoare corespunzătoare şi $\mathbf{a}$, $\mathbf{r}$, p, z, s, sc, m, ri, d, am, se şi n valorile de mai sus, în ordinarea în care sunt menţionate.

În continuare, facem următoarea presupunere *naivă*: atributele sunt **condiţional independente**, dacă se dă clasificarea.

De exemplu, avem:

$$P(V = p, Te = ri, Ti = d|\mathbf{T} = \mathbf{a}) = P(V = p|\mathbf{T} = \mathbf{a})P(Te = ri|\mathbf{T} = \mathbf{a})P(Ti = d|\mathbf{T} = \mathbf{a}),$$

$$P(Te = ri, Ti = d|\mathbf{T} = \mathbf{a}) = P(Te = ri|\mathbf{T} = \mathbf{a})P(Ti = d|\mathbf{T} = \mathbf{a}),$$

unde presupunem că $P(\mathbf{T} = \mathbf{a}) \neq 0$. Din acest exemplu se poate deduce că:

$$P(V = p|Te = ri, Ti = d, \mathbf{T} = \mathbf{a}) = P(V = p|\mathbf{T} = \mathbf{a}),$$

dacă $P(Te = ri, Ti = d, \mathbf{T} = \mathbf{a}) \neq 0$, altfel spus: probabilitatea să plouă, ştiind că traficul este aglomerat, nu depinde de temperatură sau timp. Într-adevăr, avem:

$$P(V = p|Te = ri, Ti = d, \mathbf{T} = \mathbf{a}) = \frac{P(V = p, Te = ri, Ti = d, \mathbf{T} = \mathbf{a})}{P(Te = ri, Ti = d, \mathbf{T} = \mathbf{a})}$$

$$= \frac{P(V = p, Te = ri, Ti = d|\mathbf{T} = \mathbf{a})P(\mathbf{T} = \mathbf{a})}{P(Te = ri, Ti = d|\mathbf{T} = \mathbf{a})P(\mathbf{T} = \mathbf{a})} = \frac{P(V = p|\mathbf{T} = \mathbf{a})P(Te = ri|\mathbf{T} = \mathbf{a})P(Ti = d|\mathbf{T} = \mathbf{a})}{P(Te = ri|\mathbf{T} = \mathbf{a})P(Ti = d|\mathbf{T} = \mathbf{a})}.$$

Considerăm următorul *tabel de date* obţinute în urma unor observaţii pe bulevard:

	Vreme	*Temperatură*	*Timp*	**Trafic**
1	ploaie	ridicată	amiază	**aglomerat**
2	zăpadă	scăzută	seară	**aglomerat**
3	senin	ridicată	noapte	**relaxat**
4	ploaie	scăzută	seară	**aglomerat**
5	senin	medie	amiază	**relaxat**
6	senin	scăzută	dimineaţă	**aglomerat**
7	ploaie	scăzută	amiază	**aglomerat**
8	ploaie	medie	noapte	**relaxat**
9	zăpadă	scăzută	noapte	**relaxat**
10	senin	ridicată	seară	**relaxat**
11	zăpadă	scăzută	amiază	**aglomerat**
12	ploaie	medie	seară	**aglomerat**
13	ploaie	ridicată	dimineaţă	**aglomerat**
14	zăpadă	medie	noapte	**relaxat**
15	senin	medie	dimineaţă	**relaxat**
16	ploaie	scăzută	noapte	**aglomerat**

i) Folosind datele din tabel, determinaţi probabilităţile claselor şi probabilităţile condiţionate ale atributelor, ştiind clasa.
ii) Considerăm evenimentul următor, denumit *vector de atribute*: $E = (V = p) \cap (Te = m) \cap (Ti = am)$. Alegeţi o clasă pentru E, stabilind care din următoarele probabilităţi este mai mare: $P(\mathbf{T} = \mathbf{a}|E)$ sau $P(\mathbf{T} = \mathbf{r}|E)$.

Rezolvare:

i)

$\mathbf{T} = \mathbf{a}$	$\mathbf{T} = \mathbf{r}$	$P(\mathbf{T} = \mathbf{a})$	$P(\mathbf{T} = \mathbf{r})$
9	7	$\frac{9}{16}$	$\frac{7}{16}$

V	$\mathbf{T} = \mathbf{a}$	$\mathbf{T} = \mathbf{r}$	$P(V = ...\|\mathbf{T} = \mathbf{a})$	$P(V = ...\|\mathbf{T} = \mathbf{r})$
p	6	1	$\frac{6}{9}$	$\frac{1}{7}$
z	2	2	$\frac{2}{9}$	$\frac{2}{7}$
s	1	4	$\frac{1}{9}$	$\frac{4}{7}$

Te	$\mathbf{T} = \mathbf{a}$	$\mathbf{T} = \mathbf{r}$	$P(Te = ...\|\mathbf{T} = \mathbf{a})$	$P(Te = ...\|\mathbf{T} = \mathbf{r})$
sc	6	1	$\frac{6}{9}$	$\frac{1}{7}$
m	1	4	$\frac{1}{9}$	$\frac{4}{7}$
ri	2	2	$\frac{2}{9}$	$\frac{2}{7}$

Ti	$\mathbf{T} = \mathbf{a}$	$\mathbf{T} = \mathbf{r}$	$P(Ti = ...\|\mathbf{T} = \mathbf{a})$	$P(Ti = ...\|\mathbf{T} = \mathbf{r})$
d	2	1	$\frac{2}{9}$	$\frac{1}{7}$
am	3	1	$\frac{3}{9}$	$\frac{1}{7}$
se	3	1	$\frac{3}{9}$	$\frac{1}{7}$
n	1	4	$\frac{1}{9}$	$\frac{4}{7}$

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

$$P(\mathbf{T} = \mathbf{a}|E) = \frac{P(E|\mathbf{T} = \mathbf{a})P(\mathbf{T} = \mathbf{a})}{P(E)} = \frac{P(V = p, Te = m, Ti = am|\mathbf{T} = \mathbf{a})P(\mathbf{T} = \mathbf{a})}{P(E)}$$

$$= \frac{P(V = p|\mathbf{T} = \mathbf{a})P(Te = m|\mathbf{T} = \mathbf{a})P(Ti = am|\mathbf{T} = \mathbf{a})P(\mathbf{T} = \mathbf{a})}{P(E)} = \frac{\frac{6}{9} \cdot \frac{1}{9} \cdot \frac{3}{9} \cdot \frac{9}{16}}{P(E)} = \frac{1}{P(E)} \cdot \frac{1}{72}$$

şi

$$P(\mathbf{T} = \mathbf{r}|E) = \frac{P(E|\mathbf{T} = \mathbf{r})P(\mathbf{T} = \mathbf{r})}{P(E)} = \frac{P(V = p, Te = m, Ti = am|\mathbf{T} = \mathbf{r})P(\mathbf{T} = \mathbf{r})}{P(E)}$$

$$= \frac{P(V = p|\mathbf{T} = \mathbf{r})P(Te = m|\mathbf{T} = \mathbf{r})P(Ti = am|\mathbf{T} = \mathbf{r})P(\mathbf{T} = \mathbf{r})}{P(E)} = \frac{\frac{1}{7} \cdot \frac{4}{7} \cdot \frac{1}{7} \cdot \frac{7}{16}}{P(E)} = \frac{1}{P(E)} \cdot \frac{1}{196}.$$

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

Probleme cu variabile aleatoare discrete şi caracteristici numerice

1. Cinci cartonaşe sunt numerotate cu 2, 4, 6, 8, respectiv 10. Se aleg aleator trei cartonaşe, repunând de fiecare dată înapoi în teanc cartonaşul ales. Ştiind că suma numerelor extrase este 12 şi notând cu X variabila aleatoare care indică de câte ori a fost extras numărul 2, determinaţi:
a) funcţia de repartiţie a lui X;
b) valoarea medie a lui X;
c) varianţa lui X.

R: a) $X \sim \begin{pmatrix} 0 & 1 & 2 \\ \frac{1}{10} & \frac{6}{10} & \frac{3}{10} \end{pmatrix}$, deci $F(x) = P(X \le x) = \begin{cases} 0, & x < 0, \\ \frac{1}{10}, & 0 \le x < 1, \\ \frac{7}{10}, & 1 \le x < 2, \\ 1, & x \ge 2. \end{cases}$ b) $E(X) = \frac{6}{5}$. c) $V(X) = E(X^2) - E^2(X) = \frac{18}{10} - \frac{36}{25} = \frac{9}{25}$.

2. Un sistem electronic are $n \in \mathbb{N}^*$ componente care funcţionează independent unele de altele. Fiecare componentă funcţionează cu probabilitatea $p \in (0,1)$. Fie X variabila aleatoare care indică numărul de componente funcţionale ale sistemului. Determinaţi distribuţia lui X şi apoi calculaţi valoarea sa medie şi varianţa sa.

Consecinţă: Dacă $X \sim Bino(n,p)$, atunci $E(X) = np$ şi $V(X) = np(1-p)$.

R: $E(X) = E(X_1) + \cdots + E(X_n) = np$ şi $V(X) = V(X_1) + \cdots + V(X_n) = np(1-p)$, unde $X_i \sim$ Bernoulli(p) indică funcţionarea componentei i, $i = \overline{1,n}$.

3. Un mesaj este transmis printr-un canal de comunicaţie cu perturbări. Probabilitatea ca mesajul să fie recepţionat complet este $p \in (0,1)$. Dacă mesajul nu este recepţionat complet, atunci se reia transmisia mesajului, independent de transmisiile anterioare. Fie X variabila aleatoare care indică numărul de transmisii până la prima transmisie în care mesajul este recepţionat complet. Determinaţi valoarea medie şi varianţa lui X.

Consecinţă: Dacă $X \sim Geo(p)$, atunci $E(X) = \frac{1-p}{p}$ şi $V(X) = \frac{1-p}{p^2}$.

R: $E(X) = \sum_{k=0}^{\infty} kp(1-p)^k = (1-p)\sum_{k=1}^{\infty}(k-1)p(1-p)^{k-1} + p\sum_{k=1}^{\infty}(1-p)^k = (1-p)E(X) + 1 - p$, deci $E(X) = \frac{1-p}{p}$.
$V(X) = E(X^2) - E^2(X) = E(X^2) - \frac{(1-p)^2}{p^2}$; $E(X^2) = \sum_{k=0}^{\infty} k^2 p(1-p)^k = (1-p)\sum_{k=1}^{\infty}(k^2-2k+1)p(1-p)^{k-1} + 2\sum_{k=1}^{\infty} kp(1-p)^k - p\sum_{k=1}^{\infty}(1-p)^k = (1-p)E(X^2) + 2E(X) - (1-p) = (1-p)E(X^2) + \frac{(1-p)(2-p)}{p}$, deci $E(X^2) = \frac{(1-p)(2-p)}{p^2}$ şi $V(X) = \frac{(1-p)(2-p)}{p^2} - \frac{(1-p)^2}{p^2} = \frac{1-p}{p^2}$.

4. O monedă este aruncată de n ori ($n \in \mathbb{N}^*$). Fie X variabila aleatoare care indică diferenţa dintre numărul de capete şi numărul de pajuri obţinute. Determinaţi:
i) distribuţia lui X;
ii) valoarea medie a lui X;
iii) varianţa lui X.

R: Dacă C indică numărul de capete, atunci $C \sim$ Bino$(n, \frac{1}{2})$ şi $X = C - (n - C) = 2C - n$. $E(X) = E(2C - n) = 2E(C) - n = 2 \cdot \frac{n}{2} - n = 0$. $V(X) = 4V(C) = 4 \cdot \frac{n}{4} = n$.

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 se decide cine plăteşte nota. Determinaţi valoarea medie a numărului de serii de aruncări până la seria de aruncări care va decide cine plăteşte nota.

R: $X \sim$ Geom$(\frac{3}{4})$, $E(X) = \frac{1}{3}$.

6. O persoană are două cutii de Tic Tac în buzunar. De fiecare dată când doreşte un drajeu, scoate aleatoar 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, ştiind că fiecare cutie conţinea iniţial câte n drajeuri ($n \in \mathbb{N}$).

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

7. n bile sunt puse aleator în k cutii ($k, n \in \mathbb{N}^*$), numerotate de la 1 la k, astfel: pentru fiecare bilă se alege aleator o cutie din cele k astfel încât probabilitatea de a alege cutia i să fie p_i, $i = \overline{1,k}$, unde $p_i \in (0,1)$, $i = \overline{1,k}$, sunt fixate astfel încât $p_1 + \ldots + p_k = 1$. Determinaţi valoarea medie a numărului de cutii care conţin exact m bile, $m \in \mathbb{N}$, $m \leq n$.

R: Fie $X_i = \begin{cases} 1, & \text{cutia i conţine exact } m \text{ bile,} \\ 0, & \text{altfel,} \end{cases}$ $i = \overline{1,k}$. Avem $P(X_i = 1) = C_n^m p_i^m (1-p_i)^{n-m}$, unde am folosit modelul urnei cu bila întoarsă cu 2 stări şi n extrageri. Deci $E(X) = E(\sum_{i=1}^{k} X_i) = \sum_{i=1}^{k} C_n^m p_i^m (1-p_i)^{n-m}$.

Aplicaţii ale valorii medii în teoria grafurilor

1. Fie X o variabilă aleatoare discretă care ia un număr finit de valori. Atunci $P(X \geq E(X)) \neq 0$ şi $P(X \leq E(X)) \neq 0$.

R: Fie $\{x_1, \ldots, x_n\}$ mulţimea de valori pe care le poate lua X cu probabilităţi nenule. Dacă am presupune că $P(X \geq E(X)) = 0$, atunci am avea $x_i < E(X)$, pentru fiecare $i \in \{1, \ldots, n\}$, şi astfel am obţine următoarea contradicţie:

$$E(X) = \sum_{i=1}^{n} x_i \cdot P(X = x_i) < \sum_{i=1}^{n} E(X) \cdot P(X = x_i) = E(X).$$

Similar se demonstrează cealaltă relaţie.

Definiţii: În următoarele probleme vom folosi următoarele definiţii:
- $G = (V, E)$ este un graf neorientat cu V mulţimea nodurilor/vârfurilor şi E mulţimea muchiilor;
- $n = |V| \geq 2$ şi $m = |E| \geq 1$;
- G este complet, dacă oricare două noduri sunt adiacente;
- o muchie este bicoloră, dacă nodurile ei sunt colorate diferit;
- un subgraf este monocolor, dacă toate muchiile sale sunt colorate la fel;
- $I \subseteq V$ este independentă, dacă nu există două noduri în I care să fie adiacente.

2. Demonstraţi că există o colorare cu roşu şi albastru a nodurilor lui G astfel încât cel puţin $\dfrac{m}{2}$ muchii sunt bicolore.

Idee de rezolvare: coloraţi aleator nodurile şi calculaţi valoarea medie a numărului de muchii bicolore.

R: Colorăm aleator nodurile cu ajutorul unei monede: se aruncă moneda pentru fiecare muchie; dacă obţinem *cap* colorăm muchia cu roşu, dacă obţinem *pajură* colorăm muchia cu albastru.

Fie $X_i = \begin{cases} 1, & \text{muchia } i \text{ este bicoloră,} \\ 0, & \text{altfel.} \end{cases}$ Observăm că $X = \sum_{i=1}^{m} X_i$ indică numărul de muchii bicolore. Avem $E(X) = \sum_{i=1}^{m} E(X_i) = \sum_{i=1}^{m} \frac{1}{2} = \frac{m}{2}$. Problema 1. implică $P(X \geq \frac{m}{2}) \neq 0$, deci există cel puţin o colorare cu cel puţin $\frac{m}{2}$ muchii bicolore.

3. Dacă G este complet şi avem $2C_n^k < 2^{C_k^2}$, unde $k \in \mathbb{N}^*$, atunci există o colorare cu roşu şi albastru a muchiilor astfel încât niciun subgraf complet cu k noduri nu este monocolor.

Idee de rezolvare: coloraţi aleator muchiile şi calculaţi valoarea medie a numărului de subgrafuri complete care sunt monocolore.

R: Colorăm aleator muchiile cu ajutorul unei monede (similar procedeului din problema anterioară). Fie $G_1, \ldots, G_N$ subgrafurile complete cu k noduri ale grafului G, unde $N = C_n^k$.

Fie $X_i = \begin{cases} 1, & \text{subgraful } G_i \text{ este monocolor,} \\ 0, & \text{altfel.} \end{cases}$ Observăm că $X = \sum_{i=1}^{m} X_i$ indică numărul de subgrafuri monocolore. Avem $E(X) = \sum_{i=1}^{N} E(X_i) = \sum_{i=1}^{N} 2 \cdot \left(\frac{1}{2}\right)^{C_k^2} = \frac{2C_n^k}{2^{C_k^2}} < 1$. Problema 1. implică $P(X \leq E(X)) \neq 0$, deci $P(X < 1) = P(X = 0) \neq 0$, aşadar există cel puţin o colorare astfel încât niciun subgraf complet cu k noduri nu este monocolor.

4. Demonstraţi că există o mulţime $I \subseteq V$ independentă astfel încât $|I| \geq \dfrac{n^2}{4m}$.

Idee de rezolvare: consideraţi algoritmul aleator (Turan):

```
1. Delete each vertex of G (together with its incident edges)
        independently with probability 1-1/d, where d=2m/n.
2. For each remaining edge, remove it and one of its adjacent vertices.
```

Fie X şi Y numărul de noduri, respectiv de muchii, care supravieţuiesc după prima etapă din algoritm. Calculaţi $E(X-Y)$.

R: Observăm că *output-ul* algoritmului este o mulţime de noduri independentă I.

Fie $X_i = \begin{cases} 1, & \text{nodul } i \text{ supravieţuieşte după prima etapă,} \\ 0, & \text{altfel,} \end{cases}$ $i = \overline{1,n}$, şi

$Y_j = \begin{cases} 1, & \text{muchia } j \text{ supravieţuieşte după prima etapă,} \\ 0, & \text{altfel,} \end{cases}$ $j = \overline{1,m}$. Observăm că $E(X_i) = \frac{1}{d}$ şi $E(Y_j) = \frac{1}{d^2}$, deoarece o muchie supravieţuieşte dacă şi numai dacă supravieţuiesc nodurile ei. Deci $E(X-Y) = E(X) - E(Y) = \sum_{i=1}^{n} E(X_i) - \sum_{j=1}^{m} E(Y_j) = \frac{n}{d} - \frac{m}{d^2}$. Numărul de noduri Z care supravieţuiesc după etapa a doua este cel puţin egal cu $X-Y$, deci $E(Z) \geq \frac{nd-m}{d^2} = \frac{4m}{n^2}$. Problema 1. implică $P(Z \geq \frac{4m}{n^2}) \neq 0$, deci există o mulţime I, returnată de algoritm, care are cel puţin $\frac{4m}{n^2}$ noduri.

Observaţii:

• Pentru $k, l \in \mathbb{N}^*$, numărul lui Ramsey $R(k,l)$ este definit astfel: $R(k,l)$ este cel mai mic număr natural n cu proprietatea că orice colorare cu roşu şi albastru a muchiilor grafului complet G cu n vârfuri conţine un subgraf complet cu k noduri colorat cu roşu sau un subgraf complet cu l noduri colorat cu albastru.

• Numărul lui Ramsey $R(k,l)$ se poate defini şi astfel: este cel mai mic număr de persoane pe care trebuie să le invitaţi la o petrecere astfel încât să aveţi cel puţin un grup de k persoane în care oricare două se cunosc sau cel puţin un grup de l persoane în care oricare două nu se cunosc.

• Problema 3. ne spune: dacă $2C_n^k < 2^{C_k^2}$, atunci $R(k,k) > n$.

• Sunt cunoscute doar câteva valori ale numerelor lui Ramsey. De exemplu: $R(1,1) = 1$, $R(2,2) = 2$, $R(3,3) = 6$, $R(4,4) = 18$. Numerele lui Ramsey sunt recunoscute pentru dificultatea aflării valorilor lor. Valorile $R(k,k)$ pentru $k \geq 5$ sunt încă necunoscute.

• "Erdős asks us to imagine an alien force, vastly more powerful than us, landing on Earth and demanding the value of R(5, 5) or they will destroy our planet. In that case, he claims, we should marshal all our computers and all our mathematicians and attempt to find the value. But suppose, instead, that they ask for R(6, 6). In that case, he believes, we should attempt to destroy the aliens" - din cartea Joel H. Spencer - *Ten Lectures on the Probabilistic Method* (1994).