Laboratorul 4

1. Descărcaţi următoarele liste de cuvinte cheie extrase din email-uri:
• *spam*: `http://math.ubbcluj.ro/~miancu/teaching/PS/keywords_spam.txt`
• *ham*: `http://math.ubbcluj.ro/~miancu/teaching/PS/keywords_ham.txt`
Implementaţi în Matlab o clasificare naivă Bayes care stabileşte dacă un email este *spam* sau *ham*, calculând probabilităţile corespunzătoare date de frecvenţele relative de apariţii, respectiv neapariţii, ale cuvintelor cheie în listele date (neglijaţi cuvintele din email care nu apar în liste). Testaţi programul cu următoarele email-uri:
1) "`invite your friend today to click here`"
2) "`call your friend today it's urgent thank you`".

Pentru rezolvarea acestei probleme este necesară înţelegerea exemplului de clasificare naivă Bayes din curs: `http://math.ubbcluj.ro/~hanne/teaching/proba_stat_2019.pdf`
Se pot urma paşii:
• *atributele* sunt date de cuvintele distincte din cele două fişiere: W_1 : `call`, W_2 : `click` , . . . , W_{14} : `urgent`; valorile atributelor sunt *true* sau *false*; exemplu: $W_2 =$*true* reprezintă evenimentul că email-ul are cuvântul `click`; clasele sunt date de *ham* şi *spam*; exemplu: $C = spam$ reprezintă evenimentul că email-ul este *spam*.
• se calculează probabilităţile claselor. Exemplu: probabilitatea ca un email să fie *spam* este:

$$P(C = spam) = \frac{\text{numărul de cuvinte din fişierul keywords\_spam.txt}}{\text{numărul de cuvinte din ambele fişiere}}.$$

• se calculează probabilităţile atributelor, ştiind clasa. Exemplu: probabilitatea de apariţie a lui W_1 într-un email, ştiind că email-ul este *spam*, este:

$$P(W_1 = true|C = spam) = \frac{\text{numărul de apariţii ale cuvântului } \texttt{call} \text{ în keywords\_spam.txt}}{\text{numărul de cuvinte din keywords\_spam.txt}},$$

iar probabilitatea de neapariţie a lui W_1 într-un email, ştiind că email-ul este *spam*, este:

$$P(W_1 = false|C = spam) = 1 - P(W_1 = true|C = spam).$$

• pentru vectorul de atribute E_1 dat de cuvintele din primul email, se calculează produsele probabilităţilor de mai sus, iar pe baza formulei lui Bayes şi a condiţional independenţei avem:

$$\begin{aligned} P(C = spam|E_1) &= \frac{P(E_1|C = spam)P(C = spam)}{P(E_1)} \\ &= \frac{P(W_1 = false, W_2 = true, \ldots, W_{14} = false|C = spam)P(C = spam)}{P(E_1)} \\ &= \frac{P(W_1 = false|C = spam)P(W_2 = true|C = spam)\ldots P(W_{14} = false|C = spam)P(C = spam)}{P(E_1)} \end{aligned}$$

şi

$$\begin{aligned} &P(C = ham|E_1) \\ &= \frac{P(W_1 = false|C = ham)P(W_2 = true|C = ham)\ldots P(W_{14} = false|C = ham)P(C = ham)}{P(E_1)}. \end{aligned}$$

• se compară cele două probabilităţi (adică cei doi numărători ai fracţiilor de mai sus) şi se decide clasificarea dată de probabilitatea mai mare.

2. Un număr natural aleatoriu N (mai mic decât 64) este generat conform unei reţele Bayes în care nodurile sunt variabile aleatoare B_i care indică valoarea bitului de pe poziţia i, $i = \overline{1,6}$, în reprezentarea binară a numărului N (numerotarea poziţiilor biţilor se face de la dreapta la stânga), cu următoarele probabilităţi:

B_1	$P(B_1 = ...)$
1	0,8
0	0,2

B_2	$P(B_2 = ...\|B_1 = 1)$	$P(B_2 = ...\|B_1 = 0)$
1	0,9	0,6
0	0,1	0,4

B_3	$P(B_3 = ...\|B_2 = 1, B_1 = 1)$	$P(B_3 = ...\|B_2 = 1, B_1 = 0)$	$P(B_3 = ...\|B_2 = 0, B_1 = 1)$	$P(B_3 = ...\|B_2 = 0, B_1 = 0)$
1	0,6	0,2	0,9	0,4
0	0,4	0,8	0,1	0,6

B_4	$P(B_4 = ...\|B_3 = 1)$	$P(B_4 = ...\|B_3 = 0)$
1	0,3	0,5
0	0,7	0,5

B_5	$P(B_5 = ...\|B_3 = 1)$	$P(B_5 = ...\|B_3 = 0)$
1	0,5	0,8
0	0,5	0,2

B_6	$P(B_6 = ...\|B_5 = 1, B_4 = 1)$	$P(B_6 = ...\|B_5 = 1, B_4 = 0)$	$P(B_6 = ...\|B_5 = 0, B_4 = 1)$	$P(B_6 = ...\|B_5 = 0, B_4 = 0)$
1	0,5	0,3	0,8	0,5
0	0,5	0,7	0,2	0,5

a) Simulaţi de $n \in \{500, 1000, 1500\}$ ori valoarea numărului N. Afişaţi o histogramă a valorilor obţinute.

b) Comparaţi rezultatele obţinute cu cele teoretice pentru o valoare particulară a lui N.

Exemplu:

$$\begin{aligned} P(N = 23) &= P(B_6 = 0, B_5 = 1, B_4 = 0, B_3 = 1, B_2 = 1, B_1 = 1) \\ &= P(B_6 = 0|B_5 = 1, B_4 = 0) \cdot P(B_5 = 1|B_3 = 1) \cdot P(B_4 = 0|B_3 = 1) \\ &\quad \cdot P(B_3 = 1|B_2 = 1, B_1 = 1) \cdot P(B_2 = 1|B_1 = 1) \cdot P(B_1 = 1) \\ &= 0{,}7 \cdot 0{,}5 \cdot 0{,}7 \cdot 0{,}6 \cdot 0{,}9 \cdot 0{,}8 = 0{,}10584. \end{aligned}$$

Pentru rezolvarea acestei probleme este necesară înţelegerea exemplelor de reţea Bayes din curs: `http://math.ubbcluj.ro/~hanne/teaching/proba_stat_2019.pdf`