Laboratorul 4

1. Salvaţi fişierele `keywords_spam.txt` şi `keywords_ham.txt`, care conţin liste de cuvinte cheie extrase din email-uri *spam*, respectiv *ham*.

Implementaţi în Octave/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 studierea clasificării naive Bayes din curs.

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:

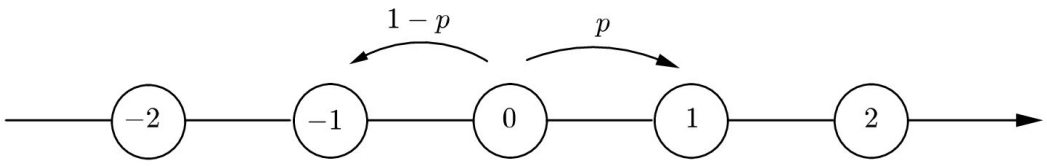
$$\begin{aligned} P(C = spam|E_1) &= \frac{P(E_1|C = spam)P(C = spam)}{P(E_1)} \\ &= \frac{P(W_1 = false, W_2 = true, \ldots, W_{14} = false|C = spam)P(C = spam)}{P(E_1)} \\ &= \frac{P(W_1 = false|C = spam)P(W_2 = true|C = spam) \ldots P(W_{14} = false|C = spam)P(C = spam)}{P(E_1)} \end{aligned}$$

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 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, iar nodul iniţial este 0:



a) Simulaţi o astfel de deplasare cu $k \in \mathbb{N}^*$ paşi, cu probabilitatea $p \in (0,1)$, şi returnaţi poziţiile curente la fiecare pas.

b) Simulaţi de $m \in \mathbb{N}^*$ ori o astfel de deplasare cu $k \in \mathbb{N}^*$ paşi, cu probabilitatea $p \in (0,1)$, şi afişaţi histograma poziţiilor finale. Care este poziţia finală cel mai des întâlnită (sau poziţiile finale cel mai des întâlnite)?

3. Un jucător de "Loto 6/49" îşi cumpără câte un bilet pentru fiecare extrage efectuată de loteria română până când reuşeşte să nimerească un bilet cu cel puţin 3 numere câştigătoare.

i) Folosind funcţiile `hygepdf` şi `geornd`, generaţi un vector x care conţine, pentru fiecare simulare, numărul de bilete necâştigătoare (care au cel mult 2 numere câştigătoare) până la primul bilet câştigător (care are cel puţin 3 numere câştigătoare).

ii) Estimaţi probabilitatea evenimentului:
"Cel puţin 10 bilete succesive sunt necâştigătoare până când jucătorul nimereşte un bilet câştigător."
Comparaţi probabilitatea estimată cu valoarea teoretică corespunzătoare, folosind funcţia `geopdf`.