5. Clasificatorul Naive Bayes

Algoritmi de învăţare generativi
Antrenarea clasificatorului Naive Bayes
Medierea Laplace
Predicţia datelor

Abordarea prezentată în acestă lucrare de laborator construieşte mai întâi un model ce descrie modul în care arată elefanţii, urmat de construcţia unui model separat ce descrie modul în care arată câinii. În final, pentru a clasifica un animal, putem calcula cât de mult se aseamănă el cu modelul ce descrie elefanţii, pe de o parte, şi cât de mult se aseamănă cu modelul ce descrie câinii, pe de altă parte.

5.1 Cerinţe

Să se scrie un script Python ce va clasifica mesaje email scrise în limba engleză ca fiind spam sau non-spam.

5.1.1 Baza de date de antrenare

Pentru a implementa algoritmul, descărcaţi datele din arhiva "lab5Data.zip".

Baza de date este împarţită în două subseturi, şi anume 700 de email-uri de antrenare şi 260 de email-uri de testare. Fiecare subset conţine 50% mesaje spam şi 50% mesaje non-spam. Adiţional, email-urile au fost procesate în felul următor:

(a) Anumite cuvinte, precum "and", "the" şi "of" sunt foarte comune în frazele din limba engleză şi nu sunt relevante în decizia de a clasifica un mesaj ca spam sau nu. Aceste cuvinte au fost înlăturate din email-uri.

(b) Cuvintele care au acelaşi înţeles, însă o terminaţie diferită, au fost ajustate în aşa fel încât să aibă aceeaşi formă. Spre exemplu, "include", "includes" şi "included" vor fi reprezentate de "include". Toate literele din corpul email-ului au fost convertite la litere mici.

(c) Numerele şi semnele de punctuaţie au fost înlăturate. Tab-urile, noile linii şi spaţiile au fost convertite la un singur spaţiu.

Mai jos se găsesc două exemple de mesaje înainte şi după procesare.

Mesajul non-spam "5-1361msg1" înainte de preprocesare

```
Subject:  Re:  5.1344 Native speaker intuitions

The discussion on native speaker intuitions has been extremely
interesting, but I worry that my brief intervention may have
muddied the waters.  I take it that there are a number of
separable issues.  The first is the extent to which a native
speaker is likely to judge a lexical string as grammatical
or ungrammatical per se.  The second is concerned with the
relationships between syntax and interpretation (although even
here the distinction may not be entirely clear cut).
```

Mesajul non-spam "5-1361msg1" după preprocesare

```
re native speaker intuition discussion native speaker intuition
extremely interest worry brief intervention muddy waters number
separable issue first extent native speaker likely judge lexical
string grammatical ungrammatical per se second concern
relationship between syntax interpretation although even here
distinction entirely clear cut
```

Mesajul spam "spmsgc19" după preprocesare

```
financial freedom follow financial freedom work ethic
extraordinary desire earn least per month work home special
skills experience required train personal support need ensure
success legitimate homebased income opportunity put back control
finance life ve try opportunity past fail live promise
```

Arhiva "lab5Data.zip" conţine fişierul "train-features-full.txt", ce reprezintă matricea de antrenare x, formată din 700 de linii şi 2500 de coloane. Fiecare linie reprezintă un email, iar fiecare coloană reprezintă un cuvânt din dicţionar. Un element dintr-o linie a matricei x reprezintă numărul de apariţii în email al cuvântului reprezentat de acel element.

Fişierul "train-labels.txt" conţine etichetele atribuite fiecărui email ($y = 1$ pentru email-

urile spam şi $y = 0$ pentru email-urile non-spam) şi are dimensiunea 700×1.

5.1.2 Antrenarea

Paşii de antrenare ai algoritmului Naive Bayes, odată ce baza de date de antrenare a fost încărcată în Python, sunt următorii:

1. calculul ϕ_y;
2. calculul fiecărui $\phi_{j|y=1}$ pentru fiecare cuvânt din dicţionar şi stocarea rezultatelor într-un vector;
3. calculul fiecărui $\phi_{j|y=0}$ pentru fiecare cuvânt din dicţionar şi stocarea rezultatelor într-un vector.

5.1.3 Testarea

Odată ce parametrii modelului au fost calculaţi, ei pot fi utilizaţi în efectuarea predicţiilor. Pentru aceasta se vor folosi fişierele "test-features-full.txt" şi "test-labels.txt", ambele având structura fişierelor "train-features-full.txt", respectiv "train-labels.txt".

Paşii utilizaţi pentru clasificarea documentelor sunt următorii:

1. pentru fiecare document din baza de date de testare se calculează $\log p(x|y = 1) + \log p(y = 1)$;
2. similar, se calculează $\log p(x|y = 0) + \log p(y = 0)$;
3. se compară cele două cantităţi de la punctele (1) şi (2) şi se ia decizia de clasificare a email-ului în spam sau non-spam, luând în considerare cea mai mare valoare.