Pràctica Intel·ligència Artificial

CONSTRUCCIÓ D'UN FILTRE DE "SPAM" MITJANÇANT EL MÈTODE "NAIVE BAYES"

La pràctica consisteix en implementar un filtre de correu "spam" fent servir els conceptes explicats a classe. Heu de:

1. Descarregar aquest fitxer. Conté uns 20000 emails triats de l'*Enron spam corpus* en una carpeta comprimida. La sotscarpeta HAM_TRAINING conté missatges "legítims" (HAM) i la sotscarpeta SPAM_TRAINING conté correu no desitjat (SPAM)

2. Implementar el filtre de correu no desitjat seguint les instruccions explicades a classe.

3. Avalueu la seva eficència. Per fer això, podeu apartar una part dels correus (5-10%) per formar un conjunt de validació. Entreneu el filtre amb els correus restants i feu servir el conjunt de validació per calcular els següents estadístics:

    a. Accuracy : Percentatge de emails classificats correctament
    b. False positive rate : Percentatge de correus legítims (HAM) classificats com a SPAM
    c. False negative rate : Percentatge de correus SPAM classificats com a HAM
    d. Total cost ratio : Té en compte que els falsos positius són en general més costosos que els falsos negatius. Podeu trobar la definició a:
    http://wiki.apache.org/spamassassin/TotalCostRatio

    La sortida hauria de ser quelcom com això:

```
-------------------- RESULTATS ----------------------
Nombre de missatges:            1908     (920H,988S)
Accuracy (%):                   94.29
False positive rate (%):        0.21     (2)
False negative rate (%):        11.47    (107)
Total cost ratio (l = 50):      4.5072
-----------------------------------------------------
```

    Fixeu-vos que per calcular el "total cost ratio" fem servir un valor de la constant lambda igual a 50.

4. Experimenteu amb els diferents paràmetres del filtre i incorporeu els refinaments que se us acudeixin per tal d'optimitzar l'eficiència (especialment el "total cost ratio")

OBSERVACIONS

La pràctica no tracta simplement d'implementar el mètode explicat a classe. Es demana que el refineu, incorporant millores, i que documenteu com aquestes millores afecten al rendiment del sistema. Aquestes millores us les podeu inventar vosaltres o bé les podeu adaptar a partir d'idees d'altres persones que trobeu a algun llibre o article o a la web. En aquest darrer cas heu d'especificar la font d'on prové la idea original.

Els correus electrònics que farem servir per aquesta pràctica **no contenen capçaleres** (from, to, cc, ...). Només el "subject" i el "body"

Podeu fer servir el llenguatge de programació que vulgueu, encara que alguns us faran la feina més senzilla que d'altres. Us aconsellem que feu servir Python ver vàries raons:

- El nombre de missatges al conjunt d'entrenament és relativament petit i la rapidesa no és un requisit clau.
- És interessant dominar Python, ja que s'està convertint en una "lingua franca" de la mineria de dades i la "data science" en general.
- És un llenguatge molt expressiu. Pel nostre problema n'hi ha prou amb amb menys de 100 línies de codi. Per exemple aquesta instrucció:

```
open("email.txt", encoding = "latin-1").read().split()
```

crea una llista amb totes les paraules del correu guardat a l'arxiu "email.txt"

El mètode "Naive Bayes" s'ha d'implementar. Hi ha una pila de llibreries de classificació de texts que ja el porten implementat. No es poden fer servir. Tampoc es pot, òbviament, simplement copiar qualsevol del miler d'implementacions que corren per la web.

Si voleu realment **aspirar a una molt bona nota**, podeu extendre la pràctica de diferents formes. Per exemple:

- Implementant un mètode de cerca heurística per determinar quina combinació de valors dels paràmetres del sistema optimitza la seva eficiència
- Implementant el mètode de "k-fold cross-validation".
- Comparant l'eficiència del filtre amb la d'altres algorismes de classificació de textes continguts en llibreries com, per exemple, scikit-learn i nltk (Python) o LingPipe i Mallet (Java) o a aplicacions de data mining com Weka i RapidMiner

Si se us acut qualsevol altra manera, comenteu-m'ho i ho discutirem

QUÊ S'HA DE LLIURAR

Cal lliurar un informe que inclogui:

- Una introducció, explicant breument l'aplicació el mètode "Naive Bayes" al filtratge de correu no desitjat. Quines són les assumpcions més importants que fa el mètode. Quins són els seus punts forts i febles.
- Una explicació detallada de com implementeu el mètode, justificant les decisions de disseny (estructura de l'aplicació, tipus i estructures de dades que es fan servir, llenguatges i llibreries utilitzades, paràmetres ajustables, assignació dels correus a una classe o l'altra, realització de càlculs,…)
- Un comentari sobre el mètode d'avaluació que heu fet servir
- Una descripció de les millores que heu incorporat a l'algorisme inicial i com han repercutit en l'eficiència del vostre filtre
- El codi comentat
- Unes conclusions que incloguin quines dificultats heu trobat i com les heu solucionat i la vostra opinió general sobre la pràctica.

COMPETICIÓ

La darrera setmana de classe es celebrarà una competició entre els grups o equips que voluntàriament es vulguin presentar. La competició consistirà en classificar un conjunt de prova format per aproximadament 1000 correus HAM i 1000 correus SPAM separats en dos directoris (HAM i SPAM). Guanyarà l'equip amb "total cost ratio" màxim. Els premis seran de 1, 0.5 i 0.25 punts sobre la nota final de la segona part de l'assignatura pel primer, segon i tercer equip, respectivament (sempre que s'aprovi l'exàmen final).