Pràctica 3 - CRI

Josep Maria Domingo Catafal - NIU 1599946

Contents

1 Introducció: explicació del problema.

La gent publica tweets constantment, i en molts casos aquests tweets reflecteixen l'estat d'ànim de la persona. Disposem d'una base de dades amb molts tweets ja classificats segons l'estat d'ànim que reflecteixen (positiu o negatiu), i l'objectiu és, a partir d'aquesta informació, entrenar un model, utilitzant ***Naive Bayes***, que permeti classificar futurs tweets.

El format del dataset es el següent:

```
ID   Contingut del tweet    Data         Sentiment (0 o 1)
-----------------------------------------------------------
16   I fell in love again   02/12/2015   1
```

2 Solució proposada

2.1 Lectura del dataset

El primer pas és llegir el dataset per poder treballar amb ell. Per fer-ho s'ha fer ús de la llibreria pandas. En aquest pas s'eliminen les columnes de ID i data, ja que no ens serveixen per res.

2.2 Divisió en conjunt de train i test

Un cop tenim les dades el seguent pas es dividir en conjunt de train i test. Per aquest pas també s'ha utilitzat padas. Per crear el conjunt d'entrenament, el que es fa es agrupar les dades segons el seu sentiment (0 o 1) i s'agafa una mostra del tamany especificat (per defecte un 70%) de cada un dels grups. D'aquesta manera ens assegurem que tindrem la mateixa proporció de positius i negatius tant en el conjunt de train com el de test. Per tal de crear el conjunt de test es simplement agafar els elements que no han sigut seleccionat per el conjunt de train.

Les particions són numpy arrays.

2.3 Implementació del classificador

Per tal d'implemental el classificador s'ha creat una classe anomenada BayesClassifier, la qual conté totes les funcions necessaries per entrenar el model i fer prediccions.

2.3.1 Entrenament

Per tal d'entrenar el model s'ha de cridar la funció fit, la qual rep les dades d'entrenament per parametre. La seva funcionalitat es cridar a la resta de funcions que ens permetran entrenar el model.

2.3.2 Calcul dels priors: P(positiu) i P(negatiu)

Per tal de calcular els priors, es tan senzill com contar quants tweets son positius i dividir-ho entre el total i el mateix per els negatius. Per fer-ho es fa em numpy, ja que las dades estan a un numpy array i aporta bon rendiment.

2.3.3 Creació dels diccionaris

S'han creat dos diccionaris, un amb les paraules dels tweets positius i un altre amb les paraules del tweets negatius. Per tal de crear aquests diccionaris, es fa servir la classe `Counter` del mòdul collections de la llibreria estandard de Python. El que es fa és dividir els tweets segons el seu sentiment, i per cada tweet, es crea una llista amb les paraules que conté, i es passa al `Counter` el qual genera un diccionari amb el nombre d'ocurrences de cada paraula. Hi ha dos contadors, un per les paraules positives i un per les paraules negatives.

El format dels diccionaris que es retornen és el següent:

```
positive_words:  {
    "word" : occurances of word given sentiment is positive,
    ...
}
negative_words:  {
    "word" : occurances of word given sentiment is negative,
    ...
}
```

2.3.4 Creació de la taula de probabilitats

Per tal de poder aplicar Naive Bayes necessitem una taula amb les probabilitats condicionades de cada paraula. La taula es genera de la forma següent: Per cada paraula en els diccionaris de l'apartat anterior, es divideix el nombre d'ocurences d'aquella paraula entre el nombre total de paraules d'aquell sentiment. És a dir si estem recorrent el diccionari de paraules positives es dividiria el nobre d'occurences d'aquella paraula entre el nombre de paraules positives.

El resultat es retorna en forma de diccionari i té el format següent:

```
{
    'word' : [P(word|negative), P(word|positive)],
    ...
}
```

2.4 Predicció

Un cop ja tenim totes les probabilitats calculades, ja podem fer prediccions aplicant Naive Bayes. El que es fa és, per cada paraula del tweet que es vol predir, multiplicar les seves probabilitats i finalment multiplicar per el la probabilitat de que sigui negatiu. Després fem el mateix però multiplcant per la probabilitat

de que sigui positiu. El resultat que sigui major dels dos ens indicarà quin és el sentiment del tweet. En cas d'empat, s'agafara el que tingui el prior més gran. En resum seria fer el següent:

$$S := \{Positiu, Negatiu\}$$

$$\arg\max_{x_i \in S} P(x_i) \cdot \prod_{i=1}^{n} P(word_i | x_i)$$

3 Problemes trobats durant el projecte i com els heu resolt.

4 Resultats

5 Conclusions i treballs futurs