Naive Bayes
Bernoulli e Multinomiale

Antonio Castellucci

January 5, 2018

1 Scopo

Implementare la tecnica di machine learning Naive Bayes,sia in versione bernoulli che multinomiale e apllicarla al data set 20 newsgroups, riportando le matrici di confusione delle due versioni di Naive Bayes.

1.1 Codice

Bernoulli.py Questo file implementa le due funzioni necessarie alla versione Bernoulli: BernoulliTrain(per l'addestramento sul TrainSet) e BernoulliCompute(che preso in ingresso un file,ne predice la classe)

Multinomial.py Questo file implementa le due funzioni necessarie alla versione Multinomiale: MultinomialTrain(per l'addestramento sul TrainSet) e MultinomialCompute(che preso in ingresso un file,ne predice la classe)

Utility.py Questo file implementa alcune funzioni che le due versioni condividono come quella per estrarre il dizionario da un insieme di documenti o quella per estrarre da una directory tutti i file presenti nelle subdirectory. Inoltre implementa due versioni distinte di una funzione (documentAnalyse per bernoulli e documentAnalyseToken per multinomiale) che serve ad estrarre l'insieme delle parole da un documento.

unknown.py Questo file insieme a quelli non citati ma presenti nella cartella serve a gestire la GUI che è stata creata con il programma PAGE. Inoltre implementa la funzione che gestisce tutto l'esperimento,ovvero createMatrix. Questa funzione,usando come identificatore il nome del test (che viene inserito dall'utente tramite la GUI) esegue in ordine i seguenti passaggi:

a. divide il DataSet in TrainSet e TestSet in base alla percentuale inserita dall'utente.

b. Se non già presente,estrae dal TrainSet il dizionario delle parole e lo salva per futuri riutilizzi.

c. in base alla modalità scelta dall'utente (Bernoulli ("b") o Multinomiale ("m") il programma verifica o meno l'esistenza della tavola delle probabilità e in caso la genera chiamando la relativa funzione di train.

d. il programma quindi predice per ogni file nel TestSet la classe di appartenenza e rilevando automaticamente la correttezza o meno della predizione costruisce la matrice di confusione.

2 DataSet

Il DataSet utilizzato per eseguire i successivi test dimostrativi è 20 NewsGroups. Esso è costituito da 20 cartelle contenenti 1 000 file ciascuna,le cartelle suddividono i vari newsgroups ma alcune di esse parlano all'incirca dello stesso argomento (es: atheism,religion e christian). Per fare i seguenti test si è usato un 20% del dataset per problemi di run-time,da ogni cartella sono stati tolti 800 file. Nel primo test i file sono stati raccolti nei 6 macro-argomenti indicati sul sito di riferimento del data set,nel secondo si è provato a distinguere i file fra tutti e 20 i newsgroups con conseguente minor precisione. Da tutti i file sono stati tolti gli Headers poichè contenevano token identificativi della classe.

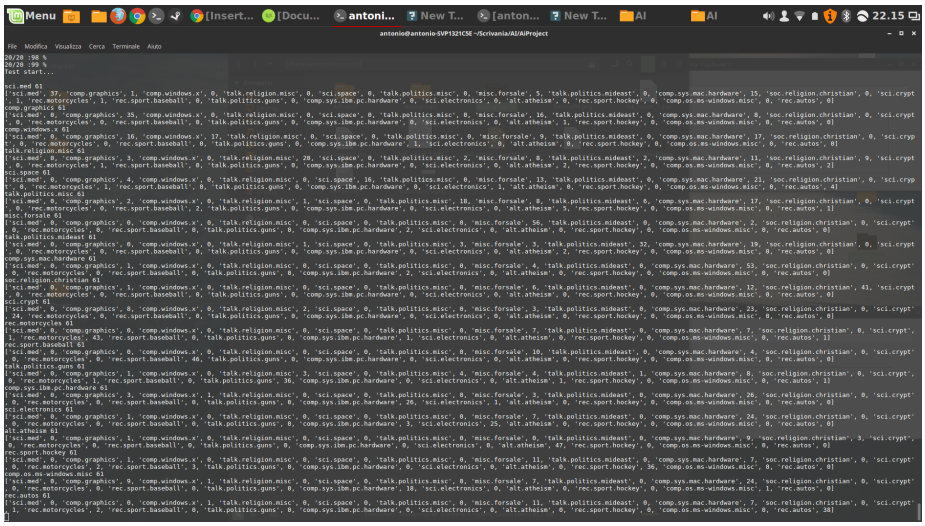
3 Dati Sperimentali

```
Affari 181
['Affari', 167, 'Politica', 0, 'Sport', 4, 'Ingegneria', 0, 'Religione', 0, 'Computer', 9]
Politica 181
['Affari', 27, 'Politica', 107, 'Sport', 10, 'Ingegneria', 9, 'Religione', 14, 'Computer', 13]
Sport 181
['Affari', 31, 'Politica', 2, 'Sport', 139, 'Ingegneria', 0, 'Religione', 1, 'Computer', 7]
Ingegneria 181
['Affari', 38, 'Politica', 7, 'Sport', 7, 'Ingegneria', 98, 'Religione', 2, 'Computer', 28]
Religione 181
['Affari', 28, 'Politica', 6, 'Sport', 3, 'Ingegneria', 4, 'Religione', 130, 'Computer', 9]
Computer 181
['Affari', 23, 'Politica', 1, 'Sport', 1, 'Ingegneria', 1, 'Religione', 0, 'Computer', 154]
```

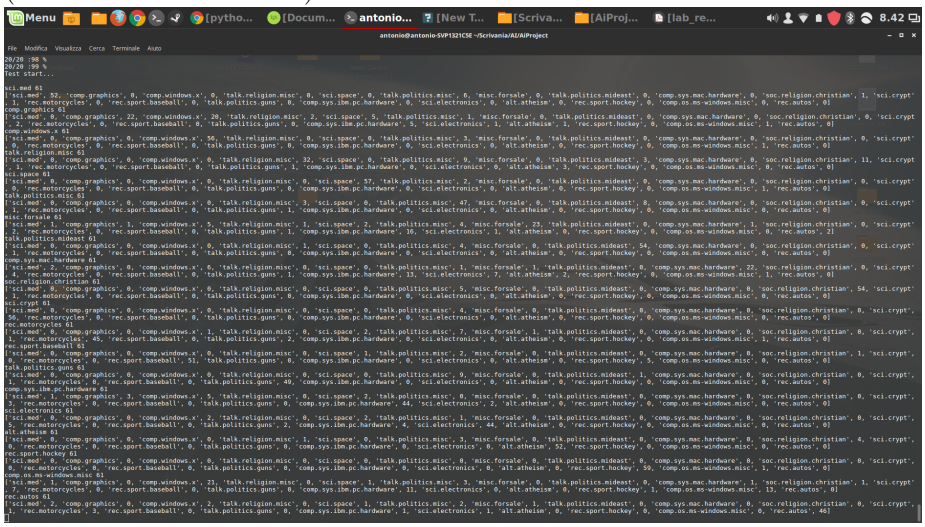
Matrice di confusione su un DataSet di 3200 elementi con TrainSet al 70% (Bernoulli-Precisione 74%).

```
Affari 181
['Affari', 138, 'Politica', 6, 'Sport', 6, 'Ingegneria', 7, 'Religione', 2, 'Computer', 21]
Politica 181
['Affari', 0, 'Politica', 164, 'Sport', 3, 'Ingegneria', 6, 'Religione', 6, 'Computer', 1]
Sport 181
['Affari', 4, 'Politica', 6, 'Sport', 163, 'Ingegneria', 3, 'Religione', 1, 'Computer', 3]
Ingegneria 181
['Affari', 2, 'Politica', 5, 'Sport', 1, 'Ingegneria', 155, 'Religione', 4, 'Computer', 13]
Religione 181
['Affari', 0, 'Politica', 21, 'Sport', 0, 'Ingegneria', 2, 'Religione', 157, 'Computer', 0]
Computer 181
['Affari', 4, 'Politica', 0, 'Sport', 1, 'Ingegneria', 9, 'Religione', 2, 'Computer', 164]
```

Matrice di confusione su un DataSet di 3200 elementi con TrainSet al 70% (Multinomiale-Precisione 87%).



Matrice di confusione su un DataSet di 3200 elementi con TrainSet al 70% (Bernoulli-Precisione 56%).



Matrice di confusione su un DataSet di 3200 elementi con TrainSet al 70% (Multinomiale-Precisione 73%).

4 Riproduzione esperimenti

Per riprodurre i dati sarà sufficente eseguire il test sui dataset presenti nella cartella DataSet che si trova nella cartella del programma. La cartella contiene già sia i vocabolari sia le tabelle con le probabilità condizionate sia per la versione Bernulli sia per la versione Multinomiale,in tal modo sarà possibile ottenere le matrici di confusione sopra riportate in breve tempo. Per non dover riestrarre

il vocabolario e riaddestrare Naive Bayes sarà necessario inserire il nome del test: "test6Classi" o "test20Classi" a seconda se si vogliono ottenere i primi o i secondi risultati.

5 Conclusioni

Il vocabolario estratto per i test riportati conta all'incirca 40 000 parole,possiamo vedere come se raggruppiamo i file per argomenti più distinti fra loro la precisione della classificazione aumenta. Dalle matrici di confusione possiamo inoltre verificare come ci aspettavamo che la versione Bernoulli è meno precisa di quella multinomiale che tuttavia risulta fino a 4 volte più lenta nella computazione. La difficoltà maggiore è stata riscontrata nella creazione del vocabolario e piccole variazioni al codice che estrae le parole dai file causano grandi differenze nell'accuratezza degli algoritmi.