Implementazione Naive Bayes

Francesco Bellocci Matricola 5941110

February 7, 2019

1 Introduzione

L'obiettivo del progetto é stato quello di implementare e in seguito confrontare, la versione Multinomiale e la versione di Bernoulli dell'algoritmo Naive Bayes ed osservare il loro comportamento nel compito di etichettare dei documenti a delle classi.

2 Esperimento

Il dataset utilizzato per effettuare l'analisi dei due classificatori é il 20 newsgroup.
Sono stati effettuati diversi test, inizialmente prendendo un numero di categorie piccole, aumentando ad ogni esperimento il numero di categorie fino ad utilizzare tutte quelle presenti nel dataset.
Per ogni test svolto, sono stati calcolati e mostrati in output dal programma:

- l'accuratezza del classificatore
- la matrice di confusione generata dal classificatore

Andiamo ad analizzare i test svolti ed i risultati ottenuti, osservando le caratteristiche e le differenze dei due classificatori.

Test no.1
Sono state considerate solo due categorie: 'alt.atheism', 'comp.graphics'.

Bernoulli

- Precisione: 95.480 %

- Matrice di confusione:

	Valori Predetti		
	alt.atheism	comp.graphics	Somma
alt.atheism	293	26	319
comp.graphics	6	383	389
Somma	299	409	708

Multinomiale

- Precisione: 98.446 %
- Matrice di confusione:

	Valori Predetti		
	alt.atheism	comp.graphics	Somma
alt.atheism	315	4	319
comp.graphics	7	382	389
Somma	322	386	708

Test no.2
Sono state considerate tre categorie:
'rec.sport.baseball','misc.forsale','soc.religion.christian'

Bernoulli

- Precisione: 91.477 %
- Matrice di confusione:

	Valori Predetti			
	rec.sport.baseball	misc.forsale	soc.religion.christian	Somma
rec.sport.baseball	390	0	0	390
misc.forsale	42	355	0	397
soc.religion.cristian	57	2	339	398
Somma	489	357	339	1185

Multinomiale

- Precisione: 98.312 %
- Matrice di confusione:

	Valori Predetti			
	rec.sport.baseball	misc.forsale	soc.religion.christian	Somma
rec.sport.baseball	386	3	1	390
misc.forsale	7	390	0	397
soc.religion.christian	6	3	389	398
Somma	399	396	390	1185

Per i Test no.3 e no.4 verrá riportato solo la precisione del classificatore a causa della grandezza della tabella.

Test no.3
Le categorie in esame sono: 'alt.atheism',
'comp.graphics','rec.sport.baseball','misc.forsale','soc.religion.christian'

Bernoulli

- Precisione: 86.424 %

Multinomiale

- Precisione: 95.351 %

Test no.4
Per questo test sono state considerate tutte le categorie presenti nel dataset 20 newsgroup.

Bernoulli

- Precisione: 65.720 %

Multinomiale

- Precisione: 80.231 %

3 Conclusione ed osservazioni

Dagli esperimenti svolti possiamo osservare che il classificatore Naive Bayes nella versione di Bernoulli si comporta bene come quello nella versione Multinomiale quando il numero delle categorie é piccolo.
All'aumentare del numero di categorie peró abbiamo che la versione Multinomiale, conserva un'accuratezza molto elevata mentre la versione di Bernoulli decresce molto velocemente, basta infatti confrontare i risultati del Test no.1 e del Test no.3 che notiamo subito un calo di prestazione del '9.056' %.
Quindi possiamo concludere affermando che entrambi i classificatori, nonostante la forte assunzione che effettuano per il calcolo della probabilitá condizionata tipiche degli algoritmi di Naive Bayes, con un piccolo numero di dati per 'allenare' il classificatore é riescono ad ottenere una precisione elevata, nel nostro caso, nel compito di etichettare i documenti ad una classe, con la differenza che la versione di Bernoulli ha un ottimo successo quando il numero di classi é piccolo mentre la versione Multinomiale mantiene elevata la sua precisione anche con un numero abbastanza 'grande' di categorie.