Confronto tra Classificatori di Testo Naive Bayes

Giuseppe Serna

February 4, 2017

1 Obiettivo

Confrontare le versioni Bernoulli e Multinomiale di Naive Bayes. A questo fine sono stati utilizzati i Dataset 20newsgroups e Reuters-21578 per produrre le learning curves e studiare l'andamento dell'errore di generalizzazione dei rispettivi classificatori al variare del numero di examples utilizzati nel training.

1.1 Tools

Librerie Nel progetto sono state utilizzate le seguenti librerie:

Scikit-learn Usato per implementare Bernoulli e Multinomial, CountVectorizer per estrarre le bag-of-words, inoltre il dataset 20newsgroups è stato scaricato e gestito trammite fetch20newsgroups.

nltk Trammite questa libreria è stato scaricato(nltk.download()) e gestito il dataset Reuters-21578 presente nel Corpus Reuters.

matplotlib Usato per creare i grafici delle learning curves

Implementazione del codice Il codice è diviso in 4 file:

Reuters Dove si suddividono i dati di train da quelli di test e si calcolano i target per ognuna delle 10 categorie pi frequenti.

newsgroupsrandom Dove viene scelto in modo casuale l' 80% del dataset 20newsgroup per il training e il restante 20% per il testing, tramitte il parametro mindf di Countvectorizer si eliminano le words che appaiono una volta sola in un documento. Infine si usa la funzione accuracy-score() per misurare la prestazione.

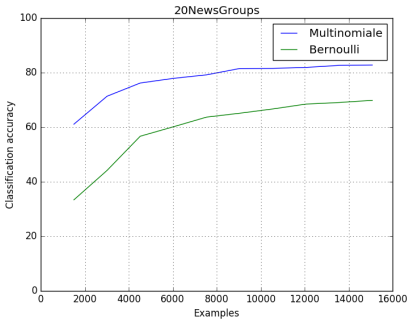
ReutersClassifier Dove viene classificato il dataset Reuters-21578 in base a una categoria, la prestazione viene misurata con la funzione f1-score() che tiene conto sia della precision che della recall del classificatore.

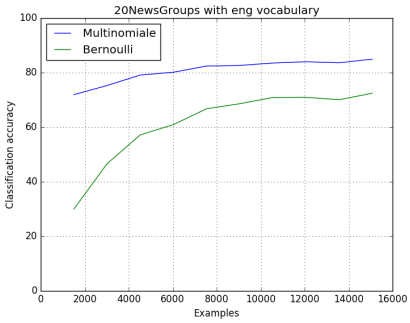
main Nel quale si impostano i parametri dei classificatori; quando si elabora 20newsgroups vengono calcolati 4 risultati e fatta la media; nel richiamare i classificatori si può decidere di rimuovere headers, oppure filtrare le words con un dizionario inglese. Nel Classificatore per Reuters invece si può solo decidere di filtrare le words con un dizionario inglese, vengono calcolati 10 risultati uno per categoria.

---

2 Risultati e Conclusioni

Si riportano qui sotto le learning curves riferite a 20NewsGroups, Nella prima figura sono state filtrate le parole che appaiono una volta e rimossi gli headers dagli articoli, si può subito notare come Bayes Multinomiale arrivi approssimativamente a un picco del 83% di accuratezza mentre Bernoulli un massimo del 71%, inoltre Bernoulli non ottiene mai risultati migliori; in definitiva si può notare che in entrambi i casi l'accuratezza migliora all'aumentare del numero di examples. Nel secondo caso si è deciso di filtrare attraverso un dizionario inglese le words, si può notare un generale anche se piccolo miglioramento rispetto al caso precedente, in particolare il classificatore Multinomiale giova di questo filtraggio quando il numero di examples è al minimo, tutto questo può essere dovuto dal fatto che si elimina in questo modo parte del rumore dovuto a words poco significativi.



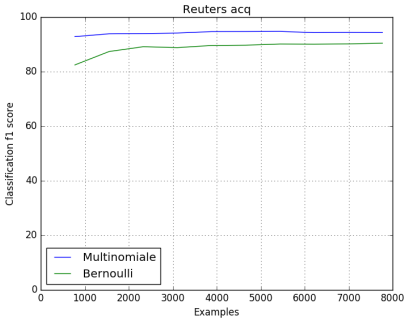


Nelle pagine seguenti sono invece riportati i risultati dei test eseguiti sul dataset Reuter-21578, in questo caso è stata eseguita una classificazione binaria per ognuna delle 10 categorie pi popolari, in questo caso la misura di prestazione F1 si rivela pi valida. in questo caso a parte il caso di acq, Bernoulli si rivela un classificatore non affidabile qualunque sia il numero di examples, tuttavia anche

il classificatore Multinomiale non risente particolarmente del numero di examples di training. Anche in questo caso è stato fatto un seguente test filtrando le words con un vocabolario inglese, in questo caso però i cambiamenti sono quasi impercettibili e i risultati non sono stati inseriti essendo non rilevanti.

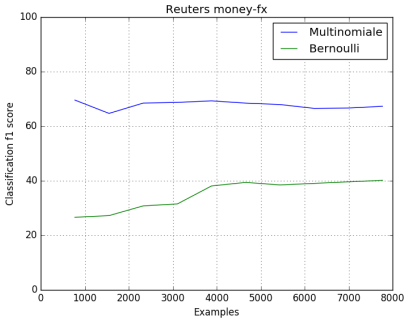
2.1 Conclusioni

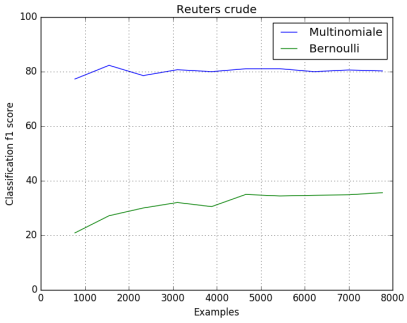
Infine si può dire che dai risultati di 20newsgroups che l'errore di generalizzazione del classificatore Multinomiale non risente molto delle variazioni di dimensioni del dataset di training, inoltre tale variazione è ancora pi ridotta se si migliora la qualità del dizionario.Bernoulli d'altra parte risente della carenza di dati di training e un miglioramento della qualità del vocabolario non produce effetti rilevanti; l'errore di generalizzazione è inversamente proporzionale alla dimensione del training set , si può concludere che Bernoulli è molto meno efficiente della versione Multinomiale ma opera in maniera discreta quando il training-set è a dimensione massima. Riguardo i risultati del dataset Reuter-21578, si può notare che l'errore di generalizzazione per il classificatore Multinomiale non dipende dalla dimensione del trainingset poichè lo score rimane pressochè invariato all'interno di ogni grafico, tuttavia lo score cambia molto a seconda della categoria. La versione di Bernoulli presenta delle lievi migliorie all'aumento del trainingset ma l'errore di generalizzazione rimane essere troppo grande perchè possa essere considerato efficace per l'utilizzo in questo dataset. Si può concludere che il classificatore Multinomiale può avere un errore di generalizzazione basso anche con un basso numero di examples di training a patto di migliorare la qualità del vocabolario, l'errore nel Bernoulli dipende in maniera più forte dalla dimensione del training set.


Reuters acq
Classification f1 score
Examples
Multinomiale
Bernoulli


Reuters earn
Classification f1 score
Examples
Multinomiale
Bernoulli

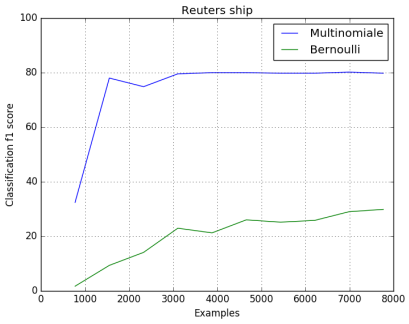

Reuters grain
Classification f1 score
Examples
Multinomiale
Bernoulli


Reuters money-fx
Classification f1 score
Examples
Multinomiale
Bernoulli

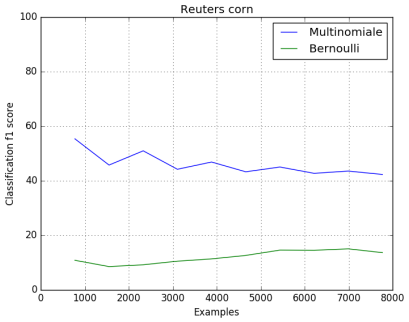

Reuters crude
Classification f1 score
Examples
Multinomiale
Bernoulli


Reuters trade
Classification f1 score
Examples
Multinomiale
Bernoulli


Reuters interest
Multinomiale
Bernoulli
Classification f1 score
Examples
0
20
40
60
80
100
0
1000
2000
3000
4000
5000
6000
7000
8000


Reuters ship
Multinomiale
Bernoulli
Classification f1 score
Examples
0
20
40
60
80
100
0
1000
2000
3000
4000
5000
6000
7000
8000


Reuters wheat
Multinomiale
Bernoulli
Classification f1 score
Examples
0
20
40
60
80
100
0
1000
2000
3000
4000
5000
6000
7000
8000


Reuters corn
Multinomiale
Bernoulli
Classification f1 score
Examples
0
20
40
60
80
100
0
1000
2000
3000
4000
5000
6000
7000
8000