Classificazione Documenti di Test sui dataset Reuters-21578 e 20newsgroups

Christian Pratellesi

7 febbraio 2020

Indice

1 Introduzione

In questo esperimento è stato eseguito un confronto tra le versioni **Bernoulli** e **Multinomiale** di *Naive Bayes* analizzando la precisione di predizione della categoria di un testo all'aumentare delle parole considerate e cercando di riprodurre i risultati ottenuti in *McCallum - Nigam 1998*. Per entrambe le versioni di Naive Bayes è stata utilizzata l'implementazione fornita da **scikit-learn**. L'analisi è stata effettuata su due datasets di documenti: **Reuters-21578** e **20newsgroups**.

1.1 Reuters-21578

È un dataset **multi-class** (i documenti sono suddivisi in 90 classi diverse) e **multi-label** (ogni documento può appartenere a più di una classe). Il numero totale di documenti è 10788, suddivisi in train e test secondo lo split **ModApte** (7769 documenti di training e 3019 documenti di testing). Per questa analisi sono stati utilizzati soltanto documenti appartenenti alle 10 classi più frequenti (acq, corn, crude, earn, grain, interest, money-fx, ship, trade, wheat), andando ad analizzare un numero totale di 9979 documenti, di cui 7192 per il training e 2787 per il testing, per un totale di 18676 parole diverse.

1.2 20newsgroups

È un dataset composto da un totale di 18846, ognuno dei quali appartiene ad uno dei 20 topics. Il dataset è gia diviso in train (11314 documenti) e test (7132 documenti) con un totale di 130107 parole diverse.

2 Il Codice

Il codice è suddiviso in due file: `Reuters.py` e `20newsgroups.py`. Il primo file analizza **Reuters-21578** mentre il secondo analizza **20newsgroups**. Quello che fa il codice, per entrambi i file, è andare a scaricare la corrispondente libreria da internet, vettorizzare i documenti andando a contare il numero di parole diverse trovate in ogni documento e creando un array di dimensione **n-documenti** x **n-parole-diverse**, creare due classificatori (**Bernoulli** e **Multinomiale**) allenati con i documenti di training ed infine predire la categoria per ogni documento di test analizzando i risultati per numero di parole crescente (selezionate utilizzando il selettore `SelectKBest`

disponibile in **scikit-learn** che seleziona le k parole più significative effettuando, in questo caso, un test del chi-quadro). In output si ha un file `.csv` contenente i valori numerici ottenuti per quanto riguarda l'accuratezza di predizione.

3 Analisi dei risultati

Una volta creati i classificatori, si sono allenati con il subset di training ed è stata effettuata la misura di accuratezza andando a confontare i la predizione con i valori da noi attesi. I risultati sono i seguenti.

3.1 Reuters-21578

Per quanto riguarda questo dataset, ci aspettavamo risultati migliori dall'implementazione Multinomiale di Naive Bayes e, se ottenuti, risultati migliori dell'implementazione di Bernoulli solo per un numero di parole ridotto. Le analisi mostrano, come in Figura 1, che inizialmente l'implementazione di Bernoulli ha un'accuratezza di predizione maggiore, superata intorno alle 2500 parole da quella dell'implementazione Multinomiale. Tuttavia, superate le 15000 parole, la nostra analisi ottiene risultati di precisione migliori nuovamente con l'implementazione di Bernoulli.

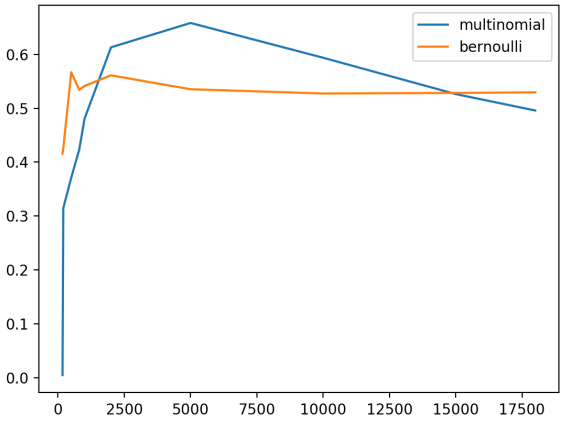

Figura 1: Preciosione di predizione sul dataset *Reuters-21578*

È stata analizzata la precisione di predizione anche per quanto riguarda la singola classe **acq**, i risultati che ci aspettavamo sono mostrati in Figura 2:

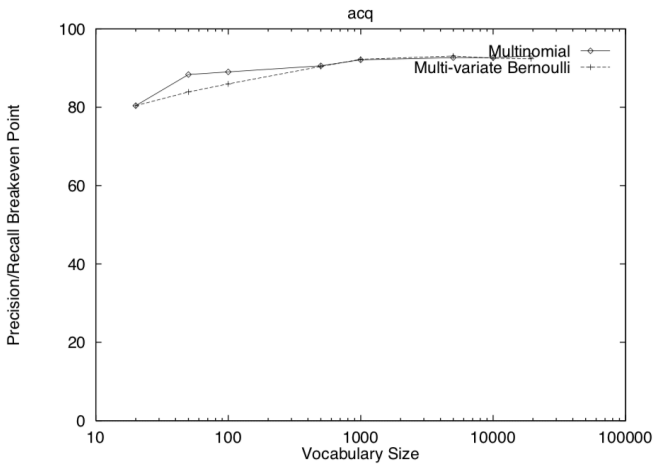

Figura 2: Precisione di predizione della classe acq in *McCallum - Nigam 1998*

I risultati ottenuti sono invece mostrati in Figura 3:

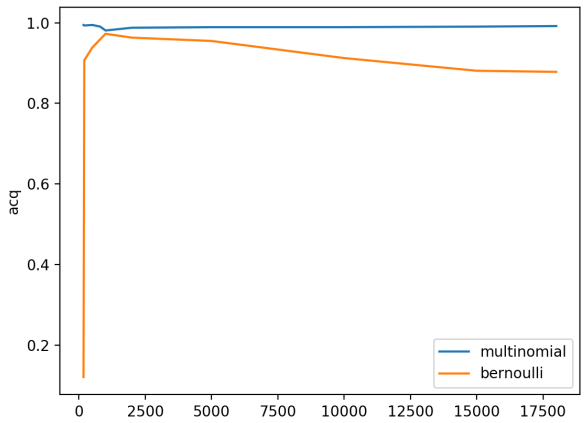

Figura 3: Preciosione di predizione sulla classe acq

I risultati ottenuti seguono quindi l'andamento inizialmente aspettato.

3.2 20newsgroups

Per quanto riguarda questo dataset, ci aspettavamo risultati di accuratezza migliori da parte di Bernoulli con un numero di parole basso. Per quanto riguarda un vocabolario più ampio, ci aspettavamo risultati migliori da parte dell'implementazione Multinomiale, come si vede in Figura 4:

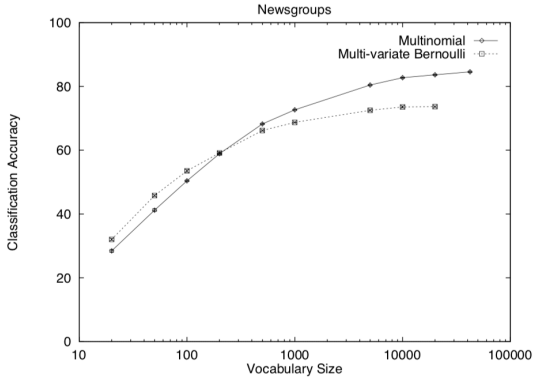

Figura 4: Preciosione di predizione sul dataset *20newsgroups* in *McCallum - Nigam 1998*

I risultati ottenuti sono stati i seguenti:

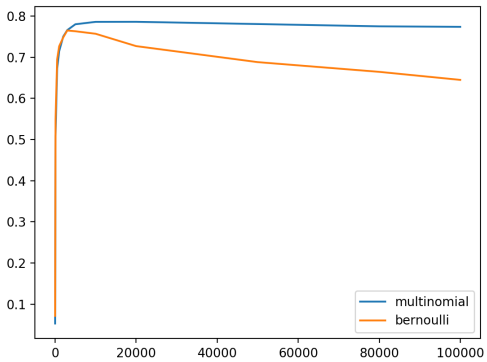

Figura 5: Preciosione di predizione sul dataset *Reuters-21578*

Come si vede in Figura 5, l'andamento è quello aspettato.