Relazione Intelligenza Artificiale

Alessio Bonacchi

February 6, 2021

Chapter 1

Obiettivo

Lo scopo del progetto è quello di implementare le versioni naive Bayes degli algoritmi per l'apprendimento e classificazione del testo:Bernoulli e Multinomiale.Gli algoritmi sono stati applicati al dataset di recensioni di film da IMDb,in modo da categorizzare queste ultime in positive e negative.

1.1 Tools Usati

Al fine di realizzare il progetto sono state utilizzate le seguenti librerie:

Sklearn: utilizzato per realizzare il bag of words con la classe *CountVectorizer* e le matrici di confusione con *metrics*.Inoltre ho utilizzato *load−files* per caricare i dataset da locale.
nltk: utilizzato per la tokenization($word-tokenize$) e lo stemming($PorterStemmer$) delle parole e per importare un set di stopwords da escludere dai documenti in analisi.

1.2 Implementazione

Di seguito saranno spiegate in modo sintetico le funzioni e classi cardine del programma.
prepareDataset: si occupa di ricevere una lista di stringhe(documenti) e ritorna una lista di liste in cui sono memorizzati per ogni riga i-esima le parole contenute nel documento i-esimo opportunamente processate.In ordine vengono eseguite operazioni di decodifica,tokenizzazione,rimozione della punteggiatura/stopwords e stemming.Tuttavia questa parte di codice è solo dimostrativa,è interessante per capire i passaggi di preprocessing che vengono fatti sui documenti.Il dataset viene preparato direttamente da **LemmaTokenizer**
setBagofWords: sia versione per multinomiale che Bernoulli.Si occupa di creare i bag of words di ogni documento.Nel caso multinomiale conta il numero

effettivo di parole presenti mentre nel caso bernoulliano denota la presenza o l'assenza delle parole(0 o 1).Questo procedimento è implementato attraverso **CountVectorizer** di *sklearn*.
report: si occupa di mostrare le matrici di incidenza e i valori di precisione,recall e accuratezza.
BernoulliClassifier e MultinomialClassifier: sono il cuore della classificazione.Si occupano del training sul dataset e la predizione sul test utilizzando rispettivamente le versioni di Bernoulli e multinomiale di naive Bayes per la classificazione.Utilizzano **calculatePrior** per calcolare le prior delle classi di documenti e due diverse versioni di **wordInClassesCounter** per poi ottenere le likelihood sulle parole.

1.3 Risultati e Conclusione

1.3.1 Risultati Empirici

I miei test sono stati eseguiti sullo stesso dataset di ImDB ma scegliendo diverse dimensioni per il train set,intesa come numero di documenti.
1)Primo risultato: 200 elementi di train 4500 parole nel vocabolario

Classificatore di Bernoulli:

Matrice di confusione:

$$\begin{pmatrix} 93 & 7 \\ 71 & 29 \end{pmatrix}$$

Accuratezza: 0.61 **Precisione sui positivi**: 0.57 **Precisione sui negativi**: 0.81

Classificatore Multinomiale:

Matrice di confusione:

$$\begin{pmatrix} 50 & 50 \\ 33 & 67 \end{pmatrix}$$

Accuratezza: 0.58 **Precisione sui positivi**: 0.57 **Precisione sui negativi**: 0.60

2)Secondo risultato: 1500 elementi di train con 12000 parole nel vocabolario

Classificatore di Bernoulli:

Matrice di confusione:

$$\begin{pmatrix} 186 & 17 \\ 188 & 24 \end{pmatrix}$$

Accuratezza: 0.51 **Precisione sui positivi**: 0.59 **Precisione sui negativi**: 0.50

Classificatore Multinomiale:

Matrice di confusione:

$$\begin{pmatrix} 102 & 101 \\ 32 & 180 \end{pmatrix}$$

Accuratezza: 0.68 **Precisione sui positivi**: 0.64 **Precisione sui negativi**: 0.76

3)Terzo risultato: 4000 elementi di train con 25000 parole nel vocabolario

Classificatore di Bernoulli:

Matrice di confusione:

$$\begin{pmatrix} 247 & 254 \\ 60 & 441 \end{pmatrix}$$

Accuratezza: 0.69 **Precisione sui positivi**: 0.63 **Precisione sui negativi**: 0.80

Classificatore Multinomiale:

Matrice di confusione:

$$\begin{pmatrix} 459 & 42 \\ 201 & 300 \end{pmatrix}$$

Accuratezza: 0.76 **Precisione sui positivi**: 0.88 **Precisione sui negativi**: 0.70

1.3.2 Conclusioni

Quello che possiamo vedere è un comportamento diverso al crescere della dimensione del dizionario: infatti l'accuratezza del classificatore multinomiale

tende ad aumentare vigorosamente via via che si aumenta il dizionario mentre il classificatore di Bernoulli aumenta si,ma di poco.In generale le performance di accuratezza del classificatore multinomiale superano quelle del classificatore di Bernoulli ad eccezione per quanto riguarda il primo risultato:è visibile infatti come quest'ultimo sia più affidabile rispetto al primo con vocabolari meno estesi,anche se,comunque, le sue performance migliorano all'aumentare del dizionario.