Indian Institute of Technology, Madras
CS5691: Pattern recognition and Machine Learning
PRML Assignment-III Report

Ashutosh Kakadiya - CS18S013,
Rajan Kumar Soni - CS18S038

October 2019

Contents

Chapter 1

Problem Model Overview

The task here is to classify the given mail into either as a spam or ham. We have implemented Multinomial Bayes classifier to attempt this problem of classification. Here is the block diagram of our Model pipeline:

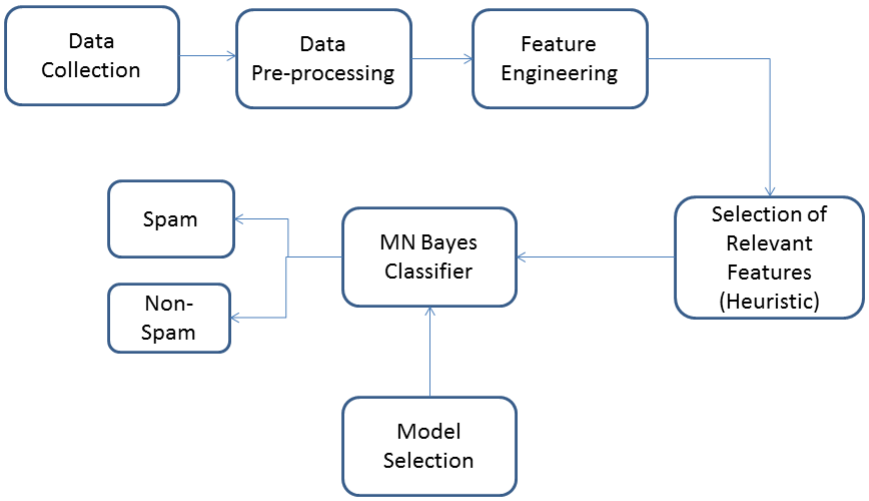

Figure 1.1: Model Pipeline

Chapter 2

Data Prepocessing

We used the spam email Enron data set which have been used as training in "Spam Filtering with Naive Bayes - Which Naive Bayes?"[2]. It consists over 30k total emails. Distribution of Spam/Ham is around $60 : 40$. The source of data set is found here Dataset link

2.1 Data Cleaning

Data cleaning is the most important and crucial step in the ML pipelines. It contains a lot of noise as well as a necessary texts and values. Addition to that most real world data sets are in unstructured format. So, for good results the good structure and removing unimportant things are also more important. We have done following things in data cleaning:

- Firstly we read the mail and remove the HTML keywords, tabs, commands and irrelevant details.
- We also removed hyperlinks, words with numbers etc.
- We used Regex expressions to remove symbols, white-space, brackets.
- Converted all the words into lowecase.
- We, I , am, to, myself, is etc stop words are very common and frequently occurring words in the text and not contain any relevant or important Things. So, we removed it using **NLTK** library.

2.2 Feature Engineering

After cleaning the text, we made a corpus of words from the cleaned mails. We made dictionary out of all words which are occurred in corpus.
We considered the frequency of word in train data set as our heuristics for feature vector. Out of around 6k words we took the 3000-4000 most occurred words as features. So, it will be very sparse vector, indeed and non zero values contain the frequency of a word occurred in given mail.

Chapter 3

Model

Based on the literature review, we came up with a decision that Multinomial naive bayes Classifier works very good when features will be discrete and goal is binary classification. In Multinomial naive bayes features are independent of each other, which is hardly true for most datasets but works well in practice!
Reason behind selecting Multinomial Nive Bayes is that feature vector contains the frequency of the word occurred in the mail. If we make feature vector only based on if the word occurred or not then we can use Naive Bayes classifier. It takes into account the number of occurrences of particular term in training documents from given class, including multiple occurrences.
Hear is the training algorithm of Multinomial Bayes Classifier.

TRAINMULTINOMIALNB($\mathbb{C}, \mathbb{D}$)
1 $V \leftarrow$ EXTRACTVOCABULARY($\mathbb{D}$)
2 $N \leftarrow$ COUNTDOCS($\mathbb{D}$)
3 **for each** $c \in \mathbb{C}$
4 **do** $N_c \leftarrow$ COUNTDOCSINCLASS($\mathbb{D}, c$)
5 $prior[c] \leftarrow N_c/N$
6 $text_c \leftarrow$ CONCATENATETEXTOFALLDOCSINCLASS($\mathbb{D}, c$)
7 **for each** $t \in V$
8 **do** $T_{ct} \leftarrow$ COUNTTOKENSOFTERM($text_c, t$)
9 **for each** $t \in V$
10 **do** $condprob[t][c] \leftarrow \frac{T_{ct}+1}{\sum_{t'}(T_{ct'}+1)}$
11 **return** $V, prior, condprob$

APPLYMULTINOMIALNB($\mathbb{C}, V, prior, condprob, d$)
1 $W \leftarrow$ EXTRACTTOKENSFROMDOC(V, d)
2 **for each** $c \in \mathbb{C}$
3 **do** $score[c] \leftarrow \log prior[c]$
4 **for each** $t \in W$
5 **do** $score[c] \mathrel{+}= \log condprob[t][c]$
6 **return** $\arg\max_{c \in \mathbb{C}} score[c]$

Chapter 4

Results

Comparisons of different classifiers on Validation accuracy can be found in following table :

Model List		
S. No.	Model Name	Validation Accuracy
1	Multinomial Naive Bayes	98.4
2	Logistic Regression	98.4
3	Perceptron	98.2
4	Support Vector Machine	98.1
5	Decision Tree Classifier	96.1
6	Linear Discriminant Analysis	94.2

We can see from the results 4 out of 6 of them approx results, So we implemented Multinomial Naive Bayes. NOTE: We have used sklearn library to select the best model to be implemented. We only implemented Multinomial Naive Bayes.

Observations:

1. Linear/Non linear models performance increases as the number of feature increase.
2. Generative and discriminative models perform similar as the data is almost linearly separable in this dimension.
3. Linear and Non linear models perform similar as the data is almost linearly separable in this dimension.

Bibliography

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[2] Vangelis Metsis, Ion Androutsopoulos, and Georgios Paliouras. Spam filtering with naive bayes - which naive bayes? 01 2006.