**Niranjan Humagain**

**Machine Learning- Project 2**

**Naïve Bayes Classifier Model for Text Classification**

**How Code Works:**

The script function reads data from trainee file estimate parameters by calculating the prior for classes use MLE and Likelihood by MAP estimation. So, each word for each label (newsgroup) will have probability of being on that newsgroup.

Now, I import test files and labels and calculate probabilities for all classes for each document. Each document will have all vocabulary words with counts. The document is classified for that class having the maximum probability. Accuracy rate is then calculated as the fraction of total number of correctly classified documents to the total number of documents in the test dataset.

Also the there is another script function to measure the words with the highest measure on which categories most rely on.

**Accuracy obtained for different beta values:**

***BETA VALUE ACCURACY RATE (%)***

|  |  |
| --- | --- |
| 0.00001 | 82.74 |
| 0.0005 | 83.30 |
| 0.002 | 83.31 |
| 0.01 | 83.61 |
| 0.5 | 82.34 |
| 1 | 80.5 |

Not too high and not too low values for beta works well for the program, as it does not let the trainee data to over fit.

**Question 1:**

Answer: Too many parameters to estimate. e.g.1000 documents, each 1000 words long, where each word comes from a 50,000 word vocabulary . This accounts to: no. Of classes \*(2^no of words -1)=billions of parameters to estimate which is impractical for most of the domains.

**Question 2**:Codes are in MATLAB scripts attached.

**Question 3:**

Yes.

As we can see confusion matrix,

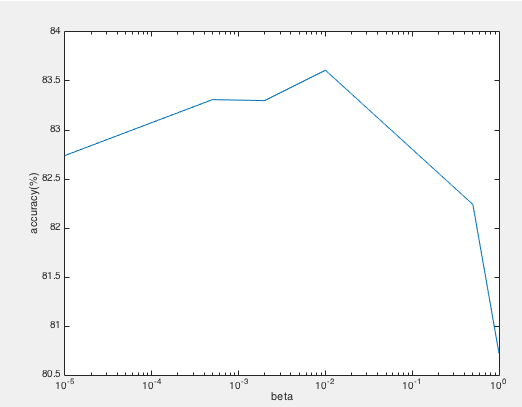
I) It has highest of 12109 documents that was supposed to be in talks.politics.misc but are categorized in talks.politics.guns.

ii) Similar, 4548 documents were originally in comp.windows.x but were categorized under comp.graphics.

iii) 5074 documents were originally in talk.religion.misc but were categorized under alt.atheism.

This statistics gives us an idea that two newsgroups that are related are more confusing. OR we can say two newsgroups that have the same root of origination had more documents miscategorized. Eg. ii) Shows that words that appeared in comp.windows.x could appear on comp.graphics because both are related to Information Technology. No. iii) Shows that talk.religion.music and alt.atheism both come under religion. There can be words that could be same appearing in both newsgroups.

**Question 4:**



The graph shows the accuracy rate drops both for higher and lower values of beta. The lower and higher values of our beta over fit our training data. This does not help to generalize new data. Hence, we got lower accuracy rate.

**Question 5:**

For each word, we can find the maximum number of it counts for different labels. After that, we choose those words, which have high frequency in one or few labels. That is the word distribution that has high peak value for only one or few labels. This tracking of information will give us the words that the classifiers most rely on. This is to say we need to find the words that have maximum likelihood as likelihood is proportional to word count.

**Question 6:**

Code is implemented in MATLAB.

**Question 7:**

The words we are using for classifying cannot provide much information on categories we want to assign to. It’s because these words are nearly equally likely to be in many of the relation categories as we saw in confusion matrix and question 3 answer. Example, the word “religion” might appear frequently in some label does not mean that label is about religious information. It’s more important to analyze the word in context, which will give meaningful information. But our model does not support this.