Name: Janani Eshwaran

Net ID: jxe130830

Assignment 2 Write-up

1) Naive Bayes:

Accuracy of Test Data without removing Stop words:

Ham Class: 96%

Spam Class: 98%

Accuracy of Test Data with removal of Stop words:

Ham Class: 95%

Spam Class: 98%

- After removing stop words there is reduction in the accuracy of the ham class while the accuracy of spam remains the constant .This is because spam files contains more amount random words and less of stop words so removing stop words doesn't alter the accuracy of spam class.
- However, in ham class, the occurrence of stop words are more as compare to spam so it causes some over-fitting of data, removing stop words gives correct and stable accuracy.
- Stop words in ham class will cause to increase the probability of that class.

2) Logistic Regression:

For 100 iteration: Initial Weight range: **0 to 3**, Learning rate: **0.001.**

Maximum accuracy achieved for lambda: **1.8** (highlighted)

Accuracy on Test Data before removing Stop words		
Lambda	Accuracy on Ham Class	Accuracy On Spam Class
1.2	59	90
1.4	60	91
1.6	65	86
1.8	72	87
2	55	89
2.2	55	87
Accuracy on Test Data after removing Stop words		
Lambda	Accuracy on Ham Class	Accuracy On Spam Class
1.2	70	86
1.4	65	90
1.6	74	85

1.8	74	88
2	62	88
2.2	66	83

- In logistic regression even after removing stop words accuracy of spam remains almost even after removing stop words.
- However in case of Ham the value of accuracy increases this is because there we are calculating the probability of that particular token in terms of weights unlike simply counting the number of occurrence of that word. Hence in case of logistic regression decrease in the accuracy after removing the stop words is giving correct value.

<u>Hard-Limit:</u>

As per the above data as the iteration increases the accuracy is going to increased so I have kept hard limit on the iteration as 100.

For 100 iteration the maximum accuracy is: 72% without removal of stop words and 74% after removal of stop words.

<u>Lambda ƛ</u>

I tried for the five above mentioned values of lambda. And I observed that as the value of lambda increases the accuracy is increases.

This is because in case of weight update formula as the lambda value increases to second term is tending and third term trying to nullify each other and weight is remains almost constant with slight fluctuation.

This is called as the convergence of weight. As the weights becoming towards the saturation accuracy goes on increasing. This is because weights are stable.

<u>Laplace Smoothing:</u>

If particular word is new to that class this will tend the data likelihood to zero and hence will give wrong results. To avoid this we are adding “1” in numerator and denominator which will help to give probability as 1 /total number of words occurring in document.

In case of **<u>logistic regression</u>** while calculating the probability with following formulas. Summation of weight has taken with log values as it is exponential to “e”.

$$P(Y=1|X) = \frac{\exp(w_0 + \sum_{i=1}^{n} w_i X_i)}{1 + \exp(w_0 + \sum_{i=1}^{n} w_i X_i)}$$

$$P(Y=0|X) = \frac{1}{1 + \exp(w_0 + \sum_{i=1}^{n} w_i X_i)}$$