Assignment2 Machine Learning – ARC180006

Assumption/Constraint:

1) In the given document only words with digit or words or numbers or ‘are considered.

2) The Stop words were filtered using NLTK library (discussed with in the class).

3) The accuracy is calculated over all the dataset for both spam and ham

The Naïve Bayes Report:

Below are the results of Naïve Bayes Classification:

Overall Accuracy of Naive Bayesian without stop words filtration is **94.35%**

Overall Accuracy of Naive Bayesian with stop words filtration is **94.14%**

Stop Words Filtration	Accuracy (Ham)	Accuracy (Spam)
No	96.84%	87.69%
Yes	96.84%	86.92%

The Accuracy of the Ham is greater than accuracy of spam, the reason is that ham prior is higher than spam prior, because there are more 3 time more ham files than spam files, so there is more chance that file be considered ham over spam. So, most files will be classified as Ham.

With elimination of stop words, the overall accuracy of Naive Bayes was slightly decreased. While the accuracy of ham dataset was same but spam accuracy was slightly decreased. It is likely that at least some of the stop words which were removed by **NLTK** package is predictive of the spam category. Its count is more in the spam files which might have played role in spam classification.

The Logistic Regression Report:

Overall Accuracy of Logistic Regression without stop words filtration is **93.31%**

Overall Accuracy of Logistic Regression with stop words filtration is **94.56%**

Learning Rate = 0.001, Lambda = 0.001, no of epoch = 300

Stop Words Filtration	Accuracy (Ham)	Accuracy (Spam)
No	94.83%	89.23%
Yes	97.84%	87.69%

Overall Accuracy of Logistic Regression without stop words filtration is **93.31%**

Overall Accuracy of Logistic Regression with stop words filtration is **94.56%**

Learning Rate = 0.001, Lambda = 0.0001, no of epoch = 300

Stop Words Filtration	Accuracy (Ham)	Accuracy (Spam)
No	94.83%	89.23%
Yes	97.12%	87.69%

Overall Accuracy of Logistic Regression without stop words filtration is **93.01%**

Assignment2 Machine Learning – ARC180006

Overall Accuracy of Logistic Regression with stop words filtration is **94.35%**

Learning Rate = 0.001, Lambda = 1.0, no of epoch = 300

Stop Words Filtration	Accuracy (Ham)	Accuracy (Spam)
No	94.83%	88.46%
Yes	97.12%	86.92%

The above result was obtained by Learning rate of 0.001, chosen as low as possible to not miss the Gradient ascent point convergence point. The regularization parameters used is 0.001 and no of epoch was constrained to 300

With elimination of stop words, the overall accuracy of was slightly increased. The accuracy of ham files was increased and accuracy of spam file detection was decreased. It should be because of no of training example of ham were greater than spam files.

I have tried 3 different values of lambda (0.001, >0.001 and <0.001) Lambda parameter among the variety of possible values the weight vector within certain range

The number of iterations chosen ranged from 100 to 500 as we increase the range to more than 300, the increase in the weight is very less for each iteration (but takes more time to converge). so for reducing time I have taken 100 iterations.