Assignment lll

Question 1: Naïve Bayes

I have used Naïve Bayes formula along with laplace smoothing: (Bag of words).

With stopwords:

Accuracy in ham test folder:

100.0

Accuracy in spam test folder:

24.615384615384617

Overall Accuracy

79.45492662473794

Question 2: Logistic Regression

Converged the value of w for 100 iterations

W= [1.3 1.3 1.3 1.3……]n times where n is the vocab length.

1) Lambda =0.01
   Accuracy with stop words = 82.5021%

2) Lambda=0.1
   Accuracy with stop words = 82.5021%

3) Lambda=10
   Accuracy with stop words = 82.5021%

Question 3: Without stopwords:

Naïve Bayes:

Accuracy in ham test folder:

99.71181556195965

Accuracy in spam test folder:

33.84615384615385

Overall Accuracy

81.76100628930818

There is a mild increase in the accuracy after removing stop words. The reason for this is that stop words do not have any meaning and does not hence help the classifier as they are equally present in ham and spam files. Hence after we remove stop words it improve the accuracy.

RESULT: There is a marginal increase in accuracy of the naïve Bayes classifier after removing stop words.

Logistics Regression:

Number of iteration:500

Test cases for different values of lambda:

4) Lambda =0.01
   Accuracy with stop words = 82.5021%
   Accuracy without stop words = 82.5021%

5) Lambda=0.1
   Accuracy with stop words = 82.5021%
   Accuracy without stop words = 82.5021%

6) Lambda=10
   Accuracy with stop words = 82.5021%
   Accuracy without stop words = 82.5021%

As we can see that the accuracy under Logistic Regression is better than Naïve Bayes.