CS6375.001 Arijeet Roy axr165030@utdallas.edu

Spam Classification: Naïve Bayes and Logistic Regression

Naïve Bayes:

- *Before removal of stop words:*
  Spam Accuracy (Ratio of number of correctly classified spam files and total number of spam files): 88.46%
  Ham Accuracy (Ratio of number of correctly classified ham files and total number of ham files): 99.13%
  Overall Accuracy (Ratio of number of correctly classified spam + ham files and total number of spam + ham files): 96.23%

- *After removal of stop words:*
  Spam Accuracy (Ratio of number of correctly classified spam files and total number of spam files): 87.69%
  Ham Accuracy (Ratio of number of correctly classified ham files and total number of ham files): 99.13%
  Overall Accuracy (Ratio of number of correctly classified spam + ham files and total number of spam + ham files): 96.02%

After removal of stop words, the overall accuracy remains approximately the same with a very slight reduction (~0.23%). This is because of the distribution of stop words being selected to filter out.

Logistic Regression:

Accuracy is observed on bases of various factors such as learning rate(eta), penalty regularization term(lambda) and number of iterations before convergence. It takes significant time to complete the given number of iterations since the dictionary contains a lot of words.

1. Number of iterations: 20
   Learning rate: 0.01
   Regularization term: 0.01

- *Before removal of stop words:*
  Spam Accuracy (Ratio of number of correctly classified spam files and total number of spam files): 75.38%
  Ham Accuracy (Ratio of number of correctly classified ham files and total number of ham files): 87.64%
  Overall Accuracy (Ratio of number of correctly classified spam + ham files and total number of spam + ham files): 84.30%

- *After removal of stop words:*
  Spam Accuracy (Ratio of number of correctly classified spam files and total number of spam files): 87.69%
  Ham Accuracy (Ratio of number of correctly classified ham files and total number of ham files): 85.91%

Overall Accuracy (Ratio of number of correctly classified spam + ham files and total number of spam + ham files): 86.40%

The value of parameter learning rate indicates how fast we approach the convergence point. The regularization factor controls the penalty we add to the weights. Since it takes more computation time to find the exact convergence point, we restrict the number of iterations and instead try with a relatively high value of learning rate.
After removal of stop words, we see a slight increment in the overall accuracy (~2%). It depends on the list of stop words we are using which restricts the vocabulary size.

Similarly, for different values of learning rate and regularization parameter

2. Number of iterations: 20
   Learning rate: 0.05
   Regularization term: 0.05

- *Before removal of stop words:*
  Spam Accuracy (Ratio of number of correctly classified spam files and total number of spam files): 69.23%
  Ham Accuracy (Ratio of number of correctly classified ham files and total number of ham files): 87.64%
  Overall Accuracy (Ratio of number of correctly classified spam + ham files and total number of spam + ham files): 82.63%

- *After removal of stop words:*
  Spam Accuracy (Ratio of number of correctly classified spam files and total number of spam files): 87.69%
  Ham Accuracy (Ratio of number of correctly classified ham files and total number of ham files): 93.67%
  Overall Accuracy (Ratio of number of correctly classified spam + ham files and total number of spam + ham files): 92.05%