CS6375.003-Machine Learning

Assignment 3 Report – Text Classification

By – Srivastchavan Rengarajan(sxr190067)

Multinomial Naive Bayes algorithm-

Testing Data:

Files in Ham Folder=348
Files in Spam Folder=130
Total files = 478

Results (before removing stop words)–

Ham Folder-

Number of emails correctly classified as Ham=340
Number of emails wrongly classified as Spam=8
Naïve Bayes model Accuracy for Ham folder=97.70%

Spam Folder-

Number of emails correctly classified as Ham=110
Number of emails wrongly classified as Spam=20
Naïve Bayes model Accuracy for Ham folder=84.62%

Total(Both Folders)-

Accuracy of Naïve Bayes model for text Classification =94.14%

Results (after removing stop words)–

Ham Folder-

Number of emails correctly classified as Ham=343
Number of emails wrongly classified as Spam=5
Naïve Bayes model Accuracy for Ham folder=98.56%

Spam Folder-

Number of emails correctly classified as Ham=108
Number of emails wrongly classified as Spam=22
Naïve Bayes model Accuracy for Ham folder=83.08%

Total Accuracy-

Accuracy of Naïve Bayes model for text Classification (All Folders) =94.35%

The Naïve Bayes algorithm has better accuracy for ham email classification than the spam email classification. This is true in both cases where stop words are present as well as removed. This means that the algorithm mostly never falsely predicts ham emails as spam which is good, but few spam emails end up with the ham emails due to the algorithm performing poorly in spam classification.

Also, the removal of stop words improved the ham email classification by around 1% whereas decreased the performance of spam classification by around the same margin. The overall accuracy remains more or less the same with a mild increase.

MCAP Logistic Regression algorithm-

Testing Data:

Files in Ham Folder=348
Files in Spam Folder=130
Total files = 478

Learning rate=0.001

Iterations hard limit=100

Various Lambda Values=0.001,0.005,0.01,0.05,0.1,0.5,1,100

Results (before removing stop words)–

	Various Lambda Values							
	0.001	0.005	0.01	0.05	0.1	0.5	1	100
Ham Accuracy	94.82	94.82	94.82	94.82	94.82	94.82	94.82	97.70
Spam Accuracy	90.00	90.00	90.00	90.00	90.00	90.00	90.00	39.23
Total Accuracy	93.51	93.51	93.51	93.51	93.51	93.51	93.51	81.79

Results (after removing stop words)–

	Various Lambda Values							
	0.001	0.005	0.01	0.05	0.1	0.5	1	100
Ham Accuracy	97.41	97.41	97.41	97.41	97.41	97.41	97.41	89.08
Spam Accuracy	90.00	90.00	90.00	90.76	90.76	90.76	90.76	93.84
Total Accuracy	95.39	95.39	95.39	95.60	95.60	95.60	95.60	90.37

The Logistic regression algorithm has better ham email classification accuracy than the spam email classification. But the difference between the two is 5-7% as compared to the Naïve bayes algorithm which had 13-14% difference. Hence, we can conclude that the Logistic regression performs well in both ham and spam classification.

The change in various values of lambda or regularization parameter from 0.001 to 1 had very little effect in the performance of the algorithm. The result in both cases is almost the same as seen in the table above. But when the lambda value is very high for example 100, there is a considerable decrease in performance in the case of test data with stop words and small decrease in performance in case of test data without stop words. Only the spam classification shows improvement for large values of lambda after removing stop words but performs very poorly if stop words are not removed.

Removing stop words improves the ham email classification accuracy but has no effect in spam email classification. The overall accuracy of the logistic regression algorithm improves 2-3% after removing the stop words.

Finally, comparing both the algorithms, Naive Bayes has better ham email classification accuracy than the Logistic regression algorithm before and after removing stop words whereas Logistic regression performs better than Naïve Bayes in spam email classification. Overall, both algorithms have nearly similar total accuracy.