Cody Herberholz
CS445
2/13/2017

Support Vector Machines

Experiment 1

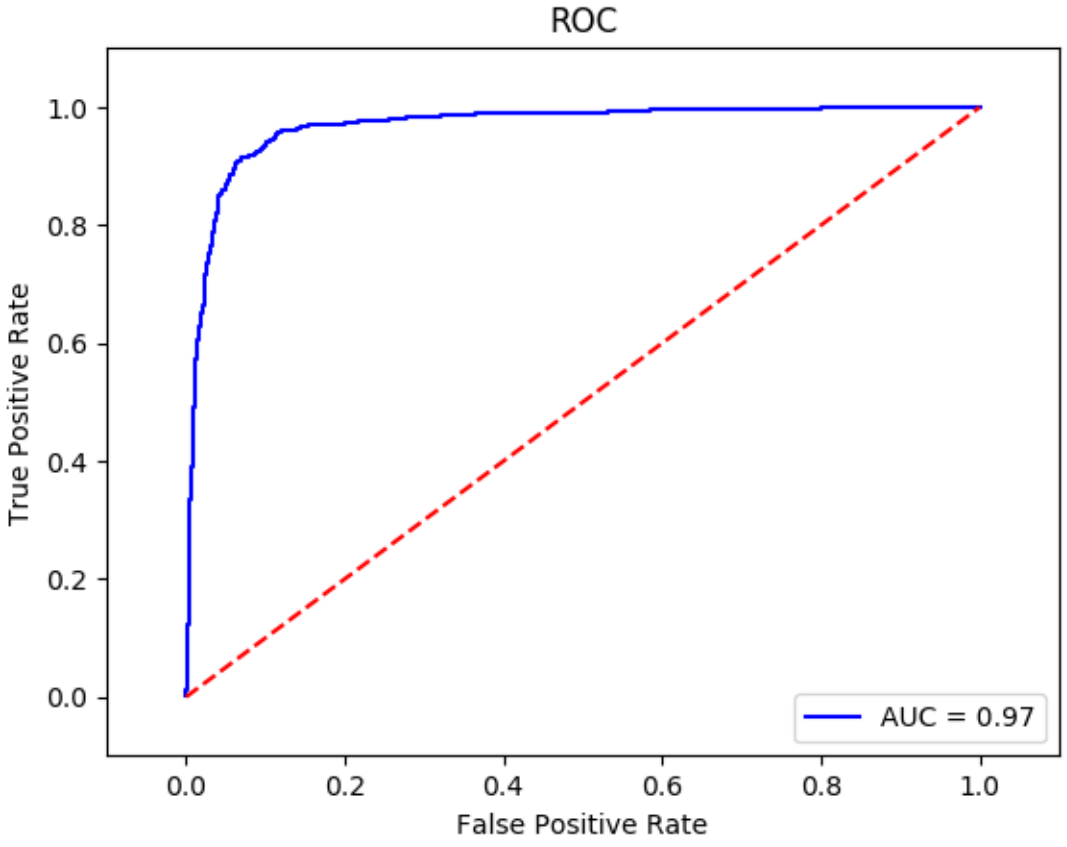
Description

The first experiment is to use a support vector machine library to provide classification data on whether or not test data from the provided spam database is actually spam or not. For this project I used the svm package from sklearn inside of python.

Accuracy- 0.926

Precision- 0.912

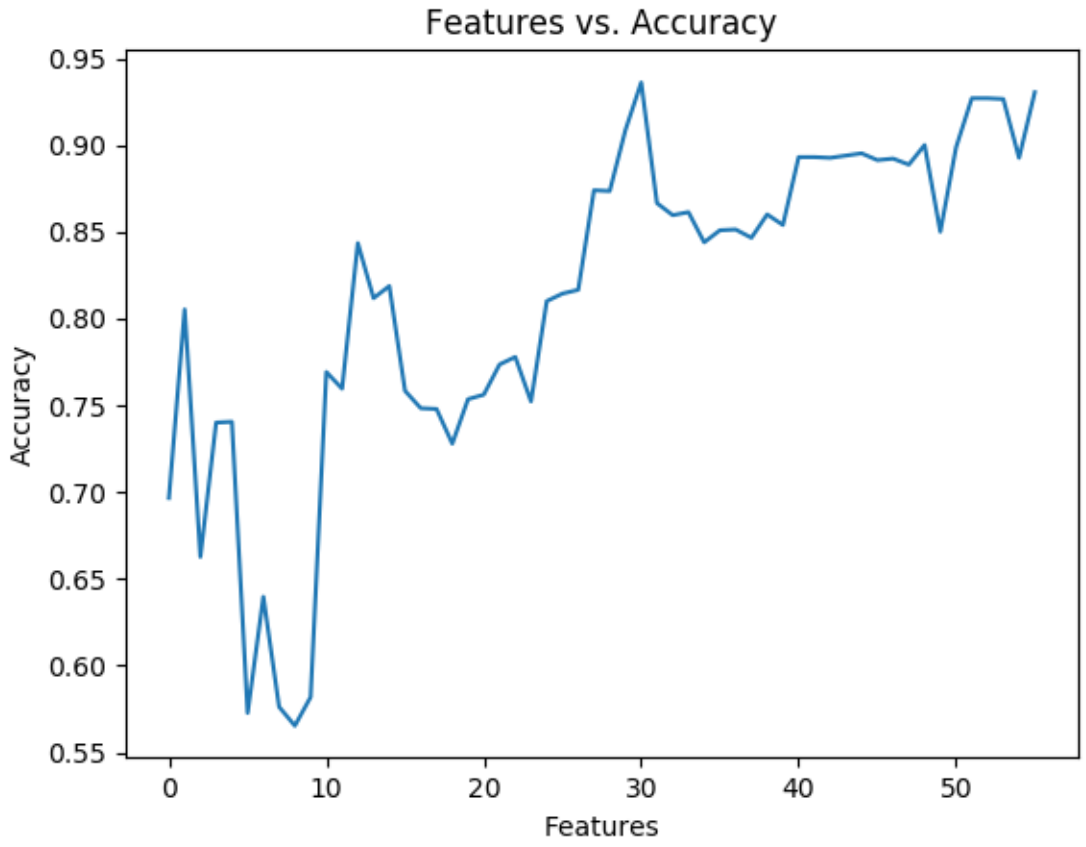
Recall – 0.898



Experiment 2

Description

The purpose of experiment 2 was to perform feature selection based upon the weight vectors. The top five features of the experiment from the model where from element 7, 25, 27, 41, and 53.



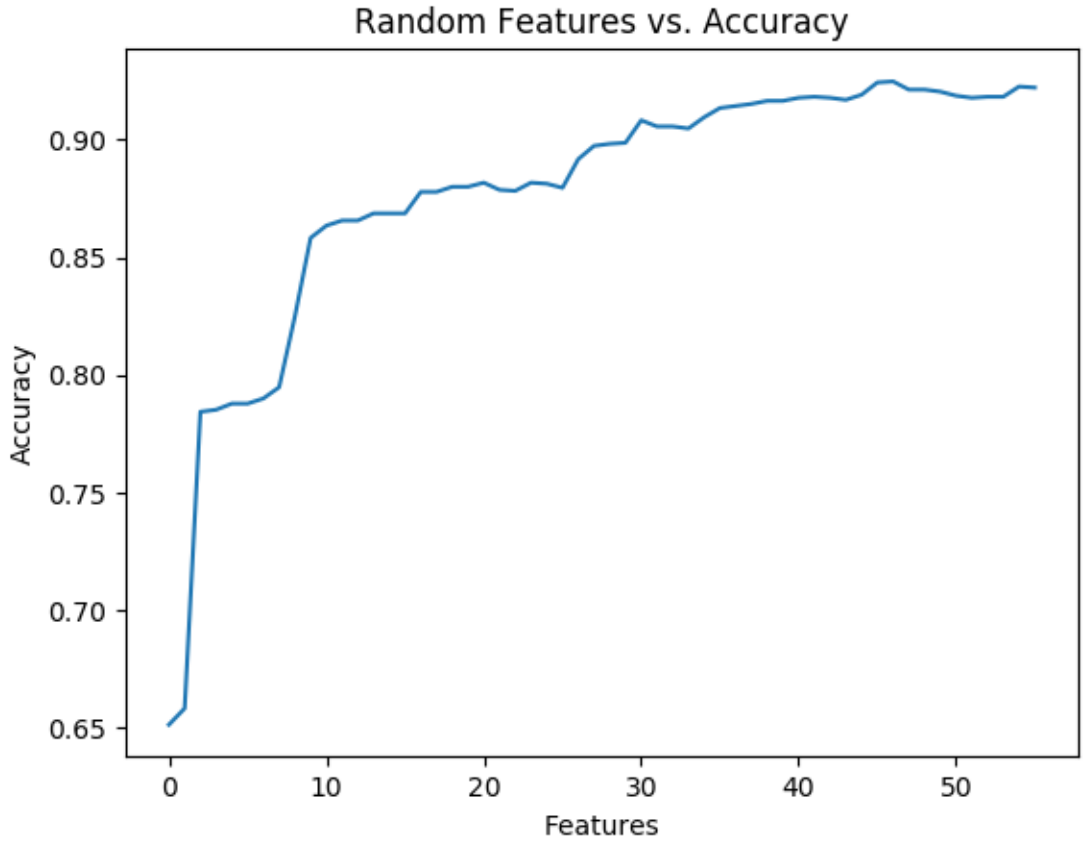
Summary

What we can take away from the plotted data is that with 2 of the best features selected we have a classifier that can correctly classify 80% of the time but when added on after, that accuracy dips quite a bit. It isn't until about 12 and 29 features that the plot peaks again and shows signs of increased accuracy. Feature selection is beneficial but I is a hyperparameter that must be tweaked in order to achieve maximum performance. This can be seen when selecting 9 of the best features as opposed to 30 features.

Experiment 3

Description

The purpose of experiment 3 was to perform feature selection randomly and then compare the results to experiment two's results.



Summary

What we can take away from the above plotted data is that when randomly choosing features is that while the accuracy is low in the beginning with a small amount of random features, the increase of accuracy is much steadier with a random selection algorithm. When compared to the previous experiment, when selecting the features with the highest weights first we get a higher accuracy then with the current plot but when adding to the data of experiment 2 the following features start acting like noise and dropping the accuracy, due to the best features already have being select. The random feature selection does not suffer as bad as a hyperparameter learning curve due to its random nature. The more features selected the better the accuracy until it finally flat lines.