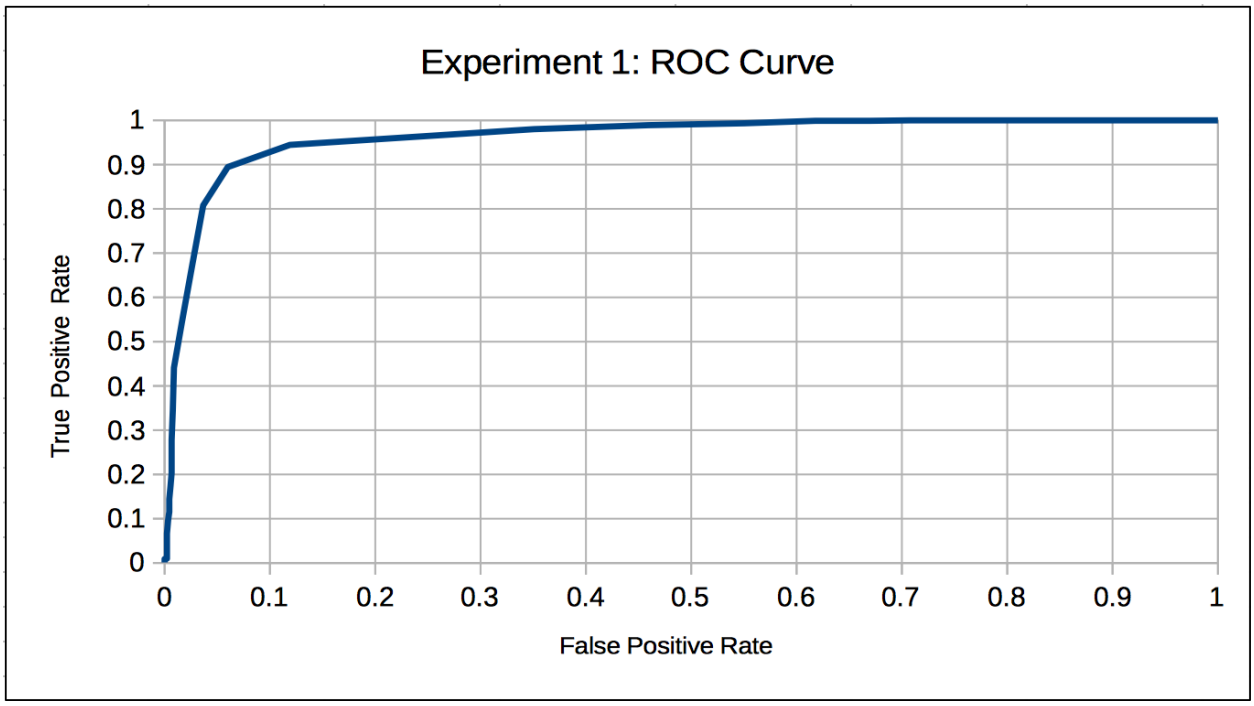
Tyler Sorg
CS445 Machine Learning
Project 3: Support Vector Machines and Spambase

Experiment 1:

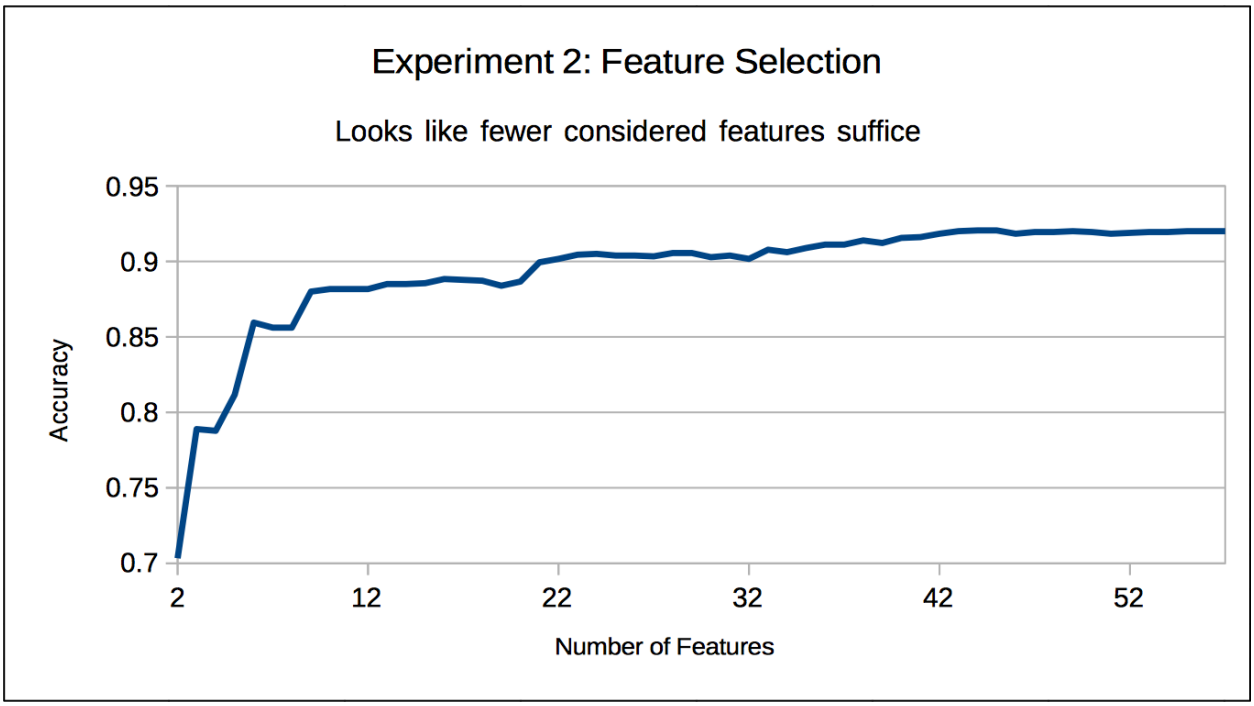
I used the scikit-learn SVM package for python. The best C value I obtained changed each time I ran the program. I'm suspicious of this part of the algorithm because the accuracies were so similar between the ten classifiers. I opted to compare the predicted label to the actual label instead of using an arbitrary threshold and the decision function scores, which could be a source of error. I used that method again in the other experiments, but I used the scores and thresholds for the rest of experiment 1. Overall, I wasted a lot of time by not knowing how to use the library's tools or which tools were available. It was educational, though.

The reported accuracy was 92%, the reported precision was 93.1%, and the reported recall was 90.7% for the final learned model.

Here is the ROC curve generated by using 200 evenly spaced thresholds:



Experiment 2:

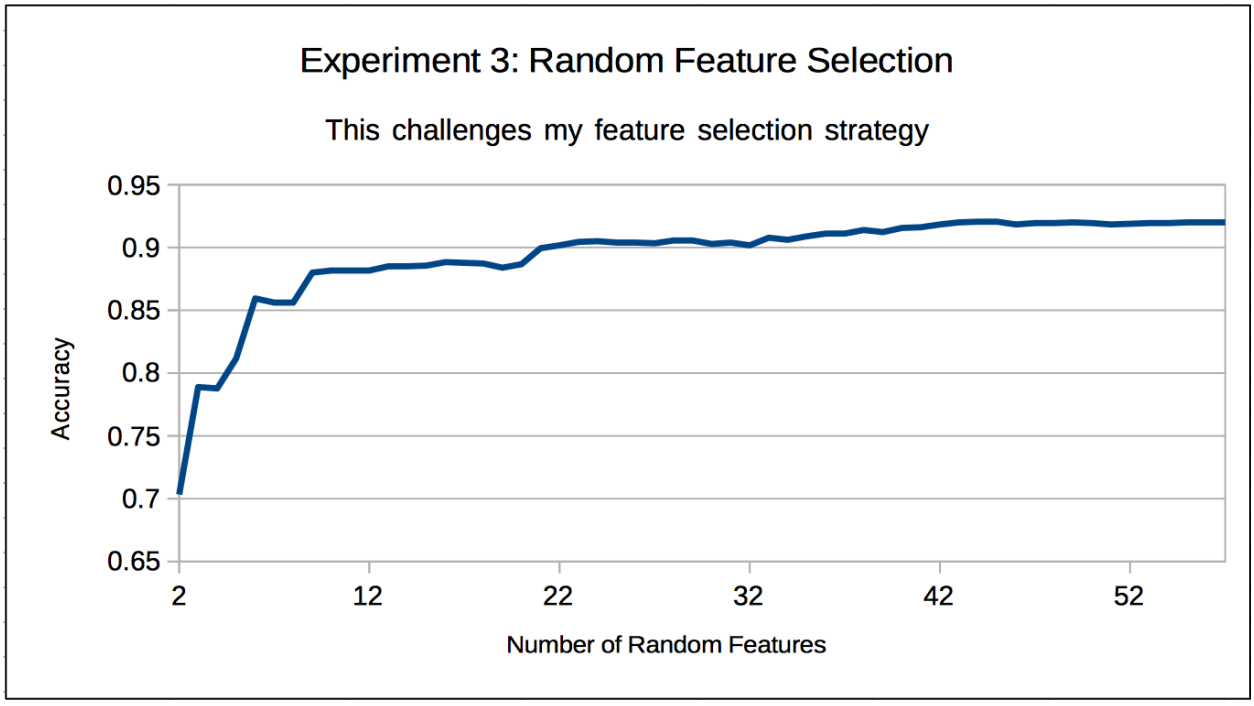


The top five features were 45, 55, 52, 24, and 6. The last one varied. I might be counting wrongly because the names could be out of order. From the URL in the assignment, I count 56,0,1,2,…,55 top to bottom:

6: word_freq_remove
24: word_freq_hp
45: word_freq_table
52: char_freq_#
55: capital_run_length_total

It is apparent that increasing the number of dimensions to consider when training has diminishing returns (at least with a linear kernel?). Limiting the ‘m’ most significant magnitudes of weights saves resources with potentially little cost to accuracy of classification. However, does choosing features with the ‘m’ largest weights make a difference? The next experiment addresses that.

Experiment 3:



The accuracies between experiments 2 and 3 are comparable. This suggests that training on fewer dimensions is what should be emphasized and not which vectors have the largest magnitudes. The larger magnitudes do not necessarily imply likelihood of spam. Perhaps the frequency of a possibly-spam word varies, and the weights reflect that?