Abbie Jones
CS 545: Spring, 2017
5/9/17

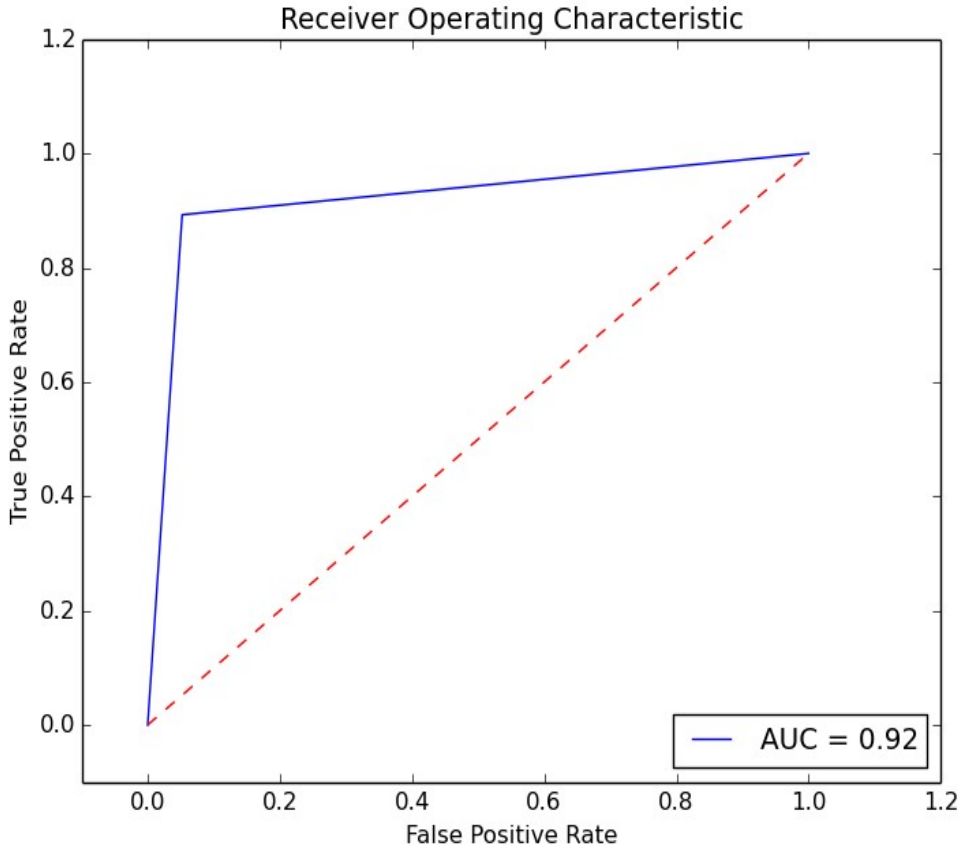
Experiment 1:

I wrote my project in Python and used **pandas** and **numpy** for data manipulation in preparation for support vector machine analysis with **scikit-learn.**

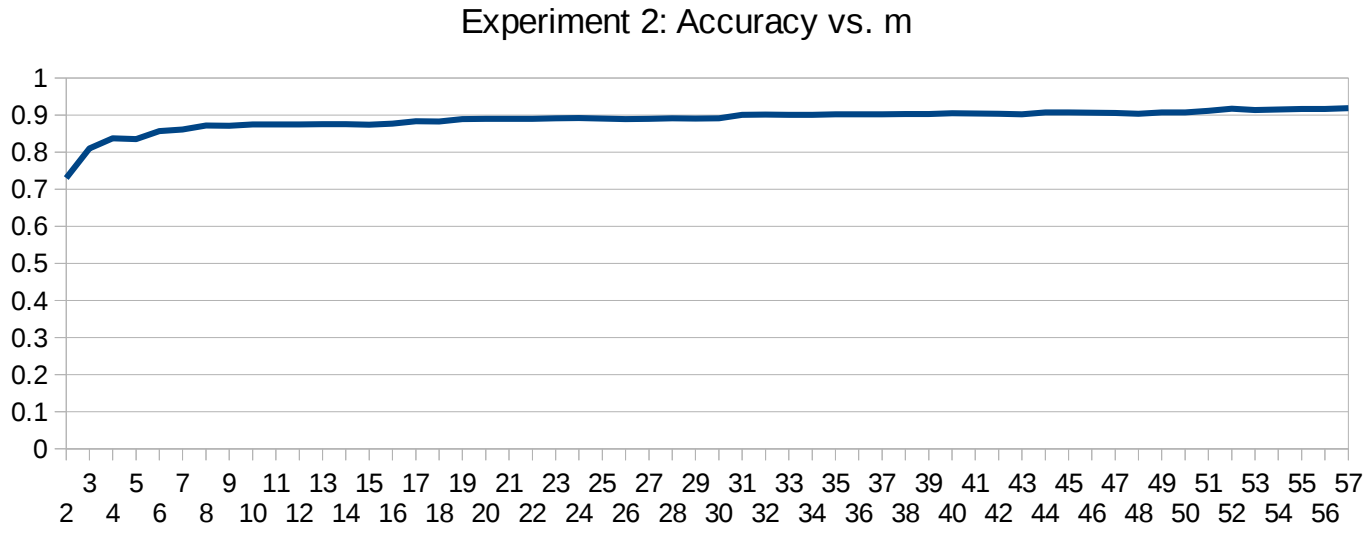
After training and testing the model on my data:

accuracy: **0.928292046936**
precision: **0.937570942111**
recall: **0.882478632479**

ROC curve:



Experiment 2:



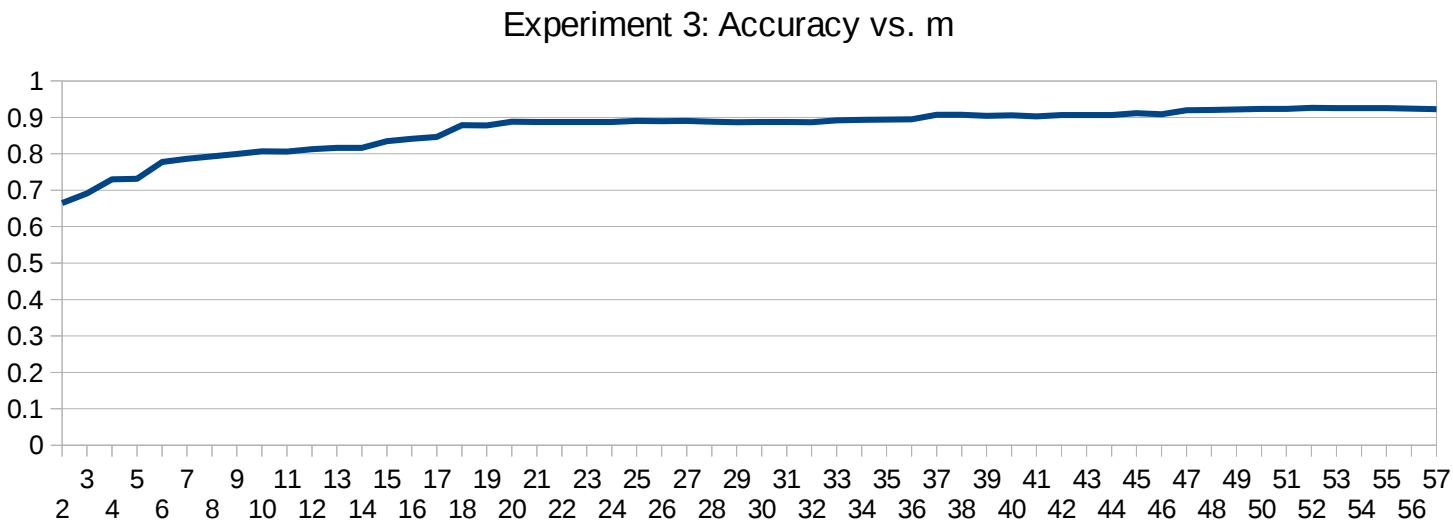
The top 5 most heavily weighted characteristics were:

1. char_freq_$
2. word_freq_remove
3. word_freq_free
4. word_freq_3d
5. char_freq_#

The first and fifth correspond to the frequency of the characters “$” and “#” being in e-mails with spam. This makes sense to me—spam is often rife with nonsensical symbols that really have no place in a relevant e-mail. The second through fourth characteristics are an overrepresentation of the words “remove”, “free”, and “3d”. The reason behind these three words being so commonly seem in spam is a little more opaque to me.

With fewer characteristics to assess the data in the training session, the accuracy of the model on the testing model is greatly reduced. With only two characteristics to evaluate the data, the accuracy rate is only at 72%. It pretty rapidly increases from 72% to around 88% as we progressively add characteristics (with about 15 characteristics to evaluate now). For the remainder, there is a slow creep up from 88% to a terminal accuracy of ~92%. It seems that after around the 15 characteristic mark, it takes a lot longer for the support vector machine to increase in accuracy. The acceleration slows down.

Experiment 3:



The difference between random feature selection and feature selection based on highest-weighted characteristic is pretty significant. Whereas with only 2 features in the hand-picked SVM run, we were at an accuracy of around 72%, we start here with random selection at around ~66%. Like before, we see a pretty significant acceleration in accuracy as we continue to add features for the first 15 or so characteristics. In fact, the acceleration is much greater here. Again we get to around 88% by the 15th or so feature added. Then the graphs look pretty similar. This makes sense to me; as the random selections grow, we probably see a lot more overlap between the features selected in the hand-picked experiment 2; they should start to match each other more and more. It also makes sense that the initial accuracies should be so different. We aren't necessarily getting a very highly weighted feature to assess our spam database in experiment 3; whereas, we know we are in experiment 2.