**CS 445 - Machine Learning**

**Homework 2: Neural Networks**

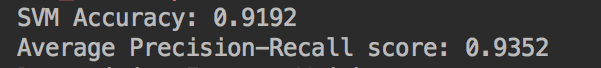
**Overview**

For this assignment, we were given the Spambase data set and the task of using an SVM to classify the collection of emails as spam or not. We were able to use existing libraries for this task, so implementing this was rather simple. By splitting the email data in half, one half for training and the other for testing, we can see that an SVM uses the features given to determine whether an email is spam or not. These features are in the form of words that occur often in spam email, and therefore act as good indicators as to the intent of the email.

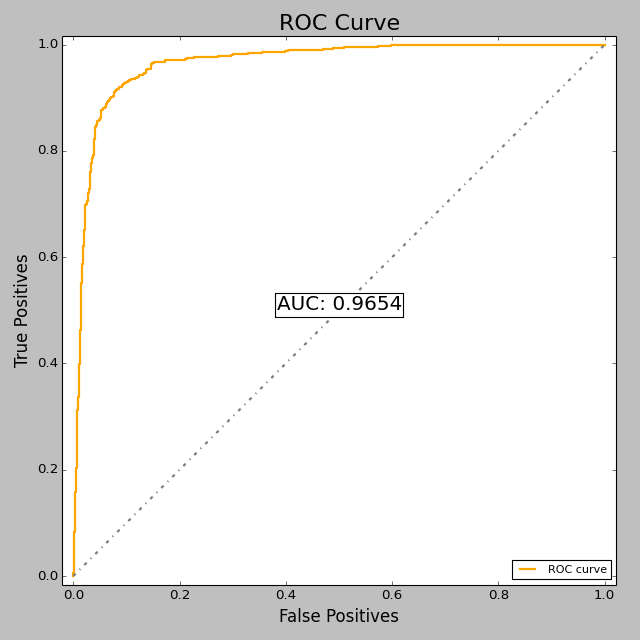
**Experiment 1**

SMV Package used: SKLearn’s SVM library

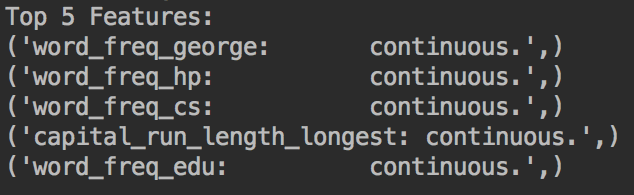
Accuracy, Precision, Recall Scores: SKLearn’s Metrics library had a Precision-Recall function



ROC Curve:

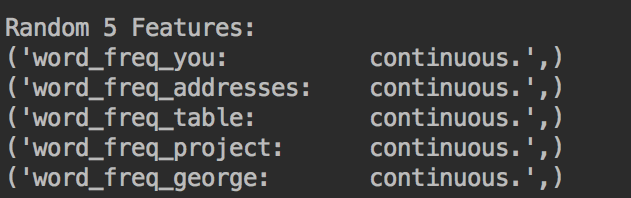
****

**Experiment 2**

****

The top 5 feature weights were as follows: “George”, “hp”, “cs”, long words in all caps, and “edu”. As seen in the plot shown below (blue line), knowing the top weighted features gives the SVM a much more accurate ability to decipher the emails, with additional features not improving the machine too much. At 10 features, the SVM has an approximate 90% accuracy. The remaining 40+ features only give the machine an additional ±4%. The slight dips in the accuracy after the first ten features indicate that too many features can essentially confuses the SVM.

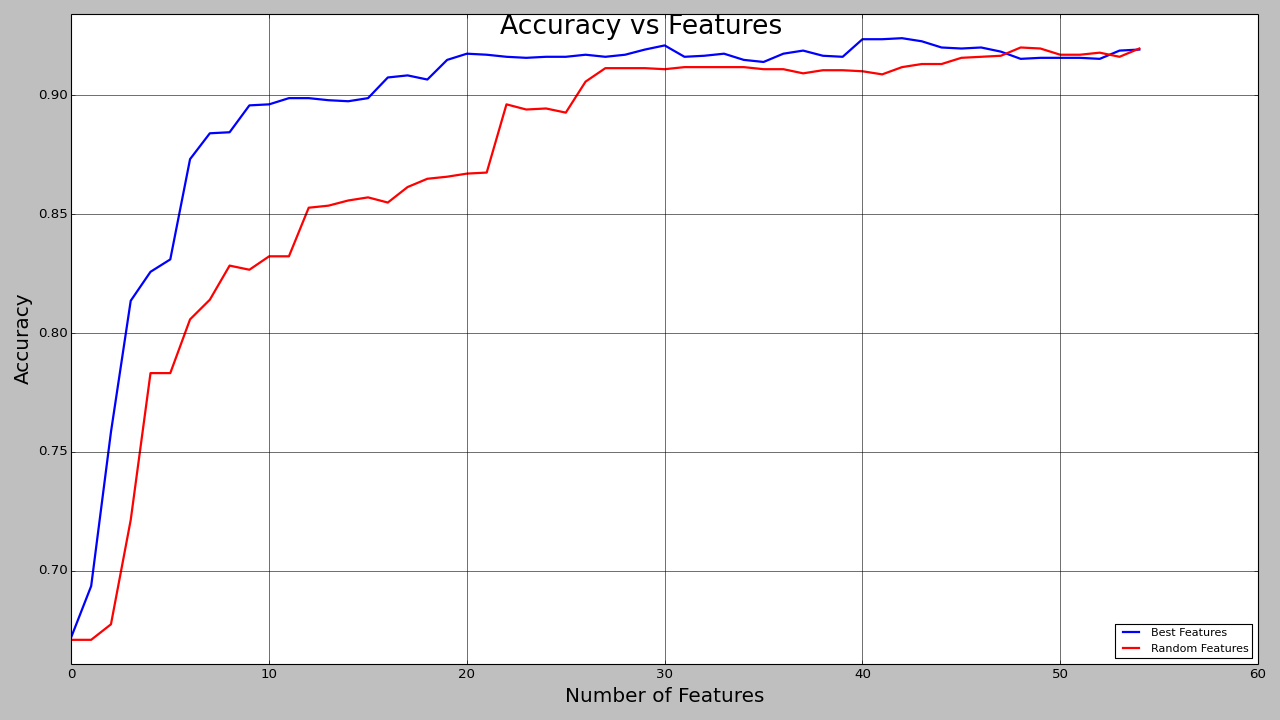
**Experiment 3**

****

Using randomly ordered features for experiment 3 (oddly enough, George showed up in the top 5 of this iteration as well), we can see that choosing the features randomly does not benefit the SVM. As shown below in the plot (via the red line), we can see that it takes the random selection an adition 15 features to reach the approximate 90% accuracy, compared to the weighted selection. And only until approximately 47 features have been accounted for do the two SVMs converge.

**SEE NEXT PAGE FOR EXPERIMENT 2 & 3 PLOTS**

**Experiments 2 and 3 Plot comparison**

****

**Conclusion**

Weighted feature selection is initially more accurate than randomly selected features, but this difference grows more negligible as more features are added to the SVM (as expected).