Homework-3

In this homework we are using spambase database to detect whether an email is spam or not.it is having 57 feature to train and test on. We have divided the data into two halves for training and testing purposes.

Experiment 1-

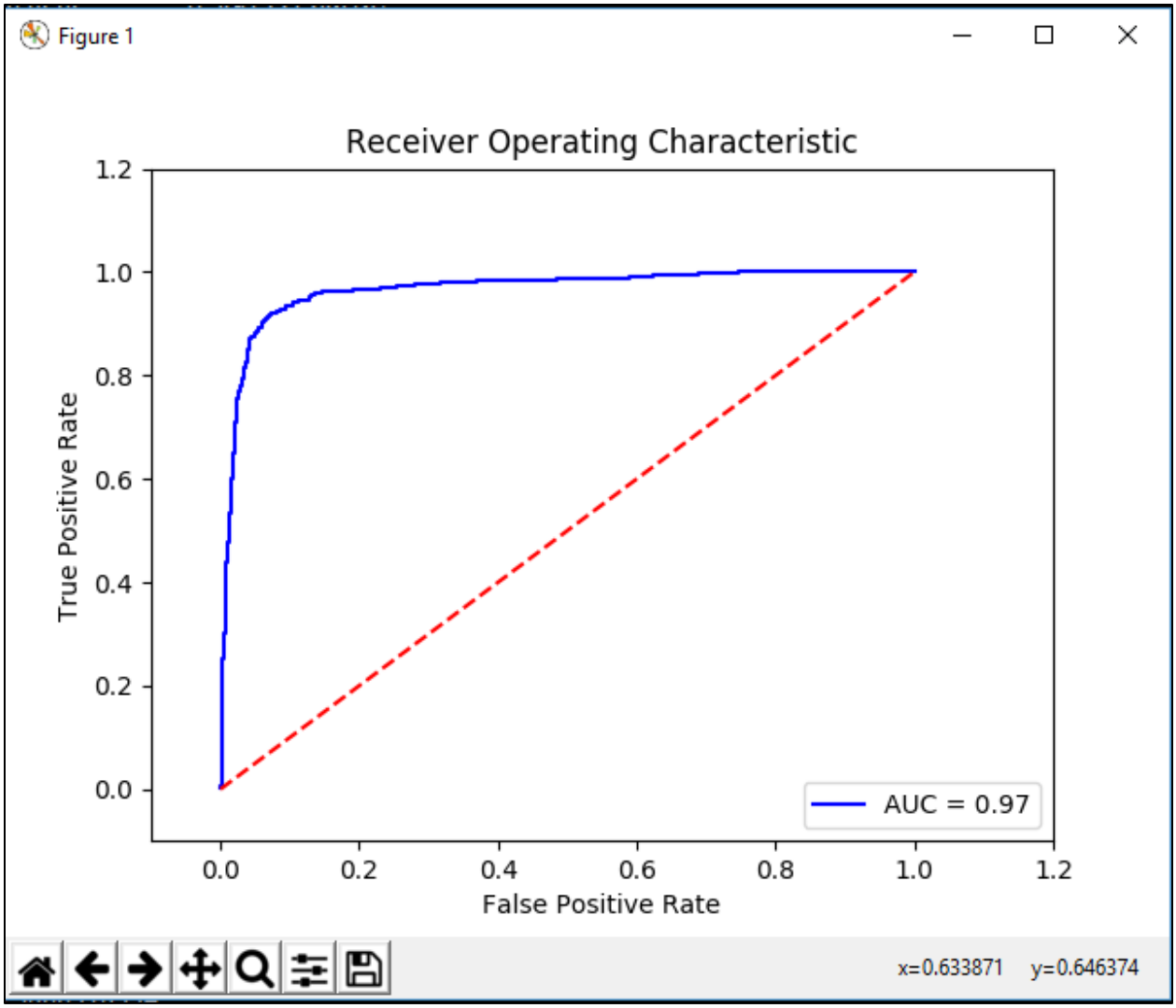
Description-In this experiment we have trained the svm with the linear kernel function and tested it, also calculated accuracy, precision and recall. Also we have plotted the ROC curve for it.

I have used scikit-learn / sklearn for python.

Accuracy=.92, precision= .91 and recall=.87

```
Experment 1 results:
Accuracy :        0.923076923077
Precision:        0.91976744186
Recall:           0.879866518354
```

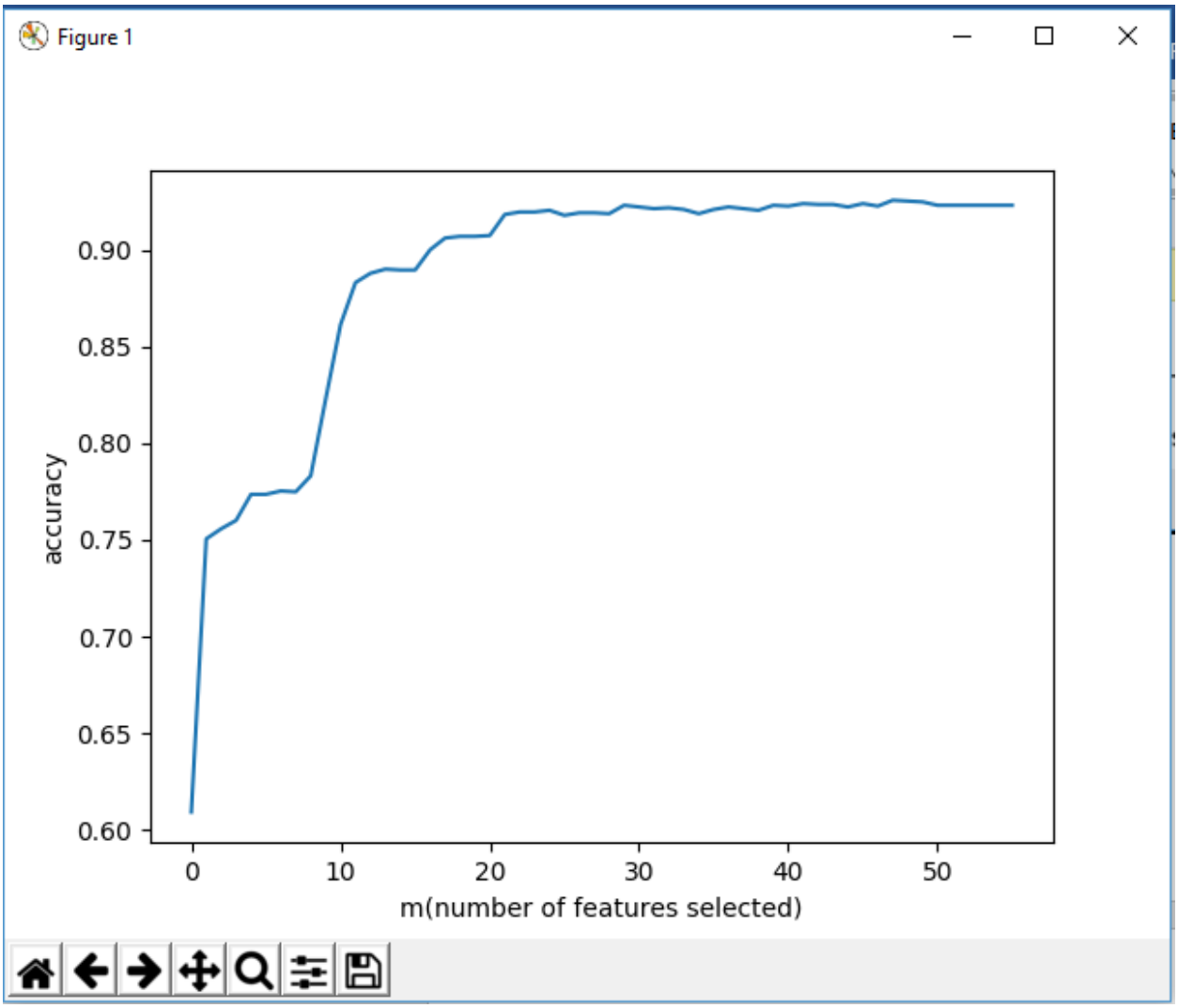
ROC Curve-



Experiment 2-

Description: In this experiment we have recreated our training and testing samples by selecting the most relevant features which are selected on the basis of the maximum weights. For features=2-57 we have calculated the accuracies and plotted it.

Plot-



Discussion of what the top 5 features were (see https://archive.ics.uci.edu/ml/machine-learning-databases/spambase/spambase.names for details of features)

These are the features index of the top 5 features-[26, 45, 24, 34, 40] These features are-
word_freq_hpl, word_freq_re, word_freq_money, word_freq_415, word_freq_direct

These are all the words i.e. how many times these words are present in the mail.for eq-most of the spam mail talks about giving money (lucky draw or something) so obviously the frequency of this word can be used to detect the mail is spam or not. Similarily others words are also words frequently used in spam.

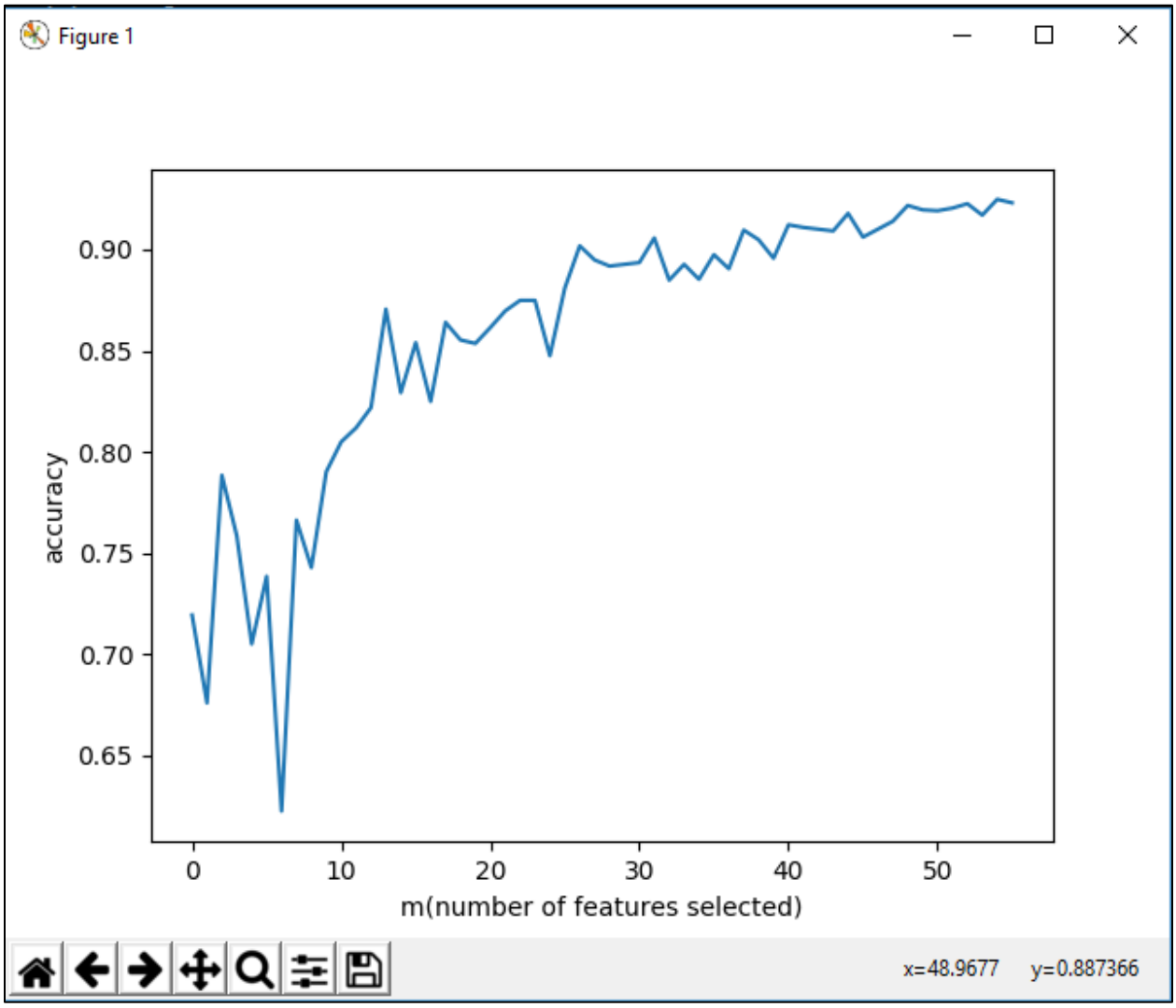
Discussion of the effects of feature selection (about 1 paragraph).

We do feature selection in large data sets in order to get rid of redundant or irrelevant features(it removes noise in the data set). The presence of irrelevant or redundant features may deviate from the optimal solution, so when we do feature selection we take into account the most relevant features for the classification to be done. In our experiment 2 we have done the selection on the basis of the weights vector, we have selected m (2-57) features with the maximum weights(more relevant) and then calculated the accuracy with it on the test data once the svm is trained on the training data

Experiment 3-

Description- In this experiment we have recreated our training and testing samples by selecting the features randomly. For features=2-57 we have calculated the accuracies and plotted it.

Plot-



Discussion of results of random feature selection vs. SVM-weighted feature selection (about 1 paragraph).

As seen from the graphs of the random and weighted feature selection, the graph is much more smoother(accuracy increasing consistently and not decreasing at any point) in case of the weighted selection whereas for random the graph is more spikier showing the in consistency in the way the accuracy is proceeding, This is happening because in case of the weighted selection we are selecting the feature on the weights(with maximum weights are selected) i.e. more relevant features are being selected but for random any feature is selected randomly even if it contributes less for the classification. Although the final accuracy achieved for both the cases were same but the initial accuracy was more in case of the feature selection with the svm-weighted selection.