CS314, Assignment 6 - Report

Uttam Kumar Reddy(200010052),Rajashekhar Reddy(200030058)

March 14, 2022

1 Problem statement

Spam email classification using Support Vector Machine: In this assignment you will use a SVM to classify emails into spam or non-spam categories. And report the classification accuracy for various SVM parameters and kernel functions.

2 Libraries Used and their purpose

- **PANDAS** : For reading the data file.
- **SVM** : For classification of data, kernel functions and testing the accuracy of the built model.

3 Methodology

3.1 Functions used from SVM

- **svm.SVC** for the kernel functions
- **model.fit()** for fitting the data into the model
- **model.predict()** to predict the output of the test data based on the training data.
- **model.score()**It comapares the output from the given model and compares it with the actual result and gives the accuracy.

3.2 Details of the SVM package used

```
from sklearn.model_selection import train_test_split
from sklearn import svm
```

4 Experimental Results

Note that the instances highlighted in green are the best instances for the given kernel function and the data set.

- RBF

RBF		
C values	Test Accuracy	Training Accuracy
100000	0.9391745112	0.9444099379
110000	0.9377262853	0.9453416149
120000	0.9377262853	0.9459627329
90000	0.9384503983	0.9440993789
85000	0.9370021723	0.9444099379
95000	0.9391745112	0.9444099379

The C value giving the best accuracy is 100000

- QUADRATIC

Quadratic		
C values	Test Accuracy	Training Accuracy
100000	0.8877624909	0.8919254658
110000	0.8892107169	0.8947204969
12000	0.8399710355	0.8447204969
90000	0.8855901521	0.8916149068
80000	0.8855901521	0.8903726708
95000	0.887038378	0.8934782609

The **C** value giving the best accuracy is 110000

- LINEAR

Linear		
C values	Test Accuracy	Training Accuracy
0.35	0.9290369298	0.9347826087
0.38	0.9290369298	0.9347826087
1	0.9283128168	0.9347826087
0.01	0.9123823316	0.9127329193
0.15	0.9268645909	0.9313664596
0.1	0.924692252	0.9304347826

The **C** value giving the best accuracy is 0.38

5 Conclusion

We can notice that the RBF kernel functions gives us the best accuracy compared to the other 2 kernel functions.
Thus, we can conclude that we can classify the spam mails with an accuracy of 93.917 percent using the RBF kernel Function.