Spam Email Classification

Assignment 6

Team -14
Tabish Khalid Halim ,
200020049
Anand Hegde ,
200020007

Department of Computer Science, IIT Dharwad

March 16, 2022

Contents

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

- **Pandas :** In order to read the data.
- **SVM :** To classify the data, kernel function and train the model and test it's accuracy.
- **Statistics :** We have used this to calculate mode of the data wherever required.
- **Numpy :** To classify the data-set into an array in python.

3 Methodology

We first tried to run SVM without any pre-processing, then we noticed that the algorithm was behaving a bit peculiarly as we tried changing the regularization hyper-parameter. Then we noticed that the last 4 features in the feature vectors were too big compared to others. Also the linear kernel was taking a lot of time to train.
So, We performed ***Mean-Normalization*** on the data and scaled all the features with $mean = 0$ and $variance = 1$.
This helped us to speed up the training process, and the Constant C which was about 10000 before for the Gaussian kernel came down to about 5. And the training of linear kernel which used to take around 5 mins to complete, is now completed in 0.001 seconds.

3.1 SVM Functions used:

- model.fit()
- model.predict()
- model.score()
- svm.SVC()

3.2 SVM Packages used :

- **from sklearn import svm** :
- **from sklearn.model_selection import train_test_split** :
- **from sklearn.preprocessing import StandardScaler** :

4 Experimental Results

RBF

C	Training Accurac	Testing Accuracy
1	0.95	0.922519913
2	0.955900621	0.925416365
3	0.95931677	0.928312817
4	0.961490683	0.931209269
5	0.963664596	0.933381608
6	0.965217391	0.93410572
7	0.967080745	0.930485156
8	0.968944099	0.928312817
9	0.969875776	0.928312817
10	0.971428571	0.92903693

Quadratic Kernel

C	Training Accuracy	Testing Accuracy
5	0.937888199	0.900796524
6	0.940993789	0.907313541
7	0.943478261	0.910209993
8	0.947204969	0.913830558
9	0.950621118	0.916727009
10	0.953726708	0.919623461
11	0.954347826	0.919623461
12	0.954968944	0.918899348
13	0.956521739	0.921071687
14	0.956521739	0.9217958

Linear Kernel

C	Training Accuracy	Testing Accuracy
10	0.941304348	0.918175235
13	0.941614907	0.918175235
16	0.942546584	0.920347574
19	0.941304348	0.920347574
22	0.941614907	0.920347574
25	0.94068323	0.919623461
28	0.94068323	0.918899348

5 Conclusion

After carefully examining the data, we have constructed the accuracy table for each of the methods {RBF Kernel,Quadratic Kernel & Linear Kernel} and found out that RBF kernel has the highest accuracy as compared to the other kernel methods.
Thus we can conclude that using the RBF function method, we can get an accuracy of 93.4% for the problem of spam email classification.

6 References

- SVM
- `https://en.wikipedia.org/wiki/Support-vector_machine`