# 16EE10056
# Soumava Paul
# Assignment 5

Spam Email Detection using SVM Classifier

Introduction

In this assignment, I have used 3 different SVM kernels, namely linear, quadratic and RBF for efficient spam email detection in the given database. The only other variable parameter was C - the generalization constant which was varied from 10^{-6} to 10^{6}. Relevant plots and tables are included in the later parts of the document.

Results from 3 SVM classifiers

Linear Classifier performed best on the test set at a setting of C = 10^{3} with a test set accuracy of 92.686% and corresponding training accuracy of 93.571%.

Quadratic Classifier performed best on the test set at a setting of C = 10^{5} with a test set accuracy of 94.424% and corresponding training accuracy of 96.491%.

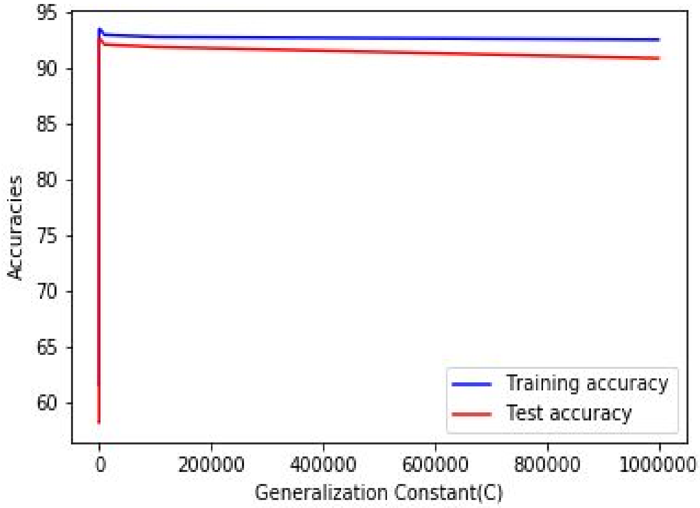
Rbf Classifier performed best on the test set at a setting of C = 10^{5} with a test set accuracy of 94.496% and corresponding training accuracy of 97.205%.

Overall, **Quadratic** and **RBF** classifiers performed almost equally well on the test set at the same setting of C - the generalization constant, RBF classifier being only marginally better. However, the quadratic classifier showed less tendency to overfit.

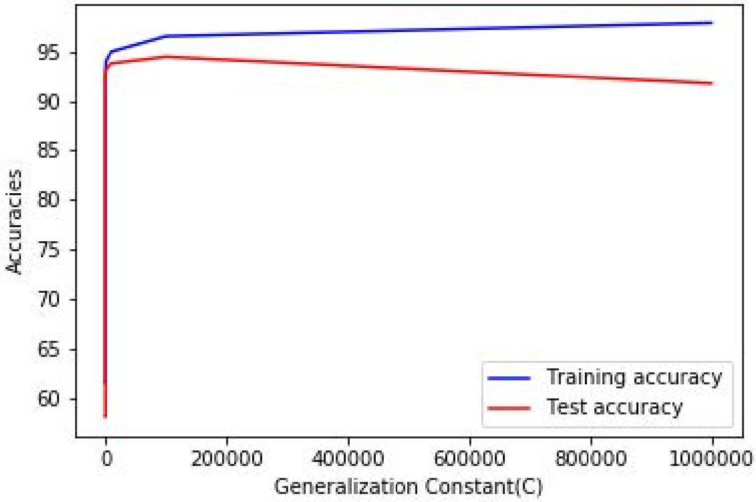
PLOTS & TABLES

Plots and Tables

Training and Test Accuracies



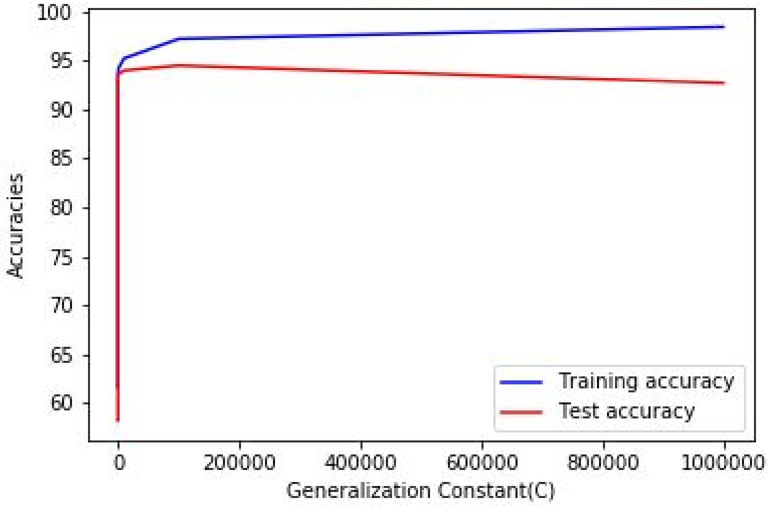
Linear Kernel



Quadratic Kernel

Plots and Tables

Training and Test Accuracies



RBF Kernel

ACCURACY TABLE FOR DIFFERENT VALUES OF C

Generalization Constant C

Acc	**Kernels**	1e-6	1e-5	1e-4	1e-3	1e-2	1e-1	1.0	10	100	10^3	10^4	10^5	10^6
Test	linear	58.22	58.22	58.22	58.22	79.87	88.63	91.74	92.11	92.54	92.69	92.18	91.96	90.95
Train		61.61	61.61	61.61	61.61	80.62	88.26	91.92	93.14	93.48	93.57	93.04	92.86	92.61
Test	quadratic	58.22	58.22	58.22	58.22	58.22	58.22	85.52	90.3	92.54	93.19	93.77	94.42	91.74
Train		61.61	61.61	61.61	61.61	61.61	62.39	85.71	90.99	92.95	94.04	94.94	96.49	97.83
Test	RBF	58.22	58.22	58.22	58.22	58.22	58.29	86.02	91.24	93.41	93.63	93.99	94.5	92.69

Training and Test Accuracies

Train		61.61	61.61	61.61	61.61	61.61	61.96	85.93	91.61	93.38	94.22	95.22	97.2	98.42

OBSERVATIONS

The generalization constant C is a penalty for the error in misclassification. A large value of C implies a low margin hyperplane and vice-versa. For higher values of C, the SVM rarely misclassifies a data point in the training set since it heavily penalizes any case of misclassification. Hence the classifier will exhibit low bias and high variance with a tendency to overfit. For very low values of C, the classifier fails to learn the complexity of the training data and hence fails to accurately classify both the train and test data. The goal is to find the ideal tradeoff between misclassification and the hyperplane margin. In this experiment, 10^5 was found to be an optimal value of C (for quadratic and RBF kernels) beyond which there was overfitting tendency.