Assignment -5

#16MI31014 #SAYAN GUHA #Assignment -5

1.Package Used-The package used for this assignment was sklearn.svm.SVC.

Support Vector Classification.

The implementation is based on libsvm. The fit time complexity is more than quadratic with the number of samples which makes it hard to scale to dataset with more than a couple of 10000 samples.

The multiclass support is handled according to a one-vs-one scheme

Methodology adapted:

1.We were given the dataset https://archive.ics.uci.edu/ml/datasets/Spambase.This dataset contained 4601 instances and 57 attributes.

2.The dataset was loaded into a pandas DataFrame and the feature vector X consisted of first 57 columns while target label y was 0/1 distinguishing for not spam/spam.

3.The dataset is then split into training(70%) and test set(30%) using sklearn.model_selection import train_test_split

4.Preprocessing using StandardScaler was done on the Dataset.

5.Then sklearn.svm.svc package was trained on our training dataset using kernel functions such as rbf,quadratic and linear an varying values of C.

6.The Experimental results are shown in the next page.Accuracies for different kernel for different C values are presented in the following table.rbf kernel with C=3 has the highest accuracy in Test set while rbf with C=100 has highest accuracy in Training Set.

Kernel/ C	1	2	3	4	5	0.1	10	100
Linear	0.9268	0.9268	0.9261	0.9261	0.9275	0.9210	0.9290	0.9254
RBF	0.9370	0.9413	0.9427	0.9406	0.9385	0.9051	0.9384	0.9217
Quadratic	0.8298	0.8754	0.8942	0.90079	0.9051	0.7139	0.9217	0.9275

ACCURACY FOR TEST SET DATA

Best value of accuracy for linear,RBF,Quadratic kernel are 0.9275(5),0.9427(3),0.9275(100)

Kernel/ C	1	2	3	4	5	0.1	10	100
Linear	0.9332	0.9329	0.9350	0.9350	0.9350	0.9288	O.9350	0.9357
RBF	0.9419	0.9490	0.9524	0.9543	0.9565	0.9099	0.9645	0.9860
Quadratic	0.8543	0.8869	0.9136	0.92204	0.9322	0.7298	0.9478	0.9711

ACCURACY FOR TRAINING SET DATA

Best value of accuracy for linear,RBF,Quadratic kernel are 0.9357(100),0.9860(100), 0.9711 (100)