Assignment 5 : Spam Email Classification using Support Vector Machines

Methodology

Libraries : Numpy, scikit-learn, Matplotlib

The email classification data was read as a numpy array. 70% of the dataset was randomly assigned to the train dataset and the rest was kept as the test dataset. In order to scale each attribute of the dataset between the values [-1,1], the **MinMaxScaler** function provided by scikit-learn was used. The test set was scaled using the same parameters used for scaling the train set.

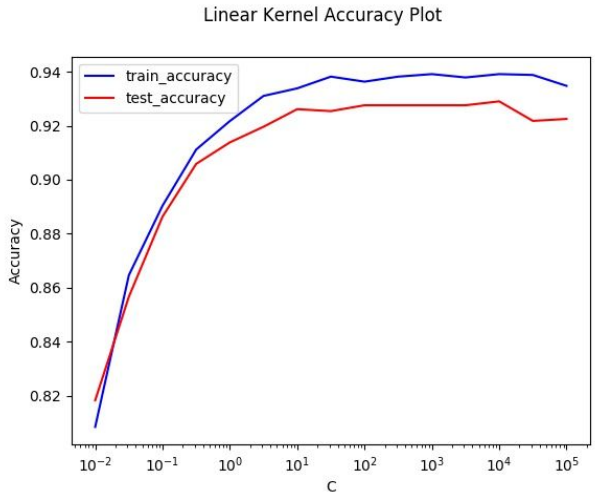
After the dataset was scaled, the SVM classifier provided by scikit-learn, **SVC**, was used. The model was trained for different values of C and the kernels were set to ***linear, poly*** *(with degree=2)* and ***rbf****.* After training the classifier, the train and test accuracy were obtained. Plots have also been generated to show the behavior of the Support Vector Machine classifier at different values of C for different kernel functions.

Experimental Results

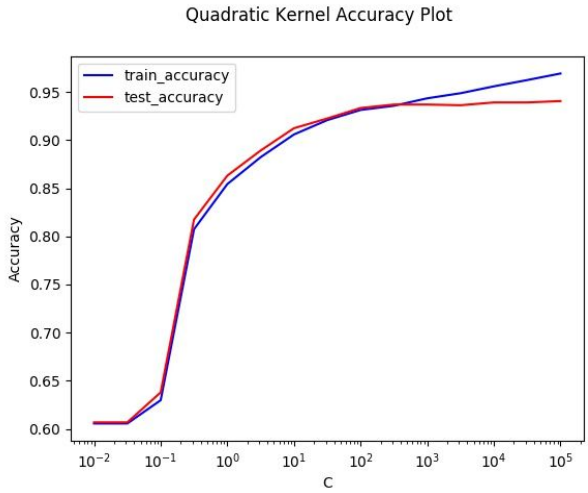
Accuracy Table

Kernel	**C**	10^{-2}	$10^{-1.5}$	10^{-1}	$10^{-0.5}$	10^{0}	$10^{0.5}$	10^{1}	$10^{1.5}$	10^{2}	$10^{2.5}$	10^{3}	$10^{3.5}$	10^{4}	$10^{4.5}$	10^{5}
Linear	Train accuracy	80.83	86.45	89.04	91.12	92.17	93.11	93.38	93.82	93.63	93.82	93.91	93.79	93.91	93.88	93.47
	Test accuracy	81.82	85.66	88.63	90.59	91.38	91.96	92.61	92.54	92.76	92.76	92.76	92.76	92.90	92.17	92.25
Quadratic	Train accuracy	60.56	60.56	62.98	80.74	85.43	88.23	90.59	92.08	93.14	93.57	94.35	94.88	95.59	96.24	96.92
	Test accuracy	60.68	60.68	63.79	81.75	86.31	88.92	91.24	92.25	93.34	93.70	93.70	93.63	93.92	93.92	94.06
RBF	Train accuracy	61.21	61.21	63.14	82.86	87.05	88.76	91.58	93.26	94.16	94.22	94.94	95.25	95.96	96.64	97.17
	Test accuracy	59.16	59.16	60.39	79.72	84.72	86.53	89.21	91.60	92.03	92.25	92.47	92.32	92.54	92.03	92.03

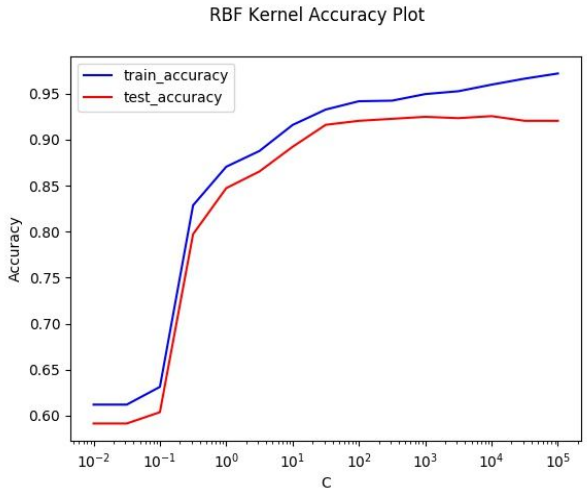
Linear Kernel Function : The maximum test accuracy = 92.90% was obtained for C = 1000.



Quadratic Kernel Function : The maximum test accuracy = 94.06% was obtained for C = 10^5 .



RBF Kernel Function : The maximum test accuracy = 92.47% was obtained for C = 1000.



Discussion

In a SVM classifier, we look for a couple of things : a hyperplane with the largest minimum margin and a hyperplane the correctly separates as many instances as possible. The parameter C is responsible for how much we desire to achieve the latter. Therefore, having a low C gives a pretty large minimum margin. Thus, the models pays less attention to how correctly the hyperplane separates the instances and thus, the test accuracy obtained is less for a smaller value of C. On the contrary, a large value of C implies small minimum margin. Therefore, the models tends to overfit the data and the accuracy tends to decrease as C becomes extremely large. Hence, we aim to obtain an optimal value of C to achieve a tradeoff between attaining the largest minimum margin and the tendency of the hyperplanes to correctly classify separate instances In our case, the maximum test accuracy obtained is 94.06% for a quadratic kernel function with $C=10^5$.