Binary Classification Based on Support Vector Machine

Divyank Garg (526005747)

Introduction: The dataset consist of two features and two class 0 and 1. Based on these two features, the model need to be fit using SVM classifier to predict the class.

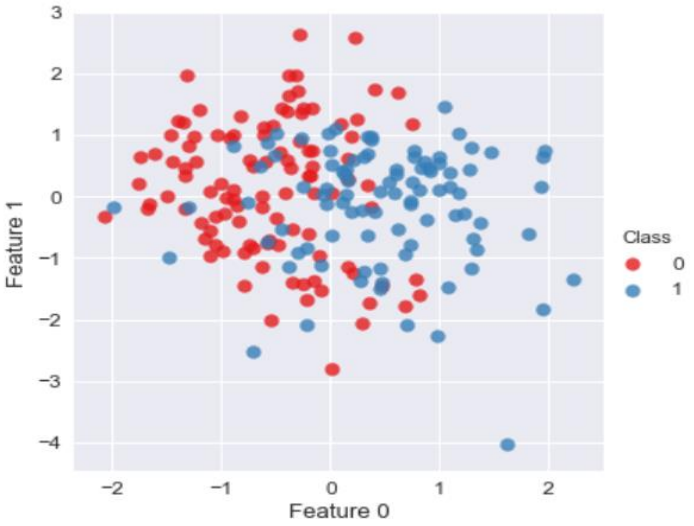

Fig-1: Scatterplot of different class

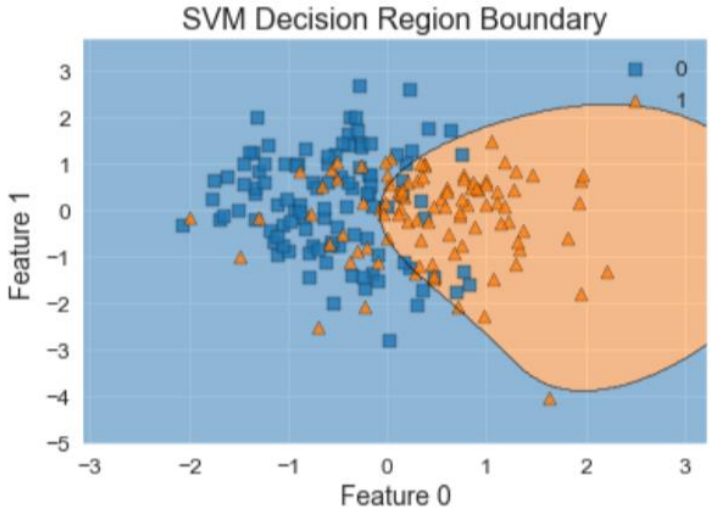

Fig-2: RBF kernel boundary in test data

Method and Explanation: Initially the data was divided into two dataset. One consist of features and other as class. Next to find out the distribution of class based on features

Using scatterplot between these features in Fig-1 it can be found that the data points for both class are jumbled and it shows that nonlinear boundary can only separate these classes and can predict accurately. But then also both linear and nonlinear were performed to know the best fit kernel for differentiating the class.

the linear kernel was taken to find out the hyperplane separating the classes along with margin. The cost function (C value) plays an important role in specifying the cost of a violation to the margin. So to find this optimal C value the cross validation was performed and found C value to be 0.120 and the CV score to be 0.768.

The sigmoid kernel was taken into account and then in this type of kernel both C and gamma value plays important role in specifying the cost of violation of margin. Using cross validation, C and gamma value found to be 0.1321 and 0.2021 respectively. The CV score was found to be 0.781.

The rbf kernel was taken into account then C value and gamma value was found to be 0.3053 and 0.2021 respectively. The CV score for this kernel was highest and found to be 0.8182.

Result: So among all the above kernel, it can be found that rbf kernel have highest CV score. Using the optimal value of C and gamma the class was predicted for test data. The boundary was made using rbf kernel to differentiate the classes as shown in Fig-1.

To check the accuracy of my model the training dataset was divided into 80% of train data and 20% test data and then found accuracy of 81% for rbf kernel. This accuracy was highest among the other three kernel, so using rbf kernel and C and gamma value the class was predicted for test data.