Report for Challenge 5 – Support Vector Machine

Vedant Mehta

1. Introduction

Challenge 5 was based on binary classification using Support Vector Machines for two features. The class labels were 0 and 1. There are four well known kernel options available in sklearn package – 'rbf', 'poly', 'linear', 'sigmoid'. Decision boundary was built using *'rbf'* – Radial Basis Function and *'poly'*- Polynomial decision boundary with varying degrees of non-linearity.

Below given figure shows the scatterplot of the training dataset on which we will try to fit the decision boundary using SVM.

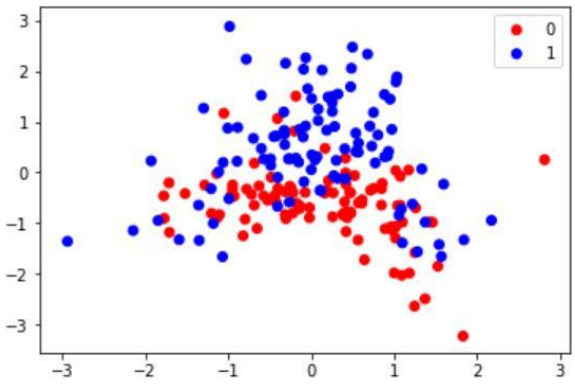

Figure 1 Scatterplot of training dataset

Looking at the scatterplot, it is pretty clear that we would get a better classification by either using rbf or polynomial kernel. Linear or sigmoid kernel won't give a good classification accuracy.

So, after splitting the dataset into training and validation set, SVC model using rbf and polynomial is fit on the training data. RBF kernel gives better result and so final prediction done using RBF kernel

2. Model Training and Accuracy

RBF kernel is given by the formula: $K(y_1, y_2) = \exp\left(-\frac{\|y_1 - y_2\|^2}{2\sigma^2}\right)$

Model built with rbf kernel with default parameter gives an accuracy of **81.5%**. While trying to tune the model, it did not increase the accuracy much, hence to avoid complexity, simple model is used for classification. While Polynomial kernel gave an accuracy of **64%** (with degree of 5) Below given image shows the decision boundary for the model built on the whole training dataset.