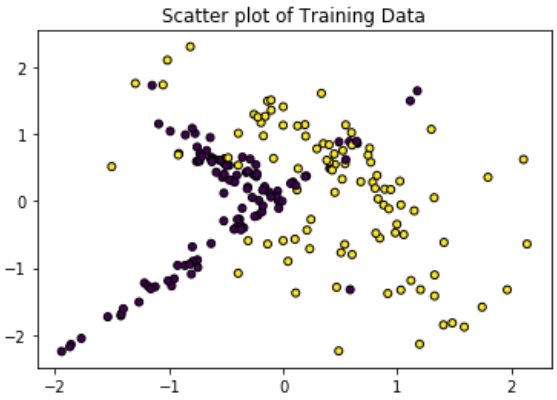
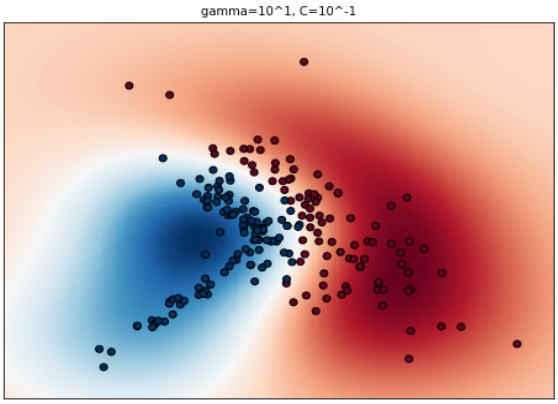
CHALLENGE 5 REPORT - BINARY CLASSIFICATION USING SVM

Introduction:

The training and testing dataset consists of 200 samples each, with 2 features and it was divided into 2 classes. The objective is to find the decision boundary for the classification of the data using Support Vector Machines(SVM).

Methodology:





As we can see from the above figure, we can be very clear that a linear hyperplane will not be able to classify them into two separate groups. Therefore we need to look for such a hyperplane in a higher dimension of space by projecting the data points in that space. In sklearn we have different options of kernel available such as RBF, Polynomial, Sigmoid and Linear. In order to evaluate the model performance with the training dataset, we go for k-fold cross validation technique. I found out that SVM model with RBF kernel (default parameters) performed the best with an accuracy score of 88%. In order to make the model performance even better, we need to fine-tune the hyperparameters of the RBF kernel.

RBF stands for Radial Basis Function. When training an SVM with the *Radial Basis Function* (RBF) kernel, two parameters must be considered: C and gamma. The parameter C, common to all SVM kernels, trades off misclassification of training examples against simplicity of the decision surface. A low C makes the decision surface smooth, while a high C aims at classifying all training examples correctly. gamma defines how much influence a single training example has. The larger gamma is, the closer other examples must be to be affected.

Optimal C and Gamma can be found using GridSearch technique. This technique is nothing but choosing each and every (C,Gamma) pairs from a range of parameter values to build the model and choose the best one among it. In this report, I have chosen a range of 10^-3 to 10^3 for both C and Gamma and found out that C=0.1 and Gamma = 10 gave the best result with an accuracy score of 92%. The Decision plot of this model is shown above in the picture. Using this model, the testing set prediction is also done and is updated in the testing csv file.