Challenge 5
Binary Classification with SVM

Priya Balasubramanian
Texas A&M University

Abstract: The task was to develop a binary classifier for the given data using Support Vector Machines. The dataset is a randomly generated set of data with two classes. The testing data is also randomly generated. The testing data classes are not specified and the model is built based on the training data provided. The Radial Basis Function kernel of SVM was used to build the classifier after experimenting with linear and polynomial SVM models of different degrees. The following is a detailed report on the process and conclusions and the model built.

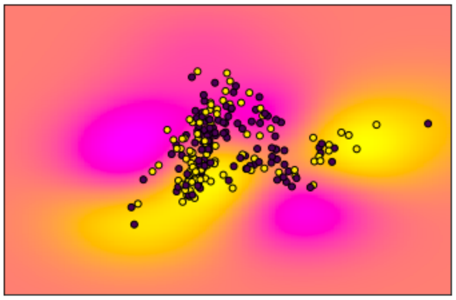
I. INTRODUCTION

A Support Vector Machine (SVM) is a hyperplane that helps in classifying the data. The hyperplane is built on the training data that is provided based on supervised learning. The hyperplane is simply a line that divides the data points into different classes if the given space is two-dimensional.

The current task of building a binary classifier for randomly generated training data and predicting the classes of the randomly generated test data points using the model built is done with the SVM method in Pythons's sklearn package.

II. MODEL FORMULATION & ANALYSIS

There are multiple options for building a classifier with SVM. These options are presented as kernels, regularization, gamma and margin. Radial Basis Function (RBF) is the chosen kernel for this particular task as the data overlaps strongly and other options like linear and polynomial fails in classifying the data appropriately. The regularization parameter used is C=2 which allows the model to only misclassify 2 data points while building the model. Margin is the separation of the hyperplane to the closest points and the gamma value which allows for consideration of the close points and the farther points while finding the separation hyperplane are all chosen to be default values as provided by the sklearn package.



These choices ended up creating a kernel with classification accuracy of close to 94% in the training data. The image on the left shows the decision boundary created on the training data by the SVM hyperplane that was modeled.

III. CONCLUSION

The overlap on the data points made the RBF kernel best suited for classification purposes. This shows clearly that some level of separation between the data points is appreciated when building a linear or polynomial kernel while an RBF kernel is better suited for building kernels for overlapping data points.