Challenge 5 – Support Vector Machines

Siddharth Ajit
Texas A&M University

I. Introduction

A. Dataset description

The training and test dataset has 200 data instances with two features. The objective is to classify a datapoint to class 0 and 1. From the training dataset, a linear classifier doesn't seem suitable. Other nonlinear classifiers may provide better classification accuracy

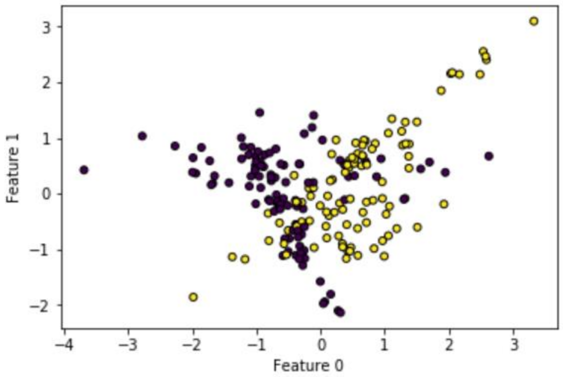

Figure 1. Feature space

B. Methodology

To identify the best kernel and the corresponding hyperparameters, cross validation misclassification rate was used. The kernels used are linear, polynomial (degree = 3) and Radial basis function. Further, hyperparameters C (Regularization parameter) and Gamma are optimized using Gridsearch CV to obtain the optimal hyper-parameters of the corresponding kernels.

$$K(\mathbf{x},\mathbf{x}') = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2}\right)$$

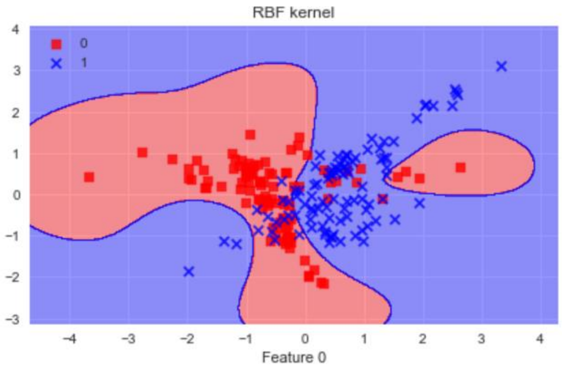

Figure 2. Decision boundary

RBF kernel function on two sample points x, x'. The term ($1/2\sigma^2$) is equivalent to γ parameter [4]. Using an RBF kernel cross validation accuracy is 0.875 with hyperparameters C = 10 and γ = 0.75, polynomial kernel of degree 3 gives an accuracy of 0.77

As expected, a nonlinear boundary performs better than a linear hyperplane in terms of misclassification rate. Further, a model trained with these parameters of RBF was used to predict the test data set.