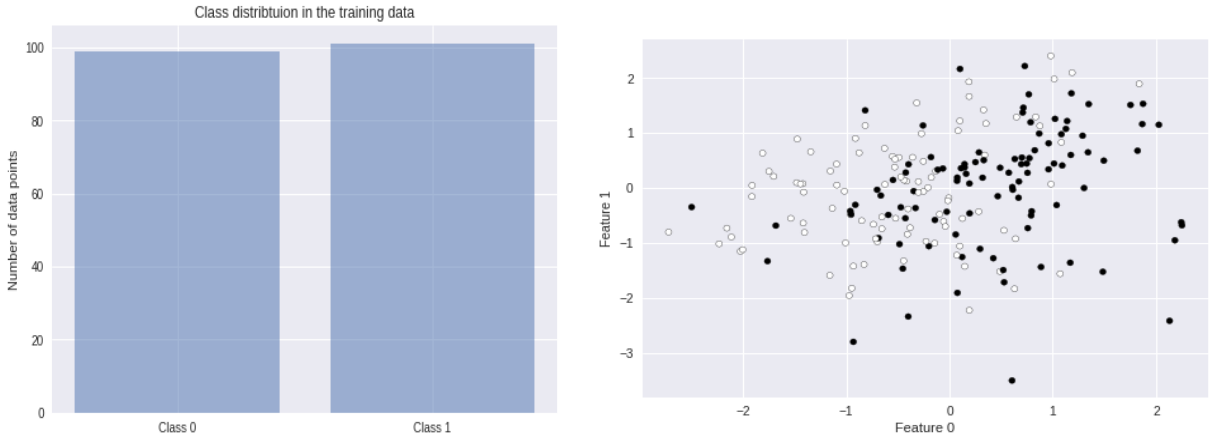
SVM Classification Challenge

-Aditya Lahiri

***Abstract*-The objective of this challenge was to classify a two-dimensional dataset using support vector machines. This challenge was tackled using linear as well as kernel support vector machines and we found that the linear and radial basis kernels have better classification accuracy on the training set over other kernel support vector classifiers. We then performed model selection and parameter tuning to find that the radial basis kernel had the highest classification accuracy on the training set and used it to predict the class labels on the testing set.**

I. Dataset

The training dataset provided for this challenge consisted of two features namely 'Feature 0' and 'Feature 1' and a binary dependent variable (0s and 1s) which specified the class labels. A testing dataset was also provided with both the features without the class labels. Both the training and testing dataset consisted of 200 data points each. The range of Feature 0 and Feature 1 was similar so we did not apply any standardization to the data. Both the class labels were almost evenly distributed in the training data (fig.1a) and the data didn't seem to be linearly separable from our visualization in fig.1.b.



1.a 1.b

Figure 1. Visualization of the training data. 1.a Distribution of class labels. 1.b Scatter plot of training data.

II. Model Selection

We train the four Support Vector Machine (SVM) classifiers using a linear, radial basis function (RBF), sigmoid and polynomial kernels with a soft margin parameter C=1 and keeping other default parameter specifications as in the Scikit Learn library in python [1]. We perform model selection using 10-Fold cross-validation and calculate the average validation accuracy for each of the classifier. The results are summarized in the table below.

SVM Kernels	Mean Cross Validation Accuracies
Linear	69%
RBF	71.5%
Sigmoid	66.4%
Polynomial	62.5%

From our initial analysis above we see that RBF and linear kernels have the highest and second highest classification accuracies respectively. In the next section, we will optimize and the tune the hyperparameter of these two classifiers and measure their accuracy to design the best classifier.

III. Parameter tuning

Finding the best classifier is essentially an optimization problem. Here we use validation classification accuracy as the objective function and we optimize this function over the soft margin parameter for the linear SVM classifier. and over both soft margin parameter and gamma parameter for the RBF SVM classifier. The gamma parameter can be thought of as the inverse of the radius of influence of training samples. Upon optimizing and performing 10-fold cross-validation on the training set using the Gridsearchcv package in the Scikit learn library [2], RBF SVM classifier was found to classify the datapoints more accurately when compared to linear SVM. The optimization and cross-validation were performed over several values of C and Gamma parameters. The optimum gamma parameter was found to be 0.1 and the soft margin parameter was found to be 5.705. These values maximized the validation accuracy which was found to be 75%. The prediction on the training set with the optimal RBF SVM classifier yielded an accuracy of 75.5%. The decision boundary plot for the training set is provided in fig.2 below.

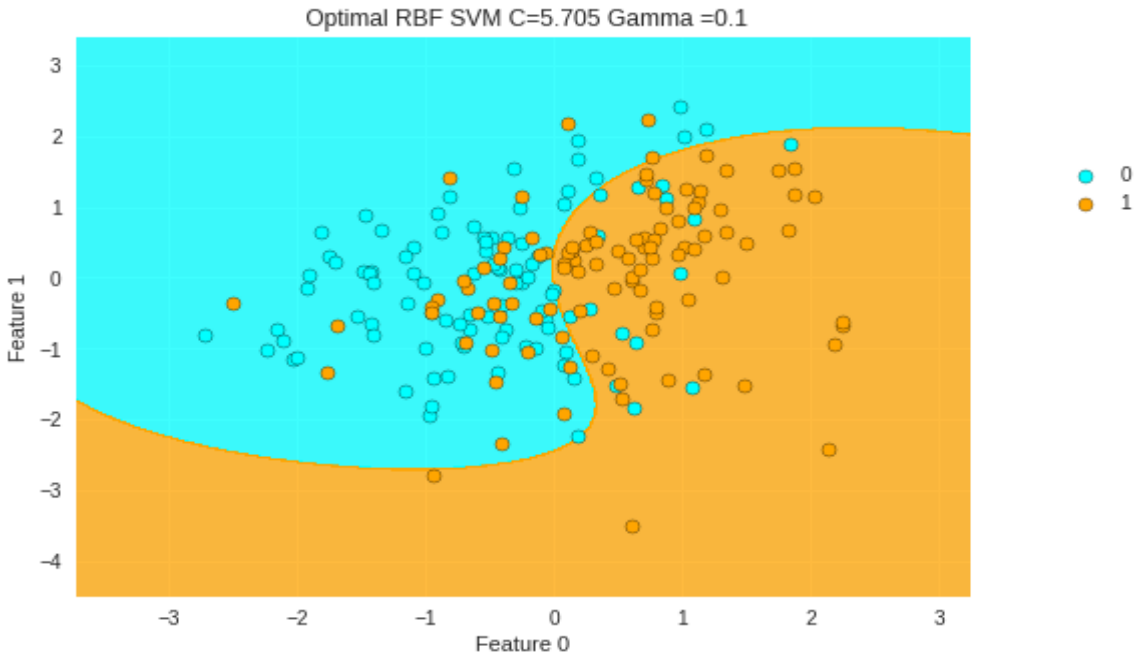


Figure 2. Visualization of the training data.

IV. References

[1] Scikit-learn.org. (2018). RBF SVM parameters — scikit-learn 0.20.0 documentation. [online] Available at: http://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html [Accessed 5 Nov. 2018].

[2]Scikit-learn.org.(2018). sklearn.model_selection.GridSearchCV—scikit-learn0.20.0 documentation.[online]Availableat:http://scikitlearn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html [Accessed 7 Nov. 2018].