ECEN 689 Challenge 5: Support Vector Machines

Sayeed Alvi

INTRODUCTION

Challenge 5 is a binary classification task based on the use of support vector machines and kernel methods. The objective is to justify the use of specific kernel for the given dataset and features. The feature matrix comprises of two features and the target variable has two classes 0 or 1.

METHODOLOGY

The first step is plot the data for visualizing its distribution as it is done in Fig 1 below. Clearly the given is not linearly separable and hence a nonlinear kernel is needed for classification between 0 or 1.

.

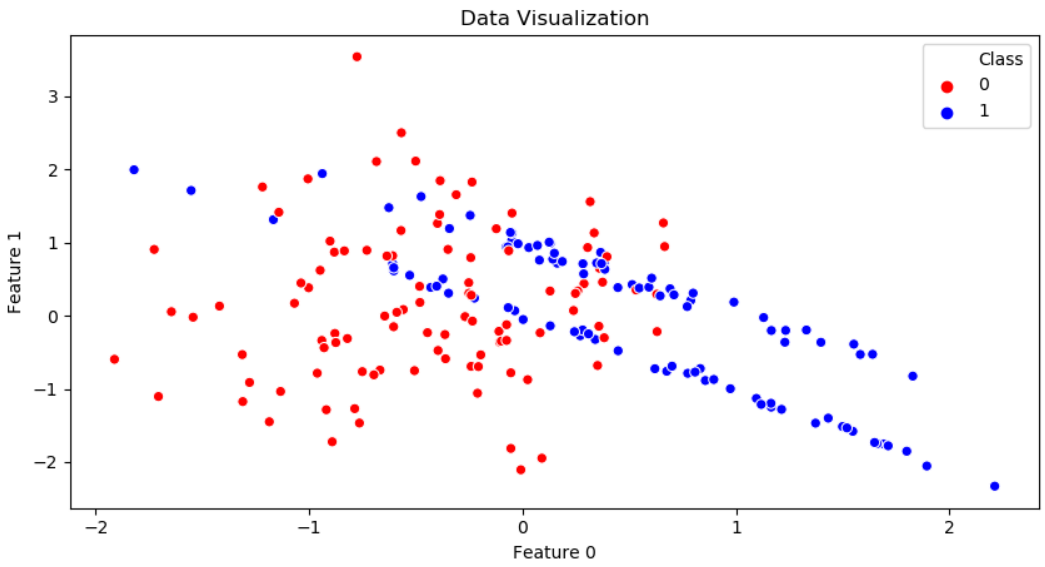

Fig 1

As there are only 200 data points (very few data) and the evaluation criteria for this challenge is categorization accuracy, cross validation would be the reasonable method to approach in the training as it avoids losing important patterns/trends in the data. Thus, for hyperparameter optimization, grid search cross validation over an exhaustive search space is applied based on the scoring criteria: Accuracy. The implementation of grid search over the set of parameters: C (np. logspace), gamma (np. linspace) and nonlinear kernels (RBF or Sigmoid) gives

RBF as the best choice of kernel with cross validation accuracy 0.86 and training accuracy over the entire dataset 0.91.

The best SVM estimator found through grid search cross validation was used on the testing data to predict the class column values. The decision region juxtaposed with the training set. is shown below in Fig 2.

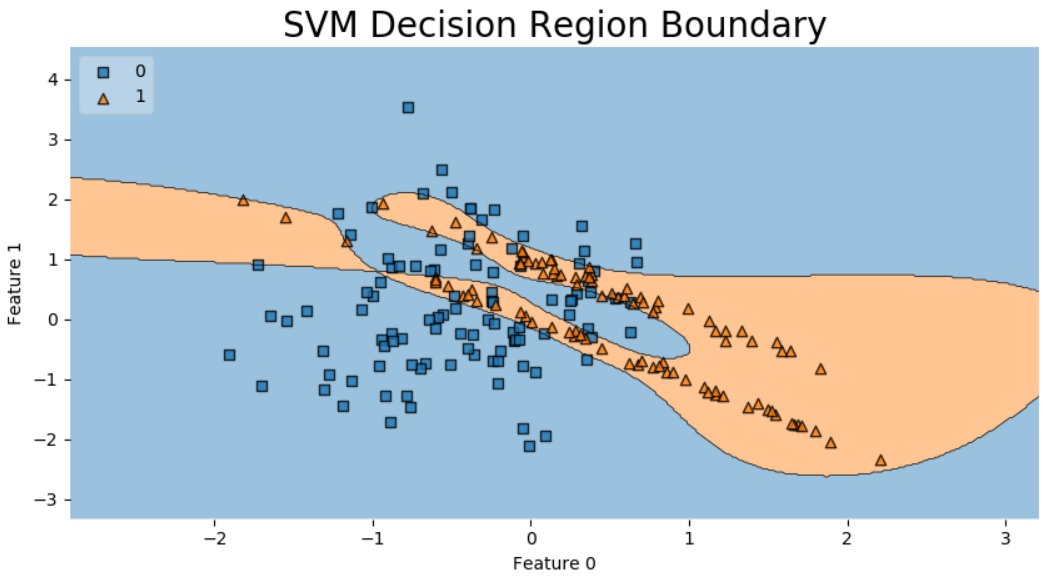

Fig 1: SVM decision boundary juxtaposed with the training data

REFERENCES

[1] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.
[2] http://rasbt.github.io/mlxtend/user_guide/plotting/plot_decision_regions/
[3] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
[4] https://en.wikipedia.org/wiki/Cross-validation_(statistics)

.