

ECEN 689: Challenge 5

Swati Ramachandran

Introduction

In Challenge 5, the main task is to build a binary classifier using SVM (Support Vector Machine). The dataset comprises of two features (Feature 0, Feature 1) and two class labels (0, 1). The aim of this report is to describe the choice of kernel used and depict the decision boundary.

Implementation

Before deciding which kernel is suitable for the problem, it is important to explore the data. Figure 1 represents the training dataset with respect to both features. It is clear that the data does not seem to be linearly separable.

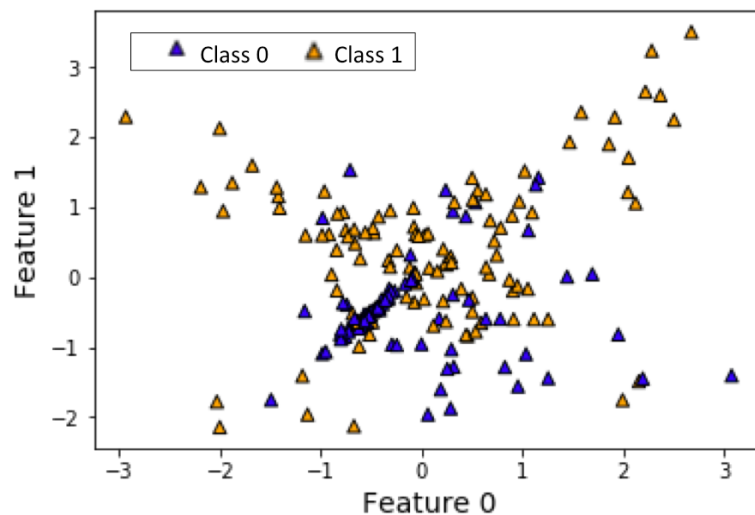


Figure 1. Representation of training dataset

There are 4 available kernels in Scikit-learn: linear, polynomial, sigmoid and rbf (radial basis function). The RBF kernel also uses a regularization parameter, C. A grid search was performed to identify the best kernel based on classification accuracy. Table 1 represents the results of different kernels based on classification accuracy using train/test split of 80-20. It is seen that RBF performed best, which is expected for datasets where the number of observations are greater than the number of features.

Kernel	Accuracy
Linear	0.725
Sigmoid	0.525
Polynomial	0.475
RBF	0.775

Table 1. Performance of different kernels

It is also seen that increasing the value of C increases the classification accuracy but only till a saturation point. $C=20$ gave the best classification accuracy and is used for predicting the test data. RBF ken

Figure 2 represents the decision boundary with the training data. This is plotted using mlxtend package [1].

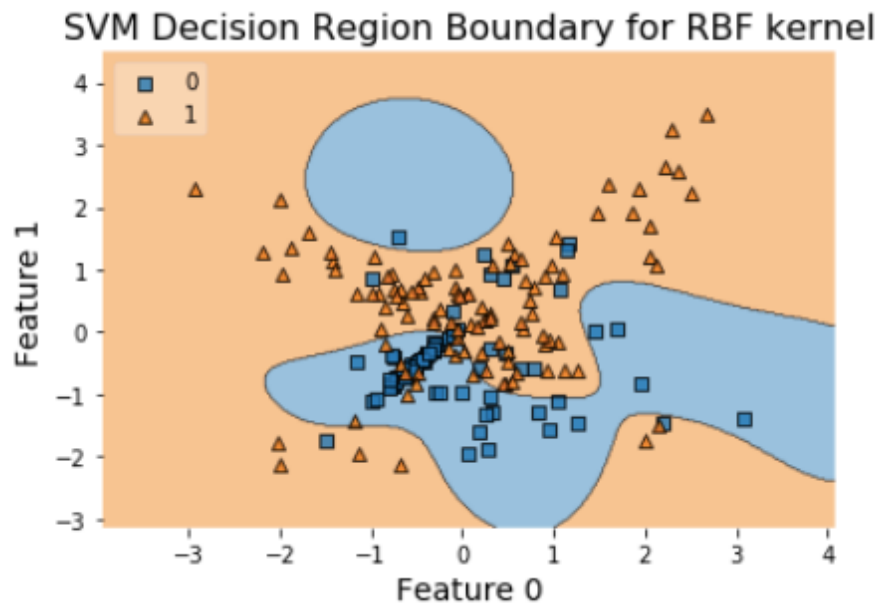


Figure 2. Decision boundary for SVM with rbf kernel

Conclusion

It is seen that grid search and parameter tuning are essential to obtain the optimal classifier. Additionally, for data set such as the one above, linear separation is not usually possible, in which case other kernels need to be applied. RBF was found to have the best classification accuracy on training set.

References

[1] Raschka, (2018). MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. Journal of Open Source Software, 3(24), 638, <https://doi.org/10.21105/joss.00638>