ECEN 689 APPLIED INFORMATION SCIENCE

CHALLENGE 5-REPORT-HARISH CHIGURUPATI

Abstract: **In this challenge, we perform the binary classification of the dataset provided into two classes (class 0 or class 1) using Support Vector Machine (SVM). We try to choose the best kernel and their corresponding parameters which classifies the dataset more accurately and plot the decision boundary.**

1. Method

Initially, we plot the training data set so that we know how the data is spread. We use the scatter plot here.

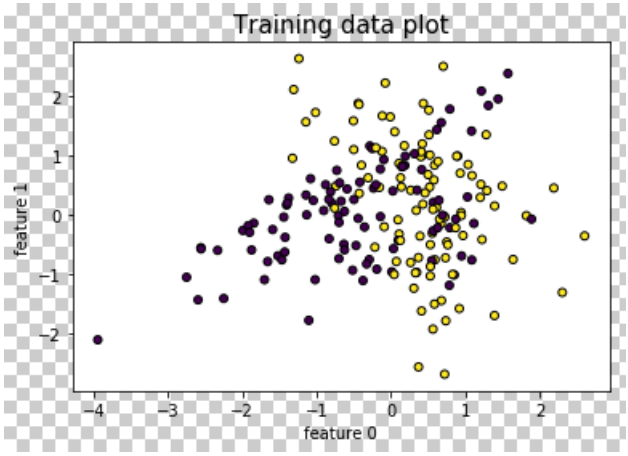

Figure 1 shows the scatter plot of the training data. On analysing the data visually, we see that the data cannot be separated linearly as the data points from the 2 classes overlap each other and we cannot define a decision boundary. In order to solve this classification problem, we implement Support Vector Machine (SVM).

We choose 4 kernels (Linear, Polynomial, Sigmoid, Radial Basis Function (RBF)) to see which performs better.

Figure 1. Scatter plot of the Training data.

2. Performance

When we employed all the 4 kernels and fitted the SVM model over the training data, we get the following results.

Kernel	Accuracy
Linear	0.73
Sigmoid	0.645
Polynomial	0.625
Radial Basis Function (RBF)	0.795

From the above table, we see that Radial Basis Function performs better when compared to other kernels.

3. Inference

The RBF kernel function on two sample points x, y is given by:

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$$

The performance of the RBF kernel can further be enhanced when we fine tune the parameters (C- Penalty parameter, gamma) relating to it. This is implemented using a Cross Validation

scheme Grid search, where we give different values for C and Gamma and give the optimum values for which the kernel performs the best. After performing the grid search, we got the value C=0.4, gamma=1.2 and we saw an improvement in the accuracy.

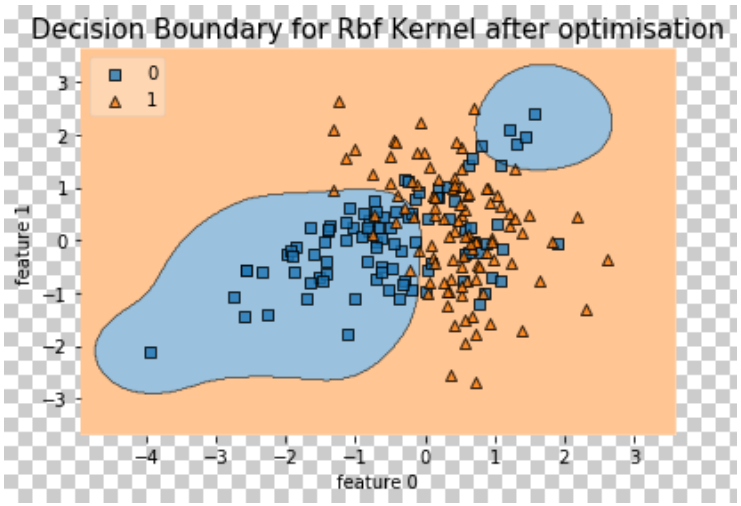

Figure 2. RBF kernel decision boundary

From the above results and graphs, we see that the data provided here is non-linear, so linear kernel cannot be used. The RBF kernel is a Squared Exponential kernel, which can be viewed as an infinite polynomial kernel which can be used to access all analytic functions. Also, the squared exponential kernels are more flexible than the other kernels where we can model more functions in its space. But the polynomial and sigmoid kernels are more fixed, hence after some point they get saturated and it does not help in classifying the data accurately. This fact is proved when we tested all the kernels simultaneously and RBF outperforms other kernels. Hence we choose RBF kernel with the same tuned parameters over the Testing dataset to predict the class.

4. References

[1]. https://en.wikipedia.org/wiki/Support_vector_machine.

[2]. Understanding Mlextend python package, https://www.pydoc.io/pypi/mlxtend-0.8.0/.

[3]. Marti A. Hearst, Support Vector machines, http://web.cs.iastate.edu/~honavar/hearst-svm.pdf