Binary SVM Classifier

Ranjith Tamil Selvan

MS, Computer Engineering
ECEN, Texas A&M University

ranjith@tamu.edu

https://github.com/CourseReps/ECEN689-Fall2018/blob/master/Students/ranjith/5challengeranjith.ipynb

I. Introduction

The task involved building a binary SVM classifier for a database generated synthetically per student, using sklearn's sklearn.datasets.make_classification. The main task involved exploring the usage of SVM classifier, finding the largest margin/decision boundary for the classification problem. The SVM classifier tries to draw a decision boundary to separate different labeled classes - the algorithm being different for linear and non-linear data. For a linear data, the SVM classification can be solved by finding a linear hyperplane that separates both the data. However, for non-linear data, we would need to explore different kernels and fine-tune the parameters influencing the classification.

II. PLOTTING THE DATA

For an idea of the generated data, the data was first plotted using *matplotlib* library. As shown in Fig. 1, the plot clearly shows that the data is non-linear.

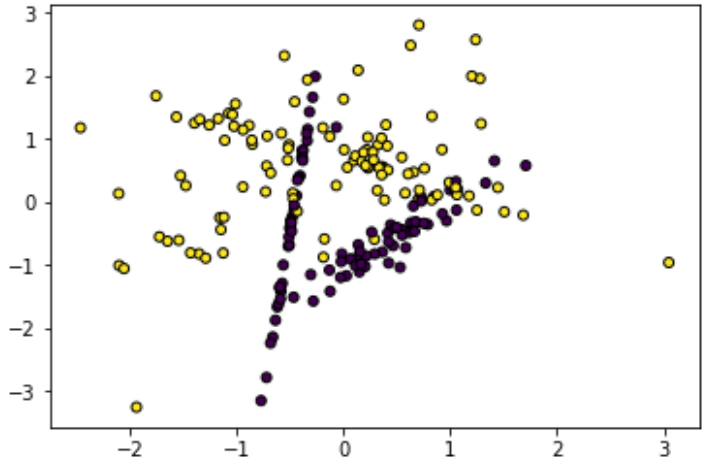

Fig. 1 Plot of generated data

III. EXPLORING CLASSES OF SVM KERNELS

The different kernels provided by *sklearn svm* library - linear, poly, sigmoid, radial-basis function were used to classify the data. For an initial analysis, I proceeded with the default-values for each kernel. The scores obtained for the different kernels are shown in TABLE 1.

TABLE I
SCORES FOR DIFFERENT SVM KERNELS

MODEL	SCORE
LINEAR	0.78
POLY	0.69
SIGMOID	0.72
RBF	0.845

A. TUNING THE PARAMETERS OF RBF KERNEL

As seen in the previous table, RBF (Radial Basis Function) kernel clearly gave the best score amongst the available kernels.

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

(where $||x-y||2$ is the squared Euclidean distance between x and y)

The challenge remaining was to identify the optimum parameters for the RBF kernel.

svm.SVC(kernel='rbf', gamma = < >, c= < >)

gamma - influences the influence of a single data point from training set. (low gamma gives low bias and high variance, high gamma gives high bias and low variance)
c - regularization parameter in SVM (low C gives high bias and lower variance, high C gives low bias and high variance).

The classifier for RBF SVC can be optimised by cross-validation, which was performed by the GridSearchCV object from sklearn. The api helps to tune the hyper-parameters of the RBF SVC classifier, namely - *gamma* and *c*.

From the outputs, it is observed that the optimum values for gamma and c were :

gamma : 0.1
c : 1000000

B. FINAL SCORE USING OPTIMUM PARAMETERS

The optimum parameters were used to remodel the RBF, and new score was obtained [TABLE 2]

TABLE 2

MODEL	SCORE
RADIAL-BASIS FUNCTION	0.91

The decision boundaries were visualized for this RBF SVM classifier with the gamma and c parameters as 0.1 and 1000000 respectively. [FIG. 2]

The SVM model was used to predict the class for the testing data provided. The plot for the predicted data is shown in [FIG. 3]

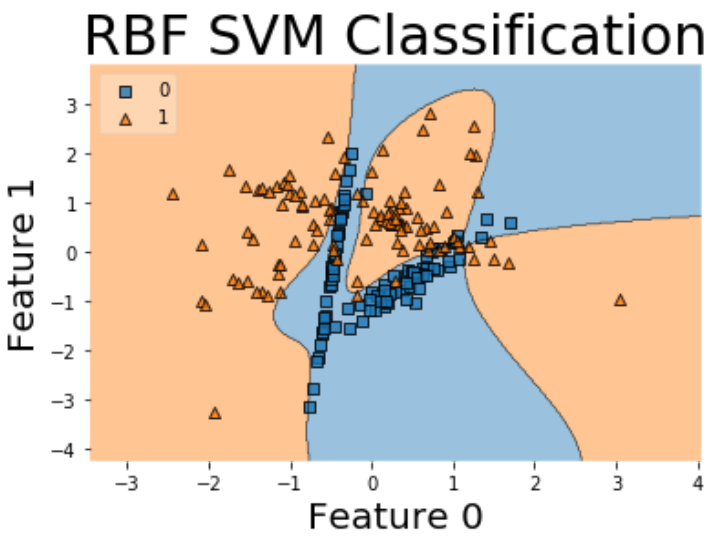

Fig. 2 RBF Decision boundary

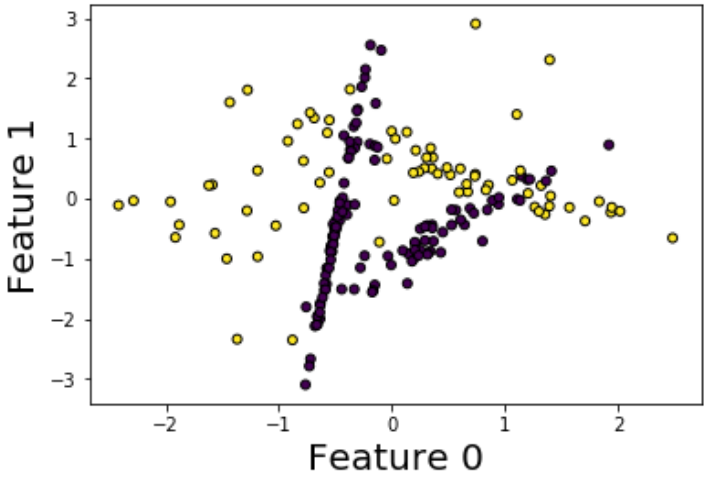

Fig. 3 Predicted Values

IV. REFERENCES

[1] https://medium.com/@LSchultebraucks/introduction-to-support-vector-machines-9f8161ae2fcb
[2] http://www.iosrjournals.org/iosr-jce/papers/Vol18-issue6/Version-3/K1806036065.pdf
[3] http://mypages.iit.edu/~jwang134/posts/Parameter-in-SVM.html
[4] http://pages.cs.wisc.edu/~matthewb/pages/notes/pdf/svms/RBFKernel.pdf
[5] http://openclassroom.stanford.edu/MainFolder/DocumentPage.php?course=MachineLearning&doc=exercises/ex8/ex8.html
[6] http://web.mit.edu/6.034/wwwbob/svm-notes-long-08.pdf
[7] https://www.ics.uci.edu/~welling/teaching/KernelsICS273B/svmintro.pdf