Double Moon Classifier using SVM

Aditya N. Govardhan

Abstract: In this report, the problem of double moon classification is analysed using Support Vector Machine (SVM) approach. The dataset is generated using random number generation functions of MATLAB. It has been found that the SVM classifier is simple and efficient in terms of parameters to be decided and its accuracy. This report aims to put light on the types of design parameters of SVM and evaluate the performance of different kernel functions.

***Index Terms*—SVM, penalty factor, kernels, non-linear classification**

I. Introduction

Double moon classification is one of the representative problems for pattern recognition in which data points are located on a plane as shown in Fig. 2. As the length d between half moons is varied, the data points can be made separable or inseparable.

Support Vector Machine (SVM) is one of the most simple and elegant approaches for classification problems. This paper analyses various types of kernels and their penalty parameters to determine accuracy of classification on test data.

II. Method

A. Data Generation

In this setup, 1400 data points are generated where each data point are (x, y) cartesian coordinates of the point. Out of these 1400 points, 1000 points are used for training the kernel and 400 points for testing. In this report, the classification problem is analyzed for d = -12.

B. SVM Design Parameters

SVM is a widely used machine learning algorithm for classification because of its simplicity. It can be designed using two parameters, the kernel function and the penalty parameter. The kernel functions return the inner product between two points in a suitable feature space. It is also known as the kernel trick. The penalty factor C determines how much stringent the classification should be on the training data. Higher the value of C, higher is the penalty and hence more accuracy in training data. However, note that this doesn't ensure high accuracy in testing data since the penalty factor is available only for training data. Hence large values of C could lead to overfitting problem.

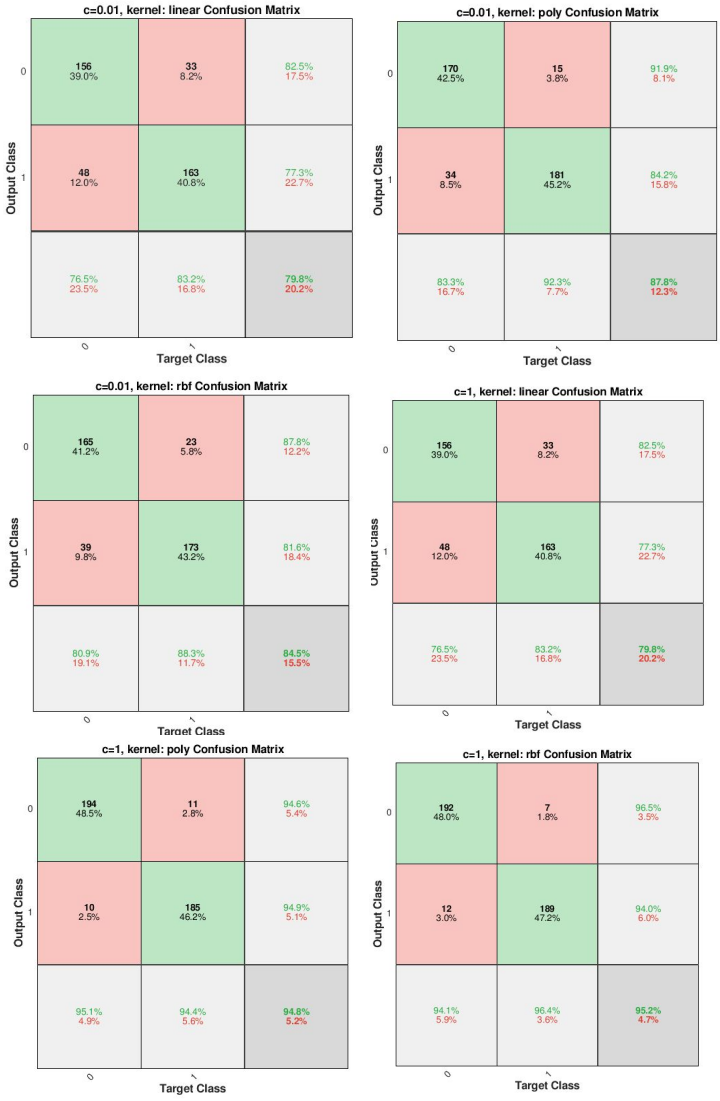

Fig. 1. Confusion matrices for c = 0.01 and c = 1

Aditya N. Govardhan is with the Arizona State University, Tempe, AZ 85281 USA (e-mail: Aditya.Govardhan@asu.edu).

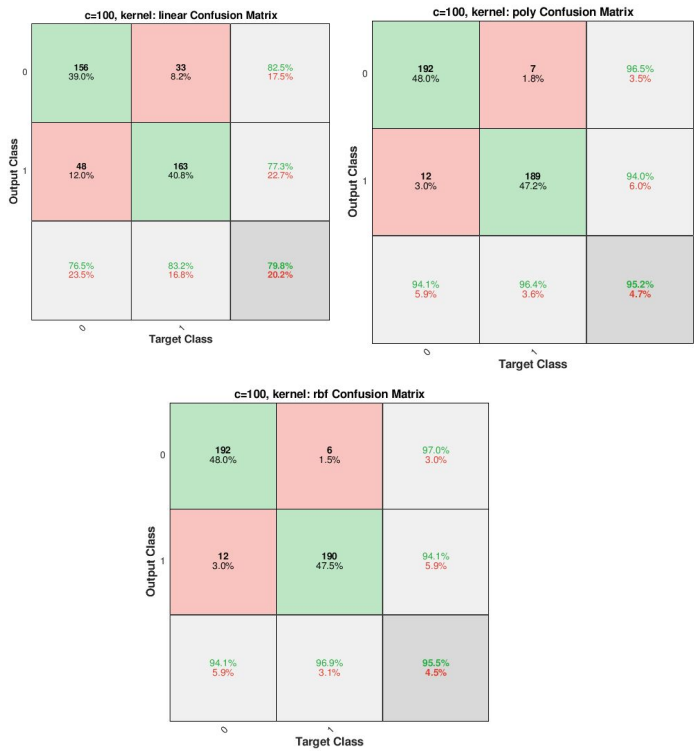

Fig. 2. Confusion matrices for c = 100

In this report, the SVM is designed for three values of the penalty parameter (0.01, 1 and 100) and three types of kernel functions (linear, polynomial and radial basis function). Thus, the SVM is analyzed for nine combinations.

III. Results

For each combination, confusion matrix and classification boundary is plotted. Fig. 1. shows the confusion matrices of the three kernels for c=0.01 and c=1. Fig. 2. shows the confusion matrices of the three kernels for the c=100. Boundary decisions for all the combinations have been plotted in Fig. 3. Highest accuracy is obtained for RBF and c=100, with accuracy of 95.5%.

IV. Conclusion

Radial basis function and polynomial function being non-linear in nature, show higher accuracy for testing dataset. As the value of penalty parameter increases, the algorithm fits the training data better. Corresponding accuracy for testing data also increases. For higher values of c, the performance is found to be degrading due to overfitting. However, it can be observed that with two design parameters, such high accuracy can be obtained.

References

[1] MATLAB Statistics and Machine Learning Toolbox

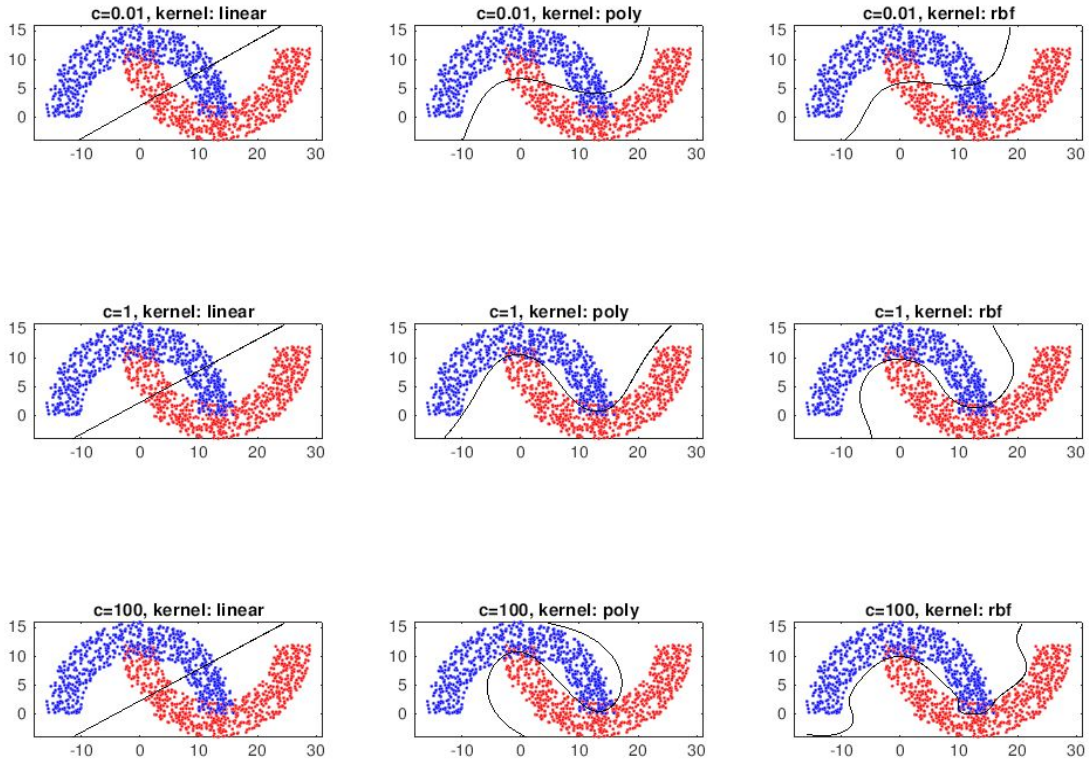

Fig. 3. Decision boundaries