**ECE 8527: Introduction to  
Machine Learning and Pattern Recognition**

# HW No. 7: Linear Discriminant Analysis

This assignment is simple ☺

Repeat HW #3 but use linear discriminant analysis to classify the data. Hopefully, if your code for HW #3 was written in a modular fashion, you will be able to reuse most of your code, and simply swap training and classification modules.

You can also use the Java applet provided to test and debug your code.

How does this performance compare to HW #3? Justify your findings.

# Exercise 1

In this exercise, we are using the same data generation function used in Homework 3 for the YingYang data. The difference between this homework and Homework 3 is that in this homework assignment we are using LDA to discriminate these two sets of data instead of using the maximum likelihood classifier.

The first step of this exercise is to generate the training data set for the LDA training purposes. This data is passed into the function to determine the classifier. Once the classifier is set, it is then passed via a set of evaluation data for the error rate of these classifiers. Both error rates are displayed after calculation in the terminal for comparison.

In additional to compute the error rate, the LDA decision surface is also drawn as an overlap on the scatter plot of the data set. The LDA result is reduced into one dimension, in this case, the x axis, but attached with values for the y axis in order to show the decision line in a two-dimensional plot. The following diagram shows a demonstration of the plot with the scatter plot of selected overlap.

The green line in the LDA for the evaluation data is the dummy line modified from the result of LDA into the two-dimensional space. This value varies based on different data sets which will be represented in the following section.

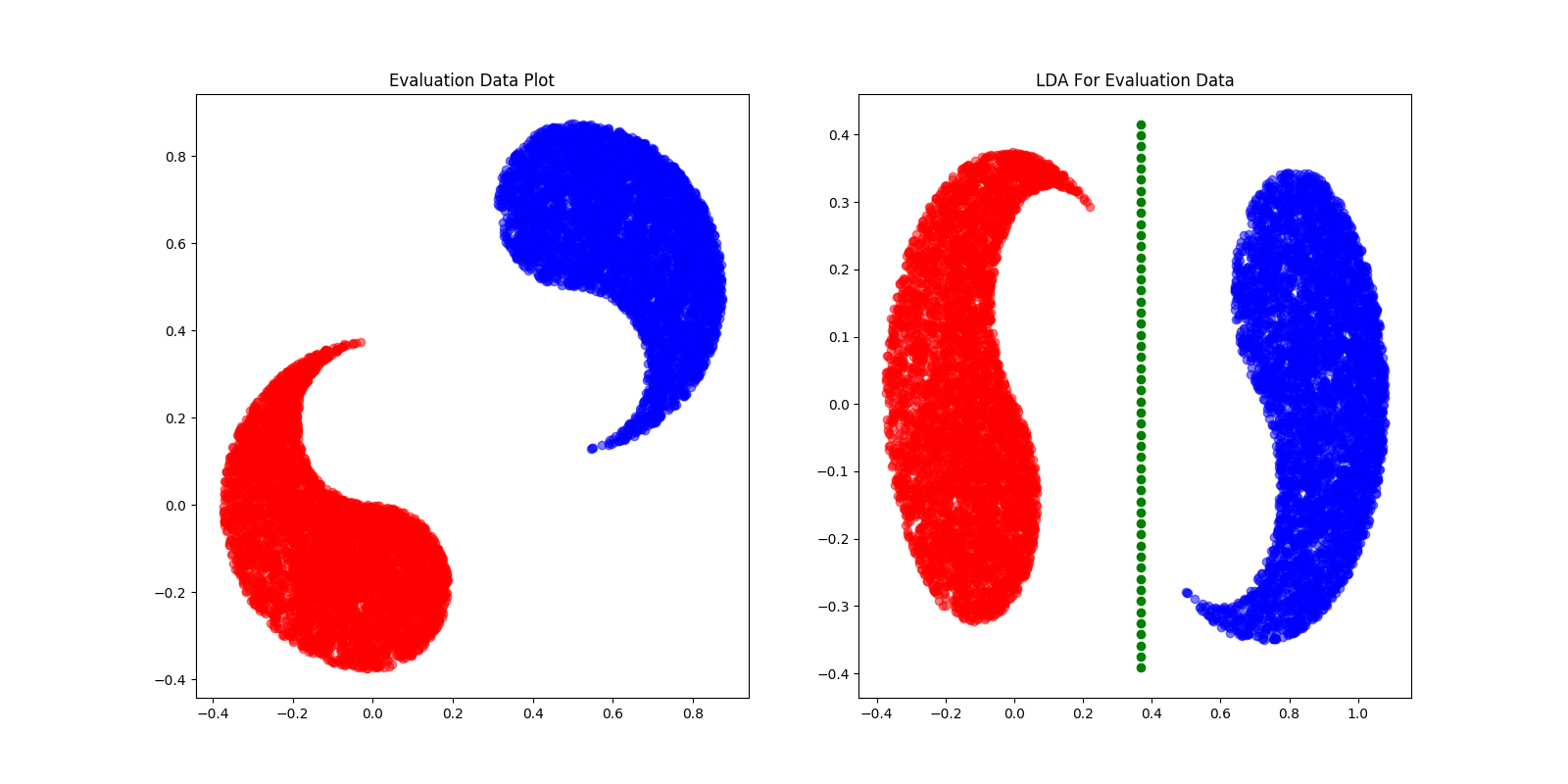


Figure Scatter Plot with Overlap at -1

# Exercise 2

In this exercise, the function that is created in the previous exercise is called multiple times for the purpose of generating different error rate for different overlap values of the data set. The function written from the previous exercise is named “athletic” for the purpose of simplification.

With a different set of values for the overlap input with the athletic function, the error rate table is generated as following with no additional parameter values such as plot and save. The error rate that it is generated is normalized to have the unit of “%” so that it would be easier to compare with each other. The following table shows the result obtained from the function.

Table Error Rate in terms of Overlap

|  |  |  |
| --- | --- | --- |
| Overlap | Train Error (%) | Evaluation Error (%) |
| -1 | 0 | 0 |
| -0.25 | 0.885 | 0.78 |
| -0.1 | 4.19 | 4.11 |
| 0 | 9.69 | 9.16 |
| 0.1 | 16.505 | 15.57 |
| 0.25 | 29.575 | 29.54 |
| 1 | 1.44 | 1.56 |

From the chart above, it shows that the error rate gets worse when the overlap value is gettering closer to 0.25 and then decreases. The following diagram shows the result output of scatter plot of the data set with an overlap of 0.25.

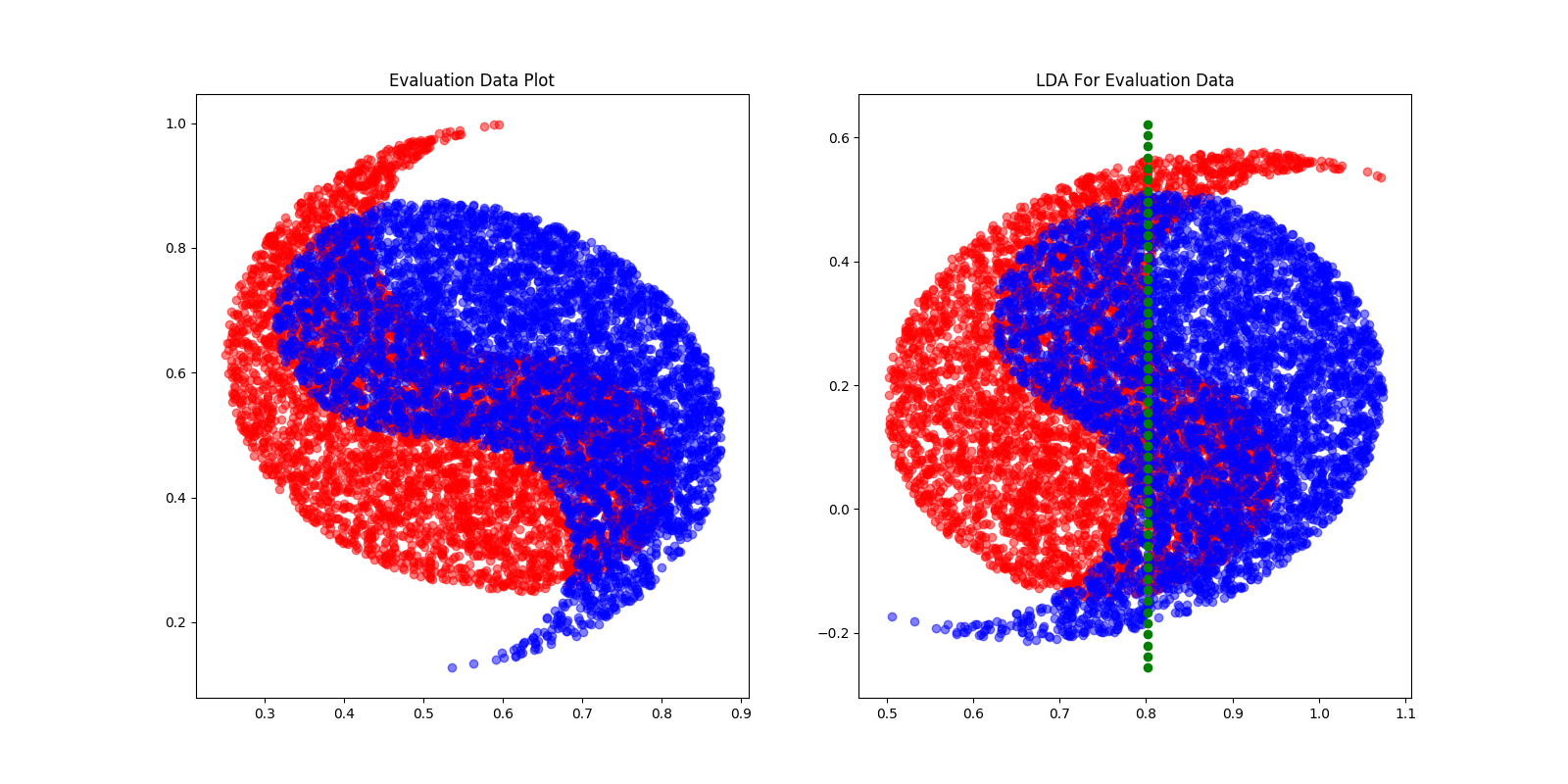


Figure Scatter Plot with Overlap at 0.25

One additional set is also generated with an overlap value of 0.45 showing the error rate of around 48%. This error rate is the worst error rate obtained among several testing values input into the athletic function for the desired error rate.

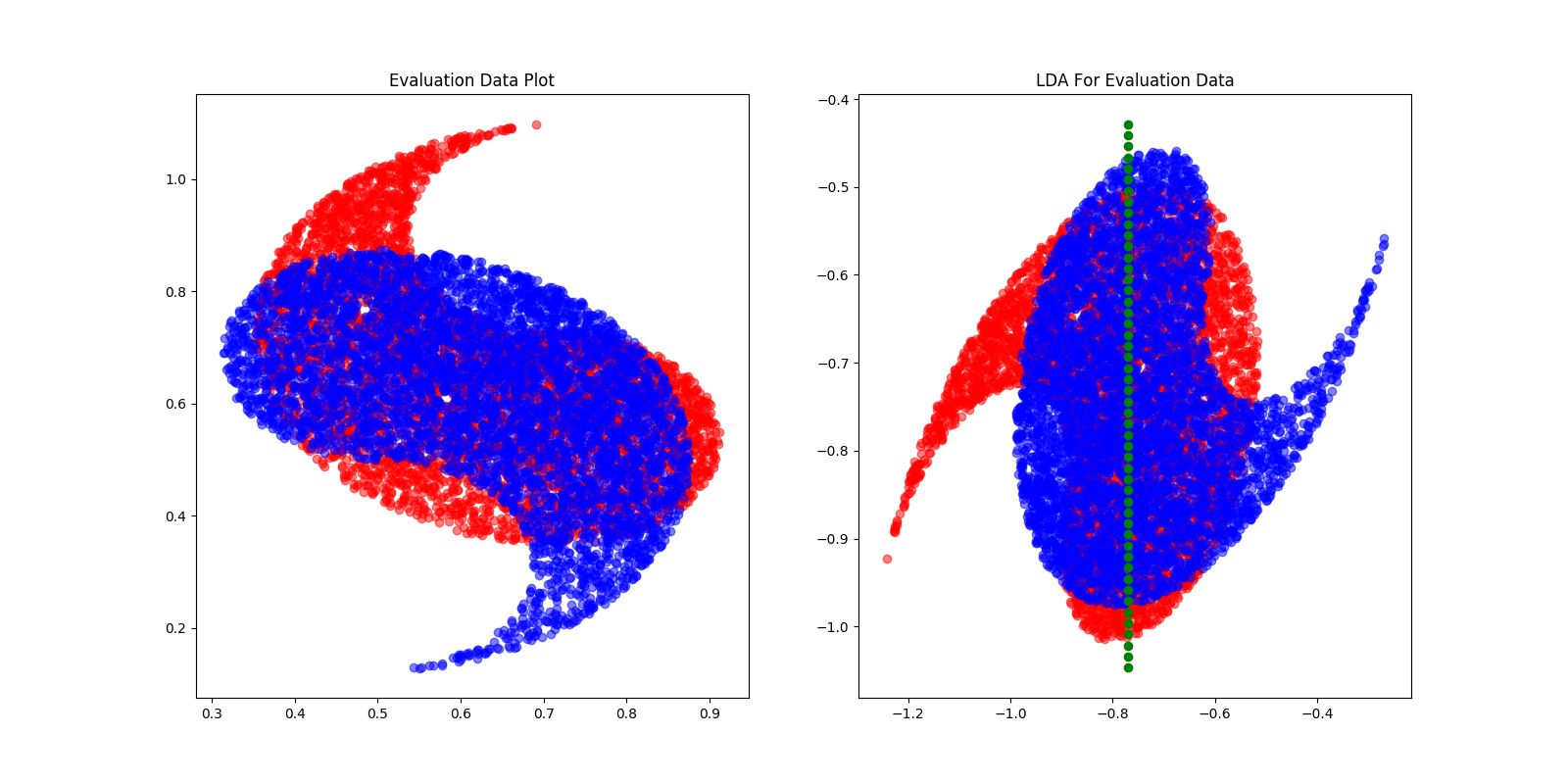


Figure Scatter Plot with Overlap of 0.45

The error rate shows in the chart above suggests that the error in between the training data and the evaluation data does not have much difference in terms of the error rate, which is within the range of 1%. With the overlap value of 0.45, these two classes are basically on top of each other, thus the classifier doesn’t have a very distinguishable axis to tell these two classes different. Based on the observation for the LDA data, the error rate seems to be reasonable.

As suggested by the maximum likelihood classifier, the error rate chart shows the same trend as it was with the maximum likelihood classifier. The error rate shows an increase as the overlap value increases from -1 to 0.45 and then decrease from there. The biggest difference in between the maximum likelihood classifier and the LDA is that when classified data for the LDA is rotated towards he most significant axis to support the reduce of dimension of the LDA process. In this specific case for the YingYang data set, there are only two classes to be classified and the majority of the overlap shows a significant axis that can be used for LDA. So in this specific case, these two classifiers relatively the same, but however, if there is no significant axis that can be used in the for the classification, the LDA would be failing while maximum likelihood classifier would have a chance not to be failing.