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

# HW No. 3: Nonlinear DEcision Surfaces

For this assignment, you will use the data generator located here:

*https://www.isip.piconepress.com/courses/temple/ece\_8527/resources/data/set\_05/yinyang.py*

This tool generates data that follows the shape of a yin yang (Taijitu) symbol. The interface is simple:

*python3 yinyang.py N0 N1 <overlap>*

where N0 is the number of points in class “0”, N1 is the number of points in class “1”, and <overlap> is a parameter that controls the overlap between the two classes. Its range is [-1,1].

The tasks to be accomplished in this homework assignment are:

1. Generate a training set with 10,000 points in each class using an overlap of -1. Train a maximum classifier likelihood on this data. Generate an independent evaluation set of 5,000 points per class. Evaluate it using your classifier trained only on the training data. Measure performance on both the training set and the evaluation set. Draw the decision surface that corresponds to your classifier. Superimpose this over a scatter plot of the data.
2. Repeat this for overlap parameter values of **-0.25**, -0.10, 0.0, 0.10, 0.25. Generate a table of performance on the training and evaluation sets as a function of this parameter.

Discuss the performance of this classifier based on what you have observed in these experiments. What could you do to improve performance?

# Exercise 1

In this exercise, the data is generated via a given function, yinyang, but modified so that this class can be imported directly into new function and output data from different classes. The result of this modification is the data output has the composition of how train data and dev data (evaluation data) look like in the previous homework exercise so that classes that was used in them can be ported and modified quite easily.

The first step of this exercise is to generate the train data set for the max likelihood classifier for individual data sets, class 0 and class 1. Once the classifier is set, it is run through the original training data as well as the newly generated dev data in order to evaluate the performance in terms of the error rate of the classifier. Both error rate is printed out in the terminal window for comparison.

In addition to compute the error rate, the decision surface is also drawn as an overlap on the scatter plot of the data point. The calculation of the decision surface basically determines where the probability of class 0 and class 1 is the same within the range of input array, which is set to obtained automatically based on the current axis limit of the plot. The contour function used in plotting is set to select only where the difference of the probability is 0. The following diagram generated from the plot will demonstrate the result of the decision surface calculation.

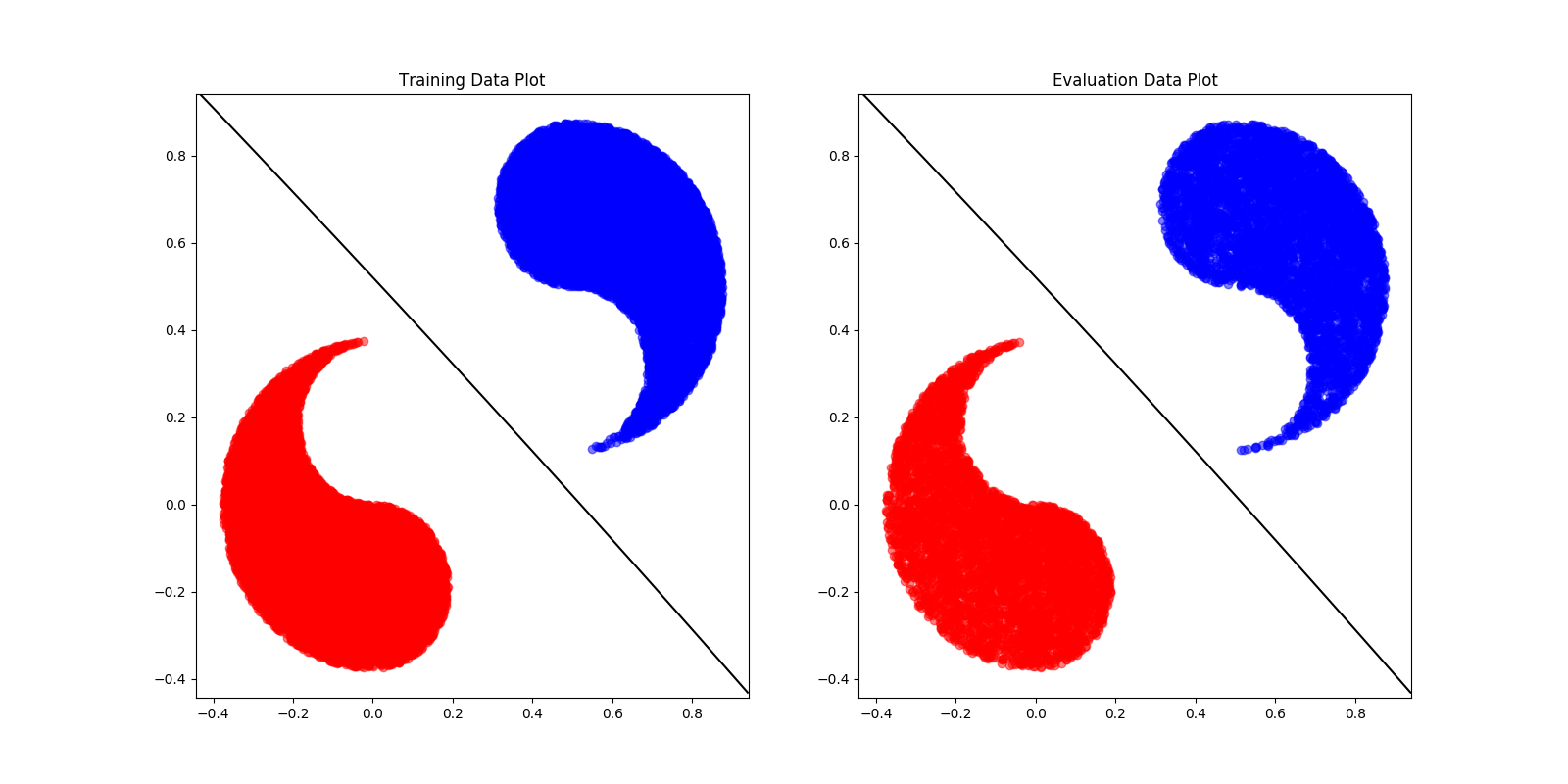


Figure 1 Scatter Plot with Overlap at -1

# Exercise 2

In this exercise, the function that is created in the previous exercise is called serval times for the purpose of generating different error rate for different overlap value of the yinyang data set. The function that is used in the previous exercise is call “athletic” in order to simplify when it is mentioned in the following explanations.

With a different set value for the overlap, the athletic function is run to generate the error rate of the train data and the dev data. With each run of the athletic function, it sets the overlap parameter of the function to the desired value and return the classified error rate of the train data and the dev data while setting all graphic and indicational parameters of the function to none to minimize the complexity of the function. The error rate that it outputs is set to in the unit of percent, %, so that it is normalized with more accuracy upon printing the storage array. The following diagram is taken from the terminal window showing the result of the computation for different values of the overlap for the yinyang data set.

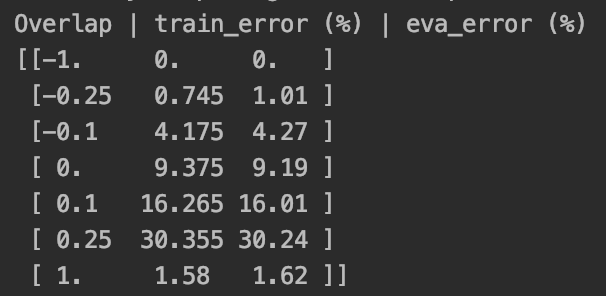


Figure 2 Error Rate in terms of Overlap

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

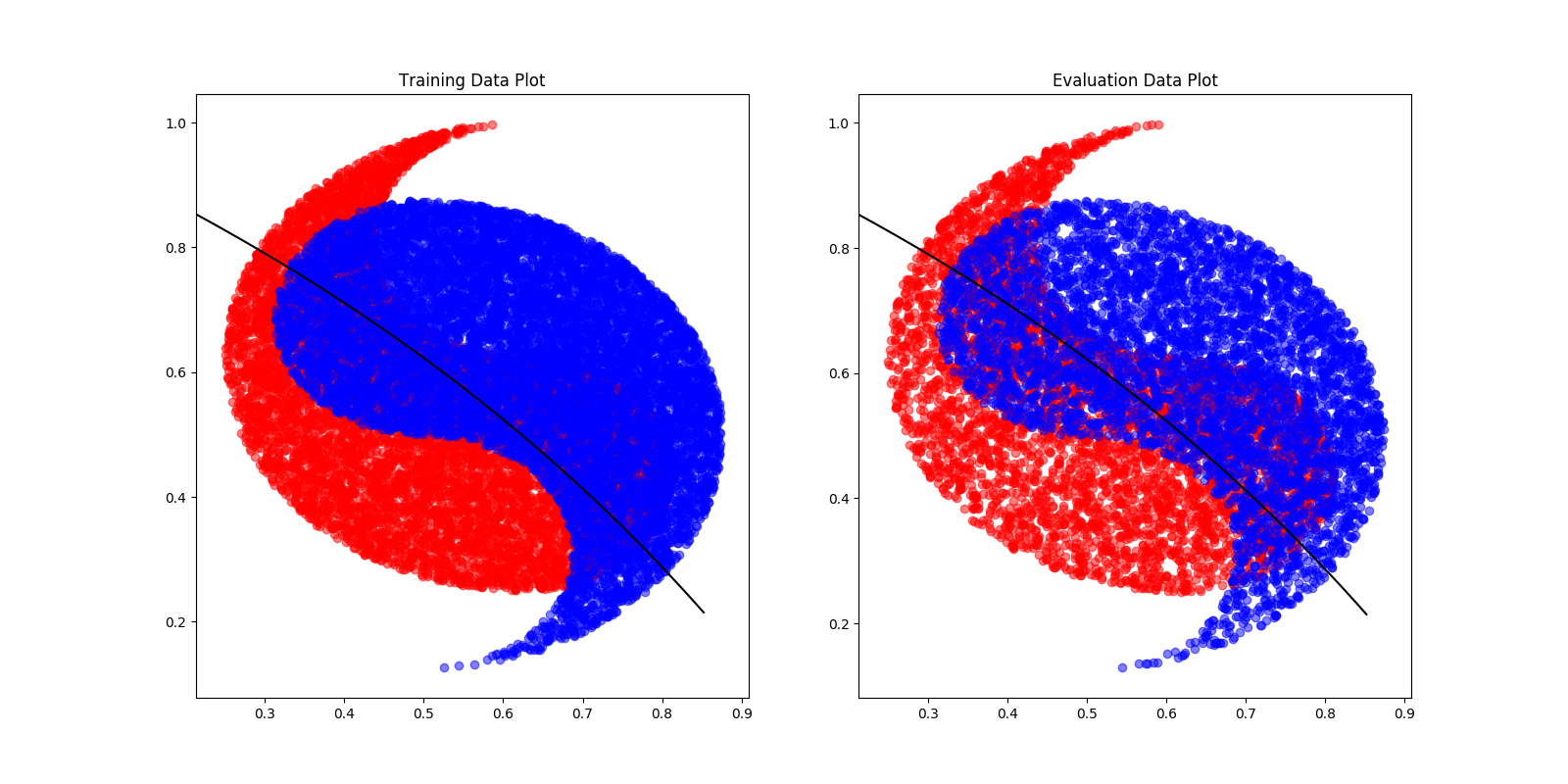


Figure 3 Scatter Plot with Overlap at 0.25

One additional set is also generated with overlap value of 0.45 showing the error rate of 48.34 % to the training data set and 48.32% towards the dev data set, which is the worst error rate obtained among several testing value for different overlap value of the data set.

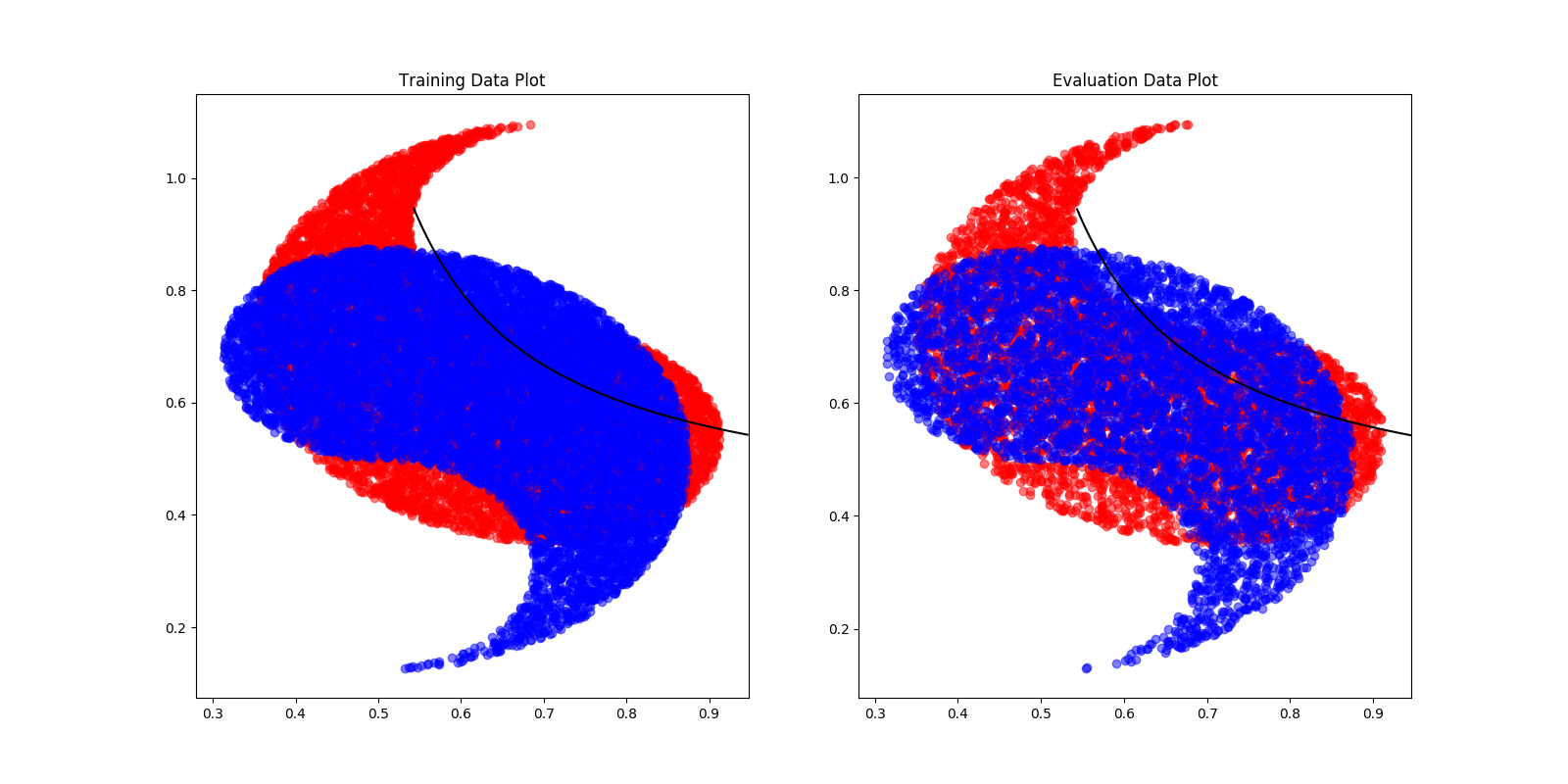


Figure 4 Scatter Plot with Overlap at 0.45

By comparing the error output of different overlap value, the chart shows that the classifier performed relatively the same with a difference of the error with in about 0.1% of each other with the expectation when the overlap value is -0.25 and its difference in error rate of about 0.3%.

The overlap value that is passed into athletic also got passed into the generation of the yinyang data set. It is the critical element that differs the distance in between two classes. As the overlap value is set to -1, it shows that these two class is very separable as can be visually seen in the scatter plot from the previous exercise while set it to 0.45 shows these tow classes are basically on top of each other.

For such a two-feature data set, the performance of the classifier can be determined on how distinguishable of these two classes. As shown above, when them are very distinguishable, when overlap is -1, the classifier shows an extraordinary performance. However, when the overlap is 0.45, where these two classes are visually on top of each other meaning not distinguishable at all, it shows a terrible performance with the error rate of 48%. As for how to improve the performance of this classifier, the most direct approach is to use the most distinguishable feature of the set and using another different, more distinguishable feature, maybe even introduce a different set of feature that is more unique than the selected two, when these two feature is visually overlapping too much. On the other hand, if limited to two features, a different classifier method should be used that takes in several unique points for classification instead of the mean value of these two classes as used in the maximum likelihood classifier of this one.