1 Motivation

The programming language is python. The basic algorithm is the nearest centroid mean algorithm. The algorithm initially compute the mean for each class. The mean is a vector which is computed by averaging two features of training data. The algorithm to classify data is to compute the Euclidean distance between feature vector and mean vector trained by the classifier.

2 Problem solution

2.1 problem a

Two figures are generated by using *PlotDecBoundaries()*. The CSV file is read by the library supported by the numpy. The data is stored into the *ndarray*. The Figure 1 shows the data points, class mean and decision boundary for synthetic data set 1. The Figure 3 shows these information for synthetic data set 2.

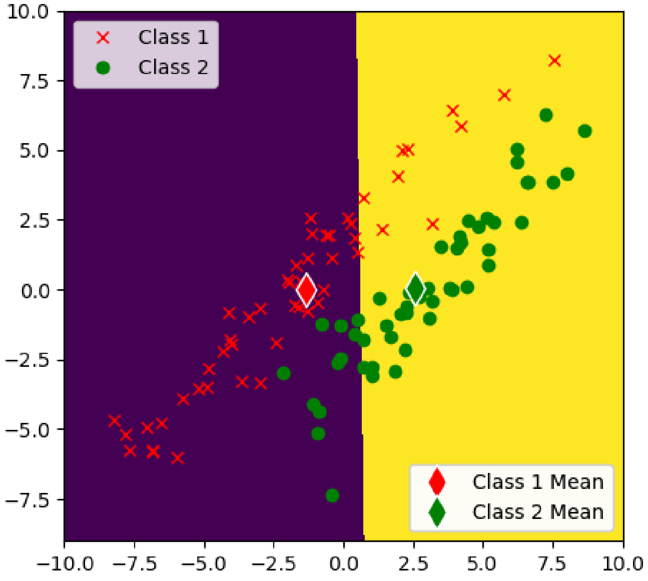

Figure 1: synthetic train1

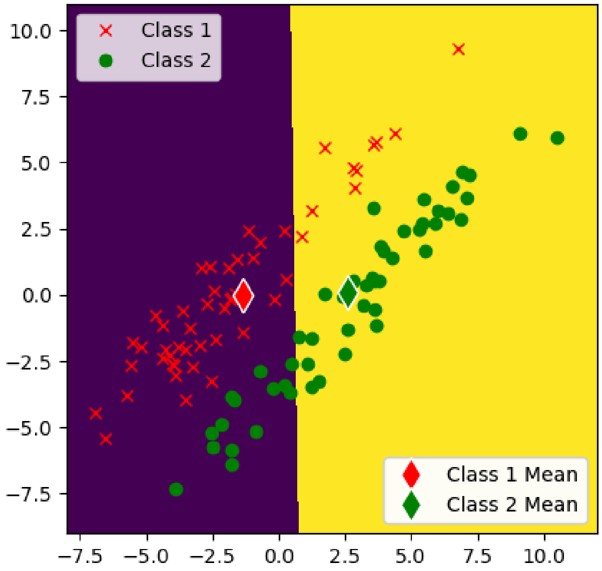

Figure 2: synthetic test1

Data	Error rate	Test samples
synthetictrain1	0.24	100
synthetictest1	0.24	100
synthetictrain2	0.04	100
synthetictest2	0.04	100

Table 1: Error rate

2.2 problem b

The error rate of both synthetic data set is shown in the table 2. The table also shows the error rate of the synthetic data set 2 is smaller than the error rate of the synthetic data set 1. The performance of data set 2 is also good in the training data set.

2.3 problem c

The Figure 5 shows the training data points, sample mean and decision boundary for wine data set.

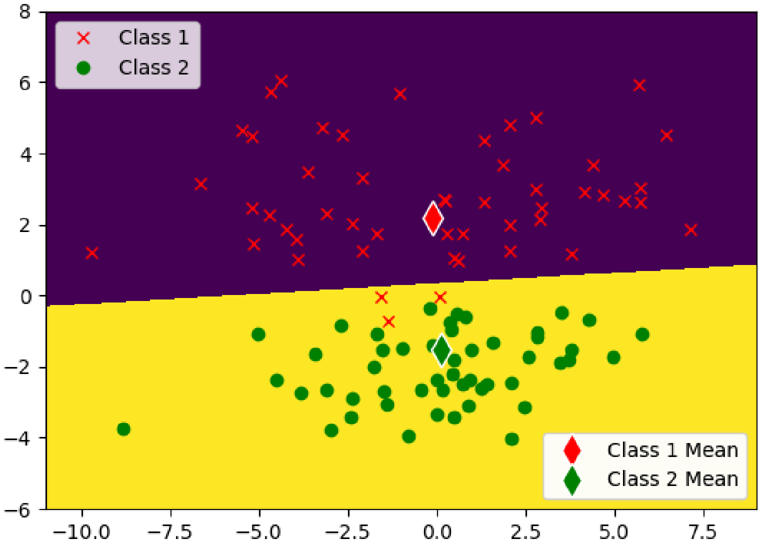

Figure 3: synthetic train2

2.4 problem d

ID	Error rate(Train Data)	Error rate(Test Data)	Feature1	Feature2
1	0.11	2	100	1
2	0.24	3	100	2
3	0.04	4	100	2
4	0.04	5	100	2

Table 2: Error rate

2.5 problem e

3 Summary

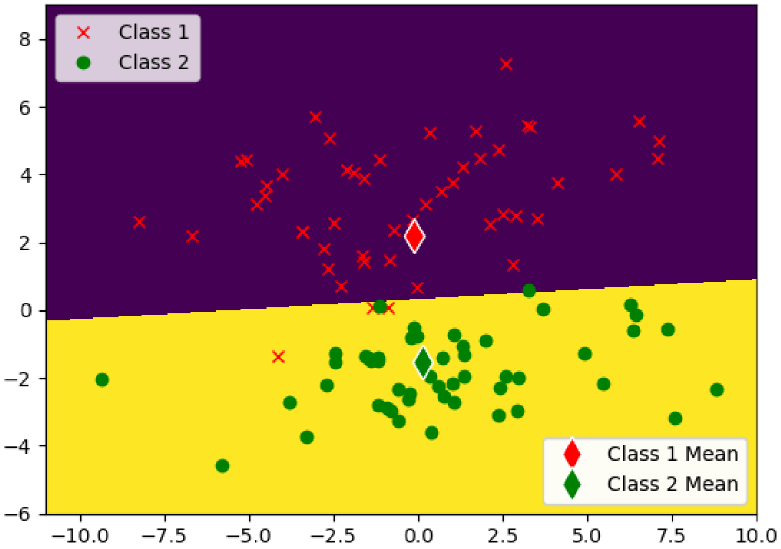

Figure 4: synthetic test2

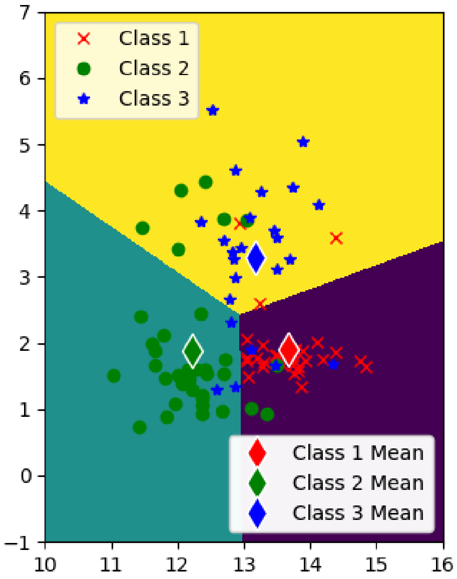

Figure 5: the data scatter plot of wine data set whose feature is 0 and 1