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.

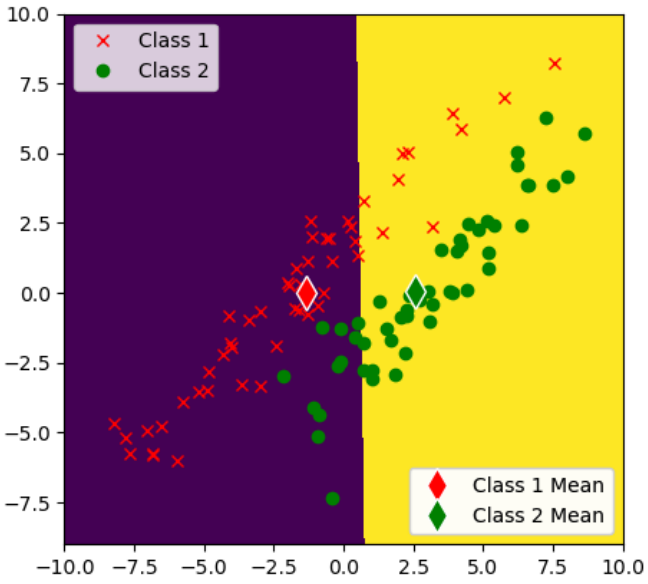

Figure 1: synthetic train1

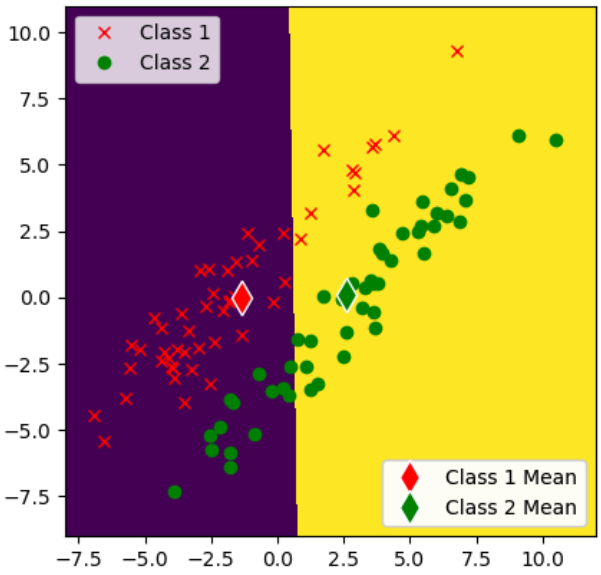

Figure 2: synthetic test1

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 1. 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. The class error is computed by running python program synthetic.py. The console display the error rate. The error rate for both data set 1 and 2 is the same when the error rate of training data and error rate of testing data is compared. If the error rate is different for test set and train set, this means some data is not collected for some reason. The train set is not complete and can not fully represent the class. As error rate in our data set is the same, there is no problem for data collection of training data set. If error

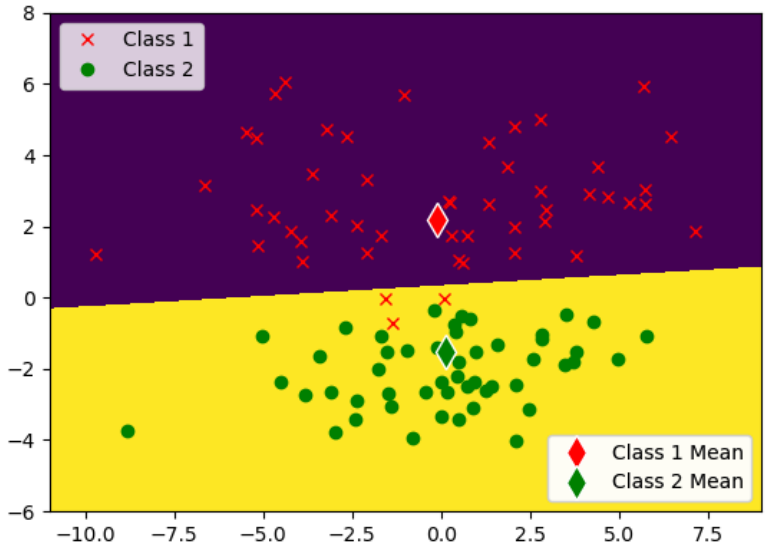

Figure 3: synthetic train2

rate is very high, this means the algorithm perform bad for classification.

2.3 problem c

The Figure 5 shows the training data points, sample mean and decision boundary for wine data set. The sample mean is produced by the feature 1 and 2 of training data. Error rate of the train data was shown in the table 2.

2.4 problem d

The Figure 6 has shown boundaries of best features. The principle to select best feature is to compare the error rate of the training data. The feature which produces the lowest error rate is selected. All combination of two features in thirteen feature was compared. The best features were selected to produce the Figure 6. The table 2 also provided some error rates of different pair of features.

2.5 problem e

The mean of error rate of all pairs of two selected feature is 0.336, and the variance of error rate is 1.285. The result was generated by program *wine.py* . From comparison between error rate of train data and test data of chosen best

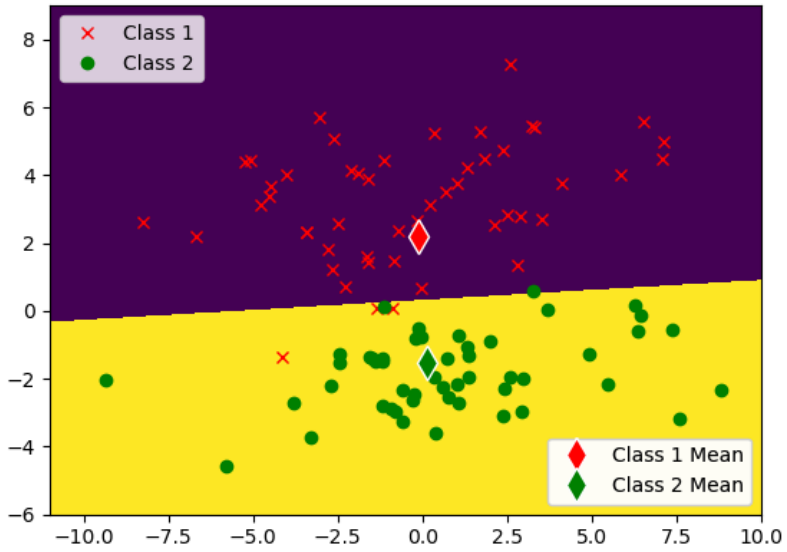

Figure 4: synthetic test2

Description	Error rate(Train Data)	Sample Number
best feature 1 and 12 (train data)	0.079	89
feature 1 and 2 (train data)	0.020	89
best feature 1 and 12 (test data)	0.742	89
feature 1 and 2 (test data)	0.224	89
feature 2 and 3 (train data)	0.393	89
feature 2 and 3 (test data)	0.730	89
feature 2 and 4 (train data)	0.393	89
feature 2 and 4 (test data)	0.674	89

Table 2: Error Rate for wine data set

pair of features, the best features seem to be over fitting for the train data set. The error rate of test data set is extremely high.

3 Summary

The nearest centroid mean algorithm performs well on synthetic data set. For real problem of wine classification, some features have small error rate for train

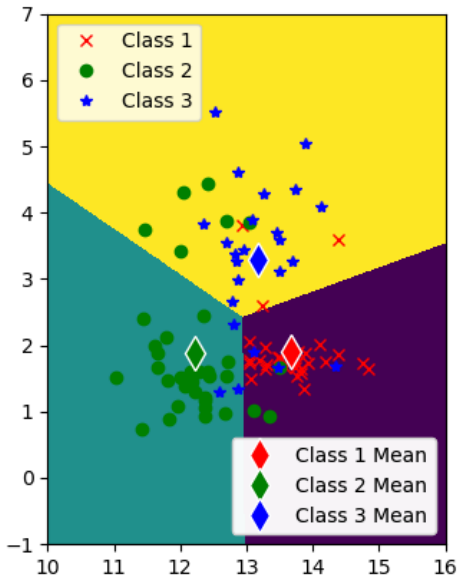

Figure 5: the data scatter plot of wine data set whose feature is 0 and 1

data. The error rate of test data of same features is high. This case presents the defect of the algorithm. The selected features are over-fitting for train data set.

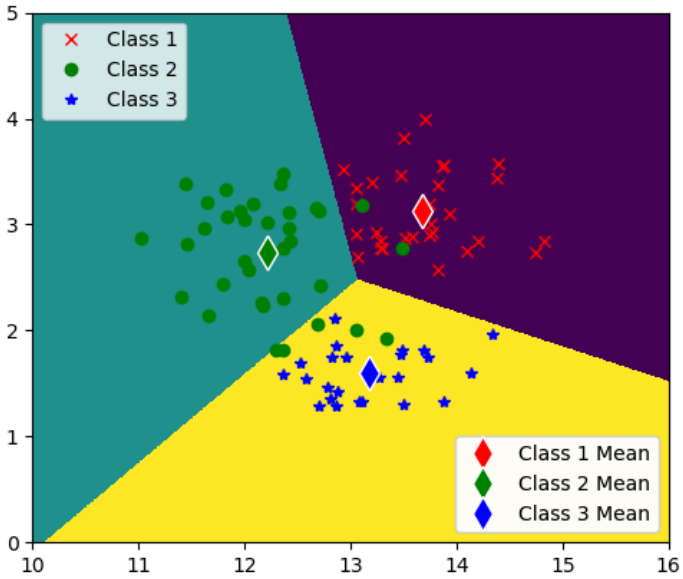

Figure 6: graph produced by the feature 1 and 12 of train data