EE559 Homework 1 (week 2)

Jingquan Yan

USC ID: 1071912676

Email: jingquan@usc.edu

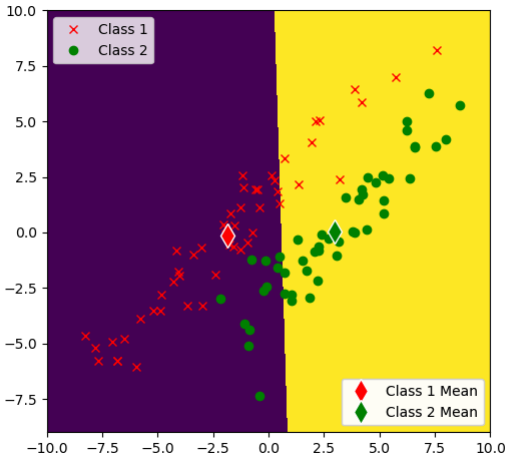
EE559 repository: Github

(a) :

> For each of the two synthetic datasets, there are in total C=2 classes and D=2 features. For each synthetic dataset: (i) train the classifier, plot the (training-set) data points, the resulting class means, decision boundaries, and decision regions (using PlotDecBoundaries()); also run the trained classifier to classify the data points from their inputs; give the classification error rate on the training set, and separately give the classification error rate on the test set. The test-set data points should never be used for training. Turn in the plots and error rates.

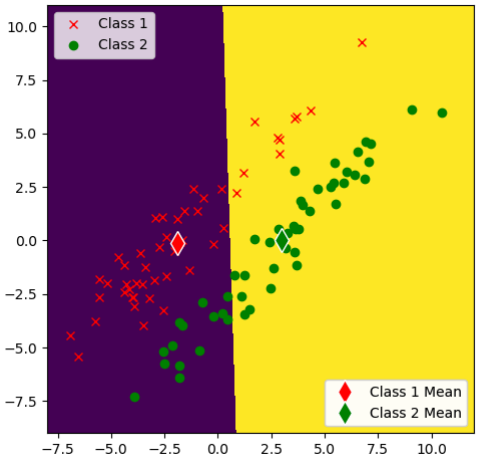
Solution:

1. Classifier plot and error rates for **training-set** 1:



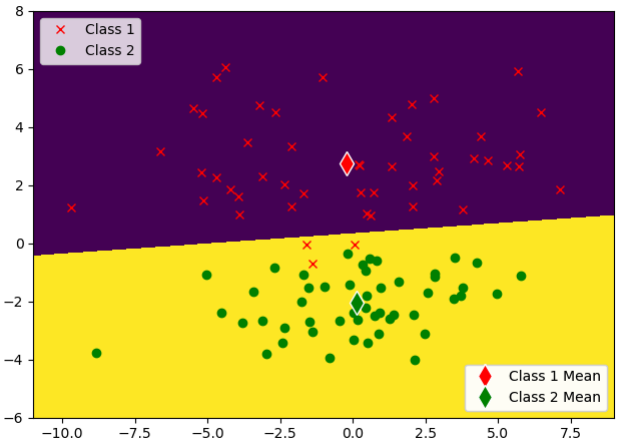

```
C:\Users\Yan\AppData\Local\Programs\Python\Python37\python.exe "C:/Git/559/HW1/datasets/Problem 1.py"
The error rate is: 0.20999999999999996
```

2. Classifier plot and error rates for **test-set** 1:



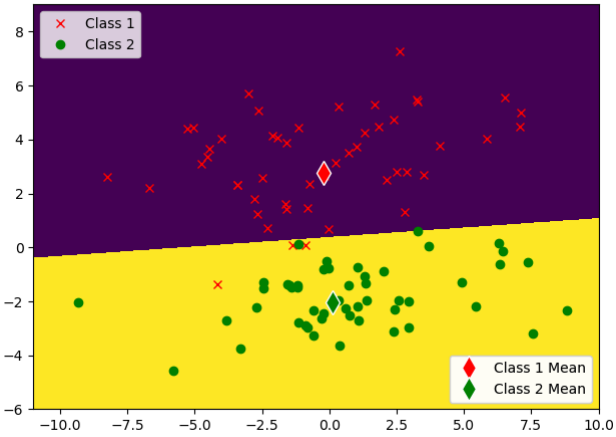
```
C:\Users\Yan\AppData\Local\Programs\Python\Python37\python.exe "C:/Git/559/HW1/datasets/Problem 1.py"
The error rate is: 0.24
```

3. Classifier plot and error rates for **training-set** 2:




```
C:\Users\Yan\AppData\Local\Programs\Python\Python37\python.exe "C:/Git/559/HW1/datasets/Problem 1.py"
The error rate is: 0.030000000000000027
```

4. Classifier plot and error rates for **test-set** 2:




```
C:\Users\Yan\AppData\Local\Programs\Python\Python37\python.exe "C:/Git/559/HW1/datasets/Problem 1.py"
The error rate is: 0.040000000000000036
```

(b)

> Is there much difference in error rate between the two synthetic datasets? Why or why not?

Answer:

There is an obvious difference in error rate between the two synthetic datasets.

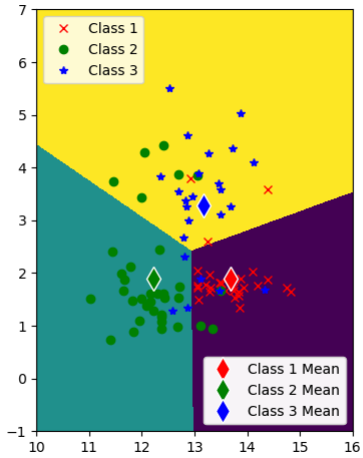
The reason is that the decision criteria only brings the average coordinate into consideration and ignores the data distribution (overall shape). Hence the error rate may vary when we have different distribution of input.

(c)

> For the wine dataset, there are in total C=3 classes (grape cultivars) and D=13 features (measured attributes of the wine). In this problem you are to use only 2 features for classification.
>
> Pick the first two features (alcohol content and malic acid content), and repeat the procedure of part (a) for this dataset.

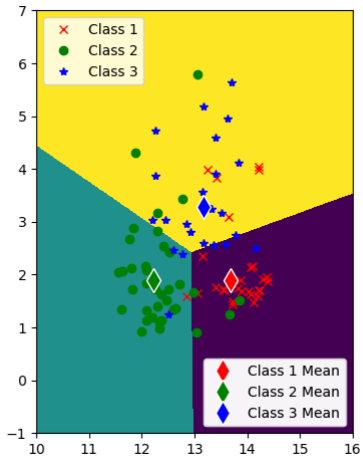
Solution:

1. Classifier plot and error rates for wine **training-set** with column 1 and 2:




```
C:\Users\Yan\AppData\Local\Programs\Python\Python37\python.exe "C:/Git/559/HW1/datasets/Problem 1.py"
The error rate is: 0.202247191011236
```

2. Classifier plot and error rates for wine **test-set** with column 1 and 2:




```
C:\Users\Yan\AppData\Local\Programs\Python\Python37\python.exe "C:/Git/559/HW1/datasets/Problem 1.py"
The error rate is: 0.2247191011235955
```

(d)

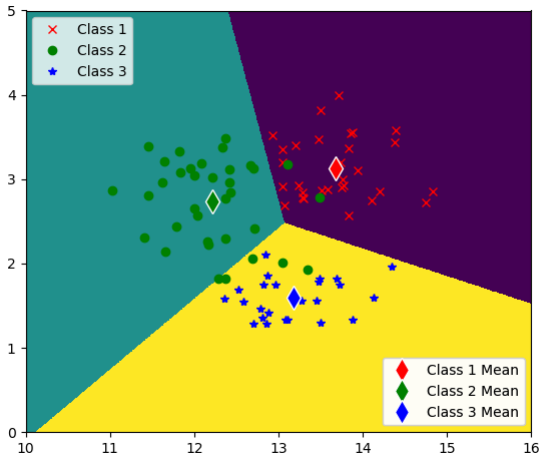
> Again for the “wine” dataset, find the 2 features among the 13 that achieve the minimum classification error on the training set. (We haven’t yet covered how to do feature selection in class, but will later in the semester. For this problem, try coming up with your own method - one that you think will give good results - and see how well it works.

Solution:

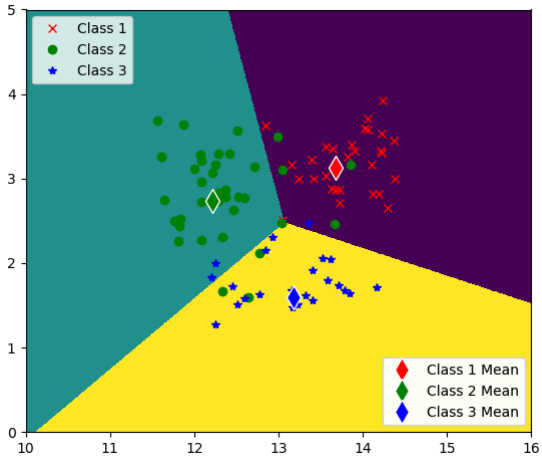
1. We use the error rate as the criteria to choose the 2 columns that could make the best classification (least error rate) of the data. After traversal all 78 combinations ($\frac{13^2-13}{2}$), the 2 features in **training data** that reaches the minimum error rate are columns 1 and 12 with error rate 0.07866:

   ```
   C:\Users\Yan\AppData\Local\Programs\Python\Python37\python.exe "C:/Git/559/HW1/datasets/Problem 1.py"
   The minimum error rate comes with columns 1 and 12 with an error-rate of 0.0786516853932584
   ```

   And the corresponding plot is:



2. To examine the generalization ability of the two columns selected above, we input the corresponding columns in the **test data**, the result is as follows:




```
C:\Users\Yan\AppData\Local\Programs\Python\Python37\python.exe "C:/Git/559/HW1/datasets/Problem 1.py"
The error rate is: 0.1235955056179775
```

(e)

For **training data**, there are differences for different pairs of features but not such large. The error rates of different pairs are around 0.1-0.5 and the mean and standard deviation are as follow:

```
The mean of error_rate is : 0.3357821953327571
The standard deviation of error_rate is : 0.1291926181039988
```

For **test data**, there are differences as well and the standard deviation of error rate is lower than the training dataset:

```
The minimum error rate comes with columns 1 and 7 with an error-rate of 0.1123595505617978
The mean of error_rate is : 0.32728320368769814
The standard deviation of error_rate is : 0.0963386128805748
```