Thomas Chung
Fang
COEN 140

Homework 2

1. Break the sample into 80% for training, and 20% for test datasets. You can choose the first 80% instances from each class for training and the rest for testing.

See Code.

2. Build an LDA classifier based on the training data. Report the training and test errors for your classifier.

After running a LDA classier on the training data where **the training data was used to calculate the mean average and the covariance matrix**, the program outputted:

```
The test error rate was 0.0 (0%)
The training error rate was 0.025 (2.5%)
The Setosa test error rate was 0.0 (0%)
The Versicolor test error rate was 0.0 (0%)
The Virginica test error rate was 0.0 (0%)
The Setosa training error rate was 0.0 (0%)
The Versicolor training error rate was 0.05 (5%)
The Virginica training error rate was 0.025 (2.5%)
```

3. Build a QDA classifier based on the training data. Report the training and test errors for your classifier.

After running a QDA classier on the training and test data, the program outputted:

```
The test error rate was 0.0 (0%)
The training error rate was 0.016666666666666666 (1.6%)
The Setosa test error rate was 0.0 (0%)
The Versicolor test error rate was 0.0 (0%)
The Virginica test error rate was 0.0 (0%)
The Setosa training error rate was 0.0 (0%)
The Versicolor training error rate was 0.05 (5%)
The Virginica training error rate was 0.0 (0%)
```

From both the LDA and QDA case, we can see that our test error rate was at 0% which indicates to us our classification model is working superbly well. Furthermore, our training error rates are

2.5% and 1.7% for LDA and QDA respectively which indicates to us that we **did not overfit our data** for the training set.

4. Is there any class linearly separable from other classes? Explain your answer based on your experiments.

For a class to be linear separable, its individual error rate has to be 0%. From the QDA example above, we see that the test error rate (the error rate which we are most interested in) has **Setosa, Versicolor, and Virginica** at 0%. So from that, it would be safe to say that all three classes are linearly separable.

5. Are any of the variables not important in classifying iris type? Explain your answer based on your experiments.

The follow output are the results of repeatedly running the LDA algorithm whilst removing one feature at a time. There are three features at all times and are removed/put back in this order: sepal width, sepal length, petal width, and petal length.

(Removed Sepal Width)

```
The test error rate was 0.0
The training error rate was 0.025
The Setosa test error rate was 0.0
The Versicolor test error rate was 0.0
The Virginica test error rate was 0.0
The Setosa training error rate was 0.0
The Versicolor training error rate was 0.05
The Virginica training error rate was 0.025
```

(Removed Sepal Length)

```
The test error rate was 0.0
The training error rate was 0.016666666666666666
The Setosa test error rate was 0.0
The Versicolor test error rate was 0.0
The Virginica test error rate was 0.0
The Setosa training error rate was 0.0
The Versicolor training error rate was 0.025
The Virginica training error rate was 0.025
```

(Removed Petal Width)

```
The test error rate was 0.03333333333333333
The training error rate was 0.041666666666666664
The Setosa test error rate was 0.0
The Versicolor test error rate was 0.0
The Virginica test error rate was 0.1
The Setosa training error rate was 0.0
The Versicolor training error rate was 0.05
The Virginica training error rate was 0.075
```

(Removed Petal Length)

```
The test error rate was 0.0
The training error rate was 0.058333333333333334
The Setosa test error rate was 0.0
The Versicolor test error rate was 0.0
The Virginica test error rate was 0.0
The Setosa training error rate was 0.0
The Versicolor training error rate was 0.075
The Virginica training error rate was 0.1
```

Based on the above results, we see that removing the sepal length/sepal width and petal length features one at time all result in 0% error rate for testing. While our training error increases to 5.8% in the case of the removal of petal length, the testing error is of more importance to us. We also see that only when we remove the petal width feature does our test error rate increases from 0% to 3.3%. So it is safe to say that sepal width, sepal length, and petal length are not important in classifying type.

This is reinforced by the results when we remove all three features and leave petal width as the only feature to use and we still get 0% error rate:

```
The test error rate was 0.0
The training error rate was 0.05
The Setosa test error rate was 0.0
The Versicolor test error rate was 0.0
The Virginica test error rate was 0.0
The Setosa training error rate was 0.0
The Versicolor training error rate was 0.05
The Virginica training error rate was 0.1
```

6. Assume the features are independent, i.e., ∑ is a diagonal matrix. Repeat 2 and 3, and report your results.

The following output is the result of enabling the indepFeatures boolean for my LDA classifier. Enabling the indepFeatures booleans forces the program in its calculateSigma() step to only include diagonal elements in its Sigma values.

```
The test error rate was 0.0
The training error rate was 0.05
The Setosa test error rate was 0.0
The Versicolor test error rate was 0.0
The Virginica test error rate was 0.0
The Setosa training error rate was 0.0
The Versicolor training error rate was 0.05
The Virginica training error rate was 0.1
```

As we can see for the above output, when we assume the features are independent for an LDA classifier, the training error rate for Virginica goes up.

When we do the same for an QDA classifier, we get the following output:

```
The test error rate was 0.0
The training error rate was 0.041666666666666664
The Setosa test error rate was 0.0
The Versicolor test error rate was 0.0
The Virginica test error rate was 0.0
The Setosa training error rate was 0.0
The Versicolor training error rate was 0.05
The Virginica training error rate was 0.075
```

As we can see, the overall training error rate goes up for the QDA cases as well. This is due to Virginica's error rate going up to 7.5%.