Assignment 2 Report

Yingnan Zhao id: 260563769

Question2

W0 and W1 is produced from the training data, first by calculating the possibility and mean for each class and the covariance matrix. Finally we get w0 and w1 fromthe equation below.

$$\underbrace{\log(p(y_0)) - \log(p(y_1)) - \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 + \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1}_{W_0} + \underbrace{\mathbf{x}^T\Sigma^{-1}(\mu_0 - \mu_1)}_{x^Tw_1} \overset{?}{>} 0$$

This is a linear decision boundary!

The accruacy, precision, recall and F1 measure of DS1_test are shown below.

```
parameter learnt
w0 = 2.01239190359
w1 = [[ 1.05511023]
 [-0.63276528]
 [-0.41290599]
 [-0.22683658]
 [-0.72655371]
 [-0.29721443]
 [ 1.24715225]
 [-1.78586254]
 [-2.14790584]
 [ 0.66975918]
 [-0.95714165]
 [-0.90663457]
 [ 1.15530344]
 [ 0.96745546]
 [-0.42961878]
 [ 0.97122599]
 [ 2.16524577]
 [-0.50713488]
 [-0.04769346]
 [-0.37327559]]
treating class 2 as positive
Accuracy = 0.955
Precision = 0.9595959595959596
Recall = 0.95
F1 Measure = 0.9547738693467336
```

As shown above, for DS1 the LDA model yields decent result.

Question 3

The distance between the test example and all the samples in the training dataset are calculated using the equation below.

$$d(x, x') = \sqrt{(x_1 - x'_1)^2 + (x_2 - x'_2)^2 + \ldots + (x_n - x'_n)^2}$$

Different K has been used to find the optimal k which has the highest F1 measure, the results are shown below.

```
k = 1
F1 Measure = 0.48124428179322964
k = 3
F1 Measure = 0.4918032786885245
k = 5
F1 Measure = 0.4873477038425492
k = 7
F1 Measure = 0.4735376044568246
k = 9
F1 Measure = 0.4626865671641791
k = 11
F1 Measure = 0.46760563380281694
k = 13
F1 Measure = 0.4619718309859155
k = 15
F1 Measure = 0.47457627118644063
k = 17
F1 Measure = 0.4711447492904447
k = 19
F1 Measure = 0.47826086956521735
treating class 2 as positive
the Best K is 3
Accuracy = 0.535
Precision = 0.5113636363636364
Recall = 0.47368421052631576
F1 Measure = 0.4918032786885245
```

The best number of neighbor is 3, it achieves an accuracy of 53.5%, precision of 51.1%, recall of 47.4% and F1 measure of 49.2% which is slightly better than chance, the classifier is considerably worse than the LDA classifier. And the F1 measure varies slightly when using different k value.

Question 5

The results below are produced by applying LDA classifier on the DS2_test produced from question4.

```
parameter learnt
w0 = 0.0303414354516
w1 = [[ 0.02985879]
 [-0.02545177]
 [ 0.00645555]
 [ 0.00826951]
 [-0.00987854]
 [ 0.00140088]
 [-0.00376674]
 [-0.02966947]
 [-0.02279957]
 [ 0.01878309]
 [ 0.01552167]
 [-0.0064863 ]
 [-0.04418118]
 [-0.010366  ]
 [ 0.01322158]
 [ 0.00740697]
 [ 0.0111111 ]
 [-0.00113356]
 [ 0.02137036]
 [-0.0070229 ]]
treating class 2 as positive
Accuracy = 0.52
Precision = 0.520066889632107
Recall = 0.5183333333333333
F1 Measure = 0.5191986644407345
```

As shown above the all of the measures of LDA droped dramatically on dataset 2 which is a mixture of three different Gaussians.

The results below are produced by applying KNN classifier on the DS2_test produced from question4 with different k value to find the optimal k.

```
k = 1
F1 Measure = 0.5154471544715448
k = 3
F1 Measure = 0.5070892410341952
k = 5
F1 Measure = 0.5063938618925832
k = 7
F1 Measure = 0.5134899912967799
k = 9
F1 Measure = 0.5127304653204565
k = 11
F1 Measure = 0.506896551724138
k = 13
F1 Measure = 0.49826989619377166
k = 15
F1 Measure = 0.5047701647875109
k = 17
F1 Measure = 0.49387040280210165
k = 19
F1 Measure = 0.5
treating class 2 as positive
the Best K is 1
Accuracy = 0.5033333333333333
Precision = 0.5
Recall = 0.5318791946308725
F1 Measure = 0.5154471544715448
```

As shown above, the optimal k is 1, it has 50.3% accruacy, 50% precision, 53.2% recall and 51.5% f1 measure. And it yields similar result as when the KNN classifier being applied to dataset 1.

Question 6

There is a huge drop in all measures when applying the LDA classifier to the dataset 2, because dataset 2 is a mixture of 3 gaussian distributions that has different means and different covariance matrix. Thus the dataset violates the key assumptions of LDA classifier, the data is not a gaussian, it is a gaussian mixture, and not all data in each class share the same covariance matrix. However, for the KNN classifier, it does not have any explicit assumptions about the dataset. Thus, the performance measures of the KNN classifier remain relatively stable.