Machine Learning Lab 3

Junming Zhang 29299527 jz1g17@ecs.soton.ac.uk

Problem 1 and 2

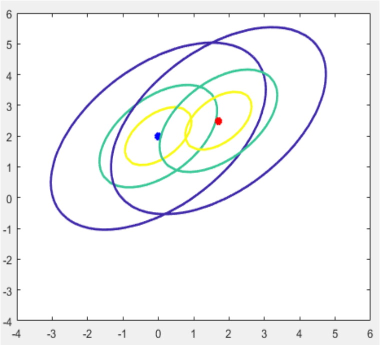


Figure 1: Two class patterns' classification

Figure 1 is the calss patterns of Gaussian distributed with $m1 = [0 \quad 2]^t$, $m2 = [1.7 \quad 2.5]^t$ and a common convariance matrix $\begin{bmatrix}2 & 1\\1 & 2\end{bmatrix}$. To draw this diagram, the density of each point in matrix will be calculated first, and then using points, density and different rate of maxmum denstity to plot the contour line with Gaussian distribution.

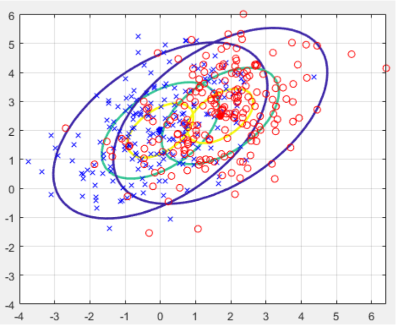


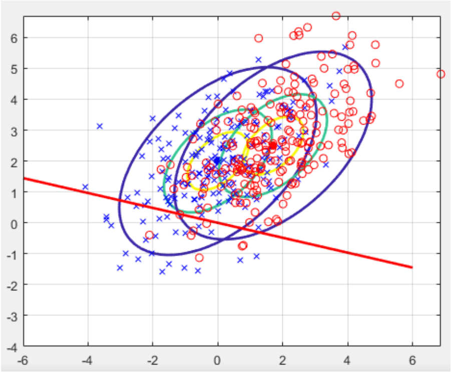
Figure 2: Samples and contour lines

In this question, 200 samples are obtained by using command:

```
mvnrnd(mean,convariance,number of samples)
```

Combine them to the Figure 1, it could find figure 1 is the contour line of these two Gaussian distibutions.

Problem 3 and 4



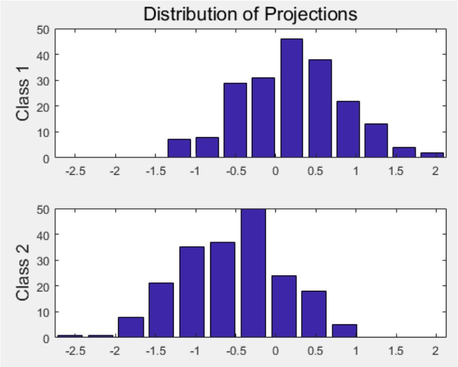


Figure 3: Fisher Linear Discriminant

Figure 4: Histograms of the distribution

In this case, Fisher Linear Discriminant use two tagged class with Gaussian distribution project to one dimentional vector which the discriminant direction is:

$$wF = \mathrm{inv}(\mathrm{C1} + \mathrm{C2}) * (\mathrm{m1} - \mathrm{m2}) = \begin{bmatrix} -0.4833 \\ -0.1167 \end{bmatrix}$$

After projection, two manifolds will have different feature. From Figure 4, it could find most projected data in Class 1 in the positive area. At meanwhile, the projected data in Class 2 most distributed in the negative area.

The fisher linear discriminant can solve the unlinear distribution question by projecting different data to one dimensional vector. Therefore, the question can be changed to linear question.

Problem 5,6 and 7

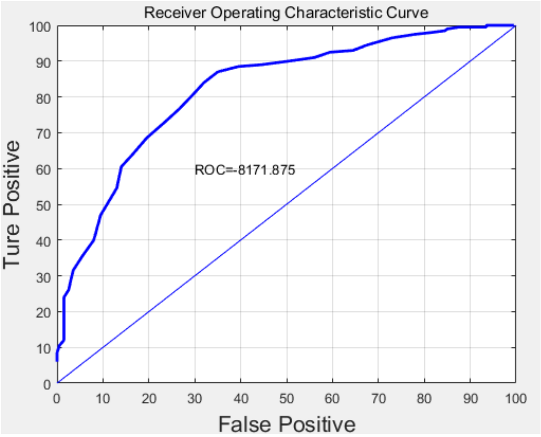


Figure 5: ROC and the area under curve

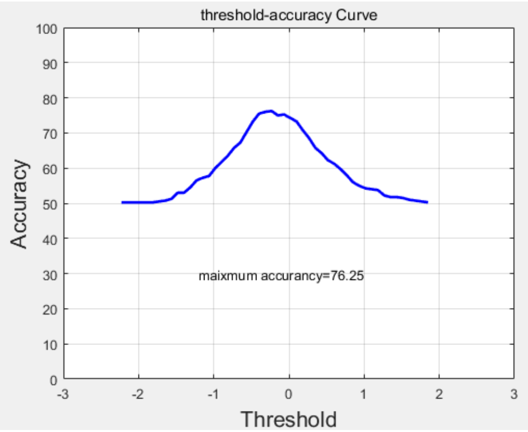


Figure 6: threshold-accruarcy curve and max accuracy

TP: True Positive FN: False Negative

FP: False Positive TN: True Negative

In problem 5, TP and TN will be used to draw the ROC curve. The probability is the key to judge those values. In this case, 50 thresholds will be used to find the points which their probability bigger than the TP and smaller than FP. Then the ROC curve can be plotted by them.

In problem 6, command *trapz()* is used to calculate the area under ROC curve, the result is shown as Figure 5 which is 8171.875(0.82).

In Problem 7, the TP and TN will be used to find the best accuracy and suitable threshold. In each iterate, find the points which their probability bigger than the TP and smaller than TN. Then the threshold-accuracy curve can be calculated. Using command max() to find the best accuracy and the corresponding threshold in this running is:

Threshold: -0.2307
Accuracy: 76.2500

25	-0.2307	76.2500

Problem 8

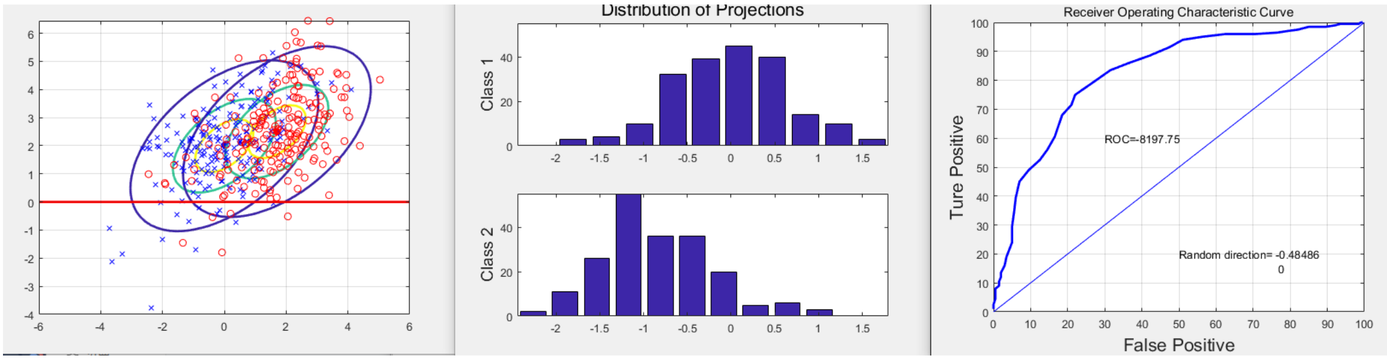


Figure 7: Random direction plotted diagram

Change the fisher discriminant coefficient to random number by using:

$$\mathrm{wF} = \mathrm{rand}(1)\&\mathrm{randi}([-1,1],2,1);$$

to generate a random 1×2 matrix between -1 to 1. Then apply it to script of problem 6. The random coefficient is shown in Figure 7 which is $\begin{bmatrix}-0.48486\\0\end{bmatrix}$ and the ROC also shows in Figure 7 which is 8197.75(0.82).

Problem 9

k-NN nearest neighbor classifier is a non-parametric method used for classification and regression. In this case, k=1, therefore the classification will be decided by the nearest point. The accuracy is 99.5025 which is higher than Fisher Discriminant Analyzer.

Problem 10

Euclidean distance:

$$L_2(x_i, x_j) = \left(\sum\nolimits_{l=1}^{n} \left|x_i^{(l)} - x_j^{(l)}\right|^2\right)^{\frac{1}{2}}$$

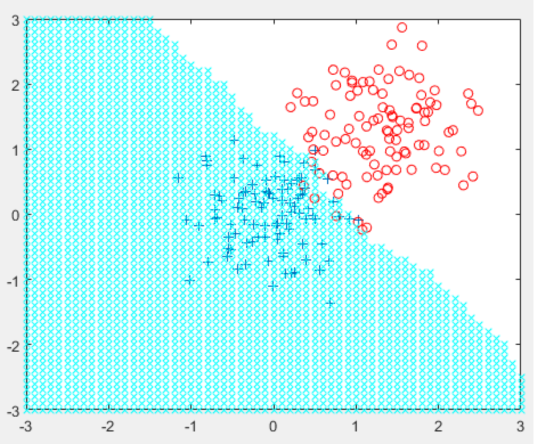


Figure 8: 11-NN Euclidean distance classfication

Accuracy = 0.955

Manhattan distance:

$$L_1(x_i, x_j) = \sum_{l=1}^{n} \left| x_i^{(l)} - x_j^{(l)} \right|$$

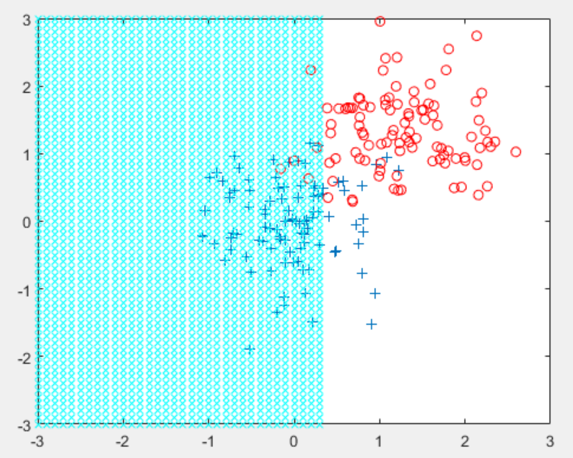


Figure 9: 11-NN Manhattan distance classfication

Accuracy = 0.88

It could find the Euclidean distance is more accuracy than the Manhattan distance.

Problem 11

Problem 12