Yixiao Zhang
31353282
yz3a19@soton.ac.uk

COMP6245 Lab 3

27th October 2019

1 Class boundaries and Posterior probabilities

Three two-class classification problems are generated by 3 group of parameters below. Each classification problem contains two Gaussian distribution and Each Gaussian distribution contains 200 data. The distributions are shown as Figure 1, Figure 2 and Figure 3. The contours on the posterior probability $P\left[\omega_1|x\right]$ are shown as Figure 4, Figure 5 and Figure 6.

Parameters group 1: $m1 = \begin{bmatrix} 0 \\ 3 \end{bmatrix}$, $m2 = \begin{bmatrix} 3 \\ 2.5 \end{bmatrix}$, $C1 = C2 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$, $P1 = P2 = 0.5$

Parameters group 2: $m1 = \begin{bmatrix} 0 \\ 3 \end{bmatrix}$, $m2 = \begin{bmatrix} 3 \\ 2.5 \end{bmatrix}$, $C1 = C2 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$, $P1 = 0.7$, $P2 = 0.3$

Parameters group 3: $m1 = \begin{bmatrix} 0 \\ 3 \end{bmatrix}$, $m2 = \begin{bmatrix} 3 \\ 2.5 \end{bmatrix}$, $C1 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$, $C2 = \begin{bmatrix} 1.5 & 0 \\ 0 & 1.5 \end{bmatrix}$, $P1 = P2 = 0.5$

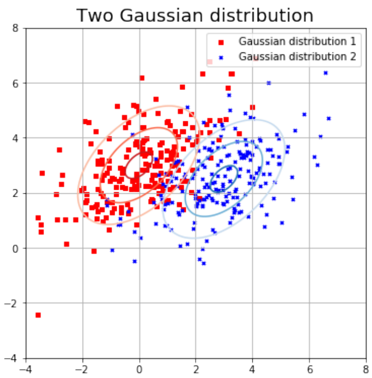

Figure 1: Distribution 1

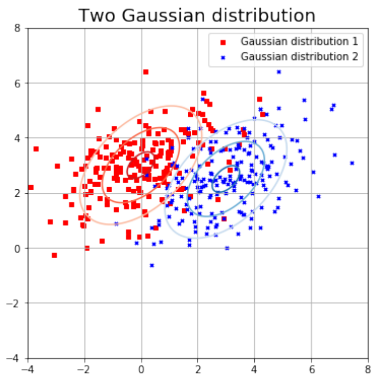

Figure 2: Distribution 2

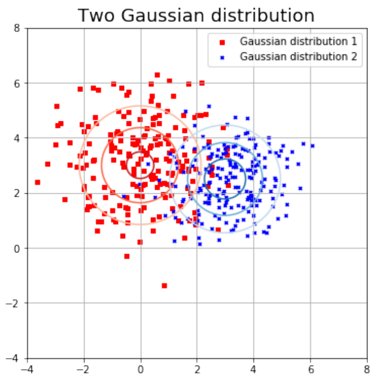

Figure 3: Distribution 3

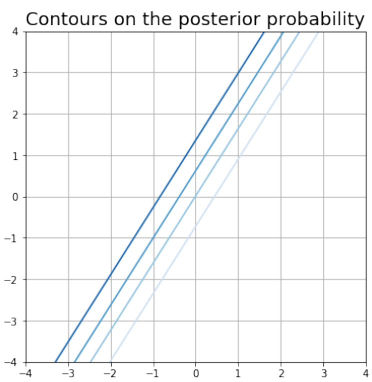

Figure 4: Posterior probability 1

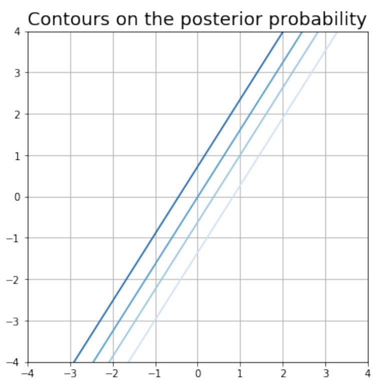

Figure 5: Posterior probability 2

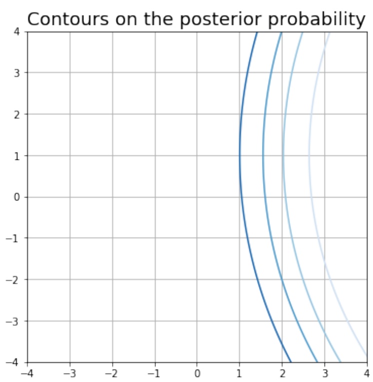

Figure 6: Posterior probability 3

From Figure 4, Figure 5 and Figure 6, the posterior probabilities are linear when covariance matrices are same, whereas the posterior probabilities are not linear when covariance matrices are not same. As well as the probabilities $P1$ and $P2$ do not influence whether the posterior probabilities are linear. In Figure 3, the distribution 1 is larger than distribution 2, so that the posterior probability become curly. In conclusion, the result of posterior probabilities are same as class boundaries.

2 Fisher LDA and ROC Curve

Two Gaussian distributions are generated by $m1 = \begin{bmatrix} 0 \\ 3 \end{bmatrix}$, $m2 = \begin{bmatrix} 3 \\ 2.5 \end{bmatrix}$, $C1 = C2 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ as Figure 7 shows. 200 samples from each distributions are plotted on Figure 7 as well. According to Fisher linear discriminant direction formula $w_f = (C_1 + C_2)^{-1}(m_1 - m_2)$, the discriminant direction is computed and drawn on Figure 7.

In order to classify the data using Fisher linear discriminant direction, these 400 samples are projected to the Fisher linear discriminant direction. These projections are put into 40 bins and shown as Figure 8. From Figure 7, the right(red) Gaussian distribution is mainly projected on the right, as well as the left(blue) Gaussian distribution is mainly projected on the left. Some of data which are at the middle of two distribution are overlap on the Figure 8 between -3 to 1.

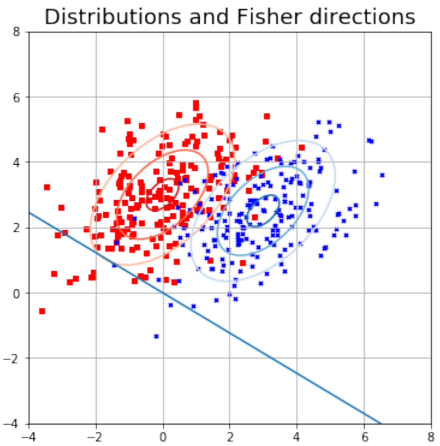

Figure 7: Distributions and Fisher Linear Discriminant direction

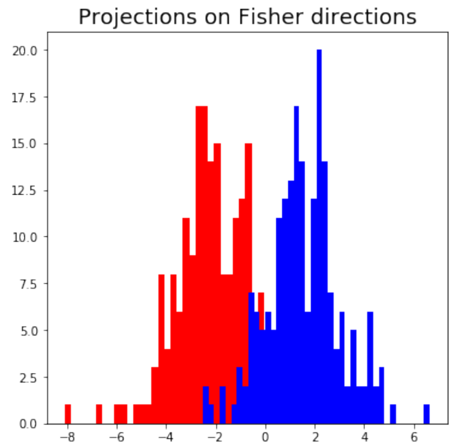

Figure 8: Histograms of distribution of projections

Figure 9 shows the Receiver Operating Characteristic (ROC) curve of the distribution in Figure 7 by sliding a decision threshold. At the left-bottom of Figure 7, every samples are judged as negative, so that both True-Positive and False-Positive are 0(0% of samples). At the right-top of Figure 7, every samples are judged as positive, so that both True-Positive and False-Positive are 100(50% of samples). When the True-Positive is as high as possible and False-Positive is as low as possible, the accuracy could be highest. The best point on Figure 8 is (89.0 17.0). The area under the ROC curve is 9348.75 which is an important value to evaluate the performance of a binary classifier.

Figure 10 shows the trend of accuracy increasing with threshold. When threshold smaller than -6 or greater than 6, every samples are classify to positive or negative, so that the accuracy is around 50%. When threshold at -0.276 which is around the middle of two distributions, accuracy achieves the peak at 86%.

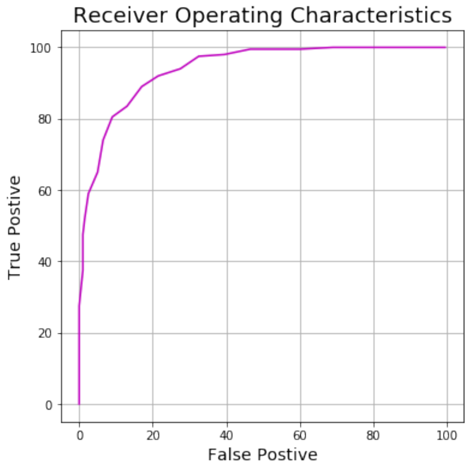

Figure 9: ROC curve

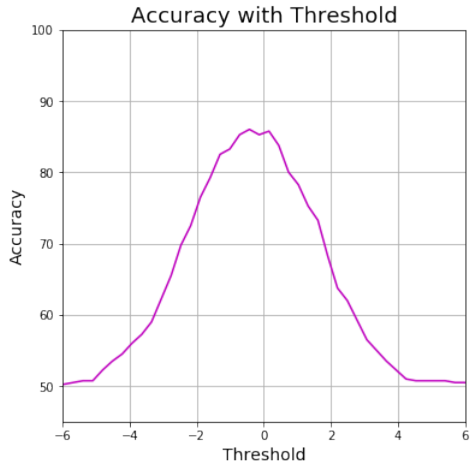

Figure 10: Classfication accuracy

Figure 11 shows 3 directions which are Fisher linear discriminant direction, a random direction, and direction connecting the means of two distributions. The ROC curves of these 3 directions are shown as Figure 12. The classification accuracy are shown as Figure 13. As different direction has different length of threshold, thresholds in Figure 13 have been nomorlised. From both Figure 12 and Figure 13, Fisher linear discriminant direction achieves 86% accuracy which is better than direction connecting the means of two distributions at around 84% accuracy and much better than random direction at around 69% accuracy.

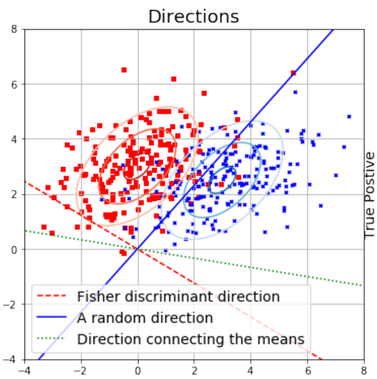

Figure 11: ROC curve

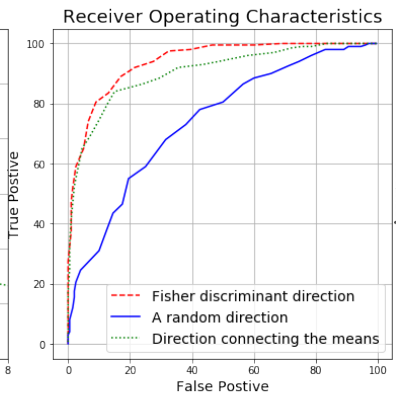

Figure 12: ROC curve

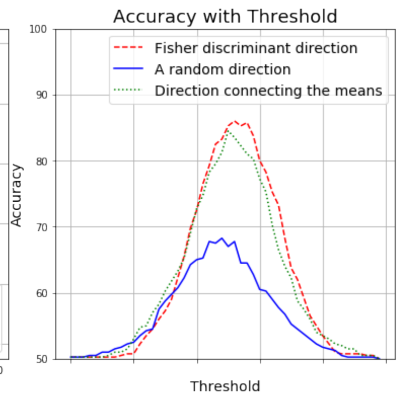

Figure 13: Classfication accuracy

3 Mahalanobis Distance

Mahalanobis distance is another distance than Euler distance which is a measure of the distance between a point and a distribution. It can be computed by the formula: $D_m(\vec{x}) = \sqrt{(\vec{x}-\vec{m})^T C^{-1}(\vec{x}-\vec{m})}$.

In most of situation, Mahalanobis distance could perform better than Euler distance for a classifier. But in some case, these two distance may has same result or the Euler distance better than Mahalanobis distance.

Figure 14 and Figure 18 show two classification results using Mahalanobis distance classifier and Euler distance classifier. These two classifier show same result at 93.25% accuracy.

In the distribution as Figure 17 and Figure 16 show, Mahalanobis distance classifier at 61.25% accuracy is not as good as Euler distance classifier which at 69% accuracy.

In case 3 as Figure 15 and Figure 19 show, Mahalanobis distance classifier at 58.25% accuracy is better than Euler distance classifier which at 55.25% accuracy.

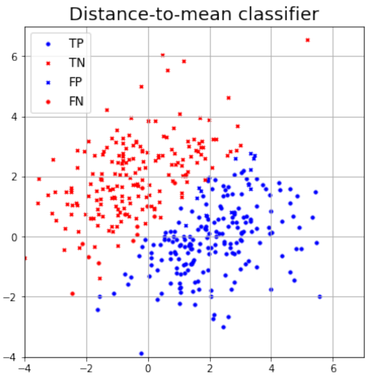

Figure 14: Case 1

(Euler distance 93.25%)

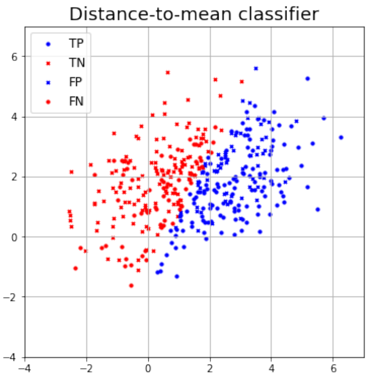

Figure 15: Case 2

(Euler distance 69%)

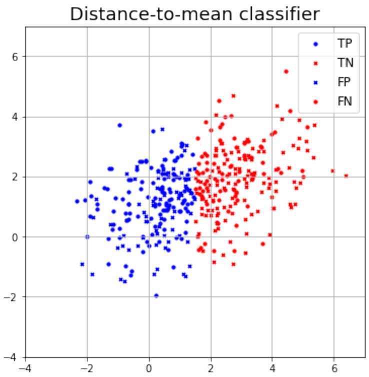

Figure 16: Case 3

(Euler distance 55.25%)

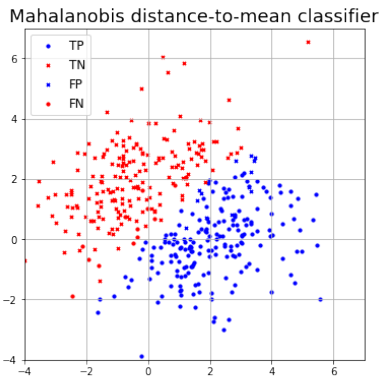

Figure 17: Case 1

(Mahalanobis distance 93.25%)

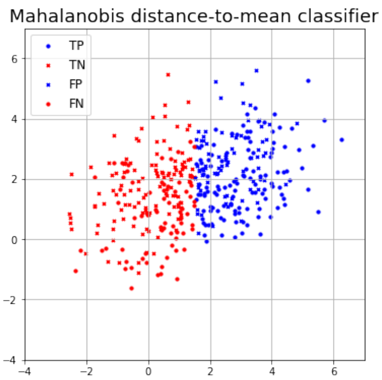

Figure 18: Case 2

(Mahalanobis distance 61.25%)

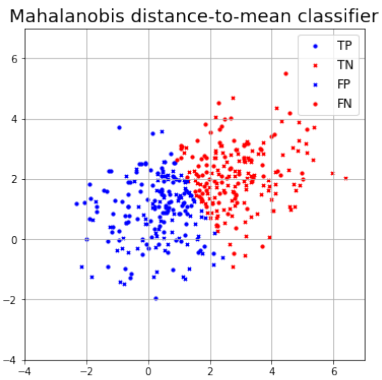

Figure 19: Case 3

(Mahalanobis distance 58.25%)