Yang Liu
3323643

Lab 3

COMP6245
yl58n21@soton.ac.uk

Class Boundaries and Posterior Probabilities

The data in each group of these two-class classification problems are generated by two sets of Gaussian distributions which contain 300 data. The distributions, contours on posterior probabilities of each group are shown on **Figure 1**, **Figure 2**, **Figure 3** respectively.
In fact, the contours on posterior probabilities are linear if each covariance matrices of Gaussian distributions are the same while they are curly if the covariance matrices are different. And the probabilities of the distribution have nothing to do with the shape of the contour.
Parameters of group 1: $C1 = C2 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$, $m1 = \begin{bmatrix} 0 & 3 \end{bmatrix}$, $m2 = \begin{bmatrix} 3 & 2.5 \end{bmatrix}$, $P1 = P2 = 0.5$
Parameters of group 2: $C1 = C2 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$, $m1 = \begin{bmatrix} 0 & 3 \end{bmatrix}$, $m2 = \begin{bmatrix} 3 & 2.5 \end{bmatrix}$, $P1 = 0.7, P2 = 0.3$
Parameters of group 3: $C1 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$, $C2 = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$, $m1 = \begin{bmatrix} 0 & 3 \end{bmatrix}$, $m2 = \begin{bmatrix} 3 & 2.5 \end{bmatrix}$, $P1 = P2 = 0.5$

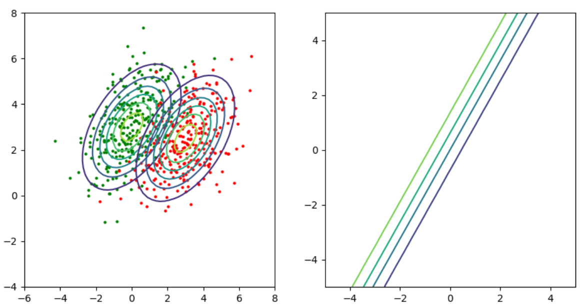

Figure 1: Distribution and contours 1

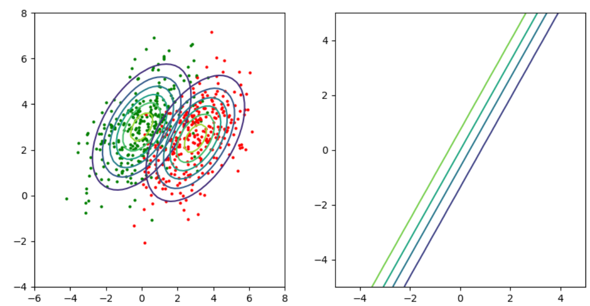

Figure 2: Distribution and contours 2

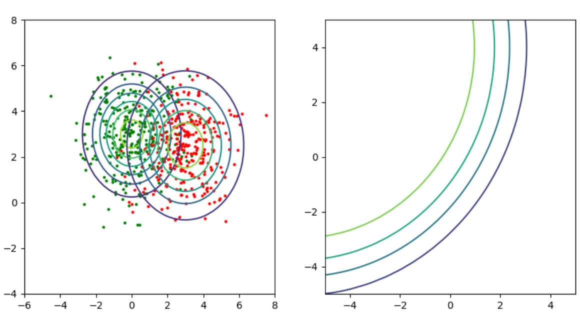

Figure 3: Distribution and contours 3

Fisher LDA and ROC Curve

The two classes that needed to be classified are generated by two Gaussian distributions with means $m1 = \begin{bmatrix} 0 & 3 \end{bmatrix}$ and $m1 = \begin{bmatrix} 3 & 2.5 \end{bmatrix}$ respectively while have the same covariance matrix $\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$.

By using $W_F = (c1+c2)^{-1}(m1-m2)$, we could plot the discriminant direction. **Figure 4** shows 300 samples each generated by these two Gaussian distributions and the discriminant direction. The total 600 samples are projected to the discriminant direction and put into 40 bins in a histogram as **Figure 5** shows.

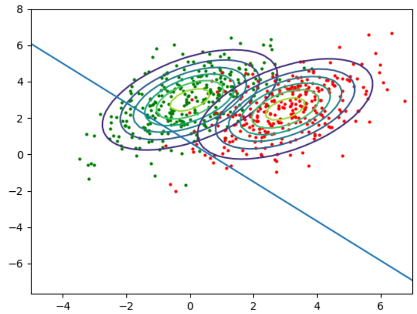

Figure 4: Distribution and Fisher discriminant direction

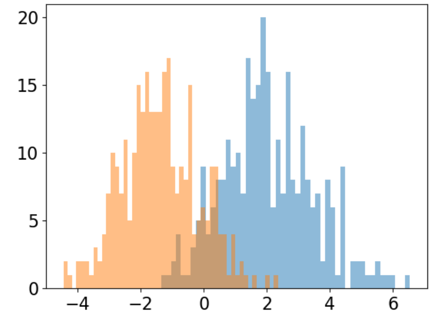

Figure 5: Distribution of projected distribution

The **Figure 6** shows ROC curve which could be drawn by using True Positive and False Positive rates of the decision thresholds. The AUC(area under curve) of this ROC curve is 0.97 which indicates the performance of this binary classifier.
The classification accuracy varies depending on the decision threshold. The accuracy could be computed by $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$. **Figure 7** shows the accuracy of each threshold. The Accuracy reaches its peak at 93.2 while the threshold is 24.

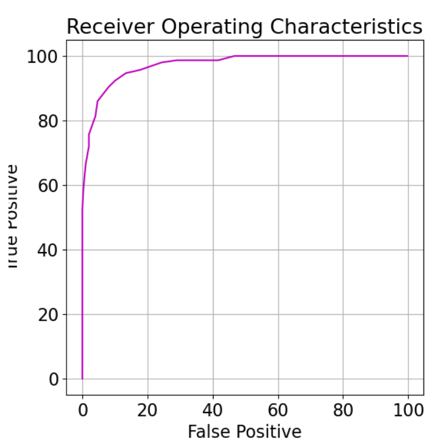

Figure 6: ROC curve

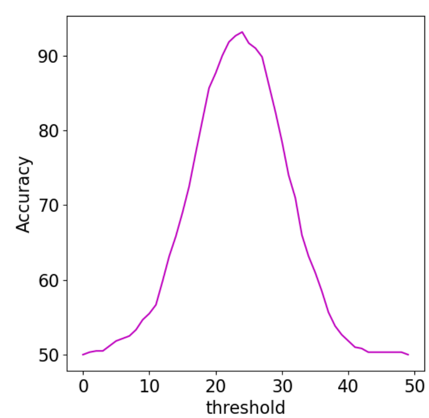

Figure 7: Accuracy of each threshold

Figure 8 shows random direction(W_F is randomly generated) and **Figure 9** shows the ROC curve for the random direction. The AUC of this ROC curve is 0.81.
Figure 10 shows direction connecting the means of the two classes and **Figure 11** shows the ROC curve for this direction. The AUC of this ROC curve is 0.90.
AUC(Area Under ROC Curve) is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. It is used to assess the performance of a model, a higher AUC generally means a better model and better performance.

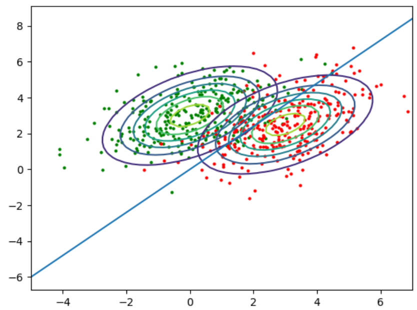

Figure 8: Distribution and random direction

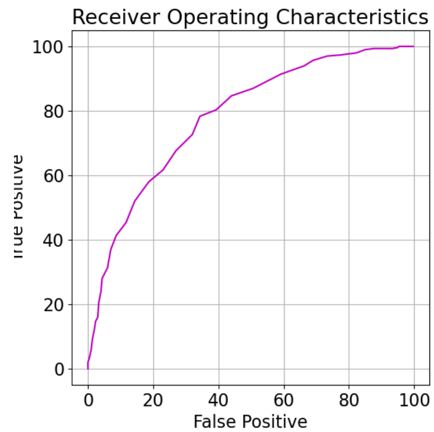

Figure 9: ROC curve

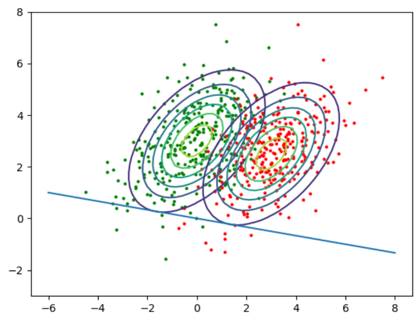

Figure 10: Distribution and means connection direction

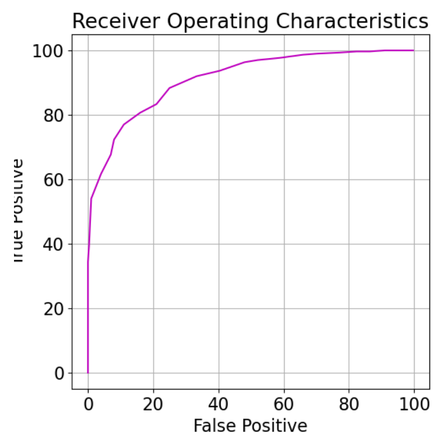

Figure 11: ROC curve

Mahalanobis Distance

Mahalanobis Distance is a measure to compute the distance between a point and a distribution. Unlike Euclidean Distance, Mahalanobis Distance takes the sacle of each dimension into account which could be seen in the formula $D_m(\vec{x}) = \sqrt{(\vec{x} - \vec{m})^T C^{-1} (\vec{x} - \vec{m})}$ where C is covariance matrix and m is mean. When the covariance matrix is a unit matrix, which means the scales of each dimension are same, the Euclidean Distance and Mahalanobis Distance are the same, but if it's not, then Mahalanobis Distance would always be better than Euclidean Distance. As for the distributions we discussed above, use the Mahalanobis Distance is better compared with Euclidean Distance since the covariance matrix is not a unit matrix.