Foundations of Machine Learning - Lab 3 Report

Johnny Joyce (jvjj1u19@soton.ac.uk)

October 29, 2019

1 Class boundaries and posterior probabilities of Gaussian classifiers

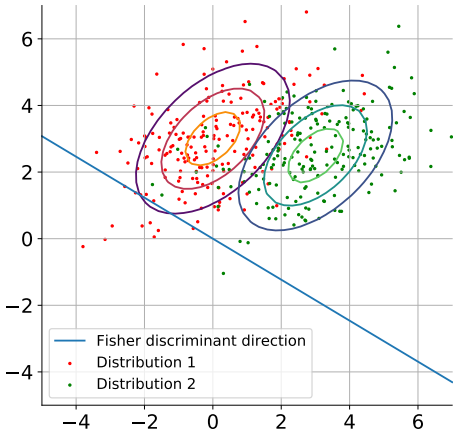


Figure 1: Samples from two bivariate Gaussian distributions and corresponding contours.

This lab utilises two bivariate Gaussian distributions (let us call them distributions 1 and 2), each with 200 respective samples. Figure 1 shows the samples from distributions 1 and 2 along with contour plots of their corresponding distributions. It also shows the direction of the Fisher discriminant (henceforth $\boldsymbol{w}_F$), which maximises the separation between means of distributions 1 and 2 and minimises the within-class separation of each distribution.

2 Fisher linear discriminant analysis

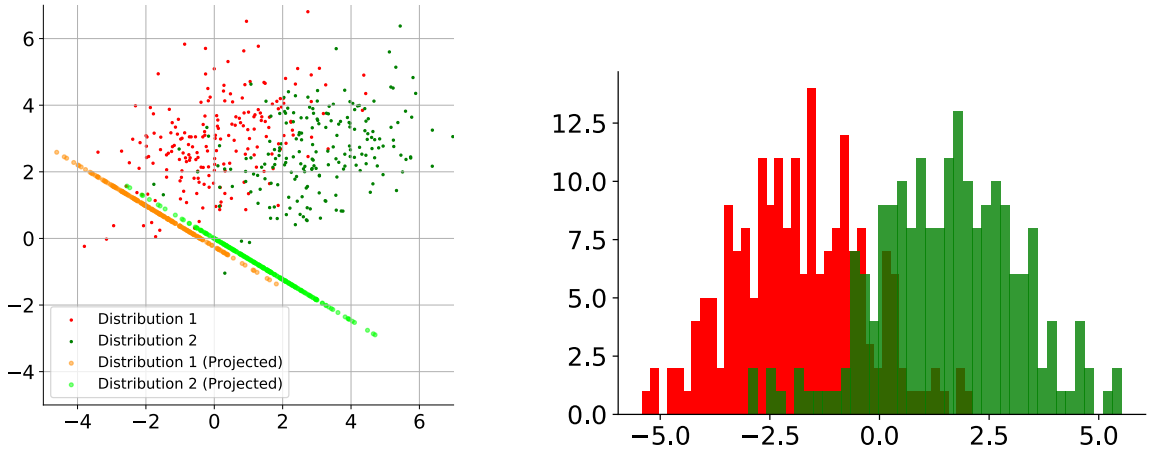
We can obtain the vector projection of any coordinate $\boldsymbol{x}$ onto $\boldsymbol{w}_F$ by calculating

$$\frac{\boldsymbol{x} \cdot \boldsymbol{w}_F}{\|\boldsymbol{w}_F\|^2} \boldsymbol{w}_F.$$

which results in the data shown in Figure 2 (a). We can also obtain the scalar projection in the direction of $\boldsymbol{w}_F$ by multiplying each vector $\boldsymbol{x}$ by $\boldsymbol{w}_F$, resulting in the histogram shown in Figure 2 (b), allowing us to see the frequency of projected values. It is interesting to note that this results in a univariate Gaussian distribution as it is the one-dimensional projection of a multivariate Gaussian distribution.

3 Receiver operating characteristic curve

Given our discriminant, we can find a receiver operating characteristic (ROC) curve. We set a threshold at the minimum of our projected values and find the number of true positives and false positives by comparing each projected value to our threshold. We gradually increase the threshold and repeat until we have a set of coordinates representing the number of true/false positives for certain thresholds. Figure 3 shows the ROC curve in the direction of $\boldsymbol{w}_F$ (in purple) and in a random direction (in red).



(a) (b)

Figure 2: (a) Vector projections of data from our distribution onto $\boldsymbol{w}_F$ (Note that the projections for each distribution are slightly separated for clarity) (b) Histogram of the scalar projections onto $\boldsymbol{w}_F$.

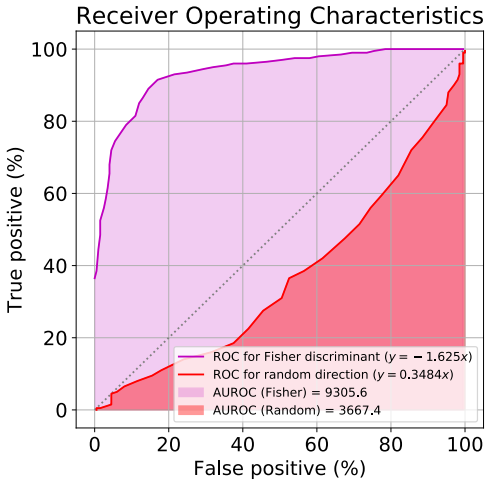


Figure 3: Plots of the receiver operating characteristics in the direction of the Fisher discriminant, as well as in a random direction for comparison.

The area under the ROC curve (AUROC), which is also shown in Figure 3 gives a measure of our model's ability to differentiate between the two distributions. Since each axis is on a scale from 0 to 100, the maximum possible value of the AUROC is $100^2 = 10,000$. This would imply that the model could always perfectly identify each coordinate without any misclassifications.

The ROC curve in the direction of $\boldsymbol{w}_F$ gave an AUROC value of 9305.6, which was very close to the maximum possible value. When projecting in the direction of the difference of the means (i.e. the vector from $\boldsymbol{\mu}_1$ to $\boldsymbol{\mu}_2$), we obtain an AUROC value of 9129.6. This is relatively high since the vector from $\boldsymbol{\mu}_1$ to $\boldsymbol{\mu}_2$ has a similar gradient to the vector which minimises the between-class variance of each distribution.

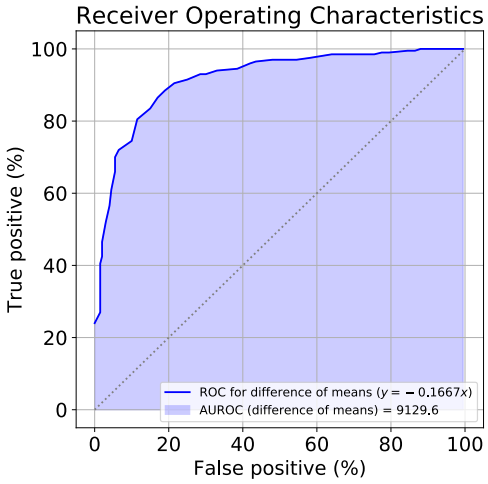
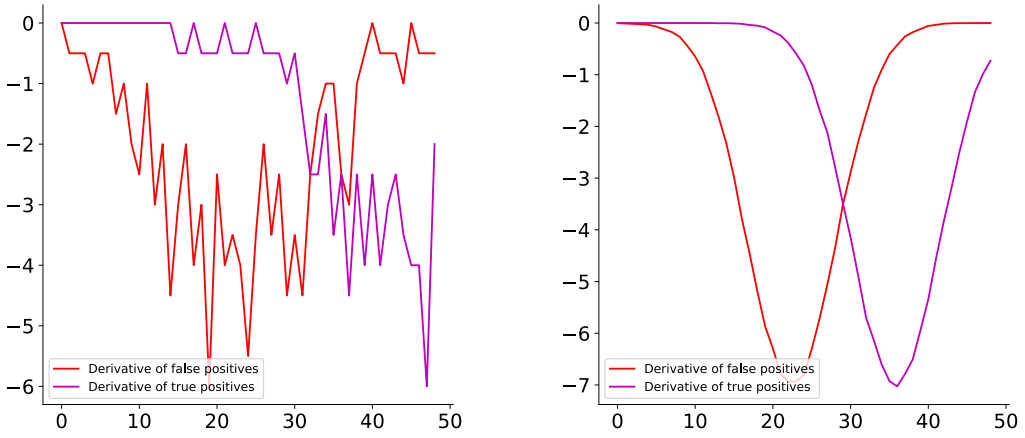


Figure 4: ROC curve obtained by projecting in the direction $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$.

It is also interesting to note that by differentiating the number of true positives and false positives, we obtain univariate Gaussian distributions, as shown in Figure 5 (a). This becomes more apparent when we increase the number of samples from each distribution from 200 to 200,000, as shown in Figure 5 (b). This result is to be expected since the number of true positives and the number of false positives each act as a cumulative probability density function. Since we obtained our data from Gaussian distributions, as our ROC threshold approaches the means of the data, the numbers of true positives and false positives increases according to a Gaussian distribution, resulting in the graphs we see.



(a) (b)

Figure 5: Derivatives of the numbers of true positives and false positives with (a) 200 samples (b) 200,000 samples

4 Mahalanobis distance

We have so far been classifying coordinates based on their Euclidean distance from the means of the respective distributions. That is, our metric has been the L^2 distance. We could instead use the Mahalanobis distance, defined by:

$$D_m(\boldsymbol{x}) \coloneqq \sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^\intercal \boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{C}$ are the mean and covariance matrix of the distribution in question, respectively. We would then classify a coordinate $\boldsymbol{x}$ based on its Mahalanobis distance from each mean and choosing the distribution that minimises the distance.

A Mahalanobis classifier accounts for the fact that the variance of the data differs depending on the direction from which we are looking at it. This results in the ROC curve seen in Figure 6, which has a corresponding AUROC value of 8748.5.

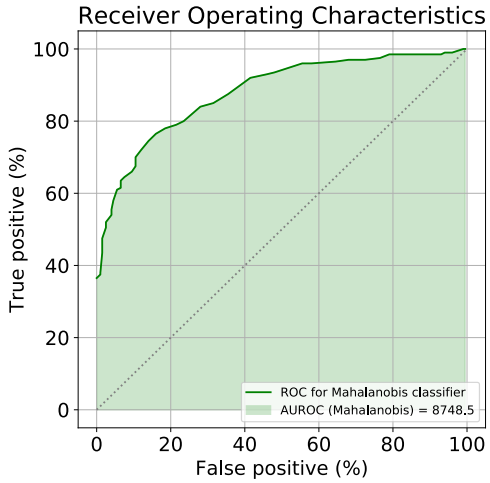


Figure 6: ROC curve using a Mahalanobis classifier.