LDA vs. PCA

Linear discriminant analysis is very similar to PCA both look for linear combinations of the features which best explain the data.

The main difference is that the Linear discriminant analysis is a supervised dimensionality reduction technique that also achieves classification of the data simultaneously.

LDA focuses on finding a feature subspace that **maximizes the separability** between the groups.

While Principal component analysis is an **unsupervised** Dimensionality reduction technique, it ignores the class label.

PCA focuses on capturing the direction of **maximum variation** in the data set.

LDA and PCA both form a new set of components.

The **PC1** the first principal component formed by PCA will account for maximum variation in the data.**PC2** does the second-best job in capturing maximum variation and so on.

The **LD1** the first new axes created by Linear Discriminant Analysis will account for capturing most variation between the groups or categories and then comes LD2 and so on.

Note In, LDA The target dependent variable can have binary or multiclass labels.

Working of Linear Discriminant Analysis

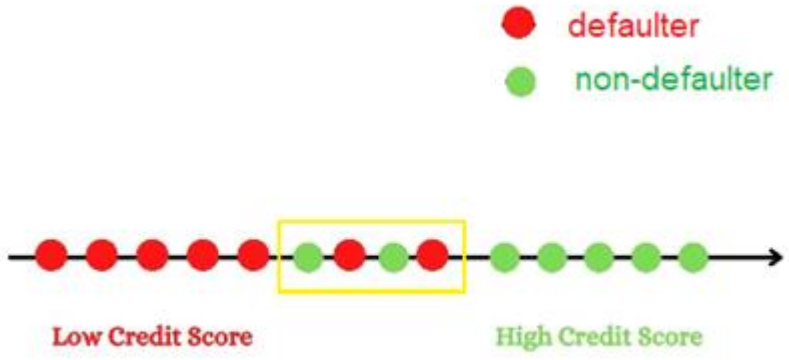
Working of the Linear Discriminant Analysis with the help of an example.

Imagine you have a credit card loan dataset with a target label consisting of two classes defaulter and non-defaulter.

Class‘ 1’ is the **defaulter** and **class ‘0’** is the **non-defaulter**.

Understanding a basic 1-D and 2-D graph before proceeding on to LDA projection

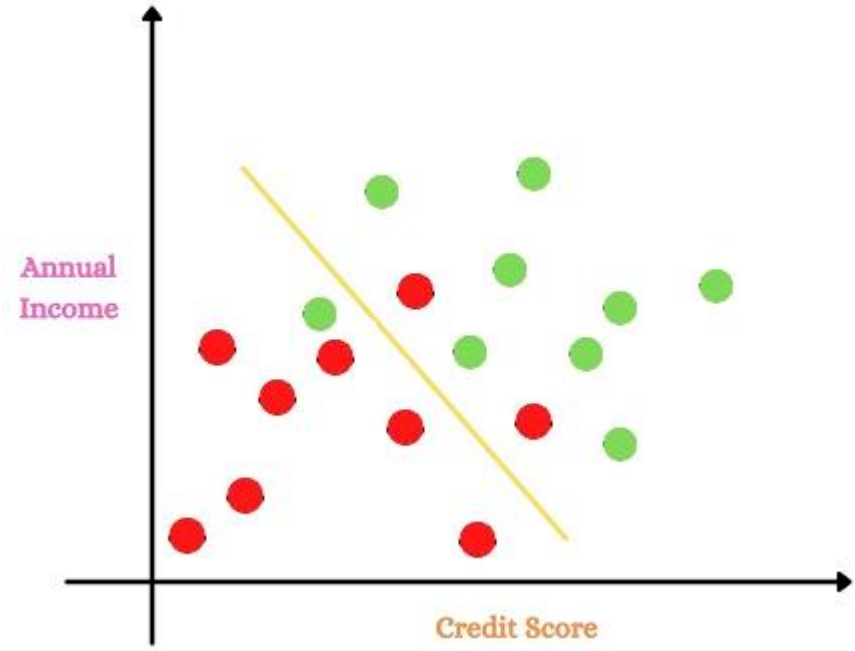
When you have just **one attribute** say the **credit score** the graph would be a 1-D graph which is a number line.



With just one attribute, although we were able to separate the categories certain points have been **over-lapped** due to no specific cut-off point. In reality, there could be many such overlapped data points when there is a large number of observations in the data set.

Let’s see what happens when we have two attributes, are we able to classify better?

Considering another attribute annual income along with the credit score.



Adding another feature, we were able to reduce the overlapping than the earlier case with a single attribute. But when we work on large data sets with many features and observations, there would still be a lot many overlapped points left.

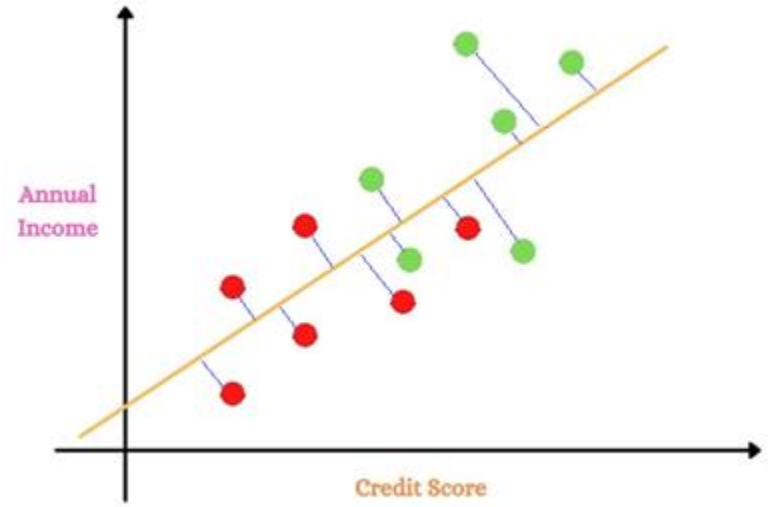
Something interesting, we noticed is adding features we were able to reduce the number of overlapped points and distinguish better. But this becomes extremely difficult to visualize when we have high dimension dataset.

This is where the implementation of LDA helps.

How does LDA project the data?

Linear Discriminant Analysis projects the data points onto new axes such that these new components **maximize** the separability among categories while keeping the variation within each of the categories at a **minimum** value.

Let us now understand in detail how LDA projects the data points.



Source: Image by author

1.LDA uses information from both the attributes and projects the data onto the new axes.

2.It projects the data points in such a way that it satisfies the criteria of maximum separation between groups and minimum variation within groups simultaneously.

Step 1:

The projected points and the new axes

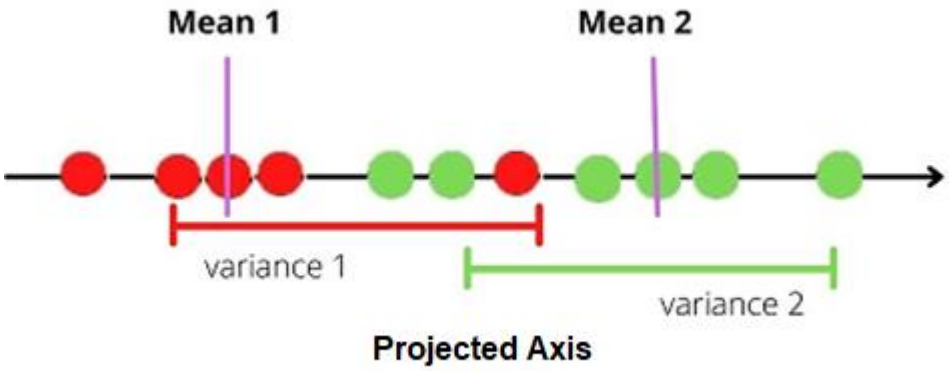
Source: Image by Author

Step-2

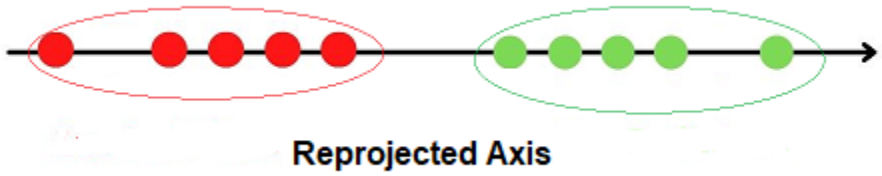
Criterion LDA applies to the projected points is as follows.

1.It maximizes the distance between the means of each category.

2. It minimizes the variation or scatter within each category represented by s^2



Source: Image



Let the mean of the category 1 defaulter be mean 1 and mean 2 be the mean of the category non-defaulter.

Similarly,

$S1^2$ be the scatter of the first category and

$S2^2$ be the scatter of the second category.

It now calculates the formula.

$(mean1-mean2)^2/(S1^2+S2^2)$

Note: Numerator is squared to avoid negative values.

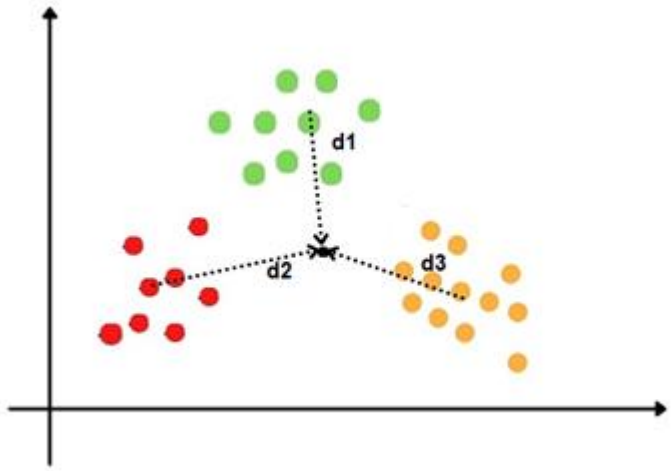
Let mean1-mean2 be represented by d.

The formula will now be.

$d^2/(S1^2+S2^2)$

Note: Ideally**,** the Larger, the numerator greater, the separation between groups. While smaller the denominator less variance within the groups.

How does LDA work when there are more than two categories?



Source: Image by Author

When there are more than two categories, LDA calculates the central point of all the categories and the distance between the central points of each category to that point.

It then projects the data onto the new axes in such a way that there is a maximum separation between the groups and minimum variation within the groups.

The formula will now be.

$(d1^2+d2^2+d3^2)/(s1^2+S2^2+S3^2)$

and now the curious part…

How does LDA make predictions?

Linear Discriminant Analysis uses Baye's theorem to estimate the probabilities.

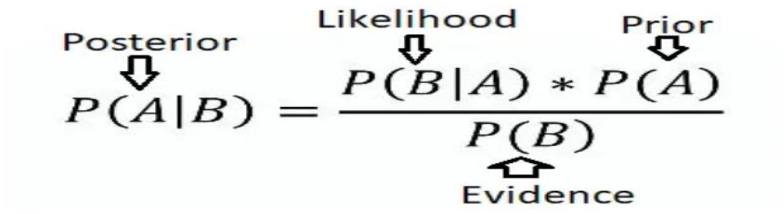
It first calculates the prior probabilities from the given data set. With the help of these prior probabilities, it calculates the posterior probabilities using Baye's theorem.

From Bayes Theorem we know that.

P(A|B)=P(B|A)*P(A)/P(B)

where A, B are events and P(B) is not equal to zero

Assume we have three classes class 0, class 1, class 2 in the data set.



Naive Bayes theorem

step 1

LDA calculates the **prior** probabilities of each of the classes P(y=0), P(y=1), P(y=2) of the data set.

Next, it calculates the conditional probabilities of the observations.

step 2

let us consider an observation x.

P(x|y=0),P(x|y=1),P(x|y=2) represent the **likelihood functions**.

step 3

LDA now calculates the **Posterior probabilities** to make predictions**.**

P(y=0|x)=P(x|y=0)*P(y=0)/P(x)

$P(y=1|x)=P(x|y=1)*P(y=1)/P(x)$

$P(y=2|x)=P(x|y=2)*P(y=2)/P(x)$

The **general equation** for a set of 'c**'** classes

Let y1, y2,.., yc be the set of 'c' classes and consider i=1,2,..,c.

$P(x|yi)$ will represent the **likelihood** function or the **conditional** probability.

$P(yi)$ will be the **prior** probability of each class in the data set. Which is nothing but the ratio of the number of observations in that particular class to the total number of observations in all the classes.

The posterior probability will be **Likelihood*Prior/Evidence.**

$P(y=yi|x)=P(x|y=yi)*P(yi)/P(x)$

Assumptions of Linear Discriminant Analysis

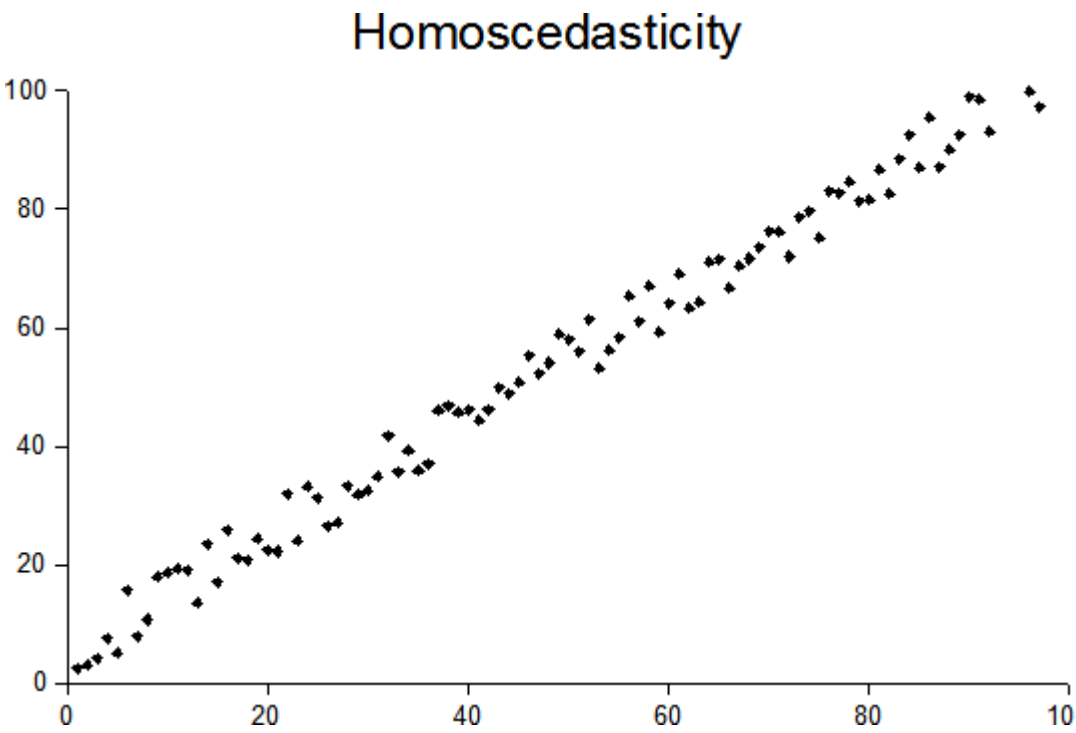
There are certain assumptions Linear Discriminant Analysis makes on the data set they are.

Assumption 1

LDA assumes that the independent variables are normally distributed for each of the categories.

Assumption 2

Homoscedascity



Plot with data showing Homogeneity of variance i.e. for each value of x the value of y has the same variance. Source: Creative Commons by Wikimedia

LDA assumes the independent variables have equal variances and covariances across all the categories. This can be tested with Box's M statistic.

In simple words, The assumption states that the variance across all variables for the categories in the data set say when y=0, y=1 has to be equal, and also the covariance between the variables has to be equal.

This assumption helps the **Linear Discriminant Analysis** to create the **linear decision boundary** between the categories.

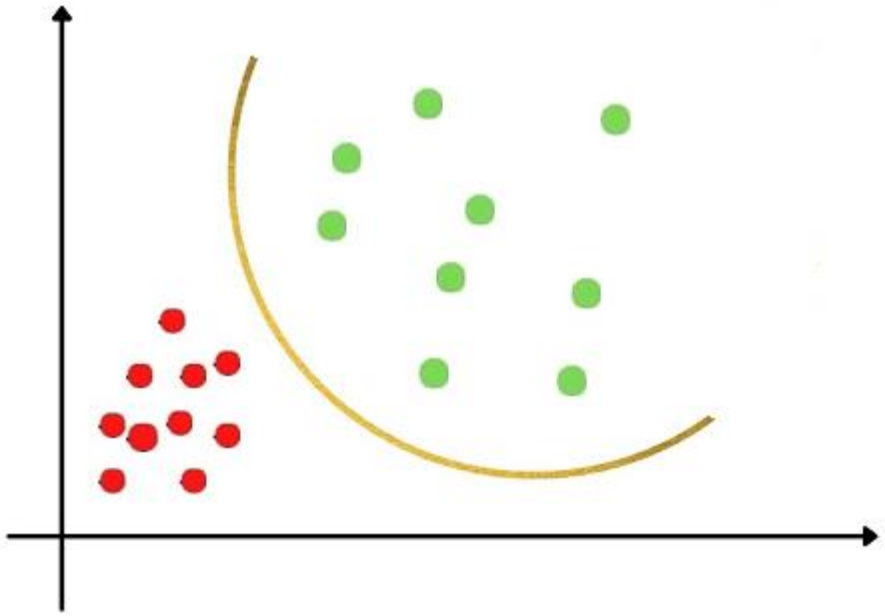
Assumption 3

Multicollinearity

The performance of prediction can decrease with the increased correlation between the independent variables.

Note: Studies show that LDA is robust to slight violations of these assumptions.

What happens when the second assumption fails?



Source: Image by Author

When this assumption fails, There is huge variation between the classes in the data set and the other variant of Discriminant analysis is used which is the **Quadratic Discriminant Analysis** (**QDA).**

In Quadratic Discriminant Analysis, The mathematical function which separates the categories will now be quadratic and not linear to achieve classification.

Applications of LDA

1. LDA is most widely used in pattern recognition tasks. For example, analyzing customer behavior patterns based on attributes.
2. Image recognition, LDA can distinguish between categories. For example Faces and not faces, objects and not objects.
3. In the medical field, To classify patients into different groups based on symptoms for a particular disease.

2. The Concept of Linear and Quadratic Discriminant Methods

The probability density function for multivariate normal distribution, or $x \sim N(\mu, \sum)$, is written as:

$$f(x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^p |\sum|}} \exp\left(-\frac{(x-\mu)^T \sum^{-1} (x-\mu)}{2}\right), \quad (1)$$

where $x = (x_1, x_2, ..., x_p)$ is the independent variable, $\mu = (\mu_1, \mu_2, ..., \mu_p)$ is the mean of the independent variables, and Σ is the covariance matrix.

2.1. Linear Doscriminant Analysis (LDA)

LDA [19] assumes that the binary classification has the equal covariance matrix $\Sigma_1 = \Sigma_2 = \Sigma$. Therefore, the equation (1) becomes:

$$\begin{aligned} &\frac{1}{\sqrt{(2\pi)^p |\sum|}} \exp\left(-\frac{(x-\mu_1)^T \sum^{-1} (x-\mu_1)}{2}\right) \pi_1 \\ &= \frac{1}{\sqrt{(2\pi)^p |\sum|}} \exp\left(-\frac{(x-\mu_2)^T \sum^{-1} (x-\mu_2)}{2}\right) \pi_2, \\ &\exp\left(-\frac{(x-\mu_1)^T \sum^{-1} (x-\mu_1)}{2}\right) \pi_1 \\ &= \exp\left(-\frac{(x-\mu_2)^T \sum^{-1} (x-\mu_2)}{2}\right) \pi_2, \end{aligned} \quad (2)$$

where π_1 and π_2 are the prior probability of the two classes and μ_1 and μ_2 are the mean of the two classes.

Using the natural logarithm from two sides of the equation, we get the simplified term:

$$\begin{aligned} &-\frac{1}{2}(x-\mu_1)^T\sum^{-1}(x-\mu_1)+\ln(\pi_1)=-\frac{1}{2}(x-\mu_2)^T\sum^{-1}(x-\mu_2)\\ &\quad+\ln(\pi_2),\\ &-\frac{1}{2}x^T\sum^{-1}x-\frac{1}{2}\mu_1^T\sum^{-1}\mu_1+\mu_1^T\sum^{-1}x+\ln(\pi_1)\\ &=-\frac{1}{2}x^T\sum^{-1}x-\frac{1}{2}\mu_2^T\sum^{-1}\mu_2+\mu_2^T\sum^{-1}x+\ln(\pi_2), \end{aligned} \tag{3}$$

where equation (3) is $x^T\sum^{-1}\mu_1=\mu_1^T\sum^{-1}x$, and when each side is multiplied by 2, we get:

$$2\left(\sum^{-1}(\mu_2-\mu_1)\right)^T x+\left((\mu_2-\mu_1)^T\sum^{-1}(\mu_2-\mu_1)\right)+2\ln\left(\frac{\pi_2}{\pi_1}\right)=0. \tag{4}$$

Equation (4) shows the form of a linear term. LDA discriminant function the two classes as:

$$\begin{aligned} \delta(x)=2\left(\sum^{-1}(\mu_2-\mu_1)\right)^T x+\left((\mu_2-\mu_1)^T\sum^{-1}(\mu_2-\mu_1)\right)\\ +2\ \ln\left(\frac{\pi_2}{\pi_1}\right)=0. \end{aligned} \tag{5}$$

The class of instance x is defined by:

$$\widehat{c}(x)=\left\{\begin{cases}1, \text{if}\,\delta(x)<0,\\ 2, \text{if}\,\delta(x)>0.\end{cases}\right) \tag{6}$$

The LDA Algorithm.

(1) Input independent variables $x = (x_1, x_2, ..., x_p)$ of n sample.

(2) Compute the mean of each class μ_1 and μ_2.

(3) Calculate the prior probability of each class π_1 and π_2.

(4) Compute the covariance matrix for each class Σ_1 and Σ_2.

(5) Approximate the pooled covariance matrix as $\widehat{\Sigma} = (n_1 - 1)\widehat{\sum}_1 + (n_2 - 1)\widehat{\sum}_2 / n - 2, , n = n_1 + n_2$.

(6) Calculate the LDA discriminant function as in equation (5).

(7) Assign the class label as in as in equation (6).

2.2. Quadratic Doscriminant Analysis (QDA)

QDA [19] for binary classification is defined as the unequal covariance matrix $\Sigma_1 \neq \Sigma_2$. Therefore, equation (2) can be adjusted in term of unequal covariance as:

$$\begin{aligned} &\frac{1}{\sqrt{(2\pi)^p \left|\sum_1\right|}} \exp\left(-\frac{(x-\mu_1)^T \sum_1^{-1} (x-\mu_1)}{2}\right)\pi_1 \\ &= \frac{1}{\sqrt{(2\pi)^p \left|\sum_2\right|}} \exp\left(-\frac{(x-\mu_2)^T \sum_2^{-1} (x-\mu_2)}{2}\right)\pi_2. \end{aligned} \tag{7}$$

The natural logarithm taken from equation (7) is:

$$
\begin{aligned}
&-\frac{p}{2}\ln(2\pi)-\frac{1}{2}\ln\left(\left|\sum\nolimits_1\right|\right)-\frac{1}{2}(x-\mu_1)^T\sum_1^{-1}(x-\mu_1)+\ln(\pi_1)\\
&=-\frac{p}{2}\ln(2\pi)-\frac{1}{2}\ln\left(\left|\sum\nolimits_2\right|\right)-\frac{1}{2}(x-\mu_2)^T\sum_2^{-1}(x-\mu_2)+\ln(\pi_2),\\
&-\frac{1}{2}\ln\left(\left|\sum\nolimits_1\right|\right)-\frac{1}{2}x^T\sum_1^{-1}x-\frac{1}{2}\mu_1^T\sum_1^{-1}\mu_1+\mu_1^T\sum_1^{-1}x+\ln(\pi_1)\\
&=-\frac{1}{2}\ln\left(\left|\sum\nolimits_2\right|\right)-\frac{1}{2}x^T\sum_2^{-1}x-\frac{1}{2}\mu_2^T\sum_2^{-1}\mu_2+\mu_2^T\sum_2^{-1}x+\ln(\pi_2).
\end{aligned}
\tag{8}
$$

We multiplied the sides of the previous equation by 2 and got:

$$
\begin{aligned}
&x^T\left(\sum\nolimits_1-\sum\nolimits_2\right)^{-1}x+2\left(\sum_2^{-1}\mu_2-\sum_1^{-1}\mu_1\right)^T x+\left(\mu_1^t\sum_1^{-1}\mu_1-\mu_2^t\sum_2^{-1}\mu_2\right)\\
&+\ln\left(\frac{\left|\sum_1\right|}{\left|\sum_2\right|}\right)+2\ \ln\left(\frac{\pi_2}{\pi_1}\right)=0.
\end{aligned}
\tag{9}
$$

To obtain Equation (8) in the quadratic form $x^TAx+b^Tx+c=0$, we brought the expression to the QDA discriminant function as:

$$
\begin{aligned}
\delta(x) \coloneqq\ & x^T\left(\sum\nolimits_1-\sum\nolimits_2\right)^{-1}x\\
&+2\left(\sum\nolimits_2^{-1}\mu_2-\sum\nolimits_1^{-1}\mu_1\right)^T x\\
&+\left(\mu_1^t\sum\nolimits_1^{-1}\mu_1-\mu_2^t\sum\nolimits_2^{-1}\mu_2\right),\\
+\ln&\left(\frac{\left|\sum_1\right|}{\left|\sum_2\right|}\right)+2\ \ln\left(\frac{\pi_2}{\pi_1}\right)=0.
\end{aligned}
\tag{10}
$$

The classification term is shown in equation (6).

The QDA Algorithm.

(1) Input independent variables $\mathbf{x}=(x_1,x_2,...,x_p)$ of n sample.

(2) Compute the mean of each class μ_1 and μ_2.

(3) Calculate the prior probability of each class π_1 and π_2.

(4) Compute the covariance matrix for each class Σ_1 and Σ_2.

(5) Calculate the QDA discriminant function as in equation (10).

(6) Assign the class label as in as in equation (6).