Math for Machine Learning

Open Doors to Data Science and Artificial Intelligence

Richard Han

CONTENTS

PREFACE

Welcome to Math for Machine Learning: Open Doors to Data Science and Artificial Intelligence. This is a first textbook in math for machine learning. Be sure to get the companion online course Math for Machine Learning here: Math for Machine Learning Online Course. The online course can be very helpful in conjunction with this book.

The prerequisites for this book and the online course are Linear Algebra, Multivariable Calculus, and Probability. You can find my online course on Linear Algebra here: Linear Algebra Course.

We will not do any programming in this book.

This book will get you started in machine learning in a smooth and natural way, preparing you for more advanced topics and dispelling the belief that machine learning is complicated, difficult, and intimidating.

I want you to succeed and prosper in your career, life, and future endeavors. I am here for you. Visit me at: Online Math Training

1 - INTRODUCTION

Welcome to Math for Machine Learning: Open Doors to Data Science and Artificial Intelligence! My name is Richard Han. This is a first textbook in math for machine learning.

Ideal student:

If you're a working professional needing a refresher on machine learning or a complete beginner who needs to learn Machine Learning for the first time, this book is for you. If your busy schedule doesn't allow you to go back to a traditional school, this book allows you to study on your own schedule and further your career goals without being left behind.

If you plan on taking machine learning in college, this is a great way to get ahead.

If you're currently struggling with machine learning or have struggled with it in the past, now is the time to master it.

Benefits of studying this book:

After reading this book, you will have refreshed your knowledge of machine learning for your career so that you can earn a higher salary.

You will have a required prerequisite for lucrative career fields such as Data Science and Artificial Intelligence.

You will be in a better position to pursue a masters or PhD degree in machine learning and data science.

Why Machine Learning is important:

- Famous uses of machine learning include:
    - Linear discriminant analysis. Linear discriminant analysis can be used to solve classification problems such as spam filtering and classifying patient illnesses.

- Logistic regression. Logistic regression can be used to solve binary classification problems such as determining whether a patient has a certain form of cancer or not.
- Artificial neural networks. Artificial neural networks can be used for applications such as self-driving cars, recommender systems, online marketing, reading medical images, speech and face recognition
- Support Vector machines. Real world applications of SVM's include classification of proteins and classification of images.

What my book offers:

In this book, I cover core topics such as:

- **Linear Regression**
- **Linear Discriminant Analysis**
- **Logistic Regression**
- **Artificial Neural Networks**
- **Support Vector Machines**

I explain each definition and go through each example step by step so that you understand each topic clearly. Throughout the book, there are practice problems for you to try. Detailed solutions are provided after each problem set.

I hope you benefit from the book.

Best regards,

Richard Han

2 – LINEAR REGRESSION

LINEAR REGRESSION

Suppose we have a set of data $(x_1, y_1), \dots, (x_N, y_N)$. This is called the training data.

Each x_i is a vector $\begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix}$ of measurements, where x_{i1} is an instance of the first input variable X_1, x_{i2} is an instance of the second input variable X_2, etc. $X_1, \dots, X_p$ are called ***features*** or ***predictors***.

$y_1, \dots, y_N$ are instances of the output variable Y, which is called the ***response***.

In linear regression, we assume that the response depends on the input variables in a linear fashion: $y = f(X) + \varepsilon$, where $f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$.

Here, ε is called the ***error term*** and $\beta_0, \dots, \beta_p$ are called ***parameters***.

We don't know the values of $\beta_0, \dots, \beta_p$. However, we can use the training data to approximate the values of $\beta_0, \dots, \beta_p$. What we'll do is look at the amount by which the predicted value $f(x_i)$ differs from the actual y_i for each of the pairs $(x_1, y_1), \dots, (x_N, y_N)$ from the training data. So we have $y_i - f(x_i)$ as the difference. We then square this and take the sum for $i = 1, \dots, N$:

$$\sum_{i=1}^{N} \left(y_i - f(x_i)\right)^2$$

This is called the ***residual sum of squares*** and denoted $RSS(\beta)$ where $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$.

We want the residual sum of squares to be as small as possible. Essentially, this means that we want our predicted value $f(x_i)$ to be as close to the actual value y_i as possible, for each of the pairs (x_i, y_i). Doing this will give us a linear function of the input variables that best fits the given training data. In the case of only one input variable, we get the best fit line. In the case of two input variables, we get the best fit plane. And so on, for higher dimensions.

THE LEAST SQUARES METHOD

By minimizing $RSS(\beta)$, we can obtain estimates $\widehat{\beta_0}, \widehat{\beta_1}, \dots, \widehat{\beta_p}$ of the parameters $\beta_0, \dots, \beta_p$. This method is called the ***least squares method***.

Let $X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & & \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix}$ and $\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$.

Then $\boldsymbol{y} - X\beta = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} - \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & & \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$

$$= \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} - \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \cdots + \beta_p x_{1p} \\ \beta_0 + \beta_1 x_{21} + \cdots + \beta_p x_{2p} \\ \vdots \\ \beta_0 + \beta_1 x_{N1} + \cdots + \beta_p x_{Np} \end{bmatrix}$$

$$= \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} - \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_N) \end{bmatrix}$$

$$= \begin{bmatrix} y_1 - f(x_1) \\ \vdots \\ y_N - f(x_N) \end{bmatrix}$$

So $(\boldsymbol{y} - X\beta)^T(\boldsymbol{y} - X\beta) = \sum_{i=1}^{N}\big(y_i - f(x_i)\big)^2 = RSS(\beta)$

$\Longrightarrow \quad RSS(\beta) = (\boldsymbol{y} - X\beta)^T(\boldsymbol{y} - X\beta)$.

Consider the vector of partial derivatives of $RSS(\beta)$:

$$\begin{bmatrix} \dfrac{\partial RSS(\beta)}{\partial \beta_0} \\ \dfrac{\partial RSS(\beta)}{\partial \beta_1} \\ \vdots \\ \dfrac{\partial RSS(\beta)}{\partial \beta_p} \end{bmatrix}$$

$$RSS(\beta) = \left(y_1 - (\beta_0 + \beta_1 x_{11} + \cdots + \beta_p x_{1p})\right)^2 + \cdots + \left(y_N - (\beta_0 + \beta_1 x_{N1} + \cdots + \beta_p x_{Np})\right)^2$$

Let's take the partial derivative with respect to β_0.

$$\frac{\partial RSS(\beta)}{\partial \beta_0} = 2\left(y_1 - (\beta_0 + \beta_1 x_{11} + \cdots + \beta_p x_{1p})\right) \cdot (-1) + \cdots + 2(y_N - (\beta_0 + \beta_1 x_{N1} + \cdots + \beta_p x_{Np})) \cdot (-1)$$

$$= -2 \cdot [1 \quad \cdots \quad 1](\boldsymbol{y} - X\beta)$$

Next, take the partial derivative with respect to β_1.

$$\frac{\partial RSS(\beta)}{\partial \beta_1} = 2\left(y_1 - (\beta_0 + \beta_1 x_{11} + \cdots + \beta_p x_{1p})\right) \cdot (-x_{11}) + \cdots + 2(y_N - (\beta_0 + \beta_1 x_{N1} + \cdots + \beta_p x_{Np})) \cdot (-x_{N1})$$

$$= -2[x_{11} \quad \cdots \quad x_{N1}] \cdot (\boldsymbol{y} - X\beta)$$

In general, $\frac{\partial RSS(\beta)}{\partial \beta_k} = -2[x_{1k} \quad \cdots \quad x_{Nk}] \cdot (\boldsymbol{y} - X\beta)$

So,

$$\begin{bmatrix} \frac{\partial RSS(\beta)}{\partial \beta_0} \\ \frac{\partial RSS(\beta)}{\partial \beta_1} \\ \vdots \\ \frac{\partial RSS(\beta)}{\partial \beta_p} \end{bmatrix} = \begin{bmatrix} -2 \cdot [1 \quad \cdots \quad 1](\boldsymbol{y} - X\beta) \\ -2[x_{11} \quad \cdots \quad x_{N1}](\boldsymbol{y} - X\beta) \\ \vdots \\ -2[x_{1p} \quad \cdots \quad x_{Np}](\boldsymbol{y} - X\beta) \end{bmatrix}$$

$$= -2 \begin{bmatrix} 1 & \cdots & 1 \\ x_{11} & \cdots & x_{N1} \\ & \vdots & \\ x_{1p} & \cdots & x_{Np} \end{bmatrix} (\boldsymbol{y} - X\beta)$$

$$= -2X^T(\boldsymbol{y} - X\beta)$$

If we take the second derivative of $RSS(\beta)$, say $\frac{\partial^2 RSS(\beta)}{\partial \beta_k \partial \beta_j}$, we get

$$\frac{\partial}{\partial \beta_j}(2\left(y_1 - (\beta_0 + \beta_1 x_{11} + \cdots + \beta_p x_{1p})\right) \cdot (-x_{1k}) + \cdots + 2\left(y_N - (\beta_0 + \beta_1 x_{N1} + \cdots + \beta_p x_{Np})\right) \cdot (-x_{Nk}))$$

$$= 2x_{1j}x_{1k} + \cdots + 2x_{Nj}x_{Nk}$$

$$= 2(x_{1j}x_{1k} + \cdots + x_{Nj}x_{Nk})$$

Note $X = \begin{bmatrix} x_{10} & x_{11} & x_{12} & \cdots & x_{1p} \\ x_{20} & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & & \\ x_{N0} & x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix}$

$$\Longrightarrow \qquad X^T X = \begin{bmatrix} x_{10} & x_{20} & \cdots & x_{N0} \\ x_{11} & x_{21} & \cdots & x_{N1} \\ \vdots & & & \\ x_{1p} & x_{2p} & \cdots & x_{Np} \end{bmatrix} \begin{bmatrix} x_{10} & x_{11} & \cdots & x_{1p} \\ x_{20} & x_{21} & \cdots & x_{2p} \\ \vdots & & & \\ x_{N0} & x_{N1} & \cdots & x_{Np} \end{bmatrix}$$

$$= (a_{jk}) \qquad \text{where } a_{jk} = x_{1j}x_{1k} + \cdots + x_{Nj}x_{Nk}$$

So $\frac{\partial^2 RSS(\beta)}{\partial \beta_k \partial \beta_j} = 2a_{jk}$

$\Longrightarrow$ The matrix of second derivatives of $RSS(\beta)$ is $2X^T X$. This matrix is called ***the Hessian***. By the second derivative test, if the Hessian of $RSS(\beta)$ at a critical point is positive definite, then $RSS(\beta)$ has a local minimum there.

If we set our vector of derivatives to $\mathbf{0}$, we get

$$-2X^T(\boldsymbol{y} - X\beta) = \mathbf{0}$$

$$\Longrightarrow \qquad -2X^T\boldsymbol{y} + 2X^T X\beta = \mathbf{0}$$

$$\Longrightarrow \qquad 2X^T X\beta = 2X^T\boldsymbol{y}$$

$$\Longrightarrow \qquad X^T X\beta = X^T\boldsymbol{y}$$

$$\Longrightarrow \qquad \beta = (X^T X)^{-1} X^T \boldsymbol{y}.$$

Thus, we solved for the vector of parameters $\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$ which minimizes the residual sum of squares $RSS(\beta)$.

So we let $\begin{bmatrix} \widehat{\beta_0} \\ \widehat{\beta_1} \\ \vdots \\ \widehat{\beta_p} \end{bmatrix} = (X^T X)^{-1} X^T \boldsymbol{y}.$

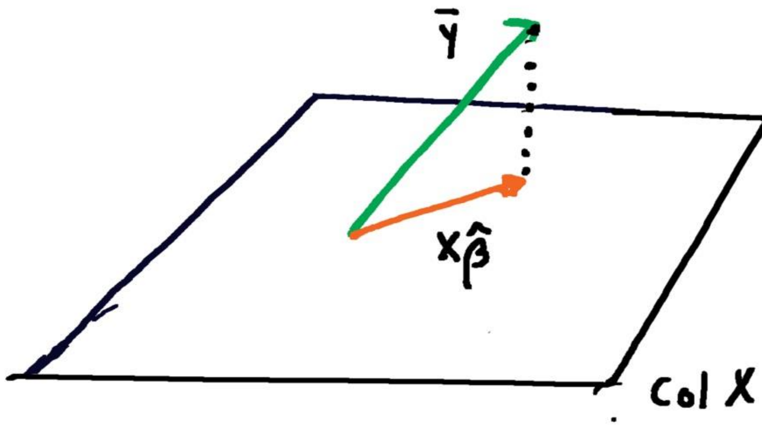
LINEAR ALGEBRA SOLUTION TO LEAST SQUARES PROBLEM

We can arrive at the same solution for the least squares problem by using linear algebra.

Let $X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & & \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix}$ and $\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$ as before, from our training data. We want a vector β such that $X\beta$ is close to $\boldsymbol{y}$. In other words, we want a vector β such that the distance $||X\beta - \boldsymbol{y}||$ between $X\beta$ and $\boldsymbol{y}$ is minimized. A vector β that minimizes $||X\beta - \boldsymbol{y}||$ is called a ***least-squares solution*** of $X\beta = \boldsymbol{y}$.

X is an N by $(p+1)$ matrix. We want a $\hat{\beta}$ in $\mathbb{R}^{p+1}$ such that $X\hat{\beta}$ is closest to $\boldsymbol{y}$. Note that $X\hat{\beta}$ is a linear combination of the columns of X. So $X\hat{\beta}$ lies in the span of the columns of X, which is a subspace of $\mathbb{R}^N$ denoted $Col\ X$. So we want the vector in $Col\ X$ that is closest to $\boldsymbol{y}$. The projection of $\boldsymbol{y}$ onto the subspace $Col\ X$ is that vector.

$proj_{Col\ X}\boldsymbol{y} = X\hat{\beta}$ for some $\hat{\beta} \in \mathbb{R}^{p+1}$.



Consider $\boldsymbol{y} - X\hat{\beta}$. Note that $\boldsymbol{y} = X\hat{\beta} + (\boldsymbol{y} - X\hat{\beta})$.

$\mathbb{R}^N$ can be broken into two subspaces $Col\ X$ and $(Col\ X)^\perp$, where $(Col\ X)^\perp$ is the subspace of $\mathbb{R}^N$ consisting of all vectors that are orthogonal to the vectors in $Col\ X$. Any vector in $\mathbb{R}^N$ can be written uniquely as $\boldsymbol{z} + \boldsymbol{w}$ where $\boldsymbol{z} \in Col\ X$ and $\boldsymbol{w} \in (Col\ X)^\perp$.

Since $\boldsymbol{y} \in \mathbb{R}^N$, and $\boldsymbol{y} = X\hat{\beta} + (\boldsymbol{y} - X\hat{\beta})$, with $X\hat{\beta} \in Col\ X$, the second vector $\boldsymbol{y} - X\hat{\beta}$ must lie in $(Col\ X)^\perp$.

$\Rightarrow$ $\quad \boldsymbol{y} - X\hat{\beta}$ is orthogonal to the columns of X.

$\Rightarrow$ $\quad X^T(\boldsymbol{y} - X\hat{\beta}) = \boldsymbol{0}$

$\Rightarrow$ $\quad X^T\boldsymbol{y} - X^TX\hat{\beta} = \boldsymbol{0}$.

$\Rightarrow$ $\quad X^TX\hat{\beta} = X^T\boldsymbol{y}$.

Thus, it turns out that the set of least-squares solutions of $X\beta = \boldsymbol{y}$ consists of all and only the solutions to the matrix equation $X^TX\beta = X^T\boldsymbol{y}$.

If X^TX is positive definite, then the eigenvalues of X^TX are all positive. So 0 is not an eigenvalue of X^TX. It follows that X^TX is invertible. Then, we can solve the equation $X^TX\hat{\beta} = X^T\boldsymbol{y}$ for $\hat{\beta}$ to get $\hat{\beta} =$

$(X^T X)^{-1} X^T \boldsymbol{y}$, which is the same result we got earlier using multi-variable calculus.

EXAMPLE: LINEAR REGRESSION

Suppose we have the following training data:

$(x_1, y_1) = (1, 1), (x_2, y_2) = (2, 4), (x_3, y_3) = (3, 4)$.

Find the best fit line using the least squares method. Find the predicted value for $x = 4$.

Solution:

Form $X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$ and $\boldsymbol{y} = \begin{bmatrix} 1 \\ 4 \\ 4 \end{bmatrix}$.

The coefficients β_0, β_1 for the best fit line $f(x) = \beta_0 + \beta_1 x$ are given by $\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = (X^T X)^{-1} X^T \boldsymbol{y}$.

$X^T = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix}$

$\Longrightarrow \quad X^T X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}$

$\Longrightarrow \quad (X^T X)^{-1} = \begin{bmatrix} 7/3 & -1 \\ -1 & 1/2 \end{bmatrix}$

$\Longrightarrow \quad (X^T X)^{-1} X^T \boldsymbol{y} = \begin{bmatrix} 7/3 & -1 \\ -1 & 1/2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \\ 4 \end{bmatrix}$

$$= \begin{bmatrix} 0 \\ 3/2 \end{bmatrix}$$

$\Longrightarrow \quad \beta_0 = 0$ and $\beta_1 = 3/2$.

Thus, the best fit line is given by $f(x) = \left(\frac{3}{2}\right) x$.

The predicted value for $x = 4$ is $f(4) = \left(\frac{3}{2}\right) \cdot 4 = 6$.

SUMMARY: LINEAR REGRESSION

- In the least squares method, we seek a linear function of the input variables that best fits the given training data. We do this by minimizing the residual sum of squares.

- To minimize the residual sum of squares, we apply the second derivative test from multi-variable calculus.

- We can arrive at the same solution to the least squares problem using linear algebra.

PROBLEM SET: LINEAR REGRESSION

1. Suppose we have the following training data:

   $(x_1, y_1) = (0, 2), (x_2, y_2) = (1, 1),$

   $(x_3, y_3) = (2, 4), (x_4, y_4) = (3, 4).$

   Find the best fit line using the least squares method. Find the predicted value for $x = 4$.

2. Suppose we have the following training data:

   $(x_1, y_1), (x_2, y_2), (x_3, y_3)$ where
   $x_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, x_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, x_3 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, x_4 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$
   and $y_1 = 1, y_2 = 0, y_3 = 0, y_4 = 2.$

   Find the best fit plane using the least squares method. Find the predicted value for $x = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$.

SOLUTION SET: LINEAR REGRESSION

1. Form $X = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$ and $\boldsymbol{y} = \begin{bmatrix} 2 \\ 1 \\ 4 \\ 4 \end{bmatrix}$.

   The coefficients β_0, β_1 for the best fit line $f(x) = \beta_0 + \beta_1 x$ are given by $\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = (X^T X)^{-1} X^T \boldsymbol{y}$.

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{bmatrix} \qquad \Rightarrow \; X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 4 & 6 \\ 6 & 14 \end{bmatrix}$$

$$\Rightarrow \; (X^T X)^{-1} = \begin{bmatrix} \frac{7}{10} & -\frac{3}{10} \\ -\frac{3}{10} & \frac{1}{5} \end{bmatrix}$$

$$\Rightarrow \; (X^T X)^{-1} X^T \boldsymbol{y} = \begin{bmatrix} \frac{7}{10} & -\frac{3}{10} \\ -\frac{3}{10} & \frac{1}{5} \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ 4 \\ 4 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{14}{10} \\ \frac{9}{10} \end{bmatrix}$$

$$\Rightarrow \; \beta_0 = \frac{14}{10} \quad \text{and } \beta_1 = \frac{9}{10}.$$

   Thus, the best fit line is given by

$$f(x) = \frac{14}{10} + \frac{9}{10} x$$

   The predicted value for $x = 4$ is $f(4) = \frac{14}{10} + \frac{9}{10} \cdot 4 = 5$.

2. Form $X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$ and $\boldsymbol{y} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 2 \end{bmatrix}$.

   The coefficients $\beta_0, \beta_1, \beta_2$ for the best fit line $f(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ are given by $\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} =$ $(X^T X)^{-1} X^T \boldsymbol{y}$.

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \qquad \Rightarrow \; X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 2 & 1 \\ 2 & 1 & 2 \end{bmatrix}$$

$$\implies (X^TX)^{-1} = \begin{bmatrix} \frac{3}{4} & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 & 0 \\ -\frac{1}{2} & 0 & 1 \end{bmatrix}$$

$$\implies (X^TX)^{-1}X^T\boldsymbol{y} = \begin{bmatrix} \frac{3}{4} & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 & 0 \\ -\frac{1}{2} & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 2 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{3}{4} & \frac{1}{4} & \frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 2 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{4} \\ \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$$

$$\implies \beta_0 = \frac{1}{4} \quad , \ \beta_1 = \frac{1}{2}, \beta_2 = \frac{1}{2}$$

Thus, the best fit plane is given by

$$f(x_1, x_2) = \frac{1}{4} + \frac{1}{2}x_1 + \frac{1}{2}x_2$$

The predicted value for $x = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$ is $f(2, 2) = 2\frac{1}{4}$.

3 – LINEAR DISCRIMINANT ANALYSIS

CLASSIFICATION

In the problem of regression, we had a set of data $(x_1, y_1), \dots, (x_N, y_N)$ and we wanted to predict the values for the response variable Y for new data points. The values that Y took were numerical, quantitative, values. In certain problems, the values for the response variable Y that we want to predict are not quantitative but qualitative. So the values for Y will take on values from a finite set of classes or categories. Problems of this sort are called ***classification problems***. Some examples of a classification problem are classifying an email as spam or not spam and classifying a patient's illness as one among a finite number of diseases.

LINEAR DISCRIMINANT ANALYSIS

One method for solving a classification problem is called ***linear discriminant analysis***.

What we'll do is estimate $\Pr(Y = k | X = x)$, the probability that Y is the class k given that the input variable X is x. Once we have all of these probabilities for a fixed x, we pick the class k for which the probability $\Pr(Y = k | X = x)$ is largest. We then classify x as that class k.

THE POSTERIOR PROBABILITY FUNCTIONS

In this section, we'll build a formula for the posterior probability $\Pr(Y = k | X = x)$.

Let $\pi_k = \Pr(Y = k)$, the prior probability that $Y = k$.

Let $f_k(x) = \Pr(X = x | Y = k)$, the probability that $X = x$, given that $Y = k$.

By Bayes' rule,

$$\Pr(Y = k | X = x) = \frac{\Pr(X = x | Y = k) \cdot \Pr(Y = k)}{\sum_{l=1}^{K} \Pr(X = x | Y = l) \Pr(Y = l)}$$

Here we assume that k can take on the values $1, \dots, K$.

$$= \frac{f_k(x) \cdot \pi_k}{\sum_{l=1}^{K} f_l(x) \cdot \pi_l}$$

$$= \frac{\pi_k \cdot f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

We can think of $\Pr(Y = k | X = x)$ as a function of x and denote it as $p_k(x)$.

So $p_k(x) = \frac{\pi_k \cdot f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$. Recall that $p_k(x)$ is the posterior probability that $Y = k$ given that $X = x$.

MODELLING THE POSTERIOR PROBABILITY FUNCTIONS

Remember that we wanted to estimate $\Pr(Y = k|X = x)$ for any given x. That is, we want an estimate for $p_k(x)$. If we can get estimates for $\pi_k, f_k(x), \pi_l$ and $f_l(x)$ for each $l = 1, \ldots, K$, then we would have an estimate for $p_k(x)$.

Let's say that $X = (X_1, X_2, \ldots, X_p)$ where $X_1, \ldots, X_p$ are the input variables. So the values of X will be vectors of p elements.

We will assume that the conditional distribution of X given $Y = k$ is the multivariate Gaussian distribution $N(\mu_k, \Sigma)$, where μ_k is a class-specific mean vector and Σ is the covariance of X.

The class-specific mean vector μ_k is given by the vector of class-specific means $\begin{bmatrix} \mu_{k1} \\ \vdots \\ \mu_{kp} \end{bmatrix}$, where μ_{kj} is the class-specific mean of X_j.

So $\mu_{kj} = \sum_{i:y_i=k} x_{ij} \Pr(X_j = x_{ij})$. Recall that $x_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}$. (For all those x_i for which $y_i = k$, we're taking the mean of their jth components.)

Σ, the covariance matrix of X, is given by the matrix of covariances of X_i and X_j.

So $\Sigma = (a_{ij})$, where $a_{ij} = Cov(X_i, X_j) \stackrel{\text{def}}{=} E[(X_i - \mu_{X_i})(X_j - \mu_{X_j})]$.

The multivariate Gaussian density is given by

$$f(x) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

for the multivariate Gaussian distribution $N(\mu, \Sigma)$.

Since we're assuming that the conditional distribution of X given $Y = k$ is the multivariate Gaussian distribution $N(\mu_k, \Sigma)$, we have that

$$\Pr(X = x|Y = k) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)}.$$

Recall that $f_k(x) = \Pr(X = x|Y = k)$.

So $f_k(x) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)}$.

Recall that $p_k(x) = \frac{\pi_k \cdot f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$.

Plugging in what we have for $f_k(x)$, we get

$$p_k(x) = \frac{\pi_k \cdot \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)}}{\sum_{l=1}^{K} \pi_l \cdot \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_l)^T\Sigma^{-1}(x-\mu_l)}}$$

$$= \frac{\pi_k \cdot e^{-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)}}{\sum_{l=1}^{K} \pi_l \cdot e^{-\frac{1}{2}(x-\mu_l)^T\Sigma^{-1}(x-\mu_l)}} \ .$$

Note that the denominator is $(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}} \sum_{l=1}^{K} \pi_l f_l(x)$ and that

$$\sum_{l=1}^{K} \pi_l f_l(x) = \sum_{l=1}^{K} f_l(x)\pi_l$$

$$= \sum_{l=1}^{K} \Pr(X = x|Y = l) \Pr(Y = l)$$

$$= \Pr(X = x).$$

So the denominator is just $(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}} \Pr(X = x)$.

Hence, $p_k(x) = \frac{\pi_k \cdot e^{-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)}}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}} Pr(X=x)}$.

LINEAR DISCRIMINANT FUNCTIONS

Recall that we want to choose the class k for which the posterior probability $p_k(x)$ is largest. Since the logarithm function is order-preserving, maximizing $p_k(x)$ is the same as maximizing $\log p_k(x)$.

Taking $\log p_k(x)$ gives $\log \frac{\pi_k \cdot e^{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)}}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}} Pr(X=x)}$

$$= \log \pi_k + \left(-\frac{1}{2}\right)(x-\mu_k)^T \Sigma^{-1}(x-\mu_k) - \log\left((2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}} \Pr(X=x)\right)$$

$$= log\, \pi_k + \left(-\frac{1}{2}\right)(x-\mu_k)^T \Sigma^{-1}(x-\mu_k) - \log C \qquad \text{where } C = (2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}} P\,r(X=x).$$

$$= \log \pi_k - \frac{1}{2}(x^T\Sigma^{-1} - \mu_k^T\Sigma^{-1})(x-\mu_k) - \log C$$

$$= \log \pi_k - \frac{1}{2}[x^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_k - \mu_k^T\Sigma^{-1}x + \mu_k^T\Sigma^{-1}\mu_k] - \log C$$

$$= log\, \pi_k - \frac{1}{2}[x^T\Sigma^{-1}x - 2x^T\Sigma^{-1}\mu_k + \mu_k^T\Sigma^{-1}\mu_k] - log\, C,$$

because $x^T\Sigma^{-1}\mu_k = \mu_k^T\Sigma^{-1}x$

Proof: $x^T\Sigma^{-1}\mu_k = \mu_k(\Sigma^{-1})^T x$

$= \mu_k^T(\Sigma^T)^{-1}x$

$= \mu_k^T\Sigma^{-1}x$ because Σ is symmetric.

$$= \log \pi_k - \frac{1}{2}x^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k - \log C$$

$$= x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log \pi_k - \frac{1}{2}x^T\Sigma^{-1}x - \log C$$

Let $\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + log\, \pi_k$.

Then $\log p_k(x) = \delta_k(x) - \frac{1}{2}x^T\Sigma^{-1}x - log\, C$.

$\delta_k(x)$ is called a ***linear discriminant function***. Maximizing $\log p_k(x)$ is the same as maximizing $\delta_k(x)$ since $-\frac{1}{2}x^T\Sigma^{-1}x - \log C$ does not depend on k.

ESTIMATING THE LINEAR DISCRIMINANT FUNCTIONS

Now, if we can find estimates for π_k, μ_k, and Σ, then we would have an estimate for $p_k(x)$ and hence for $\log p_k(x)$ and $\delta_k(x)$.

In an attempt to maximize $p_k(x)$, we instead maximize the estimate of $p_k(x)$, which is the same as maximizing the estimate of $\delta_k(x)$.

π_k can be estimated as $\widehat{\pi_k} = \frac{N_k}{N}$ where N_k is the number of training data points in class k and N is the total number of training data points.

Remember $\pi_k = \Pr(Y = k)$. We're estimating this by just taking the proportion of data points in class k.

The class-specific mean vector $\mu_k = \begin{bmatrix} \mu_{k1} \\ \vdots \\ \mu_{kp} \end{bmatrix}$, where $\mu_{kj} = \sum_{i:y_i=k} x_{ij} \Pr\left(X_j = x_{ij}\right)$.

We can estimate μ_{kj} as $\frac{1}{N_k}\sum_{i:y_i=k} x_{ij}$.

So we can estimate μ_k as $\widehat{\mu_k} = \begin{bmatrix} \frac{1}{N_k}\sum_{i:y_i=k} x_{i1} \\ \vdots \\ \frac{1}{N_k}\sum_{i:y_i=k} x_{ip} \end{bmatrix} = \frac{1}{N_k}\begin{bmatrix} \sum_{i:y_i=k} x_{i1} \\ \vdots \\ \sum_{i:y_i=k} x_{ip} \end{bmatrix}$

$$= \frac{1}{N_k}\sum_{i:y_i=k} \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}$$

$$= \frac{1}{N_k}\sum_{i:y_i=k} x_i$$

In other words, $\widehat{\mu_k} = \frac{1}{N_k}\sum_{i:y_i=k} x_i$. We estimate the class-specific mean vector by the vector of averages of each component over all x_i in class k.

Finally, the covariance matrix Σ is estimated as $\hat{\Sigma} = \frac{1}{N-K}\sum_{k=1}^{K}\sum_{i:y_i=k}(x_i - \widehat{\mu_k})\,(x_i - \widehat{\mu_k})^T$.

Recall that $\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + log\,\pi_k$.

So, $\widehat{\delta_k}(x) = x^T\hat{\Sigma}^{-1}\widehat{\mu_k} - \frac{1}{2}(\widehat{\mu_k})^T\hat{\Sigma}^{-1}\widehat{\mu_k} + log\,\widehat{\pi_k}$.

Note that $\hat{\Sigma}, \widehat{\mu_k}$, and $\widehat{\pi_k}$ only depend on the training data and not on x. Note that x is a vector and $x^T\hat{\Sigma}^{-1}\widehat{\mu_k}$ is a linear combination of the components of x. Hence, $\widehat{\delta_k}(x)$ is a linear combination of the components of x. This is why it's called a linear discriminant function.

CLASSIFYING DATA POINTS USING LINEAR DISCRIMINANT FUNCTIONS

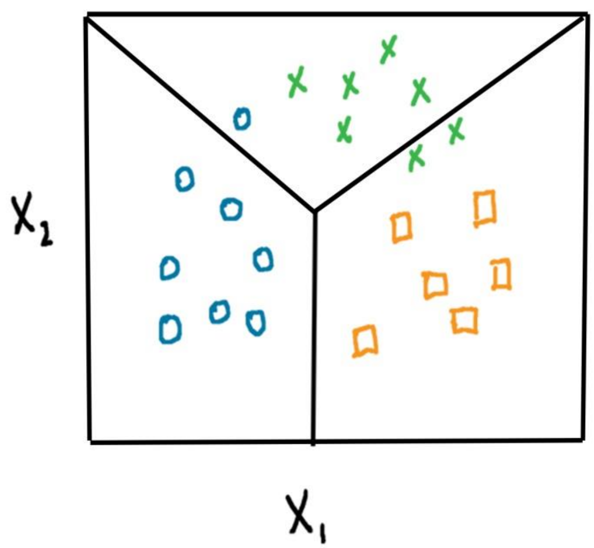
If (k_1, k_2) is a pair of classes, we can consider whether $\widehat{\delta_{k_1}}(x) > \widehat{\delta_{k_2}}(x)$. If so, we know x is not in class k_2. Then, we can compare $\widehat{\delta_{k_1}}(x) > \widehat{\delta_{k_3}}(x)$ and rule out another class. Once we've exhausted all the classes, we'll know which class x should be assigned to.

Setting $\widehat{\delta_{k_1}}(x) = \widehat{\delta_{k_2}}(x)$, we get

$$x^T\hat{\Sigma}^{-1}\widehat{\mu_{k_1}} - \frac{1}{2}\left(\widehat{\mu_{k_1}}\right)^T\hat{\Sigma}^{-1}\widehat{\mu_{k_1}} + log\,\widehat{\pi_{k_1}} = x^T\hat{\Sigma}^{-1}\widehat{\mu_{k_2}} - \frac{1}{2}\left(\widehat{\mu_{k_2}}\right)^T\hat{\Sigma}^{-1}\widehat{\mu_{k_2}} + log\,\widehat{\pi_{k_2}}.$$

This gives us a hyperplane in $\mathbb{R}^p$ which separates class $k_1$from class k_2.

If we find the separating hyperplane for each pair of classes, we get something like this:



In this example, $p = 2$ and $K = 3$.

LDA EXAMPLE 1

Suppose we have a set of data $(x_1, y_1), \ldots, (x_6, y_6)$ as follows:

$x_1 = (1,3), x_2 = (2,3), x_3 = (2,4), x_4 = (3,1), x_5 = (3,2), x_6 = (4,2),$

with $y_1 = y_2 = y_3 = k_1 = 1$ and $y_4 = y_5 = y_6 = k_2 = 2$.

Apply linear discriminant analysis by doing the following:

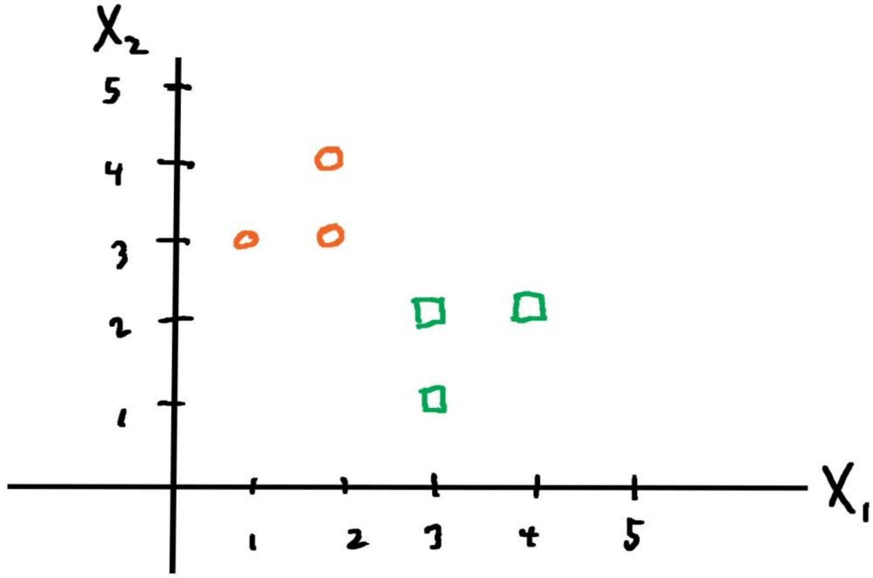
a) Find estimates for the linear discriminant functions $\delta_1(x)$ and $\delta_2(x)$.

b) Find the line that decides between the two classes.

c) Classify the new point $x = (5, 0)$.

Solution:

Here is a graph of the data points:



The number of features p is 2, the number of classes K is 2, the total number of data points N is 6, the number N_1 of data points in class k_1 is 3, and the number N_2 of data points in class k_2 is 3.

First, we will find estimates for π_1 and π_2, the prior probabilities that $Y = k_1$ and $Y = k_2$, respectively.

Then, we will find estimates for μ_1 and μ_2, the class-specific mean vectors.

We can then calculate the estimate for the covariance matrix Σ.

Finally, using the estimates $\widehat{\pi_1}, \widehat{\pi_2}, \widehat{\mu_1}, \widehat{\mu_2}, \hat{\Sigma}$, we can find the estimates for the linear discriminant functions $\delta_1(x)$ and $\delta_2(x)$.

$$\widehat{\pi_1} = \frac{N_1}{N} = \frac{3}{6} = \frac{1}{2}$$

$$\widehat{\pi_2} = \frac{N_2}{N} = \frac{3}{6} = \frac{1}{2}$$

$$\widehat{\mu_1} = \frac{1}{N_1}\sum_{i:y_i=1} x_i = \frac{1}{3}[x_1 + x_2 + x_3] = \begin{bmatrix} 5/3 \\ 10/3 \end{bmatrix}$$

$$\widehat{\mu_2} = \frac{1}{N_2}\sum_{i:y_i=2} x_i = \frac{1}{3}[x_4 + x_5 + x_6] = \begin{bmatrix} 10/3 \\ 5/3 \end{bmatrix}$$

$$\hat{\Sigma} = \frac{1}{N-K}\sum_{k=1}^{K}\sum_{i:y_i=k}(x_i - \widehat{\mu_k})(x_i - \widehat{\mu_k})^T$$

$$= \frac{1}{6-2}\sum_{k=1}^{2}\sum_{i:y_i=k}(x_i - \widehat{\mu_k})(x_i - \widehat{\mu_k})^T$$

Plugging in what we got for $\widehat{\mu_1}$ and $\widehat{\mu_2}$, we get

$$\hat{\Sigma} = \frac{1}{4}\begin{bmatrix} 4/3 & 2/3 \\ 2/3 & 4/3 \end{bmatrix} = \begin{bmatrix} 1/3 & 1/6 \\ 1/6 & 1/3 \end{bmatrix}$$

$$\Longrightarrow\ \hat{\Sigma}^{-1} = \begin{bmatrix} 4 & -2 \\ -2 & 4 \end{bmatrix}$$

$$\widehat{\delta_1}(x) = x^T\hat{\Sigma}^{-1}\widehat{\mu_1} - \frac{1}{2}(\widehat{\mu_1})^T\hat{\Sigma}^{-1}\widehat{\mu_1} + log\,\widehat{\pi_1}.$$

$$= x^T\begin{bmatrix} 0 \\ 10 \end{bmatrix} - \frac{1}{2}\left(\frac{100}{3}\right) + \log\frac{1}{2}$$

$$= 10X_2 - \frac{50}{3} + \log\frac{1}{2}$$

$$\widehat{\delta_2}(x) = x^T\hat{\Sigma}^{-1}\widehat{\mu_2} - \frac{1}{2}(\widehat{\mu_2})^T\hat{\Sigma}^{-1}\widehat{\mu_2} + log\,\widehat{\pi_2}.$$

$$= x^T\begin{bmatrix} 10 \\ 0 \end{bmatrix} - \frac{1}{2}\left(\frac{100}{3}\right) + \log\frac{1}{2}$$

$$= 10X_1 - \frac{50}{3} + \log\frac{1}{2}$$

Setting $\widehat{\delta_1}(x) = \widehat{\delta_2}(x)$

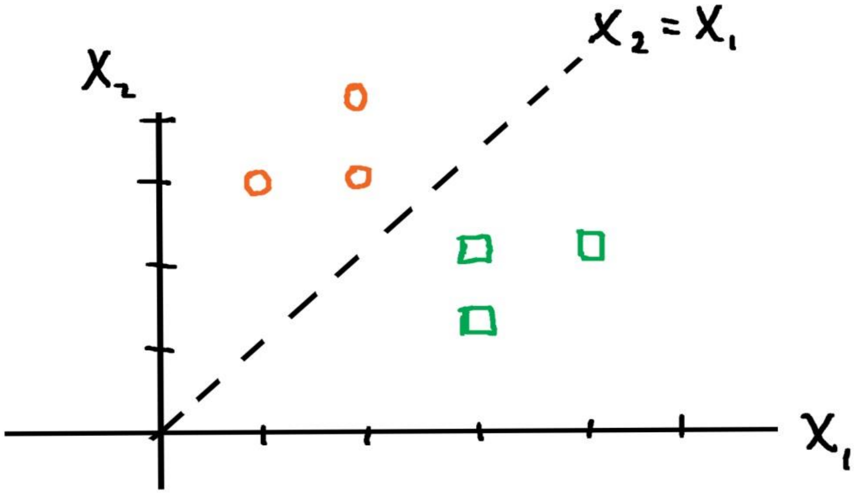
$$\Longrightarrow \qquad 10X_2 - \tfrac{50}{3} + log\,\tfrac{1}{2} = 10X_1 - \tfrac{50}{3} + log\,\tfrac{1}{2}$$

$$\Longrightarrow \qquad 10X_2 = 10X_1$$

$$\Longrightarrow \qquad X_2 = X_1.$$

So, the line that decides between the two classes is given by $X_2 = X_1$.

Here is a graph of the deciding line:



If $\widehat{\delta_1}(x) > \widehat{\delta_2}(x)$, then we classify x as of class k_1. So if x is above the line $X_2 = X_1$, then we classify x as of class k_1. Conversely, if $\widehat{\delta_1}(x) < \widehat{\delta_2}(x)$, then we classify x as of class k_2. This corresponds to x being below the line $X_2 = X_1$.

The point $(5, 0)$ is below the line; so we classify it as of class k_2.

LDA EXAMPLE 2

Suppose we have a set of data $(x_1, y_1), \ldots, (x_6, y_6)$ as follows:

$x_1 = (0, 2), x_2 = (1, 2), x_3 = (2, 0), x_4 = (2, 1), x_5 = (3, 3), x_6 = (4, 4)$,

with $y_1 = y_2 = k_1 = 1$, $y_3 = y_4 = k_2 = 2$, and $y_5 = y_6 = k_3 = 3$.

Apply linear discriminant analysis by doing the following:

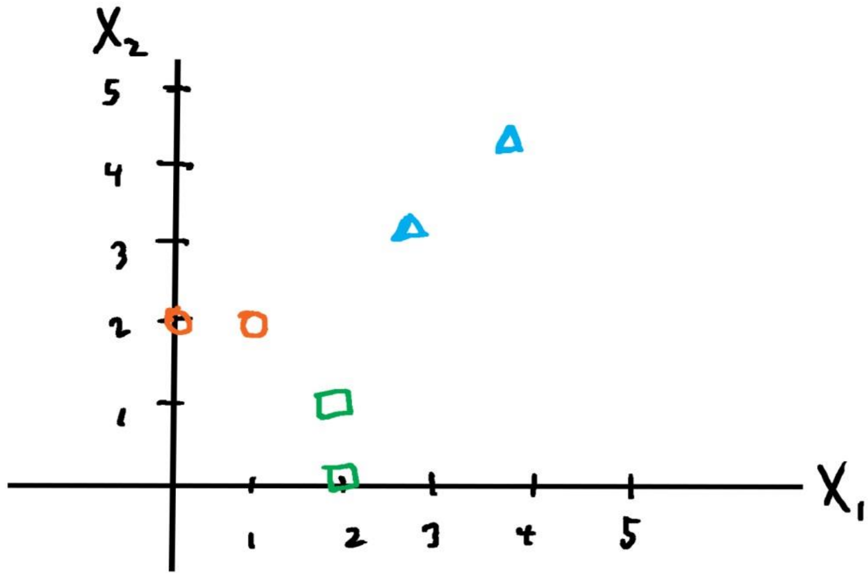
a) Find estimates for the linear discriminant functions $\delta_1(x), \delta_2(x)$, and $\delta_3(x)$.

b) Find the lines that decide between each pair of classes.

c) Classify the new point $x = (1, 3)$.

Solution:

Here is a graph of the data points:



The number of features p is 2, the number of classes K is 3, the total number of data points N is 6, the number N_1 of data points in class k_1 is 2, the number N_2 of data points in class k_2 is 2, and the number N_3 of data points in class k_3 is 2.

First, we will find estimates for π_1, π_2, π_3, the prior probabilities that $Y = k_1$, $Y = k_2$, $Y = k_3$, respectively.

Then, we will find estimates for μ_1, μ_2, μ_3, the class-specific mean vectors.

We can then calculate the estimate for the covariance matrix Σ.

Finally, using the estimates $\widehat{\pi_1}, \widehat{\pi_2}, \widehat{\pi_3}, \widehat{\mu_1}, \widehat{\mu_2}, \widehat{\mu_3}, \hat{\Sigma}$, we can find the estimates for the linear discriminant functions $\delta_1(x), \delta_2(x)$, and $\delta_3(x)$.

$$\widehat{\pi_1} = \frac{N_1}{N} = \frac{2}{6} = \frac{1}{3}$$

$$\widehat{\pi_2} = \frac{N_2}{N} = \frac{2}{6} = \frac{1}{3}$$

$$\widehat{\pi_3} = \frac{N_3}{N} = \frac{2}{6} = \frac{1}{3}$$

$$\widehat{\mu_1} = \frac{1}{N_1} \sum_{i:y_i=1} x_i = \frac{1}{2}[x_1 + x_2] = \begin{bmatrix} 1/2 \\ 2 \end{bmatrix}$$

$$\widehat{\mu_2} = \frac{1}{N_2} \sum_{i:y_i=2} x_i = \frac{1}{2}[x_3 + x_4] = \begin{bmatrix} 2 \\ 1/2 \end{bmatrix}$$

$$\widehat{\mu_3} = \frac{1}{N_3}\sum_{i:y_i=3} x_i = \frac{1}{2}[x_5 + x_6] = \begin{bmatrix} 7/2 \\ 7/2 \end{bmatrix}$$

$$\hat{\Sigma} = \frac{1}{N-K}\sum_{k=1}^{K}\sum_{i:y_i=k}(x_i - \widehat{\mu_k})(x_i - \widehat{\mu_k})^T$$

$$= \frac{1}{6-3}\begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix} = \begin{bmatrix} 1/3 & 1/6 \\ 1/6 & 1/3 \end{bmatrix}$$

$$\Rightarrow \quad \hat{\Sigma}^{-1} = \begin{bmatrix} 4 & -2 \\ -2 & 4 \end{bmatrix}$$

$$\widehat{\delta_1}(x) = x^T\hat{\Sigma}^{-1}\widehat{\mu_1} - \frac{1}{2}(\widehat{\mu_1})^T\hat{\Sigma}^{-1}\widehat{\mu_1} + log\,\widehat{\pi_1}.$$

$$= x^T\begin{bmatrix} -2 \\ 7 \end{bmatrix} - \left(\frac{13}{2}\right) + \log\frac{1}{3}$$

$$= -2X_1 + 7X_2 - \frac{13}{2} + \log\frac{1}{3}$$

$$\widehat{\delta_2}(x) = x^T\hat{\Sigma}^{-1}\widehat{\mu_2} - \frac{1}{2}(\widehat{\mu_2})^T\hat{\Sigma}^{-1}\widehat{\mu_2} + log\,\widehat{\pi_2}.$$

$$= x^T\begin{bmatrix} 7 \\ -2 \end{bmatrix} - \left(\frac{13}{2}\right) + \log\frac{1}{3}$$

$$= 7X_1 - 2X_2 - \frac{13}{2} + \log\frac{1}{3}$$

$$\widehat{\delta_3}(x) = x^T\hat{\Sigma}^{-1}\widehat{\mu_3} - \frac{1}{2}(\widehat{\mu_3})^T\hat{\Sigma}^{-1}\widehat{\mu_3} + log\,\widehat{\pi_3}.$$

$$= x^T\begin{bmatrix} 7 \\ 7 \end{bmatrix} - \left(\frac{49}{2}\right) + \log\frac{1}{3}$$

$$= 7X_1 + 7X_2 - \frac{49}{2} + \log\frac{1}{3}$$

Setting $\widehat{\delta_1}(x) = \widehat{\delta_2}(x)$

$$\Rightarrow \quad -2X_1 + 7X_2 - \frac{13}{2} + log\,\frac{1}{3} = 7X_1 - 2X_2 - \frac{13}{2} + log\,\frac{1}{3}$$

$$\Rightarrow \quad -2X_1 + 7X_2 = 7X_1 - 2X_2$$

$$\Rightarrow \quad 9X_2 = 9X_1$$

$\Longrightarrow \qquad X_2 = X_1.$

So, the line that decides between classes k_1 and k_2 is given by $X_2 = X_1$.

Setting $\widehat{\delta_1}(x) = \widehat{\delta_3}(x)$

$\Longrightarrow \qquad -2X_1 + 7X_2 - \frac{13}{2} + log\frac{1}{3} = 7X_1 + 7X_2 - \frac{49}{2} + log\frac{1}{3}$

$\Longrightarrow \qquad 18 = 9X_1$

$\Longrightarrow \qquad X_1 = 2$

So, the line that decides between classes k_1 and k_3 is given by $X_1 = 2$.

Setting $\widehat{\delta_2}(x) = \widehat{\delta_3}(x)$

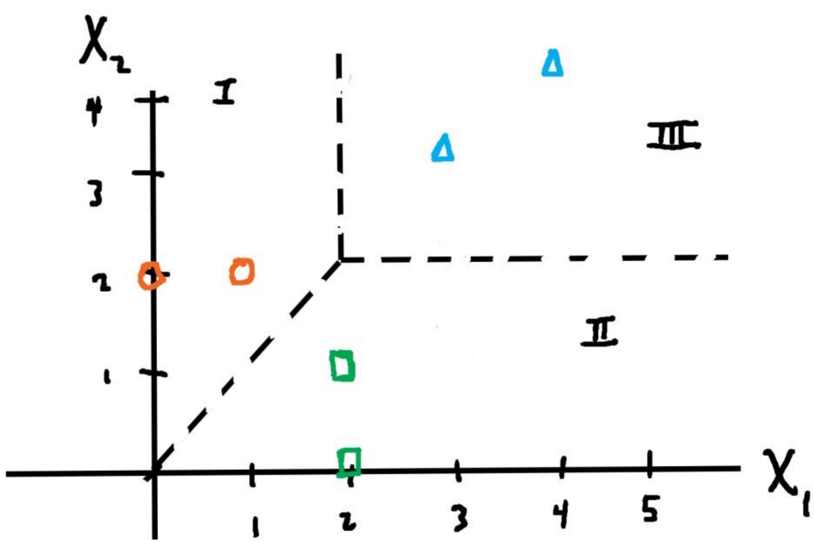
$\Longrightarrow \qquad 7X_1 - 2X_2 - \frac{13}{2} + log\frac{1}{3} = 7X_1 + 7X_2 - \frac{49}{2} + log\frac{1}{3}$

$\Longrightarrow \qquad 18 = 9X_2$

$\Longrightarrow \qquad X_2 = 2$

So, the line that decides between classes k_2 and k_3 is given by $X_2 = 2$.

Here is a graph of the deciding lines:



The lines divide the plane into 3 regions.

$\widehat{\delta_1}(x) > \widehat{\delta_2}(x)$ corresponds to the region above the line $X_2 = X_1$. Conversely, $\widehat{\delta_1}(x) < \widehat{\delta_2}(x)$ corresponds to the region below the line $X_2 = X_1$.

$\widehat{\delta_1}(x) > \widehat{\delta_3}(x)$ corresponds to the region to the left of the line $X_1 = 2$. Conversely, $\widehat{\delta_1}(x) < \widehat{\delta_3}(x)$

corresponds to the region to the right of $X_1 = 2$.

$\widehat{\delta_2}(x) > \widehat{\delta_3}(x)$ corresponds to the region below the line $X_2 = 2$. Conversely, $\widehat{\delta_2}(x) < \widehat{\delta_3}(x)$ corresponds to the region above the line $X_2 = 2$.

If $\widehat{\delta_1}(x) > \widehat{\delta_2}(x)$ and $\widehat{\delta_1}(x) > \widehat{\delta_3}(x)$, then we classify x as of class k_1. So if x is in region I, then we classify x as of class k_1. Conversely, if x is in region II, then we classify x as of class k_2; and if x is in region III, we classify x as of class k_3.

The point $(1, 3)$ is in region I; so we classify it as of class k_1.

SUMMARY: LINEAR DISCRIMINANT ANALYSIS

- In linear discriminant analysis, we find estimates $\widehat{p_k}(x)$ for the posterior probability $p_k(x)$ that $Y = k$ given that $X = x$. We classify x according to the class k that gives the highest estimated posterior probability $\widehat{p_k}(x)$.

- Maximizing the estimated posterior probability $\widehat{p_k}(x)$ is equivalent to maximizing the log of $\widehat{p_k}(x)$, which, in turn, is equivalent to maximizing the estimated linear discriminant function $\widehat{\delta_k}(x)$.

- We find estimates of the prior probability π_k that $Y = k$, of the class-specific mean vectors μ_k, and of the covariance matrix Σ in order to estimate the linear discriminant functions $\delta_k(x)$.

- By setting $\widehat{\delta_k}(x) = \widehat{\delta_{k'}}(x)$ for each pair $(k,\, k')$ of classes, we get hyperplanes in $\mathbb{R}^p$ that, together, divide $\mathbb{R}^p$ into regions corresponding to the distinct classes.

- We classify x according to the class k for which $\widehat{\delta_k}(x)$ is largest.

PROBLEM SET: LINEAR DISCRIMINANT ANALYSIS

1. Suppose we have a set of data $(x_1, y_1), \dots, (x_6, y_6)$ as follows:

   $x_1 = (1, 2), x_2 = (2, 1), x_3 = (2, 2), x_4 = (3, 3), x_5 = (3, 4), x_6 = (4, 3)$ with

   $y_1 = y_2 = y_3 = k_1 = 1$ and $y_4 = y_5 = y_6 = k_2 = 2$.

   Apply linear discriminant analysis by doing the following:

   a) Find estimates for the linear discriminant functions $\delta_1(x)$ and $\delta_2(x)$.

   b) Find the line that decides between the two classes.

   c) Classify the new point $x = (4, 5)$.

2. Suppose we have a set of data $(x_1, y_1), \dots, (x_6, y_6)$ as follows:

   $x_1 = (0, 0), x_2 = (1, 1), x_3 = (2, 3), x_4 = (2, 4), x_5 = (3, 2), x_6 = (4, 2)$ with

   $y_1 = y_2 = k_1 = 1, y_3 = y_4 = k_2 = 2$ and $y_5 = y_6 = k_3 = 3$.

   Apply linear discriminant analysis by doing the following:

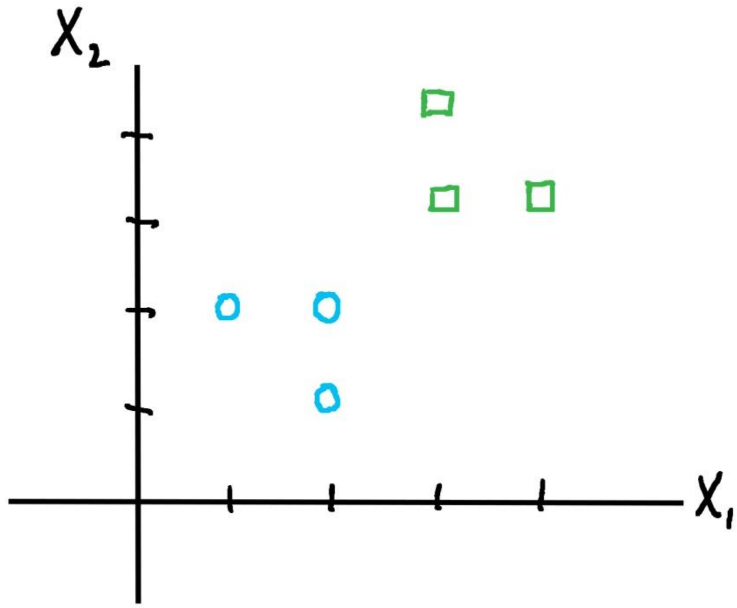
   a) Find estimates for the linear discriminant functions $\delta_1(x)$, $\delta_2(x)$ and $\delta_3(x)$.

   b) Find the lines that decide between each pair of classes.

   c) Classify the new point $x = (3, 0)$.

SOLUTION SET: LINEAR DISCRIMINANT ANALYSIS

1. Here is a graph of the data points:



The number of features p is 2, the number of classes K is 2, the total number of data points N is 6, the number N_1 of data points in class k_1 is 3, and the number N_2 of data points in class k_2 is 3.
First, we will find estimates for π_1 and π_2, the prior probabilities that $Y = k_1$ and $Y = k_2$, respectively.
Then, we will find estimates for μ_1 and μ_2, the class-specific mean vectors.
We can then calculate the estimate for the covariance matrix Σ.
Finally, using the estimates $\widehat{\pi_1}, \widehat{\pi_2}, \widehat{\mu_1}, \widehat{\mu_2}, \hat{\Sigma}$, we can find the estimates for the linear discriminant functions $\delta_1(x)$ and $\delta_2(x)$.

$$\widehat{\pi_1} = \frac{N_1}{N} = \frac{3}{6} = \frac{1}{2}$$
$$\widehat{\pi_2} = \frac{N_2}{N} = \frac{3}{6} = \frac{1}{2}$$
$$\widehat{\mu_1} = \frac{1}{N_1} \sum_{i:y_i=1} x_i = \frac{1}{3}[x_1 + x_2 + x_3] = \begin{bmatrix} \frac{5}{3} \\ \frac{5}{3} \end{bmatrix}$$

$$\widehat{\mu_2} = \frac{1}{N_2}\sum_{i:y_i=2} x_i = \frac{1}{3}[x_4 + x_5 + x_6] = \begin{bmatrix}\frac{10}{3}\\ \frac{10}{3}\end{bmatrix}$$

$$\hat{\Sigma} = \frac{1}{N-K}\sum_{k=1}^{K}\sum_{i:y_i=k}(x_i - \widehat{\mu_k})(x_i - \widehat{\mu_k})^T$$

$$= \frac{1}{6-2}\begin{bmatrix}12/9 & -6/9\\ -6/9 & 12/9\end{bmatrix} = \begin{bmatrix}1/3 & -1/6\\ -1/6 & 1/3\end{bmatrix}$$

$$\Longrightarrow \quad \hat{\Sigma}^{-1} = \begin{bmatrix}4 & 2\\ 2 & 4\end{bmatrix}$$

$$\begin{aligned}\widehat{\delta_1}(x) &= x^T\hat{\Sigma}^{-1}\widehat{\mu_1} - \frac{1}{2}\widehat{\mu_1}^T\hat{\Sigma}^{-1}\widehat{\mu_1} + \log\widehat{\pi_1}\\ &= x^T\begin{bmatrix}10\\ 10\end{bmatrix} - \frac{1}{2}\left(\frac{100}{3}\right) + \log\frac{1}{2}\\ &= 10X_1 + 10X_2 - \frac{50}{3} + \log\frac{1}{2}\end{aligned}$$

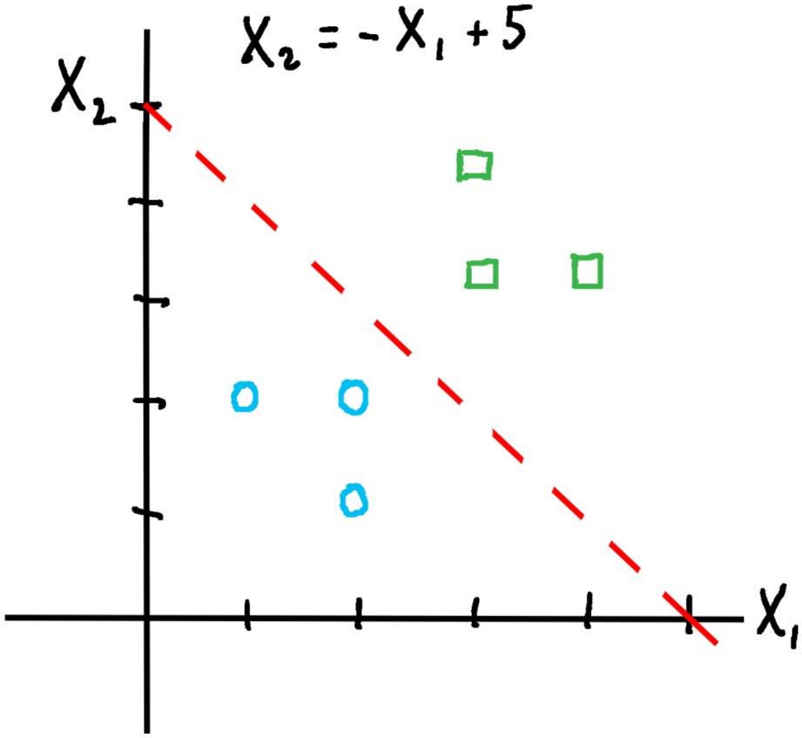
$$\begin{aligned}\widehat{\delta_2}(x) &= x^T\hat{\Sigma}^{-1}\widehat{\mu_2} - \frac{1}{2}\widehat{\mu_2}^T\hat{\Sigma}^{-1}\widehat{\mu_2} + \log\widehat{\pi_2}\\ &= x^T\begin{bmatrix}20\\ 20\end{bmatrix} - \frac{1}{2}\left(\frac{400}{3}\right) + \log\frac{1}{2}\\ &= 20X_1 + 20X_2 - \frac{200}{3} + \log\frac{1}{2}\end{aligned}$$

Setting $\widehat{\delta_1}(x) = \widehat{\delta_2}(x)$

$$\begin{aligned}&\Longrightarrow \quad 10X_1 + 10X_2 - \frac{50}{3} + \log\frac{1}{2} = 20X_1 + 20X_2 - \frac{200}{3} + \log\frac{1}{2}\\ &\Longrightarrow \quad \frac{150}{3} = 10X_1 + 10X_2\\ &\Longrightarrow \quad 50 = 10X_1 + 10X_2\\ &\Longrightarrow \quad 5 = X_1 + X_2\\ &\Longrightarrow \quad -X_1 + 5 = X_2\end{aligned}$$

So, the line that decides between the two classes is given by $X_2 = -X_1 + 5$.

Here is a graph of the decision line:



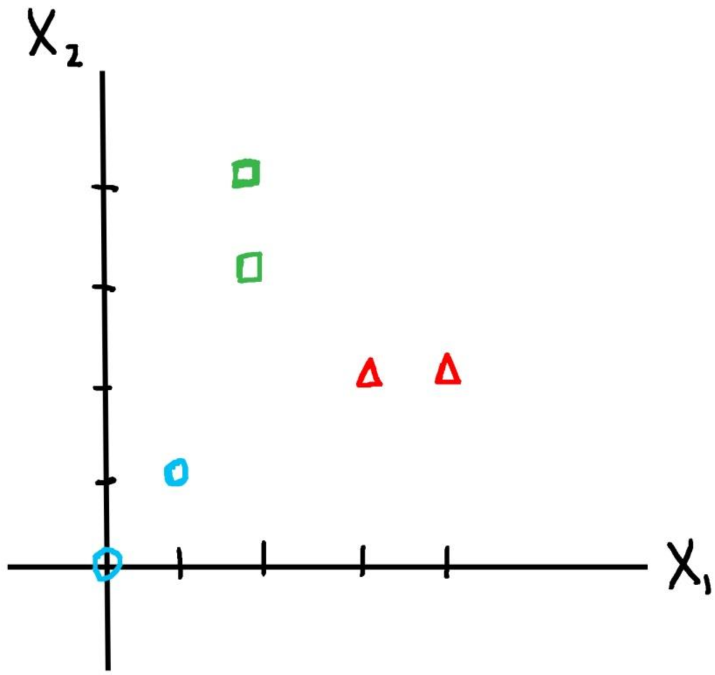
If $\widehat{\delta_1}(x) > \widehat{\delta_2}(x)$, then we classify x as of class k_1.

So if x is below the line $X_2 = -X_1 + 5$, then we classify x as of class k_1.

Conversely, if $\widehat{\delta_1}(x) < \widehat{\delta_2}(x)$, then we classify x as of class k_2. This corresponds to x being above the line $X_2 = -X_1 + 5$.

The point $(4, 5)$ is above the line; so we classify it as of class k_2.

2. Here is a graph of the data points:



The number of features p is 2, the number of classes K is 3, the total number of data points N is 6, the number N_1 of data points in class k_1 is 2, the number N_2 of data points in class k_2 is 2, and the number N_3 of data points in class k_3 is 2.

First, we will find estimates for π_1, π_2, π_3, the prior probabilities that $Y = k_1$, $Y = k_2$, $Y = k_3$, respectively.

Then, we will find estimates for μ_1, μ_2, μ_3, the class-specific mean vectors.

We can then calculate the estimate for the covariance matrix Σ.

Finally, using the estimates $\widehat{\pi_1}, \widehat{\pi_2}, \widehat{\pi_3}, \widehat{\mu_1}, \widehat{\mu_2}, \widehat{\mu_3}, \hat{\Sigma}$, we can find the estimates for the linear discriminant functions $\delta_1(x)$, $\delta_2(x)$, and $\delta_3(x)$.

$$\widehat{\pi_1} = \frac{N_1}{N} = \frac{2}{6} = \frac{1}{3}$$
$$\widehat{\pi_2} = \frac{N_2}{N} = \frac{2}{6} = \frac{1}{3}$$
$$\widehat{\pi_3} = \frac{N_3}{N} = \frac{2}{6} = \frac{1}{3}$$

$$\widehat{\mu_1} = \frac{1}{N_1}\sum_{i:y_i=1} x_i = \frac{1}{2}[x_1 + x_2] = \begin{bmatrix}1/2\\1/2\end{bmatrix}$$

$$\widehat{\mu_2} = \frac{1}{N_2}\sum_{i:y_i=2} x_i = \frac{1}{2}[x_3 + x_4] = \begin{bmatrix}2\\7/2\end{bmatrix}$$

$$\widehat{\mu_3} = \frac{1}{N_3}\sum_{i:y_i=3} x_i = \frac{1}{2}[x_5 + x_6] = \begin{bmatrix}7/2\\2\end{bmatrix}$$

$$\hat{\Sigma} = \frac{1}{N-K}\sum_{k=1}^{K}\sum_{i:y_i=k}(x_i - \widehat{\mu_k})(x_i - \widehat{\mu_k})^T$$

$$= \frac{1}{6-3}\begin{bmatrix}1 & 1/2\\1/2 & 1\end{bmatrix} = \begin{bmatrix}1/3 & 1/6\\1/6 & 1/3\end{bmatrix}$$

$$\Longrightarrow \quad \hat{\Sigma}^{-1} = \begin{bmatrix}4 & -2\\-2 & 4\end{bmatrix}$$

$$\begin{aligned}\widehat{\delta_1}(x) &= x^T\hat{\Sigma}^{-1}\widehat{\mu_1} - \frac{1}{2}\widehat{\mu_1}^T\hat{\Sigma}^{-1}\widehat{\mu_1} + \log\widehat{\pi_1}\\ &= x^T\begin{bmatrix}1\\1\end{bmatrix} - \tfrac{1}{2}(1) + \log\tfrac{1}{3}\\ &= X_1 + X_2 - \tfrac{1}{2} + \log\tfrac{1}{3}\end{aligned}$$

$$\begin{aligned}\widehat{\delta_2}(x) &= x^T\hat{\Sigma}^{-1}\widehat{\mu_2} - \frac{1}{2}\widehat{\mu_2}^T\hat{\Sigma}^{-1}\widehat{\mu_2} + \log\widehat{\pi_2}\\ &= x^T\begin{bmatrix}1\\10\end{bmatrix} - \tfrac{1}{2}(37) + \log\tfrac{1}{3}\\ &= X_1 + 10X_2 - \tfrac{37}{2} + \log\tfrac{1}{3}\end{aligned}$$

$$\begin{aligned}\widehat{\delta_3}(x) &= x^T\hat{\Sigma}^{-1}\widehat{\mu_3} - \frac{1}{2}\widehat{\mu_3}^T\hat{\Sigma}^{-1}\widehat{\mu_3} + \log\widehat{\pi_3}\\ &= x^T\begin{bmatrix}10\\1\end{bmatrix} - \tfrac{1}{2}(37) + \log\tfrac{1}{3}\\ &= 10X_1 + X_2 - \tfrac{37}{2} + \log\tfrac{1}{3}\end{aligned}$$

Setting $\widehat{\delta_1}(x) = \widehat{\delta_2}(x)$

$\Longrightarrow \quad X_1 + X_2 - \frac{1}{2} + \log\frac{1}{3} = X_1 + 10X_2 - \frac{37}{2} + \log\frac{1}{3}$

$\Longrightarrow \quad 18 = 9X_2$

$\Longrightarrow \quad 2 = X_2$

So, the line that decides between classes k_1 and k_2 is given by $X_2 = 2$.

Setting $\widehat{\delta_1}(x) = \widehat{\delta_3}(x)$

$\Longrightarrow \quad X_1 + X_2 - \frac{1}{2} + \log\frac{1}{3} = 10X_1 + X_2 - \frac{37}{2} + \log\frac{1}{3}$

$\Longrightarrow \quad 18 = 9X_1$

$\Longrightarrow \quad 2 = X_1$

So, the line that decides between classes k_1 and k_3 is given by $X_1 = 2$.

Setting $\widehat{\delta_2}(x) = \widehat{\delta_3}(x)$

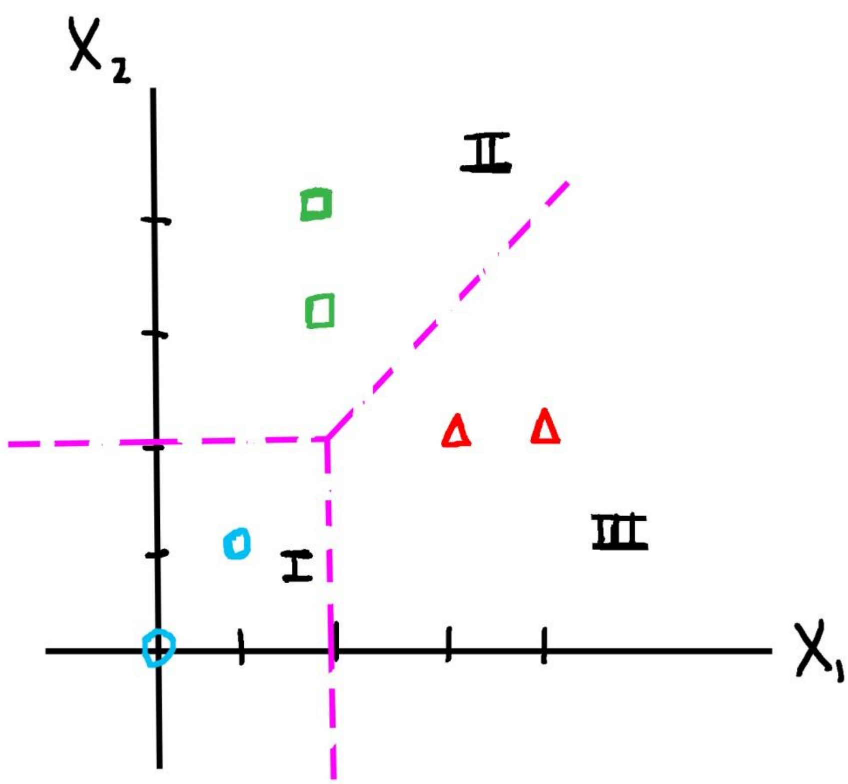
$\Longrightarrow \; X_1 + 10X_2 - \frac{37}{2} + \log\frac{1}{3} = 10X_1 + X_2 - \frac{37}{2} + \log\frac{1}{3}$

$\Longrightarrow \; 9X_2 = 9X_1$

$\Longrightarrow \; X_2 = X_1$

So, the line that decides between classes k_2 and k_3 is given by $X_2 = X_1$.

Here is a graph of the decision lines:



The lines divide the plane into 3 regions.
If x is in region I, then we classify x as of class k_1. Similarly, points in region II get classified as of k_2, and points in region III get classified as of k_3.
The point $(3, 0)$ is in region III; so we classify it as of class k_3.