Matemáticas del Aprendizaje Automático

Introducción a la analítica de datos e inteligencia artificial

Richard Han

CONTENIDO

PREFACIO

Bienvenido a Matemáticas del Aprendizaje Automático: Introducción a la analítica de datos e inteligencia artificial. Este es un texto introductorio en matemáticas para el Aprendizaje Automático. Asegúrese de obtener el curso complementario por medio del sitio web: www.onlinemathtraining.com. El curso en línea puede ser muy útil junto con este libro.

Los requisitos previos para este libro y el curso en línea son álgebra lineal, cálculo multivariable y probabilidad. Puedes encontrar mi curso en línea sobre Álgebra Lineal en el mismo sitio web.

No haremos ninguna programación en este libro.

Este libro le ayudará a comenzar con el Aprendizaje Automático de una manera suave y natural, preparándolo para temas más avanzados y disipando la creencia de que la analítica de datos e inteligencia artificial es complicado, difícil e intimidante.

Quiero que tengas éxito y prosperes en tu carrera, tu vida y tus futuros esfuerzos. Estoy aqui para ti. Visítame en: www.onlinemathtraining.com.

1 - INTRODUCCIÓN

Bienvenido a Matemáticas del Aprendizaje Automático: Introducción a la analítica de datos e inteligencia artificial Mi nombre es Richard Han. Este es un texto introductorio en matemáticas para el Aprendizaje Automático.

Estudiante ideal:

Si usted es un profesional que necesita un resumen sobre el Aprendizaje Automático o un principiante que necesita aprender Aprendizaje Automático por primera vez, este libro es para usted. Si su situación no le permite regresar a una escuela tradicional, este libro le permite estudiar según su propio horario y alcanzar sus metas profesionales sin quedarse atrás.

Si planea tomar el Aprendizaje Automático en la universidad, esta es una excelente manera de avanzar.

Si estás luchando con el Aprendizaje Automático o has luchado con él en el pasado, ahora es el momento de dominarlo.

Beneficios de estudiar este libro:

Después de leer este libro, habrá actualizado su conocimiento de la analítica de datos e inteligencia artificial para que pueda ganar un mejor salario.

Tendrá un requisito previo obligatorio para campos profesionales lucrativos, como la ciencia de datos y la inteligencia artificial.

Estará en una mejor posición para obtener una maestría o un doctorado en Aprendizaje Automático y ciencia de la información.

¿Por qué el Aprendizaje Automático es importante?:

- Los usos famosos del Aprendizaje Automático incluyen:
    - Análisis discriminante lineal. El análisis discriminante lineal puede utilizarse para resolver problemas de clasificación, como el filtrado de spam y la clasificación de enfermedades del paciente.
    - Regresión logística. La regresión logística se puede usar para resolver problemas de

clasificación binaria, como determinar si un paciente tiene cierta forma de cáncer o no.

- Redes neuronales artificiales. Las redes neuronales artificiales se pueden usar para aplicaciones tales como autos de conducción automática, sistemas de recomendación, mercadeo en línea, lectura de imágenes médicas, habla y reconocimiento facial.
- Máquinas de vectores de soporte (SVM). Las aplicaciones de los SVM incluyen la clasificación de proteínas y la clasificación de imágenes.

Lo que mi libro ofrece:

En este libro, cubro temas principales como:

- **Regresión Lineal**
- **Análisis Discriminante Lineal**
- **Regresión Logística**
- **Redes neuronales artificiales**
- **Máquinas de vectores de soporte**

Explico cada definición y completo cada ejemplo paso a paso para que entienda cada tema con claridad. A lo largo del libro, hay ejercicios para que practiques. Se proporcionan soluciones detalladas después de cada conjunto de ejercicios.

Espero que te beneficies del libro.

Atentamente,
Richard

2 – REGRESIÓN LINEAL

REGRESIÓN LINEAL

Supongamos que tenemos un conjunto de datos $(x_1, y_1), \dots, (x_N, y_N)$. Esto se llama los datos de entrenamiento.

Cada x_i es un vector $\begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix}$ de medidas, donde x_{i1} es una instancia del primer variable de entrada X_1, x_{i2} es una instancia del segundo variable de entrada X_2, etc. $X_1, \dots, X_p$ se conocen como ***caracteristicas*** or ***predictores***.

$y_1, \dots, y_N$ son instancias del variable de salida Y, que se conoce como la ***respuesta***.

En regresión lineal, suponemos que la respuesta depende de las variables de entrada de forma lineal: $y = f(X) + \varepsilon$, donde $f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$.

Aquí, ε se conoce como el ***término de error*** y $\beta_0, \dots, \beta_p$ se conoce como ***parámetros.***

No sabemos los valores de $\beta_0, \dots, \beta_p$. Pero podemos usar los datos de entrenamiento para aproximar los valores de $\beta_0, \dots, \beta_p$. Lo que haremos es mirar la cantidad por la cual el valor predicho $f(x_i)$ se difiere de la cantidad actual y_i para cada par $(x_1, y_1), \dots, (x_N, y_N)$ de los datos de entrenamiento. Asi que tenenmos $y_i - f(x_i)$ como la diferencia. Luego cuadramos esto y tomamos la suma para $i = 1, \dots, N$:

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

Esto se llama la ***suma residual de cuadrados*** y se denota como $RSS(\beta)$ donde $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$.

Queremos que la suma de cuadrados residual sea los más pequeña possible. Esencialmente, esto significa que queremos nuestro valor predicho $f(x_i)$ que sea los más cercano al valor real y_i posible, por cada uno de los pares (x_i, y_i). Hacer esto nos dará una función lineal de las variables de entrada que mejor se adapten a los datos de entrenamiento. En el caso de una sola variable de entrada, obtenemos la mejor línea de ajuste. En el caso de dos variables de entrada, obtenemos el mejor plano de ajuste. Y así sucesivamente, para dimensiones más altas.

METODO DE LOS MÍNIMOS CUADRADOS

Minimizando $RSS(\beta)$, podemos obtener estimaciones $\widehat{\beta_0}, \widehat{\beta_1}, \dots, \widehat{\beta_p}$ para los parámetros $\beta_0, \dots, \beta_p$. Este metodo se llama el ***metodo de los mínimos cuadrados***.

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

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

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

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

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

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

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

Considera el vector de derivadas parciales de $RSS(\beta)$:

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

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

Tomemos la derivada parcial con respecto a β_0.

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

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

Después, toma la derivada parcial con respecto a β_1.

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

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

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

Asi que,

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

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

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

Si tomamos la segunda derivada de $RSS(\beta)$, que es $\frac{\partial^2 RSS(\beta)}{\partial \beta_k \partial \beta_j}$, obtenemos

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

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

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

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

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

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

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

$\Longrightarrow$ La matriz de segundas derivadas de $RSS(\beta)$ es $2X^TX$. Ésta matriz se llama la ***matriz hessiana***. Por la segunda prueba derivada, si la matriz hessiana de $RSS(\beta)$ en un punto crítico es positivo definitivamente, entonces $RSS(\beta)$ tiene un mínimo local allí.

Si configuramos nuestro vector de derivados a $\mathbf{0}$, obtenemos

$$-2X^T(\boldsymbol{y} - X\beta) = \mathbf{0}$$
$$\Longrightarrow \quad -2X^T\boldsymbol{y} + 2X^TX\beta = \mathbf{0}$$
$$\Longrightarrow \quad 2X^TX\beta = 2X^T\boldsymbol{y}$$
$$\Longrightarrow \quad X^TX\beta = X^T\boldsymbol{y}$$
$$\Longrightarrow \quad \beta = (X^TX)^{-1}X^T\boldsymbol{y}.$$

Así, resolvimos para el vector de parámetros $\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$ que minimiza la suma residual de cuadrados $RSS(\beta)$.

Entonces dejamos que $\begin{bmatrix} \widehat{\beta_0} \\ \widehat{\beta_1} \\ \vdots \\ \widehat{\beta_p} \end{bmatrix} = (X^TX)^{-1}X^T\boldsymbol{y}$.

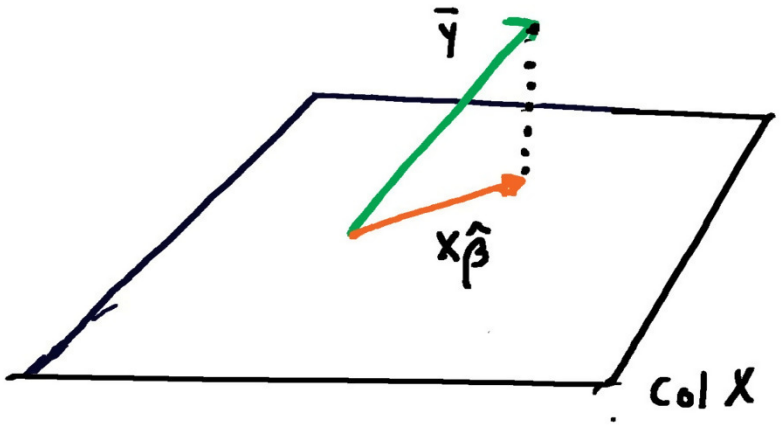
SOLUCIÓN USANDO ÁLGEBRA LINEAL PARA MÍNIMOS CUADRADOS

Podemos llegar a la misma solución para el problema de mínimos cuadrados utilizando álgebra lineal. Deja que $X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & & \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix}$ y deja que $\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$ como antes, de nuestro datos de entrenamiento. Queremos un vector β donde $X\beta$ es cercano a $\boldsymbol{y}$. En otras palabras, queremos un vector β tal que la distancia $\|X\beta - \boldsymbol{y}\|$ entre $X\beta$ y entre $\boldsymbol{y}$ está minimizado. Un vector β que minimiza $\|X\beta - \boldsymbol{y}\|$ se llama una ***solución de mínimos cuadrados*** de $X\beta = \boldsymbol{y}$.

X es una matriz con dimensiones N por $(p + 1)$. Queremos un $\hat{\beta}$ en $\mathbb{R}^{p+1}$ tal que $X\hat{\beta}$ es el más cercano a $\boldsymbol{y}$. Nota que $X\hat{\beta}$ es una combinación lineal de las columnas de X. Entonces $X\hat{\beta}$ se encuentra en el lapso de las columnas de X, que es un subespacio de $\mathbb{R}^N$ denotado como $Col\ X$. Entonces

queremos el vector en $Col\ X$ que es más cercano a $\boldsymbol{y}$. La proyección de $\boldsymbol{y}$ en el subespacio $Col\ X$ es esl vector.

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



Considera $\boldsymbol{y} - X\hat{\beta}$. Nota que $\boldsymbol{y} = X\hat{\beta} + (\boldsymbol{y} - X\hat{\beta})$.

$\mathbb{R}^N$ se puede dividir en dos subespacios $Col\ X$ y $(Col\ X)^{\perp}$, donde $(Col\ X)^{\perp}$ es el subespacio de $\mathbb{R}^N$ que consiste en todos los vectores que son ortogonales a los vectores en $Col\ X$. Cualquier vector en $\mathbb{R}^N$ puede ser escrito únicamente como $\boldsymbol{z} + \boldsymbol{w}$ donde $\boldsymbol{z} \in Col\ X$ y $\boldsymbol{w} \in (Col\ X)^{\perp}$.

Ya que $\boldsymbol{y} \in \mathbb{R}^N$, y $\boldsymbol{y} = X\hat{\beta} + (\boldsymbol{y} - X\hat{\beta})$, con $X\hat{\beta} \in Col\ X$, el segundo vector $\boldsymbol{y} - X\hat{\beta}$ debe estar en $(Col\ X)^{\perp}$.

$\Longrightarrow$ $\boldsymbol{y} - X\hat{\beta}$ es ortogonal a las columnas de X.

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

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

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

Así, resulta que el conjunto de soluciones de mínimos cuadrados de $X\beta = \boldsymbol{y}$ Consiste en todas y solo las soluciones a la ecuación matricial $X^TX\beta = X^T\boldsymbol{y}$.

Sí X^TX es positive por seguro, entonces los valores propios de X^TX son todos positivos. Así, 0 no es un valor propio de X^TX. Resulta que X^TX es invertible. Entonces, podemos resolver la ecuación $X^TX\hat{\beta} = X^T\boldsymbol{y}$ por $\hat{\beta}$ para obtener $\hat{\beta} = (X^TX)^{-1}X^T\boldsymbol{y}$, que es el mismo resultado que obtuvimos antes usando el cálculo multivariable.

EJEMPLO: REGRESIÓN LINEAL

Supongamos que tenemos los siguientes datos de entrenamiento:

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

Encuentra la mejor línea de ajuste usando el método de mínimos cuadrados Encuentra el valor predicho para $x = 4$.

Solución:

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

Los coeficientes β_0, β_1 para la mejor línea de ajuste $f(x) = \beta_0 + \beta_1 x$ son dados por $\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = (X^T X)^{-1} X^T \boldsymbol{y}$.

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

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

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

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

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

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

Así, la mejor línea de ajuste está dada por $f(x) = \left(\frac{3}{2}\right) x$.

El valor predicho para $x = 4$ es $f(4) = \left(\frac{3}{2}\right) \cdot 4 = 6$.

RESUMEN: REGRESIÓN LINEAL

- En el método de mínimos cuadrados, buscamos una función lineal de las variables de entrada que mejor se adapte a los datos de entrenamiento dados. Hacemos esto minimizando la suma residual de cuadrados.

- Para minimizar la suma de cuadrados residual, aplicamos la segunda prueba derivada del cálculo multivariable.

- Podemos llegar a la misma solución para el problema de los mínimos cuadrados utilizando álgebra lineal.

EJERCICIOS: REGRESIÓN LINEAL

1. Supongamos que tenemos los siguientes datos de entrenamiento:

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

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

   Encuentra la mejor línea de ajuste usando el método de mínimos cuadrados. Encuentra el valor predicho para $x = 4$.

2. Supongamos que tenemos los siguientes datos de entrenamiento:

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

   Encuentra el plano de mejor ajuste usando el método de mínimos cuadrados. Encuentra el valor predicho para $x = \begin{bmatrix}2\\2\end{bmatrix}$.

SOLUCIÓN: REGRESIÓN LINEAL

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

Los coeficientes β_0, β_1 para la mejor línea de ajuste $f(x) = \beta_0 + \beta_1 x$ son dados por $\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = (X^T X)^{-1} X^T \boldsymbol{y}$.

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

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

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

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

$$\Longrightarrow \; \beta_0 = \tfrac{14}{10} \quad \text{y} \quad \beta_1 = \tfrac{9}{10}.$$

Así, la mejor línea de ajuste está dada por

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

El valor predicho para $x = 4$ es $f(4) = \frac{14}{10} + \frac{9}{10} \cdot 4 = 5$.

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

Los coeficientes $\beta_0, \beta_1, \beta_2$ para la mejor línea de ajuste$f(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ son dados por $\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = (X^T X)^{-1} X^T \boldsymbol{y}$.

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

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

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

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

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

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

Así, el mejor plano de ajuste está dado por

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

El valor predicho para $x = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$ es $f(2,2) = 2\frac{1}{4}$.

3 – ANÁLISIS DISCRIMINANTE LINEAL

CLASIFICACIÓN

En el problema de la regresión, teníamos un conjunto de datos $(x_1, y_1), \dots, (x_N, y_N)$ y queríamos predecir los valores para la variable de respuesta Y para los nuevos datos. Los valores que toma Y fueron valores numericos y cuantitativos. En ciertos problemas, los valores para la variable de respuesta Y que queremos predecir no son cuantitativos sino cualitativos. Así que los valores para Y tomará los valores de un conjunto finito de clases o categorías. Problemas de este tipo se conocen como ***problemas de clasificacion***. Algunos ejemplos de un problema de clasificación son clasificar un correo electrónico como spam o no spam y clasificar la enfermedad de un paciente como uno de entre un número finito de enfermedades.

ANÁLISIS DISCRIMINANTE LINEAL (LDA)

Un método para resolver un problema de clasificación se llama ***análisis discriminante lineal***.

Lo que haremos es estimar $\Pr(Y = k|X = x)$, la probabilidad que Y es la clase k dado que la variable de entrada Xes$x$. Una vez que tenemos todas estas probabilidades para un x fijo, escojemos la clase k para lo cual la probabilidad $\Pr(Y = k|X = x)$ es más grande. Entonces clasificamos x como la clase k.

LAS FUNCIONES DE PROBABILIDAD POSTERIOR

En esta sección, construiremos una fórmula para la probabilidad posterior $\Pr(Y = k|X = x)$.

Deja que $\pi_k = \Pr(Y = k)$, la probabilidad previa de que $Y = k$.

Deja que $f_k(x) = \Pr(X = x|Y = k)$, la probabilidad que $X = x$, dado que $Y = k$.

Por la regla Bayes,

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

Aquí suponemos que k puede asumir los valores $1, \dots, K$.

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

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

Podemos pensar en $\Pr(Y = k|X = x)$ como una función de x y denotarlo como $p_k(x)$.

Entonces $p_k(x) = \frac{\pi_k \cdot f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$. Recuerda que $p_k(x)$ es la probabilidad posterior de que $Y = k$ dado

que $X = x$.

MODELANDO LAS FUNCIONES DE PROBABILIDAD POSTERIOR

Recuerda que queríamos estimar $\Pr(Y = k|X = x)$ por cualquier x. Es decir, queremos una estimación para $p_k(x)$. Si podemos obtener estimaciones para $\pi_k, f_k(x), \pi_l$ y para $f_l(x)$ por cada $l = 1, \dots, K$, entonces tendríamos un estimado para $p_k(x)$.

Digamos que $X = (X_1, X_2, \dots, X_p)$ donde $X_1, \dots, X_p$ son las variables de entrada. Así que los valores de X serán vectores de p elementos.

Supondremos que la distribución condicional de X dado por $Y = k$ es la distribución gaussiana multivariable $N(\mu_k, \Sigma)$, donde μ_k es un vector medio específico de clase y Σ es la covarianza de X.

El vector medio específico de clase μ_k está dada por el vector de los medios específicos de la clase $\begin{bmatrix} \mu_{k1} \\ \vdots \\ \mu_{kp} \end{bmatrix}$, donde μ_{kj} es el medio específico de la clase X_j.

Entonces $\mu_{kj} = \sum_{i:y_i=k} x_{ij} \Pr(X_j = x_{ij})$. Recuerda que $x_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}$. (Para todos x_i por cual $y_i = k$, estamos tomando el medio de su jth componentes.)

Σ, la matriz de covarianza de X, está dada por la matriz de covarianzas de X_i y de X_j.

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

La densidad gaussiana multivariable está dada por

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

para la distribución gaussiana multivariable distribution $N(\mu, \Sigma)$.

Dado que estamos asumiendo que la distribución condicional de X dado que $Y = k$ es la distribución gaussiana multivariable $N(\mu_k, \Sigma)$, tenemos que

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

Recuerda que $f_k(x) = \Pr(X = x|Y = k)$.

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

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

Conectando lo que tenemos para $f_k(x)$, tenemos que

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

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

Tenga en cuenta que el denominador es $(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}\sum_{l=1}^{K}\pi_l f_l(x)$ y que

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

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

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

Así que el denominador es justo $(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}\Pr(X = x)$.

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

FUNCIONES LINEALES DISCRIMINANTES

Recordemos que queremos elegir la clase k para lo cual la probabilidad posterior $p_k(x)$ es más grande. Dado que la función de logaritmo conserva la orden, maximizando $p_k(x)$ es igual a maximizando $\log p_k(x)$.

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

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

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

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

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

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

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

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

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

$= \mu_k^T\Sigma^{-1}x$ porque Σ es simétrico.

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

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

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

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

$\delta_k(x)$ se conoce como la ***funcion lineal discriminante***. Maximizando $\log p_k(x)$ es igual a maximizando $\delta_k(x)$ porque $-\frac{1}{2}x^T\Sigma^{-1}x - \log C$ no depende en k.

ESTIMACIÓN DE LAS FUNCIONES DISCRIMINANTES LINEALES

Ahora, si podemos encontrar estimaciones para $\pi_k, \mu_k,$ y Σ, entonces tendríamos un estimado para $p_k(x)$ y por lo tanto para $\log p_k(x)$ y para $\delta_k(x)$.

En un intento par maximizar $p_k(x)$, en su lugar maximizamos la estimación de $p_k(x)$, que es lo

mismo que maximizar la estimación de $\delta_k(x)$.

π_k puede ser estimando como $\widehat{\pi_k} = \frac{N_k}{N}$ donde N_k es el número de datos de entrenamiento en la clase k y N es el número total de datos de entrenamiento.

Recuerda que $\pi_k = \Pr(Y = k)$. Estamos estimando esto simplemente tomando la proporción de puntos de datos en la clase k.

El vector medio específico de la case $\mu_k = \begin{bmatrix} \mu_{k1} \\ \vdots \\ \mu_{kp} \end{bmatrix}$, donde $\mu_{kj} = \sum_{i:y_i=k} x_{ij} \Pr(X_j = x_{ij})$.

Podemos estimar μ_{kj} como $\frac{1}{N_k}\sum_{i:y_i=k} x_{ij}$.

Así podemos estimar μ_k como $\widehat{\mu_k} = \begin{bmatrix} \frac{1}{N_k}\sum_{i:y_i=k} x_{i1} \\ \vdots \\ \frac{1}{N_k}\sum_{i:y_i=k} x_{ip} \end{bmatrix} = \frac{1}{N_k}\begin{bmatrix} \sum_{i:y_i=k} x_{i1} \\ \vdots \\ \sum_{i:y_i=k} x_{ip} \end{bmatrix}$

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

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

En otros sentidos, $\widehat{\mu_k} = \frac{1}{N_k}\sum_{i:y_i=k} x_i$. Estimamos el vector medio específico de la case por el vector de promedios de cada componente sobre todo los x_i el la clase k.

Finalmente, la matriz de covarianza Σ se estimada como $\hat{\Sigma} = \frac{1}{N-K}\sum_{k=1}^{K}\sum_{i:y_i=k}(x_i - \widehat{\mu_k})\,(x_i - \widehat{\mu_k})^T$.

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

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

Nota que $\hat{\Sigma}, \widehat{\mu_k}$, y $\widehat{\pi_k}$ solo dependen de los datos de entrenamiento y no de x. Nota que x es un vector y nota que $x^T\hat{\Sigma}^{-1}\widehat{\mu_k}$ es una combinación lineal de los componentes de x. Así que, $\widehat{\delta_k}(x)$ es una combinación lineal de los componentes de x. Por eso se llama la función discriminante lineal.

CLASIFICACIÓN DE DATOS USANDO FUNCIONES DISCRIMINANTES

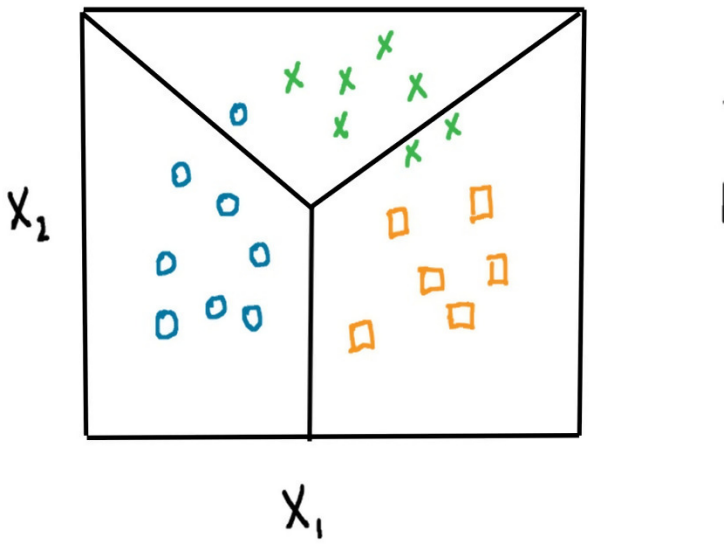
Si (k_1, k_2) es un par de clases, podemos considerar si $\widehat{\delta_{k_1}}(x) > \widehat{\delta_{k_2}}(x)$. Si es así, sabemos x no está en la clase de k_2. Después, podemos comparer si $\widehat{\delta_{k_1}}(x) > \widehat{\delta_{k_3}}(x)$ y descartar otra clase. Una vez que hayamos buscado todas las clases, sabremos qué clase x debe estar.

Ajustando $\widehat{\delta_{k_1}}(x) = \widehat{\delta_{k_2}}(x)$, nos da

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

Esto nos da un hiperplano en $\mathbb{R}^p$ que separa la clase $k_1$de la clase k_2.

Si encontramos el hiperplano de separación para cada par de clases, obtenemos algo como esto:



En este ejemplo, $p = 2$ and $K = 3$.

LDA EJEMPLO 1

Supongamos que tenemos un conjunto de datos $(x_1, y_1), \ldots, (x_6, y_6)$ como sigue:

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

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

Aplica el análisis discriminante lineal haciendo lo siguiente:

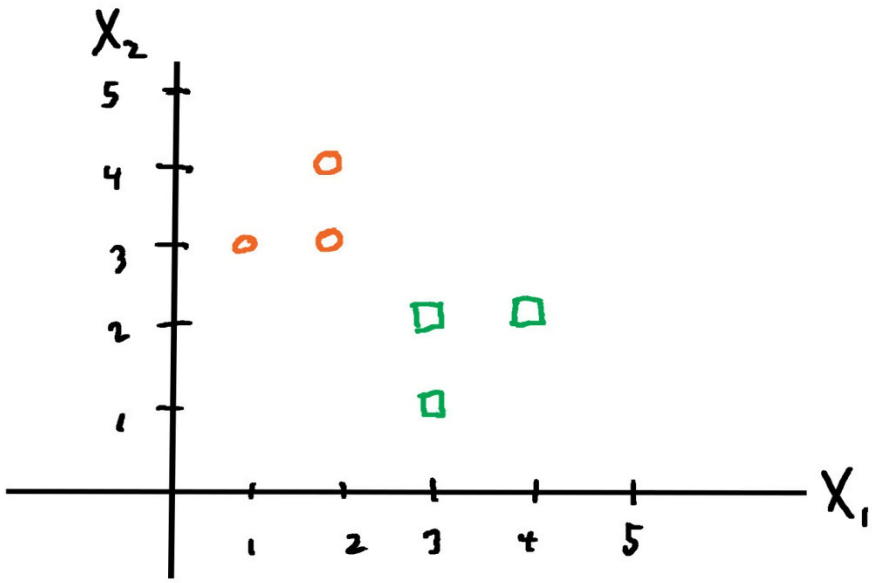
a) Encuentra estimaciones para las funciones discriminantes lineales $\delta_1(x)$ y $\delta_2(x)$.

b) Encuentra la línea que decide entre las dos clases.

c) Classifica el nuevo dato $x = (5,0)$.

Solución:

Aquí hay una gráfica de los puntos de datos:



El número de características de p es 2, el número de clases de K es 2, el número total de puntos de datos N es 6, el número N_1 de los datos de la clase k_1 es 3, y el número N_2 de los datos de la clase k_2 es 3.

Primero, encontraremos estimaciones para π_1 y π_2, las probabilidades previas de que $Y = k_1$ y de que $Y = k_2$, respectivamente.

Después, encontraremos estimaciones para μ_1 y para μ_2, los vectores medios específicos de la clase.

Luego podemos calcular la estimación de la matriz de covarianza Σ.

Finalmente, utilizando las estimaciones $\widehat{\pi_1}, \widehat{\pi_2}, \widehat{\mu_1}, \widehat{\mu_2}, \hat{\Sigma}$, podemos encontrar las estimaciones para las funciones discriminantes lineales $\delta_1(x)$ y $\delta_2(x)$.

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

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

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

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

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

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

Utilizando lo que agarramos para $\widehat{\mu_1}$ y $\widehat{\mu_2}$, tenemos

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

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

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

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

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

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

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

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

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

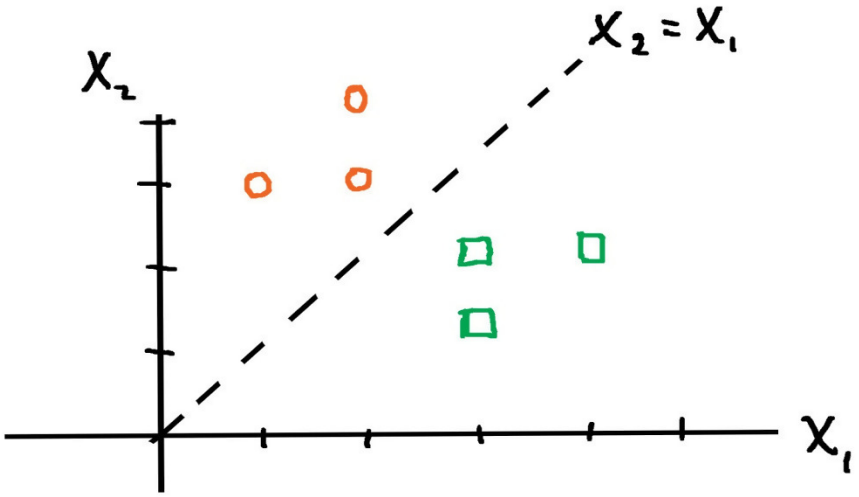
$$\Longrightarrow \quad 10X_2 - \tfrac{50}{3} + log\,\tfrac{1}{2} = 10X_1 - \tfrac{50}{3} + log\,\tfrac{1}{2}$$

$$\Longrightarrow \quad 10X_2 = 10X_1$$

$$\Longrightarrow \quad X_2 = X_1.$$

Entonces, la línea que decide entre las dos clases está dada por $X_2 = X_1$.

Aquí hay un gráfico de la línea decisiva:



Si $\widehat{\delta_1}(x) > \widehat{\delta_2}(x)$, entonces clasificamos x como una clase de k_1. Así que si x está arriba de la línea $X_2 = X_1$, clasificamos x como una clase de k_1. A la inversa, si $\widehat{\delta_1}(x) < \widehat{\delta_2}(x)$, clasificamos x como una clase de k_2. Esto corresponde a que x esté debajo de la línea $X_2 = X_1$.

El punto $(5, 0)$ está debajo de la línea, entonces lo clasificamos como clase de k_2.

LDA EJEMPLO 2

Supongamos que tenemos un conjunto de datos $(x_1, y_1), \dots, (x_6, y_6)$ como sigue:

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

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

Aplica el análisis discriminante lineal haciendo lo siguiente:

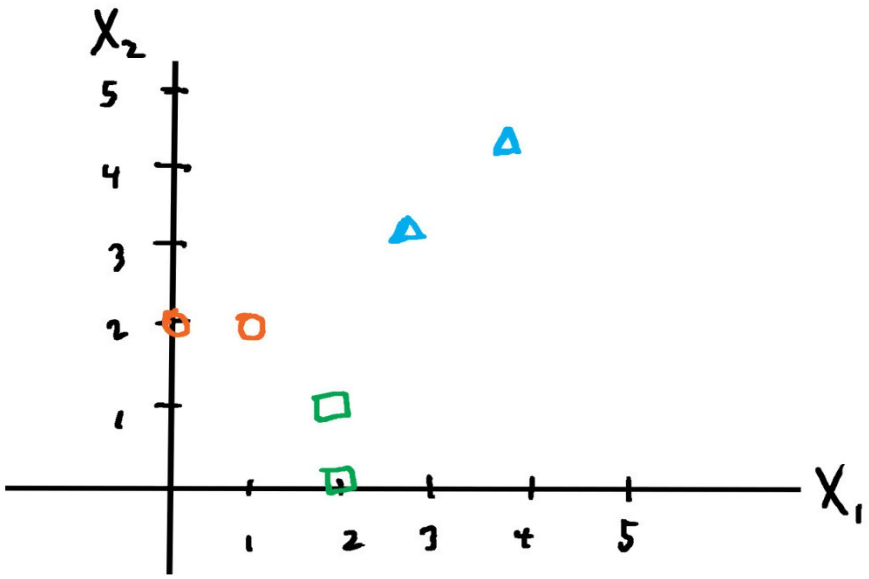
a) Encontrar estimaciones para las funciones discriminantes lineales $\delta_1(x), \delta_2(x)$, y $\delta_3(x)$.

b) Encuentra las líneas que deciden entre cada par de clases.

c) Clasifica un nuevo punto $x = (1, 3)$.

Solución:

Aquí hay una gráfica de los puntos de datos:



El número de características p es 2, el número de clases K es 3, el número total de puntos de datos N es 6, el número N_1 de datos en la clase k_1 es 2, el número N_2 de datos en la clase k_2 es 2, y el número N_3 de datos en la clase k_3 es 2.

Primero, encontraremos estimaciones para π_1, π_2, π_3, las probabilidades previas de que $Y = k_1$, $Y = k_2$, $Y = k_3$, respectivamente.

Después, encontraremos estimaciones para μ_1, μ_2, μ_3, los vectores medios específicos de la clase.

Luego podemos calcular la estimación de la matriz de covarianza Σ.

Finalmente, usando las estimaciones $\widehat{\pi_1}, \widehat{\pi_2}, \widehat{\pi_3}, \widehat{\mu_1}, \widehat{\mu_2}, \widehat{\mu_3}, \hat{\Sigma}$, podemos encontrar las estimaciones para las funciones discriminantes lineales $\delta_1(x), \delta_2(x)$, y para $\delta_3(x)$.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\Longrightarrow \qquad -2X_1 + 7X_2 = 7X_1 - 2X_2$$

$$\Longrightarrow \qquad 9X_2 = 9X_1$$

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

Entonces, la línea que decide entre las clases k_1 y k_2 está dada por $X_2 = X_1$.

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

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

$$\Longrightarrow \qquad 18 = 9X_1$$

$$\Longrightarrow \qquad X_1 = 2$$

Entonces, la línea que decide entre las clases k_1 y k_3 está dada por $X_1 = 2$.

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

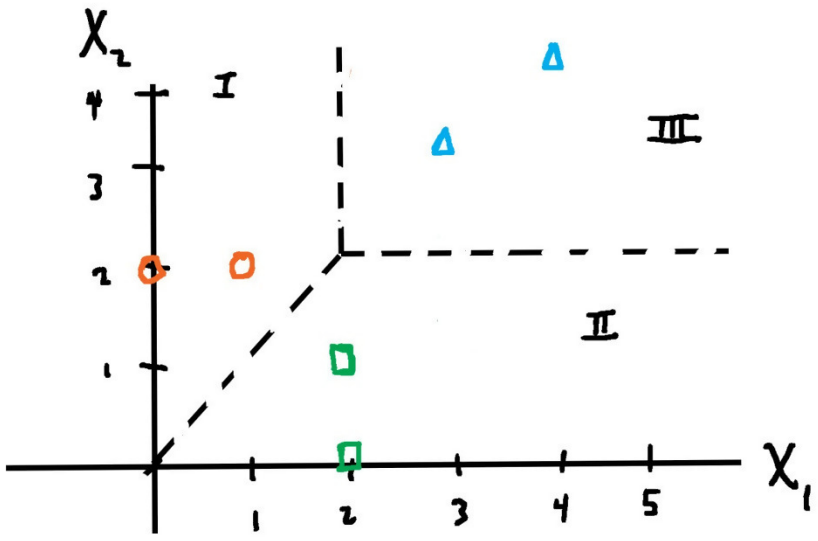
$$\Longrightarrow \qquad 7X_1 - 2X_2 - \frac{13}{2} + log\frac{1}{3} = 7X_1 + 7X_2 - \frac{49}{2} + log\frac{1}{3}$$

$$\Longrightarrow \qquad 18 = 9X_2$$

$$\Longrightarrow \qquad X_2 = 2$$

Entonces, la línea que decide entre las clases k_2 y k_3 está dada por $X_2 = 2$.

Aquí hay una gráfica de las líneas decisivas:



Las líneas dividen el plano en 3 regiones.

$\widehat{\delta_1}(x) > \widehat{\delta_2}(x)$ corresponde a la región arriba de la línea $X_2 = X_1$. Al contrario, $\widehat{\delta_1}(x) < \widehat{\delta_2}(x)$ corresponde a la región debajo de la línea $X_2 = X_1$.

$\widehat{\delta_1}(x) > \widehat{\delta_3}(x)$ corresponde a la región izquierda de la línea $X_1 = 2$. Al contrario, $\widehat{\delta_1}(x) < \widehat{\delta_3}(x)$

corresponde a la región derecha de la línea $X_1 = 2$.

$\widehat{\delta_2}(x) > \widehat{\delta_3}(x)$ corresponde a la región debajo de la línea $X_2 = 2$. Al contrario, $\widehat{\delta_2}(x) < \widehat{\delta_3}(x)$ corresponde a la región arriba de la línea $X_2 = 2$.

Si $\widehat{\delta_1}(x) > \widehat{\delta_2}(x)$ y $\widehat{\delta_1}(x) > \widehat{\delta_3}(x)$, podemos clasificar x como una clase de k_1. Así que si x está en la region I, podemos clasificar x como una clase de k_1. Al contrario, si x está el la region II, Podemos clasificar x como una clase de k_2 y si x está en la region III, podemos clasificar x como una clase de k_3.

El punto $(1, 3)$ está en la region I, entonces se clasifica como clase de k_1.

RESUMEN: ANÁLISIS DISCRIMINANTE LINEAL

- En análisis discriminante lineal, encontramos estimaciones $\widehat{p_k}(x)$ para la probabilidad posterior $p_k(x)$ que $Y = k$ dado que $X = x$. Nosotros clasificamos x segun la clase k que da la mayor probabilidad posterior estimada $\widehat{p_k}(x)$.

- Maximizando la probabilidad posterior estimada $\widehat{p_k}(x)$ es equivalente a maximizar el logartimo de $\widehat{p_k}(x)$, cual, a su vez, es equivalente a maximizar la función discriminante lineal estimada $\widehat{\delta_k}(x)$.

- Encontramos estimaciones de la probabilidad previa π_k que $Y = k$, de los vectores medios específicos de la clase μ_k, y de la matriz de covarianza Σ para estimar las funciones discriminantes lineales $\delta_k(x)$.

- Configurando $\widehat{\delta_k}(x) = \widehat{\delta_{k'}}(x)$ para cada par (k, k') de clases, tenemos hiperplanos en $\mathbb{R}^p$ que, juntos, divide $\mathbb{R}^p$ en regiones correspondientes a las distintas clases.

- Clasificamos x segun la clase k por cual $\widehat{\delta_k}(x)$ es más grande.

EJERCICIOS: ANÁLISIS DISCRIMINANTE LINEAL

1. Supongamos que tenemos un conjunto de datos $(x_1, y_1), \dots, (x_6, y_6)$ como sigue:

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

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

   Aplica el análisis discriminante lineal haciendo lo siguiente:

   a) Encuentra estimaciones para las funciones discriminantes lineales $\delta_1(x)$ y $\delta_2(x)$.

   b) Encuentra la línea que decide entre las dos clases.

   c) Clasifica un nuevo punto $x = (4,5)$.

2. Supongamos que tenemos un conjunto de datos $(x_1, y_1), \dots, (x_6, y_6)$ como sigue:

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

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

   Aplica el análisis discriminante lineal haciendo lo siguiente:

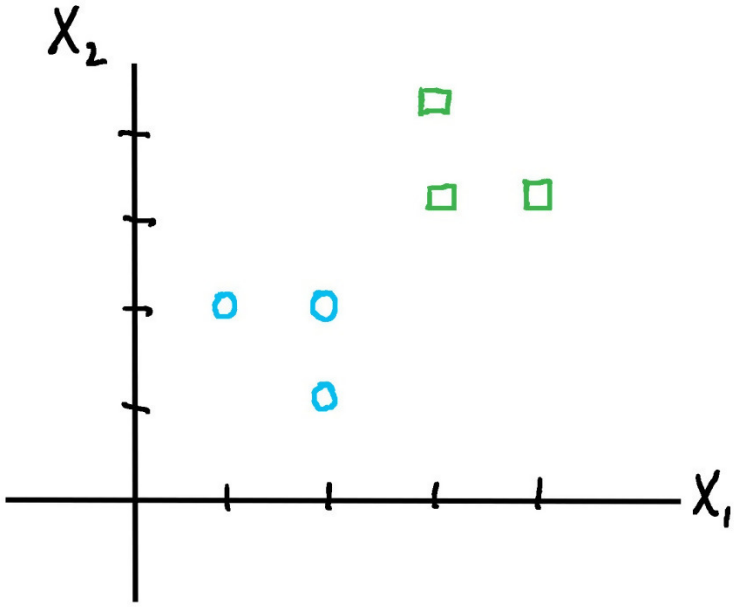
   a) Encuentra estimaciones para las funciones discriminantes lineales $\delta_1(x)$, $\delta_2(x)$ y $\delta_3(x)$.

   b) Encuentra las líneas que deciden entre cada par de clases.

   c) Clasifica un nuevo punto $x = (3,0)$.

SOLUCIONES: ANÁLISIS DISCRIMINANTE LINEAL

1. Aquí hay una gráfica de los puntos de datos:



El número de características p es 2, el número de clases K es 2, el número total de puntos de datos N es 6, el número N_1 de datos en la clase k_1 es 3, y el número N_2 de datos en la clase k_2 es 3.

Primero, encontraremos estimaciones para π_1 y π_2, las probabilidades previas de que $Y = k_1$ y $Y = k_2$, respectivamente.

Después, encontraremos estimaciones para μ_1 y μ_2, los vectores medios específicos de la clase.

Luego podemos calcular la estimación de la matriz de covarianza Σ.

Finalmente, utilizando las estimaciones $\widehat{\pi_1}, \widehat{\pi_2}, \widehat{\mu_1}, \widehat{\mu_2}, \hat{\Sigma}$, podemos encontrar las estimaciones para las funciones discriminantes lineales $\delta_1(x)$ y $\delta_2(x)$.

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

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

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

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

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

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

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

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

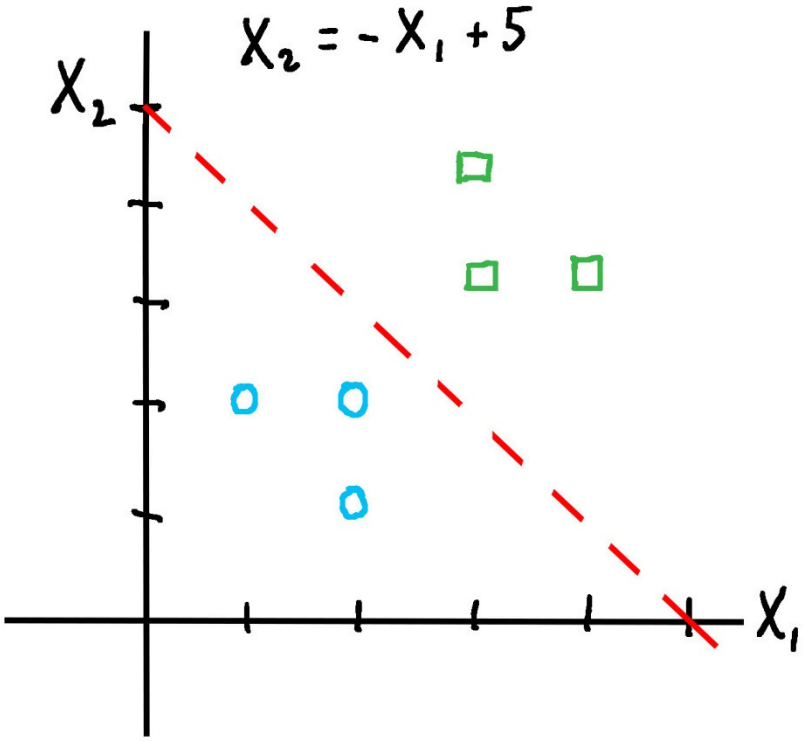
$$\widehat{\delta_2}(x) = x^T\hat{\Sigma}^{-1}\widehat{\mu_2} - \frac{1}{2}\widehat{\mu_2}^T\hat{\Sigma}^{-1}\widehat{\mu_2} + \log\widehat{\pi_2}$$
$$= x^T\begin{bmatrix}20\\20\end{bmatrix} - \tfrac{1}{2}\left(\tfrac{400}{3}\right) + \log\tfrac{1}{2}$$
$$= 20X_1 + 20X_2 - \tfrac{200}{3} + \log\tfrac{1}{2}$$

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

$$\Longrightarrow \quad 10X_1 + 10X_2 - \tfrac{50}{3} + \log\tfrac{1}{2} = 20X_1 + 20X_2 - \tfrac{200}{3} + \log\tfrac{1}{2}$$
$$\Longrightarrow \quad \tfrac{150}{3} = 10X_1 + 10X_2$$
$$\Longrightarrow \quad 50 = 10X_1 + 10X_2$$
$$\Longrightarrow \quad 5 = X_1 + X_2$$
$$\Longrightarrow \quad -X_1 + 5 = X_2$$

Entonces, la línea que decide entre las dos clases está dada por $X_2 = -X_1 + 5$.

Aquí hay un gráfico de la línea de decisión::



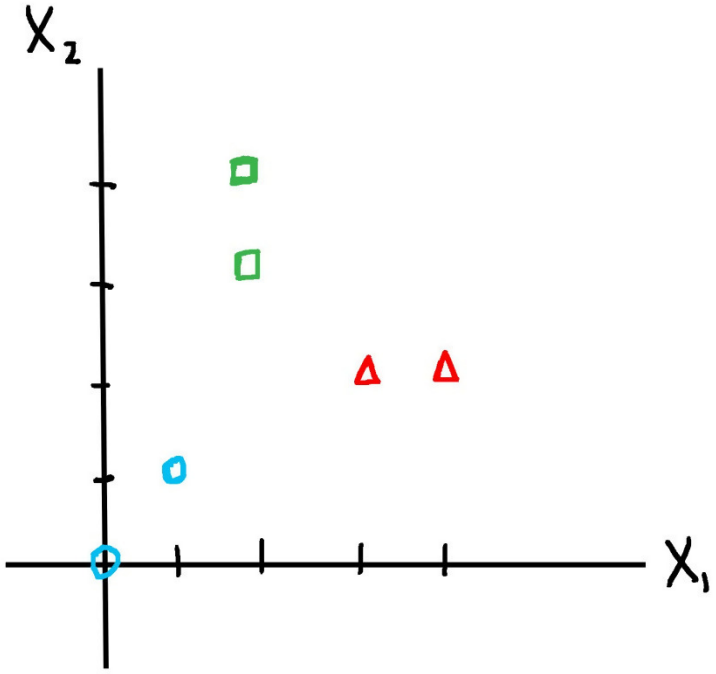
Si $\widehat{\delta_1}(x) > \widehat{\delta_2}(x)$, entonces clasificamos x como una clase de k_1.

Así que si x está debajo de la línea $X_2 = -X_1 + 5$, clasificamos x como una clase de k_1.

Al contrario, si $\widehat{\delta_1}(x) < \widehat{\delta_2}(x)$, clasificamos x como una clase de k_2. Esto corresponde a x estando arriba de la línea $X_2 = -X_1 + 5$.

El punto $(4, 5)$ está arriba de la línea, entonces lo clasificamos como una clase de k_2.

2. Aquí hay una gráfica de los puntos de datos:



El número de características p es 2, el número de clases K es 3, el número total de puntos de datos N es 6, el número N_1 de datos en la clase k_1 es 2, el número N_2 de datos en la clase k_2 es 2, y el número N_3 de datos en la clase k_3 es 2.

Primero, encontraremos estimaciones para π_1, π_2, π_3, las probabilidades previas de que $Y = k_1$, $Y = k_2$, $Y = k_3$, respectivamente.

Después, encontraremos estimaciones para μ_1, μ_2, μ_3, los vectores medios específicos de la clase.

Luego podemos calcular la estimación de la matriz de covarianza Σ.

Finalmente, utilizando las estimaciones $\widehat{\pi_1}, \widehat{\pi_2}, \widehat{\pi_3}, \widehat{\mu_1}, \widehat{\mu_2}, \widehat{\mu_3}, \hat{\Sigma}$, podemos encontrar las estimaciones para las funciones discriminantes lineales $\delta_1(x)$, $\delta_2(x)$, y $\delta_3(x)$.

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

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

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

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

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

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

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

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

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

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

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

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

$$\Longrightarrow \quad 18 = 9X_2$$

$$\Longrightarrow \quad 2 = X_2$$

Así, la línea que decide entre clases k_1 y k_2 es dado por $X_2 = 2$.

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

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

$\Longrightarrow \quad 18 = 9X_1$

$\Longrightarrow \quad 2 = X_1$

Así, la línea que decide entre clases k_1 y k_3 es dado por $X_1 = 2$.

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

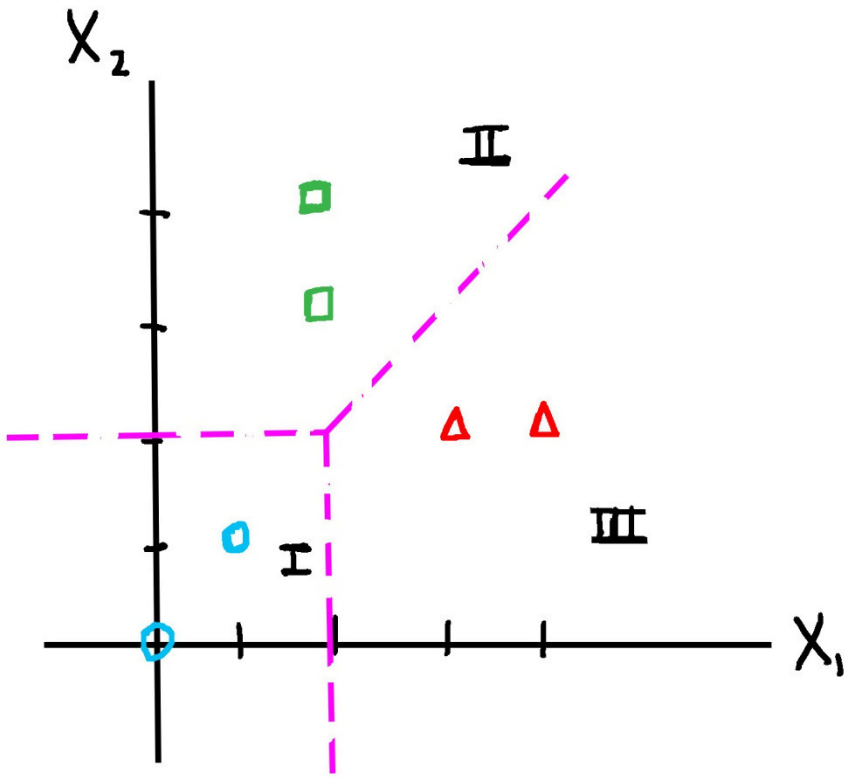
$\Longrightarrow \; X_1 + 10X_2 - \frac{37}{2} + \log\frac{1}{3} = 10X_1 + X_2 - \frac{37}{2} + \log\frac{1}{3}$

$\Longrightarrow \; 9X_2 = 9X_1$

$\Longrightarrow \; X_2 = X_1$

Así, la línea que decide entre clases k_2 y k_3 es dado por $X_2 = X_1$.

Aquí hay una gráfica de las líneas de decisión:



Las líneas dividen el plano en 3 regiones.
Si x está en la region I, podemos clasificar x como una clase de k_1. Del mismo modo, puntos en la region II estarán clasificados como parte de la clase k_2, y puntos en la region III estarán clasificados como parte de la clase k_3.

El punto $(3, 0)$ está en la region III, entonces lo clasificamos como una clase de k_3.