Aprendizaje Estadístico Supervisado

Natalia da Silva

2024

Problemas de clasificación

Vamos a repasar métodos para clasificación cubiertos en el Capítulo 4 del ISLR

$$Y = f(X) + \epsilon$$

- Cuando la variable de respuesta Y es categórica o cualitativa, el problema es de clasificación.
- Muchas veces los métodos para problemas de clasificacion se enfocan en predecir la probabilidad de cada clase y en base a ella clasifican.

Problemas de clasificación

- Las variables cualitativas toman sus valores en un conjunto no ordenado C tal que :

Si la respuesta Y = es color de ojos, entonces

$$\mathrm{Y} \in \{\mathrm{marron}, \mathrm{azul}, \mathrm{verde}, \mathrm{otro}\}$$

Problemas de clasificación

- En el capítulo 2 vimos que el error test $\text{Ave}(\text{I}(y_0 \neq \hat{y_0}))$ es minimizado en promedio por el clasificador de Bayes, que es el que asigna cada observación a la clase más probable dado el valor de sus predictoras.
- Con dos clases $\text{P}(\text{Y} = \text{g}/\text{X} = x_0) > 0.5$ predigo clase 1 y clase 2 en otro caso.
- En general en un problema de clasificación queremos aproximarnos a el calsificador de Bayes. Ya que no conocemos la distribución condicional de Y dado X la aproximamos de distintas formas.

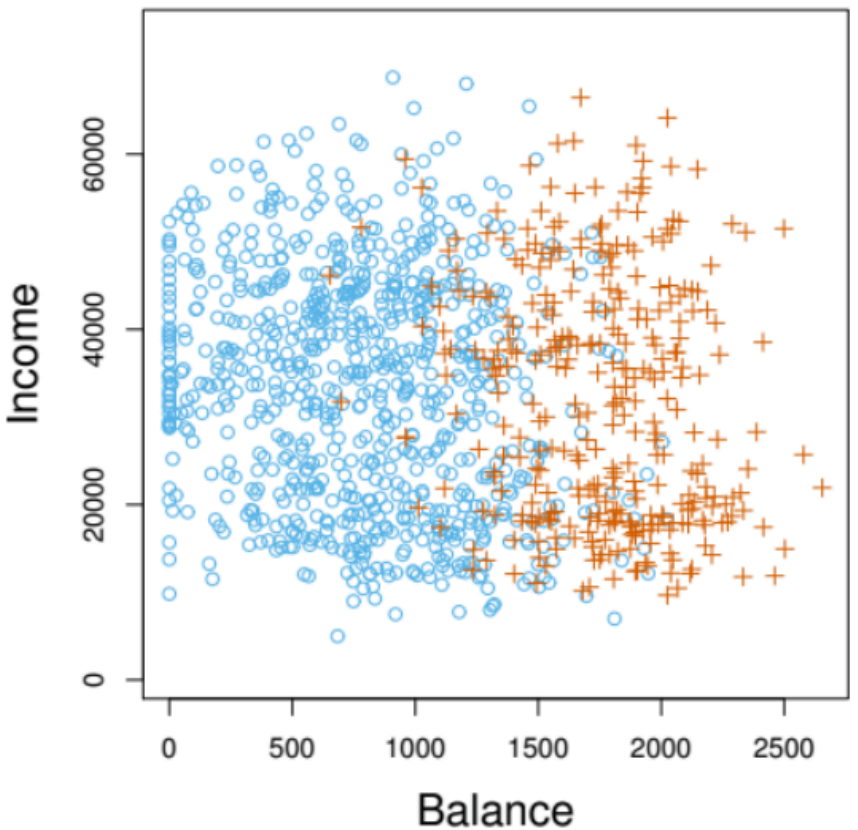
Vecino más cercano para clasificación

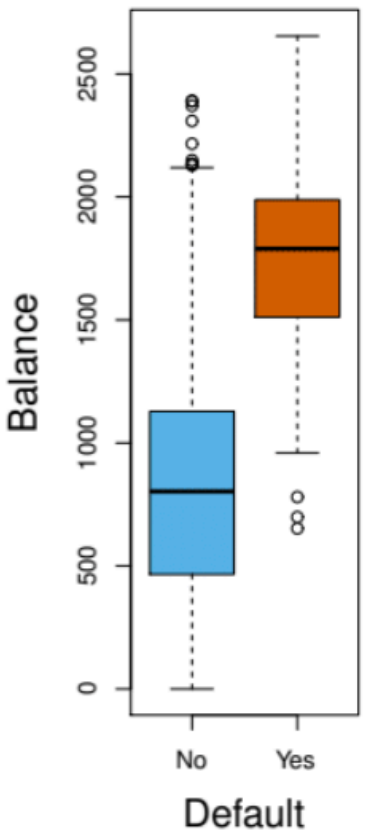
Vecino más cercano para el problema de clasificación directamente aproxima esta solución

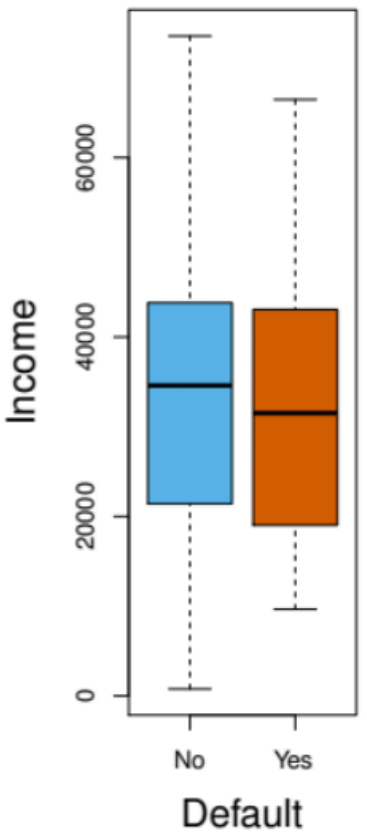
$$\hat{f}(X) = \arg\max_{g \in G_r} \sum_{x_i \in N_k(x)} I(y_i = g)$$

- La probabilidad se estima con una proporción en la muestra
- En vez de condicionar en un punto lo hace en un vecindario

Ejemplo: impago de tarjeta de crédito







¿Podemos usar un modelo de regresión lineal ?

- Suponemos que para el problema de clasificasión de impago de la tarjeta de crédito codficamos

$$Y = \begin{cases} 0 & \text{Pago} \\ 1 & \text{Impago} \end{cases}$$

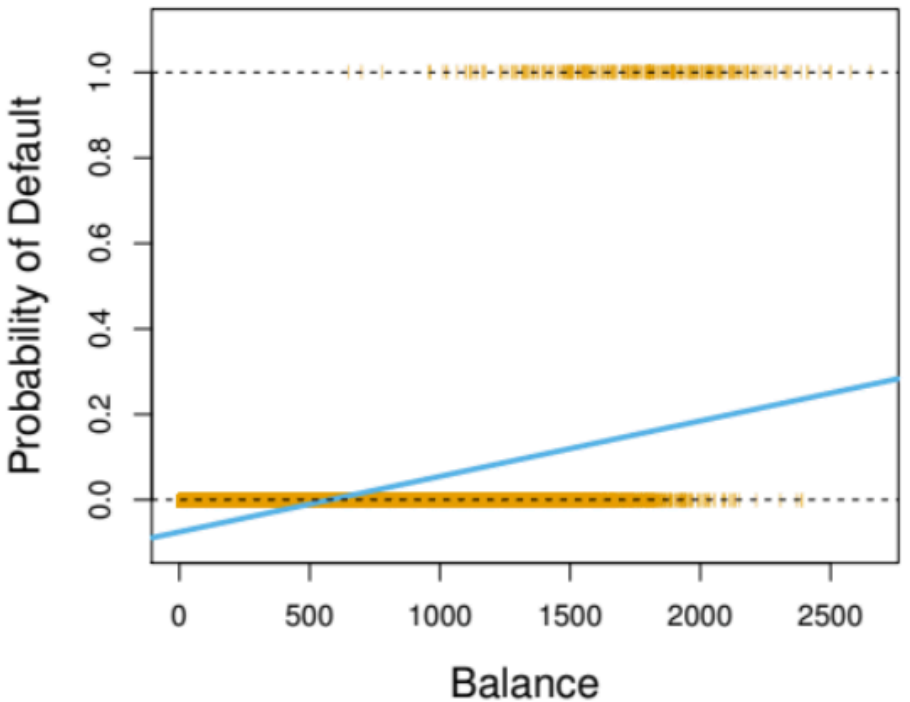
¿Podemos simplemente ajustar una regresión lineal de Y en X y clasificar como **Impago** si $Y > 0.5$?

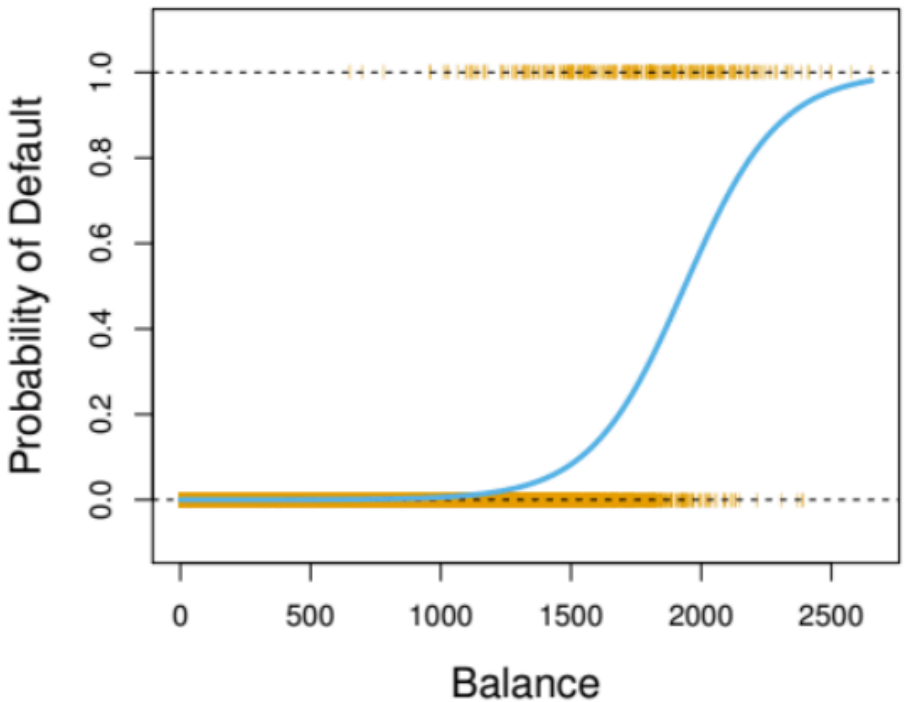
¿Podemos usar un modelo de regresión lineal ?

¿Podemos simplemente ajustar una regresión lineal de Y en X y clasificar como **Impago** si $\mathrm{Y} > 0.5$?

- En este caso de respuesta binaria, una regresión lineal hace un trabajo razonable como clasificador y es equivalente a hacer un análisis discriminante lineal.
- A nivel poblacional $\mathrm{E}(\mathrm{Y}/\mathrm{X} = \mathrm{x}) = \mathrm{P}(\mathrm{Y} = 1/\mathrm{X} = \mathrm{x})$ podemos pensar que la regresión es perfecta para este problema.
- Sin embargo, la regresión lineal puede producir probabilidades menores que cero o mayores que uno. Por lo que una regresión logística será más apropiado en este caso.

Regresión lineal vs Logistica





- Las marcas naranjas indican la respuesta Y que puede ser 0 o 1.
- Vemos que la regresión lineal no estima bien $P(Y = 1/X = x)$.
- La regresión logística parece apropiada para esta tarea.

Continuando con regresión lineal

Ahora suponemos que tenemos una variable de respuesta con tres posibles valores. Un paciente se presenta en una sala de emergencias y tenemos que clasificar su estado de acuerdo a los síntomas.

$$Y = \begin{cases} 1 & \text{infarto} \\ 2 & \text{sobredosis} \\ 3 & \text{ataque de epilepsia} \end{cases}$$

De acuerdo a esta codificación se sugiere un órden y de hecho implica que la diferencia entre infarto y sobredosis es la misma que entre sobredosis y ataque de epilepsia.

Continuando con regresión lineal

¿La regresión lineal es paropiada en este caso?

Continuando con regresión lineal

- En este contexto de clasificación para clases múltiples la regresión lineal no es apropiada.
- Regresión logística para clases múltiples o análisis discriminante son más apropiados en este caso.

Regresión Logística

- $p(X) = P(Y = 1/X = x)$ y vamos a usar la variable balance para predecir el impago de la tarjeta de crédito.
- Usamos la regresión logística:

$$p(X) = P(Y = 1/X = x) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

- Es fácil de ver que no importa los valores de β_0 y β_1 o X, $p(X)$ va a tener un valor enter 0 y 1

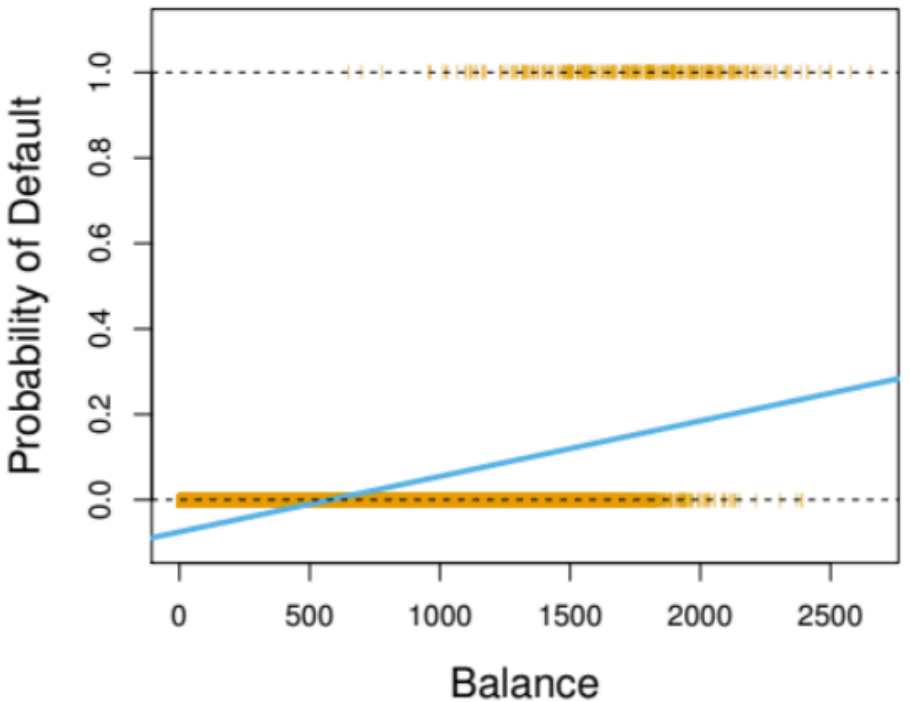
Regresión Logística

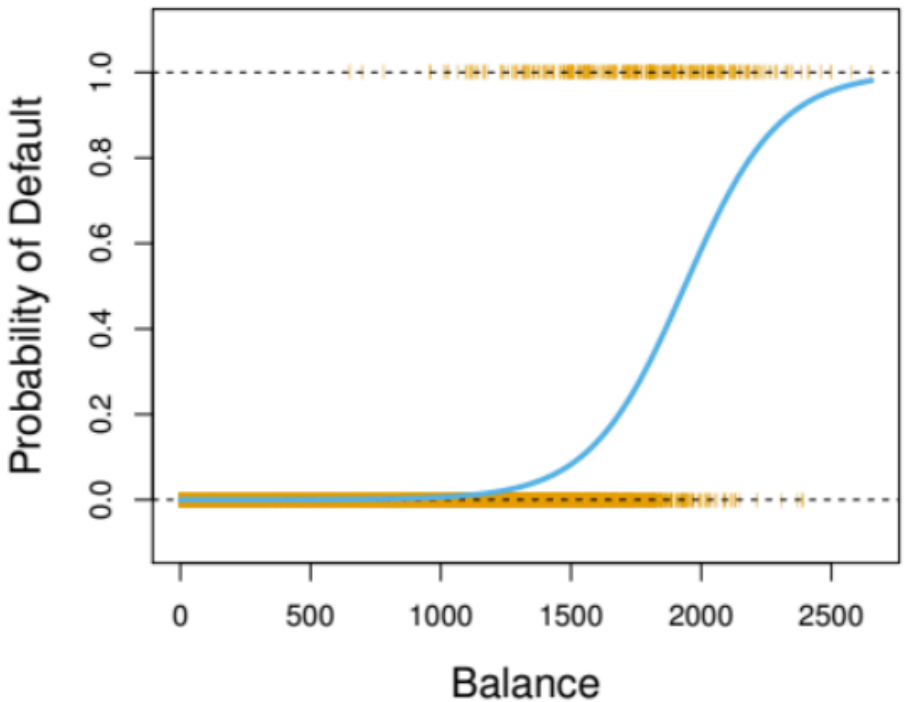
Reordenando llegamos a:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

- Esta transformación monótona es llamada **log odd** o transformación logística de $p(X)$
- El efecto de cada variable explicativa es lineal en el logaritmo de los odds
- El ood ratio es el ratio de dos probabilidades, la que ocurra el evento (p(X)) y que no ocurra (1-p(X)).
- Si da 2 significa que es dos veces más probable que ocurra el evento a que no ocurra.

Regresión Logística





- La regresión logística asegura que nuestra estimacion para $p(X)$ está entre 0 y 1.

Máxima verosimilitud

Se usa máxima verosimilitud para estimar los parámetros

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i))$$

- Esta verosimilitud nos da la probabilidad de observar 0 y 1 en los datos.
- β_0 y β_1 son los que maximizan la verosimilitud en los datos observados.

Ejemplo: resultados modelo logístico

	Coefficient	Std. Error	Z-statistic	P-value
`Intercept`	-10.6513	0.3612	-29.5	< 0.0001
`balance`	0.0055	0.0002	24.9	< 0.0001

Obteniendo predicciones

¿Cuál es la probabilidad estimada de que alguien con un balance de 1000 no pague la tarjeta de crédito?

$$p(X) = \frac{e^{\hat{\beta_0}+0.0055*X}}{1+e^{\hat{\beta_0}+0.0055*X}} = \frac{e^{10.65+0.0055*1000}}{1+e^{10.65+0.0055*1000}} = 0.006$$

¿Y con un balance de 2000 como cambia la proabilidad de impago?

$$p(X) = \frac{e^{\hat{\beta_0}+0.0055*X}}{1+e^{\hat{\beta_0}+0.0055*X}} = \frac{e^{10.65+0.0055*2000}}{1+e^{10.65+0.0055*2000}} = 0.586$$

Ahora ajustamos la regresión logística con var estudiante

	Coefficient	Std. Error	Z-statistic	P-value
`Intercept`	-3.5041	0.0707	-49.55	< 0.0001
`student[Yes]`	0.4049	0.1150	3.52	0.0004

- $\hat{P}(\text{Impago/Estudiante} = \text{Si}) = \frac{e^{-0.35+0.40*1}}{1+e^{-0.35+0.40*1}} = 0.043$
- $\hat{P}(\text{Impago/Estudiante} = \text{No}) = \frac{e^{-0.35+0.40*0}}{1+e^{-0.35+0.40*0}} = 0.029$

Regresión Logística múltiple

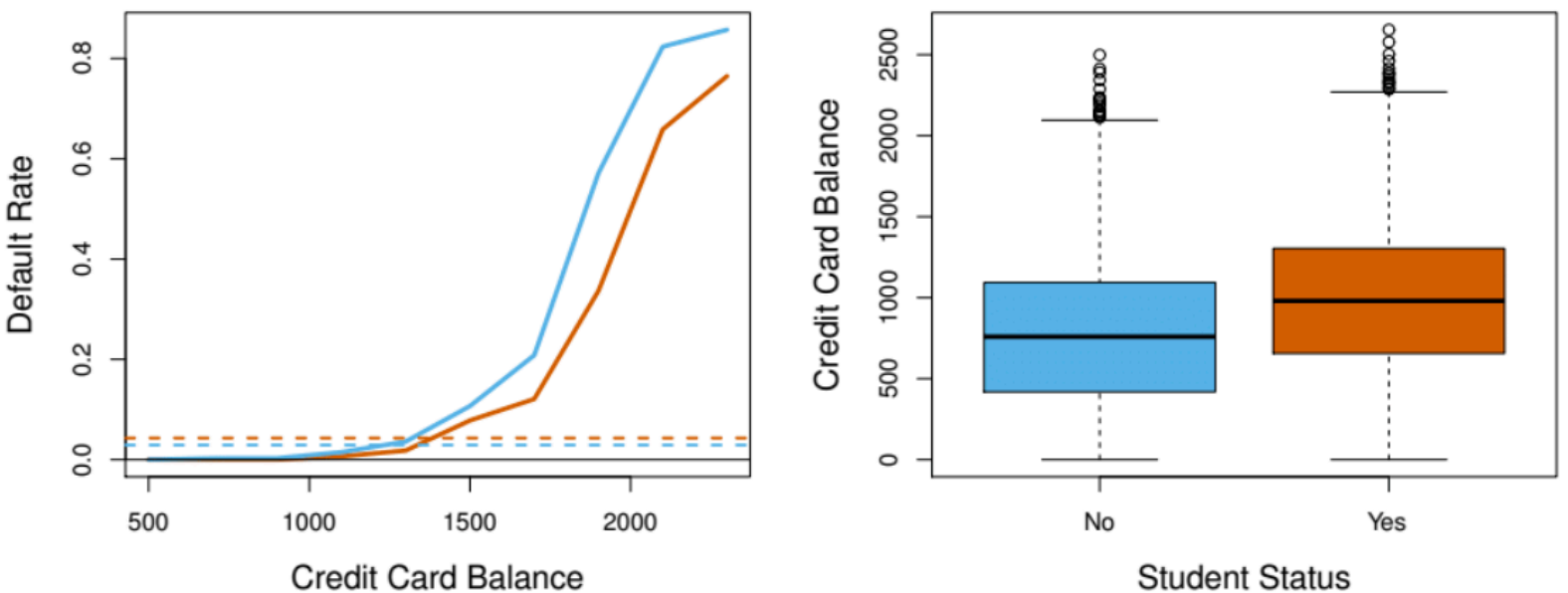
$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots \beta_p X_p$$

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2 \ldots \beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2 \ldots \beta_p X_p}}$$

	Coefficient	Std. Error	Z-statistic	P-value
`Intercept`	-10.8690	0.4923	-22.08	< 0.0001
`balance`	0.0057	0.0002	24.74	< 0.0001
`income`	0.0030	0.0082	0.37	0.7115
`student[Yes]`	-0.6468	0.2362	-2.74	0.0062

¿Porqué el coeficiente de estudiante es negativo y antes era positivo?

Confundente

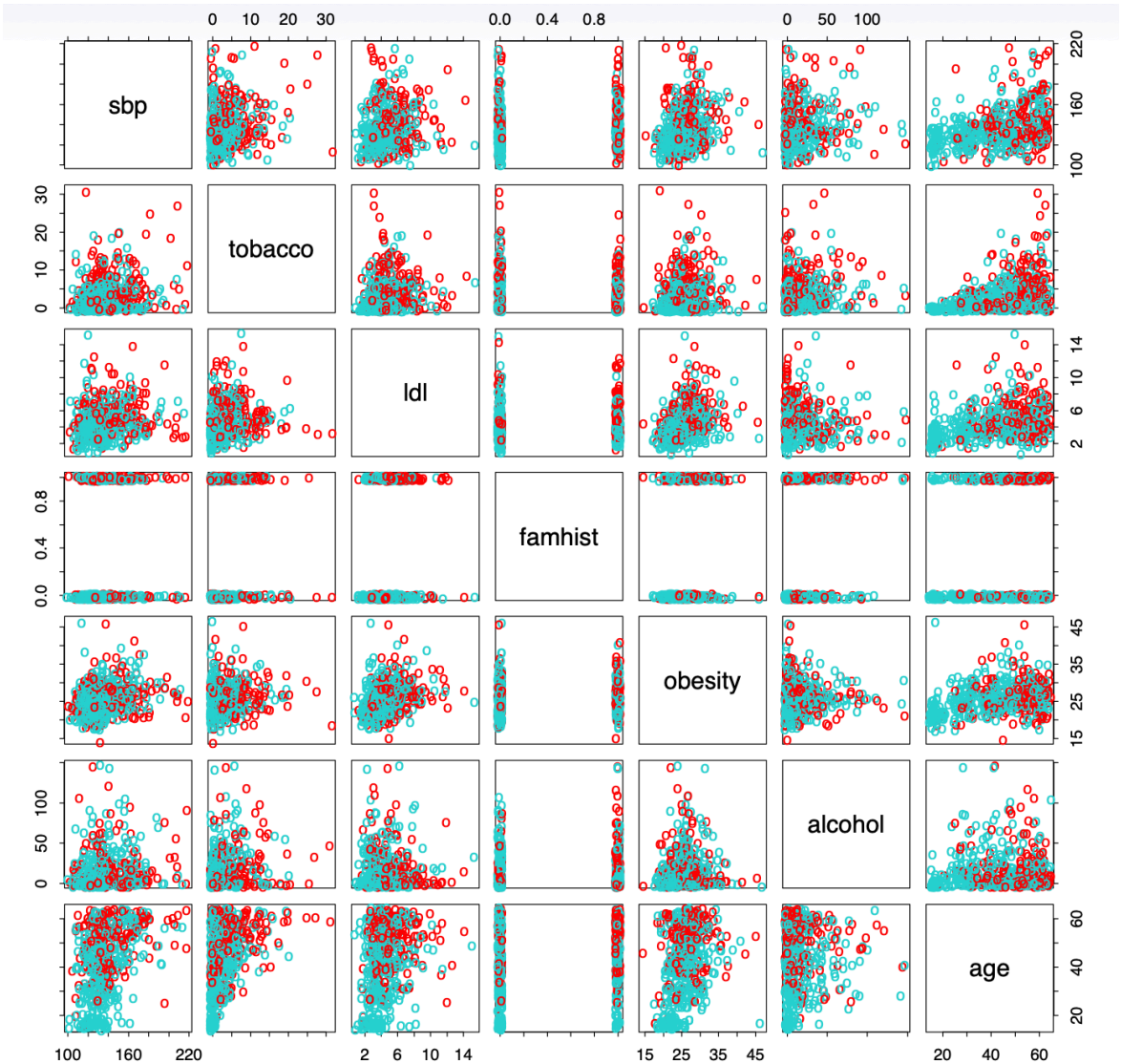


- Los estudiantes tienden a tener un balance mayor a los no estudiantes, por lo que su tasa marginal de impagos es más alta que la de los no estudiantes.
- Para cada nivel de balance la tasa de impago de los estudiantes es menor que las de no estudiantes.
- La regresión logística múltiple puede aclarar esto.

Ejemplo: Enfermedades cardiovasulares en África

- 160 casos de MI (infarto de miocardio) y 302 contro (hombres entre 15-64), en África en los 80's.
- Muy alta prevalencia en la región: 5.1%.
- Medida en 7 predictores (factor de riesgo), se presenta en el scatterplot matrix.
- Objetivo identificar fortalezas relativas y posibles factores de riesgo
- Parte de un estudio de intervensión con el objetivo de educar en dietas más saludables.

Ejemplo: Enfermedades cardiovasulares en África



Respuesta en rojo (casos de MI) y controles en celeste. `famhist` variable binaria, 1 indica casos familiares de MI.

Ejemplo: Enfermedades cardiovasulares en África

```
Call:
glm(formula = chd ~ ., family = binomial, data = heart)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -4.1295997  0.9641558  -4.283 1.84e-05 ***
sbp              0.0057607  0.0056326   1.023  0.30643
tobacco          0.0795256  0.0262150   3.034  0.00242 **
ldl              0.1847793  0.0574115   3.219  0.00129 **
famhistPresent   0.9391855  0.2248691   4.177 2.96e-05 ***
obesity         -0.0345434  0.0291053  -1.187  0.23529
alcohol          0.0006065  0.0044550   0.136  0.89171
age              0.0425412  0.0101749   4.181 2.90e-05 ***

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 483.17  on 454  degrees of freedom
AIC: 499.17
```

Regresión logística

- Hay 160 casos y 302 controles, $\tilde{\pi} = 0.35$. La prevalencia de MI es de $\pi = 0.05$
- Con la muestra de casos y controles podemos estimar los parámetros β_j precisamente si el modelo es apropiado.

- A menudo los casos son raros y los tomamos a todos, son más de 5 veces el número de control

Regresión logística con múltiples clases

Con más de dos clases el modelo de regresión logística se puede generalizar como sigue:

$$P(Y = k/X) = \frac{e^{\beta_{ok}+\beta_{1k}X_1+\beta_{2k}X_2\ldots\beta_{pk}X_p}}{\sum_{l=1}^{K} e^{\beta_{ol}+\beta_{1l}X_1+\beta_{2l}X_2\ldots\beta_{pl}X_p}}$$

Aquí hay una función lineal para cada clase

GLM (Modelo lineal generalizado)

Una forma general de describir un subconjunto amplio de modelos donde se encuentra el modelo lineal, la regresión logística entre otros.

Un GLM tiene 3 elementos basicos:

- Distribucion de la **familia exponencial** para la variable de respuesta
- Un **predictor lineal**, donde se incluyen las variables explicativas
- Una **funcion link** que vincule el predictor lineal con la media de la respuesta

Familia exponencial

La funcion de densidad tiene la forma

$$p(y|\theta,\varphi) = \exp\left\{ \frac{(y\theta - b(\theta))}{a(\varphi)} + c(y,\varphi) \right\}$$

- θ parámetro canónico
- φ parámetro de dispersión
- φ NO es un parametro a estimar

Ejemplos:

Normal, Poisson, Binomial, Gamma, Exponencial, etc

Predictor lineal y funcion link

El predictor lineal se forma como $x_i^\top \beta$, la funcion link conecta $E(y_i) = \mu_i$ con $x_i^\top \beta$.

$$
\begin{aligned}
y_i &\sim N(\mu_i, \sigma^2) \\
g(\mu_i) &= x_i^\top \beta
\end{aligned}
$$

- distribucion normal
- link identidad

Familia exponencial Bernoulli

$y \sim \text{Bernoulli}(p)$

$$
\begin{aligned}
p(y|p) &= p^y(1-p)^{1-y} \\
&= e^{y\log(p)+(1-y)\log(1-p)} \\
&= e^{y\log(\frac{p}{1-p})+\log(1-p)} \\
&= e^{\frac{y\log(\frac{p}{1-p})+\log(1-p)}{1}+0}
\end{aligned}
$$

$\theta = \log(\frac{p}{1-p}), \varphi = 1$. Donde a(.) es la identidad , $b(\theta) = \log(1+e^{\theta})$ y $c(.) = 0$

Pertenece a la familia exponencial

GLM Bernoulli

- $Y \sim \text{Bernoulli}(p(X))$
- Predictor lineal $\eta = \beta_0 + \beta_1 X1 + \beta_2 X_2 \ldots \beta_p X_p$
- $g(p(X)) = \log(\frac{p(X)}{1-p(X)})$
- $p(X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2 \ldots \beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2 \ldots \beta_p X_p}}$

Análisis discriminante

- La regresión logística directamente modela la $\mathrm{P}(\mathrm{Y} = \mathrm{k}/\mathrm{X} = \mathrm{x})$ usando la función logística.
- Ahora consideraremos una aproximación menos directa para aproximar las probabilidades de cada clase.
- En esta aproximación necesitamos la distribución de los predictores separadamente para cada una de las clases (para cada valor de Y)
- Luego usamos el teorema de Bayes para estimar $\mathrm{P}(\mathrm{Y} = \mathrm{k}/\mathrm{X} = \mathrm{x})$

Análisis discriminante

Aquí la aproximación es modelar la distribución de X en cada clase separadamente y entonces usar el teorema de Bayes para obtener $\mathrm{P(Y/X)}$

- Cuando usamos la distribución Normal para cada clase esto nos lleva al análisis de discriminante lineal o cuadrático.
- Sin embargo esta aproximación es bastante general y otras distribuciones pueden ser utilizadas. Nos enfocaremos en la distribución normal

Teorema de Bayes para clasificación

$$P(Y = k/X = x) = \frac{P(X = x/Y = k)P(Y = k)}{P(X = x)}$$

Teorema de Bayes para clasificación

$$P(Y = k/X = x) = \frac{P(X = x/Y = k)P(Y = k)}{P(X = x)}$$

- $P(Y = k/X = x)$ pobabilidad posterior que una observación pertenezca a la clase k dado el valor del predictor
- $P(Y = k)$ probabilidad previa que una observación aleatoriamente seleccionada venga de la clase k
- $P(X = x/Y = k)$ es la densidad de X para la clase k

Teorema de Bayes para clasificación

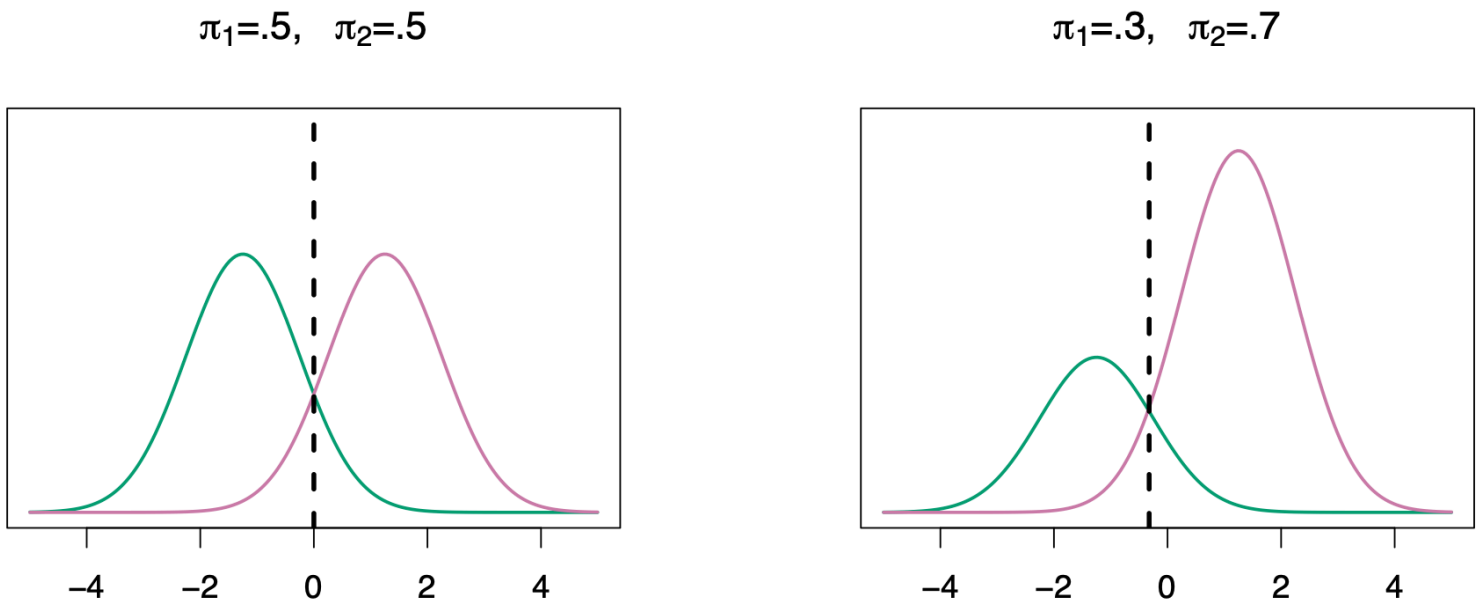
Para análisis discriminante se escribe un poco distinto:

$$P(Y = k/X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

- $f_k(x) = P(X = x/Y = k)$ es la densidad de X para la clase k
- Aquí usaremos densidades normales separadamente para cada clase.
- $\pi_k = P(Y = k)$ es la probabilidad marginal para la clase k o la previa

La idea es que en vez de estimar la probabilidad posterior se van a estima $\pi_k(x)$ y $f_k(x)$ y se van a remplazar en la ecuación. Se buscan distintas formas de aproximar el clasificador de Bayes estimando $f_k(x)$

Clasificación a la densidad más alta



- Clasificamos un nuevo punto de acuerdo a que densidad es más alta.
- Cuando las previas son distintas tomamos esta también en cuenta y comparamos $\pi f_k(x)$.
- En el panel derecho favorecemos la rosada por eso la banda de decisión se corre a las izquierda.

¿Porqué discriminante lineal?

- Cuando las clases están bien separadas la estimación de los parámetros por regresión logística son sorprendentemente intestables. El análisis discriminante lineal no sufre de este problema.
- Si n es pequeño y la distribución de los predictores X son aproximadamente normales para cada clase, el modelo de discriminante lineal es más estable que la regresión logística,
- El análisis de discriminante lineal es popular cuanto tenemos más de dos clases ya que nos permite un análisis visual en bajas dimensiones.

Análisis discriminante lineal con $p = 1$

La densidadn Normal:

$$f_k(x) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{\frac{-1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}$$

- μ_k la media de la clase k, σ_k^2 su varianza y asumiremos que $\sigma_k = \sigma$

Análisis discriminante lineal con $p = 1$

Remplazando en la ecuación de Bayes:

$$P(Y = k/X = x) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2}}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}\left(\frac{x-\mu_l}{\sigma}\right)^2}}$$

Función discriminante

Para clasificar en $X = x$ necesitamos ver que probabilidad $p_k(x)$ es la mayor.

Tomando logaritmos y descartando términos que no dependen de k esto es equivalente a asignar x a la clase con el mayor valor en el discriminante $\delta_k(x)$:

$$\delta_k(x) = x. \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

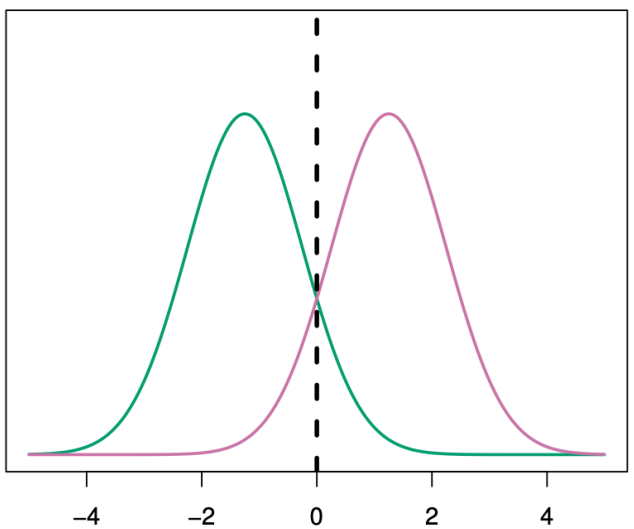
Notar que $\delta_k(x)$ es una función lineal en x

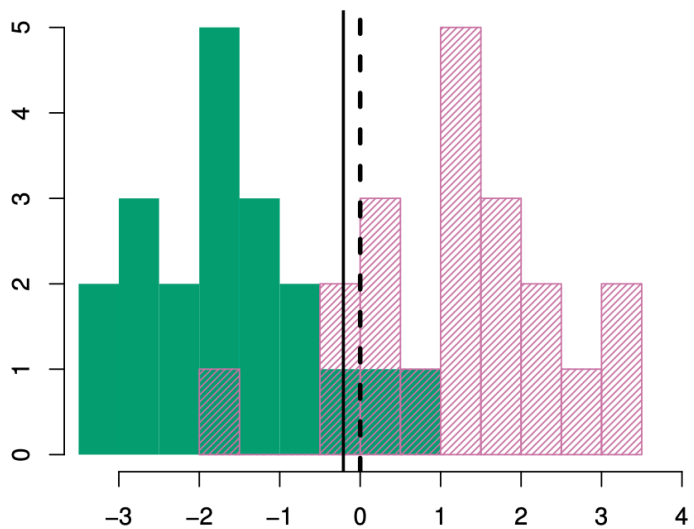
Función discriminante

- Si $K = 2$ y $\pi_1 = \pi_2 = 0.5$ el clasificador de Bayes asigna una observación a la clase 1 si $2x(\mu_1 + \mu_2) > \mu_1^2 - \mu_2^2$ y a las clase 2 en otro caso.
- La banda de decisión Bayesiana es el punto para el cuál $\delta_1(x) = \delta_2(x)$ se puede mostrar que esto se da en:

$$x = \frac{\mu_1 + \mu_2}{2}$$

Ejemplo





Example with $\mu_1 = -1.5$, $\mu_2 = 1.5$, $\pi_1 = \pi_2 = 0.5$, and $\sigma^2 = 1$.

- Tipicamente no conocemos los valores de los parámetros, sólo contamos con el conjunto de entrenamiento En este caso estimamos los parámetros y los remplazamos en la regla

Estimamos los parámetros

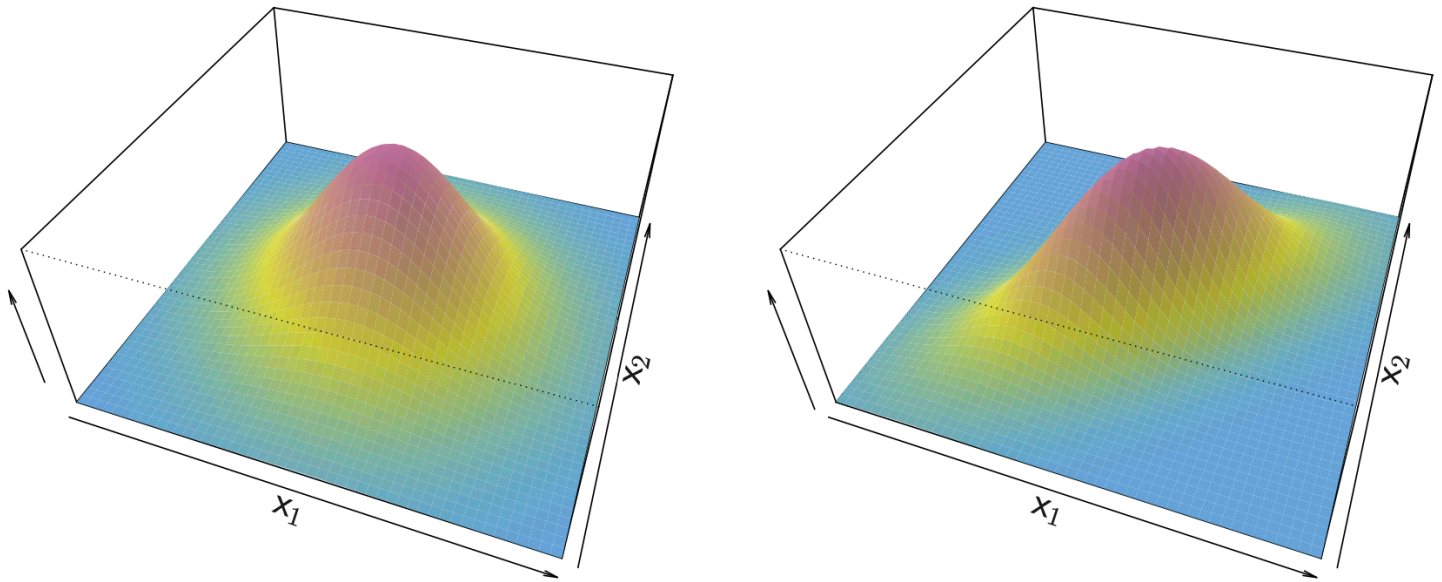
El análisis discriminante lineal estima el clasificador de Bayes remplazando los valores estimados de μ_k, π_k y σ^2 en la ecuación de $\delta_k(x)$

- $\hat{\pi_k} = \frac{n_k}{n}$
- $\hat{\mu_k} = \frac{1}{n_k} \sum_{i:y_i=k} x_i$
- $\hat{\sigma^2} = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu_k})^2$
- $= \sum_{k=1}^{K} \frac{n_k - 1}{n-K} \hat{\sigma^2}_k$

$$\hat{\delta_k}(x) = x. \frac{\hat{\mu_k}}{\sigma^2} - \frac{\hat{\mu_k^2}}{2\sigma^2} + \log(\hat{\pi_k})$$

Se llama discriminante lineal porque $\hat{\delta_k}(x)$ es una función lineal en x

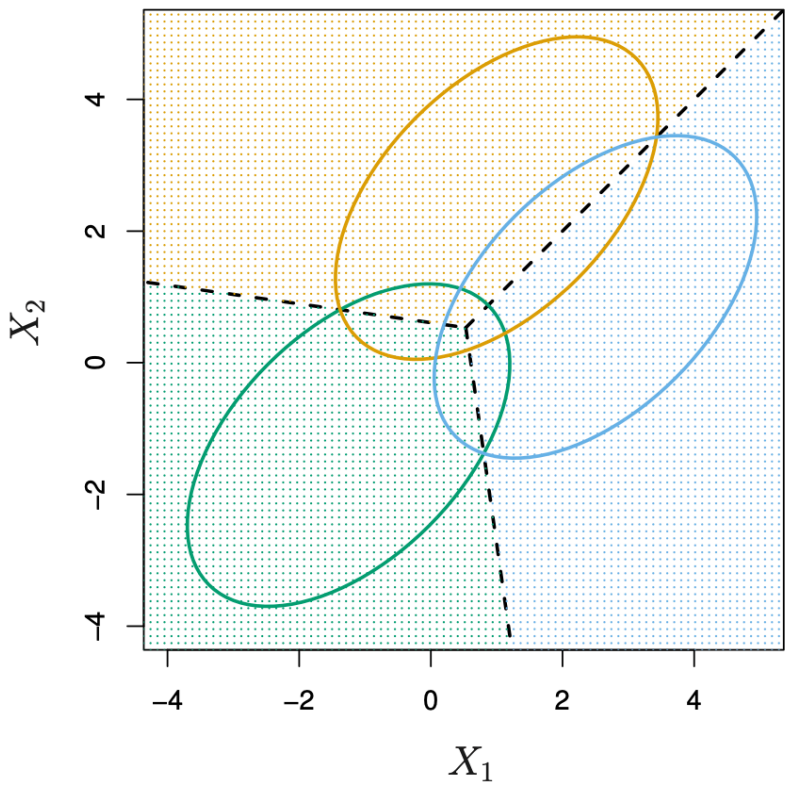
Discriminante lineal para $p > 1$

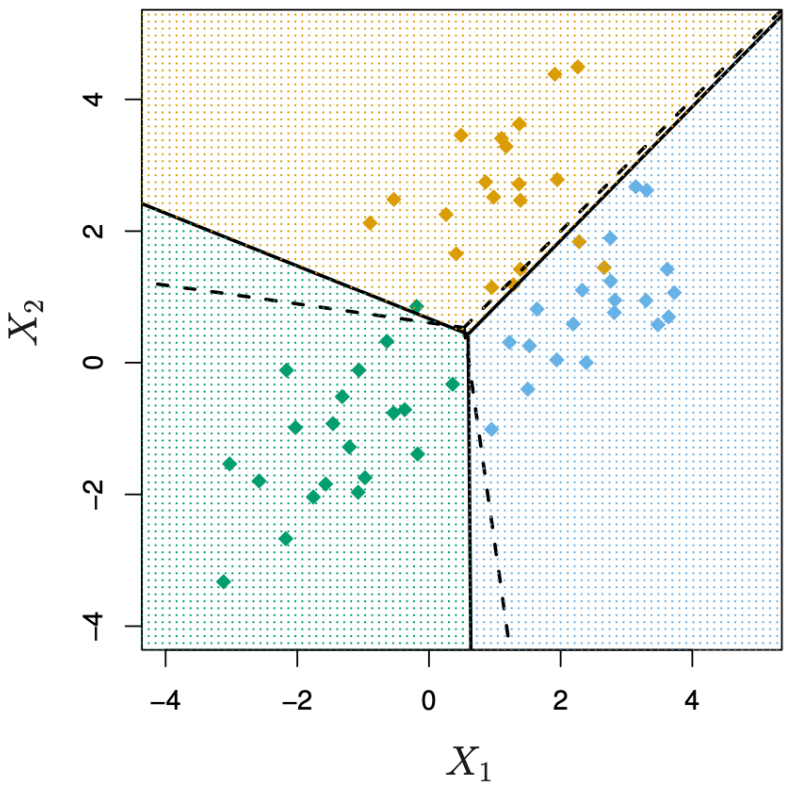


$$\text{Density: } f(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \mathbf{\Sigma}^{-1}(x-\mu)}$$

Función discriminante: $\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k)$ A pesar de su forma compleja $\delta_k(x)$ queda una función lineal de x

$p = 2$ y $K = 3$





- Aquí $\pi_1 = \pi_2 = \pi_3 = 1/3$
- Las líneas rayadas con las bandas de decisión Bayesianas, cuando son conocidas llevan al error de clasificación más pequeño