

# Ciencias de Datos con R: Fundamentos Estadísticos

Ana M. Bianco, Jemina García y Mariela Sued

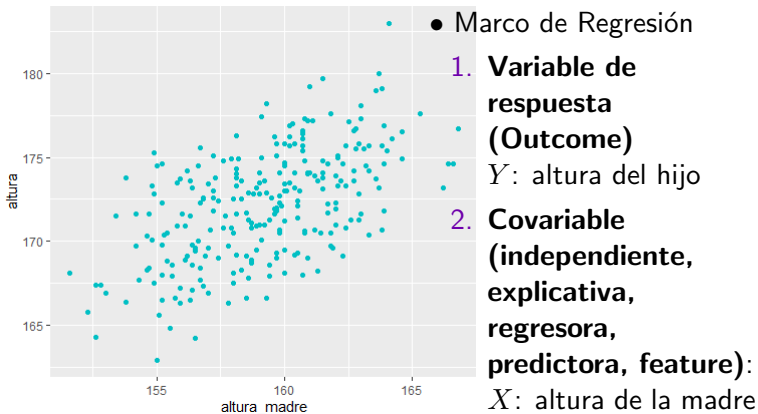
Clases 12 - Clasificación

# Problema de Alturas

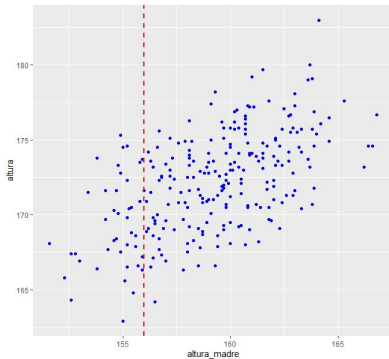
	A	B	C	D
1	altura	genero	contextura_madre	altura_madre
2	172.7	M	mediana	159.8
3	158.5	F	mediana	160.3
4	162.6	F	mediana	160.5
5	174.1	M	mediana	159.8
6	168.3	M	mediana	158.3
7	170.9	M	bajita	155.8
8	160.5	F	alta	163.5
9	169.3	M	mediana	160.5
10	172	M	mediana	159.7
11	173.5	M	bajita	155.8
12	171.5	M	mediana	157.7
13	158.8	F	mediana	160.9
14	159.5	F	alta	162.7
15	173.7	M	alta	163.6
16	160.7	F	alta	162.7
17	175.3	M	bajita	154.9
18	174.5	M	alta	163.2

- Variables
  - $Y$ : Altura (cuantitativa)
  - $X_1$ : Género
  - $X_2$ : Contextura de la madre
  - $X_3$ : Altura de la madre

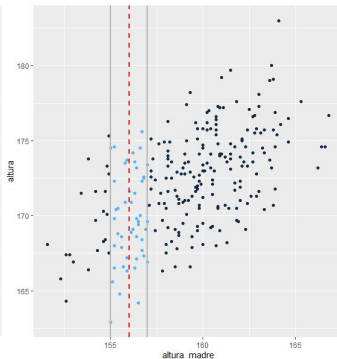
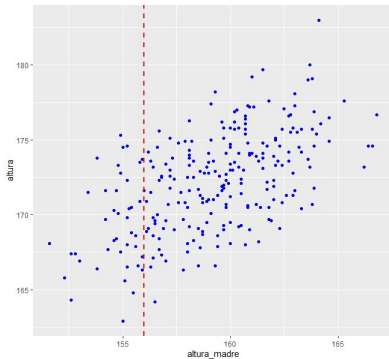
# Género Masculino: Problema de Predicción



# Estimación No Paramétrica de la Regresión



# Estimación No Paramétrica de la Regresión



## Otro escenario: ¿Le damos un crédito?

Se quiere predecir si un cliente de un banco pagará un crédito.

Variables registradas:

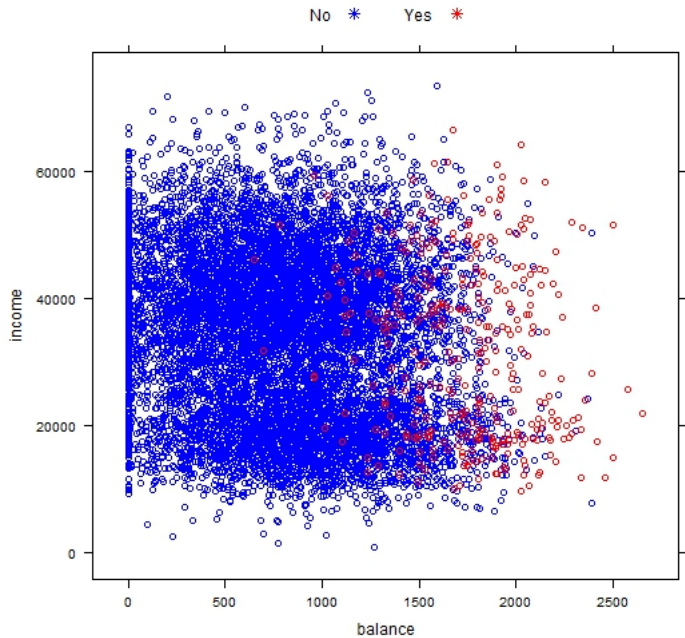
- $X_1$ : **balance** (saldo tarjeta)
- $X_2$ : **income** (ingreso anual)
- $X_3$ : **student** (si - no)

$$\bullet Y = \begin{cases} 1 & \text{(Yes) default} \\ 0 & \text{(No) c.c.} \end{cases}$$

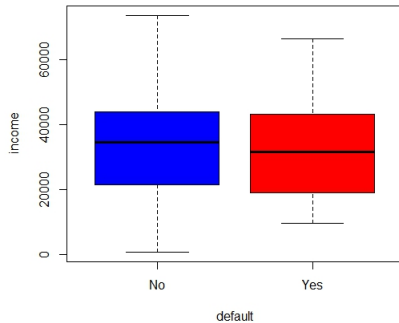
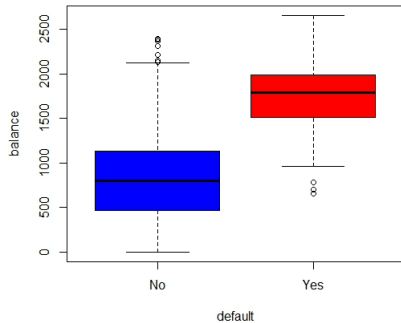
# Paquete ISLR

```
> Default
```

	default	student	balance	income
1	No	No	729.52650	44361.625
2	No	Yes	817.18041	12106.135
3	No	No	1073.54916	31767.139
4	No	No	529.25060	35704.494
5	No	No	785.65588	38463.496
6	No	Yes	919.58853	7491.559
7	No	No	825.51333	24905.227
8	No	Yes	808.66750	17600.451
9	No	No	1161.05785	37468.529
10	No	No	0.00000	29275.268
11	No	Yes	0.00000	21871.073
12	No	Yes	1220.58375	13268.562
13	No	No	237.04511	28251.695
14	No	No	606.74234	44994.556
15	No	No	1112.96840	23810.174
16	No	No	286.23256	45042.413
17	No	No	0.00000	50265.312
18	No	Yes	527.54018	17636.540
19	No	No	485.93686	61566.106
20	No	No	1095.07274	26464.631
21	No	No	228.95255	50500.182
22	No	No	954.26179	32457.509
23	No	No	1055.95660	51317.883
24	No	No	641.98439	30466.103
25	No	No	773.21172	34353.314







# Predicción

- $\mathbf{X} = (X_1, X_2, \dots, X_p)$ : vector de covariables
- $Y$ : respuesta

# Predicción

- $\mathbf{X} = (X_1, X_2, \dots, X_p)$ : vector de covariables
- $Y$ : respuesta  $\longrightarrow Y = \begin{cases} \text{cuantitativa} & \text{Regresión} \\ \text{cualitativa} & \text{Clasificación} \end{cases}$

# Predicción - Learning

- $Y$ : respuesta o outcome
- $\mathbf{X} = (X_1, X_2, \dots, X_p)$ : vector de variables predictoras
- Datos
- **Aprender**: construir a partir de los datos un procedimiento que permita predecir el valor de la variable respuesta para cierto valor de la variable predictora.

$$\longrightarrow Y = \begin{cases} \text{cuantitativa} & \text{Regresión} \\ \text{cualitativa} & \text{Clasificación} \end{cases}$$

# Diccionario - Wasserman & Hastie

$$\mathbf{X} = (X_1, X_2, \dots, X_p) \in \mathcal{X} \quad Y \in \{0, 1\}, \quad Y \in \mathcal{Y}$$

Estadística	Computer Science	Significado
Clasificación Covariables Clasificador	Aprendizaje Supervisado Features Hypothesis	Predecir $Y$ categórica con $\mathbf{X}$ las $X_i$ s map $h = \mathcal{X} \rightarrow \mathcal{Y}$

## Clasificación: Toy example - El patito feo

- Dos posibles hospedadores:

$$Y = \begin{cases} 1 & \text{rechazador} \\ 0 & \text{aceptador.} \end{cases}$$

- Se colocan  $n = 8$  huevos parasitarios en nido.
- $X$  = número de huevos removidos.
- Si se remueven 5 huevos; ¿de qué clase de nido diría que se trata?
- Si se remueven 3 huevos; ¿de qué clase de nido diría que se trata?

# Clasificador

Clasificador: Regla (de clasificación) que asigna a  $x \in \{0, 1, \dots, 8\}$   
un tipo de hospedador:  $\{0(A), 1(R)\}$

$x$	0	1	2	3	4	5	6	7	8
-----	---	---	---	---	---	---	---	---	---

# Clasificador

Clasificador: Regla (de clasificación) que asigna a  $x \in \{0, 1, \dots, 8\}$   
un tipo de hospedador:  $\{0(A), 1(R)\}$

$x$	0	1	2	3	4	5	6	7	8
clasificador 1	0	0	0	0	1	1	1	1	1



# Clasificador

Clasificador: Regla (de clasificación) que asigna a  $x \in \{0, 1, \dots, 8\}$   
un tipo de hospedador:  $\{0(A), 1(R)\}$

$x$	0	1	2	3	4	5	6	7	8
clasificador 1	0	0	0	0	1	1	1	1	1

Otro clasificador

$x$	0	1	2	3	4	5	6	7	8
clasificador 2	0	1	0	1	1	1	1	1	1

# Binarizamos

- Dos posibles hospedadores:

$$Y = \begin{cases} 1 & \text{rechazador} \\ 0 & \text{aceptador} . \end{cases}$$

- Se colocan  $n = 8$  huevos parasitarios en nido.
- $X$  = número de huevos removidos.
- Clasificador: Regla que asigna a  $x \in \{0, 1, \dots, 8\}$  un posible valor  $y \in \{0, 1\}$ .

$x$	0	1	2	3	4	5	6	7	8
clasificador	0	0	0	0	1	1	1	1	1

Table: Ejemplo de clasificador

# Clasificador

- Información disponible  $x \in \mathcal{X}$ .
- Posibles *etiquetas*. Caso binario  $\mathcal{Y} = \{0, 1\}$
- Posibles *etiquetas*. Caso general  $\mathcal{Y} = \{y_1, \dots, y_k\}$
- Clasificador: Regla que asigna a  $x \in \mathcal{X}$  un posible valor  $y \in \mathcal{Y}$ .

# Clasificación: Marco Teórico

- $X \in \mathcal{X}, Y \in \mathcal{Y}$ .
- $(X, Y)$  vector aleatorio, con puntual  $p_{XY}$ .
- Clasificador: Regla (de clasificación) que asigna a cada  $x \in \mathcal{X}$  un elemento  $y \in \mathcal{Y}$

Clasificador  $g : \mathcal{X} \rightarrow \mathcal{Y}$

- Error de Clasificación Medio (verdadero - poblacional) del clasificador  $g$

$$L(g) = \mathbb{P}(g(X) \neq Y)$$

- Objetivo (teórico): Encontrar  $g$  que minimice el error medio de clasificación

$$g^{op}$$

# Tarea 1: Volvamos a El Patito Feo

clasificador 1:

$x$	0	1	2	3	4	5	6	7	8
$g_1(x)$	0	0	0	0	1	1	1	1	1

Calcular

$$L(g_1) = \mathbb{P}(g_1(X) \neq Y)$$

# Resolución de Tarea 1: Volvamos a El Patito Feo

clasificador 1:

$x$	0	1	2	3	4	5	6	7	8
$g_1(x)$	0	0	0	0	1	1	1	1	1

¿Acertó o no?

$Y/X$	0	1	2	3	4	5	6	7	8
0	✓	✓	✓	✓	x	x	x	x	x
1	x	x	x	x	✓	✓	✓	✓	✓

# Resolución de Tarea 1: Volvamos a El Patito Feo

clasificador 1:

$x$	0	1	2	3	4	5	6	7	8
$g_1(x)$	0	0	0	0	1	1	1	1	1

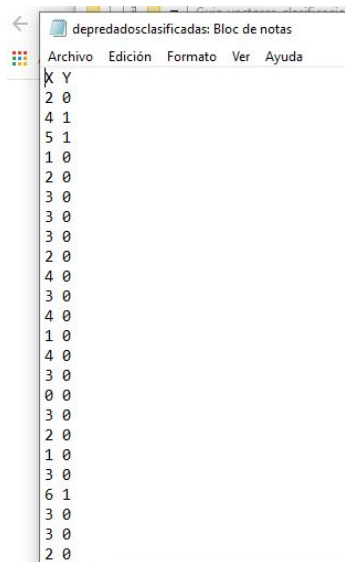
Conjunta  $(X, Y)$

$Y/X$	0	1	2	3	4	5	6	7	8	
0	0.0519	0.1779	0.2668	0.2287	0.1225	0.0420	0.0090	0.0011	0.0001	
1	0.0000	0.0000	0.0001	0.0009	0.0046	0.0147	0.0294	0.0336	0.0168	

Si sumamos las probabilidades que están en violeta, obtenemos

Error de Clasificación de  $g_1 = 0.1757$

# En la práctica, ¿cómo hacemos?



X	Y
2	0
4	1
5	1
1	0
2	0
3	0
3	0
3	0
2	0
4	0
3	0
4	0
1	0
4	0
3	0
0	0
3	0
2	0
1	0
3	0
6	1
3	0
3	0
2	0

- Datos:

$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$



# En la práctica, ¿cómo hacemos?

- $X \in \mathcal{X}, Y \in \mathcal{Y}$ .
- Clasificador:  $g : \mathcal{X} \rightarrow \mathcal{Y}$

Datos:  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

- **Error de Clasificación Empírico** del clasificador  $g$ :  
proporción de pares mal clasificados según  $g$ .

en R:  $\text{mean}(g(x) \neq y)$

## En la práctica, ¿cómo hacemos?

- $X \in \mathcal{X}, Y \in \mathcal{Y}$ .
- Clasificador:  $g : \mathcal{X} \rightarrow \mathcal{Y}$

Datos:  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

- **Error de Clasificación Empírico** del clasificador  $g$ :  
proporción de pares mal clasificados según  $g$ .

en R:  $\text{mean}(g(x) \neq y)$

en matemática:  $\hat{L}_n(g) = \frac{1}{n} \sum_{i=1}^n I_{\{g(x_i) \neq y_i\}}$

Notación:  $I_A := I_A(u) = \begin{cases} 1 & \text{si } u \in A \\ 0 & \text{si } u \notin A \end{cases}$

## Tarea 2: Volvamos a El Patito Feo

clasificador 1:

$x$	0	1	2	3	4	5	6	7	8
$g_1(x)$	0	0	0	0	1	1	1	1	1

Computar el Error de Clasificación Empírico de la regla de clasificación  $g_1$  en los datos del archivo

**depredadosclasificadas.txt**

## Clasificación - Ejemplo

$Y$	0	1
$p_Y$	0.90	0.10

## Clasificación - Ejemplo

$Y$	0	1
$p_Y$	0.90	0.10

- $X \in \mathcal{X}$ ,  $Y \in \mathcal{Y}$ ,  $(X, Y)$  vector aleatorio, con puntual  $p_{XY}$ .

$Y$	0	1
$p_{Y X=2}$	0.80	0.20

$Y$	0	1
$p_{Y X=3}$	0.30	0.70

¿Cómo clasificaríamos con esta información?

## Clasificación - Ejemplo

$Y$	0	1
$p_Y$	0.90	0.10

- $X \in \mathcal{X}$ ,  $Y \in \mathcal{Y}$ ,  $(X, Y)$  vector aleatorio, con puntual  $p_{XY}$ .

$Y$	0	1
$p_{Y X=2}$	0.80	0.20

$Y$	0	1
$p_{Y X=3}$	0.30	0.70

¿Cómo clasificaríamos con esta información?

- Clasificador  $g$ :

$x$	.	.	.	2	3	.	.	.	.
clasificador	.	.	.	0	1	.	.	.	.

## $g^{op}$ Optimo: Regla de Bayes - Caso binario

$$g^{op}(x) = \begin{cases} 1 & \text{si } \mathbb{P}(Y = 1 \mid X = x) \geq \mathbb{P}(Y = 0 \mid X = x) \\ 0 & \text{si } \mathbb{P}(Y = 0 \mid X = x) > \mathbb{P}(Y = 1 \mid X = x) \end{cases}$$

## $g^{op}$ Optimo: Regla de Bayes - Caso binario

$$g^{op}(x) = \begin{cases} 1 & \text{si } \mathbb{P}(Y = 1 \mid X = x) \geq \mathbb{P}(Y = 0 \mid X = x) \\ 0 & \text{si } \mathbb{P}(Y = 0 \mid X = x) > \mathbb{P}(Y = 1 \mid X = x) \end{cases}$$

Teorema:  $\mathbb{P}(g^{op}(X) \neq Y) \leq \mathbb{P}(g(X) \neq Y)$  , para todo  $g$ .



## $g^{op}$ Optimo: Regla de Bayes - Caso binario

$$g^{op}(x) = \begin{cases} 1 & \text{si } \mathbb{P}(Y = 1 \mid X = x) \geq \mathbb{P}(Y = 0 \mid X = x) \\ 0 & \text{si } \mathbb{P}(Y = 0 \mid X = x) > \mathbb{P}(Y = 1 \mid X = x) \end{cases}$$

Teorema:  $L(g^{op}) = \mathbb{P}(g^{op}(X) \neq Y) \leq \mathbb{P}(g(X) \neq Y) = L(g)$  , para todo  $g$ .

## $g^{op}$ Optimo: Regla de Bayes - Caso binario

$$g^{op}(x) = \begin{cases} 1 & \text{si } \mathbb{P}(Y = 1 \mid X = x) \geq \mathbb{P}(Y = 0 \mid X = x) \\ 0 & \text{si } \mathbb{P}(Y = 0 \mid X = x) > \mathbb{P}(Y = 1 \mid X = x) \end{cases}$$

## $g^{op}$ Optimo: Regla de Bayes - Caso binario

$$g^{op}(x) = \begin{cases} 1 & \text{si } \mathbb{P}(Y = 1 \mid X = x) \geq \mathbb{P}(Y = 0 \mid X = x) \\ 0 & \text{si } \mathbb{P}(Y = 0 \mid X = x) > \mathbb{P}(Y = 1 \mid X = x) \end{cases}$$

Notemos que

$$\mathbb{P}(Y = 1 \mid X = x) \geq \mathbb{P}(Y = 0 \mid X = x) \Leftrightarrow \mathbb{P}(Y = 1 \mid X = x) \geq 1/2$$

$$g^{op}(x) = \begin{cases} 1 & \text{si } \mathbb{P}(Y = 1 \mid X = x) \geq 1/2 \\ 0 & \text{c. c.} \end{cases}$$

## $g^{op}$ Optimo: Regla de Bayes - Caso binario

$$g^{op}(x) = \begin{cases} 1 & \text{si } \mathbb{P}(Y = 1 \mid X = x) > \mathbb{P}(Y = 0 \mid X = x) \\ 0 & \text{si c. c.} \end{cases}$$

De Bayes, tenemos que

$$\mathbb{P}(Y = y \mid X = x) = \frac{\mathbb{P}(X = x \mid Y = y) \mathbb{P}(Y = y)}{\mathbb{P}(X = x)}$$

$$g^{op}(x) = \begin{cases} 1 & \text{si } \mathbb{P}(X = x \mid Y = 1) \mathbb{P}(Y = 1) > \mathbb{P}(X = x \mid Y = 0) \mathbb{P}(Y = 0) \\ 0 & \text{si } \mathbb{P}(X = x \mid Y = 0) \mathbb{P}(Y = 0) > \mathbb{P}(X = x \mid Y = 1) \mathbb{P}(Y = 1) \end{cases}$$

## $g^{op}$ Optimo: Regla de Bayes - Caso binario

Si  $X|Y = 0 \sim f_0$  y  $X|Y = 1 \sim f_1$ ,  
como tenemos

$$g^{op}(x) = \begin{cases} 1 & \text{si } \mathbb{P}(X = x | Y = 1) \mathbb{P}(Y = 1) > \mathbb{P}(X = x | Y = 0) \mathbb{P}(Y = 0) \\ 0 & \text{si } \mathbb{P}(X = x | Y = 0) \mathbb{P}(Y = 0) > \mathbb{P}(X = x | Y = 1) \mathbb{P}(Y = 1) \end{cases}$$

resulta

$$g^{op}(x) = \begin{cases} 1 & \text{si } f_1(x) \mathbb{P}(Y = 1) > f_0(x) \mathbb{P}(Y = 0) \\ 0 & \text{si c. c.} \end{cases}$$

## Estimación $g^{op}$ Optimo: Modelos discriminativos

$$g^{op}(x) = \begin{cases} 1 & \text{si } \mathbb{P}(Y = 1 \mid X = x) \geq 1/2 \\ 0 & \text{si c. c.} \end{cases}$$

Aquí se estima

- $\mathbb{P}(Y = 1 \mid X = x)$ .

## Estimación $g^{op}$ Optimo: Modelos generativos

$$g^{op}(x) = \begin{cases} 1 & \text{si } f_1(x)\mathbb{P}(Y = 1) > f_0(x)\mathbb{P}(Y = 0) \\ 0 & \text{si c. c.} \end{cases}$$

Se estiman

- $f_0$  y  $f_1$
- $\mathbb{P}(Y = 1)$  o  $\mathbb{P}(Y = 0)$ .

## $g^{op}$ Optimo: Regla de Bayes - Caso binario

$$g^{op}(x) = \begin{cases} 1 & \text{si } \mathbb{P}(Y = 1 \mid X = x) > \mathbb{P}(Y = 0 \mid X = x) \\ 0 & \text{si } \mathbb{P}(Y = 0 \mid X = x) > \mathbb{P}(Y = 1 \mid X = x). \end{cases}$$

En otras palabras,

$$g^{op}(x) = \begin{cases} 1 & \text{si } p_{X|Y=1}(x) \mathbb{P}(Y = 1) > p_{X|Y=0}(x) \mathbb{P}(Y = 0) \\ 0 & \text{si } p_{X|Y=0}(x) \mathbb{P}(Y = 0) > p_{X|Y=1}(x) \mathbb{P}(Y = 1) \end{cases}$$



# Todo muy lindo, pero.... ¿quién me da la regla de clasificación?

Estimación de  $g^{op}$

- $X$  v.a. discreta
- Posibles *etiquetas*. Caso binario  $\mathcal{Y} = \{0, 1\}$
- Clasificador: Regla que asigna a cada  $x \in \mathcal{X}$  un elemento  $y \in \mathcal{Y}$
- $g^{op}$  Optimo: Regla de Bayes - Caso binario

$$g^{op}(x) = \begin{cases} 1 & \text{si } \mathbb{P}(Y = 1 \mid X = x) \geq 1/2, \\ 0 & \text{si } c. c. \end{cases}$$

# Todo muy lindo, pero.... ¿quién me da la regla de clasificación?

Estimación de  $g^{op}$

- $X$  v.a. discreta
- Posibles *etiquetas*. Caso binario  $\mathcal{Y} = \{0, 1\}$
- Clasificador: Regla que asigna a cada  $x \in \mathcal{X}$  un elemento  $y \in \mathcal{Y}$
- $g^{op}$  Optimo: Regla de Bayes - Caso binario

$$g^{op}(x) = \begin{cases} 1 & \text{si } \mathbb{P}(Y = 1 \mid X = x) \geq 1/2, \\ 0 & \text{si c. c.} \end{cases}$$

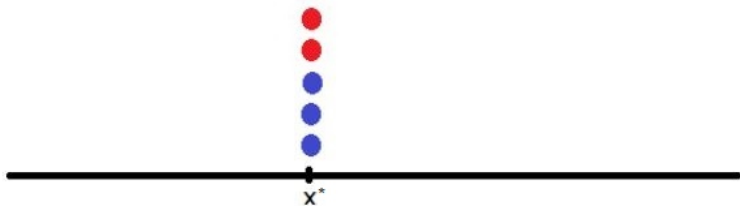
Datos:  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

Usaremos el método *Plug-in*, pero....

¿cómo podríamos estimar  $\mathbb{P}(Y = 1 \mid X = x)$  o  $\mathbb{P}(Y = 0 \mid X = x)$  ?

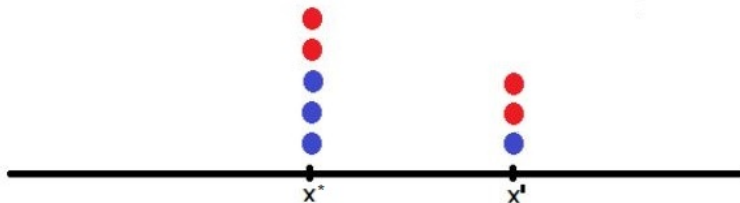
# Un método democrático: regla de la mayoría

$X$  discreta



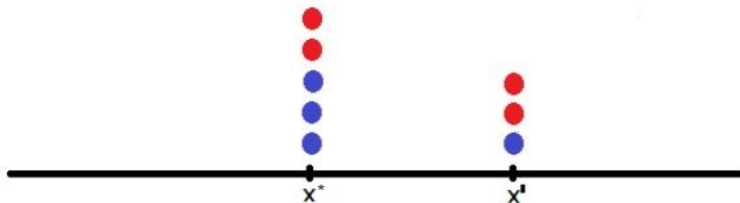
# Un método democrático: regla de la mayoría

$X$  discreta



# Un método democrático: regla de la mayoría

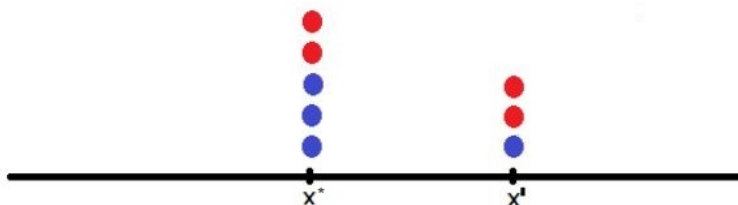
$X$  discreta



$$\frac{\sum_{i=1}^n y_i I_{\{x_i=x\}}}{\sum_{i=1}^n I_{\{x_i=x\}}}$$

# Un método democrático: regla de la mayoría

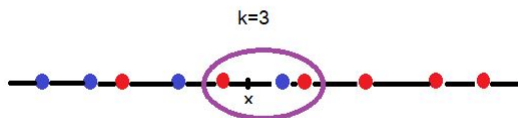
$X$  discreta



$$\frac{\sum_{i=1}^n y_i I_{\{x_i=x\}}}{\sum_{i=1}^n I_{\{x_i=x\}}} \leftrightarrow \frac{\sum_{i=1}^n y_i I_{\{|x_i-x|=0\}}}{\sum_{i=1}^n I_{\{|x_i-x|=0\}}}$$

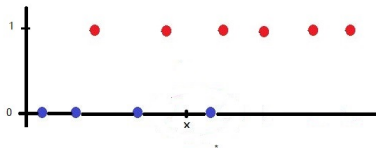
# Un método democrático: regla de la mayoría

$X$  continua



# Un método democrático: regla de la mayoría

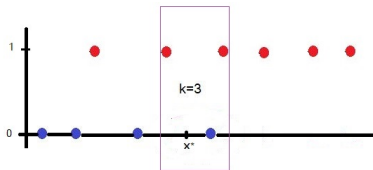
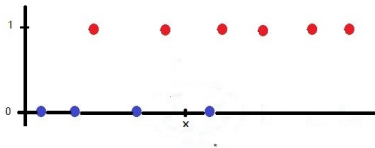
$X$  continua





# Un método democrático: regla de la mayoría

$X$  continua



## $k$ —Vecinos más cercanos ( $k$ NN: $k$ -nearest neighbours)

El método de  $k$ —Vecinos más cercanos es uno de los métodos existentes para estimar la distribución condicional de  $Y$  dado  $X$  y para después clasificar una observación en la clase con la mayor probabilidad estimada.

- Elegimos  $k$  un entero positivo y un punto  $x$  para clasificar.
- El clasificador  $k$ NN identifica el conjunto de los  $k$  puntos más cercanos a  $x$ . Sea  $N_x$  dicho conjunto.
- Estima a  $P(Y = 1 \mid X = x)$  por la fracción de puntos en  $N_x$  cuya etiqueta es igual a 1:

$$\hat{\mathbb{P}}(Y = 1 \mid X = x) = \frac{1}{k} \sum_{i \in N_x} \mathcal{I}_{\{y_i=1\}}$$

- Análogamente estimamos  $P(Y = 0 \mid X = x)$

¿Cómo elegimos el parámetro  $k$ ?

## Otra forma: estimación por núcleos

Otra manera de estimar a  $P(Y = 1 \mid X = x)$  podría ser considerar un entorno  $(x - h, x + h)$  y repetir el procedimiento anterior.

- Elegimos  $h > 0$  y un punto  $x$  para clasificar.
- El clasificador identifica en el intervalo  $(x - h, x + h)$  los puntos con etiqueta 1 y 0
- Estima a  $P(Y = 1 \mid X = x)$  por la fracción de puntos en  $(x - h, x + h)$  cuya etiqueta es igual a 1:

$$\hat{\mathbb{P}}(Y = 1 \mid X = x) = \frac{\sum_{i=1}^n y_i I_{(x-h, x+h)}(x_i)}{\sum_{i=1}^n I_{(x-h, x+h)}(x_i)}$$

¿Cómo elegimos el parámetro  $h$ ?

...quedan pendientes algunas preguntas que iremos resolviendo...