Clasificación

Ana M. Bianco & Paula M. Spano

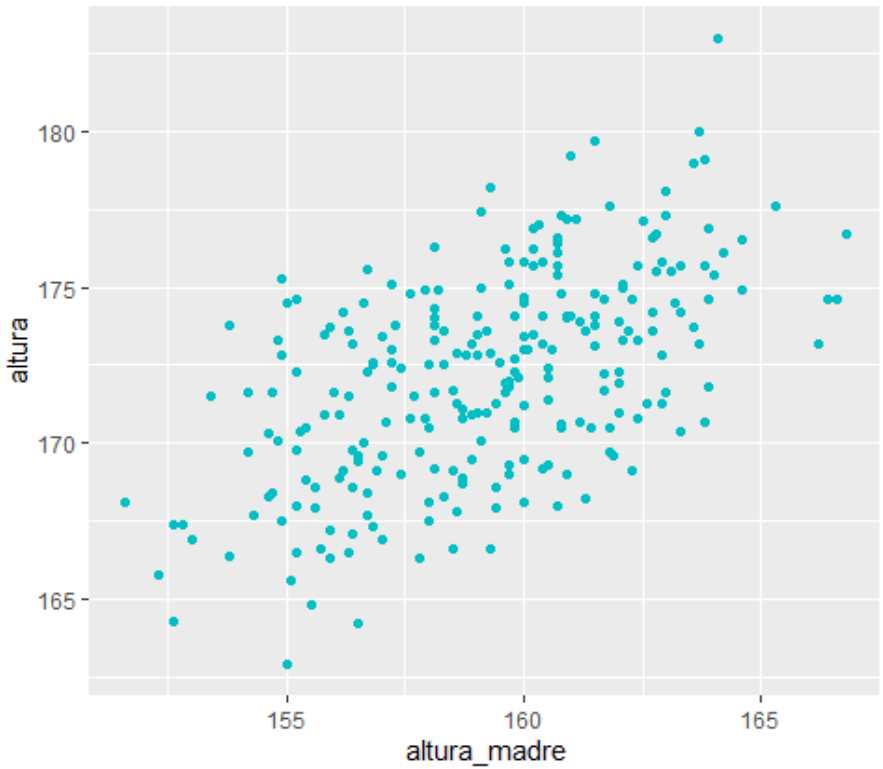
Introducción al Aprendizaje Estadístico

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

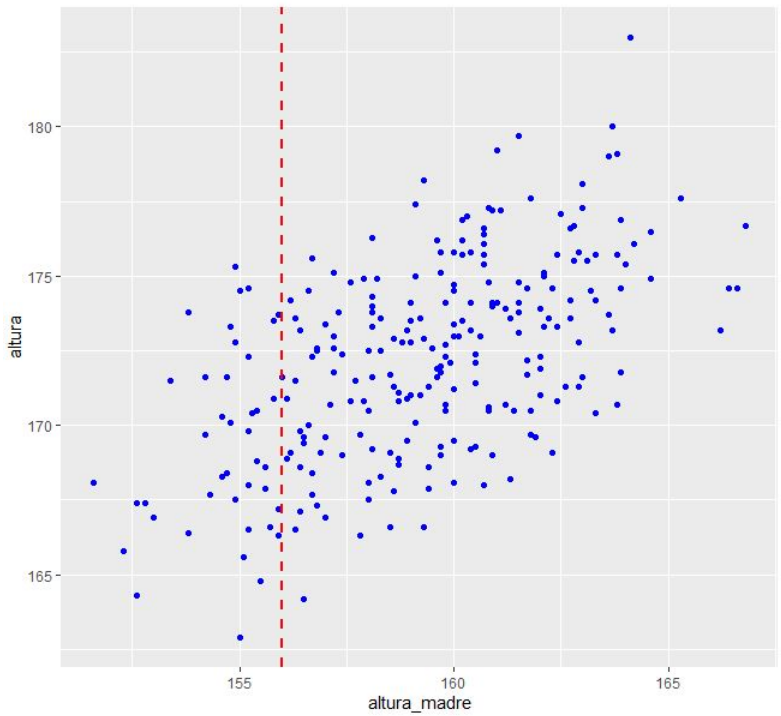
- Variables presentes
  - Y: Altura (cuantitativa)
  - X_1: Género
  - X_2: Contextura de la madre
  - X_3: Altura de la madre

Género Masculino: Problema de Predicción

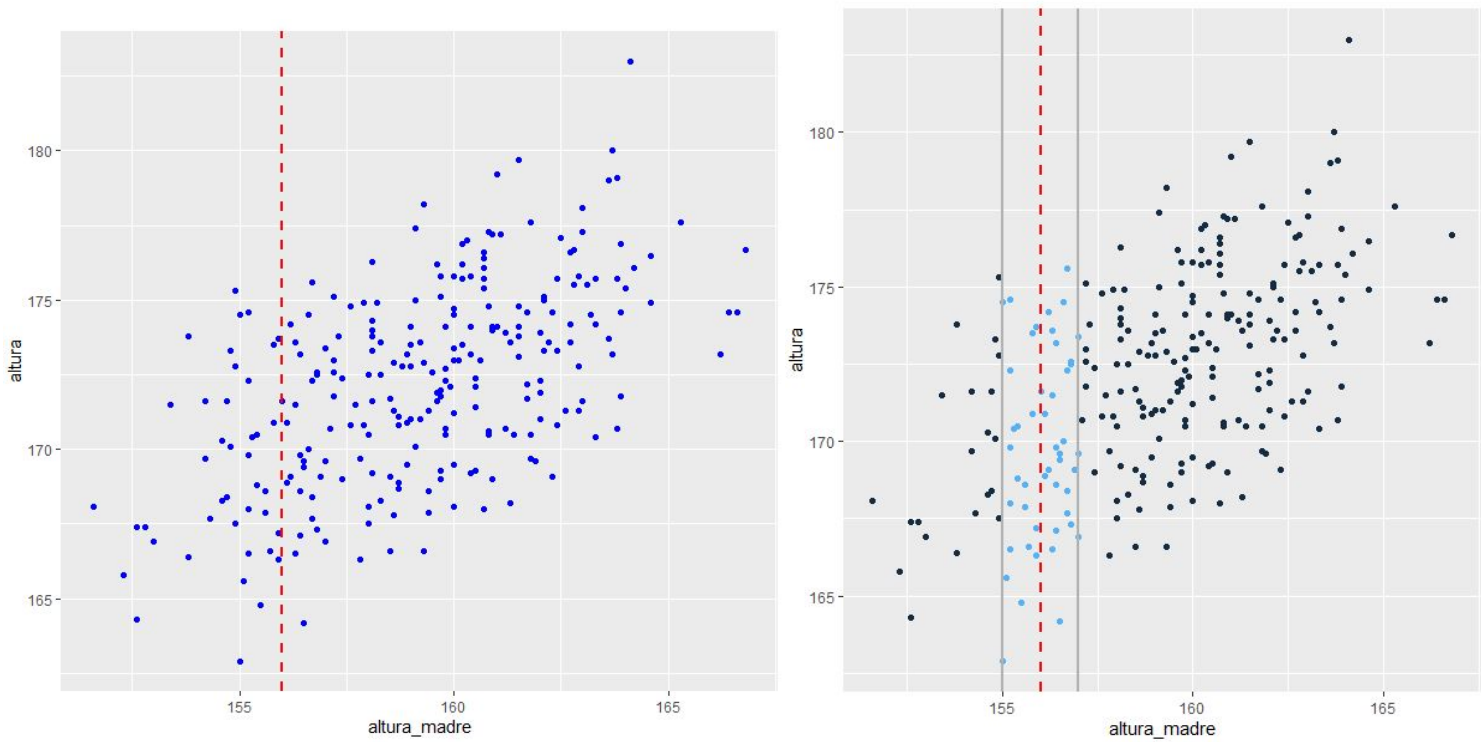


- Marco de Regresión
  1. **Variable de respuesta (Outcome)**
     Y: altura del hijo
  2. **Covariable (independiente, explicativa, regresora, predictora, feature)**:
     X: altura de la madre

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.

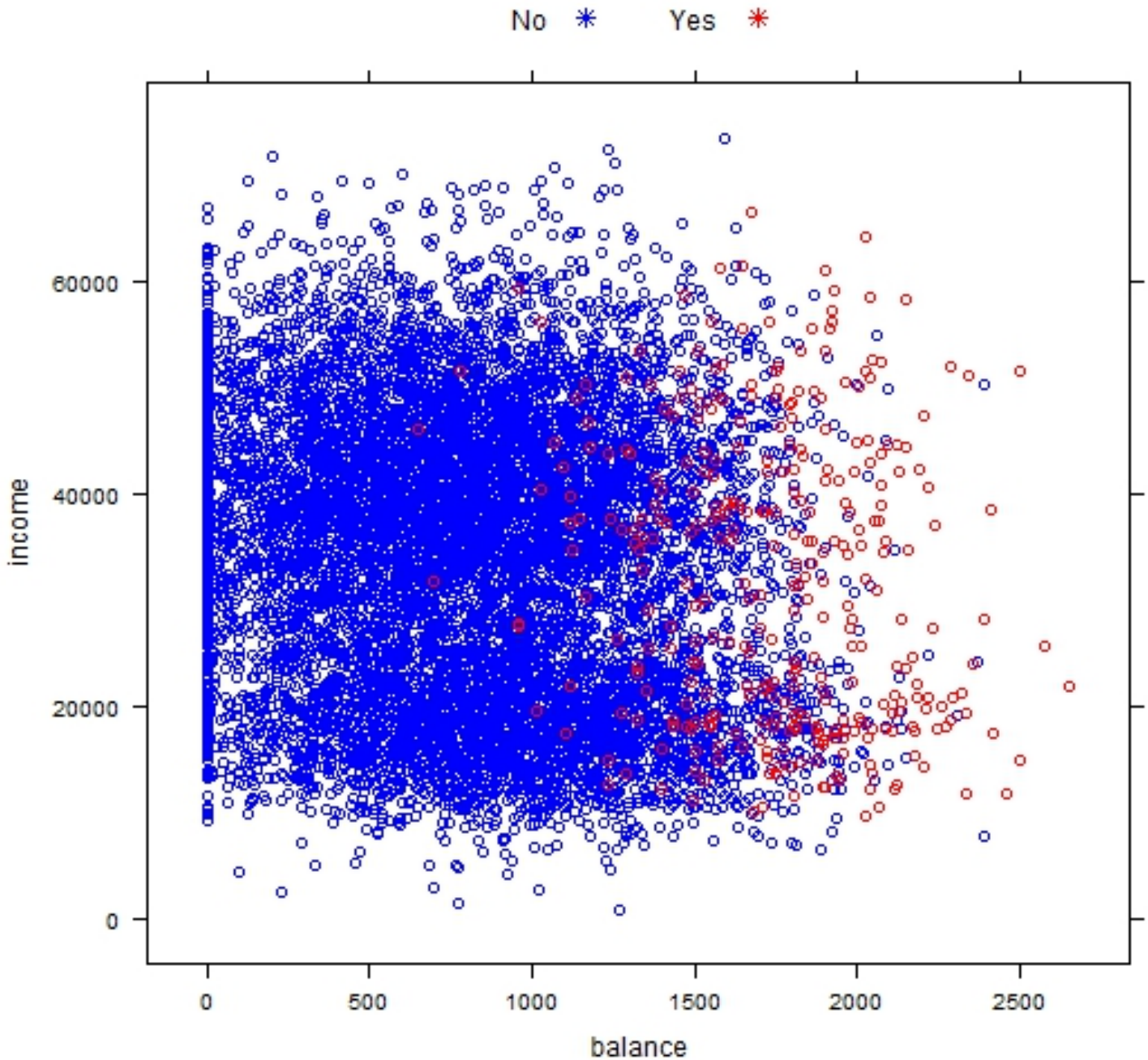
Variables registradas:

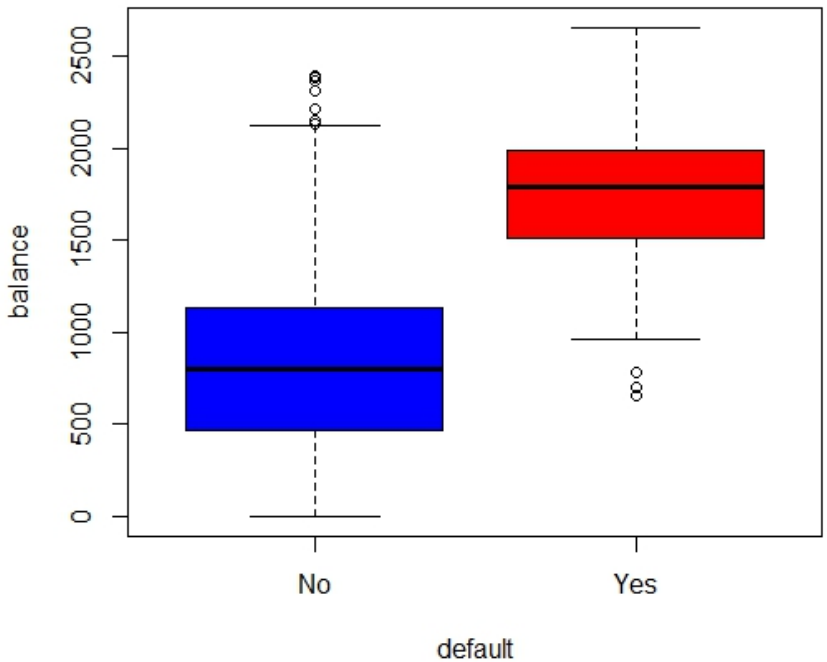
- X_1: **balance** (saldo tarjeta)
- X_2: **income** (ingreso anual)
- X_3: **student** (si - no)

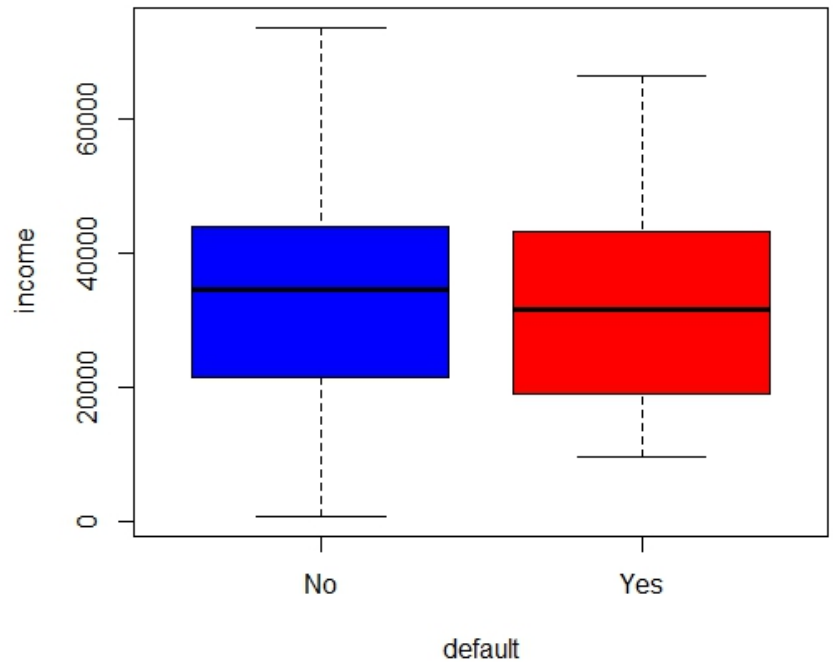
- $Y = \begin{cases} 1 & \text{(Yes)} \quad \text{default} \\ 0 & \text{(No)} \quad \text{c.c.} \end{cases}$

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
```


No
Yes
income
60000
40000
20000
0
0
500
1000
1500
2000
2500
balance


balance
0
500
1000
1500
2000
2500
No
Yes
default


income
0
20000
40000
60000
No
Yes
default

Predicción

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

Predicción

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

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 $x \in \{0, 1, \ldots, 8\}$ 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 $x \in \{0, 1, \ldots, 8\}$ un tipo de hospedador: $\{0(A), 1(R)\}$

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

Clasificador

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

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

Otro clasificador

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

Clasificador

- Información disponible $x \in \mathcal{X}$.
- Posibles *etiquetas*. Caso binario $\mathcal{Y} = \{0, 1\}$
- Posibles *etiquetas*. Caso general $\mathcal{Y} = \{y_1, \ldots, y_k\}$
- 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 $y \in \mathcal{Y}$

$$\text{Clasificador} \quad g : \mathcal{X} \to \mathcal{Y}$$

- 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

$$g^{op}$$

Volvamos a El Patito Feo

clasificador I:

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

Calcular

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

cuando la conjunta (X, Y) es

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	0.0167

Volvamos a El Patito Feo

clasificador I:

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	$\surd$	$\surd$	$\surd$	$\surd$	×	×	×	×	×
1	×	×	×	×	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$

Volvamos a El Patito Feo

clasificador I:

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

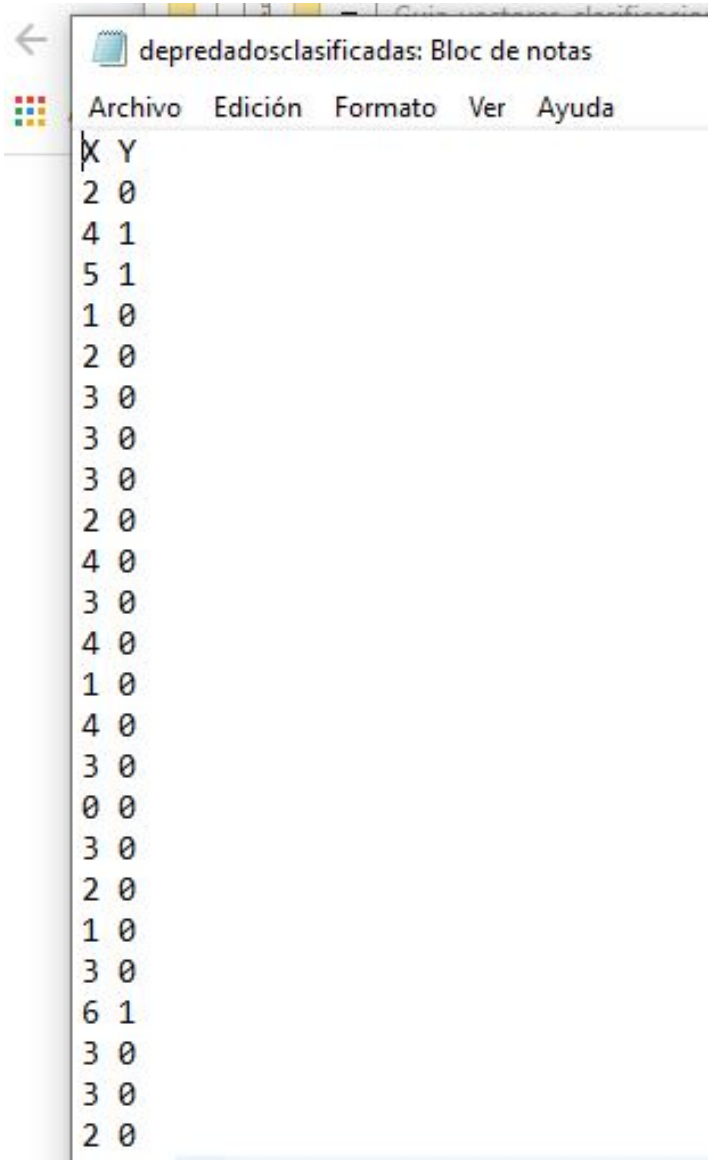
Conjunta (X, Y)

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	0.0167

Si sumamos las probabilidades que están en rojo, obtenemos

Error de Clasificación de $g_1 = 0.1757$

En la práctica, ¿cómo hacemos?


depredadosclasificadas: Bloc de notas

Archivo Edición Formato Ver Ayuda

```
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:

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

En la práctica, ¿cómo hacemos?

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

$$\underline{\text{Datos}}: (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

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

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

En la práctica, ¿cómo hacemos?

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

$$\underline{\text{Datos}}: (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

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

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

$$\text{en R:} \quad \text{mean}(g(x) ! = y)$$

Tarea: 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

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

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

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

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

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

$$\begin{aligned} L(g) &= \mathbb{P}\left(g(X) \neq Y\right) = E_{XY}[\mathcal{I}_{(g(X)!=Y)}] \\ &= E_X E_{Y|X}[\mathcal{I}_{(g(X)!=Y)}] \\ &= E_X[\mathcal{I}_{(g(X)=0)}\mathbb{P}(Y=1 \mid X=x) + \mathcal{I}_{(g(X)=1)}\mathbb{P}(Y=0 \mid X=x)]. \end{aligned}$$

Luego, para minimizar $L(g)$ es suficiente con minimizar para todo x

$$\mathcal{I}_{(g(x)=0)}\mathbb{P}(Y=1 \mid X=x) + \mathcal{I}_{(g(x)=1)}\mathbb{P}(Y=0 \mid X=x)$$

Teniendo en cuenta que las dos indicadoras son mutuamente excluyentes, alcanza con tomar

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

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

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

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

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

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

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

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

Notemos que

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

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

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

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

De Bayes, tenemos que

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

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

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

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

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

resulta

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

Estimadores Plug–in

Otra vez en el jardín de los senderos que se bifurcan...

Estimación g^{op} Optimo: Métodos discriminativos

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

Aquí se estima

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

Estimación g^{op} Optimo: Métodos generativos

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

Se estiman

- f_0 y f_1
- $p_1 = \mathbb{P}(Y=1)$ o $p_0 = \mathbb{P}(Y=0)$.

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

Estimación de g^{op} bajo el primer enfoque.

- 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

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

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

Estimación de g^{op} bajo el primer enfoque.

- 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

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

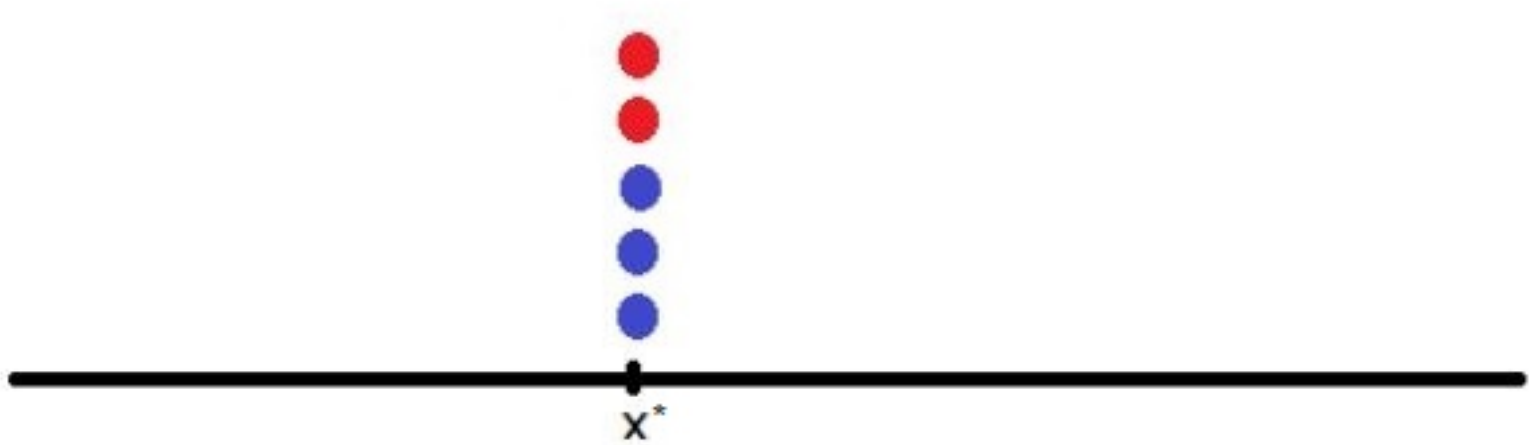
Datos: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

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

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

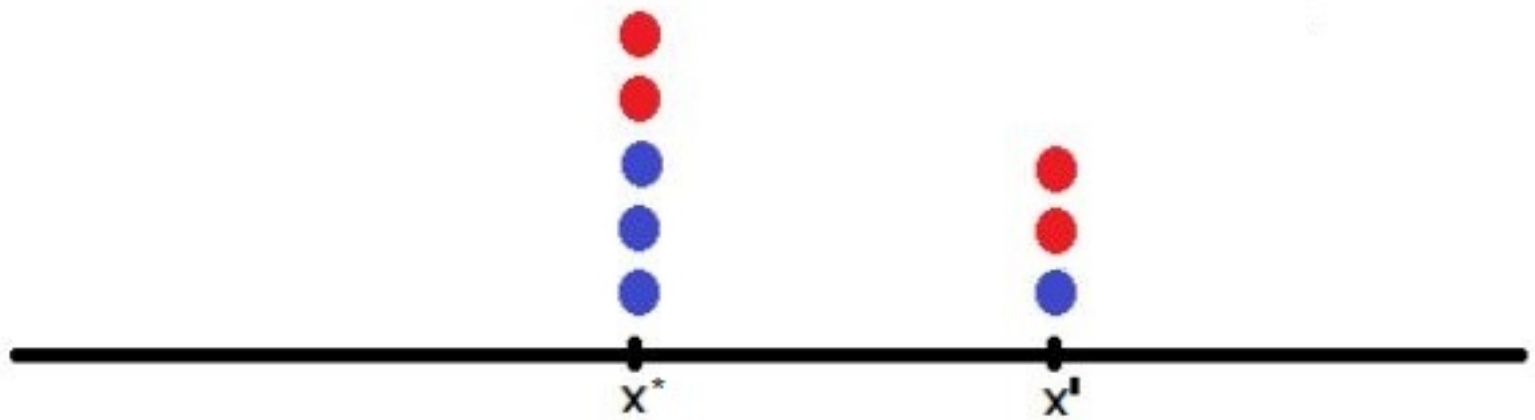
Un método democrático: regla de la mayoría

X discreta



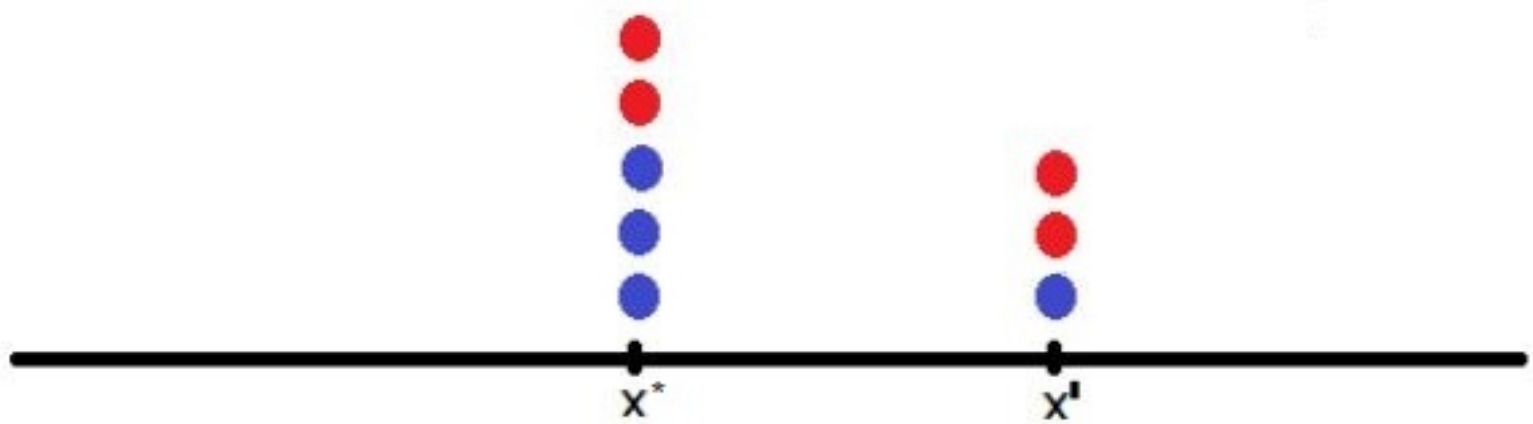
Un método democrático: regla de la mayoría

X discreta



Un método democrático: regla de la mayoría

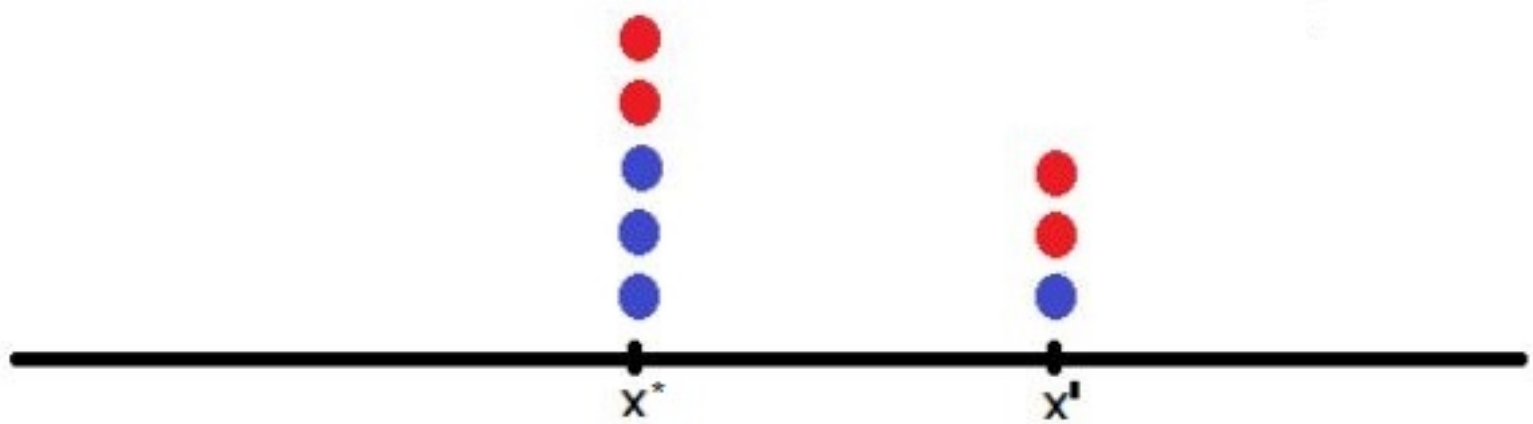
X discreta



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

Un método democrático: regla de la mayoría

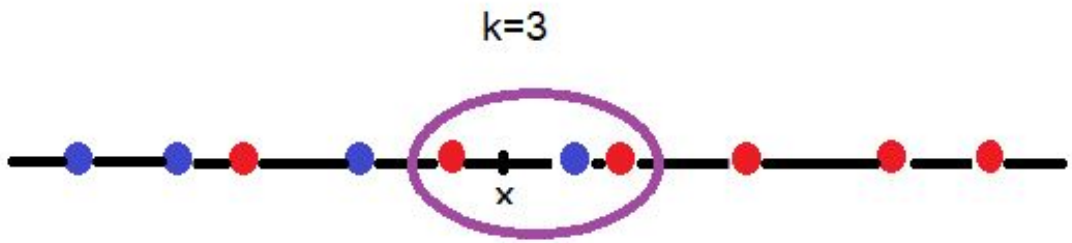
X discreta



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

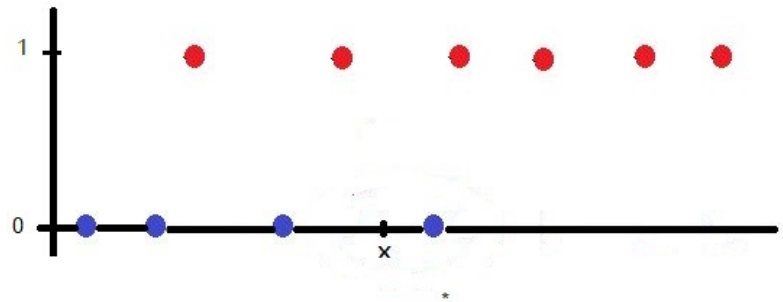
Un método democrático: regla de la mayoría

X continua



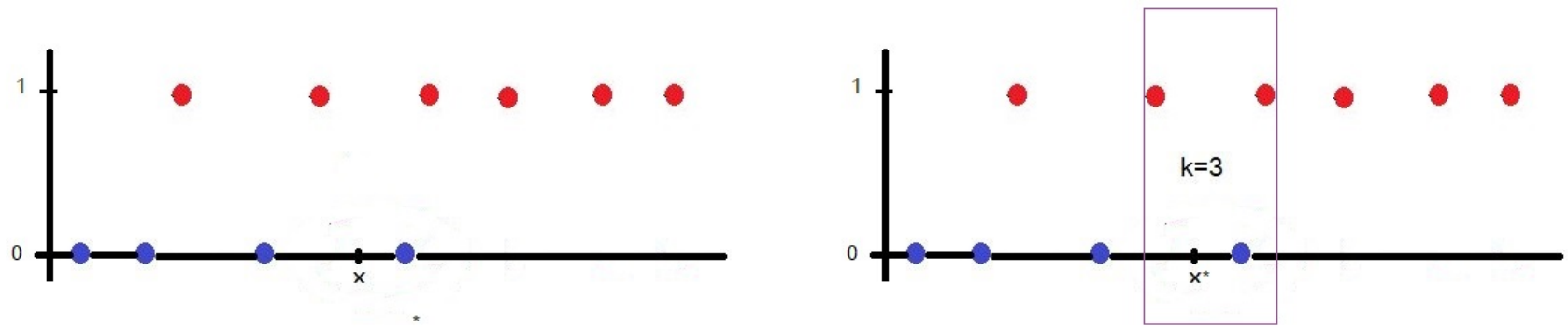
Un método democrático: regla de la mayoría

X continua



Un método democrático: regla de la mayoría

X continua



$k-$Vecinos más cercanos (kNN: 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 kNN identifica el conjunto de los k puntos más cercanos a x. Sea N_x dicho conjunto.
- Estima a $\mathbb{P}(Y = 1 \mid X = x)$ por la fracción de puntos en N_x cuya etiqueta es igual a 1:

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

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

¿Cómo elegimos el parámetro k?

Otra forma: estimación por núcleos

Otra manera de estimar a $\mathbb{P}(Y = 1 \mid X = x)$ 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 $\mathbb{P}(Y = 1 \mid X = x)$ por la fracción de puntos en $(x - h, x + h)$ cuya etiqueta es igual a 1:

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

¿Cómo elegimos el parámetro h?

Estimación g^{op} Optimo: Métodos generativos

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

Se estiman

- f_0 y f_1
- $p_1 = \mathbb{P}(Y=1)$ o $p_0 = \mathbb{P}(Y=0)$.

Regla plug–in de Bayes

- (X, Y) vector aleatorio
- Datos: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

- Estimaciones de p_1 y p_0: $\widehat{p}_1 = \frac{n_1}{n}$ y $\widehat{p}_0 = \frac{n_0}{n}$, donde $n_i = \#\{y_i = i\}, i = 0, 1$.

Regla plug–in de Bayes

- (X, Y) vector aleatorio
- Datos: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

- Estimaciones de p_1 y p_0: $\widehat{p}_1 = \frac{n_1}{n}$ y $\widehat{p}_0 = \frac{n_0}{n}$, donde $n_i = \#\{y_i = i\}, i = 0, 1$.

- Estimamos $f_1(x)$ mediante $\widehat{f}_1(x)$.

- Estimamos $f_0(x)$ mediante $\widehat{f}_0(x)$.

- Hacemos un plug–in en la regla $g^{op}(x)$:

$$\widehat{g}(x) = \begin{cases} 1 & \text{si } \widehat{f}_1(x)\, \widehat{p}_1 \geq \widehat{f}_0(x)\, \widehat{p}_0, \\ 0 & \text{si } c.c. \end{cases}$$

¿Cómo obtenemos las densidades estimadas $\widehat{f}_1(x)$ y $\widehat{f}_0(x)$?

¿Cómo obtenemos las densidades estimadas $\widehat{f}_1(x)$ y $\widehat{f}_0(x)$?

Opción 1: Enfoque No Paramétrico

- Estimamos $f_1(x)$ no paramétricamente con $\widehat{f}_{1,h_1}(x)$
- Estimamos $f_0(x)$ no paramétricamente con $\widehat{f}_{0,h_0}(x)$
- Hacemos un plug–in en la regla $g^{op}(x)$:

$$\widehat{g}_{h_0,h_1}(x) = \begin{cases} 1 & \text{si } \widehat{f}_{1,h_1}(x)\,\widehat{p}_1 \geq \widehat{f}_{0,h_0}(x)\,\widehat{p}_0, \\ 0 & \text{si } c.c. \end{cases}$$

- $\widehat{g}_{h_0,h_1}$, $h_0, h_1 =?$

¿Cómo elegimos k, h o (h_0, h_1)?

En cada caso, tenemos una $\widehat{g}_t$, ¿cómo elegimos el parámetro o vector de parámetros t?

¿Cómo elegimos k, h o (h_0, h_1)?

En cada caso, tenemos una $\widehat{g}_t$, ¿cómo elegimos el parámetro o vector de parámetros t?

Splitting the data

do as well as possible within a given class of rules... This is achieved by splitting the data into a training sequence and a testing sequence. (DGL)

Test Error

Podemos repetir el mismo razonamiento que hicimos para regresión:

- $\widehat{r}_{\mathcal{D}_n}$: $\widehat{r}$ estimador construido a partir de $\mathcal{D}_n$

- Test Error

$$\text{Error}_{\mathcal{D}_n} = \mathbb{E}\left[\{Y_{\text{new}} - \widehat{r}_{\mathcal{D}_n}(X_{\text{new}})\}^2 \mid \mathcal{D}_n\right]$$

Test Error

Podemos repetir el mismo razonamiento que hicimos para regresión:

- $\widehat{g}_{\mathcal{D}_n}$: $\widehat{g}$ regla construida a partir de $\mathcal{D}_n$

- Test Error

$$\text{Error}_{\mathcal{D}_n} = \mathbb{E}\left[\{Y_{\text{new}} - \widehat{g}_{\mathcal{D}_n}(X_{\text{new}})\}^2 \mid \mathcal{D}_n\right]$$

Test Error

Podemos repetir el mismo razonamiento que hicimos para regresión:

- $\widehat{g}_{\mathcal{D}_n}$: $\widehat{g}$ regla construida a partir de $\mathcal{D}_n$
- Test Error

$$\text{Error}_{\mathcal{D}_n} = \mathbb{E}\left[\{Y_{\text{new}} - \widehat{g}_{\mathcal{D}_n}(X_{\text{new}})\}^2 \mid \mathcal{D}_n\right]$$

- equivalente a

$$\mathbb{P}\left(\widehat{g}_{\mathcal{D}_n}(X_{\text{new}}) \neq Y_{\text{new}} \mid \mathcal{D}_n\right)$$

Muchos Datos – Menos Datos

Consideramos las mismas estrategias que ya mencionamos para estimación de la densidad y regresión.

Cross Validation: Leave one out - Fórmulas

t: tunning parameter

$$\mathsf{CV}(t) = \widehat{L(t)} = \frac{1}{n}\sum_{i=1}^{n} L_i(t)$$

- Regresión:

$$L_i(t) = \{Y_i - \widehat{r}_t^{(-i)}(X_i)\}^2$$

$$t_{\mathsf{opt}} = \mathsf{argmin}\widehat{L(t)}$$

- Clasificación:

$$L_i(t) = \{Y_i - \widehat{g}_t^{(-i)}(X_i)\}^2 = \mathcal{I}_{\{\widehat{g}_t^{(-i)}(X_i) \neq Y_i\}}$$

Cross Validation: Leave one out - Fórmulas

t: tunning parameter

$$\mathsf{CV}(t) = \widehat{L(t)} = \frac{1}{n}\sum_{i=1}^{n} L_i(t)$$

- Clasificación:

$$L_i(t) = \mathcal{I}_{\{\widehat{g}_t^{(-i)}(X_i) \neq Y_i\}}$$

Cross Validation: K folders - Fórmulas

t: tunning parameter

$$\widehat{L(t)} = \frac{1}{K}\sum_{k=1}^{K} L_k(t)$$

- Regresión:

$$L_k(t) = \frac{1}{|\mathcal{T}_k^c|}\sum_{j\in\mathcal{T}_k^c}\{Y_j - \widehat{r}_{t,\mathcal{T}_k}(X_i)\}^2$$

$$t_{\text{opt}} = \text{argmin}\widehat{L(t)}$$

- Clasificación:

$$L_k(t) = \frac{1}{|\mathcal{T}_k^c|}\sum_{j\in\mathcal{T}_k^c}\mathcal{I}_{\widehat{g}_{t,\mathcal{T}_k}(X_j)\neq Y_j}$$

Algunas otras cuestiones en Regresión y Clasificación

- ¿Qué pasa en mayor dimensión? Maldición de la Dimensión
- Clasificación: alternativa inocente Bayes Naive
- Métodos Paramétricos