

# Clasificación

Mathieu Kessler

Departamento de Matemática Aplicada y Estadística  
Universidad Politécnica de Cartagena

Cartagena

# El problema de la clasificación

- Basándonos en el valor de características, queremos clasificar cada individuo en una determinada categoría.

# El problema de la clasificación

- Basándonos en el valor de características, queremos clasificar cada individuo en una determinada categoría.
- Empezaremos con la clasificación en dos categorías posibles.

# El problema de la clasificación

- Basándonos en el valor de características, queremos clasificar cada individuo en una determinada categoría.
- Empezaremos con la clasificación en dos categorías posibles.
- Las características:  $x_0, x_1, \dots, x_k$ ; consideraremos la variable “respuesta” y dicotómica: toma valores 0 ó 1.

# El problema de la clasificación

- Basándonos en el valor de características, queremos clasificar cada individuo en una determinada categoría.
- Empezaremos con la clasificación en dos categorías posibles.
- Las características:  $x_0, x_1, \dots, x_k$ ; consideraremos la variable “respuesta” y dicotómica: toma valores 0 ó 1.
- Ejemplos:

# El problema de la clasificación

- Basándonos en el valor de características, queremos clasificar cada individuo en una determinada categoría.
- Empezaremos con la clasificación en dos categorías posibles.
- Las características:  $x_0, x_1, \dots, x_k$ ; consideraremos la variable “respuesta” y dicotómica: toma valores 0 ó 1.
- Ejemplos:
  - 1 Queremos clasificar los emails en SPAM o NO SPAM.

# El problema de la clasificación

- Basándonose en el valor de características, queremos clasificar cada individuo en una determinada categoría.
- Empezaremos con la clasificación en dos categorías posibles.
- Las características:  $x_0, x_1, \dots, x_k$ ; consideraremos la variable “respuesta” y dicotómica: toma valores 0 ó 1.
- Ejemplos:
  - 1 Queremos clasificar los emails en SPAM o NO SPAM.
  - 2 Queremos clasificar operaciones de compra online en FRAUDULENTA o NO FRAUDULENTA.

# El problema de la clasificación

- Basándonos en el valor de características, queremos clasificar cada individuo en una determinada categoría.
- Empezaremos con la clasificación en dos categorías posibles.
- Las características:  $x_0, x_1, \dots, x_k$ ; consideraremos la variable “respuesta” y dicotómica: toma valores 0 ó 1.
- Ejemplos:
  - 1 Queremos clasificar los emails en SPAM o NO SPAM.
  - 2 Queremos clasificar operaciones de compra online en FRAUDULENTA o NO FRAUDULENTA.
  - 3 Queremos clasificar tumores en BENIGNO o MALIGNO.



# El problema de la clasificación

- Basándonos en el valor de características, queremos clasificar cada individuo en una determinada categoría.
- Empezaremos con la clasificación en dos categorías posibles.
- Las características:  $x_0, x_1, \dots, x_k$ ; consideraremos la variable “respuesta” y dicotómica: toma valores 0 ó 1.
- Ejemplos:
  - 1 Queremos clasificar los emails en SPAM o NO SPAM.
  - 2 Queremos clasificar operaciones de compra online en FRAUDULENTA o NO FRAUDULENTA.
  - 3 Queremos clasificar tumores en BENIGNO o MALIGNO.
  - 4 Queremos clasificar alumnos de nuevo ingreso en “EGRESARÁ” o “ABANDONARÁ”.

## Ejemplo: predicción del abandono

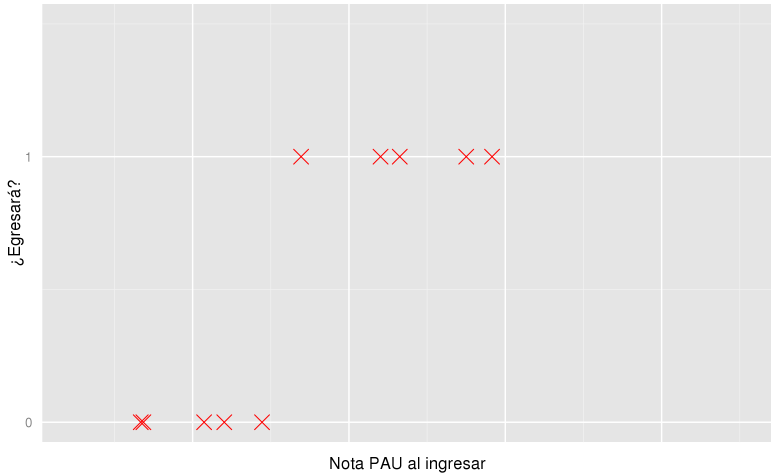
Queremos predecir el abandono de un alumno en función de su nota de PAU al ingresar.

## Ejemplo: predicción del abandono

Codificamos:  $y = 1 \leftrightarrow$  “Egresará”,  $y = 0 \leftrightarrow$  “Abandonará”.

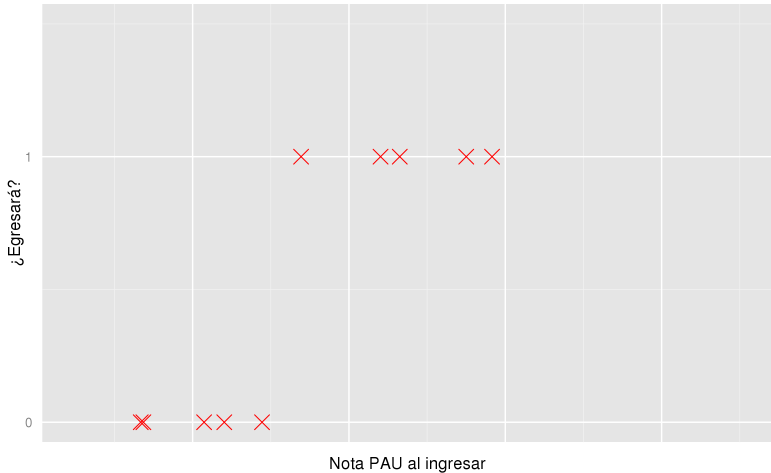
# Ejemplo: predicción del abandono

Codificamos:  $y = 1 \leftrightarrow$  “Egresará”,  $y = 0 \leftrightarrow$  “Abandonará”.



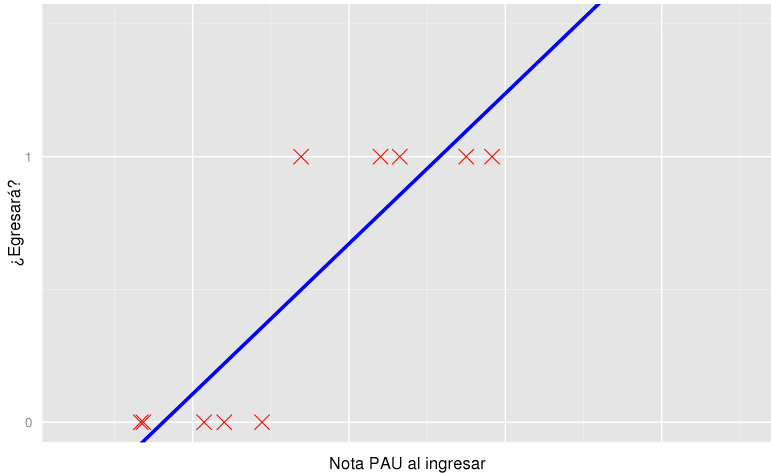
# Ejemplo: predicción del abandono

Si usamos regresión lineal, la recta ajustada  $y = h_{\theta}(x)$  es



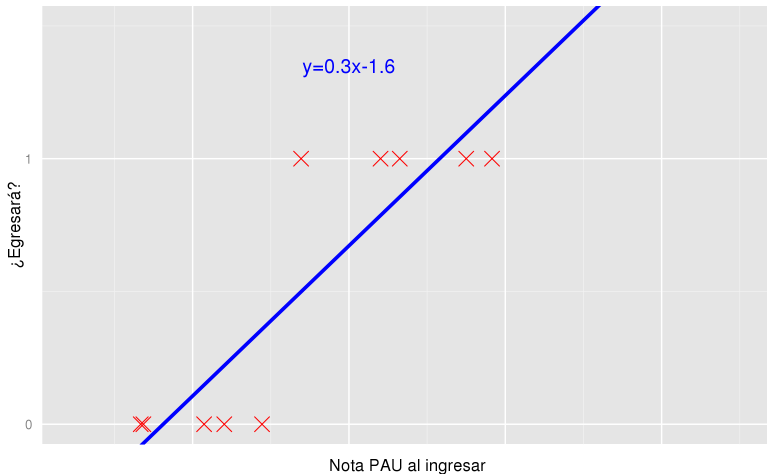
# Ejemplo: predicción del abandono

Si usamos regresión lineal, la recta ajustada  $y = h_{\theta}(x)$  es



# Ejemplo: predicción del abandono

Si usamos regresión lineal, la recta ajustada  $y = h_{\theta}(x)$  es



## Ejemplo: predicción de abandono

¿ Podemos usar esta recta ajustada  $y = 0.3x - 1.6$  para hacer predicción ante un nuevo alumno que ingresa?

¡Sí! Podemos aprovecharla para definir una regla de decisión:



## Ejemplo: predicción de abandono

¿ Podemos usar esta recta ajustada  $y = 0.3x - 1.6$  para hacer predicción ante un nuevo alumno que ingresa?

¡Sí! Podemos aprovecharla para definir una regla de decisión:

- 1 Sustituimos la nota PAU de ingreso del nuevo alumno en la ecuación ajustada. Obtenemos  $\Rightarrow \hat{y}$ .

## Ejemplo: predicción de abandono

¿ Podemos usar esta recta ajustada  $y = 0.3x - 1.6$  para hacer predicción ante un nuevo alumno que ingresa?

¡Sí! Podemos aprovecharla para definir una regla de decisión:

- 1 Sustituimos la nota PAU de ingreso del nuevo alumno en la ecuación ajustada. Obtenemos  $\Rightarrow \hat{y}$ .
- 2 Si  $\hat{y} > 0.5$ , predecimos que egresará.

## Ejemplo: predicción de abandono

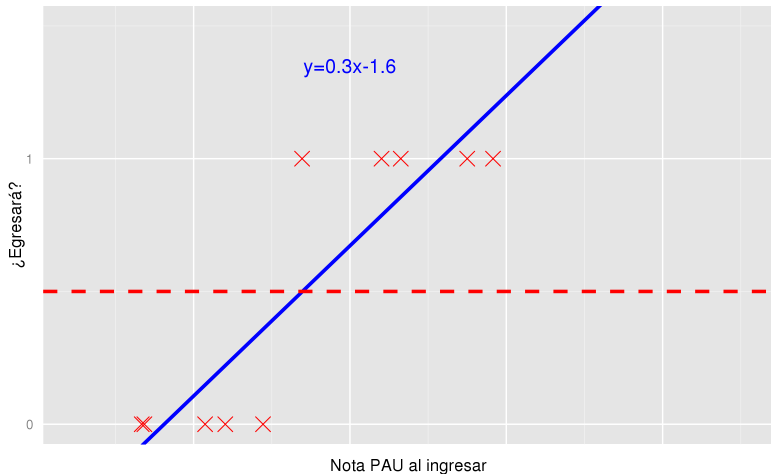
¿ Podemos usar esta recta ajustada  $y = 0.3x - 1.6$  para hacer predicción ante un nuevo alumno que ingresa?

¡Sí! Podemos aprovecharla para definir una regla de decisión:

- 1 Sustituimos la nota PAU de ingreso del nuevo alumno en la ecuación ajustada. Obtenemos  $\Rightarrow \hat{y}$ .
- 2 Si  $\hat{y} > 0.5$ , predecimos que egresará.
- 3 Si  $\hat{y} < 0.5$ , predecimos que abandonará.

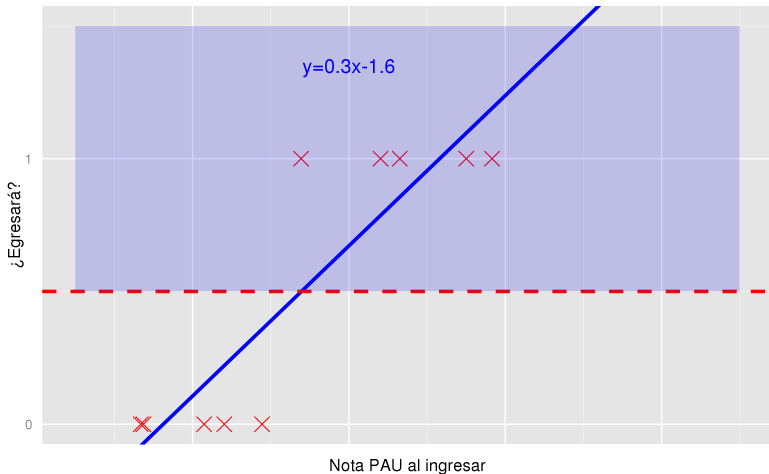
# Ejemplo: predicción del abandono

Gráficamente, nuestro criterio de decisión sobre  $\hat{y}$ :



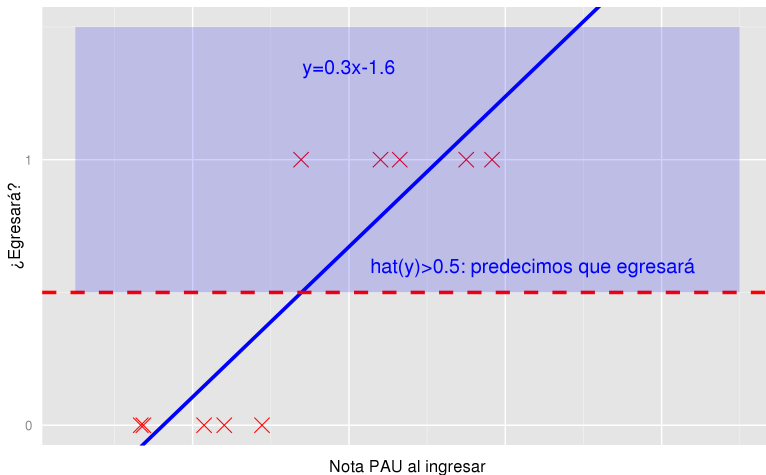
# Ejemplo: predicción del abandono

Gráficamente, nuestro criterio de decisión sobre  $\hat{y}$ :



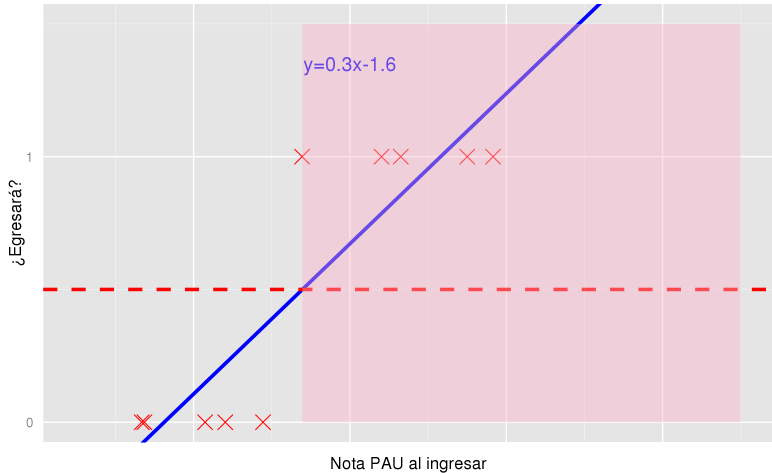
# Ejemplo: predicción del abandono

Gráficamente, nuestro criterio de decisión sobre  $\hat{y}$ :



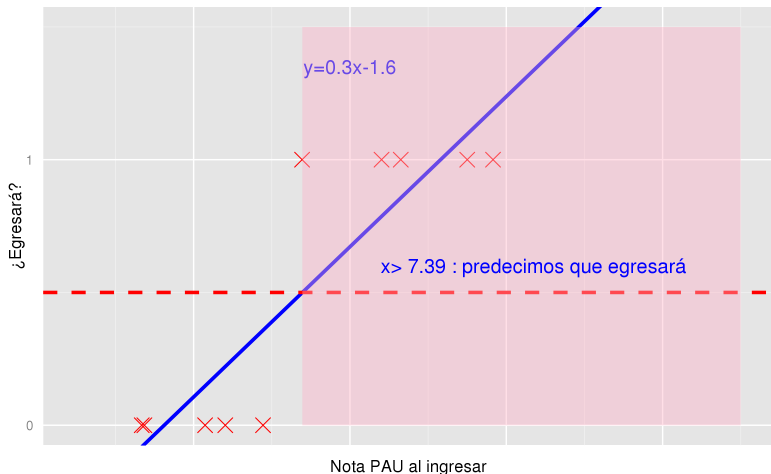
# Ejemplo: predicción del abandono

Lo traducimos en términos de  $x$ , :



# Ejemplo: predicción del abandono

Decir  $\hat{y} = 0.3x - 1.6 > 0.5$  es equivalente a  $x > 7.39$ ,





# Ejemplo: predicción del abandono

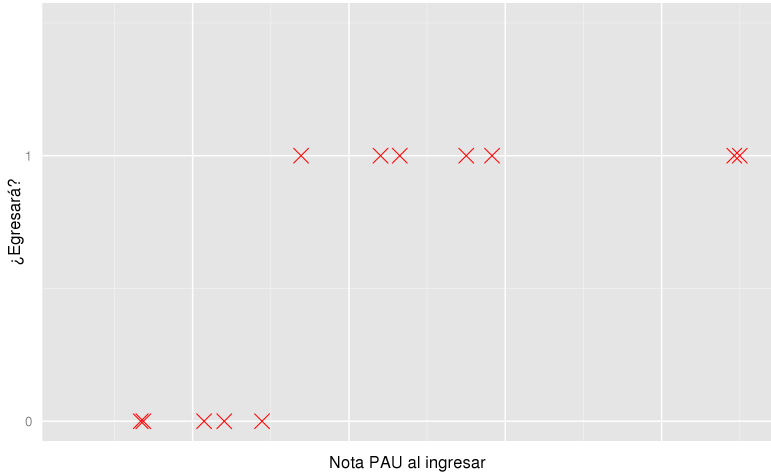
Pero nuestro criterio de decisión es sensible a datos atípicos:

# Ejemplo: predicción del abandono

Supongamos que tenemos estos puntos adicionales

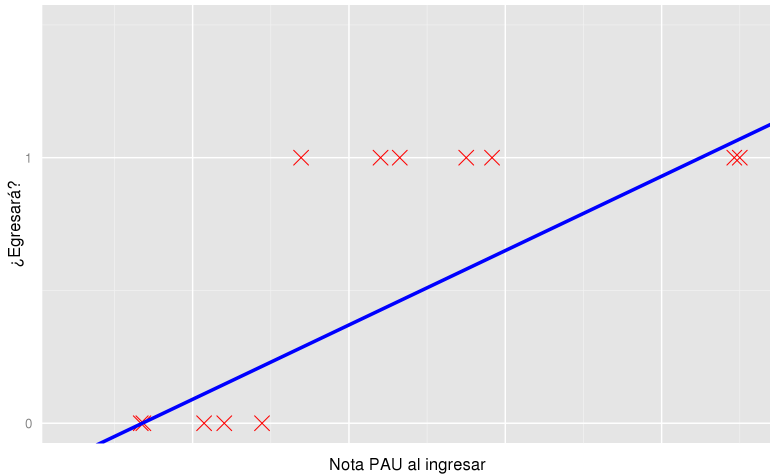
# Ejemplo: predicción del abandono

Supongamos que tenemos estos puntos adicionales



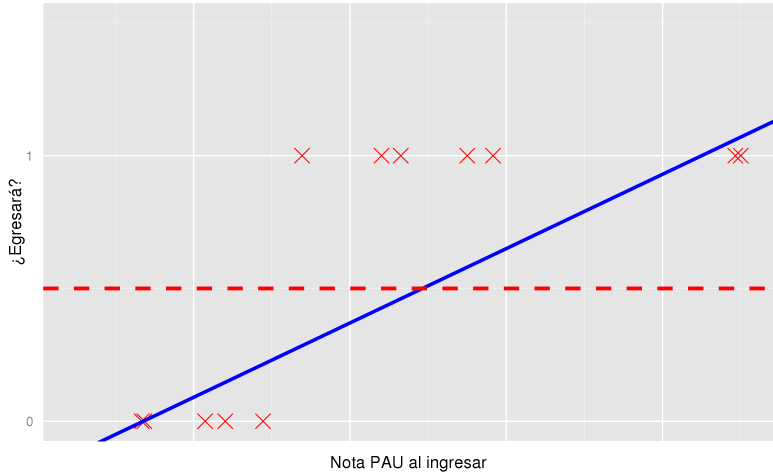
# Ejemplo: predicción del abandono

El ajuste cambia bastante



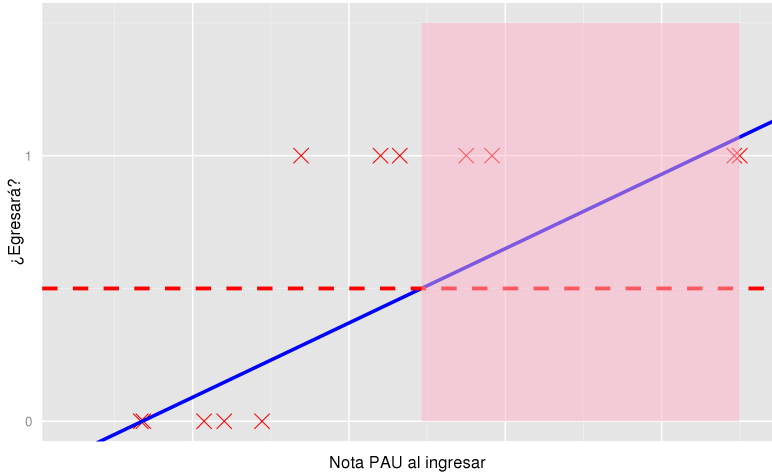
# Ejemplo: predicción del abandono

Y nuestro criterio de decisión es inadecuado



# Ejemplo: predicción del abandono

Y nuestro criterio de decisión es inadecuado



Además al ajustar una recta a nuestros datos binarios, el modelo  $h_{\theta}(x)$  puede tomar valores superiores a 1, o negativos...

# Regresión logística

Además al ajustar una recta a nuestros datos binarios, el modelo  $h_{\theta}(x)$  puede tomar valores superiores a 1, o negativos...

Pasamos a una función no lineal para ajustar estos datos binarios:

Usaremos como base la función logística:

$$g(z) = \frac{1}{1 + e^{-z}}.$$



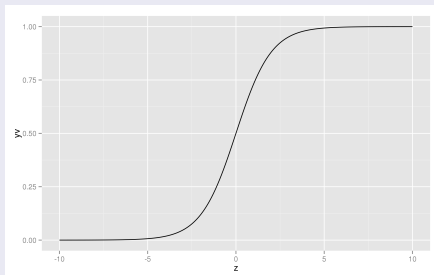
# Regresión logística

Además al ajustar una recta a nuestros datos binarios, el modelo  $h_{\theta}(x)$  puede tomar valores superiores a 1, o negativos...

Pasamos a una función no lineal para ajustar estos datos binarios:

Usaremos como base la función logística:

$$g(z) = \frac{1}{1 + e^{-z}}.$$



Ajustaremos a los datos binarios la hipótesis:

$$h_{\theta}(x) = g(x^T \theta)$$

Ajustaremos a los datos binarios la hipótesis:

$$h_{\theta}(x) = g(x^T \theta)$$

- 1** Si tenemos una única característica  $x_1$ :

$$h_{\theta}(x) = g(x^T \theta) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1)}},$$

Ajustaremos a los datos binarios la hipótesis:

$$h_{\theta}(x) = g(x^T \theta)$$

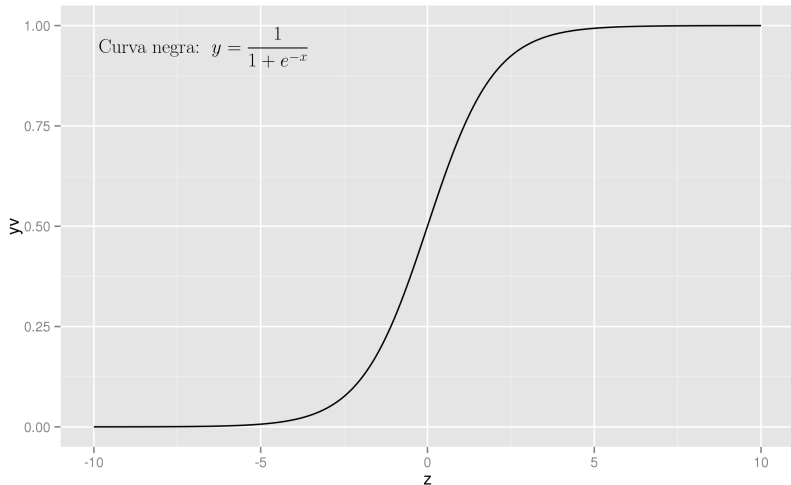
- 1 Si tenemos una única característica  $x_1$ :

$$h_{\theta}(x) = g(x^T \theta) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1)}},$$

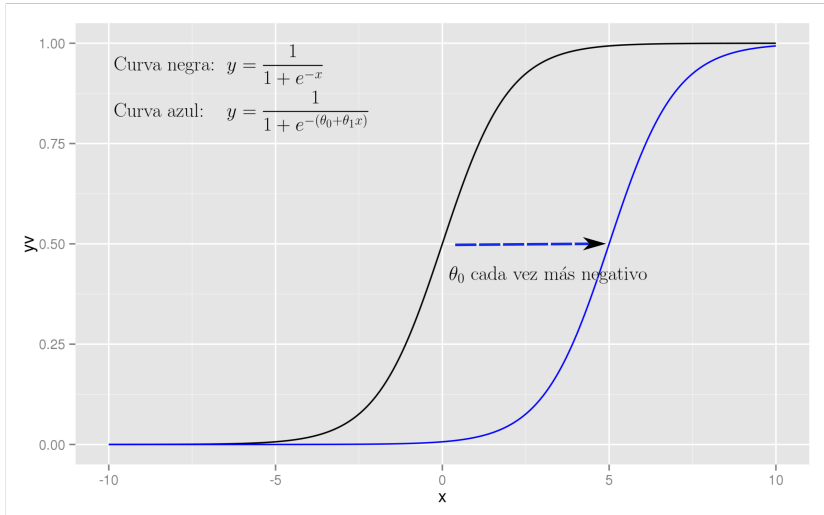
- 2 si tenemos  $k$  características  $x_1, x_2, \dots, x_k$ .

$$h_{\theta}(x) = g(x^T \theta) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1 + \dots + \theta_k x_k)}},$$

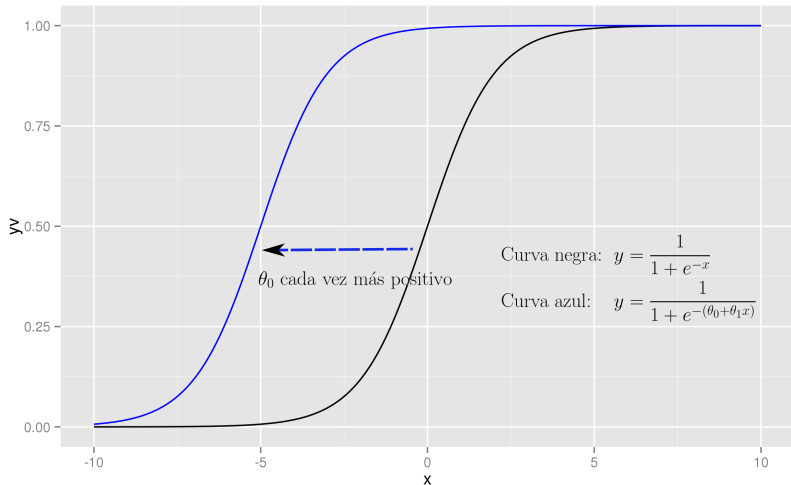
# En el caso de una característica, efecto de variar $\theta$



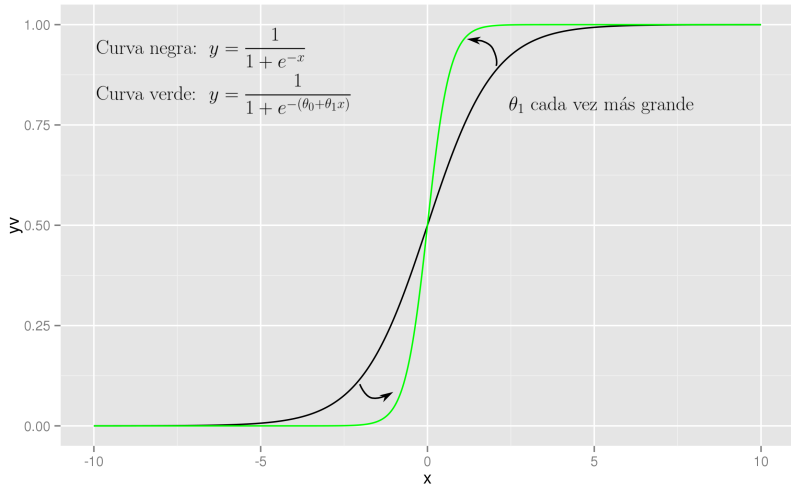
# En el caso de una característica, efecto de variar $\theta$



# En el caso de una característica, efecto de variar $\theta$

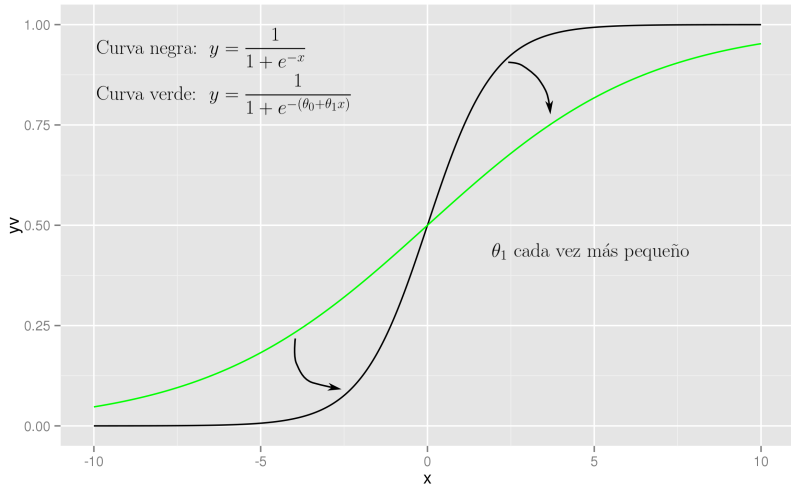


# En el caso de una característica, efecto de variar $\theta$





# En el caso de una característica, efecto de variar $\theta$



## Interpretación

El valor de  $h_{\theta}(x)$  es la probabilidad de que  $y$  tome el valor 1, para ese vector de características  $x$ , si los parámetros del ajuste son  $\theta$ .

Es

$$h_{\theta}(x) = \mathbb{P}(y = 1|x; \theta),$$

es decir la probabilidad de que  $y = 1$  condicionado a  $x$  y  $\theta$ .

## Interpretación

El valor de  $h_{\theta}(x)$  es la probabilidad de que  $y$  tome el valor 1, para ese vector de características  $x$ , si los parámetros del ajuste son  $\theta$ .  
Es

$$h_{\theta}(x) = \mathbb{P}(y = 1|x; \theta),$$

es decir la probabilidad de que  $y = 1$  condicionado a  $x$  y  $\theta$ .

Si, dado una nota media PAU, encontramos  $h_{\theta}(x) = 0.7$ , le diremos al alumno que tiene 70% de probabilidad de acabar egresando...

- Una vez entrenada nuestra regresión logística, tendremos el modelo ajustado  $h_{\hat{\theta}}(x)$ .

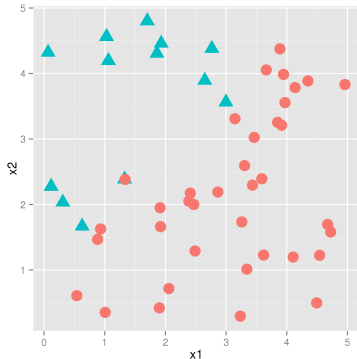
- Una vez entrenada nuestra regresión logística, tendremos el modelo ajustado  $h_{\hat{\theta}}(x)$ .
- Recordad que hemos decidido usar para clasificar la regla de decisión:
  - Si  $h_{\theta}(x) \geq 0.5$ , clasificamos  $\hat{y}$  como 1.
  - Si  $h_{\theta}(x) < 0.5$ , clasificamos  $\hat{y}$  como 0.

- Una vez entrenada nuestra regresión logística, tendremos el modelo ajustado  $h_{\hat{\theta}}(x)$ .
- Recordad que hemos decidido usar para clasificar la regla de decisión:
  - Si  $h_{\theta}(x) \geq 0.5$ , clasificamos  $\hat{y}$  como 1.
  - Si  $h_{\theta}(x) < 0.5$ , clasificamos  $\hat{y}$  como 0.
- Pero  $h_{\theta}(x) = g(x^T \theta)$ , por lo que
  - $h_{\theta}(x) \geq 0.5 \Leftrightarrow x^T \theta \geq 0$  y
  - $h_{\theta}(x) < 0.5 \Leftrightarrow x^T \theta < 0$ .

- Una vez entrenada nuestra regresión logística, tendremos el modelo ajustado  $h_{\hat{\theta}}(x)$ .
- Recordad que hemos decidido usar para clasificar la regla de decisión:
  - Si  $h_{\theta}(x) \geq 0.5$ , clasificamos  $\hat{y}$  como 1.
  - Si  $h_{\theta}(x) < 0.5$ , clasificamos  $\hat{y}$  como 0.
- Pero  $h_{\theta}(x) = g(x^T \theta)$ , por lo que
  - $h_{\theta}(x) \geq 0.5 \Leftrightarrow x^T \theta \geq 0$  y
  - $h_{\theta}(x) < 0.5 \Leftrightarrow x^T \theta < 0$ .
- Así que, en realidad, hemos especificado así una región de decisión cuya frontera es

$$x^T \theta = 0.$$

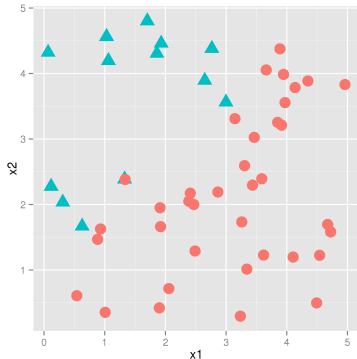
# Región de decisión ilustrada con dos características



Dos características,  $x_1$  y  $x_2$   
azul  $\leftrightarrow y = 1$ ; rojo  $\leftrightarrow y = 0$



# Región de decisión ilustrada con dos características



Dos características,  $x_1$  y  $x_2$   
azul  $\leftrightarrow y = 1$ ; rojo  $\leftrightarrow y = 0$   
Tenemos:

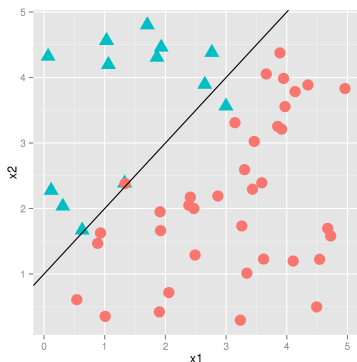
$$\theta = (\theta_0, \theta_1, \theta_2),$$

$$x = (1, x_1, x_2),$$

por lo que la frontera de la región  
es

$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 = 0.$$

# Región de decisión ilustrada con dos características



Dos características,  $x_1$  y  $x_2$   
azul  $\leftrightarrow y = 1$ ; rojo  $\leftrightarrow y = 0$   
Tenemos:

$$\theta = (\theta_0, \theta_1, \theta_2),$$

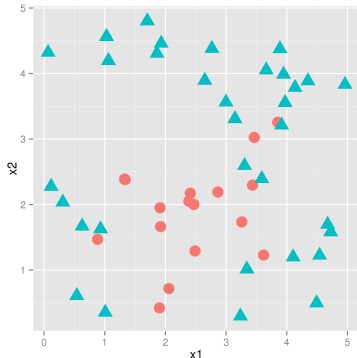
$$x = (1, x_1, x_2),$$

por lo que la frontera de la región  
es

$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 = 0.$$

Es una recta.

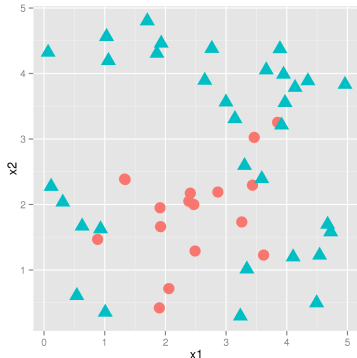
# Región de decisión ilustrada con dos características



Si introducimos potencias de grado superior de  $x_1$  y  $x_2$ , podemos obtener fronteras no lineales...

- $\theta = (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5),$
- $x = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2),$

# Región de decisión ilustrada con dos características



Si introducimos potencias de grado superior de  $x_1$  y  $x_2$ , podemos obtener fronteras no lineales...

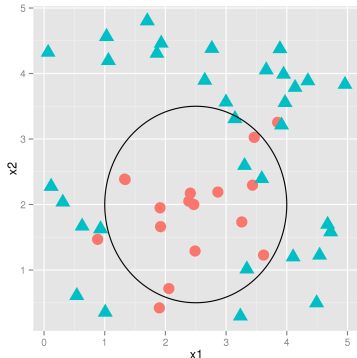
$$\theta = (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5),$$

$$x = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2),$$

por lo que la frontera de la región es

$$\theta_0 + \theta_1x_1 + \theta_2x_2 + \theta_3x_1^2 + \dots = 0.$$

# Región de decisión ilustrada con dos características



Si introducimos potencias de grado superior de  $x_1$  y  $x_2$ , podemos obtener fronteras no lineales...

$$\blacksquare \theta = (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5),$$

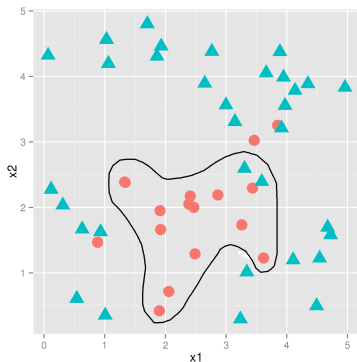
$$\blacksquare x = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2),$$

por lo que la frontera de la región es

$$\theta_0 + \theta_1x_1 + \theta_2x_2 + \theta_3x_1^2 + \dots = 0.$$

Por ejemplo un círculo...

# Región de decisión ilustrada con dos características



Y si introducimos potencias de grado aun superior  $x_1^3$  y  $x_2^3 \dots, x_1^5$  etc.. podemos obtener fronteras más complejas...

Conjunto de entrenamiento: los datos que tendremos se presentarán en la forma siguiente:

$Y$	$X_1$	$X_2$	$\dots$	$X_k$
$y_1$	$x_{11}$	$x_{12}$	$\dots$	$x_{1k}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$y_n$	$x_{n1}$	$x_{n2}$	$\dots$	$x_{nk}$

Cada fila representa un individuo, cada columna una variable o característica para ese individuo.

Los valores  $y_1, y_2$ , etc... son valores binarios (0 ó 1).

Conjunto de entrenamiento: los datos que tendremos se presentarán en la forma siguiente:

$Y$	$X_1$	$X_2$	$\dots$	$X_k$
$y_1$	$x_{11}$	$x_{12}$	$\dots$	$x_{1k}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$y_n$	$x_{n1}$	$x_{n2}$	$\dots$	$x_{nk}$

Cada fila representa un individuo, cada columna una variable o característica para ese individuo.

Los valores  $y_1, y_2$ , etc... son valores binarios (0 ó 1).

Usaremos la notación

$$x_{i\bullet} = (x_{i0}, x_{i1}, \dots, x_{ik})^T$$

para denotar el vector de características del individuo número  $i$   
(hemos incluido  $x_{i0} = 1$ .)



- Buscamos entrenar un algoritmo de regresión logística, es decir encontrar el “mejor” vector de parámetros  $\theta$ , aprendiendo de nuestro conjunto de entrenamiento.

# La función de coste

- Buscamos entrenar un algoritmo de regresión logística, es decir encontrar el “mejor” vector de parámetros  $\theta$ , aprendiendo de nuestro conjunto de entrenamiento.
- Necesitamos una función de coste que mida la calidad del ajuste al conjunto de entrenamiento, pero que posea también buenas propiedades para la minimización (convexidad).

# La función de coste

- Buscamos entrenar un algoritmo de regresión logística, es decir encontrar el “mejor” vector de parámetros  $\theta$ , aprendiendo de nuestro conjunto de entrenamiento.
- Necesitamos una función de coste que mida la calidad del ajuste al conjunto de entrenamiento, pero que posea también buenas propiedades para la minimización (convexidad).
- Por ello, introducimos la función de coste siguiente

$$J(\theta) = \frac{1}{n} \sum_{i=1}^n \text{coste}(h_{\theta}(x_{i\bullet}), y_i)$$

# La función de coste

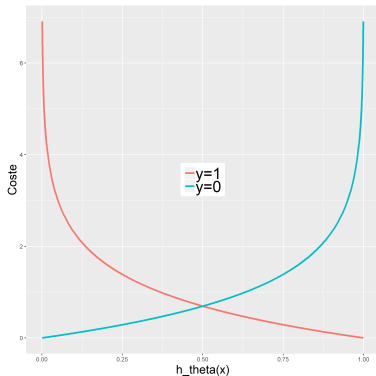
- Buscamos entrenar un algoritmo de regresión logística, es decir encontrar el “mejor” vector de parámetros  $\theta$ , aprendiendo de nuestro conjunto de entrenamiento.
- Necesitamos una función de coste que mida la calidad del ajuste al conjunto de entrenamiento, pero que posea también buenas propiedades para la minimización (convexidad).
- Por ello, introducimos la función de coste siguiente

$$J(\theta) = \frac{1}{n} \sum_{i=1}^n \text{coste}(h_{\theta}(x_i), y_i)$$

donde

$$\text{coste}(h_{\theta}(x), y) = \begin{cases} -\log(h_{\theta}(x)) & \text{si } y = 1, \\ -\log(1 - h_{\theta}(x)) & \text{si } y = 0, \end{cases}$$

# La función de coste



Este es el perfil de la función de coste.  
Por lo tanto:

- Si  $y = 1$ , cuando  $h_{\theta}(x) \rightarrow 0$ ,  $\text{coste}(h_{\theta}(x), y) \rightarrow \infty$ .
- Si  $y = 1$ , cuando  $h_{\theta}(x) = 1$ ,  $\text{coste}(h_{\theta}(x), y) = 0$ .
- Si  $y = 0$ , cuando  $h_{\theta}(x) \rightarrow 1$ ,  $\text{coste}(h_{\theta}(x), y) \rightarrow \infty$ .
- Si  $y = 0$ , cuando  $h_{\theta}(x) = 0$ ,  $\text{coste}(h_{\theta}(x), y) = 0$ .

Nuestra función de coste será por lo tanto:

$$J(\theta) = \frac{1}{n} \sum_{i=1}^n \text{coste}(h_{\theta}(x_{i\bullet}), y_i)$$

donde

$$\text{coste}(h_{\theta}(x), y) = \begin{cases} -\log(h_{\theta}(x)) & \text{si } y = 1, \\ -\log(1 - h_{\theta}(x)) & \text{si } y = 0, \end{cases},$$

lo que podemos escribir de manera más rápida como

$$J(\theta) = -\frac{1}{n} \sum_{i=1}^n \{y_i \log(h_{\theta}(x_{i\bullet})) + (1 - y_i) \log(1 - h_{\theta}(x_{i\bullet}))\}$$

Deducimos

$$\begin{aligned}\nabla_{\theta} J(\theta) &= -\frac{1}{n} \sum_{i=1}^n \{y_i \nabla_{\theta} \log(h_{\theta}(x_{i\bullet})) + (1 - y_i) \nabla_{\theta} \log(1 - h_{\theta}(x_{i\bullet}))\} \\ &= \frac{1}{n} \sum_{i=1}^n \{x_{i\bullet} \cdot (h_{\theta}(x_{i\bullet}) - y_i)\}.\end{aligned}$$

Deducimos

$$\begin{aligned}\nabla_{\theta} J(\theta) &= -\frac{1}{n} \sum_{i=1}^n \{y_i \nabla_{\theta} \log(h_{\theta}(x_{i\bullet})) + (1 - y_i) \nabla_{\theta} \log(1 - h_{\theta}(x_{i\bullet}))\} \\ &= \frac{1}{n} \sum_{i=1}^n \{x_{i\bullet} \cdot (h_{\theta}(x_{i\bullet}) - y_i)\}.\end{aligned}$$

Si usamos la matriz de diseño  $\mathbf{X}$ , obtenemos en forma compacta:

$$\nabla_{\theta} J(\theta) = \frac{1}{n} \mathbf{X}^T \cdot (\mathbf{H}_{\theta} - \mathbf{y}),$$

donde  $\mathbf{H}$  denota el vector columna:

$$\mathbf{H}_{\theta} = \begin{pmatrix} h_{\theta}(x_{1\bullet}) \\ h_{\theta}(x_{2\bullet}) \\ \vdots \\ h_{\theta}(x_{n\bullet}) \end{pmatrix}$$



# Nota: comparación con la implementación para regresión múltiple

Recordad que, para la regresión múltiple, el gradiente era:

$$\nabla_{\theta} J(\theta) = \frac{1}{n} \mathbf{X}^T \cdot (\mathbf{X}\theta - y).$$

mientras que para la regresión logística, acabamos de establecer que el gradiente es:

$$\nabla_{\theta} J(\theta) = \frac{1}{n} \mathbf{X}^T \cdot (\mathbf{H}_{\theta} - y),$$

Es muy similar, si queremos programar el algoritmo del gradiente, sólo requiere una pequeña modificación de nuestro código...

Podemos aprovechar la clase `LogisticRegression` del submódulo `linear_model` si no queremos usar algoritmo del gradiente.

Podemos aprovechar la clase `LogisticRegression` del submódulo `linear_model` si no queremos usar algoritmo del gradiente.

- Las ventajas:
  - No es necesario especificar  $\alpha$
  - Puede ser más rápido.

Podemos aprovechar la clase `LogisticRegression` del submódulo `linear_model` si no queremos usar algoritmo del gradiente.

- Las ventajas:
  - No es necesario especificar  $\alpha$
  - Puede ser más rápido.
- Sin embargo, en el caso en que haya muchas características, puede ser más eficiente usar algoritmo del gradiente (`scikit-learn` también dispone de procedimientos para ello.)

# Si queremos clasificar en más de dos categorías...

Si queremos clasificar cada individuo en más de dos categorías, usaremos la técnica del “One versus all”.

# Si queremos clasificar en más de dos categorías...

Si queremos clasificar cada individuo en más de dos categorías, usaremos la técnica del “One versus all”.

Ejemplo con tres categorías A, B y C.

Supongamos que queremos clasificar cada individuo en A, B o C.

# Si queremos clasificar en más de dos categorías...

Si queremos clasificar cada individuo en más de dos categorías, usaremos la técnica del “One versus all”.

Ejemplo con tres categorías A, B y C.

Supongamos que queremos clasificar cada individuo en A, B o C.

- Entrenamos una regresión logística para clasificar en “A” o “no A”.  $\Rightarrow$  obtenemos modelo ajustado  $h_{\hat{\theta}_A}(x)$ .

# Si queremos clasificar en más de dos categorías...

Si queremos clasificar cada individuo en más de dos categorías, usaremos la técnica del “One versus all”.

Ejemplo con tres categorías A, B y C.

Supongamos que queremos clasificar cada individuo en A, B o C.

- Entrenamos una regresión logística para clasificar en “A” o “no A”.  $\Rightarrow$  obtenemos modelo ajustado  $h_{\hat{\theta}_A}(x)$ .
- Entrenamos una regresión logística para clasificar en “B” o “no B”.  $\Rightarrow$  obtenemos modelo ajustado  $h_{\hat{\theta}_B}(x)$ .



# Si queremos clasificar en más de dos categorías...

Si queremos clasificar cada individuo en más de dos categorías, usaremos la técnica del “One versus all”.

Ejemplo con tres categorías A, B y C.

Supongamos que queremos clasificar cada individuo en A, B o C.

- Entrenamos una regresión logística para clasificar en “A” o “no A”.  $\Rightarrow$  obtenemos modelo ajustado  $h_{\hat{\theta}_A}(x)$ .
- Entrenamos una regresión logística para clasificar en “B” o “no B”.  $\Rightarrow$  obtenemos modelo ajustado  $h_{\hat{\theta}_B}(x)$ .
- Entrenamos una regresión logística para clasificar en “C” o “no C”.  $\Rightarrow$  obtenemos modelo ajustado  $h_{\hat{\theta}_C}(x)$ .

# Si queremos clasificar en más de dos categorías...

Si queremos clasificar cada individuo en más de dos categorías, usaremos la técnica del “One versus all”.

Ejemplo con tres categorías A, B y C.

Supongamos que queremos clasificar cada individuo en A, B o C.

- Entrenamos una regresión logística para clasificar en “A” o “no A”.  $\Rightarrow$  obtenemos modelo ajustado  $h_{\hat{\theta}_A}(x)$ .
- Entrenamos una regresión logística para clasificar en “B” o “no B”.  $\Rightarrow$  obtenemos modelo ajustado  $h_{\hat{\theta}_B}(x)$ .
- Entrenamos una regresión logística para clasificar en “C” o “no C”.  $\Rightarrow$  obtenemos modelo ajustado  $h_{\hat{\theta}_C}(x)$ .
- Dado un nuevo individuo, calculamos las tres probabilidades predichas  $h_{\hat{\theta}_A}(x)$ ,  $h_{\hat{\theta}_B}(x)$  y  $h_{\hat{\theta}_C}(x)$ .

# Si queremos clasificar en más de dos categorías...

Si queremos clasificar cada individuo en más de dos categorías, usaremos la técnica del “One versus all”.

## Ejemplo con tres categorías A, B y C.

Supongamos que queremos clasificar cada individuo en A, B o C.

- Entrenamos una regresión logística para clasificar en “A” o “no A”.  $\Rightarrow$  obtenemos modelo ajustado  $h_{\hat{\theta}_A}(x)$ .
- Entrenamos una regresión logística para clasificar en “B” o “no B”.  $\Rightarrow$  obtenemos modelo ajustado  $h_{\hat{\theta}_B}(x)$ .
- Entrenamos una regresión logística para clasificar en “C” o “no C”.  $\Rightarrow$  obtenemos modelo ajustado  $h_{\hat{\theta}_C}(x)$ .
- Dado un nuevo individuo, calculamos las tres probabilidades predichas  $h_{\hat{\theta}_A}(x)$ ,  $h_{\hat{\theta}_B}(x)$  y  $h_{\hat{\theta}_C}(x)$ .
- Clasificamos el individuo en la categoría que tiene la probabilidad predicha más alta...

# Medir la calidad de la predicción para una clasificación binaria

El primer indicador que podemos usar es la **tasa de acierto**, es decir el porcentaje de individuos clasificados correctamente.

# Medir la calidad de la predicción para una clasificación binaria

El primer indicador que podemos usar es la **tasa de acierto**, es decir el porcentaje de individuos clasificados correctamente.

## Ejemplo

Consideremos el problema de predecir si un tumor es benigno o maligno basándonos en unas imágenes médicas.

# Medir la calidad de la predicción para una clasificación binaria

El primer indicador que podemos usar es la **tasa de acierto**, es decir el porcentaje de individuos clasificados correctamente.

## Ejemplo

Consideremos el problema de predecir si un tumor es benigno o maligno basándonos en unas imágenes médicas.

De un total de 100 tumores, de los cuáles 5 son malignos y 95 benignos, mi algoritmo se ha equivocado en 1 maligno y 5 benignos.

# Medir la calidad de la predicción para una clasificación binaria

El primer indicador que podemos usar es la **tasa de acierto**, es decir el porcentaje de individuos clasificados correctamente.

## Ejemplo

Consideremos el problema de predecir si un tumor es benigno o maligno basándonos en unas imágenes médicas.

De un total de 100 tumores, de los cuáles 5 son malignos y 95 benignos, mi algoritmo se ha equivocado en 1 maligno y 5 benignos.

$$\text{Tasa de acierto} = \frac{94}{100} = 94\%.$$

# Medir la calidad de la predicción para una clasificación binaria

El primer indicador que podemos usar es la **tasa de acierto**, es decir el porcentaje de individuos clasificados correctamente.

## Ejemplo

Consideremos el problema de predecir si un tumor es benigno o maligno basándonos en unas imágenes médicas.

De un total de 100 tumores, de los cuáles 5 son malignos y 95 benignos, mi algoritmo se ha equivocado en 1 maligno y 5 benignos.

$$\text{Tasa de acierto} = \frac{94}{100} = 94\%.$$

Sin embargo, tiene sus limitaciones: si mi decisión hubiera sido sencillamente declarar todos como benignos, cuál habría sido mi tasa de acierto?



# Medir la calidad de la predicción para una clasificación binaria

El primer indicador que podemos usar es la **tasa de acierto**, es decir el porcentaje de individuos clasificados correctamente.

## Ejemplo

Consideremos el problema de predecir si un tumor es benigno o maligno basándonos en unas imágenes médicas.

De un total de 100 tumores, de los cuáles 5 son malignos y 95 benignos, mi algoritmo se ha equivocado en 1 maligno y 5 benignos.

$$\text{Tasa de acierto} = \frac{94}{100} = 94\%.$$

Sin embargo, tiene sus limitaciones: si mi decisión hubiera sido sencillamente declarar todos como benignos, cuál habría sido mi tasa de acierto?

$$\text{Tasa de acierto} = \frac{95}{100} = 95\%.$$

# Precisión y sensibilidad ( “recall” )

Por ello, introducimos dos indicadores que debemos considerar conjuntamente:

# Precisión y sensibilidad ( “recall” )

Por ello, introducimos dos indicadores que debemos considerar conjuntamente:

## Precisión

Es la proporción de aciertos ( $y = 1$ ) entre los que he clasificado como “positivos” ( $\hat{y} = 1$ ).

# Precisión y sensibilidad ( "recall" )

Por ello, introducimos dos indicadores que debemos considerar conjuntamente:

## Precisión

Es la proporción de aciertos ( $y = 1$ ) entre los que he clasificado como "positivos" ( $\hat{y} = 1$ ).

## Sensibilidad "Recall"

Es la proporción de aciertos ( $\hat{y} = 1$ ) entre todos los que son positivos "positivos" ( $y = 1$ ).

# Precisión y sensibilidad ( "recall" )

Por ello, introducimos dos indicadores que debemos considerar conjuntamente:

## Precisión

Es la proporción de aciertos ( $y = 1$ ) entre los que he clasificado como "positivos" ( $\hat{y} = 1$ ).

## Sensibilidad "Recall"

Es la proporción de aciertos ( $\hat{y} = 1$ ) entre todos los que son positivos "positivos" ( $y = 1$ ).

Para el problema anterior: De un total de 100 tumores, de los cuáles 5 son malignos y 95 benignos, mi algoritmo se ha equivocado en 1 maligno y 5 benignos.

# Precisión y sensibilidad ( “recall” )

Por ello, introducimos dos indicadores que debemos considerar conjuntamente:

## Precisión

Es la proporción de aciertos ( $y = 1$ ) entre los que he clasificado como “positivos” ( $\hat{y} = 1$ ).

## Sensibilidad “Recall”

Es la proporción de aciertos ( $\hat{y} = 1$ ) entre todos los que son positivos “positivos” ( $y = 1$ ).

Para el problema anterior: De un total de 100 tumores, de los cuáles 5 son malignos y 95 benignos, mi algoritmo se ha equivocado en 1 maligno y 5 benignos.

$$precision = 4/9, \quad recall = 4/5$$

# Precisión y sensibilidad ( "recall" )

Por ello, introducimos dos indicadores que debemos considerar conjuntamente:

## Precisión

Es la proporción de aciertos ( $y = 1$ ) entre los que he clasificado como "positivos" ( $\hat{y} = 1$ ).

## Sensibilidad "Recall"

Es la proporción de aciertos ( $\hat{y} = 1$ ) entre todos los que son positivos "positivos" ( $y = 1$ ).

Para el problema anterior: De un total de 100 tumores, de los cuáles 5 son malignos y 95 benignos, mi algoritmo se ha equivocado en 1 maligno y 5 benignos.

$$precision = 4/9, \quad recall = 4/5$$

Si los declaro todos como benignos:

$$precision = \text{no existe}, \quad recall = 0/5 = 0$$

# Matriz de confusión

Se suele presentar los resultados del algoritmo en forma de matriz, llamada matriz de confusión.



# Matriz de confusión

Se suele presentar los resultados del algoritmo en forma de matriz, llamada matriz de confusión.

Para el problema anterior: De un total de 100 tumores, de los cuáles 5 son malignos y 95 benignos, mi algoritmo se ha equivocado en 1 maligno y 5 benignos.

# Matriz de confusión

Se suele presentar los resultados del algoritmo en forma de matriz, llamada matriz de confusión.

Para el problema anterior: De un total de 100 tumores, de los cuáles 5 son malignos y 95 benignos, mi algoritmo se ha equivocado en 1 maligno y 5 benignos.

$y \backslash \hat{y}$	0	1
0	90	5
1	1	4

# Matriz de confusión

Se suele presentar los resultados del algoritmo en forma de matriz, llamada matriz de confusión.

Para el problema anterior: De un total de 100 tumores, de los cuáles 5 son malignos y 95 benignos, mi algoritmo se ha equivocado en 1 maligno y 5 benignos.

$y \backslash \hat{y}$	0	1
0	90	5
1	1	4

Matriz de confusión:

$$\begin{pmatrix} 90 & 5 \\ 1 & 4 \end{pmatrix}$$

# Precisión y sensibilidad

La precisión y la sensibilidad van en sentido contrario: si aumenta la precisión, baja la sensibilidad y al revés.

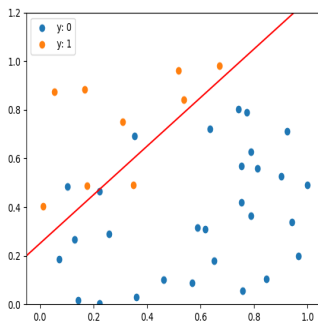
# Precisión y sensibilidad

La precisión y la sensibilidad van en sentido contrario: si aumenta la precisión, baja la sensibilidad y al revés.  
Se busca un equilibrio.

# Precisión y sensibilidad

La precisión y la sensibilidad van en sentido contrario: si aumenta la precisión, baja la sensibilidad y al revés.

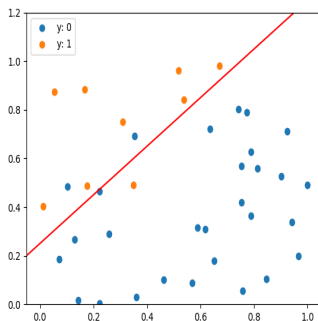
Se busca un equilibrio. Dos características y una frontera de decisión lineal:



# Precisión y sensibilidad

La precisión y la sensibilidad van en sentido contrario: si aumenta la precisión, baja la sensibilidad y al revés.

Se busca un equilibrio. Dos características y una frontera de decisión lineal:

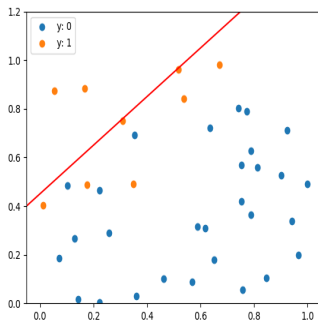


Tenemos una precisión de 80% y una sensibilidad de 8/9, (89%).

# Precisión y sensibilidad

La precisión y la sensibilidad van en sentido contrario: si aumenta la precisión, baja la sensibilidad y al revés.

Se busca un equilibrio. Dos características y una frontera de decisión lineal. Si aumento la ordenada al origen de la frontera:

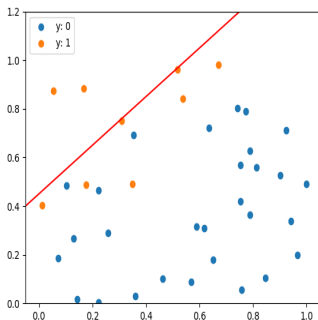




# Precisión y sensibilidad

La precisión y la sensibilidad van en sentido contrario: si aumenta la precisión, baja la sensibilidad y al revés.

Se busca un equilibrio. Dos características y una frontera de decisión lineal. Si aumento la ordenada al origen de la frontera:

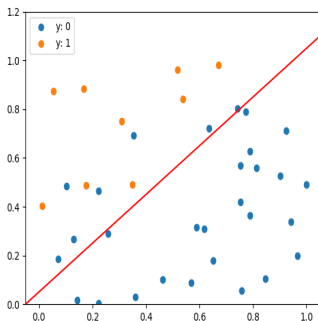


Tenemos una precisión de 100% y una sensibilidad de  $2/9$  (22%)

# Precisión y sensibilidad

La precisión y la sensibilidad van en sentido contrario: si aumenta la precisión, baja la sensibilidad y al revés.

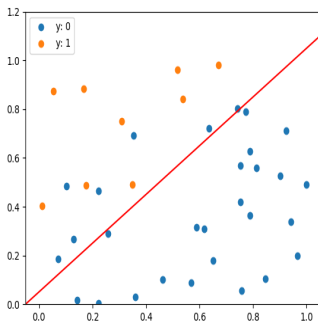
Se busca un equilibrio. Dos características y una frontera de decisión lineal. Si disminuyo la ordenada al origen de la frontera:



# Precisión y sensibilidad

La precisión y la sensibilidad van en sentido contrario: si aumenta la precisión, baja la sensibilidad y al revés.

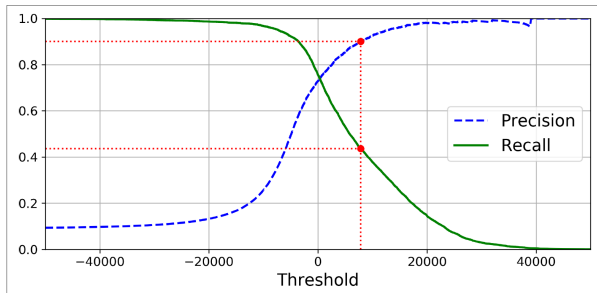
Se busca un equilibrio. Dos características y una frontera de decisión lineal. Si disminuyo la ordenada al origen de la frontera:



Tenemos una precisión de 64% y una sensibilidad de 100%

Una típica situación:

Una típica situación:



Fuente: <https://jaehyeong.github.io/2020/02/29/LSTM-Autoencoder-for-Anomaly-Detection/>