# Regresión logística

- Hipótesis
  - Función signo
  - Función logística
- Función de coste
  - 0-1
  - Logística
- Evaluación
  - Matriz de confusión y métricas derivadas
  - Análisis ROC
  - Curva de aprendizaje

# Clasificación

## Ejemplo en el sector eléctrico

- La empresa está impulsando la transición al vehículo eléctrico. Para ello tiene una política de marketing que se dirige captar nuevos usuarios entre sus clientes.
  - Dado el coste de dicha política, la acción comercial no se dirige a todos los clientes de manera universal, sino sólo a aquéllos que son potenciales compradores.
  - ¿El nuevo cliente es uno de ellos?
- Para dilucidarlo se plantea distintas cuestiones previas

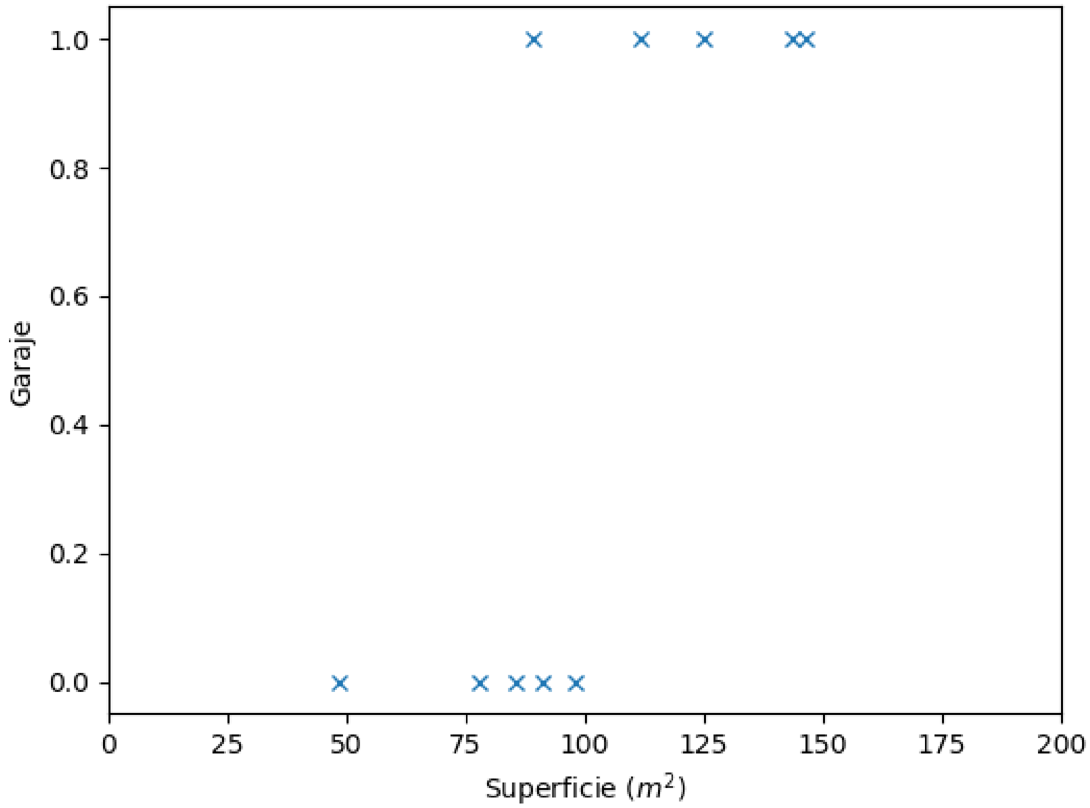
# Clasificación

## Ejemplo en el sector eléctrico

- La primera cuestión es saber si el cliente tiene o no garaje propio. Ese dato no consta en sus bases de datos, por lo que realiza un muestreo entre sus clientes.
- De este muestreo obtiene dos datos
    1. Superficie de la vivienda ($m^2$)
    2. Tiene garaje (Si/No)
- Con esa información quiere inferir sin un determinado cliente tiene garaje en función del tamaño de su vivienda (dato que sí consta en sus bases de datos)

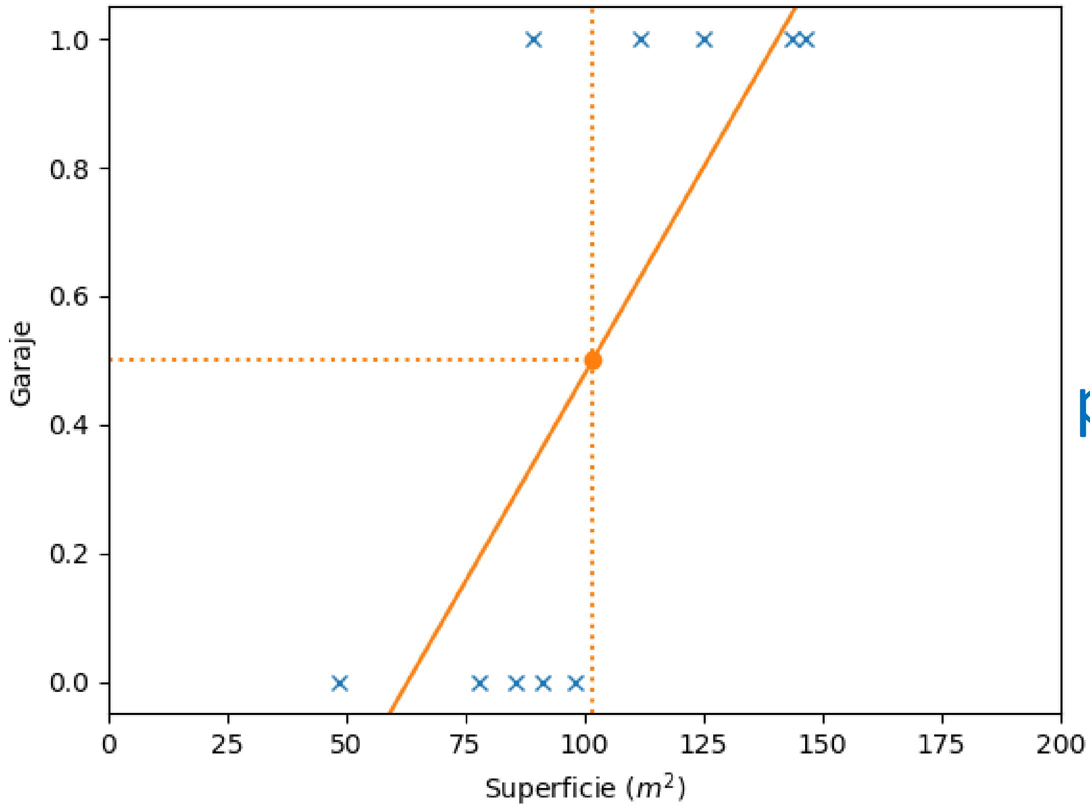
# Regresión logística

## 3. Formulación de hipótesis



# Regresión logística

## 3. Formulación de hipótesis

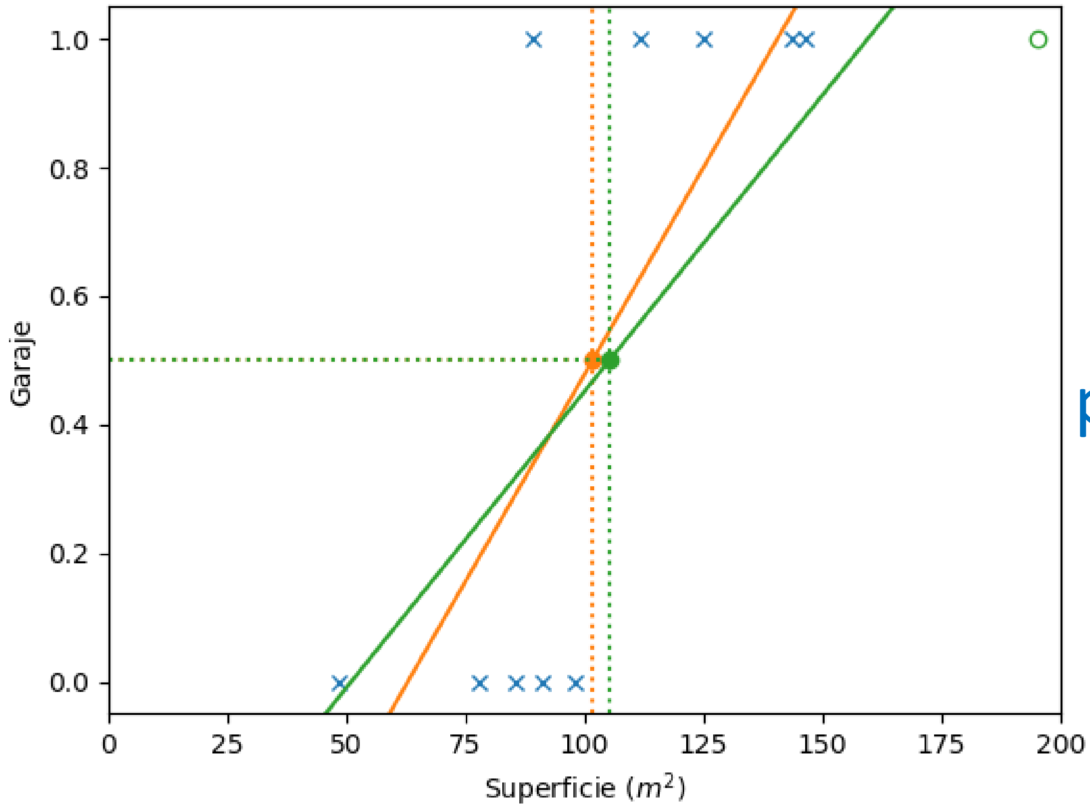



$w_0, w_1$ calculados por regresión lineal

$$h_w(x) = w_0 + w_1 x$$

# Regresión logística

## 3. Formulación de hipótesis

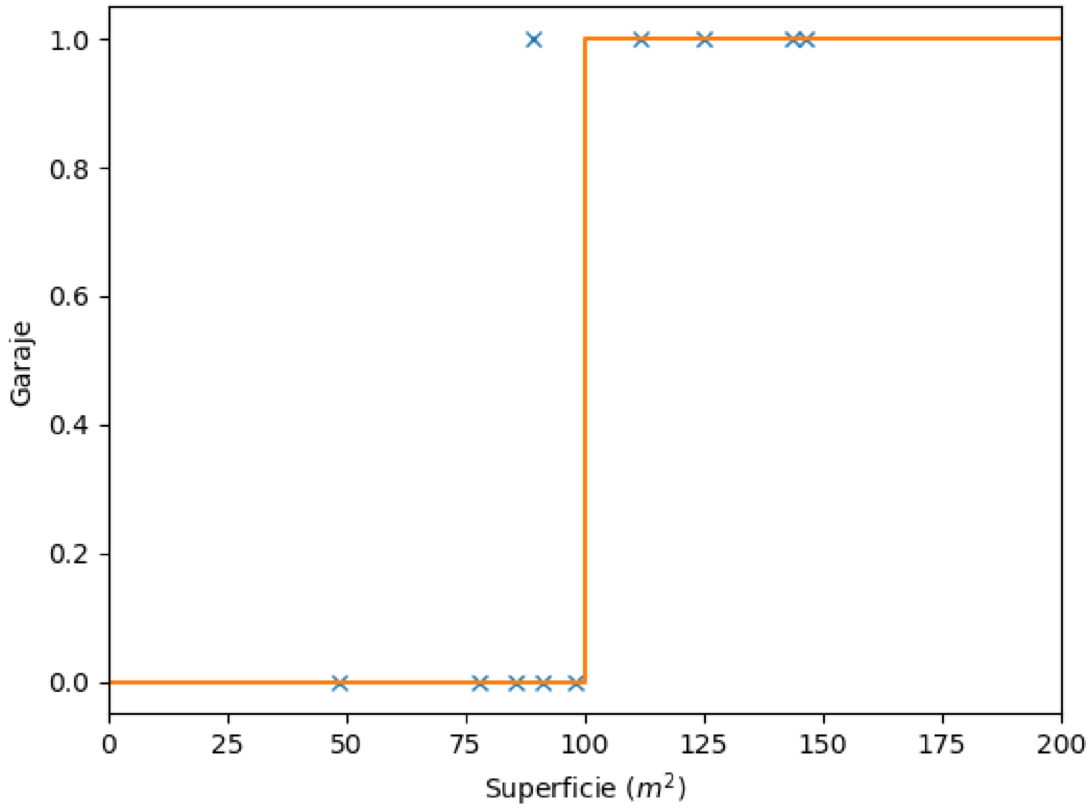



$w_0, w_1$ calculados por regresión lineal

$$h_w(x) = w_0 + w_1 x$$

# Regresión logística

## 3. Formulación de hipótesis




$w_0 = -100$

$$h_w(x) = \frac{1 + \mathrm{sign}(x + w_0)}{2}$$

# Regresión logística

## 5. Optimización del coste

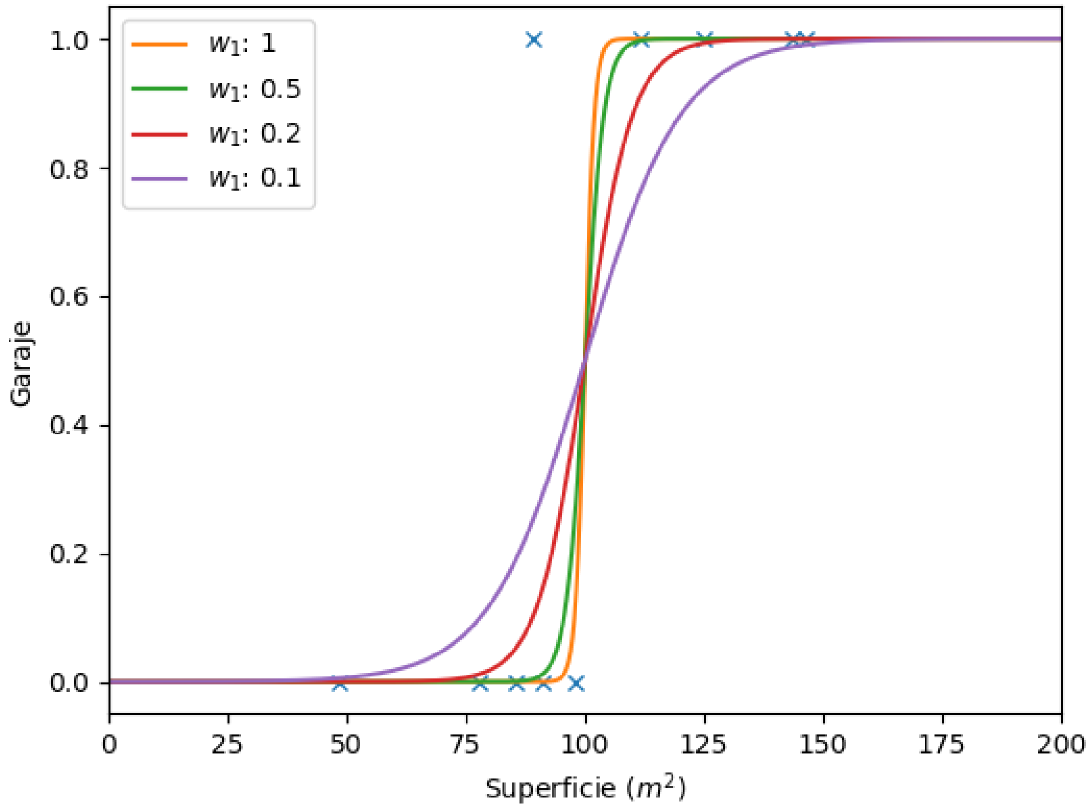
$$h_w(x) = \frac{1 + \mathrm{sign}(x + w_0)}{2}$$

$$\frac{d}{dw} J(h_w(x), y) = 0$$

La función $\mathrm{sign}(\mathrm{x})$ no es derivable (en el origen)

# Regresión logística

## 3. Formulación de hipótesis




$$\frac{w_0}{w_1} = -100$$

Función logística

$$h_w(x) = \frac{1}{1 + e^{-z}}$$

$$z \equiv w_0 + w_1 x$$

# Regresión logística

## 3. Formulación de hipótesis

Cuota (odds) $$o \equiv \frac{p}{1-p}$$ $p$: probabilidad de que tenga garaje

Log-odds $$Ln(o) = z = w_0 + w_1 x$$ Unida de medida: logit (logistic unit)

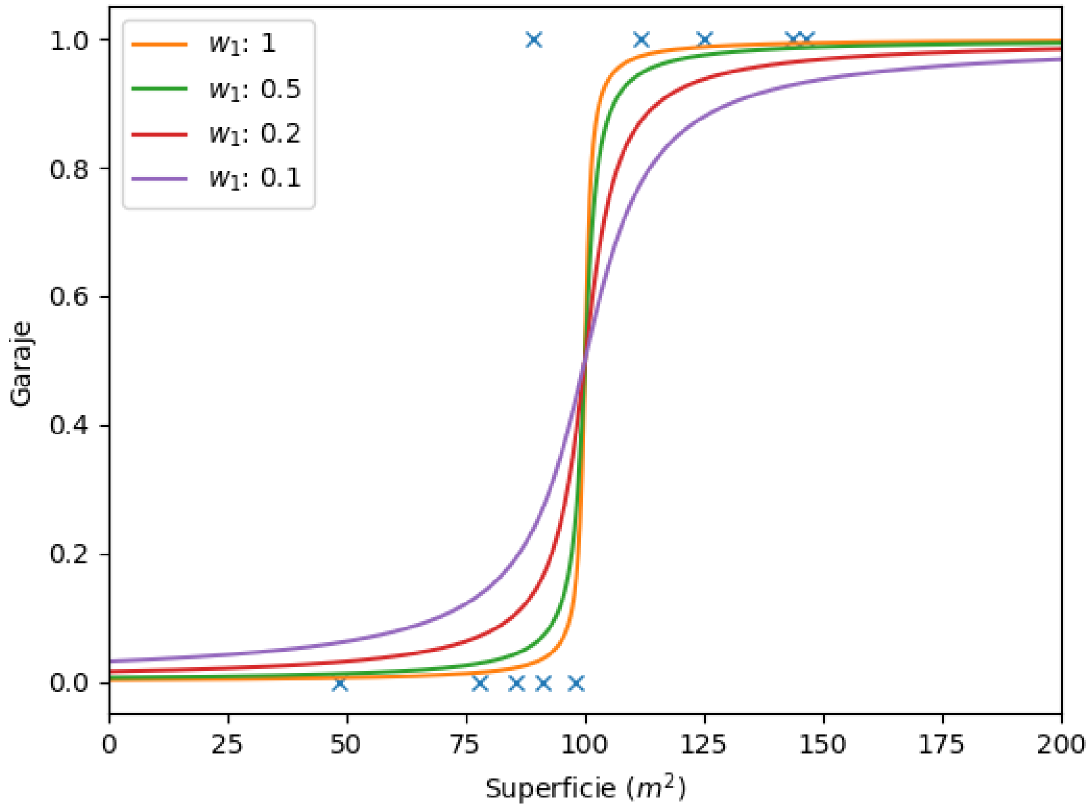
Puede usarse otra función no lineal $z = f(x)$ (p.e. polinómica)

$$o = e^z = e^{(w_0 + w_1 x)}$$

$$p = \frac{o}{1+o} = \frac{1}{1+o^{-1}} = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-(w_0+w_1 x)}}$$

# Regresión logística

## 3. Formulación de hipótesis




$$\frac{w_0}{w_1} = -100$$

Otras sigmoides

$$h_w(x) = \frac{1}{2} + \frac{\mathrm{atan}(z)}{\pi}$$

$$z \equiv w_0 + w_1 x$$

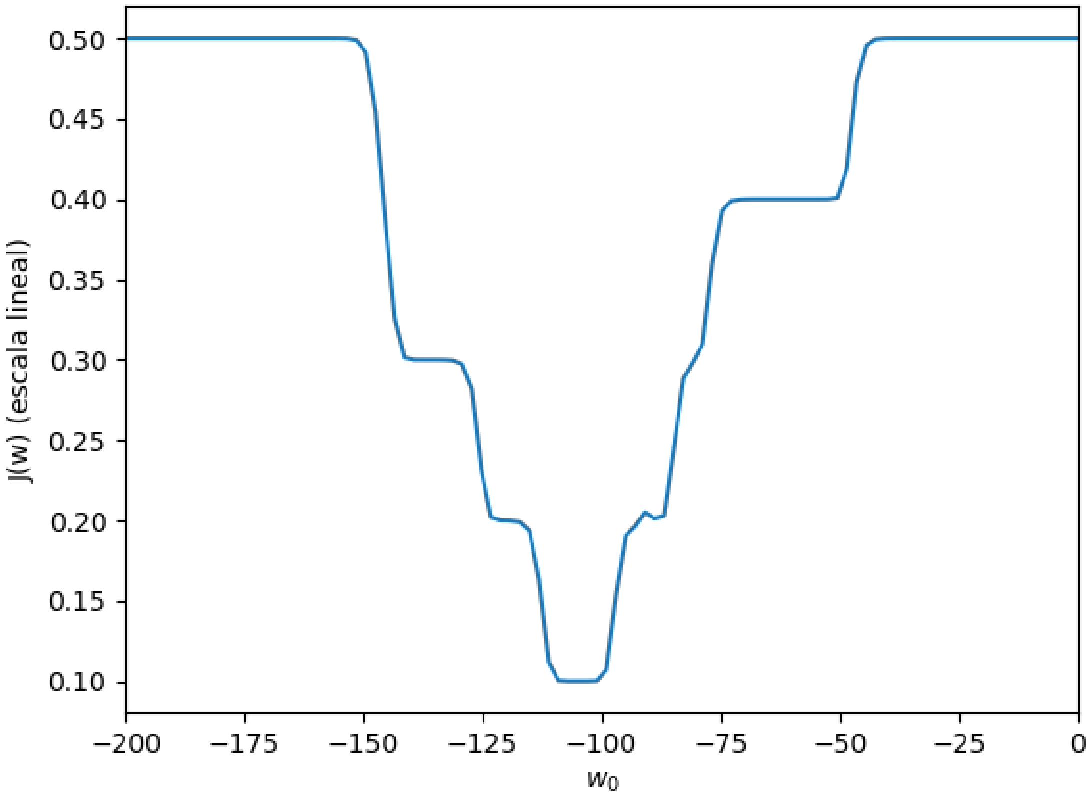
# Regresión logística

- Hipótesis
  - Función signo
  - Función logística
- Función de coste
  - 0-1
  - Logística
- Evaluación
  - Matriz de confusión y métricas derivadas
  - Análisis ROC
  - Curva de aprendizaje

# Regresión logística

## 4. Elección de la función de coste

Hipótesis:
Función
logística




$w_1 = 1$

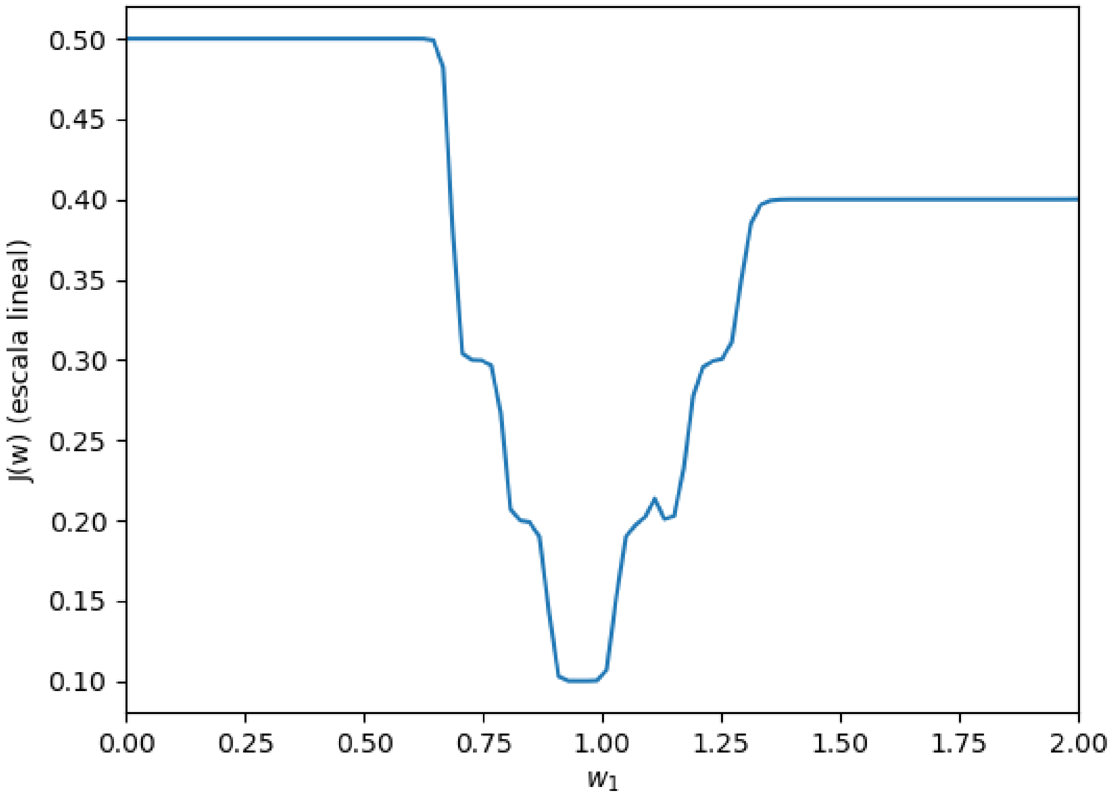
Coste
cuadrático
No convexo

$$J(h_w(x), y) = \frac{1}{n}\sum_{i=1}^{n}\left(h_w\left(x^{(i)}\right) - y^{(i)}\right)^2$$

# Regresión logística

## 4. Elección de la función de coste

Hipótesis:
Función
logística




$w_0 = -100$

Coste
cuadrático
No convexo

$$J(h_w(x), y) = \frac{1}{n}\sum_{i=1}^{n}\left(h_w\left(x^{(i)}\right) - y^{(i)}\right)^2$$

# Regresión logística

## 4. Elección de la función de coste

### Función de coste 0-1

$$\hat{y} = \begin{cases} 1, h_w(x) \geq t_h \\ 0, h_w(x) < t_h \end{cases}$$

$$t_h = 0.5$$

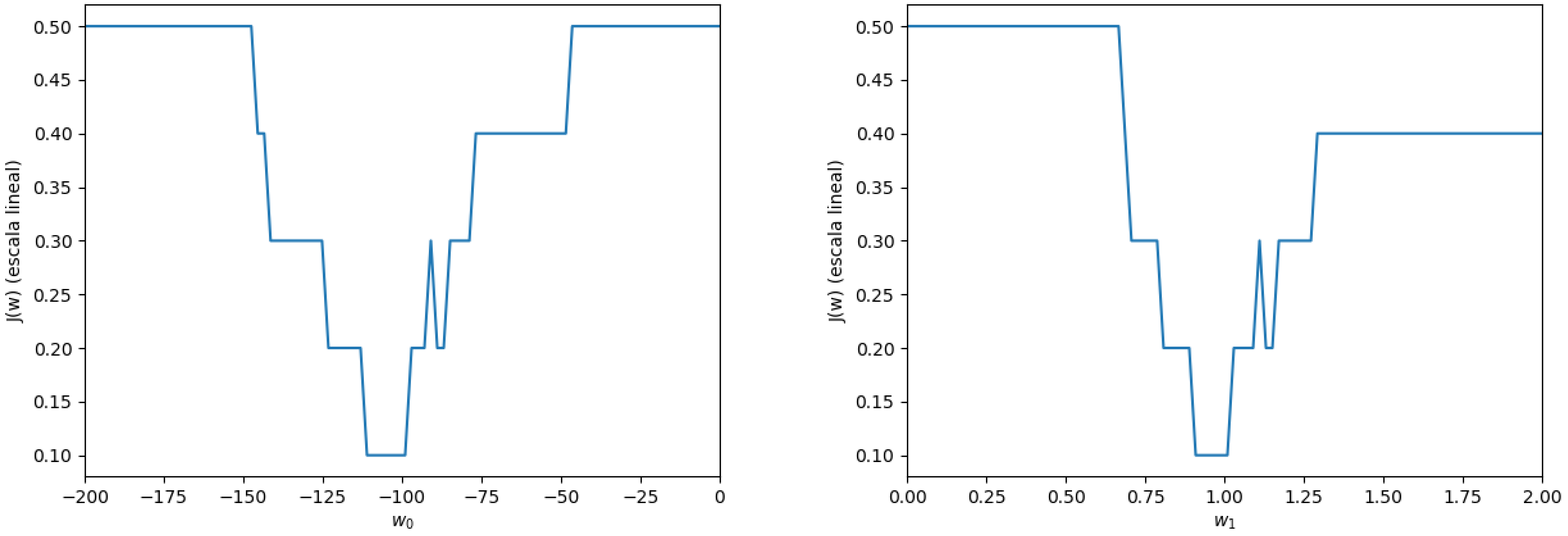
$$I(u) \equiv \begin{cases} 1, u = True \\ 0, u = False \end{cases}$$

$$J(h_w(x), y) = \frac{1}{n} \sum_{i=1}^{n} I\left(\hat{y}^{(i)} \neq y^{(i)}\right) = ACC$$

$$ACC \equiv \frac{\#aciertos}{n}$$

# Regresión logística

## 4. Elección de la función de coste




Coste 0-1

No convexo

# Regresión logística

## 4. Elección de la función de coste

### Función de coste logística (cros-entropía)

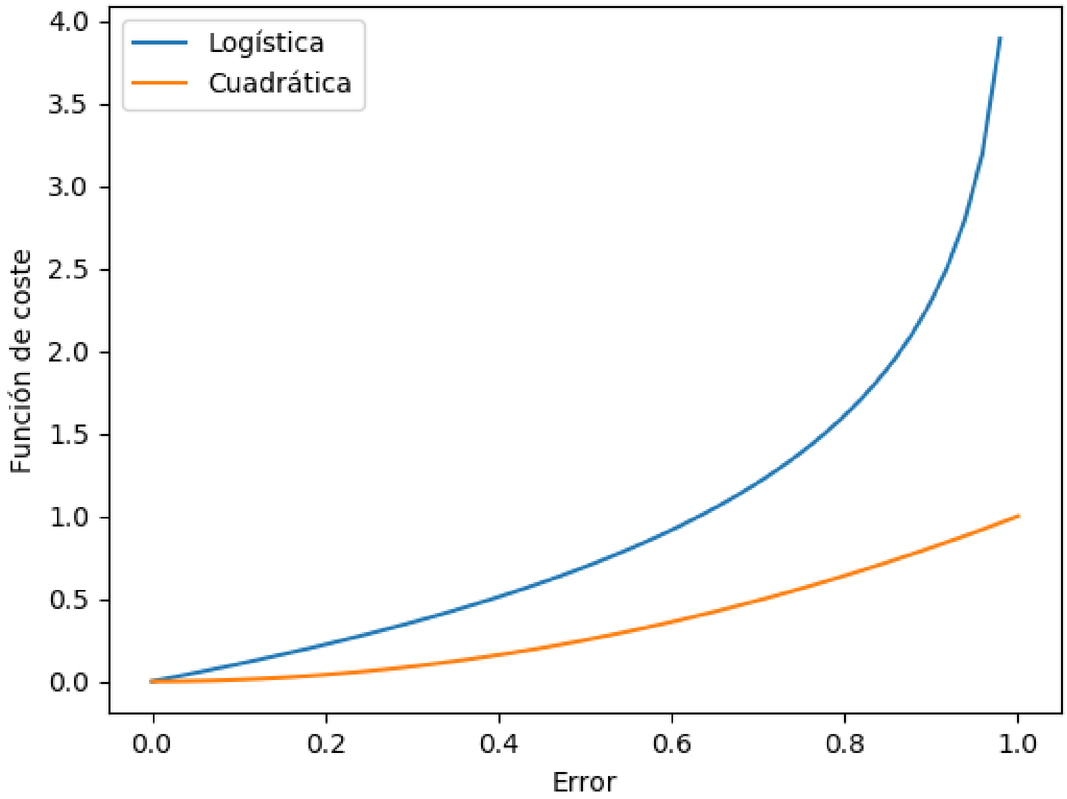
$$e(h_w(x), y) = |h_w(x) - y|$$

$$L(h_w(x), y) = -Ln(1 - e)$$

$$J(h_w(x), y) = \frac{1}{n} \sum_{i=1}^{n} L\left(h_w(x^{(i)}), y^{(i)}\right)$$

# Regresión logística

## 4. Elección de la función de coste




$$L(h_w(x), y) = -Ln(1 - e)$$

# Regresión logística

## 4. Elección de la función de coste

### Función de coste logística (cros-entropía)

$$e = |h_w(x) - y| = \begin{cases} 1 - h_w(x), \forall y = 1 \\ h_w(x), \forall y = 0 \end{cases}$$

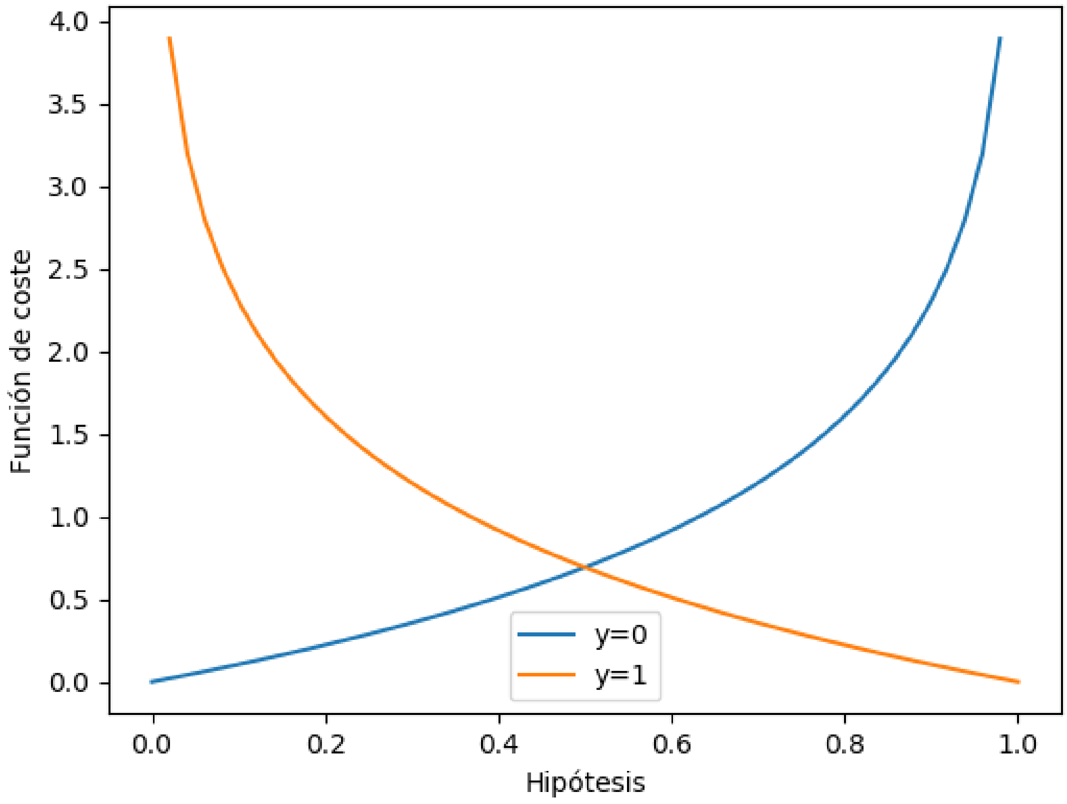
$$L(h_w(x), y) = -Ln(1 - e) = \begin{cases} -Ln\big(h_w(x)\big), \forall y = 1 \\ -Ln\big(1 - h_w(x)\big), \forall y = 0 \end{cases}$$

$$L(h_w(x), y) = -yLn\big(h_w(x)\big) - (1 - y)Ln\big(1 - h_w(x)\big)$$

$$J(h_w(x), y) = \frac{1}{n}\sum_{i=1}^{n} L\big(h_w(x^{(i)}), y^{(i)}\big)$$

# Regresión logística

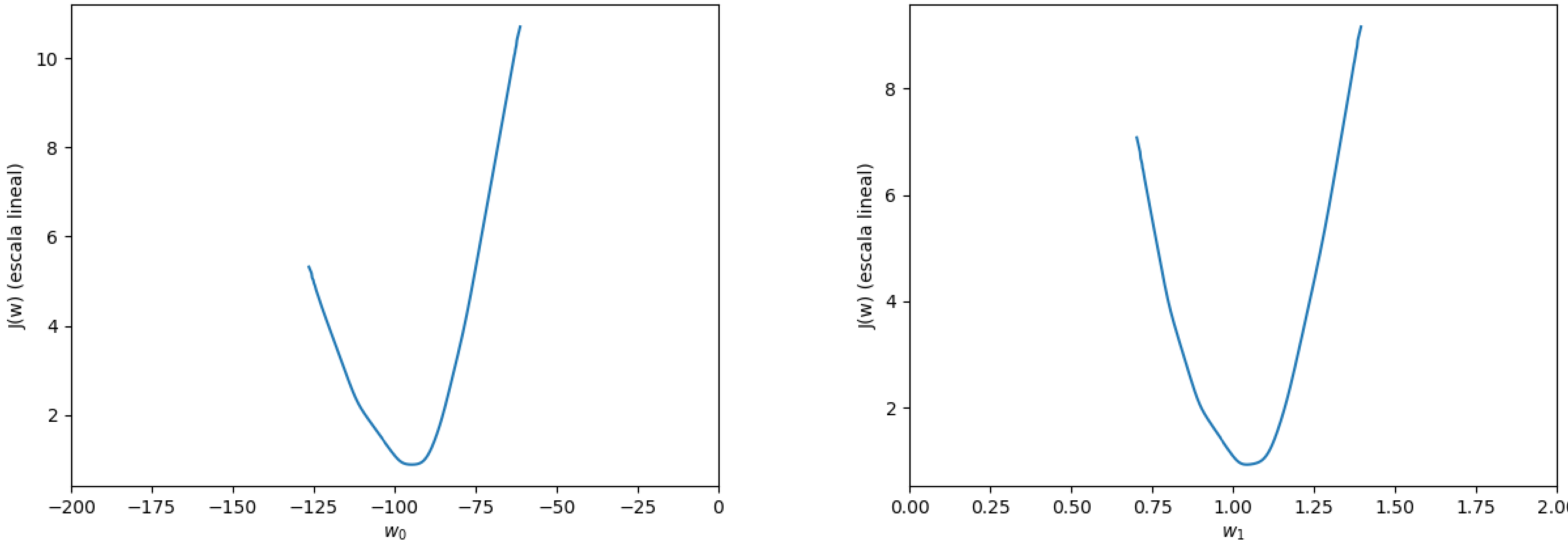
## 4. Elección de la función de coste




$$L(h_w(x), y) = -yLn\big(h_w(x)\big) - (1-y)Ln\big(1 - h_w(x)\big)$$

# Regresión logística

## 4. Elección de la función de coste

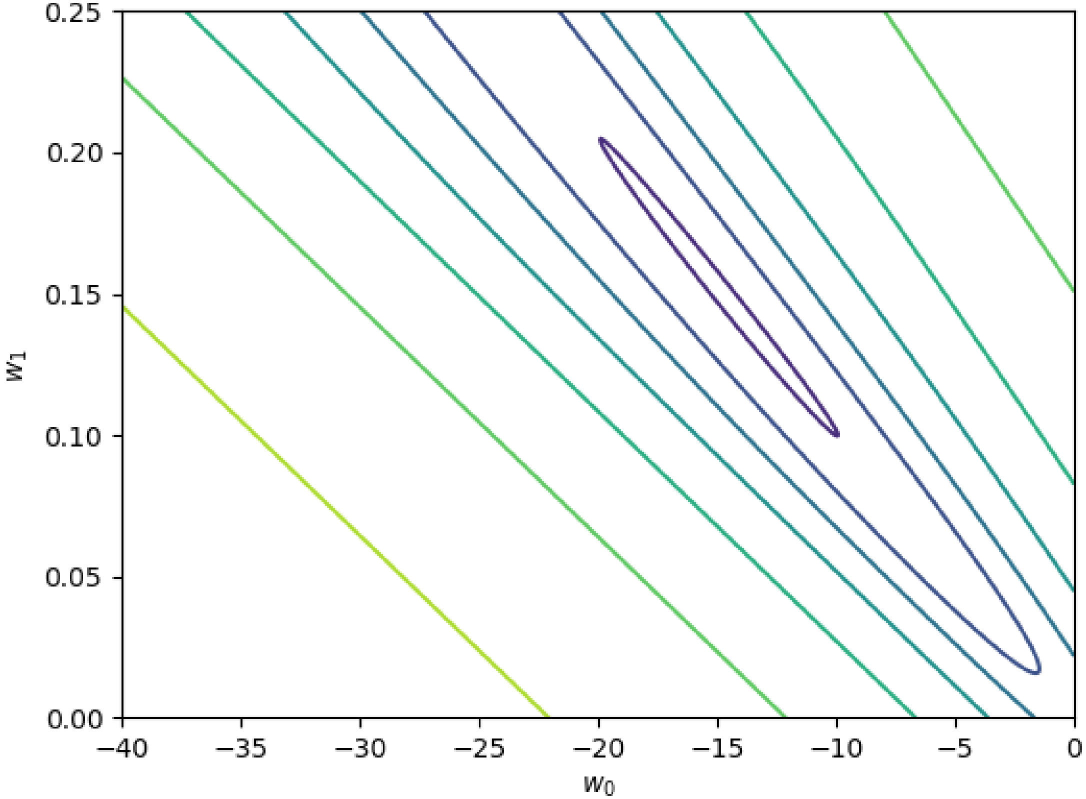



Función de coste logística (cros-entropía)

CONVEXA

# Regresión logística

## 4. Elección de la función de coste




Función de coste logística (cros-entropía)

CONVEXA

# Regresión logística

## 4. Elección de la función de coste

### Función de coste logística con regularización

$$L(h_w(x), y) = -yLn\big(h_w(x)\big) - (1-y)Ln\big(1-h_w(x)\big)$$

$$J(h_w(x), y) = \frac{1}{n}\sum_{i=1}^{n} L\big(h_w(x^{(i)}), y^{(i)}\big) + \lambda\|w\|_2^2$$

# Regresión logística

## 5. Optimización del coste

### Normal Equation

$$w^* = \arg\min_{w} J(h_w(x), y)$$

$$\nabla_w J(h_w(x), y) = 0$$

$$\nabla_w \frac{1}{n} \sum_{i=1}^{n} Loss\left(h_w(x^{(i)}), y^{(i)}\right) = 0$$

$$\sum_{i=1}^{n} \nabla_w Loss\left(h_w(x^{(i)}), y^{(i)}\right) = 0$$

# Regresión logística

## 5. Optimización del coste

### Normal Equation

$$\sum_{i=1}^{n} \nabla_w - Ln\left(1 + \left(2y^{(i)} - 1\right)\left(h_w\left(x^{(i)}\right) - y^{(i)}\right)\right) = 0$$

$$\sum_{i=1}^{n} \frac{\left(2y^{(i)} - 1\right)\left(\nabla_w h_w\left(x^{(i)}\right) - y^{(i)}\right)}{1 + (2y^{(i)} - 1)(h_w(x^{(i)}) - y^{(i)})} = 0$$

$$h_w(x) = \frac{1}{1 + e^{-(w_0 + w_1 x)}}$$

# Regresión logística

## 5. Optimización del coste

### Normal Equation

$$\nabla_w h_w = \begin{bmatrix} \frac{\partial h_w}{\partial w_0} & \frac{\partial h_w}{\partial w_1} \end{bmatrix}$$

$$\frac{\partial h_w}{\partial w_0} = \frac{-1}{\left(1 + e^{-(w_0 + w_1 x)}\right)^2} e^{-(w_0 + w_1 x)} (-1)$$

$$\frac{\partial h_w}{\partial w_1} = \frac{-1}{\left(1 + e^{-(w_0 + w_1 x)}\right)^2} e^{-(w_0 + w_1 x)} (-x)$$

# Regresión logística

## 5. Optimización del coste

### Normal Equation

$$\begin{cases} \displaystyle\sum_{i=1}^{n} \frac{\left(2y^{(i)}-1\right)\left(\frac{\partial h_w\left(x^{(i)}\right)}{\partial w_0}-y^{(i)}\right)}{1+\left(2y^{(i)}-1\right)\left(h_w\left(x^{(i)}\right)-y^{(i)}\right)}=0 \\ \displaystyle\sum_{i=1}^{n} \frac{\left(2y^{(i)}-1\right)\left(\frac{\partial h_w\left(x^{(i)}\right)}{\partial w_1}-y^{(i)}\right)}{1+\left(2y^{(i)}-1\right)\left(h_w\left(x^{(i)}\right)-y^{(i)}\right)}=0 \end{cases}$$

# Regresión logística

## 5. Optimización del coste

### Normal Equation

$$\begin{cases} \displaystyle\sum_{i=1}^{n} \frac{\left(2y^{(i)}-1\right)\left(\dfrac{e^{-\left(w_0+w_1x^{(i)}\right)}}{\left(1+e^{-\left(w_0+w_1x^{(i)}\right)}\right)^2}-y^{(i)}\right)}{1+\left(2y^{(i)}-1\right)\left(\dfrac{1}{1+e^{-\left(w_0+w_1x^{(i)}\right)}}-y^{(i)}\right)} = 0 \\ \displaystyle\sum_{i=1}^{n} \frac{\left(2y^{(i)}-1\right)\left(\dfrac{x^{(i)}e^{-\left(w_0+w_1x^{(i)}\right)}}{\left(1+e^{-\left(w_0+w_1x^{(i)}\right)}\right)^2}-y^{(i)}\right)}{1+\left(2y^{(i)}-1\right)\left(\dfrac{1}{1+e^{-\left(w_0+w_1x^{(i)}\right)}}-y^{(i)}\right)} = 0 \end{cases}$$

# Regresión logística

## 5. Optimización del coste

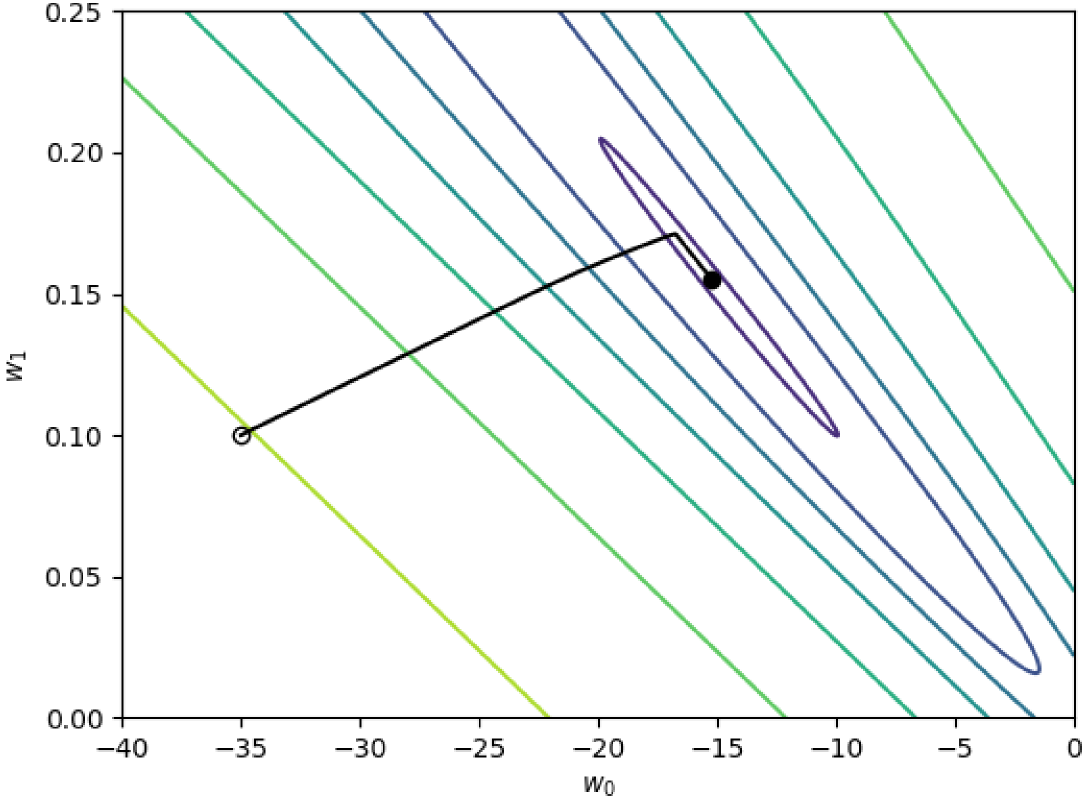
### Normal Equation

$$
\begin{cases}
\displaystyle\sum_{i=1}^{n} \frac{\left(2y^{(i)}-1\right)\left(\dfrac{e^{-\left(w_0+w_1x^{(i)}\right)}}{\left(1+e^{-\left(w_0+w_1x^{(i)}\right)}\right)^2}-y^{(i)}\right)}{1+\left(2y^{(i)}-1\right)\left(\dfrac{1}{1+e^{-\left(w_0+w_1x^{(i)}\right)}}-y^{(i)}\right)}=0\\[2em]
\displaystyle\sum_{i=1}^{n} \frac{\left(2y^{(i)}-1\right)\left(\dfrac{e^{-\left(w_0+w_1x^{(i)}\right)}}{\left(1+e^{-\left(w_0+w_1x^{(i)}\right)}\right)^2}-y^{(i)}\right)}{1+\left(2y^{(i)}-1\right)\left(\dfrac{1}{1+e^{-\left(w_0+w_1x^{(i)}\right)}}-y^{(i)}\right)}=0
\end{cases}
$$

Sin solución analítica

# Regresión logística

## 5. Optimización del coste




$$w_0^* = -15.2$$
$$w_1^* = 0.155$$

Gradient Descent

# Regresión logística

- Hipótesis
  - Función signo
  - Función logística
- Función de coste
  - 0-1
  - Logística
- Evaluación
  - Matriz de confusión y métricas derivadas
  - Análisis ROC
  - Curva de aprendizaje

# Regresión logística

## 6. Evaluación del resultado

$n_{train} = 10$

$t_h = 0.5$

| Cliente | Superficie (m²) $x^{(i)}$ | Garaje $y^{(i)}$ | Predicción $h^{(i)}$ | Garaje $\hat{y}^{(i)}$ |
|---|---|---|---|---|
| 11 | 139 | S | 0.998 | S |
| 12 | 54 | N | 0.001 | N |
| 13 | 96 | S | 0.428 | N |
| 14 | 95 | N | 0.372 | N |
| 15 | 132 | S | 0.994 | S |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

$$h_w(x) = \frac{1}{1 + e^{-(w_0 + w_1 x)}}$$

$$\hat{y} = \begin{cases} 0, \forall h_w(x) < t_h \\ 0, \forall h_w(x) \geq t_h \end{cases}$$

# Regresión logística

## 6. Evaluación del resultado

### Matriz de confusión

| | | Clase calculada | | |
|---|---|---|---|---|
| | | ***P*** | ***N*** | **Elementos** |
| **Clase real** | ***P*** | **TP** | **FN** | $n_P$ |
| | ***N*** | **FP** | **TN** | $n_N$ |
| **Estimaciones** | | $e_P$ | $e_N$ | $n$ |

| | | Clase calculada | | |
|---|---|---|---|---|
| | | ***P*** | ***N*** | **Elementos** |
| **Clase real** | ***P*** | 545 | 46 | 591 |
| | ***N*** | 77 | 322 | 399 |
| **Estimaciones** | | 622 | 368 | 990 |

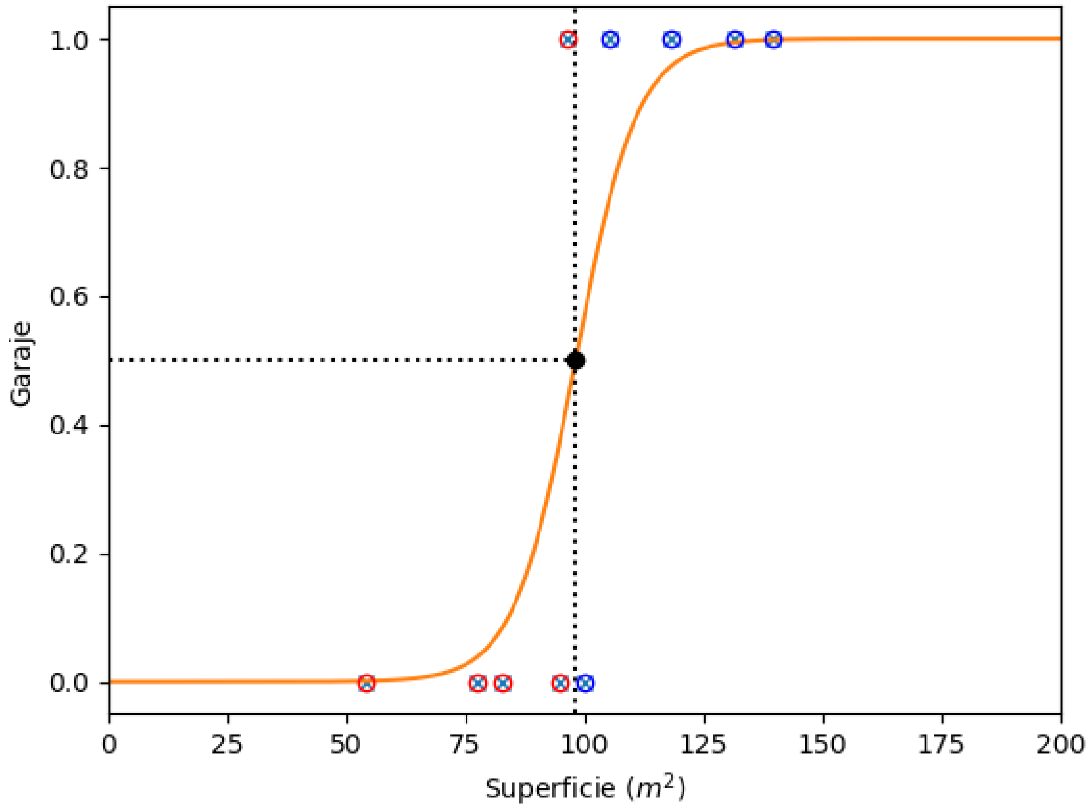
# Regresión logística

## 6. Evaluación del resultado

| Nombre | Acrónimo | Expresión | Valor |
|---|---|---|---|
| Sensitivity | SNS | $\frac{TP}{TP+FN}$ | 0.9222 |
| Specificity | SPC | $\frac{TN}{TN+FP}$ | 0.8070 |
| Precision | PRC | $\frac{TP}{TP+FP}$ | 0.8762 |
| Negative Predictive Value | NPV | $\frac{TN}{TN+FN}$ | 0.8750 |
| Accuracy | ACC | $\frac{TP+TN}{TP+TN+FP+FN}$ | 0.8758 |
| $F_1$ Score | $F_1$ | $2\frac{SNS \cdot PRC}{SNS+PRC}$ | 0.8986 |
| Geometric Mean | GM | $\sqrt{SNS \cdot SPC}$ | 0.8627 |

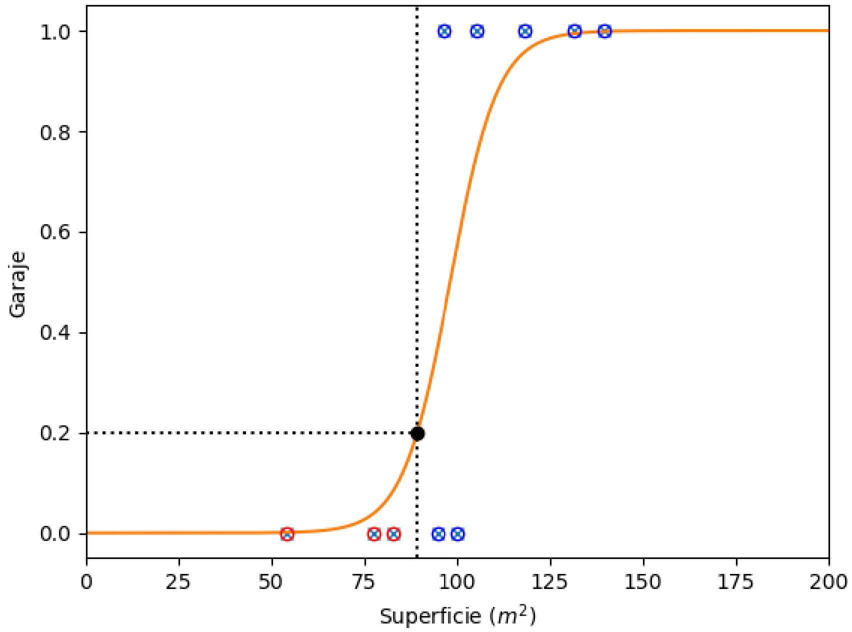
# Regresión logística

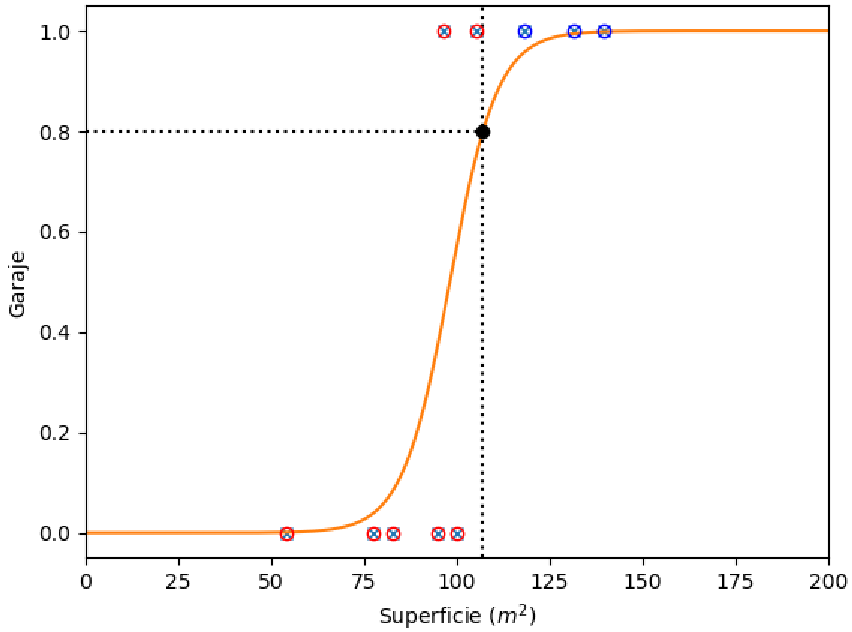
## 6. Evaluación del resultado



# Regresión logística

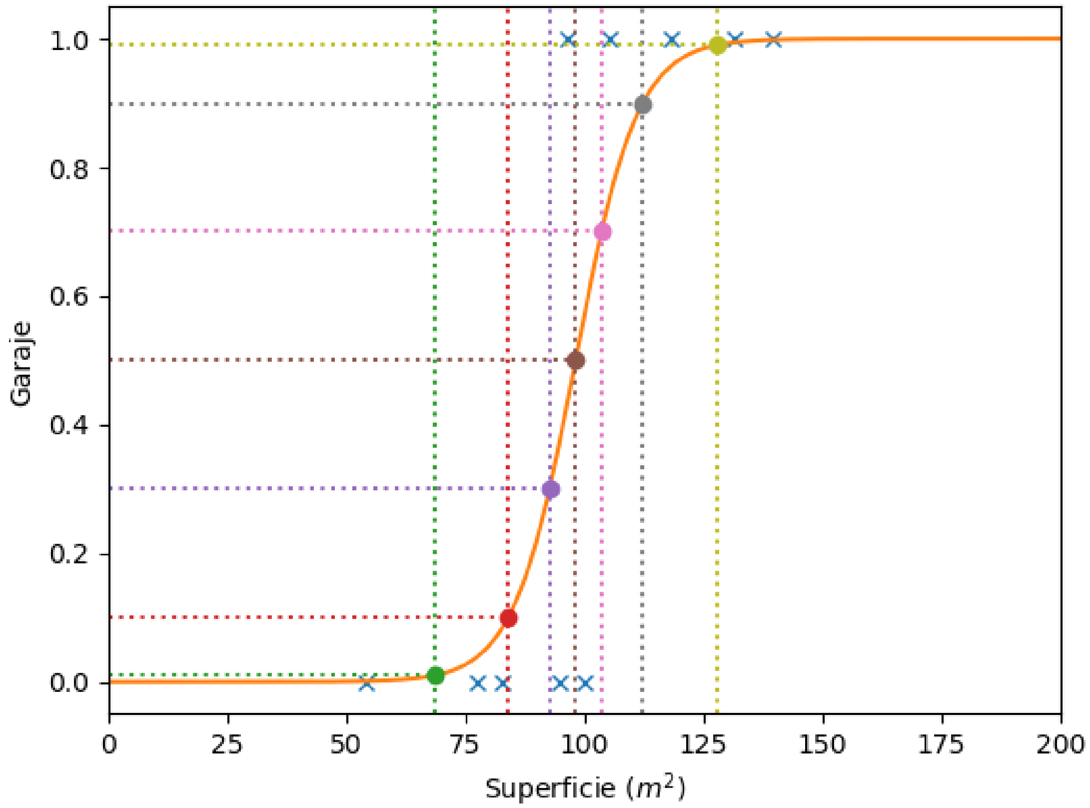
## 6. Evaluación del resultado



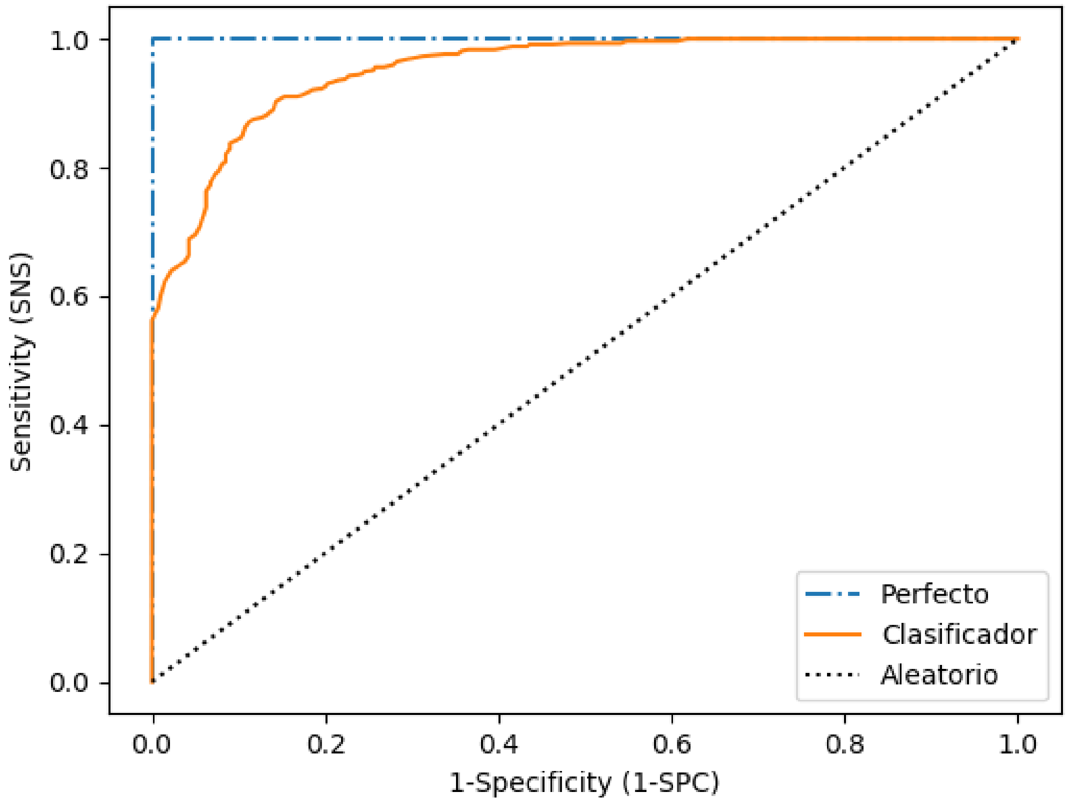
# Regresión logística

## 6. Evaluación del resultado



# Regresión logística

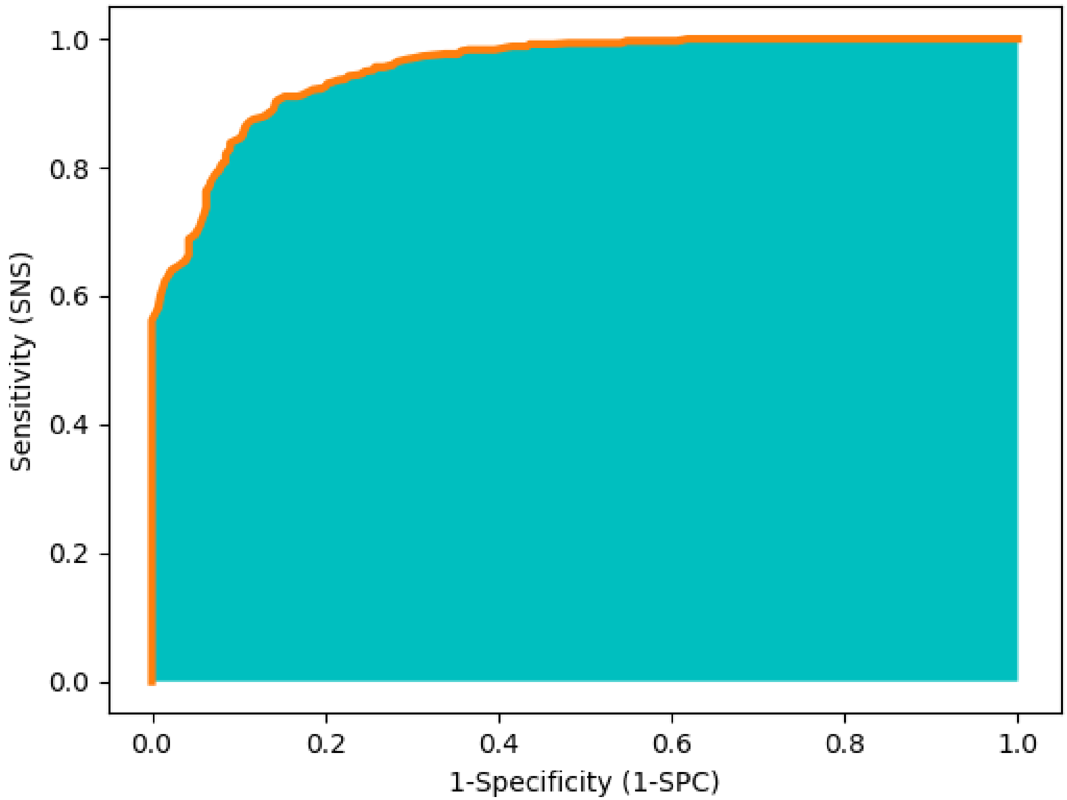
## 6. Evaluación del resultado




Receiver operating characteristic (ROC)

# Regresión logística

## 6. Evaluación del resultado

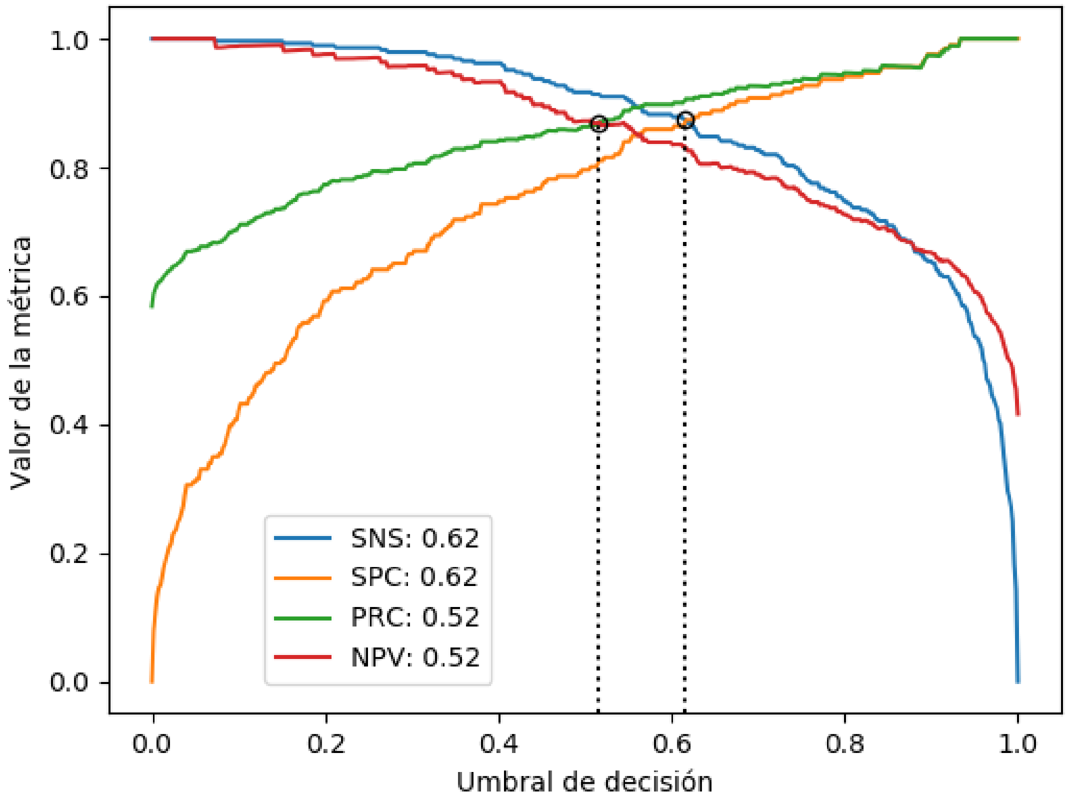



Area Under Curve (AUC)

$$AUC = 0.9519$$

# Regresión logística

## 6. Evaluación del resultado

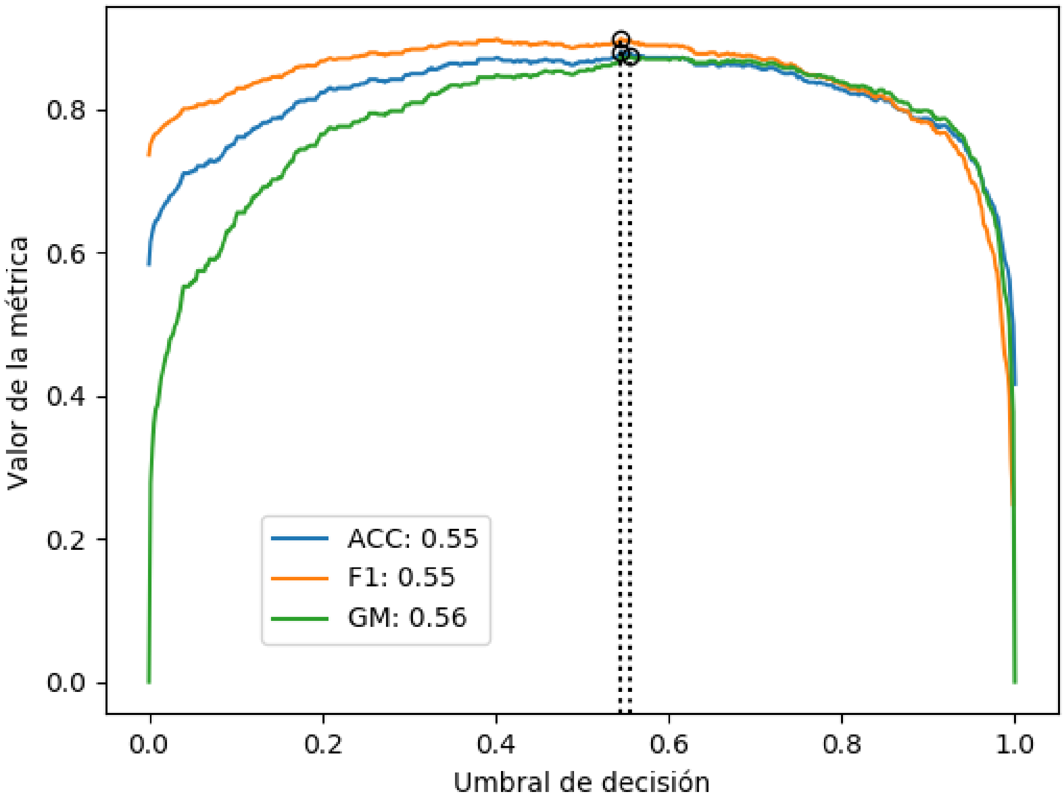



Elección del umbral de decisión

Datos de validación

# Regresión logística

## 6. Evaluación del resultado

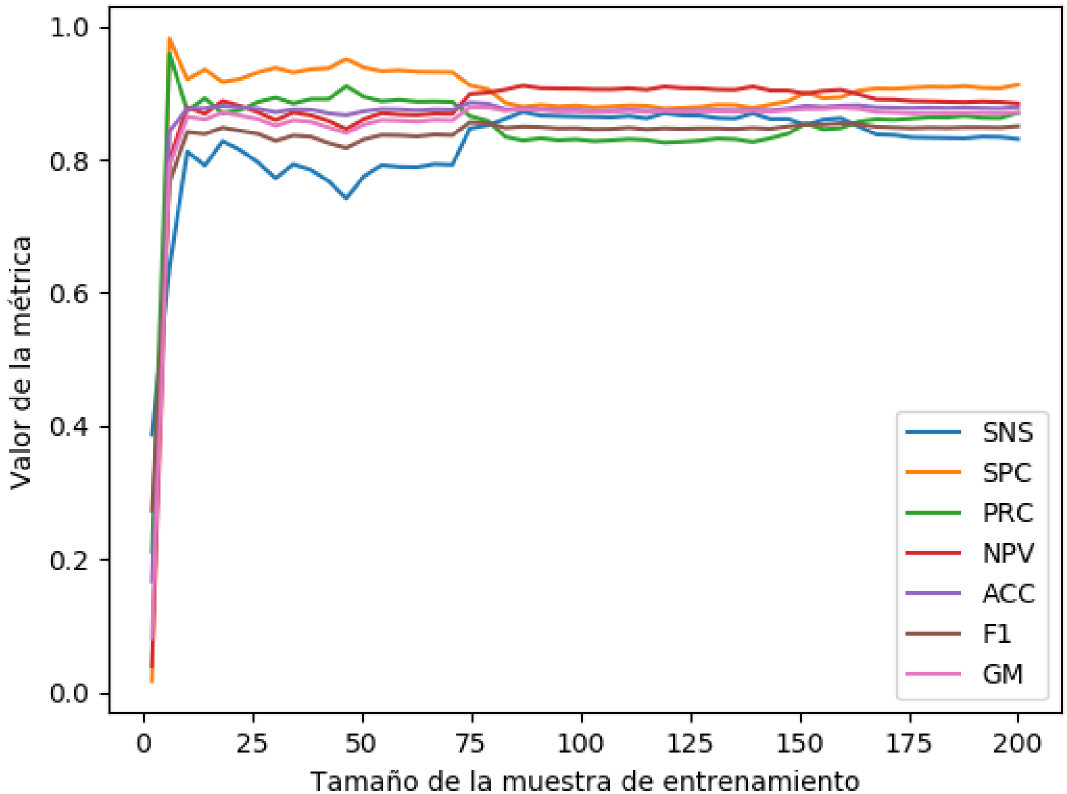



Elección del umbral de decisión

Datos de validación

# Regresión logística

## 6. Evaluación del resultado




Curva de aprendizaje