Regresión Logística

Luis Zúñiga

Agenda

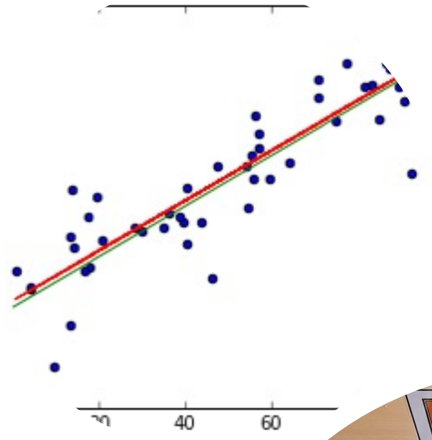
1. Introducción
2. El problema de clasificación
3. Definición de la hipótesis
4. Estimación de parámetros

Introducción

Los modelos de aprendizaje que vimos anteriormente permiten resolver problemas de **regresión**.

Ahora, vamos a considerar un modelo de **clasificación**: regresión logística.

(¿No les hace ruido el nombre?)



El Problema de Clasificación

El problema de clasificación se reduce a **asignar cada elemento del conjunto de datos a una de dos o más clases**. Por el momento, vamos a facilitar las cosas y **consideraremos únicamente dos clases**. Por ejemplo:

- Email: Spam / No es spam
- Transacciones en línea: Válida / Fraudulenta
- Tumores: Maligno / Benigno

El Problema de Clasificación

Usualmente, las clases se representan de una manera sencilla mediante valores numéricos:

$$y \in \{0,1\}$$

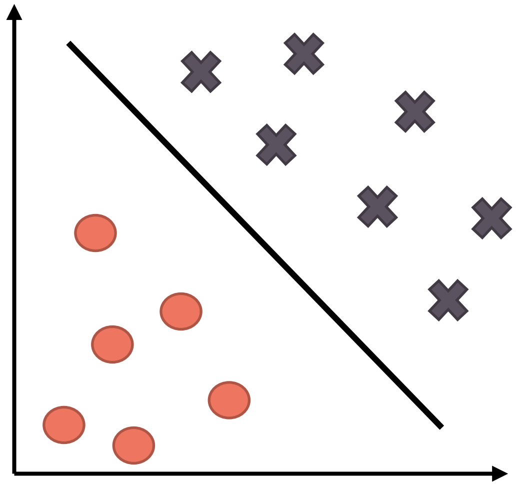
donde

- 0: clase negativa (e.g., no es spam)
- 1: clase positiva (e.g., es spam)

Otra vez: vamos a pensar con clasificación binaria.

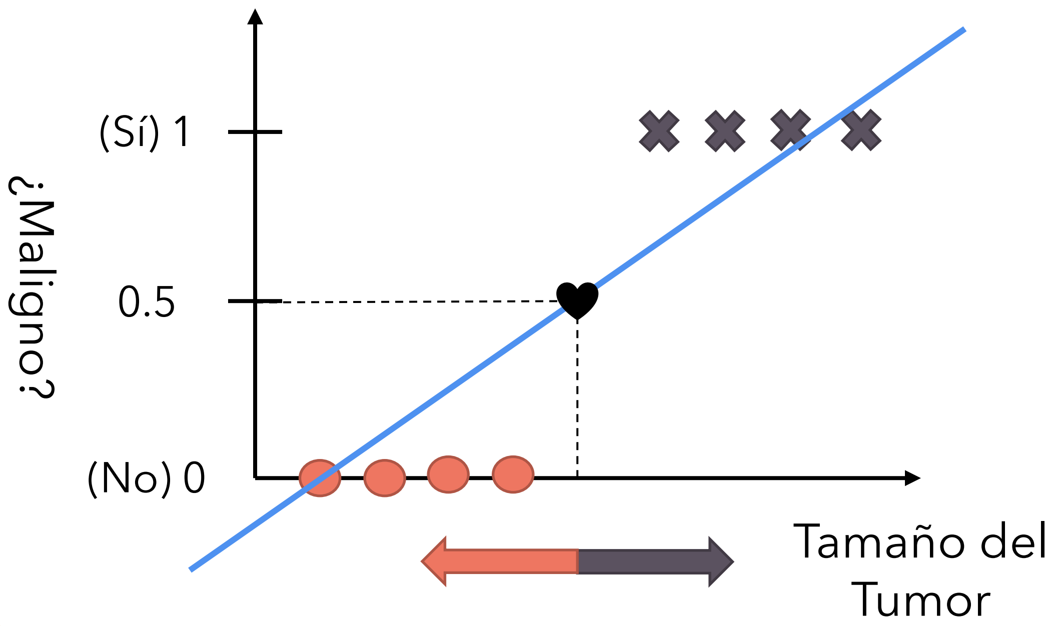
El Problema de Clasificación

¿Cómo separaría las dos clases?



Vamos a usar esta idea más adelante.

Primera Idea: Usemos lo que Tenemos



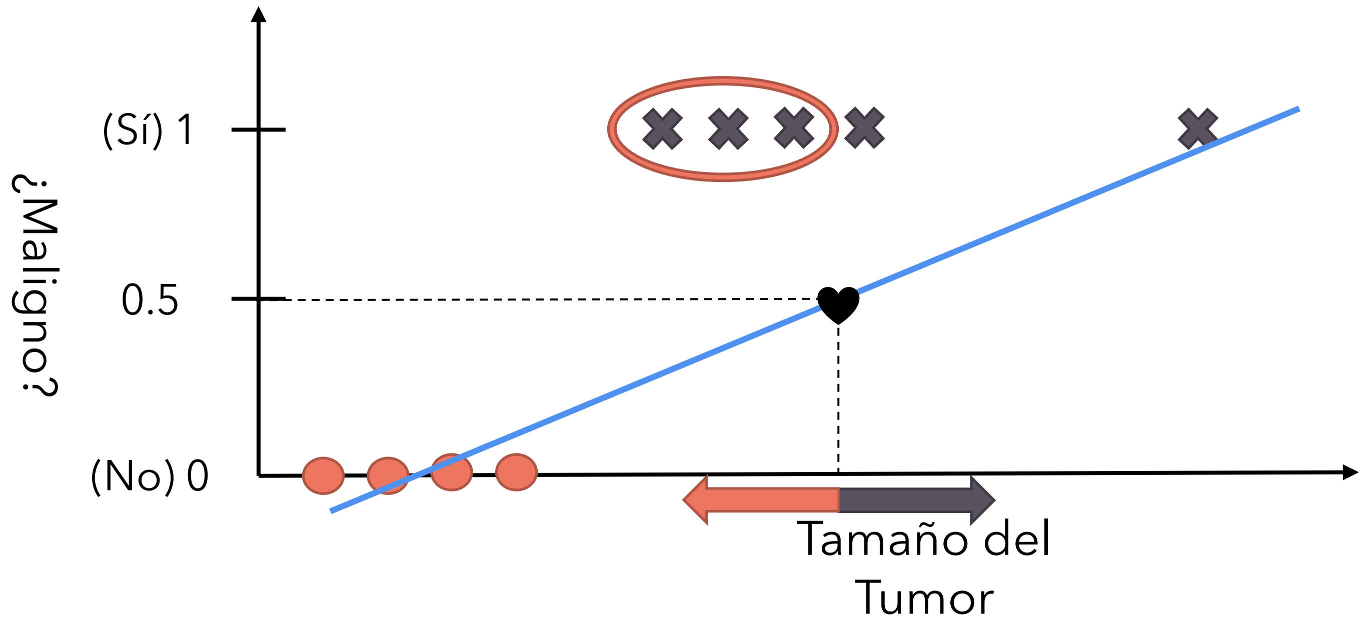
Vamos a utilizar lo que sabemos hasta el momento. Usemos un modelo de regresión lineal simple

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

Para separar los puntos, vamos a definir un valor límite de tal manera que:

$$\begin{cases} h_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq 0.5 \,, y = 1 \\ h_{\boldsymbol{\theta}}(\boldsymbol{x}) < 0.5 \,, y = 0 \end{cases}$$

Primera Idea: Usemos lo que Tenemos



Vamos a utilizar lo que sabemos hasta el momento. Usemos un modelo de regresión lineal simple

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

Para separar los puntos, vamos a definir un valor límite de tal manera que:

$$\begin{cases} h_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq 0.5\,, y = 1 \\ h_{\boldsymbol{\theta}}(\boldsymbol{x}) < 0.5\,, y = 0 \end{cases}$$

Primera Idea: Usemos lo que Tenemos

- En práctica, regresión lineal es un algoritmo que rara vez se adapta a nuestras necesidades.

- Además, al obtener valores en $\mathbb{R}$, se puede dar el caso
$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) > 1 \text{ o } h_{\boldsymbol{\theta}}(x) < 0$$
cuando se requiere que
$$y = 0 \text{ o } y = 1$$

Modelo de Regresión Logística

Se busca que $0 \leq h_\theta \leq 1$. Con este fin, se propone usar la siguiente función:

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

donde

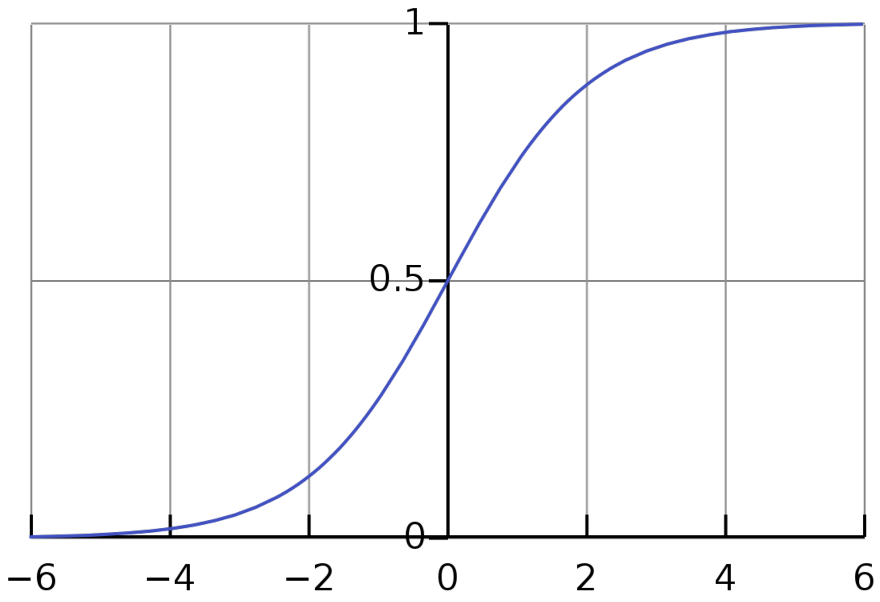
$$g(z) = \frac{1}{1 + e^{-z}}$$

es decir,

$$h_{\boldsymbol{\theta}} = \frac{1}{1 + e^{-\Theta^T \boldsymbol{x}}}$$

A esta función se le conoce como **función sigmoide** o **función logística**.

¿Cómo se ve? ¡A graficar!



Modelo de Regresión Logística

¿Cómo un valor en $\mathbb{R}$ puede determinar un valor en $y = \{0,1\}$?

$h_{\boldsymbol{\theta}}(\boldsymbol{x})$ = estimación de la probabilidad que $y = 1$ para el vector de datos $\boldsymbol{x}$.

Por ejemplo, si $h_{\boldsymbol{\theta}}(\boldsymbol{x}) = 0.7$, el paciente tiene un 70% de probabilidad de que el tumor sea maligno.

Límite de Decisión

¿Cómo se decide qué clase se le asigna a cada dato?

Hipótesis

$$h_{\boldsymbol{\theta}} = g(\Theta^T \boldsymbol{x}) = P(y = 1|x; \boldsymbol{\theta})$$

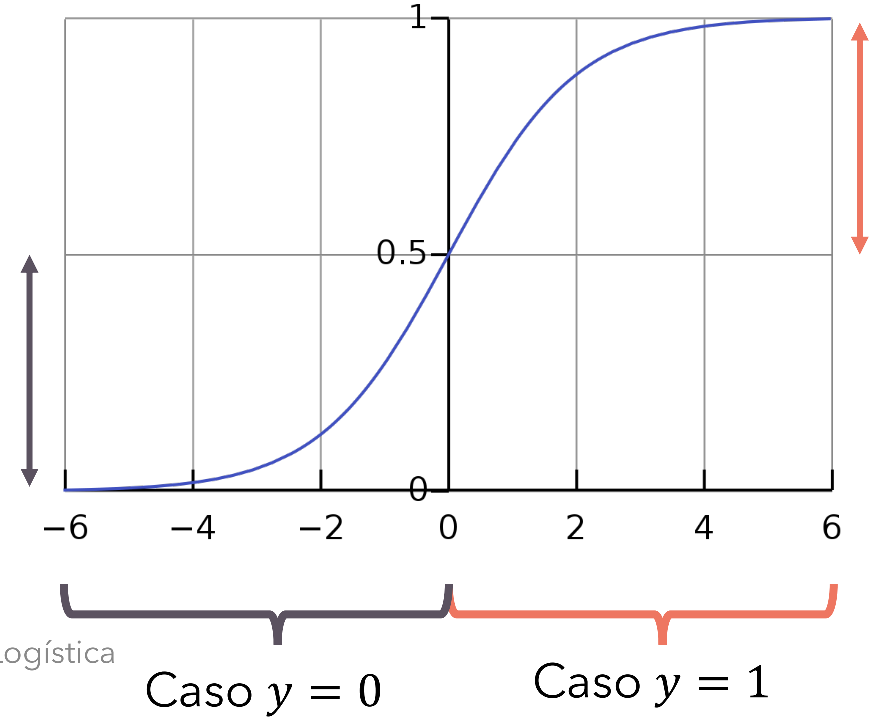
$$g(z) = \frac{1}{1 + e^{-z}}$$

Vamos a establecer la siguiente regla:

$$y = 1 \text{ si } h_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq 0.5$$

y

$$y = 0 \text{ si } h_{\boldsymbol{\theta}}(\boldsymbol{x}) < 0.5$$



Función de Decisión

De nuestro análisis anterior, si

$$y = 1 \ \text{ si } \ h_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq 0.5$$

implica que $g(z) \geq 0.5$ cuando $z = \Theta^T \boldsymbol{x} \geq 0$. De forma similar, si

$$y = 0 \ \text{ si } \ h_{\boldsymbol{\theta}}(\boldsymbol{x}) < 0.5$$

implica que $g(z) < 0.5$ cuando $z = \Theta^T \boldsymbol{x} < 0$.

Función de Decisión

Actividad

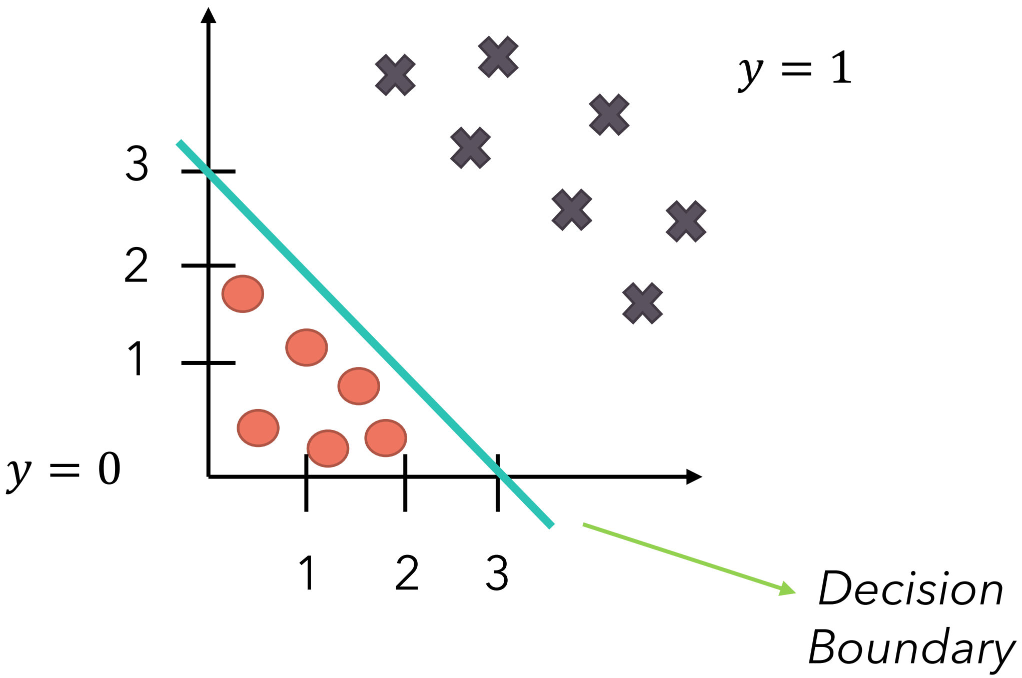
Supongamos que tenemos la siguiente función:

$$h_{\boldsymbol{\theta}}(\mathbf{x}) = \mathrm{g}(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

donde

$$\boldsymbol{\theta} = (-3,1,1).$$

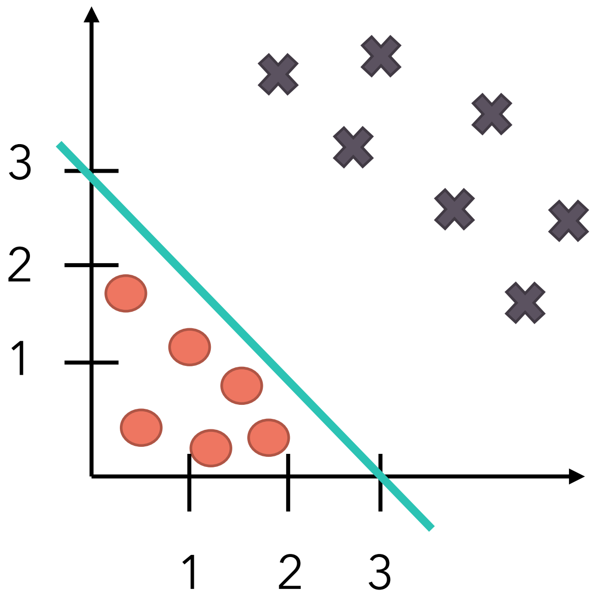
¿Cómo se ve la ecuación que determina el caso $y = 1$ $(h_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq 0.5)$?



$$-3 + x_1 + x_2 \geq 0$$

$$x_1 + x_2 \geq 3$$

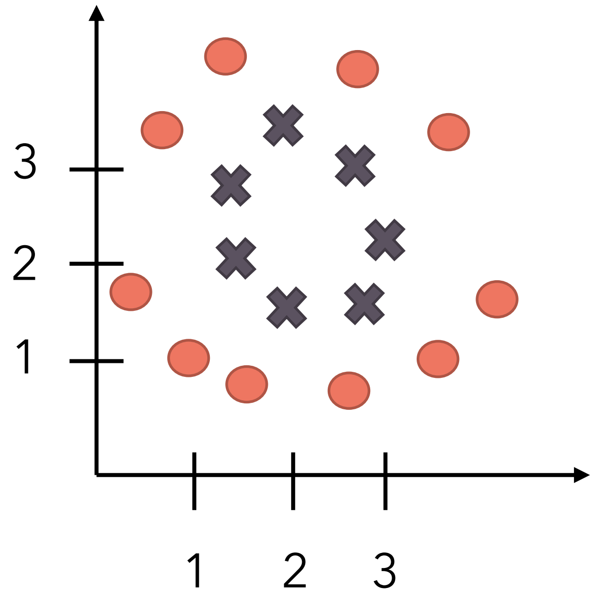
Función de Decisión



¿Qué define nuestra función de separación?

- Este es un ejemplo ideal, rara vez se presenta el caso de la separación lineal.
- Todo depende de la forma de los datos, ya que estos definen los parámetros de la función de decisión.
- Pueden utilizar cualquier función, no solo funciones lineales.

Función de Decisión



¿Cómo sería la función de decisión en este caso?

Para recapitular

Se busca que $0 \leq h_\theta \leq 1$. Con este fin, se propone usar la siguiente función:

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

donde

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

es decir,

$$h_{\boldsymbol{\theta}} = \frac{1}{1 + e^{-\Theta^T \boldsymbol{x}}}$$

- z nos permite obtener valores entre 0 y 1, similar a una función de probabilidad.
- z puede ser cualquier función, inclusive no lineal.
- z dicta la función que va a separar los puntos de ambas clases.

Estimación de Parámetros

Conjunto de Datos:

$$\{(x_i, y_i)_{i=1}^{n}\}$$

Hipótesis:

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{1 + e^{-\Theta^{\mathrm{T}}\mathbf{x}}}$$

***n* ejemplos**

$$x \in \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad x_0 = 1, \quad y \in \{0,1\}$$

Función de Costo

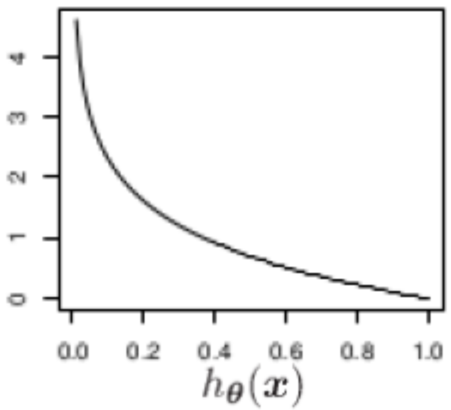
$$\mathrm{Costo}(h_{\boldsymbol{\theta}}(x), \mathrm{y}) = \begin{cases} -\log\big(h_{\boldsymbol{\theta}}(\boldsymbol{x})\big) \text{ si } y = 1 \\ -\log\big(1 - h_{\boldsymbol{\theta}}(\boldsymbol{x})\big) \text{ si } y = 0 \end{cases}$$

Estimación de Parámetros

Recuerden que *y* siempre es 0 o 1.

$$\operatorname{cost}\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}),\ y\right) = \begin{cases} -\log(h_{\boldsymbol{\theta}}(\boldsymbol{x})) & \text{if } \ y = 1 \\ -\log(1 - h_{\boldsymbol{\theta}}(\boldsymbol{x})) & \text{if } \ y = 0 \end{cases}$$

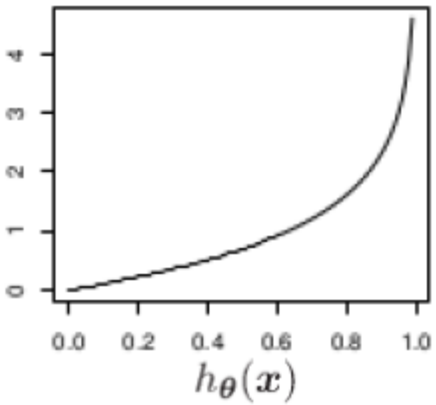
if $y = 1$



if $h_{\boldsymbol{\theta}}(\boldsymbol{x}) = 1$
then cost $= 0$

if $h_{\boldsymbol{\theta}}(\boldsymbol{x}) \to 0$
then cost $\to \infty$
predicted
prob$(y = 1 \mid \boldsymbol{x}; \boldsymbol{\theta}) = 0$
but $y = 1$

if $y = 0$



if $h_{\boldsymbol{\theta}}(\boldsymbol{x}) = 0$
then cost $= 0$

if $h_{\boldsymbol{\theta}}(\boldsymbol{x}) \to 1$
then cost $\to \infty$
predicted
prob$(y = 0 \mid \boldsymbol{x}; \boldsymbol{\theta}) = 0$
but $y = 0$

Estimación de Parámetros

Función de Costo

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

Actividad

La función de costo se puede juntar en un sola expresión. Encontrar dicha expresión.

La función de costo se puede reducir a **una expresión**:

$$\text{Costo}(h_{\boldsymbol{\theta}}(x), \mathrm{y}) = -y \log\big(h_{\boldsymbol{\theta}}(\boldsymbol{x})\big) - (1 - y) \log(1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}))$$

Estimación de Parámetros

Hipótesis:

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{1 + e^{-\Theta^{\mathrm{T}}\mathbf{x}}}$$

La función de costo se ve (finalmente) de la siguiente manera:

$$J(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n}[y^{(i)}\log\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)})\right) + (1 - y^{(i)})\log\left(1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)})\right)]$$

Aplicando gradiente descendiente:

$$\theta_j = \theta_j - \eta\frac{\partial}{\partial\theta_j}J(\boldsymbol{\theta})$$

donde

$$\frac{\partial}{\partial\theta_j}J(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) - y^{(i)}\right)x_j^{(i)}$$

Es decir:

$$\theta_j = \theta_j - \eta\frac{1}{n}\sum_{i=1}^{n}\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) - y^{(i)}\right)x_j^{(i)}$$

¡Lo mismo que el modelo de regresión lineal! Solo diferente $h_{\boldsymbol{\theta}}(\boldsymbol{x})$.

Estimación de Parámetros

Hipótesis:

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{1 + e^{-\Theta^{\mathrm{T}}\mathbf{x}}}$$

Tarea

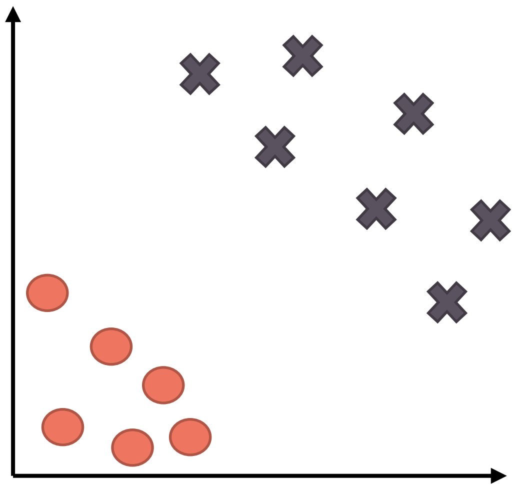
Demostrar que

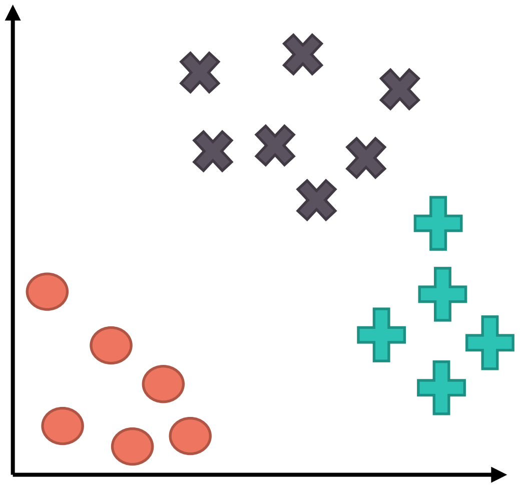
$$\frac{\partial}{\partial \theta_j} J(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right) x_j^{(i)}$$

donde:

$$J(\boldsymbol{\theta}) = -\frac{1}{n} \sum_{i=1}^{n} \left[y^{(i)} \log\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right)\right) + \left(1 - y^{(i)}\right) \log\left(1 - h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right)\right)\right]$$

Clasificación Multiclase

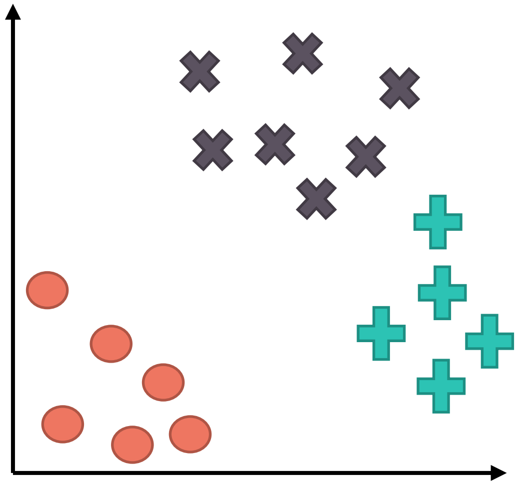


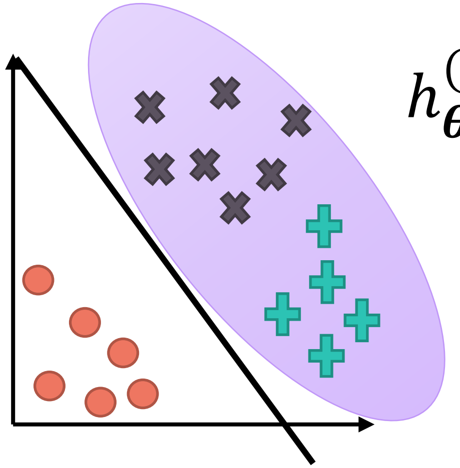


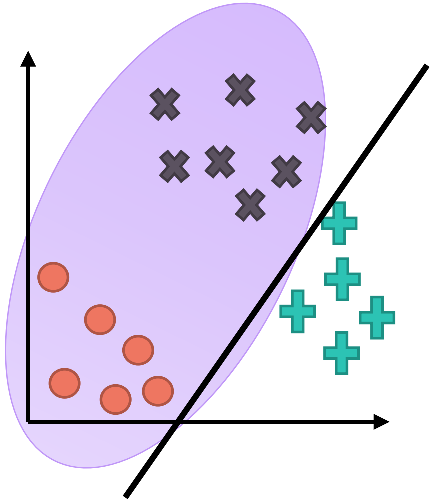
Clasificación Multiclase

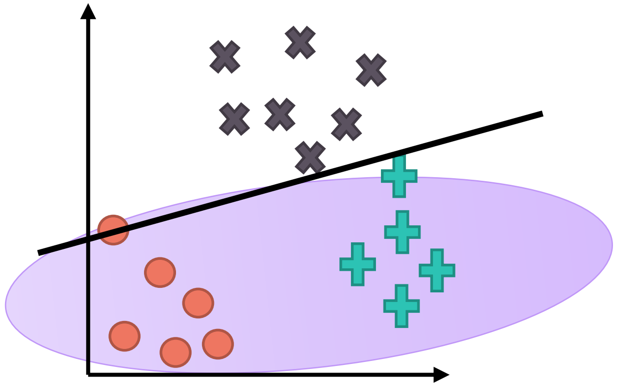
$$h_{\boldsymbol{\theta}}^{(i)}(\boldsymbol{x}) = \mathrm{P}(\mathrm{y} = \mathrm{i}|\mathrm{x}; \boldsymbol{\theta})$$
$$y = \max_{i} h_{\boldsymbol{\theta}}^{(i)}(\boldsymbol{x})$$

One-vs-all (one-vs-rest)









Gracias

Luis Zúñiga

p40887@correo.uia.mx

https://lzun.github.io