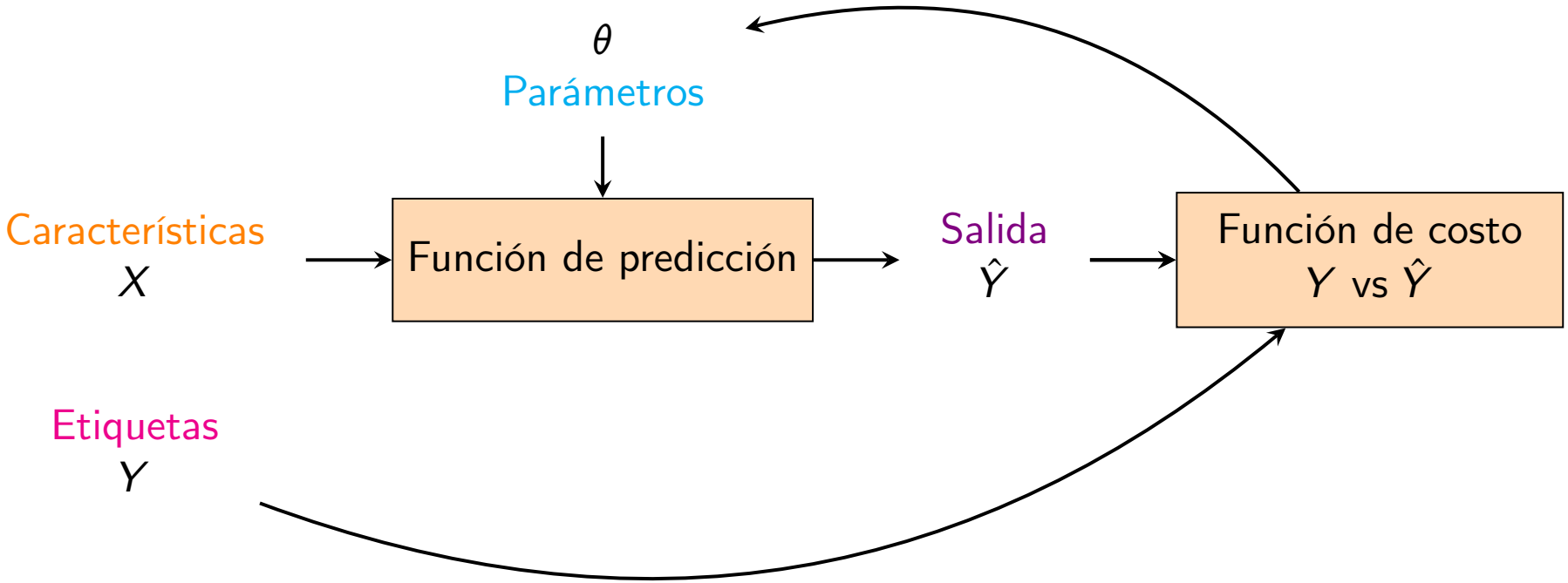
Procesamiento de Lenguaje Natural

Olivia Gutú y Julio Waissman

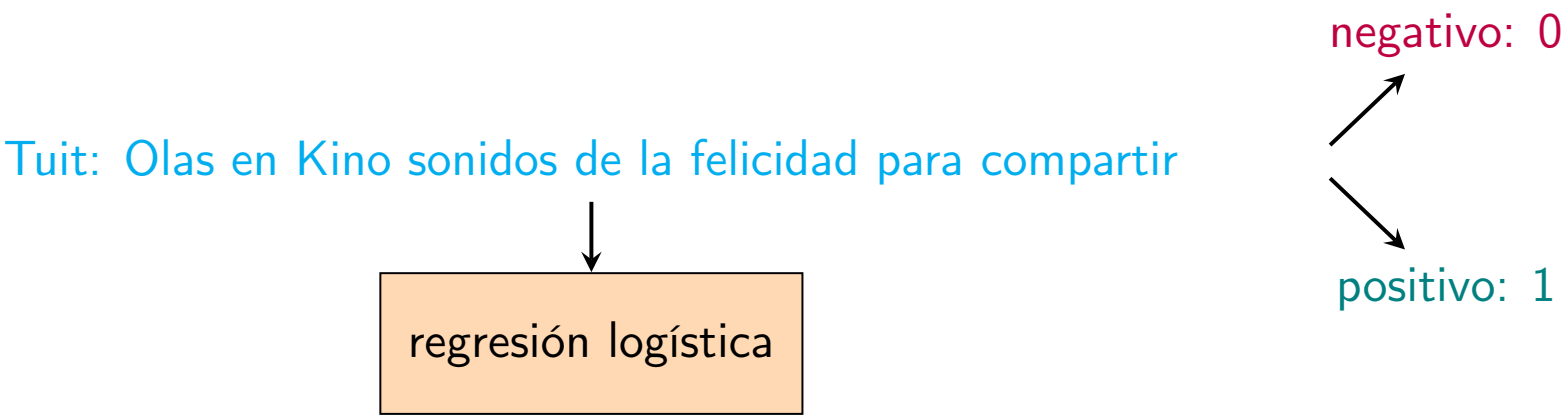
Maestría en Ciencia de Datos
Semana 1: Análisis de sentimientos y regresión logística

ML: flujo de entrenamiento

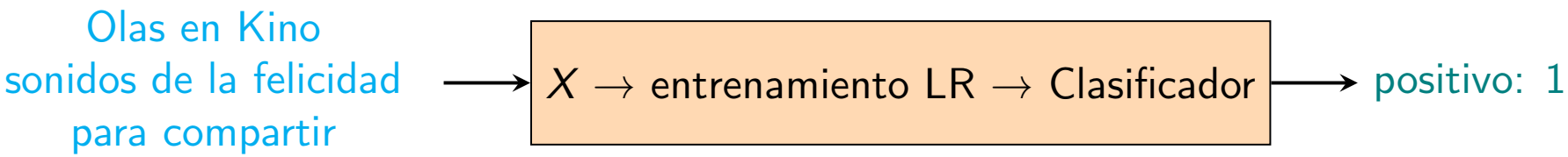


Análisis de sentimiento

Objetivo: predecir si un texto (e.g. un tuit) tiene un *sentimiento* negativo o positivo.



Análisis de sentimiento



Vocabulario

Tuits:

$\text{tuit}_1, \text{tuit}_2, \ldots, \text{tuit}_m$ $\longrightarrow$

Olas en Kino sonidos de la felicidad para compartir

$\vdots$

Estoy triste el día de hoy

$V = \{\text{Olas, en, Kino, sonidos, de, la, felicidad, para, } \ldots, \text{ triste, día, hoy}\}$

Extracción de características

Olas en Kino sonidos de la felicidad para compartir

V:	Olas	en	Kino	sonidos	de	la	felicidad	para	...	triste	día
	↓	↓	↓	↓	↓	↓	↓	↓	...	↓	↓
	1	1	1	1	1	1	1	1		0	0

¡demasiados ceros! (representación dispersa)

Problemas con la representación dispersa

- A cada tuit le asigno un vector en $\mathbb{R}^n$, n es la cardinalidad de V (*Quijote*: 23,000, el coreano: 1,100,373, en la red: !uf!).
- La regresión logística tendría que ajustar $\theta = (\theta_0, \theta_1, \ldots \theta_n) \in \mathbb{R}^{n+1}$.
- Entrenamiento costoso y largo tiempo de predicción.

Extracción por frecuencia

frec(w, c) frecuencia de la palabra en la clase

V	PosFrec(1)	NegFrec(0)
Olas	4	0
en	3	3
Kino	5	1
de	6	6
la	3	3
felicidad	5	0
⋮	⋮	⋮
triste	1	7
día	1	1
hoy	0	1

Extracción por frecuencia

$$X^m = [\ 1,\ \sum_w \text{frec}(w,1),\ \sum_w \text{frec}(w,0)\ ]$$

X^m	1	$\sum_w \text{frec}(w,1)$	$\sum_w \text{frec}(w,0)$
Carac. del tuit m.	sesgo	suma de frec. pos	suma de frec. neg.

En las sumas w recorre las palabras que contiene el tuit m (una sola vez)

Extracción por frecuencia

Tuit$_m$: Olas hoy felicidad hoy

$X^m = [1, 9, 1]$

V	PosFrec(1)	NegFrec(0)
Olas	4	0
en	3	3
Kino	5	1
de	6	6
la	3	3
felicidad	5	0
⋮	⋮	⋮
triste	1	7
día	1	1
hoy	0	1

Pre-procesamiento de los datos

Tuit: Vista desde San Carlos Sonora el día de hoy @webcamsdemexico @VisitSonora http://visitsonora.mx/destinos/playas/san-carlos/

- signos de puntuación y *stop words* (palabras vacías)
- borrar *urls* y *handles*
- *steamming* (lexematización) y *lowecasing* (todo en minúscula)

Tuit procesado: [vista, san, carlos, sonora, día, hoy]

Pre-procesamiento de los datos: stop words

- preposiciones
- conjunciones y disyunciones
- verbos copulativos y auxiliares
- palabras muy comunes

a
acá
ahí
al
algún/a/o/s
⋮
voy
vuestra/o/s
y
ya
yo

Pre-procesamiento de los datos: steamming

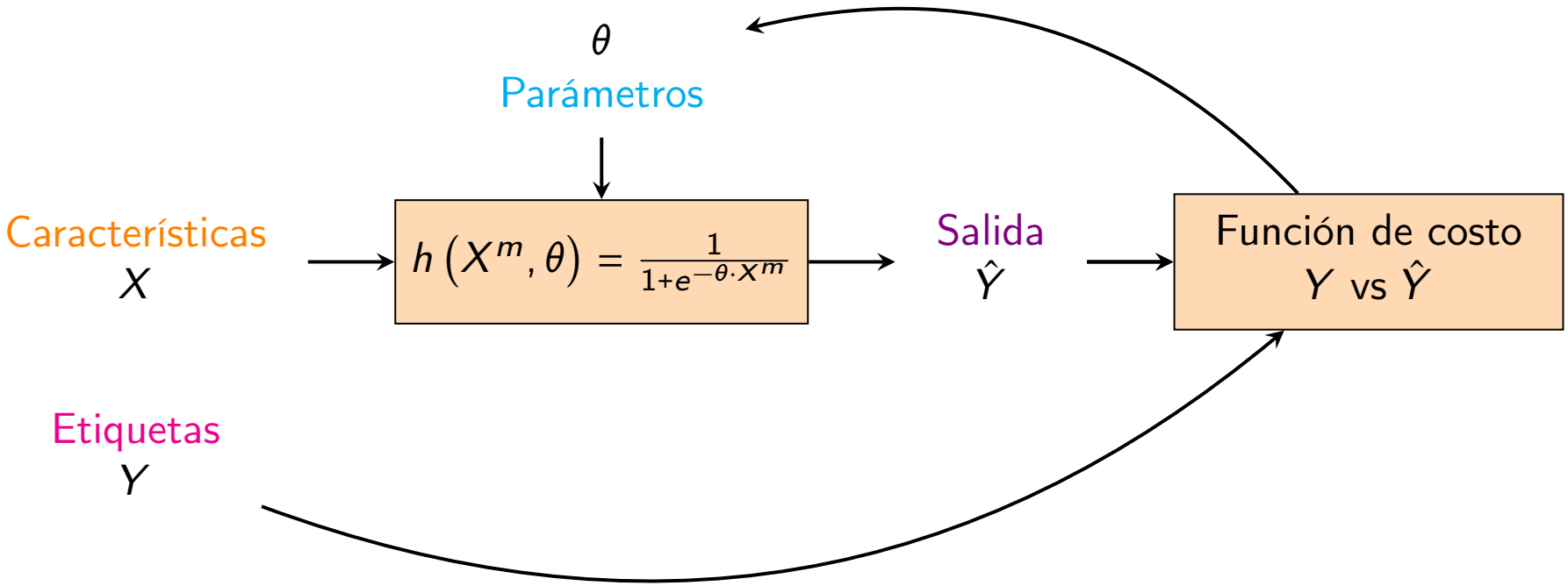
e.g. `http://stemmer-es.sourceforge.net`

che	che
checa	chec
checar	chec
checo	chec
checoslovaquia	checoslovaqui
chedraoui	chedraoui
chefs	chefs
cheliabinsk	cheliabinsk
chelo	chel
chía	chi
chiapaneca	chiapanec
chica	chic
chichimecas	chichimec

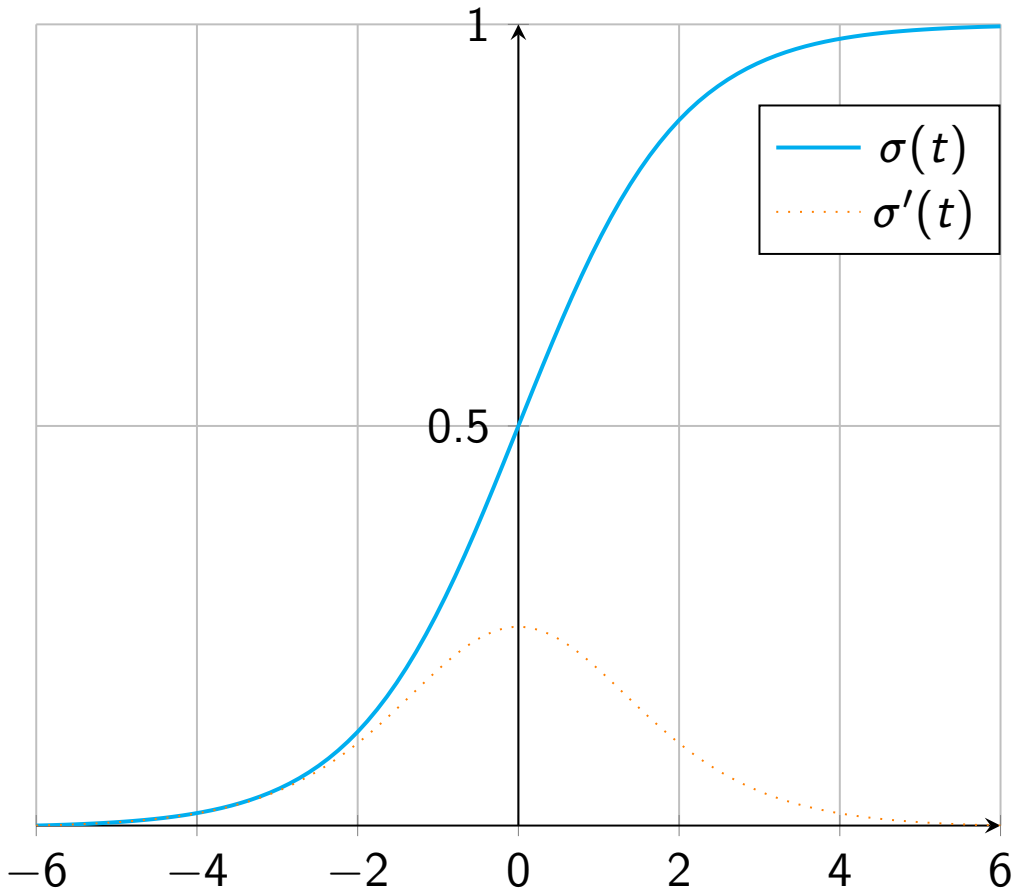
Matriz de características

$$\begin{bmatrix} X^1 \\ X^2 \\ \vdots \\ X^m \end{bmatrix} \rightarrow \begin{bmatrix} 1 & X_1^{(1)} & X_2^{(1)} \\ 1 & X_1^{(2)} & X_2^{(2)} \\ \vdots & \vdots & \vdots \\ 1 & X_1^{(m)} & X_2^{(m)} \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 3 & 2 \\ 1 & 4 & 2 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 6 \end{bmatrix}$$

Regresión logística



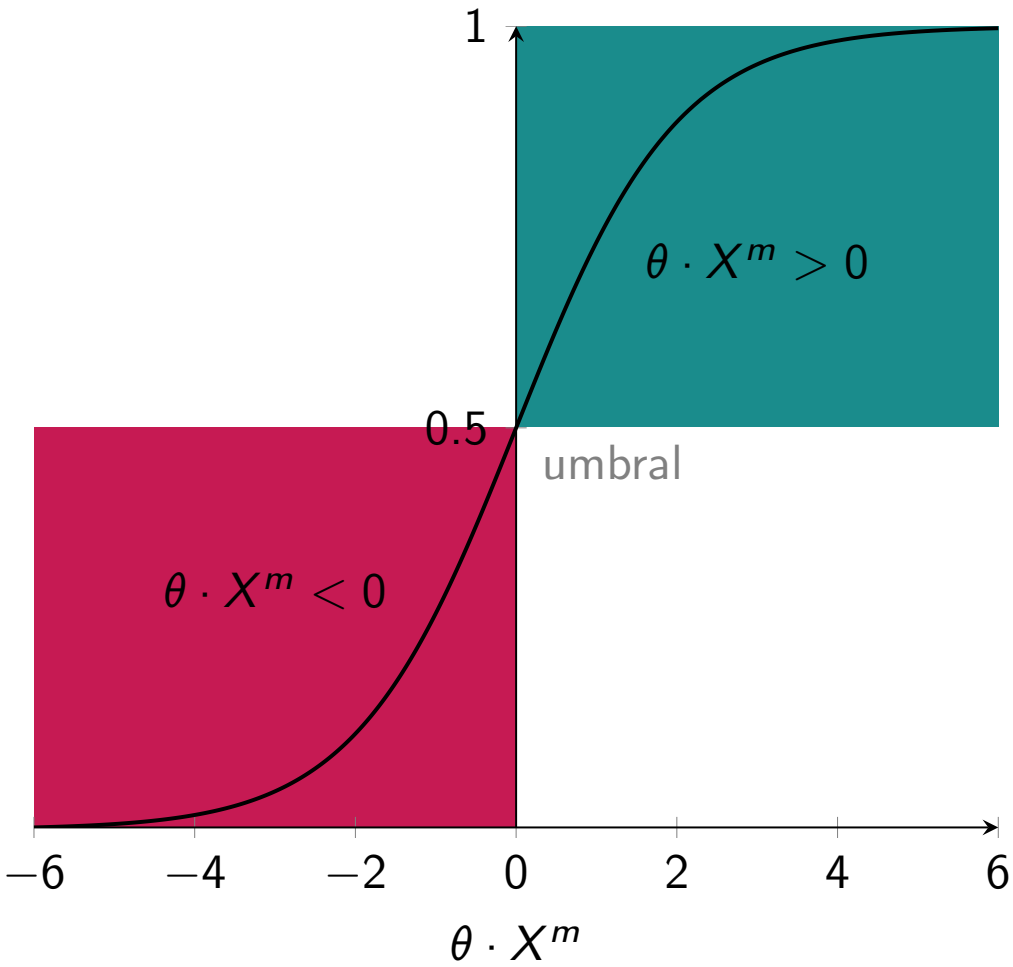
Regresión logística: función logística



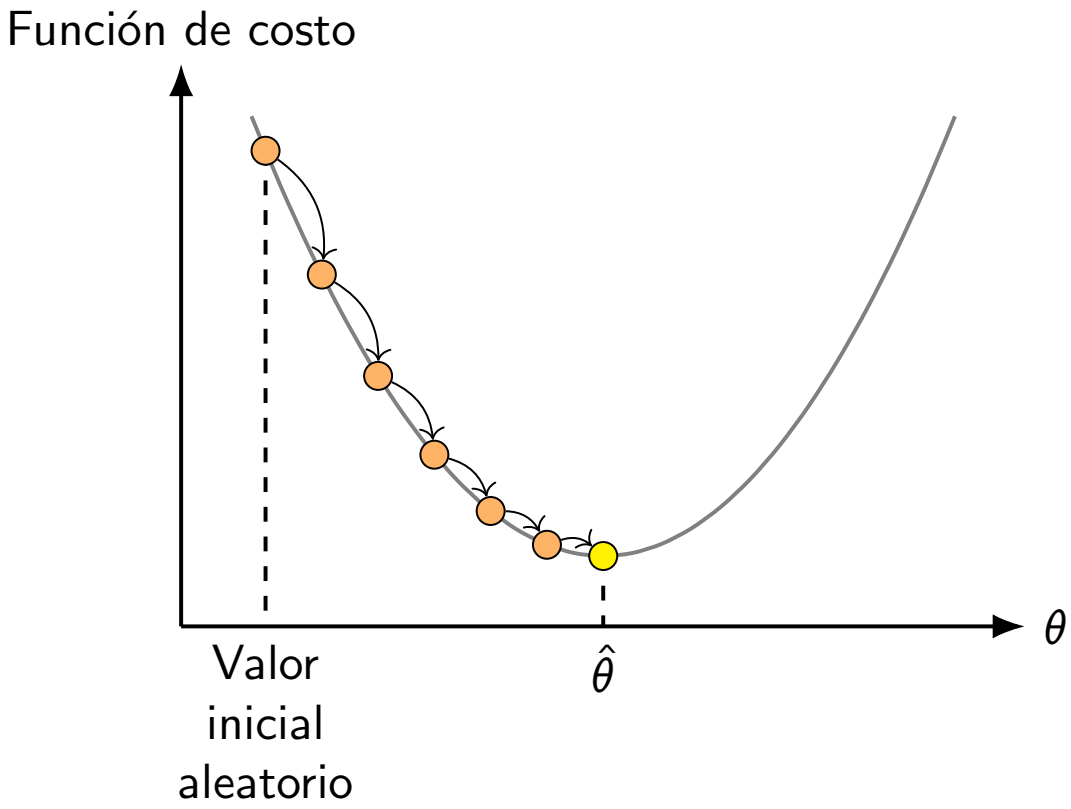
$$\sigma(t) = \frac{1}{1+e^{-t}}$$

Regresión logística: umbral

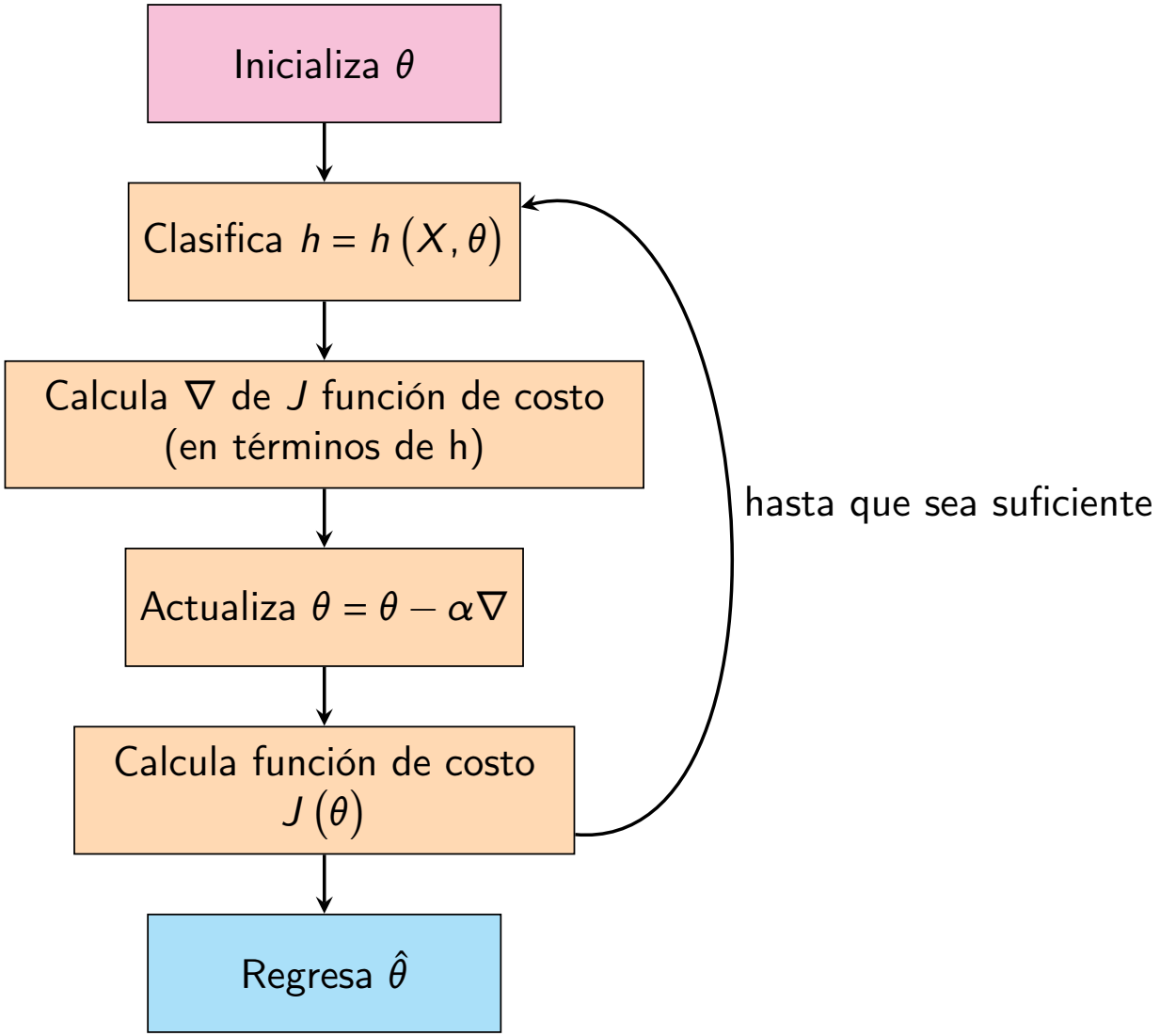
¿Cómo decir a qué clase pertenece un nuevo tuit?



Regresión logística: entrenamiento



Regresión logística: entrenamiento



Regresión logística: validación

- X_{val} una matriz de tuits codificados previamente clasificados.
- Y_{val} un vector con las correspondientes clasificaciones.
- para cada renglón X_{val}^m de la matriz X_{val} calcular:

$$\text{pred}(X_{\text{val}}^m) = \begin{cases} 0 & \text{si } h(X_{\text{val}}^m, \hat{\theta}) < 0.5; \\ 1 & \text{si } h(X_{\text{val}}^m, \hat{\theta}) \geq 0.5. \end{cases}$$

- se saca el porcentaje (*accuracy*) de casos para los cuales:

$$\text{pred}(X_{\text{val}}^m) = {Y_{\text{val}}}^m$$

Regresión logística: función de costo

Si hay n tuits de entrenamiento, $J(\theta) = \sum_{m=1}^{n} J_m(\theta)$ donde:

$$J_m(\theta) = Y^m \log h(X^m, \theta) + (1 - Y^m) \log(1 - h(X^m, \theta))$$