

H = conjunto de hipótesis

$$\hat{y} = h_{\theta}(x) = w_1 x_1 + \dots + w_n x_n + b$$

$$x = (x_1, \dots, x_n) \in \mathbb{R}^n$$

$$y \in \mathbb{R}$$

$$\theta = (w_1, \dots, w_n, b) \in \mathbb{R}^{n+1}$$

$$\begin{array}{c} \left[\begin{array}{cccc} x_1^{(1)} & \cdots & x_n^{(1)} \\ \vdots & & \vdots \\ x_1^{(m)} & \cdots & x_n^{(m)} \end{array} \right] \quad \left[\begin{array}{c} y^{(1)} \\ \vdots \\ y^{(m)} \end{array} \right] \\ \times \quad \quad \quad \quad Y \\ (m, n) \quad (m, 1) \end{array} \quad \theta = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \\ b \end{bmatrix} \quad (n+1, 1)$$

Queremos una θ que minimice el error MSE

$$\theta^* = \arg \min_{\theta \in \mathbb{R}^{n+1}} \frac{1}{2M} \sum_{i=1}^m (y^{(i)} - [x^{(i)} - 1] \begin{bmatrix} w \\ b \end{bmatrix})^2$$

$$X_e = \begin{bmatrix} x_1^{(1)} & \cdots & x_n^{(1)} & 1 \\ \vdots & & \vdots & \vdots \\ x_1^{(m)} & \cdots & x_n^{(m)} & 1 \end{bmatrix}$$

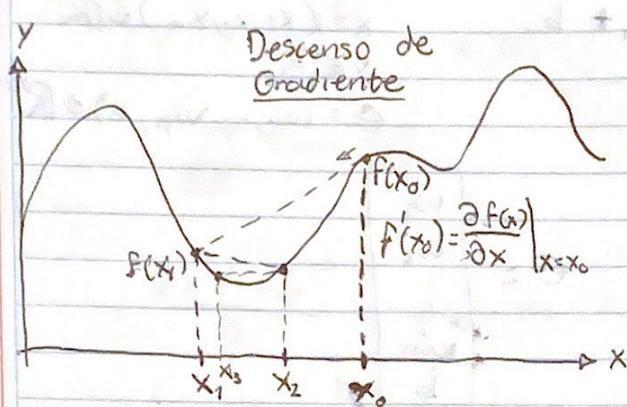
$$\hat{Y} = X_e \theta$$

$$E = Y - \hat{Y}$$

$$E_m(\theta) = \frac{1}{M} E^T E$$

$$\theta^* = \underbrace{\left(\begin{array}{cc} X_e^T & X_e \end{array} \right)^{-1}_{(n+1, m) (m, n+1)}}_{(n+1, n+1)} \underbrace{X_e^T Y}_{(n+1, 1)}$$

$f(x)$ $f : \mathbb{R} \rightarrow \mathbb{R}$



→ Forma genérica para encontrar mínimos.

η = Paso de aprendizaje
o learning rate

$$x_1 \leftarrow x_0 - \eta f'(x_0) \Rightarrow x_{k+1} \leftarrow x_k - \eta f'(x_k)$$

$$x_2 \leftarrow x_1 - \eta f'(x_1)$$

Cuando son varios parámetros

$$\theta_{k+1} \leftarrow \theta_k - \eta \nabla_{\theta} f(\theta) \Big|_{\theta=\theta_k}$$

$\nabla f(\theta_k)$

$$b_{k+1} \leftarrow b_k - \eta \frac{\partial E_{in}(w_k, b_k)}{\partial b}$$

$$w_{j, k+1} \leftarrow w_{j, k} - \eta \frac{\partial E_{in}(w_k, b_k)}{\partial w_j}$$

Recordatorio:

$$\theta_{k+1} \leftarrow \theta_k - \eta \nabla J(\theta_k)$$

$$h_\theta(x) = w^T x + b = [x^T, 1] \begin{bmatrix} w \\ b \end{bmatrix} = [x^T, 1] \theta$$

Queremos minimizar $\rightarrow \theta^* = w^*, b^* = \arg \min_{\theta \in \mathbb{R}^{n+1}} \frac{1}{2M} \sum_{i=1}^M (y^{(i)} - [x^T, 1] \theta)^2$

$$\frac{\partial E_{in}(\theta)}{\partial \theta_j} = \frac{1}{2M} \sum_{i=1}^M -2(y^{(i)} - \hat{y}^{(i)}) x_j^{(i)}$$

$$\frac{\partial E_{in}(\theta)}{\partial b} = \frac{1}{2M} \sum_{i=1}^M -2(y^{(i)} - \hat{y}^{(i)})$$

$$\nabla J(\theta_k) = \begin{bmatrix} \sum_{i=1}^M (y^{(i)} - \hat{y}^{(i)}) x_1^{(i)} \\ \vdots \\ \sum_{i=1}^M (y^{(i)} - \hat{y}^{(i)}) x_n^{(i)} \\ \sum_{i=1}^M (y^{(i)} - \hat{y}^{(i)}) \end{bmatrix}$$

$$\triangleright = -\frac{1}{M} \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(m)} \\ x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(m)} \\ \vdots & \vdots & \ddots & \vdots \\ x_n^{(1)} & x_n^{(2)} & \dots & x_n^{(m)} \\ 1 & 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} y^{(1)} - \hat{y}^{(1)} \\ y^{(2)} - \hat{y}^{(2)} \\ \vdots \\ y^{(m)} - \hat{y}^{(m)} \end{bmatrix}$$

E_{in} = función de costo

$$\theta \leftarrow \theta - \eta \nabla E_{in}(\theta)$$

$$\theta \leftarrow \theta + \frac{n}{M} X_e^T (Y - X_e \theta)$$

Programa de python: (de lo que acabamos de ver)

```
def dg_lin(X, Y, w[], b[], lr, max-epochs, e-tol):
    M = X.shape[0]
    w = w[] . copy()
    b = b[] . copy()
    hist = []
    for _ in range(max-epochs):
        Y_est = X @ w + b
        Err = Y - Y_est
        hist.append(np.square(Err).mean())
    # Como llamar la función
    X, Y
    X.shape = [m, n], Y.shape = [m, ]
    w = np.zeros(X.shape[-1])
    b = 0
    w-n, b-n, hist = dg_lin(X, Y, w, b, 0.1, 50, 1e-4)
    grad_w = -(1/M) X.T @ Err
    d_b = Err.mean()
    w -= lr * grad_w
    b -= lr * d_b
    if np.abs(grad_w).max() < e-tol:
        break
    return w, b, hist
```



Descenso de gradiente en problemas lineales:

$$\hat{Y} = Xw + b$$

$$Err = Y - \hat{Y}$$

$$E_{in} = \frac{1}{M} Err^T Err$$

$$\nabla_w E_{in} = -\frac{1}{M} X^T Err$$

$$\frac{\partial E_{in}}{\partial w} = -\frac{1}{M} \sum_{i=1}^M Err^{(i)}$$

$$w \leftarrow w - lr \nabla_w E_{in}$$

$$b \leftarrow b - lr \frac{\partial E_{in}}{\partial b}$$

$$\theta^* = \arg \min_{\theta} E_{in}(\theta)$$

$$f \approx h^* \rightarrow E_{in} \approx 0$$

$$\begin{array}{c} \uparrow \\ f \approx h^* \end{array} \rightarrow E_{in} \approx E_{out}$$

$$E_{out} \approx 0$$

Probablemente Aproximadamente Correcto
 $\Pr(|E_{out} - E_{in}| \geq \epsilon) \geq S$

Cuando se cumple

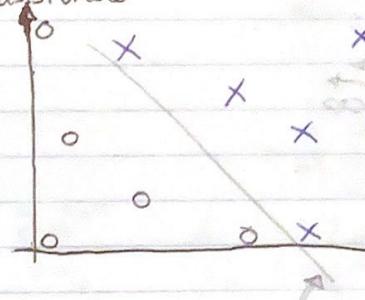
$$E_{in} \approx E_{out} \checkmark$$

pero no cumple

$$E_{in} \approx 0$$

Entonces es un modelo de alto sesgo.

asistencia



Mi hipótesis es...

$$h_{\theta} = \text{sign}(w_1 x_1 + w_2 x_2 + b)$$

$$x \in \mathbb{R}^n$$

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

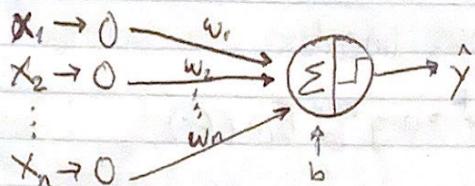
$$H = \{h_{\theta} \mid h_{\theta} = \text{sign}(w^T x + b), w \in \mathbb{R}^n, b \in \mathbb{R}, \theta = (w, b)\}$$

$$w_1 x_1 + w_2 x_2 + b = 0$$

$$\text{Loss}(y, \hat{y}) = \begin{cases} 1 & \text{si } y \neq \hat{y} \\ 0 & \text{en este caso} \end{cases} = \max(-y \hat{y}, 0) \quad 0/1 - \text{loss}$$

$$E_{in}(w, b) = \frac{1}{M} \sum_{i=1}^M \text{loss}(y^{(i)}, \text{sign}(w^T x^{(i)} + b))$$

A



Perceptron

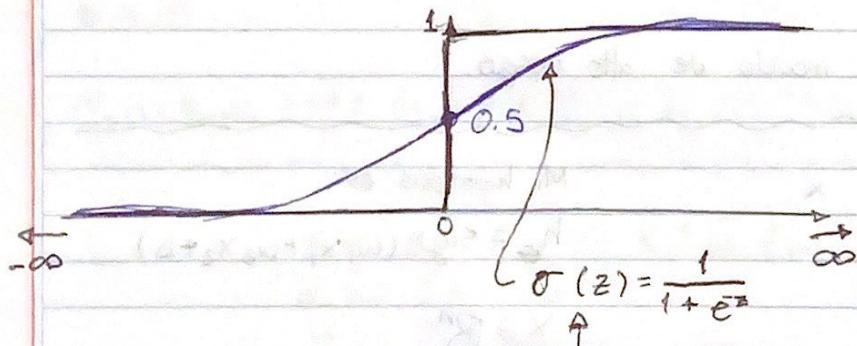
Para pasar de discreto a continuo:

$$a = \Pr^{\text{Probabilidad}}(Y=1 | X=x; \theta)$$

$$\hat{Y} = \begin{cases} 1 & \text{si } a > h \\ -1 & \text{en otro caso} \end{cases}$$

$$a = f(w_1x_1 + w_2x_2 + \dots + w_nx_n + b)$$

$$z = w_1x_1 + w_2x_2 + \dots + w_nx_n + b = x^T w + b = x^T e^\theta$$



Sigmoid, logística

A esto se le llama Regresión Logística.

Para utilizar la Regresión Logística, tenemos que saber como aprender.

$$\begin{matrix} \begin{bmatrix} x_1^{(1)} & \cdots & x_n^{(1)} \\ x_1^{(2)} & \cdots & x_n^{(2)} \\ \vdots & & \vdots \\ x_1^{(n)} & \cdots & x_n^{(n)} \end{bmatrix} & \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix} & \begin{bmatrix} a^{(1)} \\ a^{(2)} \\ \vdots \\ a^{(n)} \end{bmatrix} \end{matrix}$$

$\Rightarrow X \quad Y \quad A$

$y^{(i)} \in \{-1, 1\}$

Estos
son
mis
ejemplos

$$E_{in}(w, b) = \frac{1}{M} \sum_{i=1}^M \text{loss}(a^{(i)}, \hat{a}^{(i)})$$

Un error cuadrático podría funcionar pero eso ~~no~~ es lo que estamos buscando.

$$\text{loss}(a^{(i)}, \hat{a}^{(i)}) = \begin{cases} -\log(\hat{a}^{(i)}) & \text{si } a^{(i)} = 1 \\ -\log(1-\hat{a}^{(i)}) & \text{si } a^{(i)} = 0 \end{cases}$$

$$\text{loss}(a^{(i)}, \hat{a}^{(i)}) = -a^{(i)} \log(\hat{a}^{(i)}) - (1-a^{(i)}) \log(1-\hat{a}^{(i)})$$

Le sacamos el gradiente a lo anterior.

$$w \leftarrow w - lr \Delta_w E_{in}(w, b)$$

$$b \leftarrow b - lr \frac{\partial}{\partial b} E_{in}(w, b)$$

$$\frac{\partial}{\partial w_j} E_{in}(w, b) = \frac{\partial}{\partial w_j} \frac{1}{M} \sum_{i=1}^M -a^{(i)} \log(\hat{a}^{(i)}) - (1-a^{(i)}) \log(1-\hat{a}^{(i)})$$

donde

$$\hat{a}^{(i)} = \frac{1}{1+e^{-z^{(i)}}}, \quad z^{(i)} = w_1 x_1^{(i)} + \dots + w_n x_n^{(i)} + b$$

$$= \frac{1}{M} \sum_{i=1}^M -\frac{a^{(i)}}{\hat{a}^{(i)}} \frac{\partial \hat{a}^{(i)}}{\partial w_j} + \frac{1-a^{(i)}}{1-\hat{a}^{(i)}} + \frac{\partial \hat{a}^{(i)}}{\partial w_j}$$

$$\frac{\partial \hat{a}^{(i)}}{\partial w_j} = \frac{\partial}{\partial w_j} \frac{1}{1+e^{-z^{(i)}}}$$

$$= \frac{\partial}{\partial z^{(i)}} (1+e^{-z^{(i)}})^{-1} \frac{\partial z^{(i)}}{\partial w_j}$$

$$= (1+e^{-z^{(i)}})^{-2} e^{-z^{(i)}} x_j^{(i)}$$

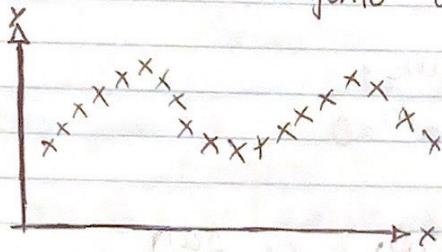
$$= \left(\frac{1}{(1+e^{-z^{(i)}})^2} - \frac{1}{(1+e^{-z^{(i)}})^2} \right) x_j^{(i)}$$

$$= \left(\frac{1}{1+e^{-z^{(i)}}} - \frac{1}{1+e^{-z^{(i)}}} \right) x_j^{(i)} = (\hat{a}^{(i)} - \hat{a}^{(i)2}) x_j^{(i)}$$

A

$$= \hat{\alpha}^{(i)} (1 - \hat{\alpha}^{(i)}) x_j^{(i)}$$

Si tenemos un conjunto de datos, y queremos aplicar regresión



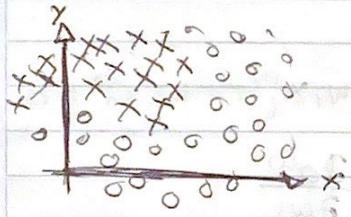
Creamos una función:

$$\phi: \mathbb{R}^n \rightarrow \mathbb{R}^m$$
$$x \in \mathbb{R}$$

$$x' = \phi(x) = (x, x^2, x^3, x^4)$$

$$x' = \phi(x) = (\phi_1(x), \phi_2(x), \dots, \phi_n(x))$$

Ahora si tenemos algo como lo siguiente



Entonces vamos a querer sacar la expansión polinomial.

$$x = (x_1, x_2)$$

$$\phi(x) = (x_1, x_2, x_1^2, x_1 x_2, x_2^2)$$

$$\phi(x) = (x_1, x_2, x_1^2, x_1 x_2, x_2^2, x_1^3, x_1^2 x_2, x_1 x_2^2, x_2^3)$$

$$x = (x_1, x_2, x_3, x_4)$$

$$\phi(x) = (x_1, x_2, x_3, x_4, x_1^2, x_1 x_2,$$

La dimensión b_c se saca contando todos los parámetros libres.

$$h(x) = w^T \phi(x) + b \quad w, \phi(x) \in \mathbb{R}^n, \quad K \in \mathbb{R}$$

Aquí la dimensión b_c es igual a $n+1$ ya que la b también cuenta como parámetro libre, junto con las w 's.

$$\|w\|_2 = \sqrt{w^T w} \rightarrow \text{Es la norma } L_2$$

$$\|w\|_1 = \sum_{j=1}^n |w_j| \rightarrow \text{Es la norma } L_1$$

$$h(x) = w^T \phi(x) + b \quad w, \phi(x) \in \mathbb{R}^n, \quad K \in \mathbb{R}$$

bajo $\underbrace{\sum_{j=1}^n w_j^2 \leq C}_{w^T w \leq C}$ \rightarrow Regularización

$$w^T w \leq C$$

$$[w^T \phi + (d-w) \cdot n] \cdot \phi = \phi^T \phi + d \cdot \phi - w \cdot \phi$$

$$w^T \phi + (d-w) \cdot n = [\sum_{j=1}^n \phi_j + (d-w) \cdot n]$$

$$[(w-\phi)^T \phi + (d-w) \cdot n] \cdot \phi = d \cdot \phi - w \cdot \phi$$

A

$$\frac{\partial E_{in}(w, b)}{\partial w_j} = \frac{1}{M} \sum_{i=1}^M \left(\frac{a^{(i)}}{\hat{a}^{(i)}} - \frac{1-a^{(i)}}{1-\hat{a}^{(i)}} \right) \hat{a}^{(i)} (1-\hat{a}^{(i)}) x_j^{(i)}$$

$$\begin{aligned}\frac{\partial E_{in}(w, b)}{\partial w_j} &= \frac{\partial}{\partial w_j} \frac{1}{M} \sum_{i=1}^M \left[-a^{(i)} \ln(\hat{a}^{(i)}) - (1-a^{(i)}) \ln(1-\hat{a}^{(i)}) \right] \\ &= \frac{1}{M} \sum_{i=1}^M -\frac{a^{(i)}}{\hat{a}^{(i)}} \frac{\partial \hat{a}^{(i)}}{\partial w_j} + \frac{1-a^{(i)}}{1-\hat{a}^{(i)}} \frac{\partial \hat{a}^{(i)}}{\partial w_j} \\ &= \frac{1}{M} \sum_{i=1}^M \left[-a^{(i)} (1-\hat{a}^{(i)}) + (1-a^{(i)}) \hat{a}^{(i)} \right] x_j^{(i)} \\ &= \frac{1}{M} \sum_{i=1}^M \left[-a^{(i)} + a^{(i)} \cancel{\hat{a}^{(i)}} + \hat{a}^{(i)} - a^{(i)} \cancel{\hat{a}^{(i)}} \right] x_j^{(i)} \\ \therefore \frac{\partial E_{in}(w, b)}{\partial w_j} &= \frac{1}{M} \sum_{i=1}^M (a^{(i)} - \hat{a}^{(i)}) x_j^{(i)}\end{aligned}$$

$$\begin{aligned}w &\leftarrow w - lr \quad X^T (A \cdot \hat{A}) \\ b &\leftarrow b - \frac{lr}{M} \sum_{i=1}^M (a^{(i)} - \hat{a}^{(i)})\end{aligned}$$

$$w^*, b^* = \arg \min_{w, b} E_{in}(w, b)$$

bajo

$$\sum_{j=1}^M w_j^2 \leq C$$

regularizado

$$w_{rg}^*, b_{rg}^* = \arg \min \left[E_{in}(w, b) + \frac{\lambda}{M} \sum_{j=1}^n w_j^2 \right]$$

$$\frac{\partial}{\partial w_j} \left[E_{in}(w, b) + \frac{\lambda}{M} \sum_{j=1}^n w_j^2 \right] = -\frac{1}{M} \sum_{i=1}^n (a^{(i)} - \hat{a}^{(i)}) x_j^{(i)} + \frac{2\lambda}{M} w_j$$

$$w_{rg}^*, b_{rg}^* = \arg \min \left[E_{in}(w, b) + \frac{\lambda}{M} \text{regw}(w) \right]$$

$$\text{reg}_U(w) = \begin{cases} \sum_{j=1}^n w_j^2 = \|w\|_e^2 \\ \sum_{j=1}^n |w_j| \end{cases}$$

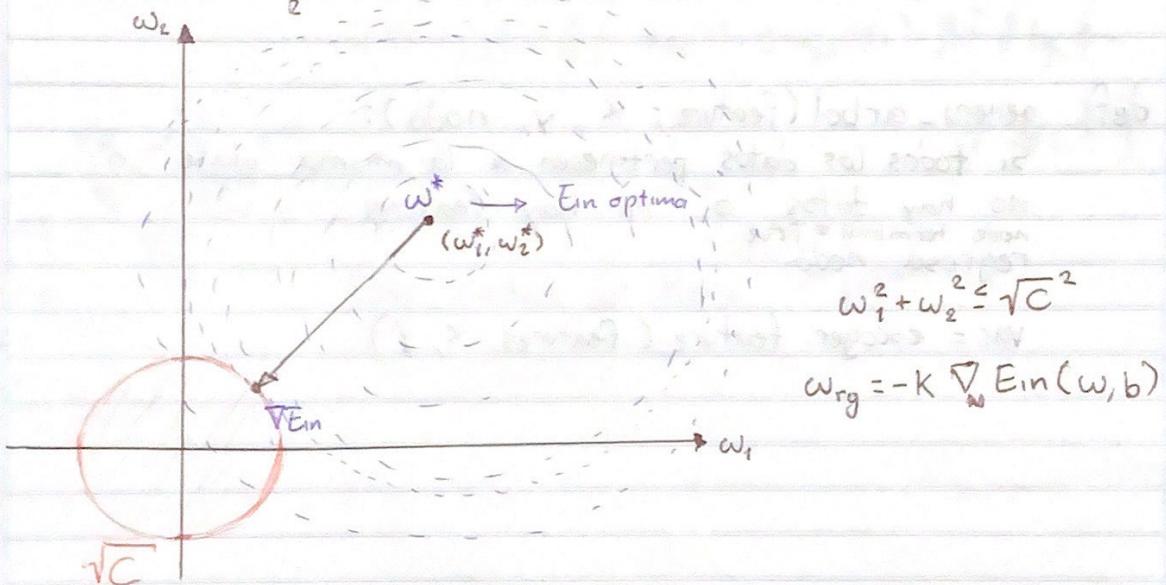
$$w^*, b^* = \arg \min_{w, b} E_{in}(w, b)$$

bajo

$$\sum_{j=1}^n w_j^2 \leq C$$

$$w^T w \leq C$$

$$\|w\|_2^2 \leq C$$



$$w_{rg} = -K \nabla E_{in}(w, b)$$