UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Licenciatura en Ciencia de Datos

Introducción al Aprendizaje Profundo

Redes densas poco profundas

Profesores:

Berenice & Ricardo Montalvo Lezama

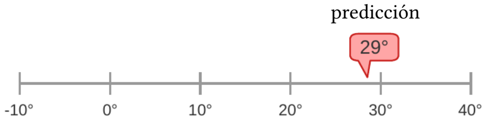
Marzo 2021

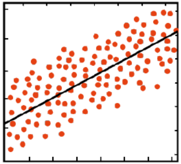
Contenido basado en el curso de AP del Dr. Gibran Fuentes Pineda del PCIC

Regresión y clasificación

Regresión: $y \in \mathbb{R}$

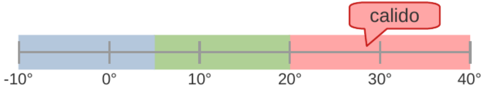
¿Cuál será la temperatura mañana?

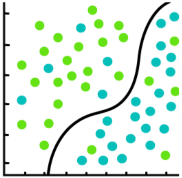




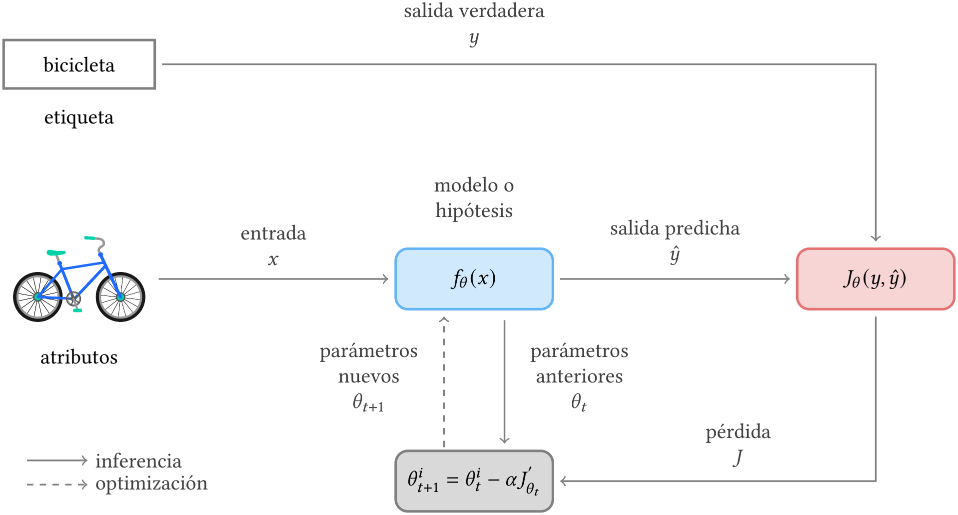
Clasificación: $y \in \{1, \ldots, k\}$

¿Cómo será el día de mañana?

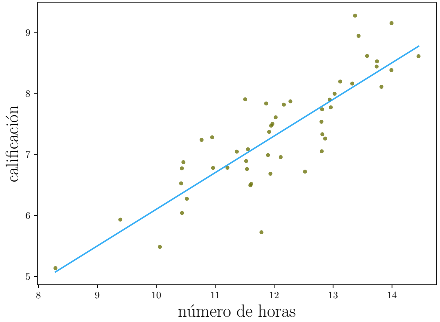


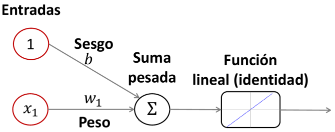


Aprendizaje supervisado



Regresión lineal simple: hipótesis

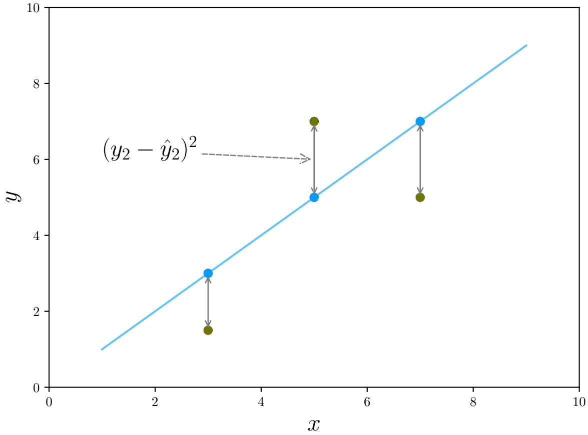




- Hipótesis:

$$\hat{y} = f_\theta(x) = xw + b \qquad \theta = \{w, b\}$$

Regresión lineal simple: pérdida



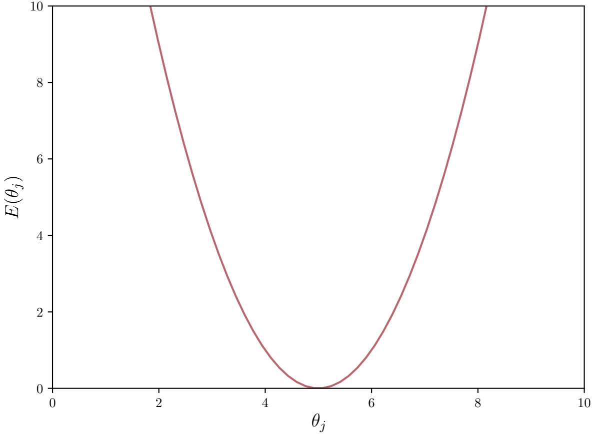
- Hipótesis:

$$\hat{y} = f_\theta(x) = xw + b \qquad \theta = \{w, b\}$$

- Función de error: error cuadrático medio

$$J_\theta = \frac{1}{2m} \sum_{i=1}^{n} (y - \hat{y})^2$$

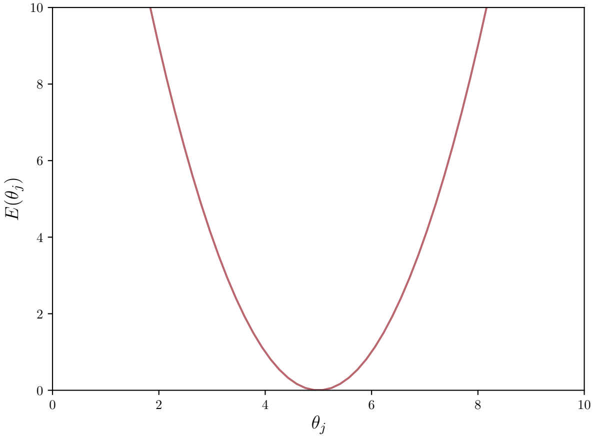
Descenso por gradiente



Repetir hasta converger:

$$\theta_{t+1} = \theta_t - \alpha \nabla J_{\theta_t}$$

Descenso por gradiente para regresión lineal



Repetir hasta converger:

$$w := w - \alpha \frac{\partial}{\partial w} J_\theta$$

$$b := b - \alpha \frac{\partial}{\partial b} J_\theta$$

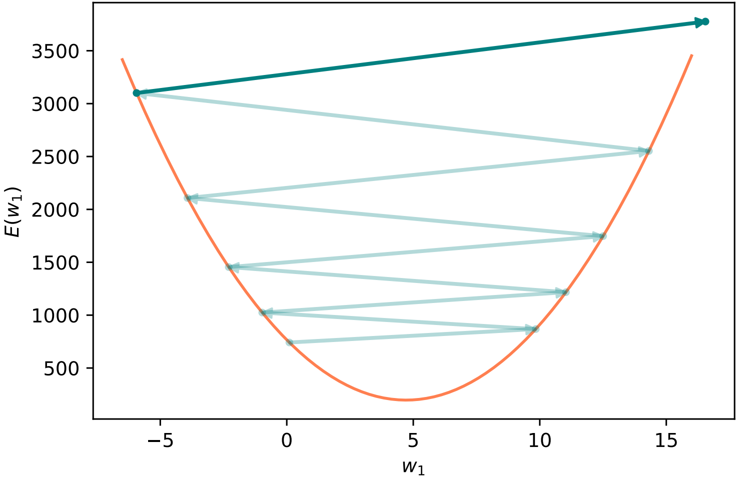
Cálculo de la derivadas:

$$\frac{\partial}{\partial w} J_\theta = \frac{1}{m} \sum_{i=1}^{n} (\hat{y}^{(i)} - y^{(i)}) \cdot x_j^{(i)}$$

$$\frac{\partial}{\partial b} J_\theta = \frac{1}{m} \sum_{i=1}^{n} (\hat{y}^{(i)} - y^{(i)})$$

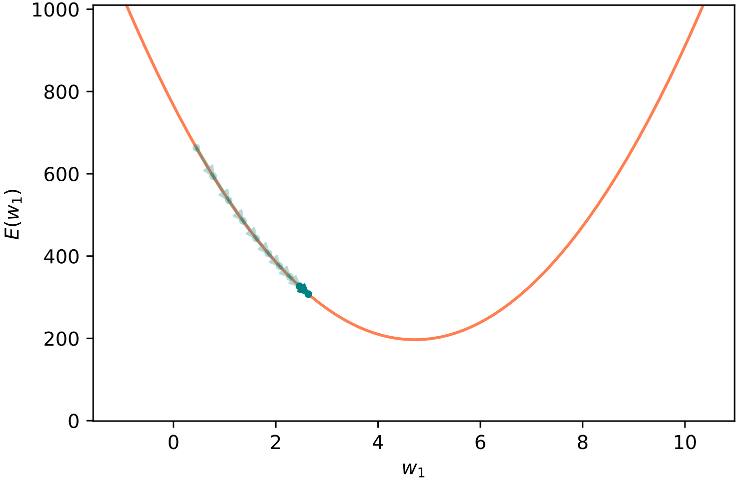
Hiperparámetro: tasa de aprendizaje

- Tasa de aprendizaje muy grande.



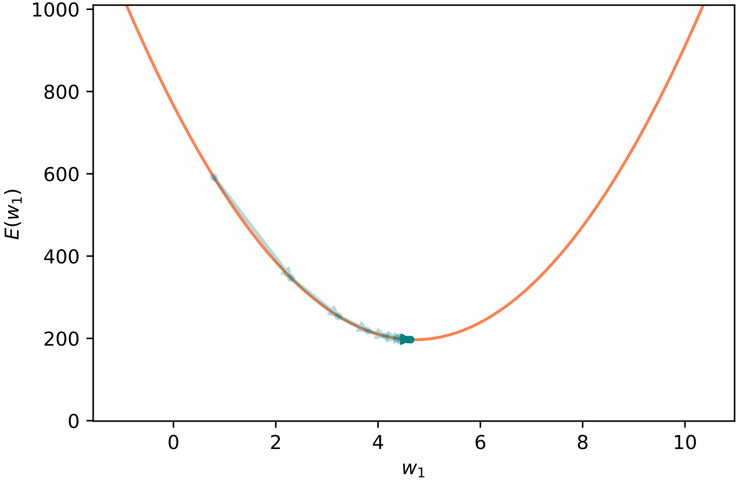
Hiperparámetro: tasa de aprendizaje

- Tasa de aprendizaje muy pequeña.

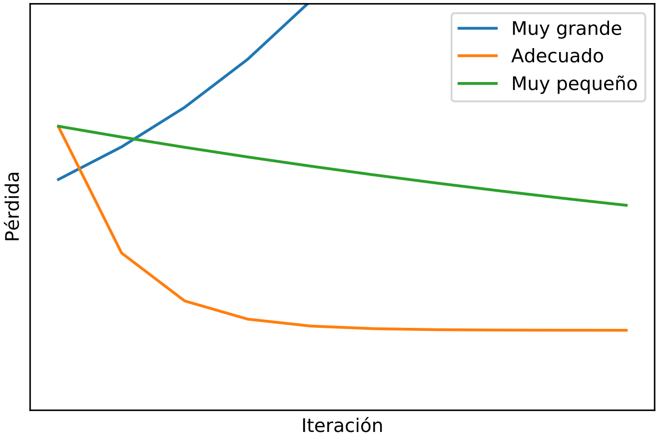


Hiperparámetro: tasa de aprendizaje

- Tasa de aprendizaje adecuada.

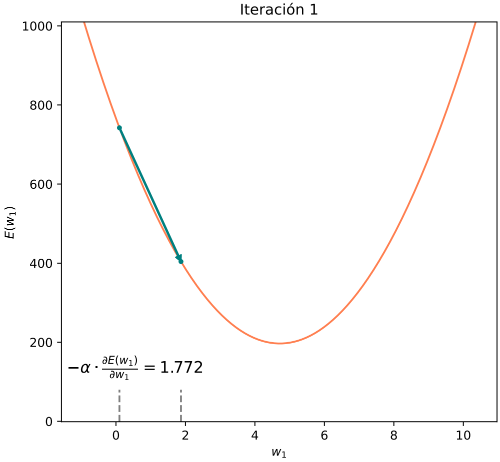


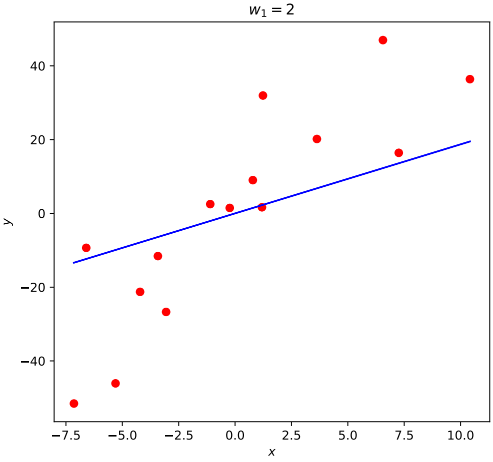
Sensibilidad a tasa de aprendizaje α



Ejemplo I del descenso por gradiente para regresión lineal

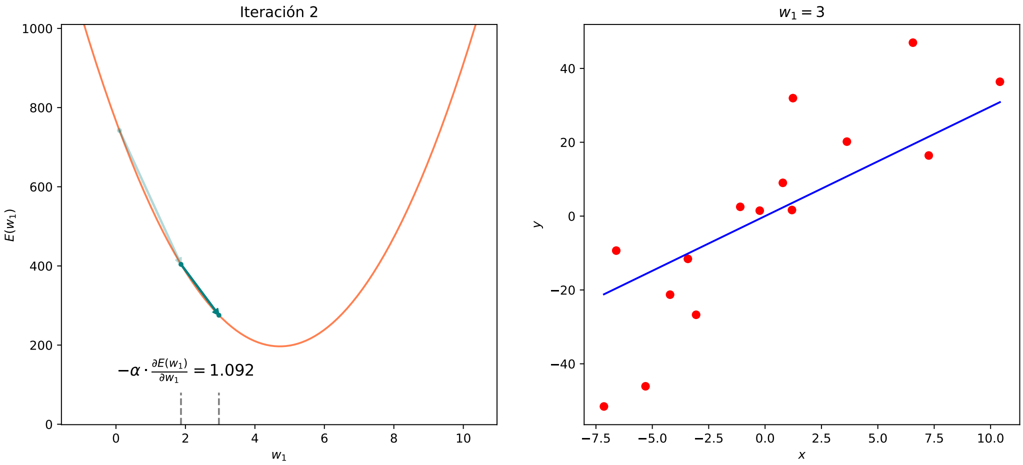
- Inicializando w_1 con un valor mayor al que minimiza la función de pérdida





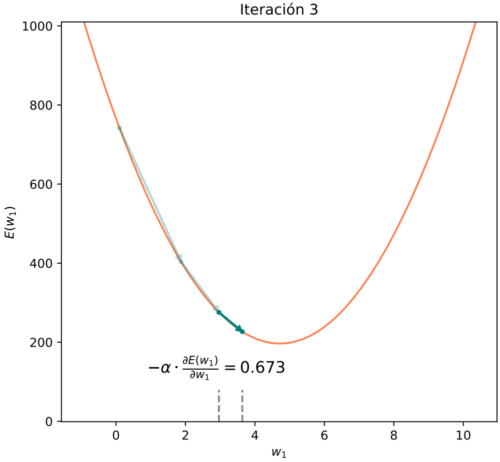
Ejemplo I del descenso por gradiente para regresión lineal

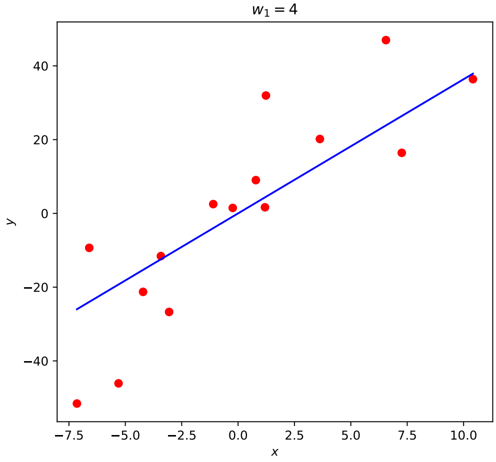
- Inicializando w_1 con un valor menor al que minimiza la función de pérdida



Ejemplo I del descenso por gradiente para regresión lineal

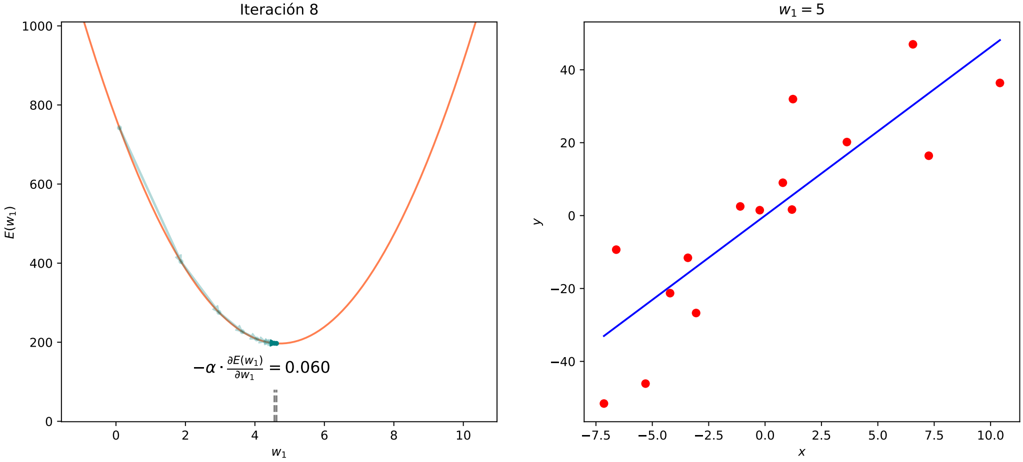
- Inicializando w_1 con un valor menor al que minimiza la función de pérdida





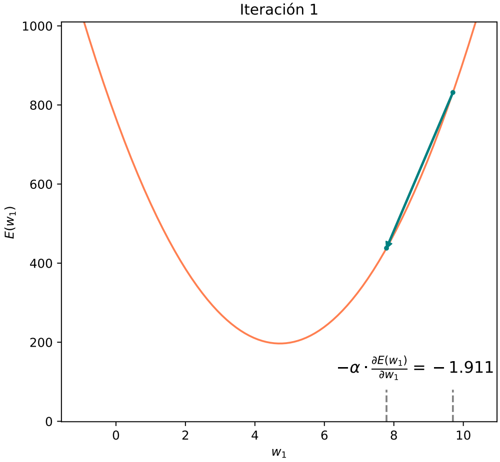
Ejemplo I del descenso por gradiente para regresión lineal

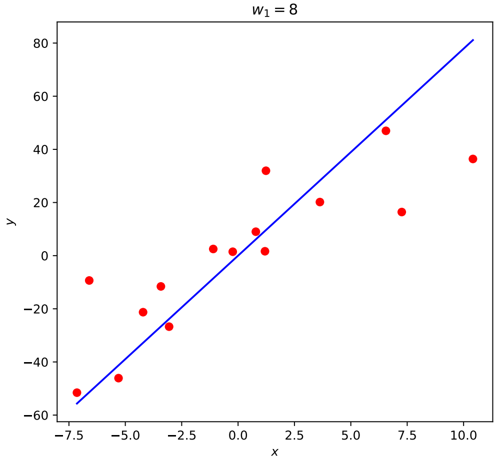
- Inicializando w_1 con un valor menor al que minimiza la función de pérdida



Ejemplo II del descenso por gradiente para regresión lineal

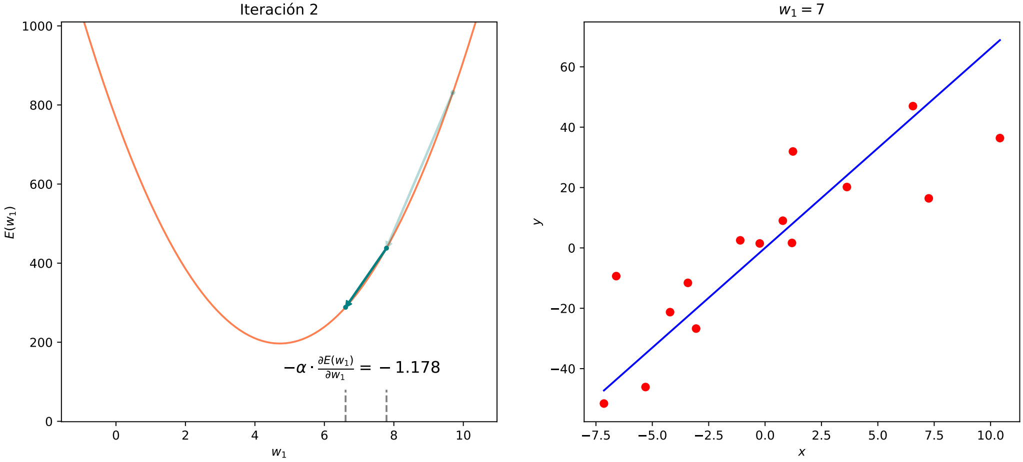
- Inicializando w_1 con un valor mayor al que minimiza la función de pérdida





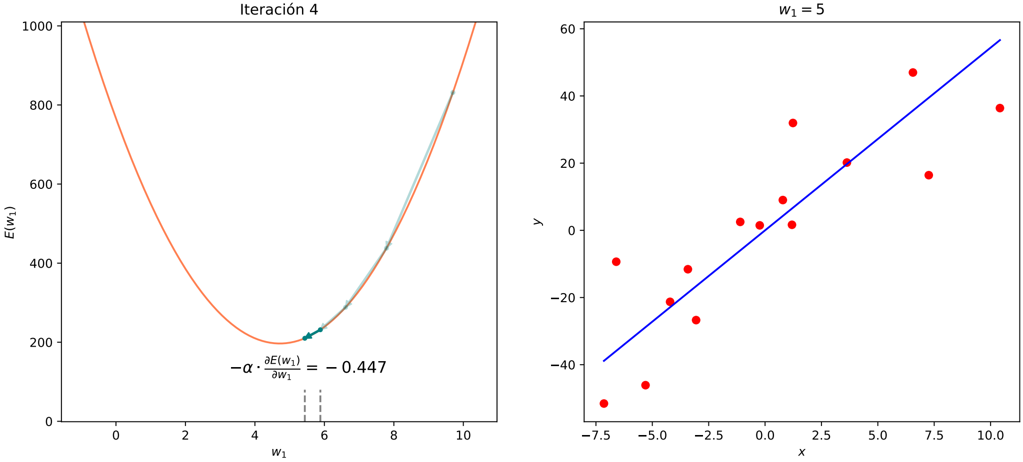
Ejemplo II del descenso por gradiente para regresión lineal

- Inicializando w_1 con un valor mayor al que minimiza la función de pérdida



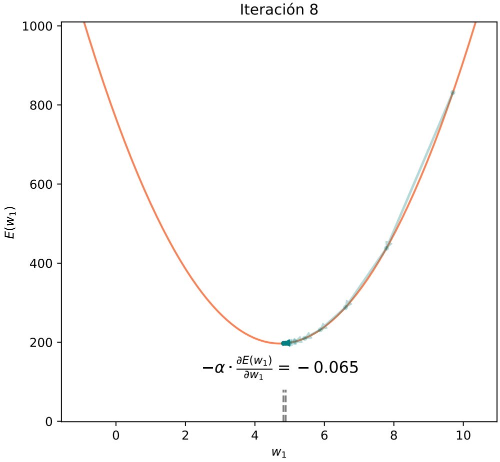
Ejemplo II del descenso por gradiente para regresión lineal

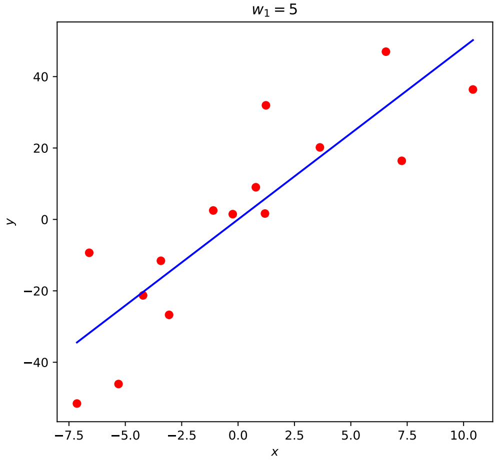
- Inicializando w_1 con un valor mayor al que minimiza la función de pérdida



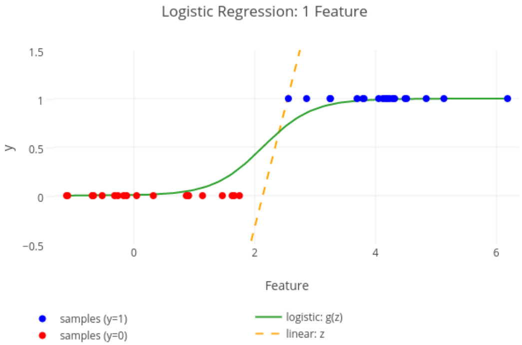
Ejemplo II del descenso por gradiente para regresión lineal

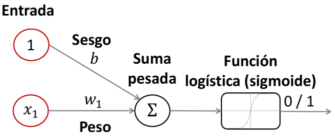
- Inicializando w_1 con un valor mayor al que minimiza la función de pérdida





Regresión logística: hipótesis

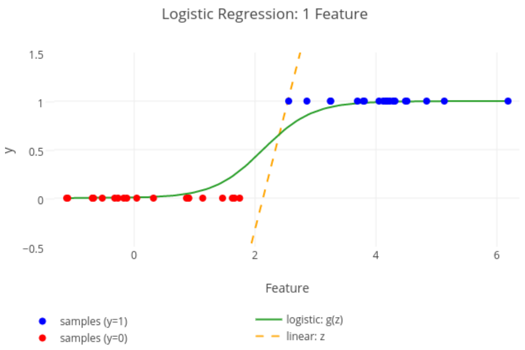


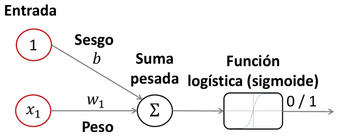


- Hipótesis:

$$P(y|x_i,\theta) = \hat{y} = \sigma(xw+b) = \frac{1}{1+e^{-(xw+b)}}$$

Regresión logística: pérdida





- Hipótesis:

$$P(y|x_i, \theta) = \hat{y} = \sigma(xw + b) = \frac{1}{1 + e^{-(xw+b)}}$$

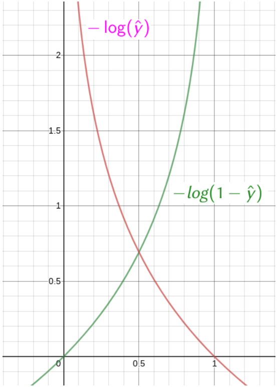
- Función de error: entropía cruzada binaria.

$$J_\theta = -\frac{1}{m}\sum_{i=1}^{m}(y\log(\hat{y}) + (1-y)log(1-\hat{y}))$$

Pérdida: entropía cruzada binaria

$$J_\theta = (y)(-\log(\hat{y})) + (1-y)(-log(1-\hat{y}))$$

Etiqueta y	Predicción $\hat{y}$	Entropía binaria	Pérdida
0	0.9	2.303	alta
0	0.1	0.105	baja
1	0.9	0.105	baja
1	0.1	2.303	alta



Descenso por gradiente para regresión logística

```
Repetir hasta converger:
```

$$w := w - \alpha \frac{\partial}{\partial w} J_\theta$$

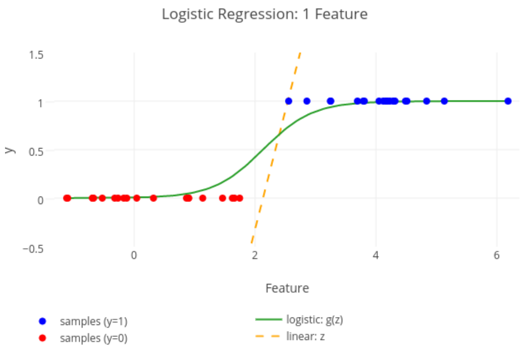
$$b := b - \alpha \frac{\partial}{\partial b} J_\theta$$

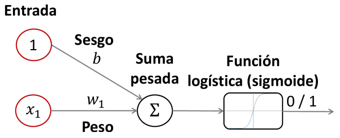
Cálculo de la derivadas:

$$\frac{\partial}{\partial w} J_\theta = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}^{(i)} - y^{(i)}) \cdot x_j^{(i)}$$

$$\frac{\partial}{\partial b} J_\theta = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}^{(i)} - y^{(i)})$$

Regresión logística: métrica





- Modelo:

$$P(y|x_i,\theta) = \hat{y} = \sigma(xw+b) = \frac{1}{1+e^{-(xw+b)}}$$

- Función de error: entropía cruzada binaria.

$$J_\theta = -\frac{1}{m}\sum_{i=1}^{m}(y\log(\hat{y}) + (1-y)log(1-\hat{y}))$$

- Optimización de la función de error:

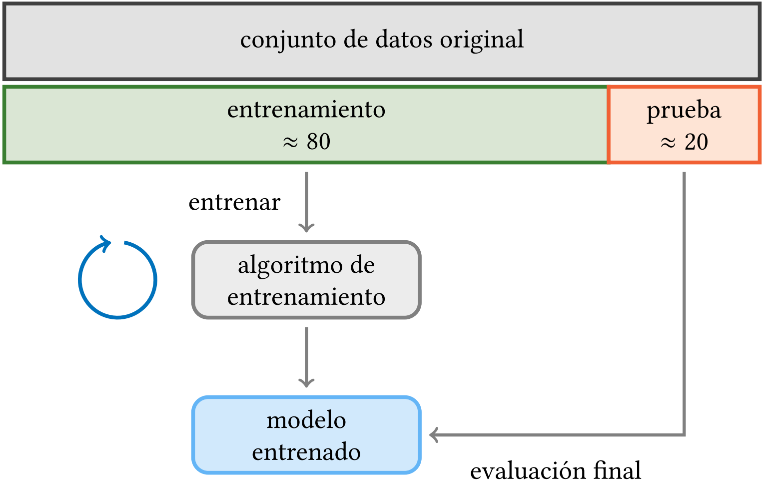
$$w := w - \alpha\frac{\partial}{\partial w}J_\theta$$

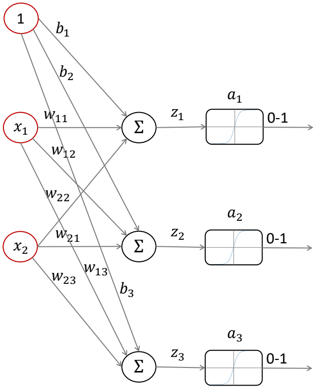
$$b := b - \alpha\frac{\partial}{\partial b}J_\theta$$

- Métrica: exactitud.

$$Ex = \frac{\#\text{ predicciones correctas}}{\#\text{ total de predicciones}}$$

Partición de datos





- Función de pérdida: entropía cruzada binaria de cada categoría

$$J_\theta(y_k, \hat{y}_k) = -\sum_{i=1}^{N} \left[y_k^{(i)} \log \hat{y}_k^{(i)} + (1 - y_k^{(i)}) \log (1 - \hat{y}_k^{(i)}) \right]$$

- Neuronas de la capa de salida tienen una función de activación *softmax* compartida, dada por

$$softmax(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}, i = 1, \ldots, K$$