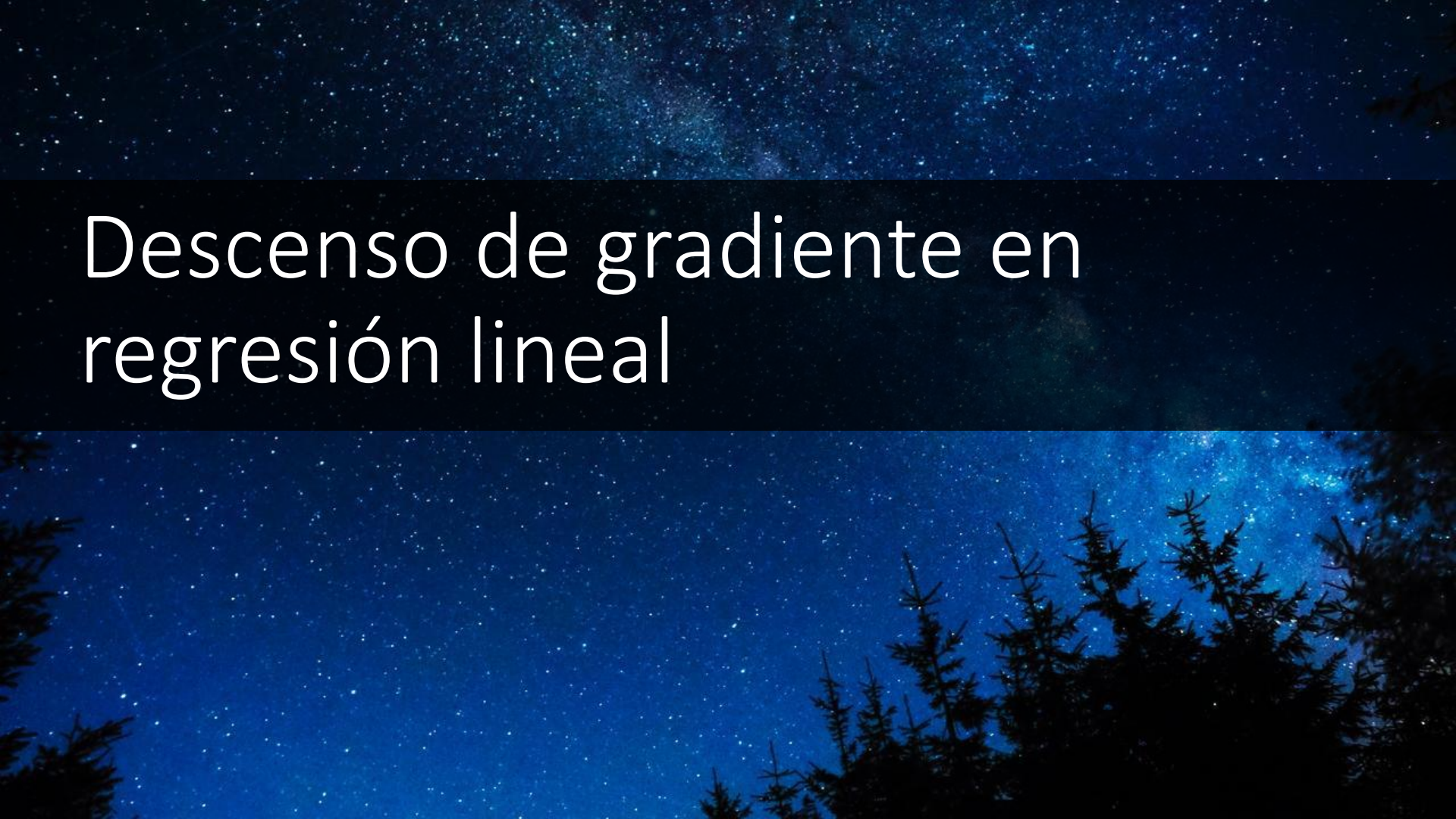
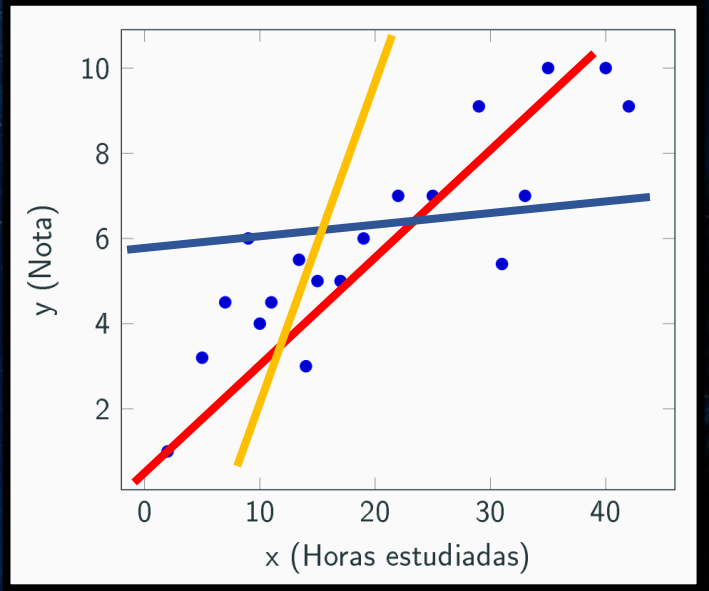
Descenso de gradiente en regresión lineal



Optimización para Regresión Lineal

Problema:

- Datos: ejemplos (x_i, y_i)
- Parámetros: m y b
- Función de error: E (cuadr. Medio)
  - Encontrar m y b que minimicen E(m,b,x,y)
  - E es derivable respecto de m y b



Optimización para Regresión Lineal

Métodos clásicos (analíticos)

- Cálculo: $\frac{\partial E}{\partial b} = 0$ y $\frac{\partial E}{\partial m} = 0$, despejo m y b.
- Álgebra lineal: $Y = mX + b = \begin{pmatrix} 1 & x \end{pmatrix} \begin{pmatrix} b \\ m \end{pmatrix}$, proyecto.
- Probabilidades: $y = mx + b + e$ con $e \sim \mathcal{N}(0, \sigma)$, estimo m y b con MLE

Ventajas

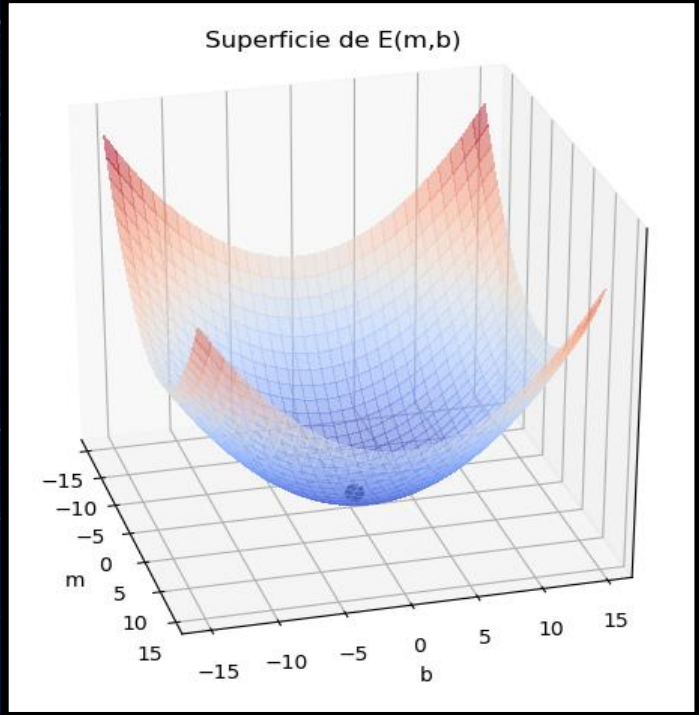
- Solución analítica, simple.

Desventajas

- Poco eficiente con muchos datos
- Problemas numéricos

Alternativa

- Descenso de gradiente



Desc. Gradiente para Regresión Lineal

Nuestro modelo de Regresión Lineal:

$$y = mx + b$$

Modelo lineal

$$E = \frac{1}{n}\sum_{i}^{n}(y'_i - y_i)^2$$

Error Cuadrático Medio

Entrenamiento:

$$w_i = w_i - \alpha \frac{\delta}{\delta w_i} E(w_1, w_2)$$

$$para\ i = \{1,2\}$$

Desc. Gradiente para Regresión Lineal

Entrenamiento:

- $m = m - \alpha \frac{\delta}{\delta \mathrm{m}} \mathrm{E}(\mathrm{m}, b)$
- $b = b - \alpha \frac{\delta}{\delta b} \mathrm{E}(\mathrm{m}, b)$

$$\frac{\delta E}{\delta m} E(m, b) = \frac{2}{n} \sum_{i}^{n} (y'_i - y_i) x_i$$

$$\frac{\delta E}{\delta b} E(m, b) = \frac{2}{n} \sum_{i}^{n} (y'_i - y_i)$$

Entrenamiento por Descenso de gradiente

Algoritmo:

- Comenzar con **m** y **b** aleatorios (o sensatos).
- Iterar hasta converger:
  - Calcular: $\boldsymbol{\Delta \mathbf{E}} = \left(\frac{\delta E}{\delta m}, \frac{\delta E}{\delta b}\right)$
  - Actualizar **m:** $m - \alpha \frac{\delta E}{\delta m}$
  - Actualizar **b:** $b - \alpha \frac{\delta E}{\delta b}$
  - Calcular el Error para los nuevos parámetros *m* y *b*

Cuestiones prácticas

¿Qué valores iniciales de m y b establecer?

- Afectan a la optimización.
- Aprovechar la experticia del dominio.

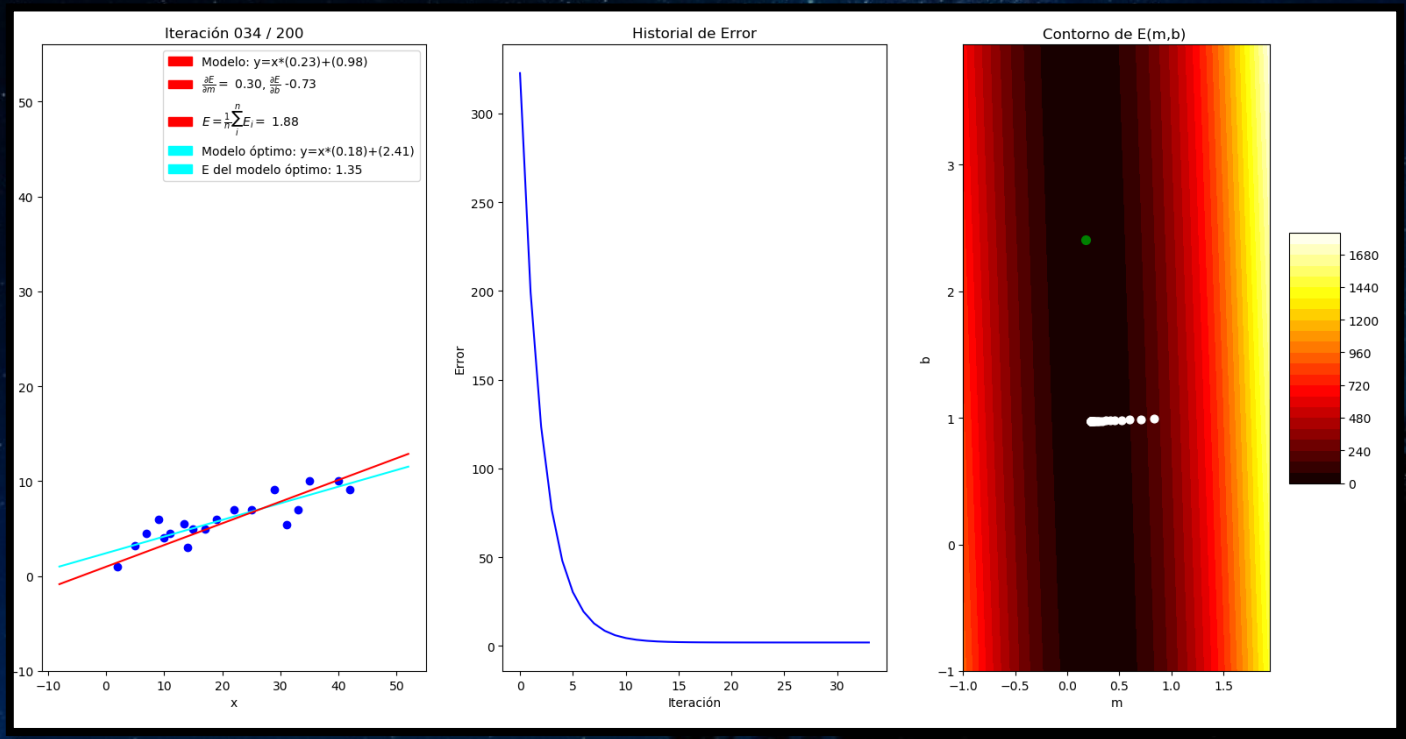
  En el ejemplo de las notas, ¿qué valores serían sensatos?

Normalización de variables

- Escala de notas de 0 al 10 vs 0 al 100. ¿Afecta al descenso?

Normalización

Variables sin normalizar. La curva de error tiene diferentes escalas para cada parámetro. Difícil encontrar el mínimo. Mucho tiempo.



Normalización

Original

Horas	Nota
2	1
5	3.2
7	4.5
9	6
10	4
11	4.5
13.4	5.5
14	3
15	5

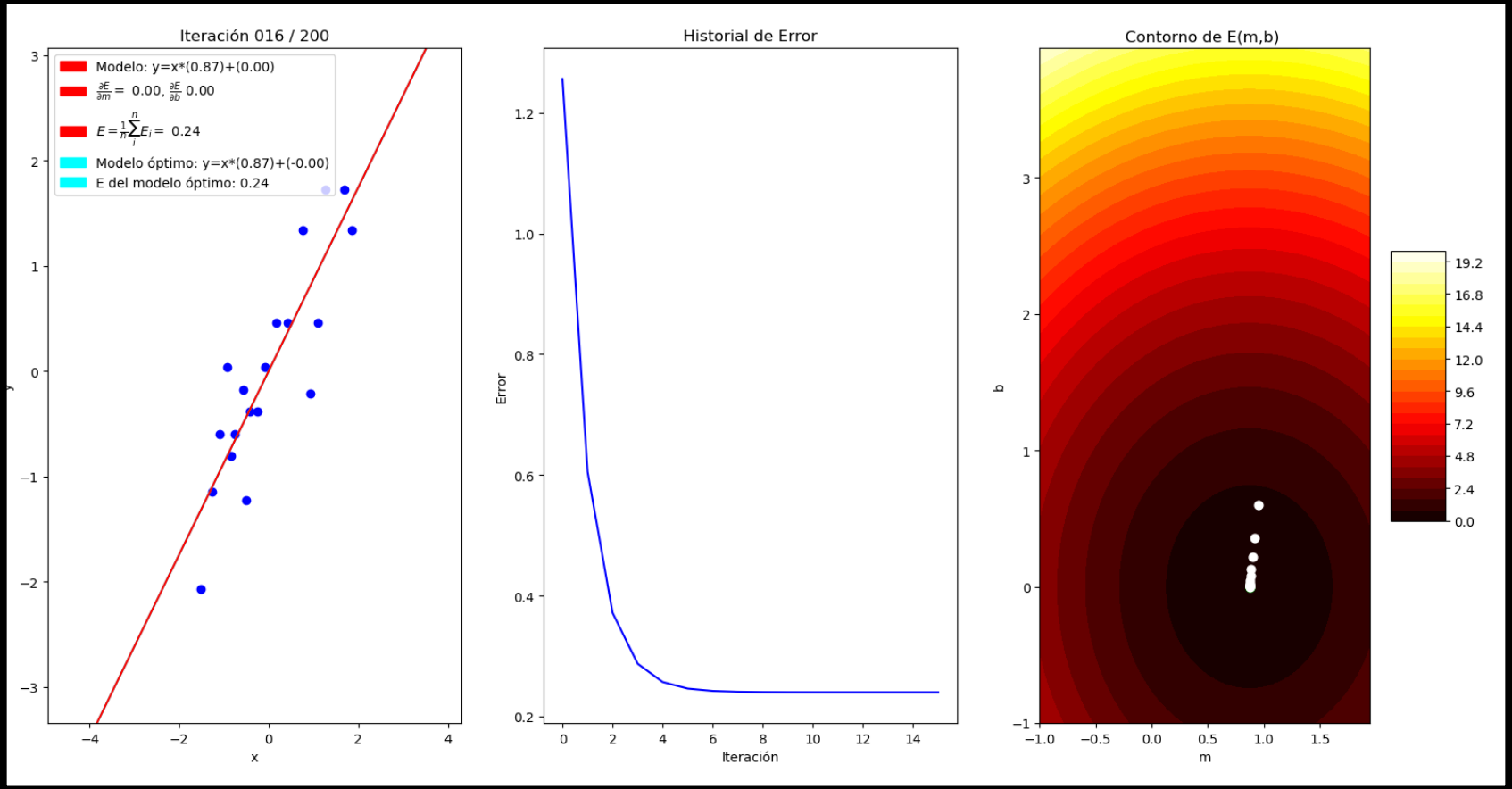
Normalización μ/σ

Horas	Nota
-1.75	-2.03
-1.06	-0.58
-0.60	0.28
-0.14	1.27
0.09	-0.05
0.32	0.28
0.87	0.94
1.01	-0.71
1.24	0.61

Normalización min/max

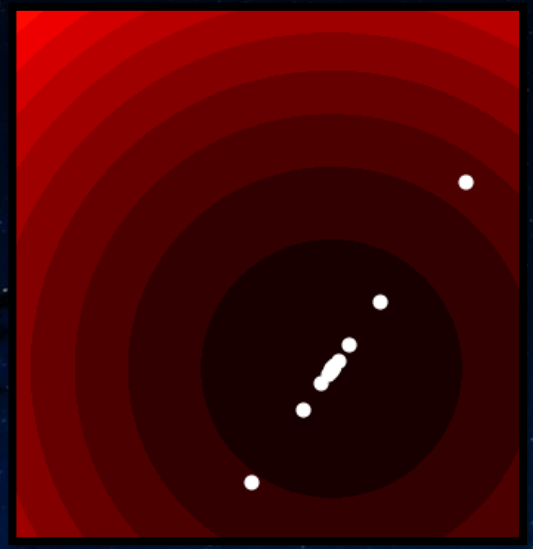
Horas	Nota
0	0
0.23	0.44
0.38	0.7
0.54	1
0.62	0.6
0.69	0.7
0.88	0.9
0.92	0.4
1.00	0.8

Entrenamiento con datos normalizados

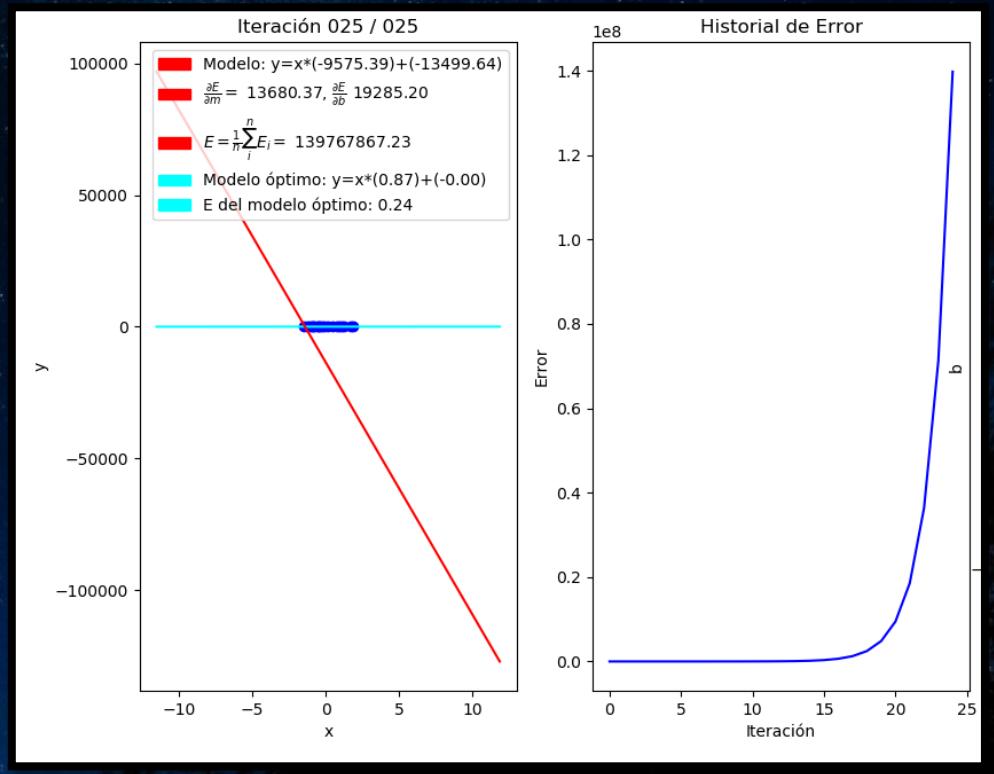


Divergencia con α muy grande

Alfa grande, muy cerca de divergir



Alfa muy grande, el algoritmo diverge



Resumen Regresión Lineal

- RL asume que la relación entre x e y es lineal (con un poco de ruido).
- El modelo que optimizamos es la función: y = mx + b
- Minimización de Error Cuadrático Medio: $\boldsymbol{E = \frac{1}{n}\sum_{i}^{n}(y'_i - y_i)^2}$
- Dado un dataset, se pueden encontrar m y b óptimos de varias maneras:
  - Las clásicas tienen soluciones analíticas (en forma cerrada).
  - Descenso de gradiente es iterativo pero sirve para varios tipos de modelo.
- Regresión lineal es un modelo de caja blanca; podemos interpretar m y b.

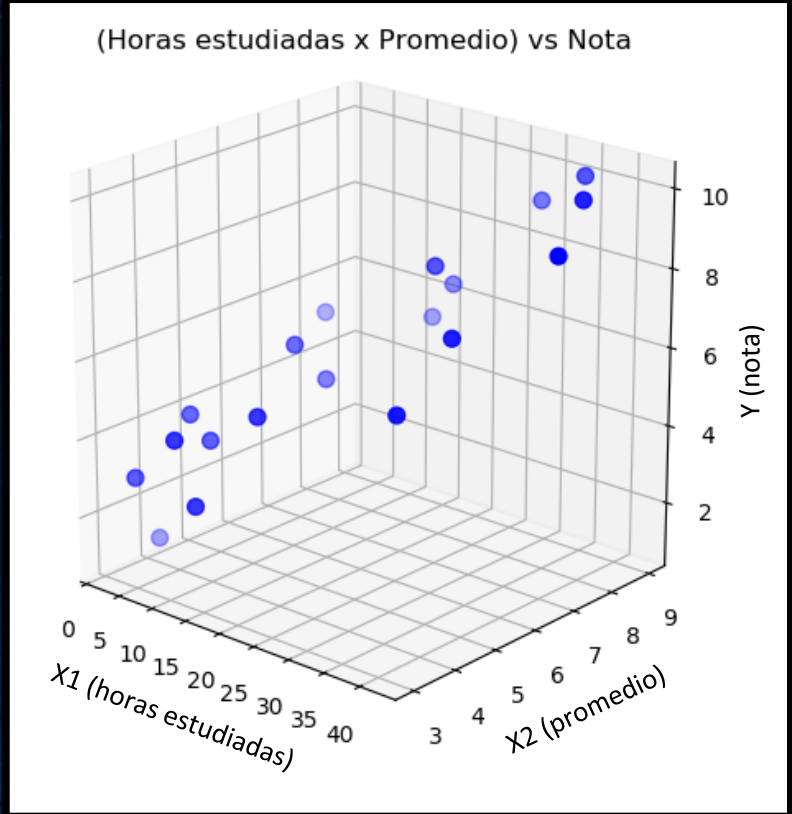
Regresión Lineal ND

Hasta ahora:

- Datos en 1D. Una variable.
- $x_i \in \mathbb{R}$.

¿Qué sucede si agregamos más información de cada ejemplo?

- $(x_i \in \mathbb{R}^d)$.
- Ejemplo, *X= (hora de estudio, promedio)*



Regresión Lineal ND

DATOS: X= (X1, X2)

Target: Y

Estudio (x1)	Promedio (X2)	Nota (Y)
2	4	1
5	3	3,2
7	4	4,5
9	7	6
10	4	4
11	3	4,5
13,4	5	5,5
14	3	3

Regresión Lineal ND

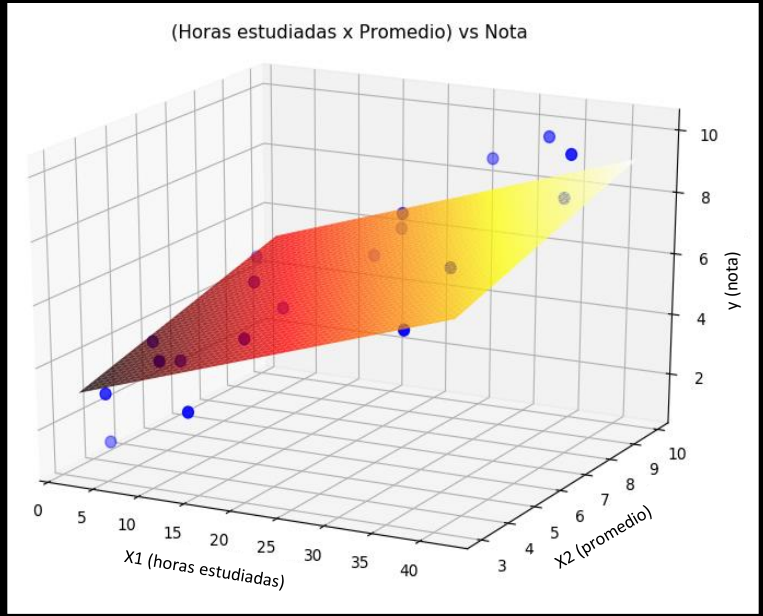
Modelo: $y = f(x1, x2) = w_1 x_1 + w_2 x_2 + b$

Dos pendientes : w_1, w_2

(coeficientes o pesos)

$$E = \frac{1}{n}\sum_{i}^{n} E_i = \frac{1}{n}\sum_{i}^{n} (y'_i - y_i)^2$$

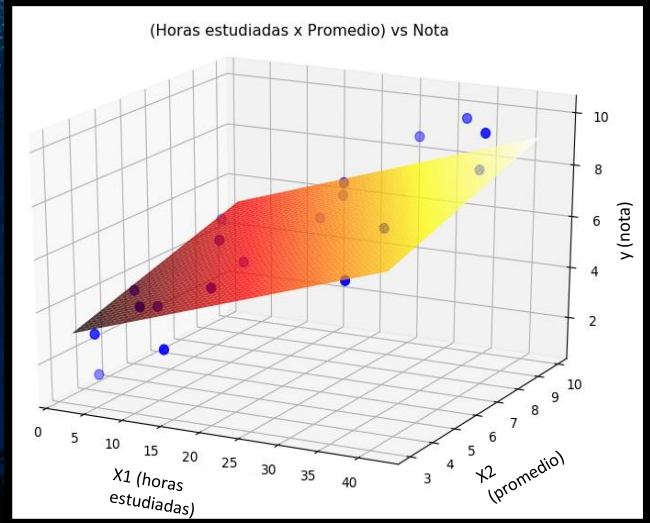
$$E = \frac{1}{n}\sum_{i}^{n} \left((w_1 x_1^i + w_2 (w_1 x_2^i + b) - y_i \right)^2$$



Regresión Lineal ND

¿Qué cambia?

- Prácticamente nada.
- Ahora tenemos 3 parámetros a optimizar (w_1, w_2, b).
- 3 Derivadas.
- Ya no es posible graficar el Error ☹
- Difícil encontrar buenos parámetros manualmente.



Derivadas del error con 2 variables

$$\Delta E = (\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, \frac{\partial E}{\partial b})$$

$$\frac{\partial E}{\partial w_1} = \frac{2}{n}\sum_{i}^{n}(y_i - f(x_i))x_{i,1}$$

$$\frac{\partial E}{\partial w_2} = \frac{2}{n}\sum_{i}^{n}(y_i - f(x_i))x_{i,2}$$

$$\frac{\partial E}{\partial b} = \frac{2}{n}\sum_{i}^{n} y_i - f(x_i)$$

Modelo Lineal con M variables

Modelo: hiperplano en R^m

m coeficientes: $w_1, w_2, \dots, w_m$

Función del modelo

$$y = f(x_1, x_2, \dots, x_m) = w_1 x_1 + w_2 x_2 + \cdots + w_m x_m + b$$

Derivadas parciales del error.

$$\boldsymbol{\Delta \mathbf{E}} = \left(\frac{\delta E}{\delta x^0}, \dots, \frac{\delta E}{\delta x^m} \right)$$

$$\frac{\delta E}{\delta x^j} = \frac{2}{n} \sum_{i}^{n} (y'_i - y_i) x_i^j \qquad , para\ j \in \{0..m\},\ x^0 = 1$$