Aprendizaje automatizado

Descenso por gradiente

Gibran Fuentes Pineda

Febrero 2023

Método alternativo: descenso por gradiente

- Algoritmo iterativo de primer orden que va moviendo los parámetros hacia donde el error descienda más rápido en el vecindario

$$\boldsymbol{\theta}^{[t+1]} = \boldsymbol{\theta}^{[t]} - \alpha \nabla E(\boldsymbol{\theta}^{[t]})$$

  donde

$$\nabla E(\boldsymbol{\theta}^{[t]}) = \left[\frac{\partial E}{\partial \theta_0^{[t]}}, \cdots, \frac{\partial E}{\partial \theta_d^{[t]}} \right]$$

- A α se le conoce como tasa de aprendizaje

Gradiente para regresión lineal

- Gradiente de la función de error de suma de errores cuadráticos respecto a los parámetros está dado por

$$\nabla E(\boldsymbol{\theta}) = \nabla \left[\frac{1}{2} \cdot \sum_{i=1}^{n} \left\{ f_{\boldsymbol{\theta}}(\mathsf{x}^{(i)}, \boldsymbol{\theta}) - y^{(i)} \right\}^2 \right] = \mathsf{X}^\top (f_{\boldsymbol{\theta}}(\mathsf{X}, \boldsymbol{\theta}) - \mathsf{y})$$

- donde X es la matriz de diseño

Algoritmo del descenso por gradiente para regresión lineal

1. Asignar valores aleatorios a los parámetros $\boldsymbol{\theta}$

2. Repetir hasta que converja

$$\theta_0 \leftarrow \theta_0 - \alpha \overbrace{\sum_{i=1}^{n} \left(f_{\boldsymbol{\theta}}(\mathsf{x}^{(i)}) - y^{(i)} \right)}^{\frac{\partial E(\theta_0)}{\partial \theta_0}}$$

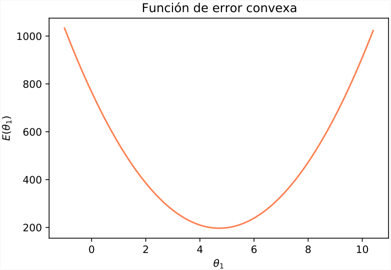
$$\theta_j \leftarrow \theta_j - \alpha \underbrace{\sum_{i=1}^{n} \left(f_{\boldsymbol{\theta}}(\mathsf{x}^{(i)}) - y^{(i)} \right) \cdot x_j^{(i)}}_{\frac{\partial E(\theta_j)}{\partial \theta_j}}$$

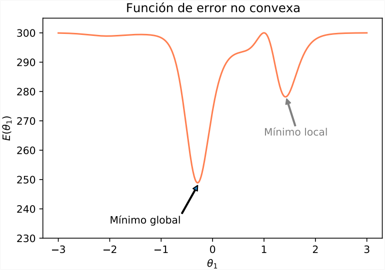
(Actualización simultánea de θ_0 y todos los θ_j)

Función de error convexa vs no convexa

- Cuando $E(\boldsymbol{\theta})$ es convexa, la solución puede converger al mínimo global

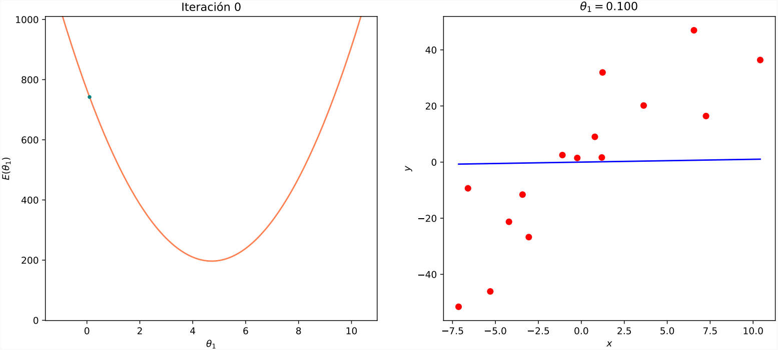
- Cuando $E(\boldsymbol{\theta})$ no es convexa, la solución puede converger a cualquier mínima





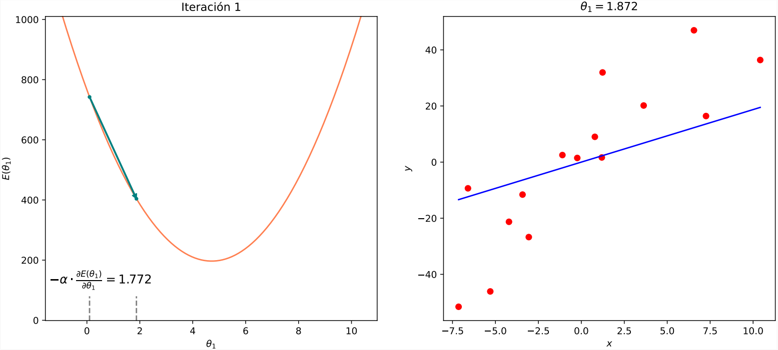
Ejemplo del descenso por gradiente (GD)

- Inicializando θ_1 con un valor menor al que minimiza la función de pérdida



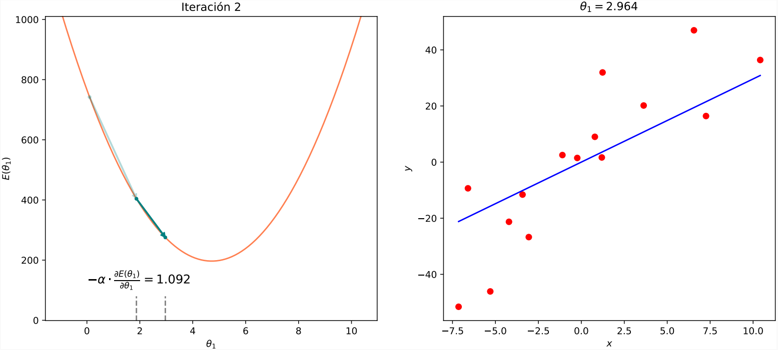
Ejemplo del algoritmo de descenso por gradiente (GD)

- Inicializando θ_1 con un valor menor al que minimiza la función de pérdida



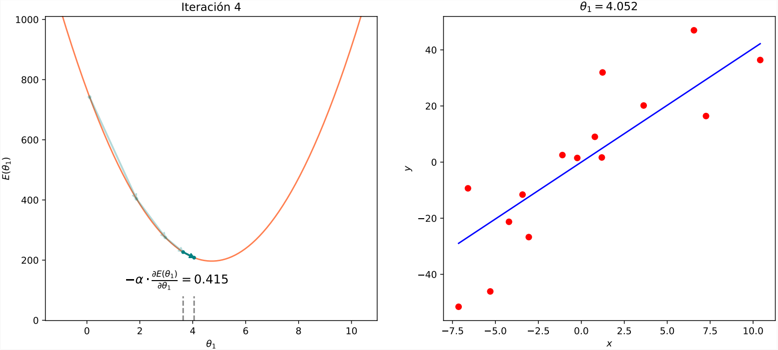
Ejemplo del algoritmo de descenso por gradiente (GD)

- Inicializando θ_1 con un valor menor al que minimiza la función de pérdida



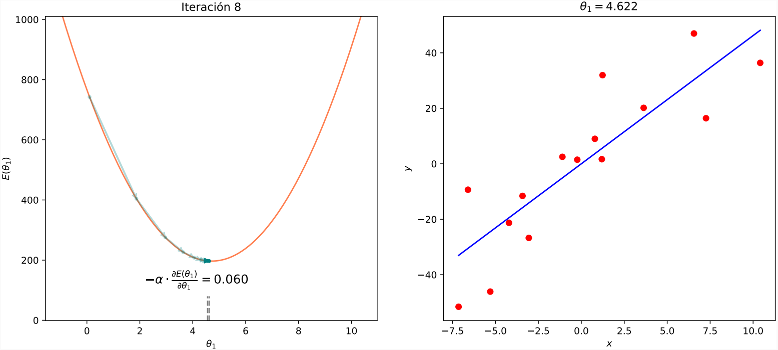
Ejemplo del descenso por gradiente (GD)

- Inicializando θ_1 con un valor menor al que minimiza la función de pérdida



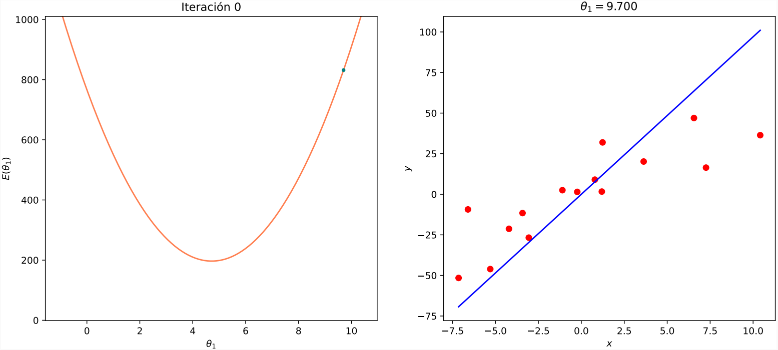
Ejemplo del descenso por gradiente (GD)

- Inicializando θ_1 con un valor menor al que minimiza la función de pérdida



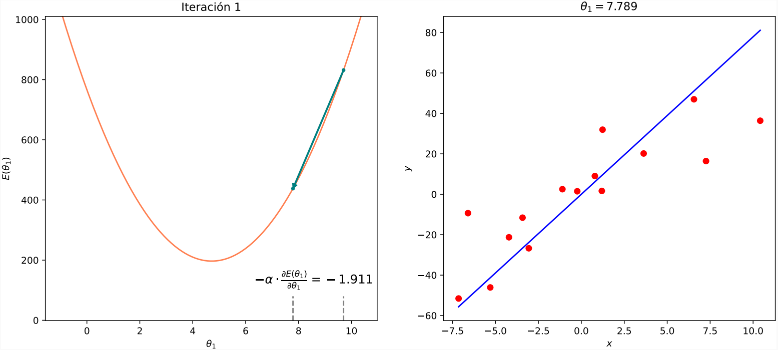
Ejemplo del descenso por gradiente (GD)

- Inicializando θ_1 con un valor mayor al que minimiza la función de pérdida



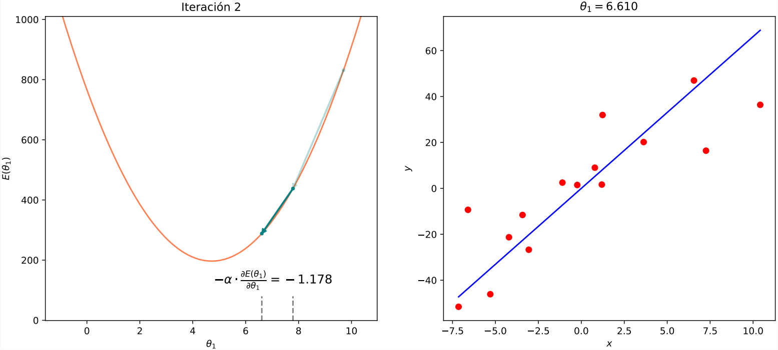
Ejemplo del descenso por gradiente (GD)

- Inicializando θ_1 con un valor mayor al que minimiza la función de pérdida



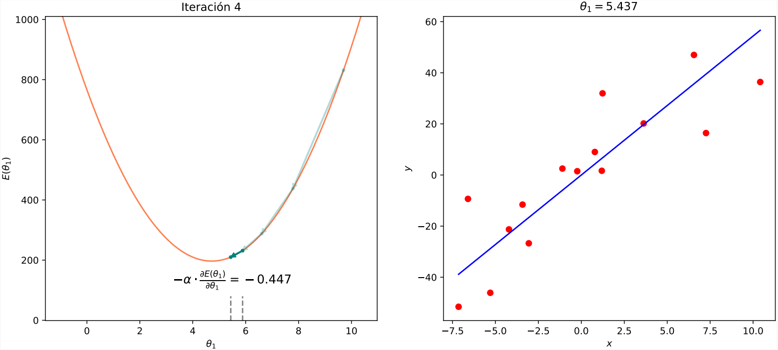
Ejemplo del descenso por gradiente (GD)

- Inicializando θ_1 con un valor mayor al que minimiza la función de pérdida



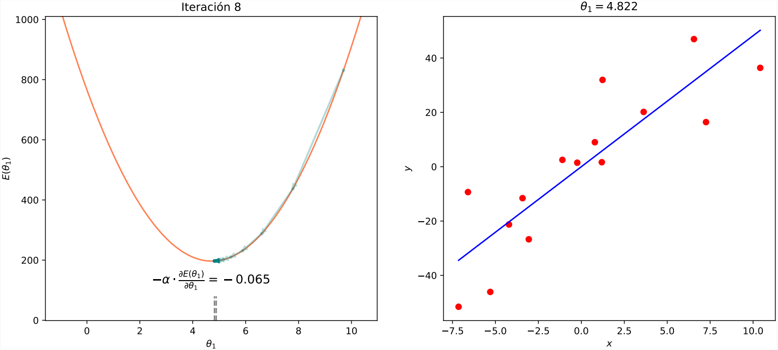
Ejemplo del descenso por gradiente (GD)

- Inicializando θ_1 con un valor mayor al que minimiza la función de pérdida

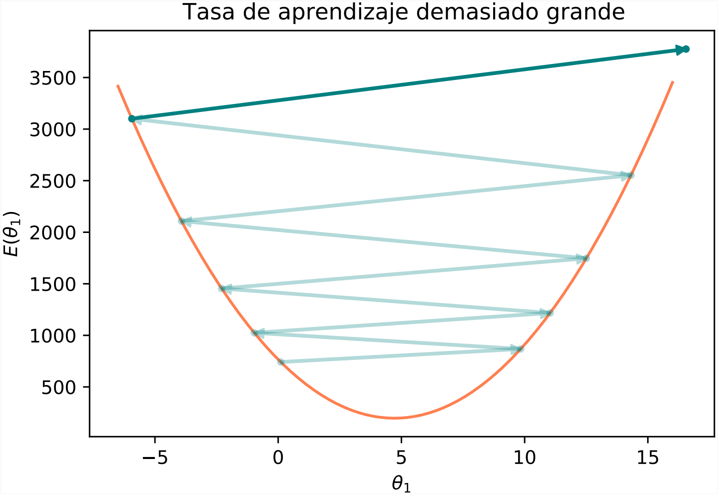


Ejemplo del descenso por gradiente (GD)

- Inicializando θ_1 con un valor mayor al que minimiza la función de pérdida

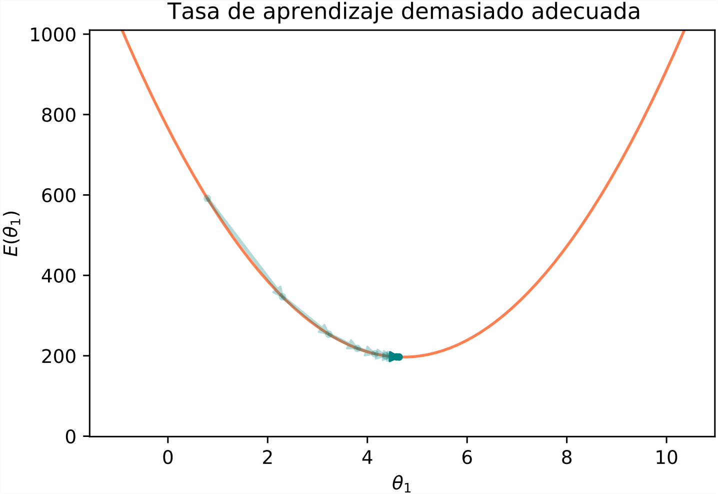


Sensibilidad a tasa de aprendizaje α

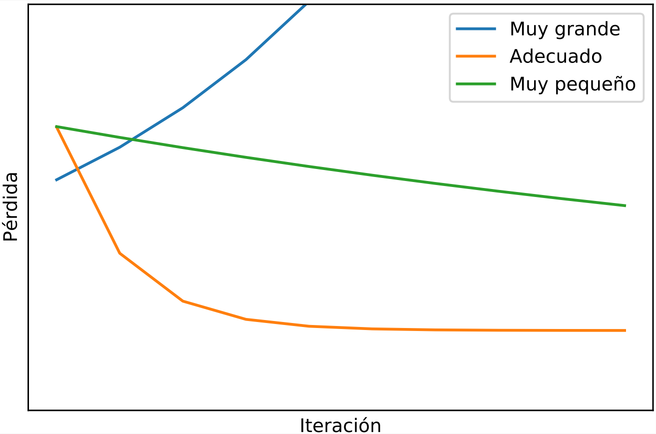


Sensibilidad a tasa de aprendizaje α

Sensibilidad a tasa de aprendizaje α



Sensibilidad a tasa de aprendizaje α



Escalando características

- **El problema**: los valores de las características pueden estar en rangos de valores muy diferentes

Escalando características

- **El problema**: los valores de las características pueden estar en rangos de valores muy diferentes
- **La estrategia**: Normalizar los rangos tal que todas las características contribuyan proporcionalmente a la distancia

Escalando características

- **El problema**: los valores de las características pueden estar en rangos de valores muy diferentes
- **La estrategia**: Normalizar los rangos tal que todas las características contribuyan proporcionalmente a la distancia
- **Diferentes métodos**:

$$x' = \frac{x - min(x_{1:n})}{max(x_{1:n}) - min(x_{1:n})} \qquad \text{(Re-escalado)}$$

$$x' = \frac{x - \bar{x}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x^{(i)} - \bar{x})^2}} \qquad \text{(Estandaziración)}$$

$$x' = \frac{x}{\|x\|} \qquad \text{(Magnitud unitaria)}$$

Descenso por gradiente estocástico

- Aproximación estocástica de GD: estima $\nabla E(\boldsymbol{\theta}^{[t]})$ y actualiza parámetros con un subconjunto $\mathcal{B}$ de ejemplos de entrenamiento
    - $|\mathcal{B}|$ es un hiperparámetro
    - Es común dividir y ordenar aleatoriamente el conjunto de n ejemplos de entrenamiento en k minilotes ($|\mathcal{B}| \times k \approx n$)