Cálculo Matricial

① Gradiente da ~~constante~~ é zero → $f(\vec{x}) = Cte \Rightarrow \nabla f = \vec{0}$

② $f(\vec{x}) = \underbrace{\vec{x}^T}_{1\times n} \cdot \underbrace{\vec{a}}_{n\times 1} \rightarrow \nabla f = \vec{a}$

vetor constante

③ $f(\vec{x}) = \underbrace{\vec{x}^T}_{1\times n} \underbrace{A}_{n\times n} \underbrace{\vec{x}}_{n\times 1} \Rightarrow \nabla f = (A + A^T)\vec{x}$

se $A = A^T \Rightarrow \nabla f = 2 \cdot A\vec{x}$

Note que: $\hat{y}^T y = y^T \hat{y}$

pois: $\underbrace{\hat{y}^T}_{1\times m} \underbrace{y}_{m\times 1}$ é matriz 1×1

Logo $(\hat{y}^T y) = (\hat{y}^T y)^T$ pois matriz 1×1 transposta é ela mesma

$\hat{y}^T y = y^T \hat{y}$

Coisas de matriz: $(A+B)^T = A^T + B^T$

$(AB)^T = B^T A^T$

Aplicando as regras do cálculo matricial ao MSE, temos:

$$\boxed{\nabla MSE = \frac{1}{m}\left(2X^T X\theta - 2X^T y\right)}$$

$$\nabla MSE = \vec{0} \Rightarrow \frac{1}{m}\left(2X^T X\theta - 2X^T y\right) = \vec{0} \Rightarrow$$

$$\Rightarrow 2X^T X\theta = 2X^T y \Rightarrow \boxed{\theta_{opt} = (X^T X)^{-1} X^T y}$$

Equação normal

Quanto custa calcular θ_{opt}?

$$\theta_{opt} = \underbrace{(X^T X)}_{O(mn^2)}{}^{-1} \cdot \underbrace{X^T y}_{O(mn)} \qquad O(n^3)$$

<u>Modelo linear para regressão</u>

$$\hat{y} = \theta_0 + \theta_1 \cdot x_1 + \theta_2 x_2 + \ldots + \theta_n x_n$$

<u>notação matricial</u>:

$$\hat{y} = X\theta \text{, onde: } \hat{y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_m \end{bmatrix} \quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

$$\boxed{MSE = \frac{1}{m}(X\theta - y)^T(X\theta - y)} = \boxed{\frac{1}{m}\left(\theta^T X^T X\theta - 2\theta^T X^T y + y^T y\right)}$$

$$\boxed{\nabla MSE = \frac{1}{m}\left(2XX^T\theta - 2X^Ty\right)}$$

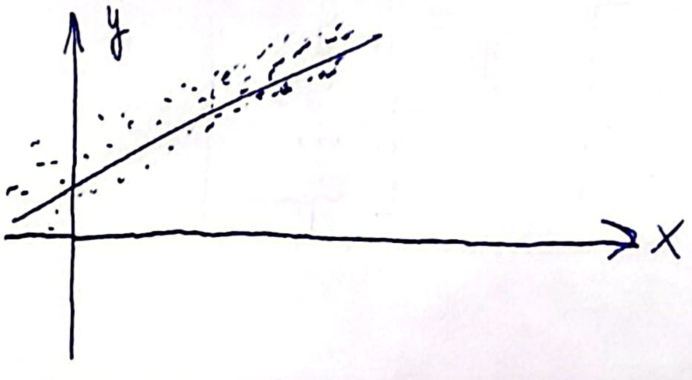
<u>Equação normal:</u>

$$\nabla MSE = 0 \Rightarrow \frac{1}{m}\left(2X^TX\theta - 2X^Ty\right) = \vec{0}$$

$$\boxed{\theta = (X^TX)^{-1} \cdot X^Ty}$$

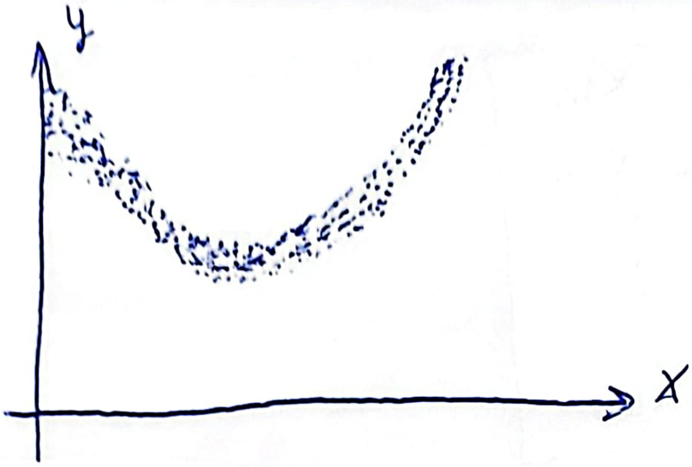
<u>Regressão linear + feature engineering:</u>

1)



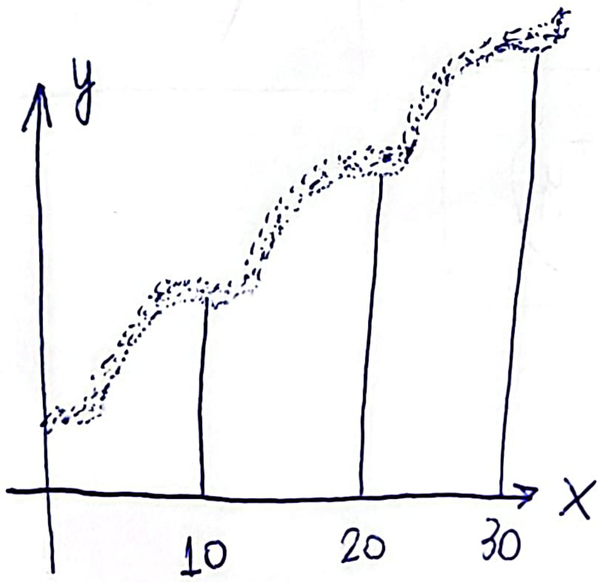
$$\hat{y} = \theta_0 + \theta_1 X$$

2)



$$\hat{y} = \theta_0 + \theta_1 x + \theta_2 x^2$$

3)



$$\hat{y} = \theta_0 + \theta_1 x + \theta_2 \cdot \cos\left[\left(\pi/5\right)(x)\right] + \theta_3 \cdot \sin\left[\left(\pi/5\right)(x)\right]$$

<u>notação matricial:</u>

1) $\Theta = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}; \; X = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \cdots & \cdots \\ 1 & x_{m1} \end{bmatrix}; \; y = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_m \end{bmatrix}$

2) $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}$; $X = \begin{bmatrix} 1 & x_{11} & x_{11}^2 \\ 1 & x_{21} & x_{21}^2 \\ \dots & \dots & \dots \\ 1 & x_{m1} & x_{m1}^2 \end{bmatrix}$

3) $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$; $X = \begin{bmatrix} 1 & x_{11} & \cos(\dots) & \sin(\dots) \\ 1 & x_{21} & \cos(\dots) & \sin(\dots) \\ \dots & \dots & \dots & \dots \\ 1 & x_{m1} & \cos(\dots) & \sin(\dots) \end{bmatrix}$

Intuição geométrica:

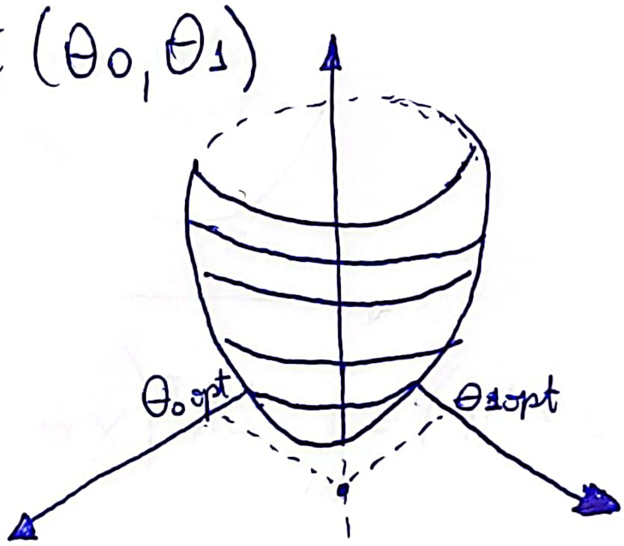
$$MSE(\theta) = \frac{1}{m}\left(\theta^T X^T X \theta - 2\theta^T X^T y + y^T y\right)$$

Ato de fé: $MSE(\theta)$ é um paraboloide em θ

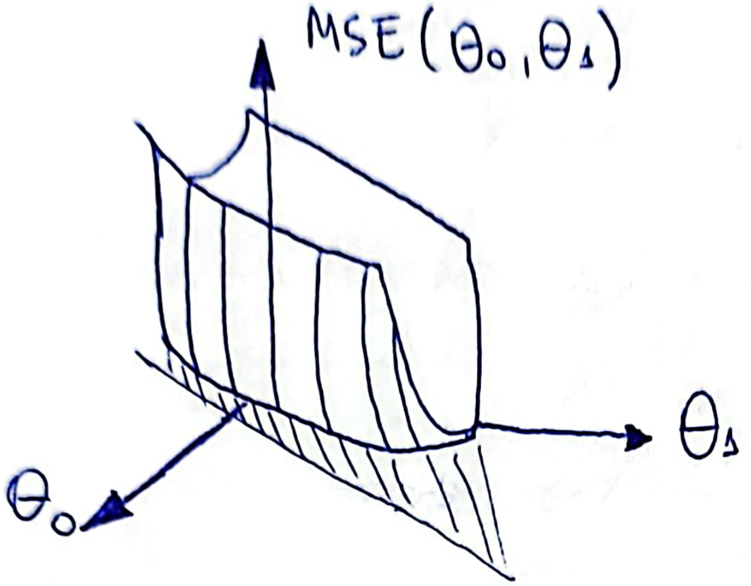
Exemplo:

$$\hat{y} = \theta_0 + \theta_1 x$$

$MSE(\theta_0, \theta_1)$



Exemplo "degenerado":



Gradient Descent

1) Chuta $\theta^{(0)}$ ← iteração
   $i=0$

2) Enquanto não "acabou": $\theta^{(i+1)} = \theta^{(i)} - \eta \nabla MSE(\theta^{(i)})$

   η ← taxa de aprendizado

3) Retorna o último θ

Lembrando que:

$$\nabla MSE = \frac{1}{m}\left(2X^T X\theta - 2X^T y\right) = \frac{1}{m} 2X^T(X\theta - y)$$

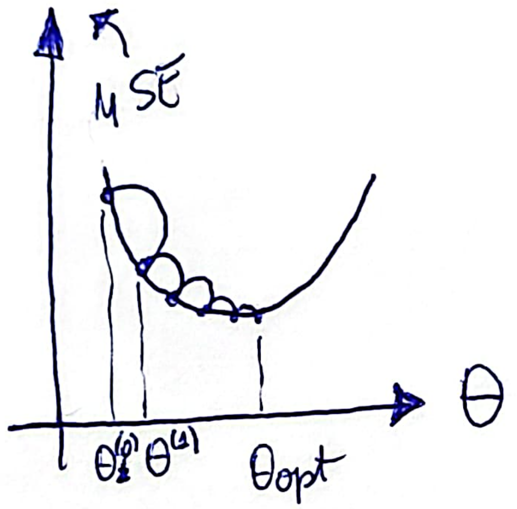
Critério de parada

Se $\|\Theta^{(i+1)} - \Theta^{(i)}\|$ < tolerância ou se i > máximo_de_iterações

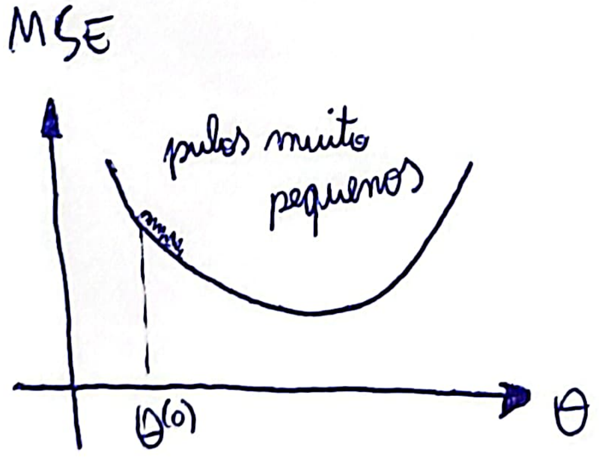
pode parar!

intuição geométrica do gradient descent:

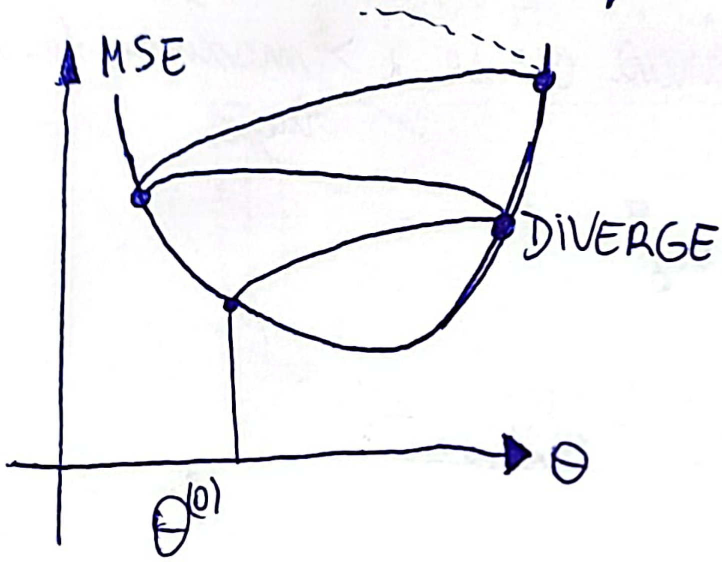
1)



2) Taxa de aprendizagem muito pequena



3) Taxa de aprendizado muito grande



Modelos lineares

- conjunto de dados:

$$D = \{ (\vec{x_1}, y_1), (\vec{x_2}, y_2), \dots, (\vec{x}_n, y_n) \}$$

onde: $\vec{x_i} = (f_1, f_2, \dots, f_n)$, $f_k \in \mathbb{R}$

$y_i \in \mathbb{R}$ (regressão)

ou

$y_i \in \{\text{conjunto de classes}\}$ (classificação)

- notação matricial:

$$X_i = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix} \text{ (matriz-coluna)}$$

$$X = \begin{bmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_m^T \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

$$X \in M_{m,n}(\mathbb{R})$$

modelo : $h_\theta : \mathbb{R}^n \to \mathbb{R}$ (regressão)

↳ parâmetros treináveis $\{\text{classes}\}$ (classificação)

denota previsão do modelo

$$\hat{y} = h_\theta(x) \rightarrow \hat{y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_m \end{bmatrix}$$

→ matriz-coluna (vetor) das previsões

erro de regressão

$$\varepsilon_i = \hat{y}_i - y_i$$

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{bmatrix} = \begin{bmatrix} \hat{y}_1 - y_1 \\ \hat{y}_2 - y_2 \\ \vdots \\ \hat{y}_m - y_m \end{bmatrix} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_m \end{bmatrix} - \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

$$\boxed{\varepsilon = \hat{y} - y}$$

erro médio quadrático (MSE)

$$MSE = \frac{1}{m} \sum_{i=1}^{m} \varepsilon_i^2 = \boxed{\frac{1}{m} \varepsilon^T \cdot \varepsilon}$$

Agora, o modelo linear

$$h_\theta(x) = \theta_0 + \theta_1 \underbrace{x_1} + \theta_2 \underbrace{x_2} + \dots + \underbrace{x_n}_{\text{features}} \cdot \theta_n$$

features

- Treinar o modelo:

• Descobrir $\Theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \dots \\ \theta_n \end{bmatrix}$ que minimiza o MSE

Um modelo de ML:

1) Dados
2) Função de erro
3) Modelo
4) Algoritmo de treinamento

Modelo Linear na forma matricial

Defina: $x_i' = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}$

Com isso: $\hat{y} = \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \dots + \theta_n x_n$

$$\boxed{\hat{y} = x'^T \cdot \Theta}$$

Portanto, defina:

$$X = \begin{bmatrix} 1 & — & x_1^T & — \\ 1 & — & x_2^T & — \\ \vdots & & \vdots & \\ 1 & — & x_m^T & — \end{bmatrix} \quad , \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_m \end{bmatrix}$$

$$\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \cdots \\ \hat{y}_m \end{bmatrix} = \begin{bmatrix} 1 & — & x_1^T & — \\ 1 & — & x_2^T & — \\ \cdots & & \cdots & \\ 1 & — & x_m^T & — \end{bmatrix} \cdot \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \cdots \\ \theta_n \end{bmatrix} = X \cdot \theta$$

Juntando as partes:

$$\varepsilon = \hat{y} - y \ ; \ \hat{y} = X\theta \ ; \ MSE = \frac{1}{m} \cdot \varepsilon^T \cdot \varepsilon$$

Logo:

$$MSE = \frac{1}{m} \cdot \left(\hat{y}^T \hat{y} - 2\hat{y}^T y + y^T y \right) =$$

$$= \frac{1}{m}\left((X\theta)^T (X\theta) - 2(X\theta)^T y + y^T y \right) = \boxed{\frac{1}{m}\left(\theta^T X^T X \theta - 2\theta^T X^T y + y^T y \right)}$$

$$\nabla MSE = \begin{bmatrix} \dfrac{\partial MSE}{\partial \theta_0} \\ \partial MSE / \partial \theta_1 \\ \partial MSE / \partial \theta_2 \\ \cdots \\ \partial MSE / \partial \theta_n \end{bmatrix} = 0$$