Modelo linear:

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

Nosso objetivo: achar os θ's tal que nosso erro de treinamento seja o menor possível.

Conjunto de treinamento: X, y

$$X = \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & & | \end{bmatrix} \qquad y = \begin{bmatrix} | \\ y \\ | \end{bmatrix}$$

exemplo 1, feature 1

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \qquad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

exemplo m, feature n

$$\hat{y}^{(i)} = \theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + \dots + \theta_n x_n^{(i)}$$

previsão p/ i-ésima amostra de treino

<u>IDEIA</u>: Quero escrever as previsões p/ todas as amostras de treinamento de uma forma bem compacta, com matrizes.

$$\hat{y} = \begin{bmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \vdots \\ \hat{y}^{(m)} \end{bmatrix} = \begin{bmatrix} \theta_0 + \theta_1 x_1^{(1)} + \dots + \theta_n x_n^{(1)} \\ \theta_0 + \theta_1 x_1^{(2)} + \dots + \theta_n x_n^{(2)} \\ \vdots \\ \theta_0 + \theta_1 x_1^{(m)} + \dots + \theta_n x_n^{(m)} \end{bmatrix}$$

$$= \begin{bmatrix} \theta_0 \cdot 1 + \theta_1 x_1^{(1)} + \dots + \theta_n x_n^{(1)} \\ \theta_0 \cdot 1 + \theta_1 x_1^{(2)} + \dots + \theta_n x_n^{(2)} \\ \vdots \\ \theta_0 \cdot 1 + \theta_1 x_1^{(m)} + \dots + \theta_n x_n^{(m)} \end{bmatrix}$$

$$= \underbrace{\begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \dots & x_n^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \dots & x_n^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(m)} & x_2^{(m)} & \dots & x_n^{(m)} \end{bmatrix}}_{\underset{\sim}{X}} \underbrace{\begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix}}_{\theta}$$

"X's" aumentada

$\Rightarrow$ todas as previsões : $\hat{y} = \underset{\sim}{X}\,\theta$

Erro de treinamento

$$\varepsilon^{(i)} = \hat{y}^{(i)} - y^{(i)}$$

$$\varepsilon = \begin{bmatrix} \varepsilon^{(1)} \\ \varepsilon^{(2)} \\ \vdots \\ \varepsilon^{(m)} \end{bmatrix} = \begin{bmatrix} \hat{y}^{(1)} - y^{(1)} \\ \hat{y}^{(2)} - y^{(2)} \\ \vdots \\ \hat{y}^{(m)} - y^{(m)} \end{bmatrix} = \begin{bmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \vdots \\ \hat{y}^{(m)} \end{bmatrix} - \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

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

$$SSE = \left[\varepsilon^{(1)}\right]^2 + \left[\varepsilon^{(2)}\right]^2 + \dots + \left[\varepsilon^{(m)}\right]^2$$

$$= \begin{bmatrix} \varepsilon^{(1)} & \varepsilon^{(2)} & \dots & \varepsilon^{(m)} \end{bmatrix} \begin{bmatrix} \varepsilon^{(1)} \\ \varepsilon^{(2)} \\ \vdots \\ \varepsilon^{(m)} \end{bmatrix}$$

$$= \varepsilon^T \cdot \varepsilon = (\hat{y} - y)^T (\hat{y} - y)$$

$$= (\underset{\sim}{X}\theta - y)^T (\underset{\sim}{X}\theta - y)$$

$$= \left([\underset{\sim}{X}\theta]^T - y^T\right)(\underset{\sim}{X}\theta - y)$$

$$= (\theta^T \underset{\sim}{X}^T - y^T)(\underset{\sim}{X}\theta - y)$$

$$SSE = \theta^T \underset{\sim}{X}^T \underset{\sim}{X} \theta - 2\theta^T \underset{\sim}{X}^T y + y^T y$$

$$y = ax^2 + bx + c$$

$$\frac{\partial y}{\partial x} = 2ax + b$$

$$\frac{\partial SSE}{\partial \theta} = \nabla_\theta SSE = 2\underset{\sim}{X}^T \underset{\sim}{X}\theta - 2\underset{\sim}{X}^T y = 0$$

$$\Rightarrow 2\underset{\sim}{X}^T \underset{\sim}{X}\theta = 2\underset{\sim}{X}^T y \Rightarrow (\underset{\sim}{X}^T \underset{\sim}{X})\theta = \underset{\sim}{X}^T y$$

$$\Rightarrow \boxed{\theta^* = (\underset{\sim}{X}^T \underset{\sim}{X})^{-1} \underset{\sim}{X}^T y}$$

Equação normal!

4

Com regularização Ridge:

$$\text{Perda} = \underline{SSE} + \alpha \underline{\sum_{j=0}^{n} \theta_j^2}$$

$$\Rightarrow \text{Perda} = 2\theta^T (\underset{\sim}{X}^T \underset{\sim}{X} + \alpha I)\theta - 2\underset{\sim}{X}^T y + y^T y$$

$$\frac{\partial \text{Perda}}{\partial \theta} = 0 \Rightarrow 2(\underset{\sim}{X}^T \underset{\sim}{X} + \alpha I)\theta - 2\underset{\sim}{X}^T y = 0$$

$$\Rightarrow \theta^* = (\underset{\sim}{X}^T \underset{\sim}{X} + \alpha I)^{-1} \underset{\sim}{X}^T y$$

Equação normal da reg. Ridge.

SVD: $X = USV^T$

U, V unitarias

$S = \begin{bmatrix} \Sigma \\ \hdashline 0 \end{bmatrix}$ bloco diagonal não neg. / bloco zeros

Pois: $\underset{\sim}{X} = USV^T$

$$\underset{\sim}{X}^T \underset{\sim}{X} = VS^T \underbrace{\cancel{U^T U}}_{I} SV$$

$$= V \left[\Sigma^T \,\vdots\, 0^T \right] \begin{bmatrix} \Sigma \\ \hdashline 0 \end{bmatrix} V = V\Sigma^2 V^T$$

~~Σ² quadrado~~

$$(\underset{\sim}{X}^T \underset{\sim}{X})^{-1} = (V\Sigma^2 V^T)^{-1} = V\Sigma^{-2} V^T$$

$$\begin{bmatrix} 1/\sigma_1^2 & & & 0 \\ & 1/\sigma_2^2 & & \\ & & \ddots & \\ 0 & & & 1/\sigma_n^2 \end{bmatrix}$$

$$(ABC)^{-1} = C^{-1} B^{-1} A^{-1}$$

Se X cols colinear $\Rightarrow \sigma_n = 0$

$\Rightarrow \Sigma^{-2}$ não existe!

Já com regularização Ridge: $\alpha > 0$

$$(\underset{\sim}{X}^T \underset{\sim}{X} + \alpha I)$$

$$= (V S^T \overset{I}{\cancel{U^T U}} S V^T + \alpha I)$$

$$= (V \Sigma^2 V^T + \alpha I)$$

$$= (V \Sigma^2 V^T + V(\alpha I) V^T)$$

$$= V(\Sigma^2 + \alpha I) V^T$$

$$(\underset{\sim}{X}^T \underset{\sim}{X} + \alpha I)^{-1} = \left[V(\Sigma^2 + \alpha I) V^T \right]^{-1}$$

$$= V \underbrace{(\Sigma^2 + \alpha I)^{-1}}_{} V^T$$

$$\begin{bmatrix} (\sigma_1^2 + \alpha) & & & 0 \\ & (\sigma_2^2 + \alpha) & & \\ & & \ddots & \\ 0 & & & (\sigma_n^2 + \alpha) \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} \frac{1}{(\sigma_1^2 + \alpha)} & & & \\ & \frac{1}{(\sigma_2^2 + \alpha)} & & \\ & & \ddots & \\ & & & \frac{1}{(\sigma_n^2 + \alpha)} \end{bmatrix}$$