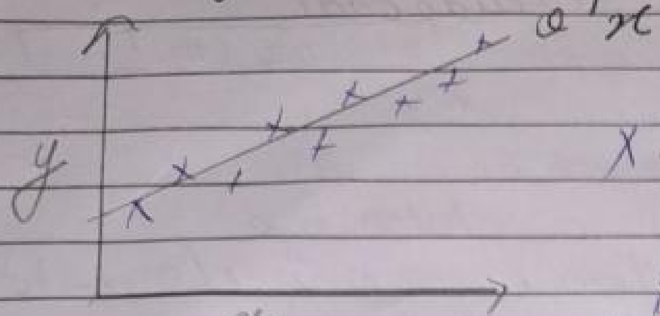
Regression - Closed Form Solution



$$\theta^* = (X^T X)^{-1} X^T y$$

$$X = \begin{bmatrix} | & \\ - x^i - & \\ | & \end{bmatrix}_{m \times (n+1)}$$

m = no. of ex

n = no. of feature

x^i_j i^{th} ex j^{th} feature

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_n \end{bmatrix}_{(n+1 \times 1)} \qquad y = \theta_1 x + \theta_0$$

Proof. $\text{loss} = \frac{1}{2} \sum_{i=1}^{m} (y_{pred} - y)^2$

$$X = \begin{bmatrix} x^1 & 1 \\ \vdots & 1 \\ - x^i - & 1 \\ x^m & 1 \end{bmatrix}_{m \times (n+1)} \quad y = \begin{bmatrix} y^1 \\ y^2 \\ y^m \end{bmatrix}_{(m) \times 1} \quad y_{pred} = h_\theta(x) = \sum_{i=0}^{n} \theta_i x_i^j$$

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}_{(n+1) \times 1}$$

we introduce one more column of 1's in x as

$h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \dots \theta_i x_i$

$x_0 = 1$

$= \theta^T x = \sum_{i=0}^{m} \theta_i x_i$

Proof

$$\text{loss} = \underbrace{\frac{1}{2} (X\theta - y)^2}_{\text{matrix notation}} = \frac{1}{2} \sum_{i=1}^{m} (y^i_{pred} - y^i)^2$$

Date :

X_0^1 = 1st example 0th feature

$$\begin{bmatrix} x_0^1 & x_1^1 & x_2^1 \\ x_0^2 & x_1^2 & x_2^2 \\ x_0^3 & x_1^3 & x_2^3 \\ \vdots \\ x_0^m & x_1^m & x_2^m \end{bmatrix} = X \qquad \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_m \end{bmatrix} = \theta$$

$\Downarrow$ 1

$$X\theta = \begin{bmatrix} \theta_0 x_0^1 + \theta_1 x_1^1 + \theta_2 x_2^1 \\ \theta_0 x_0^2 + \theta_1 x_1^2 + \theta_2 x_2^2 \\ \vdots \end{bmatrix}$$

$$= \begin{bmatrix} \theta_0 + \theta_1 x_1^1 + \theta_2 x_2^1 \\ \theta_0 * \theta_1 x_1^2 + \theta_2 x_2^2 \\ \vdots \end{bmatrix}$$

$$X\theta = \begin{bmatrix} y_{pred}^1 \\ y_{pred}^2 \\ \vdots \\ y_{pred}^m \end{bmatrix}$$

$$(\underbrace{X\theta}_{y_{pred}} - \underbrace{y}_{y_{actual}})^2 = \text{loss}$$

$$\left(\begin{bmatrix} \\ \\ \end{bmatrix}_{m \times 1} - \begin{bmatrix} \\ \\ \end{bmatrix}_{m \times 1} \right)^2 = \text{Scalar}$$

square of a vector is scalar

Page No.
Date :

$$J(\theta) = (X\theta - y)^2 \qquad z^2 = z^T z$$

$$= (X\theta - y)^T (X\theta - y)$$ loss in matrix notation

$$= (\theta^T X^T - y^T)(X\theta - y)$$

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

$$J(\theta) = \theta^T X^T X \theta - \theta^T X^T y - y^T X \theta + y^T y$$

minimize $J(\theta)$

$$\nabla_\theta J(\theta) = 0$$

computing first derivative of $J(\theta)$

$$\nabla_\theta J(\theta) = \nabla_\theta (\overset{③\,\text{rule}}{\theta^T X^T X \theta} - \theta^T X^T y - y^T X \theta + \cancel{y^T y}) \to 0$$

$$= 2X^T X \theta - \nabla_\theta (\theta^T X^T y + \theta^T X^T y)$$

$$= 2X^T X \theta - \nabla_\theta (2\theta^T X^T y) \quad ②\,\text{rule}$$

$$= 2X^T X \theta - 2X^T y = 0 \text{ for minima}$$

$$= X^T X \theta = X^T y$$

$$(X^T X)^{-1} X^T X \theta = (X^T y)(X^T X)^{-1}$$

$$(1)\ \theta = (X^T X)^{-1} X^T y$$

$$\theta = (X^T X)^{-1} X^T y$$

Rules of matrix calculus

① $\nabla_\theta a^T \theta = a$ ② $\nabla_\theta \theta^T a = a$

③ $\nabla_\theta \theta^T A \theta = 2A\theta$

Date :

Gradient Descent

→ iterative approach

→ figure out learning rate

→ slow if n is small

→ useful for big dataset

→ minibatch is useful for very large dataset

Closed form solution

→ small dataset

→ get directly solution

→ expensive if dataset is large