Normal equation

we will know that the normal equation give us the best fit model parameter (or feature weights) θ through the equation

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

- Our target is to prove and understand this formular
- This require you to know a little bit about Linear Algebra. Don't worry, I will go through some important things before I come to prove.

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

2) If A is symetric :

→ $A^T = A$

3) give you a scalar $\alpha = y^T A x$

we know that because α is scalar, we have : $\alpha = \alpha^T$

→ From that:

- $y^T A x = x^T A^T y$
- partial derivative :

$$\frac{\partial \alpha}{\partial x} = y^T . A \qquad \frac{\partial \alpha}{\partial y} = x^T A^T$$

4) given scalar : $\alpha = x^T A x$

- A is not symetric

$$\frac{\partial \alpha}{\partial x} = x^T (A + A^T)$$

- A is symetric

$$\frac{\partial \alpha}{\partial x} = 2 x^T A$$

Revise: $y = \theta_0 x_0 + \theta_1 x_1 + \dots + \theta_n x_n = \theta^T x$

with $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}_{n \times 1}$; $x = \begin{bmatrix} x_0 = 1 \\ x_1 \\ \vdots \\ x_n \end{bmatrix}_{n \times 1}$

→ with many data, we write it in the form

$$\hat{y} = X\theta$$

- $X = \begin{bmatrix} x_0^1 & x_1^1 & \dots & x_n^1 \\ x_0^2 & x_1^2 & \dots & x_n^2 \\ \vdots & \vdots & & \vdots \\ x_0^m & x_1^m & \dots & x_n^m \end{bmatrix}_{m \times n}$ · $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}_{n \times 1}$ · $\hat{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}_{m \times 1}$

- n is number of features
- m is number of samples
- x_j^i is the j^{th} feature in i^{th} sample

→ We want to minimize the cost function.

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left(\theta^T x^{(i)} - y^{(i)} \right)^2$$

Prove • We can drop $\frac{1}{m}$ in $J(\theta)$ for easier to minimize because it is constant and do not have impact when you want to minimize $J(\theta)$

• Rewritten $\sum_{i=1}^{m} (\theta^T x^{(i)} - y^{(i)})^2$ in matrix form:

$$\sum_{i=1}^{m} (\theta^T x^{(i)} - y^{(i)})^2$$

$$= \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)})^2$$

$$= \underbrace{\begin{bmatrix} (\hat{y}^{(1)} - y^{(1)}) & (\hat{y}^{(2)} - y^{(2)}) & \cdots & (\hat{y}^{(m)} - y^{(m)}) \end{bmatrix}}_{1 \times m} \underbrace{\begin{bmatrix} \hat{y}^{(1)} - y^{(1)} \\ \hat{y}^{(2)} - y^{(2)} \\ \vdots \\ y^{(m)} - y^{m} \end{bmatrix}}_{m \times 1}$$

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

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

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

→ From now we have $J(\theta) = \theta^T X^T X \theta - \theta^T X^T y - y^T X \theta + y^T y$

• We want to find θ minimize $J(\theta)$ or $\theta = \arg\min_{\theta} J(\theta)$

take derivative of $J(\theta)$ with respect to θ

this symetrix because $X^TX = (X^TX)^T$

$$\frac{\partial J}{\partial \theta} = \frac{\partial}{\partial \theta}\left(\theta^T X^T X\theta - \theta^T X^T y - y^T X\theta + y^T y\right)$$

(4) (3) (3)

→ check out the Linear Algebra above to understand this step

$$= 2\theta^T X^T X - y^T X - y^T X + 0$$

$$= 2\theta^T X^T X - 2y^T X$$

- Find the critical point to have the θ that minimize $J(\theta)$

$$\frac{\partial J}{\partial \theta} = 0$$

$$\Leftrightarrow \quad 2\theta^T X^T X = 2y^T X$$

$$\Leftrightarrow \quad (\theta^T X^T X)^T = (y^T X)^T$$

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

$$\Rightarrow \quad \theta = (X^T X)^{-1} X^T y \quad \text{(prove)}$$

- Conclusion: $\theta = (X^TX)^{-1}X^Ty$ that is best fit for model