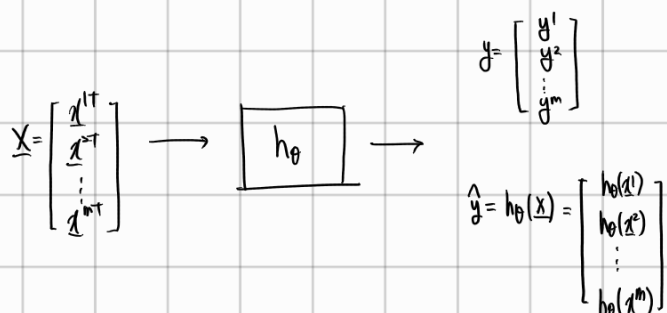


[ML_04] Linear Regression

Data Representation

Feature vector: $\underline{x} = [x_1, x_2, \dots, x_n]^T$

Training dataset: $\underline{D} = \{(\underline{x}^i, y^i)\}_{i=1}^m = \{(\underline{x}^1, y^1), \dots, (\underline{x}^m, y^m)\}$



Linear Model Representation

$\underline{x} \triangleq [x_0, x_1, x_2, \dots, x_n]^T$, $\underline{\theta} = [\theta_0, \theta_1, \dots, \theta_n]^T$

For a single training example

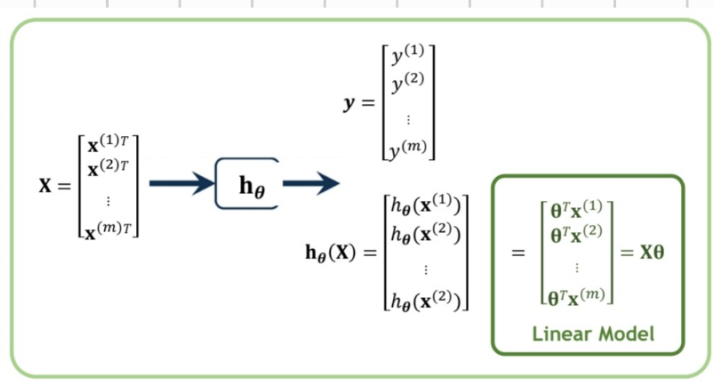
$$h_\theta(\underline{x}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$$

$$= [\theta_0 \ \theta_1 \ \theta_2 \ \dots \ \theta_n] \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \underline{\theta}^T \underline{x}$$

For a batch of training examples

$$h_\theta(\underline{X}) = \begin{bmatrix} \underline{\theta}^T \underline{x}^1 \\ \underline{\theta}^T \underline{x}^2 \\ \vdots \\ \underline{\theta}^T \underline{x}^m \end{bmatrix} = \underline{X} \underline{\theta} \rightarrow \underline{X} = \begin{bmatrix} \underline{x}^1 \\ \underline{x}^2 \\ \vdots \\ \underline{x}^m \end{bmatrix} = \begin{bmatrix} x_0^1 & x_1^1 & x_2^1 & \dots & x_n^1 \\ x_0^2 & x_1^2 & x_2^2 & \dots & x_n^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_0^m & x_1^m & x_2^m & \dots & x_n^m \end{bmatrix}$$

$\underline{a} \underline{b}^T = \underline{a}^T \underline{b}$



MSE cost for Linear Model

classic form:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_\theta(\underline{x}^i) - y^i)^2$$

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

Vector form:

$$J(\theta) = \frac{1}{2m} \|\underline{h}_\theta(\underline{X}) - \underline{y}\|_2^2$$

$$= \frac{1}{2m} \|\underline{X} \underline{\theta} - \underline{y}\|_2^2 = \frac{1}{2m} (\underline{X} \underline{\theta} - \underline{y})^T (\underline{X} \underline{\theta} - \underline{y})$$

Gradient of classic form

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m} \sum_{i=1}^m (\underline{\theta}^T \underline{x}^i - y^i) x_j^i$$

Gradient of vector form

$$\nabla J(\theta) = \frac{1}{m} \sum_{i=1}^m (\underline{\theta}^T \underline{x}^i - y^i) \underline{x}^i$$

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

$$\nabla_{\underline{\theta}} \|\underline{X} \underline{\theta} - \underline{y}\|_2^2$$

$$\|\underline{X} \underline{\theta} - \underline{y}\|_2^2 = (\underline{X} \underline{\theta} - \underline{y})^T (\underline{X} \underline{\theta} - \underline{y})$$

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

$$= \underline{\theta}^T \underline{X}^T \underline{X} \underline{\theta} - \underline{y}^T \underline{X} \underline{\theta} - \underline{\theta}^T \underline{X}^T \underline{y} + \underline{y}^T \underline{y}$$

$$= \underline{\theta}^T \underline{X}^T \underline{X} \underline{\theta} - 2 \underline{y}^T \underline{X} \underline{\theta} + \underline{y}^T \underline{y}$$

$$\nabla_{\underline{\theta}} \|\underline{X} \underline{\theta} - \underline{y}\|_2^2 = \nabla_{\underline{\theta}} (\underline{\theta}^T \underline{X}^T \underline{X} \underline{\theta}) - 2 \nabla_{\underline{\theta}} (\underline{y}^T \underline{X} \underline{\theta})$$

$$= 2 \underline{X}^T \underline{X} \underline{\theta} - 2 \underline{X}^T \underline{y}$$

$$= 2 \underline{X}^T (\underline{X} \underline{\theta} - \underline{y})$$

$$\nabla_{\underline{a}} (\underline{a}^T \underline{A} \underline{a})$$

$$= 2 \underline{A}^T \underline{a}$$

$$\nabla_{\underline{a}} (\underline{b}^T \underline{A} \underline{a})$$

$$= \nabla_{\underline{a}} \underline{c}^T \underline{a} = \underline{c}$$

$$= \underline{A}^T \underline{b}$$

Parameter Update by GD

classic form

$$\theta_j = \theta_j - \alpha \frac{\partial J(\theta)}{\partial \theta_j} = \theta_j - \alpha \frac{1}{m} \sum_{i=1}^m (\theta^T x^i - y^i) x_j^i$$

vector form

$$\underline{\theta} = \underline{\theta} - \alpha \nabla J(\underline{\theta}) = \underline{\theta} - \alpha \frac{1}{m} \underline{X}^T (\underline{X} \underline{\theta} - \underline{y})$$

Gradient Descent vs Normal Equation

$$\nabla J(\underline{\theta}) = \frac{1}{m} \underline{X}^T (\underline{X} \underline{\theta} - \underline{y})$$

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

- | | |
|------------------------------|---------------------------|
| - Need to choose alpha | - No need to choose alpha |
| - Needs many iterations | - No need to iterate |
| - $O(kn^2)$ | - $O(n^3)$ |
| - works well when n is large | - slow if n is very large |

Feature Normalization

Feature scaling

- Feature 간의 size 차이가 크면 cost function 이 특정 feature의 영향을 많이 받게 됨.
- 이것은 모든 feature를 공평하게 고려하려는 의도와 맞지 않음
- 그래서 feature 간의 size를 맞추기 위해 scaling 진행
- 각 feature의 Max 값을 나누어줌
- 즉, 각 feature의 Max 값을 1로 통일해줄 수 있음