[Week-2] Machine Learning

Multivariate Linear Regression & Normal Equation

Joohyung Kang

Contents

Multivariate Linear Regression

- **Multiple Features (Variables)**
  - **Single variable Linear Regression**

Size (feet²) x	Price ($1000) y
2104	460
1416	232
1534	315
852	178
…	…

Hypothesis:

$$h_\theta(x) = \theta_0 + \theta_1 x$$

Parameters:

$$\theta_0, \theta_1$$

Cost Function:

$$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2 = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x_i) - y_i)^2$$

Multivariate Linear Regression

❖ Multiple Features (Variables)

- **Single variable Linear Regression**
  - **Q. 집의 크기만으로 정확한 가격을 책정할 수 있을까?**
    - ✓ **No!** 집의 가치를 판단할 여러 **정보(Feature)**들이 필요!
      - * Ex. Size, Number of bedrooms, Number of floors, Age of home

Size (feet²)	Number of bedrooms	Number of floors	Age of home (years)	Price ($1000)
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
…	…	…	…	…

Multivariate Linear Regression

❖ Multiple Features (Variables)

▪ **Multivariate Linear Regression**

- **여러 개의 Input Feature를 통해 Linear Regression ➔ 선형 모델**
  - ✓ Training Data에 다양한 특성을 나타내는 변수들이 존재
    - * Ex. Size, Number of bedrooms, Number of floors, Age of home

Hypothesis:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \cdots + \theta_n x_n$$

$$= [\theta_0 \;\; \theta_1 \;\; \theta_2 \cdots \theta_n] \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$$= \;\; \boldsymbol{\theta}^{\boldsymbol{T}} \boldsymbol{x}$$

Multivariate Linear Regression

- ❖ **Multiple Features (Variables)**
  - **Multivariate Linear Regression**
    - **Ex. Housing price prediction**
      - ✓ Input Features ➔ Size, No. bedrooms, No. floors, Age of home

$$h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$$

Size (feet²) x_0	Number of bedrooms x_1	Number of floors x_2	Age of home (years) x_3	Price ($1000)
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
...	...	...	...	...

Gradient Descent for Multiple Variables

Multivariate Linear Regression

Multivariate Linear Regression

❖ Gradient Descent for Multiple Variables

▪ Definition of Mathematic

- **Gradient의 특성을 이용해 최적의 파라미터 θ를 구함**
  - ✓ Vector $\boldsymbol{\theta} = \theta_0, \theta_1, \theta_2, \theta_3, \cdots, \theta_n$

Hypothesis:

$$h_\theta(x) = \theta^T x = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

Parameters:

$$\theta$$

Cost Function:

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

$$= \frac{1}{2m}\sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)})^2 = \frac{1}{2m}\sum_{i=1}^{m}((\sum_{j=0}^{n}\theta_j x_j^{(i)}) - y^{(i)})^2$$

Multivariate Linear Regression

❖ Gradient Descent for Multiple Variables

- **Definition of Mathematic**

Cost Function:

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

Gradient Descent: repeat until convergence {

$$\theta_j := \theta_j - \alpha \boxed{\frac{1}{m}\sum_{i=1}^{m}\left(\theta^T x^{(i)} - y^{(i)}\right) \cdot x_j^{(i)}} \qquad for \quad \mathrm{j} := 0 \cdots n$$

}

$$\frac{\partial}{\partial\theta_j} J(\theta_j)$$

Multivariate Linear Regression

❖ Gradient Descent for Multiple Variables

- **Definition of Mathematic**

Repeat until convergence {

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(\theta^T x^{(i)} - y^{(i)} \right) \cdot x_j^{(i)} \qquad for \ \ \mathrm{j} := 0 \cdots n$$

}

Repeat until convergence {

$$x_0^{(i)} = 1$$

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(\theta^T x^{(i)} - y^{(i)} \right) \cdot \boxed{x_0^{(i)}} \quad h_\theta(x) = \boxed{\theta_0 x_0} + \theta_1 x_1 + \cdots + \theta_n x_n$$

$$\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(\theta^T x^{(i)} - y^{(i)} \right) \cdot x_1^{(i)}$$

$$\cdots$$

}

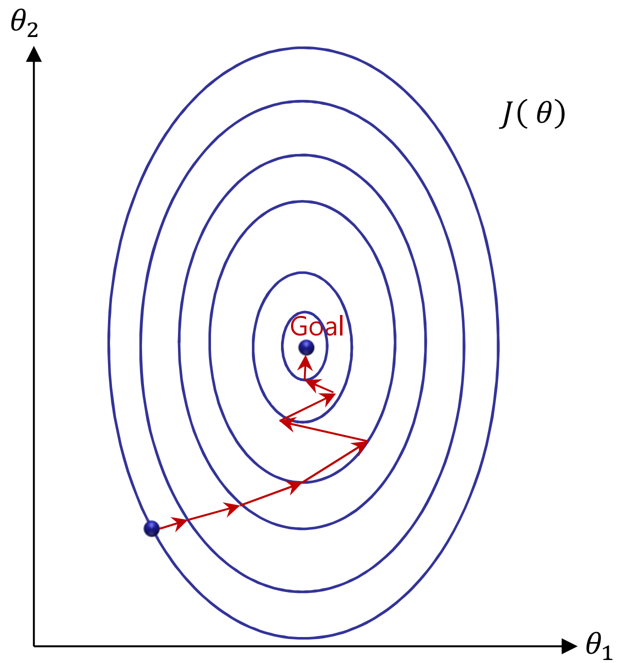
Gradient Descent – Feature Scaling

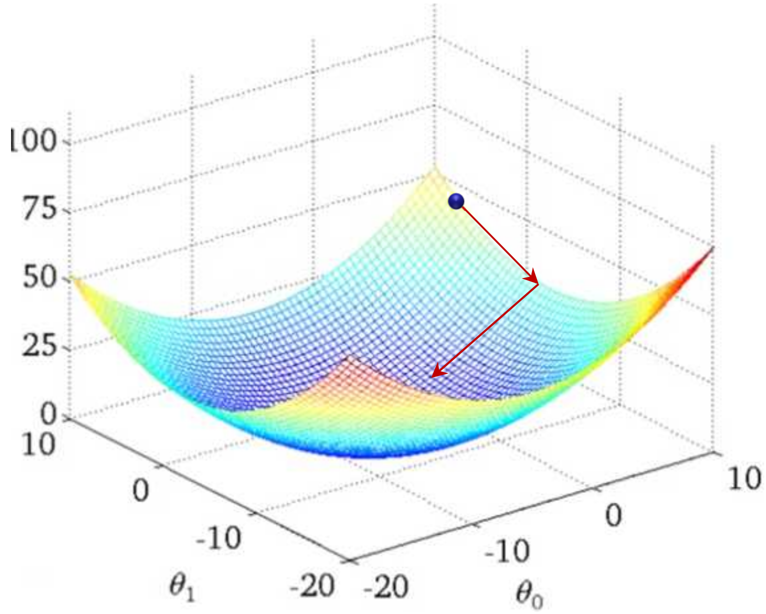
Multivariate Linear Regression

Multivariate Linear Regression

❖ Gradient Descent in Practice – Feature Scaling

- **Multiple Variables**
  - **만약, Feature간 데이터 크기의 차이가 크다면?**
    - ✓ Ex. $x_1 = Range\ (0 \sim 2000)$
      $x_2 = Range\ (1 \sim 5)$

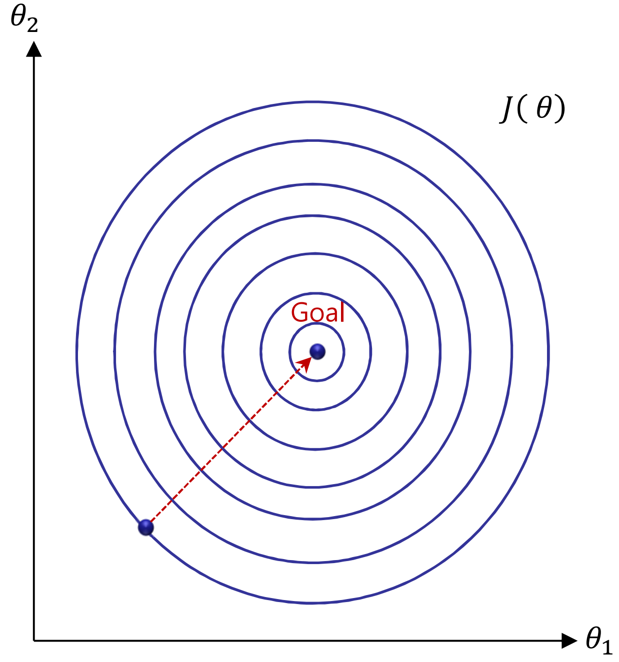


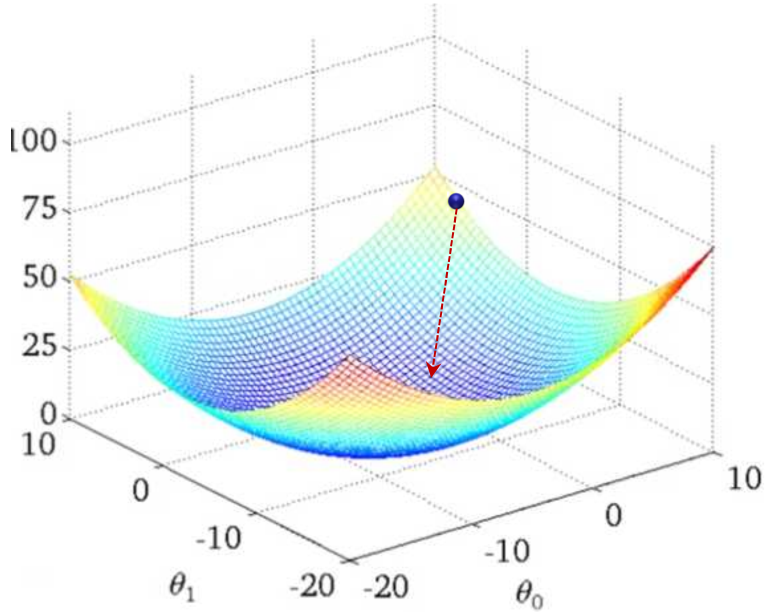


Multivariate Linear Regression

❖ **Gradient Descent in Practice – Feature Scaling**

- **Multiple Variables**
  - **만약, Feature간 데이터 크기의 차이가 비슷하다면?**
    - ✓ Ex. $x_1 = Range\ (0 \sim 1)$
      $x_2 = Range\ (0 \sim 1)$





Multivariate Linear Regression

❖ Gradient Descent in Practice – Feature Scaling

- **Multiple Variables for Gradient Descent**
  - **Feature간 데이터 크기의 차이에 따라서 극소점을 찾는 시간이 달라짐**
    - ✓ Uniform ➔ Short time
    - ✓ Non Uniform ➔ Long time

※ ***따라서, Feature간 데이터의 크기를 Scaling할 필요성이 존재***

- **Scaling Methods**
  - **Get every feature into approximately a $-1 \leq x_i \leq 1$ range.**
    - ✓ Method-1: Feature Scaling

$$x_i := \frac{x_i}{s_i} \qquad for \quad s_i = \max(x) - \min(x)$$

    - ✓ Method-2: Mean Normalization

$$x_i := \frac{x_i - u_i}{s_i} \qquad for \quad s_i = \max(x) - \min(x)$$

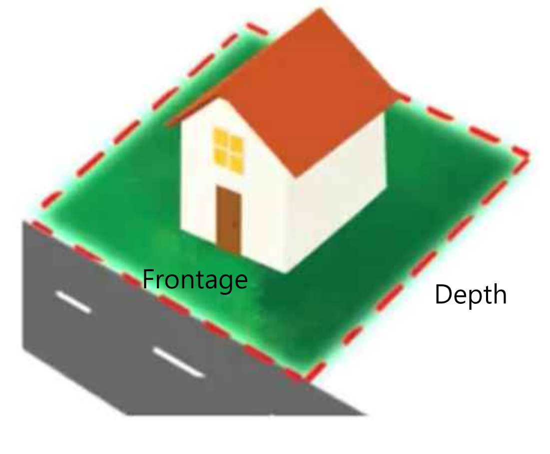
Features and Polynomial Regression

Multivariate Linear Regression

Multivariate Linear Regression

❖ Features

- **Training Data의 적절한 Feature 선택 방법**
  - **선택된 Feature가 불필요하지는 않은가?**
    - ✓ Ex. Housing prices prediction: Features ➔ *Frontage* and *depth*
      x_1 x_2



Hypothesis:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

Features: Frontage, Depth ➔ *Area*

New Feature: Area = Frontage * Depth

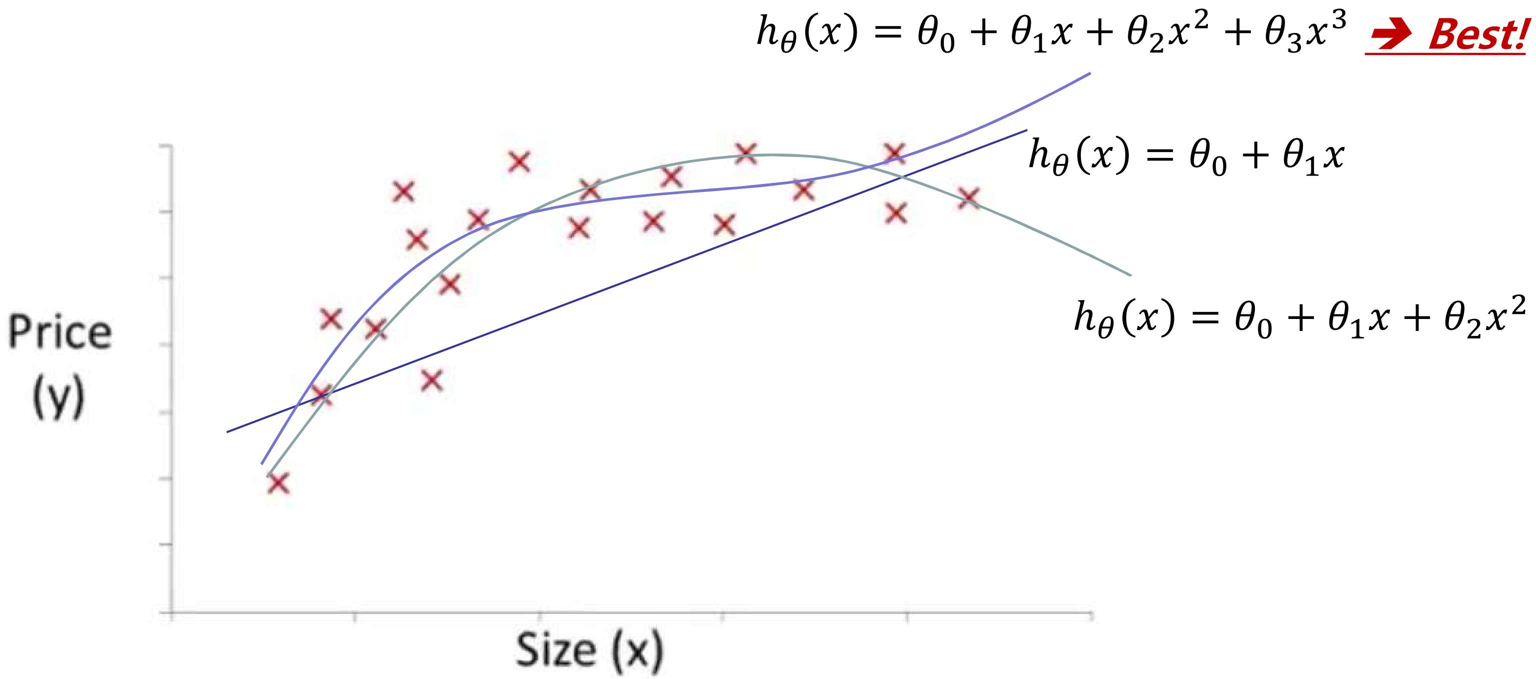
Hypothesis':

$$h_\theta(x)' = \theta_0 + \theta_1 x$$

※ ***즉, 적절한 Feature 선택에 따라 주어진 문제를 간단하게 해결할 수 있음!***

Multivariate Linear Regression

- ❖ **Polynomial Regression**
  - ▪ **Training Data에 대한 적절한 Model 선택**
    - • **Polynomial Regression**
      - ✓ 다항식 형태의 Non-Linear Regression (n차 함수 ex. 2차, 3차 함수)



Multivariate Linear Regression

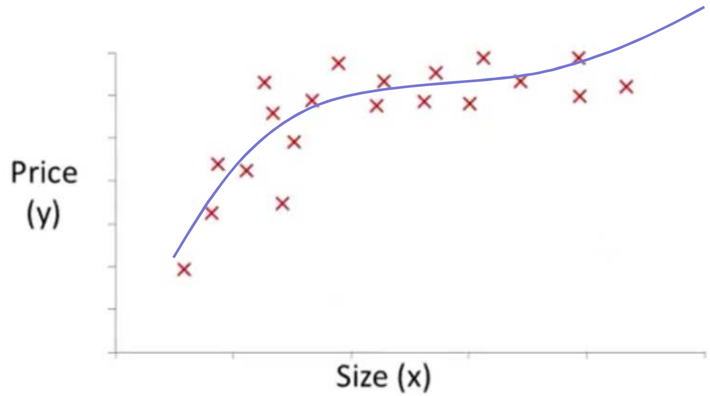
❖ Polynomial Regression

- **Training Data에 대한 적절한 Model 선택**
  - **Polynomial Regression**
    - ✓ Cubic Function: 3차 다항식 함수

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$

$$\begin{cases} x_1 = x \\ x_2 = x^2 \\ x_3 = x^3 \end{cases}$$

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$$



❖ Feature Scaling

$$x_1 = 1 \sim 1{,}000$$

$$x_2 = 1 \sim 1{,}000{,}000$$

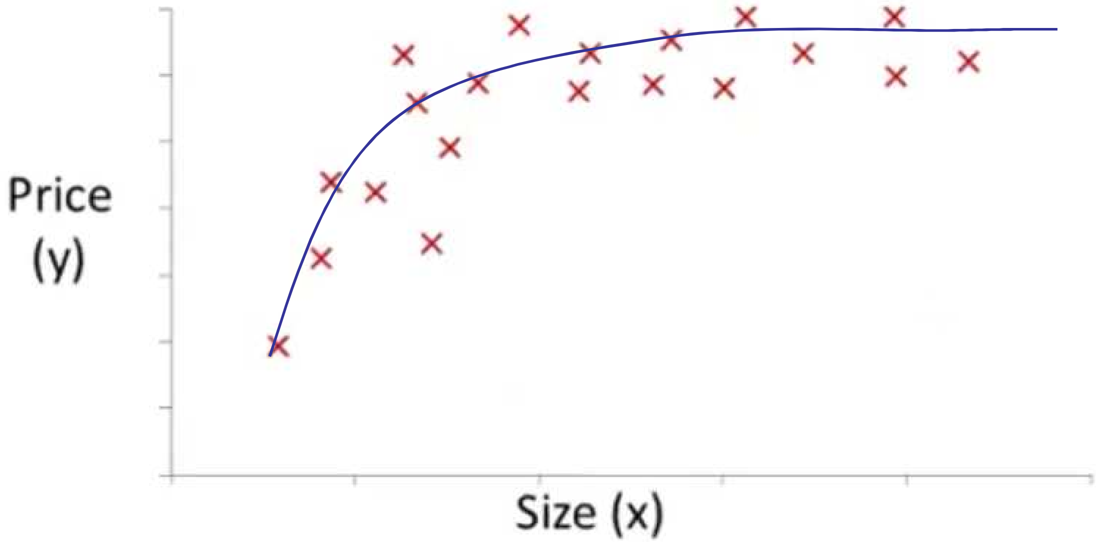
$$x_3 = 1 \sim 1{,}000{,}000{,}000$$

Multivariate Linear Regression

❖ Polynomial Regression

- **Training Data에 대한 적절한 Model 선택**
  - **Polynomial Regression**
    - ✓ Square Root Function

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 \sqrt{x}$$



Normal Equation

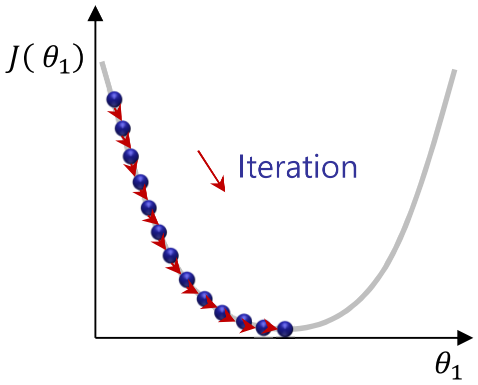
Computing Parameters Analytically

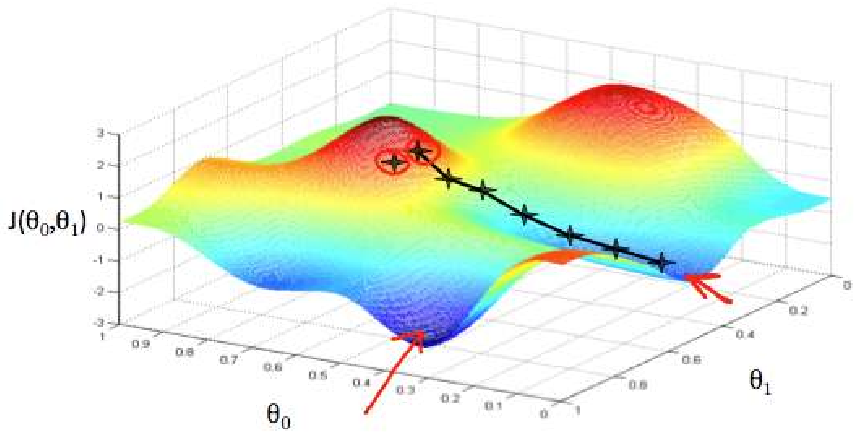
Computing Parameters Analytically

❖ **Normal Equation**

- **Gradient Descent for Cost Function Optimization**
  - **Gradient의 특성을 이용해 최적의 파라미터 θ를 구함**
    - ✓ 파라미터를 반복적으로 갱신하여 Cost가 적은 파라미터를 찾음
      - ➔ **Iteration 발생!**
    - ✓ 또한, 함수의 기울기와 Learning rate에 따라 Iteration 횟수가 가변
    - ✓ **Ex.** ***If Learning rate is too small, gradient descent can be slow.***

※ ***Normal Equation은 최적의 파라미터를 단번에 구할 수 있음!***





Computing Parameters Analytically

❖ Normal Equation

- **Definition of Mathematic**
  - **편미분 방정식을 통하여 파라미터에 대한 최적의 해를 구함**
    - ✓ 즉, 파라미터의 Cost Function $J(\theta)$이 **"0"**이 되는 지점을 구함

Cost Function:

$$J(\theta_0, \theta_1, \theta_2, \cdots, \theta_n) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

Solve for:

$$\frac{\partial}{\partial \theta_j} J(\theta) = 0 \qquad (for \quad every \quad j)$$

Normal Equation:

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

Computing Parameters Analytically

❖ Normal Equation

- **Definition of Mathematic:** ***Least Square Error***
  - Method 1. ➔ Analytic(분석적 방법)
  - Method 2. ➔ Algebraic(대수적 방법)

Computing Parameters Analytically

❖ Normal Equation

- **Definition of Mathematic**
  - **Ex. Housing prices prediction**

Examples: $m = 4$.

Size (feet²) x_1	Number of bedrooms x_2	Number of floors x_3	Age of home (years) x_4	Price ($1000) y
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178

Matrix ➔

$$\underbrace{\begin{bmatrix} 1 & 2104 & 5 & 1 & 45 \\ 1 & 1416 & 3 & 2 & 40 \\ 1 & 1534 & 3 & 2 & 30 \\ 1 & 852 & 2 & 1 & 36 \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{bmatrix}}_{\mathbf{X}} = \underbrace{\begin{bmatrix} 460 \\ 232 \\ 315 \\ 178 \end{bmatrix}}_{\mathbf{B}}$$

(Design Matrix)

$$AX = B$$

$$A^T AX = A^T B$$

$$\underline{X = (A^T A)^{-1} A^T B}$$

Computing Parameters Analytically

❖ Normal Equation

- **Non-Invertible**
  - **Feature matrix ➔ Singular or Degenerate**
    - ✓ Redundant features(Linearly dependent)
    - ✓ Too many features ($m \leq n$)
      - ➔ Delete some feature or use regularization

Computing Parameters Analytically

❖ Normal Equation

▪ Gradient Descent vs. Normal Equation

- **m** training examples, **n** features

Gradient Descent	Normal Equation
• Need to choose learning rate	• No need to choose learning rate
• Needs many iterations	• Don't need to iterate
• Works well even when n is large	• Need to compute $(A^T A)^{-1}$
	• Slow if n is very large