Sid

Multiple features (variables)

n = # of features

$x^{(i)}$ = input features of i^{th} training example - vector

$x_j^{(i)}$ = value of feature j in i^{th} training example

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

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

$$x = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix} \qquad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}$$

$$\begin{bmatrix} \theta_0 & \theta_1 & \cdots & \theta_n \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$$h_\theta(x) = \theta^T \cdot x$$

Hypothesis : $h_\theta(x) = \theta^T \cdot x = \theta_0 \underbrace{x_0}_{x_0 = 1} + \theta_1 x_1 + \cdots + \theta_n x_n$

Parameters : $\vec{\theta}$ ($n+1$ - dimensional vector)

Cost Function : $J(\vec{\theta}) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$

$$= \frac{1}{2m} \sum_{i=1}^{m} \left(\theta^T x^{(i)} - y^{(i)}\right)^2 \quad \text{or} \quad \frac{1}{2m} \sum_{i=1}^{m} \left[\left(\sum_{j=0}^{n} \theta_j x_j^{(i)}\right) - y^{(i)}\right]^2$$

Gradient Descent : $\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\vec{\theta})$

$$: \theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right) x_j^{(i)}$$

Feature Scaling $-1 \le x_i \le 1$, $x_0 = 1$

Mean normalizing : Replace x_i w/ $x_i - \mu_i$

$$x_1 \leftarrow \frac{x_1 - \mu_1}{s_1}$$

$J(\theta)$ should decrease after every iteration

Convergence if $J(\theta)$ decreased by less than 10^{-3} / iteration

Normal Equation : Solve for $\vec{\theta}$ analitically

$$X = \begin{bmatrix} & \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} \end{bmatrix}$$

$m \times (n+1)$ m-dimensional vector

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

```
pinv(X'*X)* X'*y
```

m examples : $(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})$

n features

$$\vec{x}^{(i)} = \begin{bmatrix} x_0^{(i)} \\ x_1^{(i)} \\ \vdots \\ x_n^{(i)} \end{bmatrix} \qquad X = \begin{bmatrix} \text{---} (x^{(1)})^T \text{---} \\ \text{---} (x^{(2)})^T \text{---} \\ \vdots \\ \text{---} (x^{(m)})^T \text{---} \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

Gradient Descent

- choose α
- many iterations
- ✓ for large n

Normal Equation

- no need to choose α
- no iterations
- need to compute $(X^T X)^{-1}$
  $O(n^3)$
- × for large n