§ Machine Learning — Week 2

1. $x_j^{(i)}$: the value of feature j in the i^{th} training example

   $x^{(i)}$: the input (features) of the i^{th} training examples

   m : the number of training examples

   n : the number of features

eg.

$$X = \begin{bmatrix} 1 & 34 & 78 \\ 1 & 30 & 43 \\ \vdots & \vdots & \vdots \\ 1 & 94 & 89 \end{bmatrix}$$

$m \leftarrow$ training examples

n

$x_0 \quad x_1 \quad x_2 \leftarrow$ features

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

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

$$= \begin{bmatrix} \theta_0 & \theta_1 & \theta_2 & \cdots & \theta_n \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix} = \theta^T x$$

2. gradient descent algorithm

① $n = 1$:

Repeat {

$$\theta_0 := \theta_0 - \alpha \underbrace{\frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})}_{\frac{\partial}{\partial \theta_0} J(\theta)}$$

$$\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x^{(i)}$$

}

☆ update θ_0 , θ_1 at the same time

@ $n \geq 1$:

Repeat { ☆ simultaneously update θ_j for $j = 0, \cdots, n$

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

}

$$\Rightarrow \begin{cases} \theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \underline{x_0^{(i)}} = 1 \\ \theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_1^{(i)} \\ \cdots \end{cases}$$

3. feature scaling

$$x_i := \frac{x_i - \mu_i}{s_i}$$

μ_i → average

s_i → {range of value : max − min
standard deviation

4. learning rate

① if α is too small : slow convergence

② if α is too large : $J(\theta)$ may not decrease on every iteration
⇒ may not converge

choose α : try $\cdots, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, \cdots$

3× bigger →

4. normal equation

$$x^{(i)} = \begin{bmatrix} x_0^{(i)} \\ x_1^{(i)} \\ x_2^{(i)} \end{bmatrix} \in \mathbb{R}^{n+1} \Rightarrow X = \begin{bmatrix} \cdots (x^{(1)})^T \cdots \\ \cdots (x^{(2)})^T \cdots \\ \vdots \end{bmatrix}$$

↙ design matrix

$$\begin{bmatrix} \vdots \\ x_n^{(i)} \end{bmatrix}$$

$$\underbrace{\begin{bmatrix} \vdots \\ \cdots \; (x^{(m)})^T \; \cdots \end{bmatrix}}_{m \times (n+1)}$$

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

normal equations (no feature scaling)

5. Compare gradient descent with normal equation

gradient descent:

① need to chose α

② need many iterations

③ $O(kn^2)$

④ Works well when n is large

normal equation

① don't need to chose α

② don't need to iterate

③ $O(n^3)$, need to calculate inverse of $X^T X$

④ slow when n is very large

if $X^T X$ is noninvertible:

① Redundant features, where two features are very closely related. (eg. linearly dependent)

② Too many features. (eg. $m \leq n$) → delete some features

6. Some more information about the normal equation (optional)

(1) matrix derivatives:

① $\nabla_A f(A) = \begin{bmatrix} \frac{\partial f}{\partial A_{11}} & \cdots & \frac{\partial f}{\partial A_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial A_{m1}} & \cdots & \frac{\partial f}{\partial A_{mn}} \end{bmatrix}$

eg $f(A) = \frac{3}{2}A_{11} + 5A_{12}^2 + A_{21}A_{22}$

$\Rightarrow \nabla_A f(A) = \begin{bmatrix} \frac{3}{2} & 10A_{12} \\ A_{22} & A_{21} \end{bmatrix}$

② $trA = \sum_{i=1}^{n} A_{ii}$ ⟶ trace operator

- $trAB = trBA$
- $trABC = trCAB = trBCA$
- $trABCD = trDABC = trCDAB = trBCDA$
- $trA = trA^T$
- $tr(A+B) = trA + trB$
- $tr\,aA = a\,trA$
- $\nabla_A trAB = B^T$
- $\nabla_A trAB A^T C = CBA + C^T AB^T$
- $\nabla_A |A| = |A|(A^{-1})^T$

(2) normal equation

$$X = \begin{bmatrix} -(x^{(1)})^T- \\ -(x^{(2)})^T- \\ \vdots \\ -(x^{(m)})^T- \end{bmatrix} \qquad y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

$\Rightarrow X\theta - \vec{y} = \begin{bmatrix} (x^{(1)})^T\theta \\ \vdots \\ (x^{(m)})^T\theta \end{bmatrix} - \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix} = \begin{bmatrix} h_\theta(x^{(1)}) - y^{(1)} \\ \vdots \\ h_\theta(x^{(m)}) - y^{(m)} \end{bmatrix}$

$$X\theta - \vec{y} = \begin{bmatrix} (x^{(1)})^T\theta \\ (x^{(2)})^T\theta \\ \vdots \\ (x^{(m)})^T\theta \end{bmatrix} - \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix} = \begin{bmatrix} h_\theta(x^{(1)}) - y^{(1)} \\ h_\theta(x^{(2)}) - y^{(2)} \\ \vdots \\ h_\theta(x^{(m)}) - y^{(m)} \end{bmatrix}$$

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

$\because \nabla_A^T \operatorname{tr} ABA^TC = B^TA^TC^T + BA^TC$

$\therefore \nabla_\theta J(\theta) = \nabla_\theta \frac{1}{2}(X\theta - \vec{y})^T(X\theta - \vec{y}) = X^TX\theta - X^T\vec{y}$

$\because$ minimize $J(\theta)$

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

$\Rightarrow \underbrace{X^TX\theta = X^T\vec{y}}_{\text{normal equations}} \quad \Rightarrow \theta = (X^TX)^{-1}X^T\vec{y}$