COMP2261 ARTIFICIAL INTELLIGENCE / MACHINE LEARNING

Multivariate Linear Regression

-- Normal Equation

Dr SHI Lei

Durham University

learning lab

Normal Equation Intuition

Normal Equation Intuition

Cost Function

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

To find $\theta_0, \theta_1, \dots \theta_n$ that minimise J

Normal Equation Intuition

Gradient Descent

Repeat until $\boxed{\text{convergence}}$ {

$$\theta_0 := \theta_0 - \alpha \frac{\partial J}{\partial \theta_0}(\boldsymbol{\theta})$$

$$\theta_1 := \theta_1 - \alpha \frac{\partial J}{\partial \theta_1}(\boldsymbol{\theta})$$

$$\dots$$

$$\theta_n := \theta_1 - \alpha \frac{\partial J}{\partial \theta_n}(\boldsymbol{\theta})$$

} simultaneously update $\theta_0, \dots, \theta_n$

Normal Equation Intuition

Gradient Descent

Repeat until convergence {

$$\theta_0 := \theta_0 - \alpha \frac{\partial J}{\partial \theta_0}(\boldsymbol{\theta})$$

$$\theta_1 := \theta_1 - \alpha \frac{\partial J}{\partial \theta_1}(\boldsymbol{\theta})$$

$$\cdots$$

$$\theta_n := \theta_1 - \alpha \frac{\partial J}{\partial \theta_n}(\boldsymbol{\theta})$$

learning rate

partial derivative

}

Normal Equation Intuition

Univariate Linear Regression

Cost Function

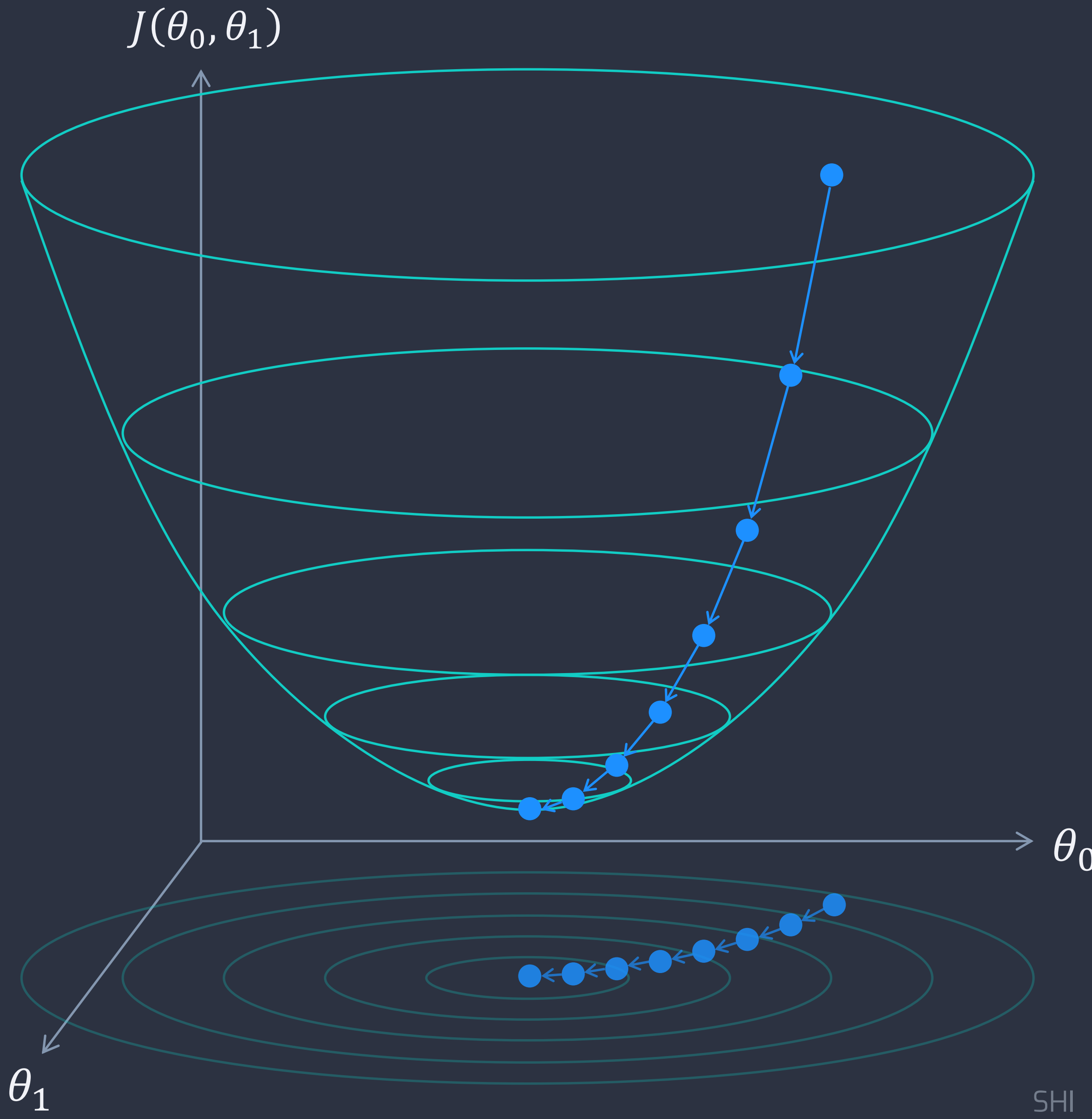
$$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

Normal Equation Intuition

Univariate Linear Regression

Cost Function

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



Normal Equation Intuition

Univariate Linear Regression

Cost Function

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

$$\frac{\partial}{\partial\theta_0}J(\theta_0,\theta_1) = \frac{1}{m}\sum_{i=1}^{m}(\theta_0 + \theta_1 \cdot x^{(i)} - y^{(i)}) = 0$$

$$\frac{\partial}{\partial\theta_1}J(\theta_0,\theta_1) = \frac{1}{m}\sum_{i=1}^{m}(\theta_0 + \theta_1 \cdot x^{(i)} - y^{(i)}) \cdot x^{(i)} = 0$$

Normal Equation Intuition

Univariate Linear Regression

Cost Function

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

$$\begin{cases} \frac{1}{m}\sum_{i=1}^{m}\left(\theta_0 + \theta_1 \cdot x^{(i)} - y^{(i)}\right) = 0 \\ \frac{1}{m}\sum_{i=1}^{m}\left(\theta_0 + \theta_1 \cdot x^{(i)} - y^{(i)}\right) \cdot x^{(i)} = 0 \end{cases} \Longrightarrow \begin{cases} \theta_0 = ? \\ \theta_1 = ? \end{cases}$$

Normal Equation Intuition

Univariate Linear Regression

Cost Function

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

$$\begin{cases} \cancel{\frac{1}{m}}\sum_{i=1}^{m}\left(\theta_0 + \theta_1 \cdot x^{(i)} - y^{(i)}\right) = 0 \\ \cancel{\frac{1}{m}}\sum_{i=1}^{m}\left(\theta_0 + \theta_1 \cdot x^{(i)} - y^{(i)}\right) \cdot x^{(i)} = 0 \end{cases} \Longrightarrow \begin{cases} \sum_{i=1}^{m}\left(\theta_0 + \theta_1 \cdot x^{(i)} - y^{(i)}\right) = 0 \\ \sum_{i=1}^{m}\left(\theta_0 + \theta_1 \cdot x^{(i)} - y^{(i)}\right) \cdot x^{(i)} = 0 \end{cases}$$

Normal Equation Intuition

Univariate Linear Regression

Cost Function

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

$$\left\{ \begin{array}{l} \boxed{\sum_{i=1}^{m}} \left(\boxed{\theta_0} + \theta_1 \cdot x^{(i)} - y^{(i)} \right) = 0 \\ \\ \sum_{i=1}^{m} \left(\theta_0 + \theta_1 \cdot x^{(i)} - y^{(i)} \right) \cdot x^{(i)} = 0 \end{array} \right. \Longrightarrow \left\{ \begin{array}{l} \boxed{\theta_0 m} + \theta_1 \sum_{i=1}^{m} x^{(i)} - \sum_{i=1}^{m} y^{(i)} = 0 \\ \\ \theta_0 \sum_{i=1}^{m} x^{(i)} + \theta_1 \sum_{i=1}^{m} {x^{(i)}}^2 - \sum_{i=1}^{m} x^{(i)} y^{(i)} = 0 \end{array} \right.$$

Normal Equation Intuition

Univariate Linear Regression

Cost Function

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

$$\begin{cases} \boxed{\sum_{i=1}^{m}} \left(\theta_0 + \boxed{\theta_1 \cdot x^{(i)}} - y^{(i)}\right) = 0 \\ \sum_{i=1}^{m} \left(\theta_0 + \theta_1 \cdot x^{(i)} - y^{(i)}\right) \cdot x^{(i)} = 0 \end{cases} \Longrightarrow \begin{cases} \theta_0 m + \boxed{\theta_1 \sum_{i=1}^{m} x^{(i)}} - \sum_{i=1}^{m} y^{(i)} = 0 \\ \theta_0 \sum_{i=1}^{m} x^{(i)} + \theta_1 \sum_{i=1}^{m} {x^{(i)}}^2 - \sum_{i=1}^{m} x^{(i)} y^{(i)} = 0 \end{cases}$$

Normal Equation Intuition

Univariate Linear Regression

Cost Function

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

$$\begin{cases} \boxed{\sum_{i=1}^{m}}\left(\theta_0 + \theta_1 \cdot x^{(i)} - \boxed{y^{(i)}}\right) = 0 \\ \sum_{i=1}^{m}\left(\theta_0 + \theta_1 \cdot x^{(i)} - y^{(i)}\right) \cdot x^{(i)} = 0 \end{cases} \Longrightarrow \begin{cases} \theta_0 m + \theta_1 \sum_{i=1}^{m} x^{(i)} - \boxed{\sum_{i=1}^{m} y^{(i)}} = 0 \\ \theta_0 \sum_{i=1}^{m} x^{(i)} + \theta_1 \sum_{i=1}^{m} {x^{(i)}}^2 - \sum_{i=1}^{m} x^{(i)} y^{(i)} = 0 \end{cases}$$

Normal Equation Intuition

Univariate Linear Regression

Cost Function

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

$$\left\{\begin{aligned} &\theta_0 m + \theta_1 \boxed{\sum_{i=1}^{m}} x^{(i)} - \boxed{\sum_{i=1}^{m}} y^{(i)} = 0 \\ &\theta_0 \boxed{\sum_{i=1}^{m}} x^{(i)} + \theta_1 \boxed{\sum_{i=1}^{m}} {x^{(i)}}^2 - \boxed{\sum_{i=1}^{m}} x^{(i)} y^{(i)} = 0 \end{aligned}\right. \Longrightarrow \left\{\begin{aligned} &\theta_0 m + \theta_1 \boxed{\Sigma} x^{(i)} - \boxed{\Sigma} y^{(i)} = 0 \\ &\theta_0 \boxed{\Sigma} x^{(i)} + \theta_1 \boxed{\Sigma} {x^{(i)}}^2 - \boxed{\Sigma} x^{(i)} y^{(i)} = 0 \end{aligned}\right.$$

Normal Equation Intuition

Univariate Linear Regression

Cost Function

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

$$\begin{cases} \theta_0 m + \theta_1 \Sigma x^{(i)} - \boxed{\Sigma y^{(i)}} = 0 \\ \theta_0 \Sigma x^{(i)} + \theta_1 \Sigma {x^{(i)}}^2 - \boxed{\Sigma x^{(i)} y^{(i)}} = 0 \end{cases} \Longrightarrow \begin{cases} \theta_0 m + \theta_1 \Sigma x^{(i)} = \boxed{\Sigma y^{(i)}} \\ \theta_0 \Sigma x^{(i)} + \theta_1 \Sigma {x^{(i)}}^2 = \boxed{\Sigma x^{(i)} y^{(i)}} \end{cases}$$

Normal Equation Intuition

Univariate Linear Regression

Cost Function

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

$$\begin{cases} \theta_0 \boxed{m} + \theta_1 \boxed{\Sigma x^{(i)}} = \boxed{\Sigma y^{(i)}} \\ \theta_0 \boxed{\Sigma x^{(i)}} + \theta_1 \boxed{\Sigma {x^{(i)}}^2} = \boxed{\Sigma x^{(i)} y^{(i)}} \end{cases} \quad \overset{\text{Cramer's Rule}}{\Longrightarrow} \quad \begin{bmatrix} \boxed{m} & \boxed{\Sigma x^{(i)}} \\ \boxed{\Sigma x^{(i)}} & \boxed{\Sigma {x^{(i)}}^2} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} = \begin{bmatrix} \boxed{\Sigma y^{(i)}} \\ \boxed{\Sigma x^{(i)} y^{(i)}} \end{bmatrix}$$

coefficient matrix variable vector

Normal Equation Intuition

Univariate Linear Regression

Cost Function

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

$$\begin{bmatrix} m & \Sigma x^{(i)} \\ \Sigma x^{(i)} & \Sigma {x^{(i)}}^2 \end{bmatrix}\begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} = \begin{bmatrix} \Sigma y^{(i)} \\ \Sigma x^{(i)} y^{(i)} \end{bmatrix} \Longrightarrow$$

$$\theta_0 = \frac{\begin{vmatrix} \Sigma y^{(i)} & \Sigma x^{(i)} \\ \Sigma x^{(i)} y^{(i)} & \Sigma {x^{(i)}}^2 \end{vmatrix}}{\begin{vmatrix} m & \Sigma x^{(i)} \\ \Sigma x^{(i)} & \Sigma {x^{(i)}}^2 \end{vmatrix}} = \frac{\Sigma y^{(i)} \Sigma {x^{(i)}}^2 - \Sigma x^{(i)} \Sigma x^{(i)} y^{(i)}}{m \Sigma {x^{(i)}}^2 - (\Sigma x^{(i)})^2}$$

$$\theta_1 = \frac{\begin{vmatrix} m & \Sigma y^{(i)} \\ \Sigma x^{(i)} & \Sigma x^{(i)} y^{(i)} \end{vmatrix}}{\begin{vmatrix} m & \Sigma x^{(i)} \\ \Sigma x^{(i)} & \Sigma {x^{(i)}}^2 \end{vmatrix}} = \frac{m \Sigma x^{(i)} y^{(i)} - \Sigma y^{(i)} \Sigma x^{(i)}}{m \Sigma {x^{(i)}}^2 - (\Sigma x^{(i)})^2}$$

Cramer's Rule

Normal Equation

for Multivariate Linear Regression

Normal Equation for Multivariate Linear Regression

Cost Function

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

To find $\theta_0, \theta_1, \dots \theta_n$ that minimise J

Normal Equation for Multivariate Linear Regression

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

$$\frac{\partial J}{\partial \theta_0}(\boldsymbol{\theta}) = 0$$

$$\frac{\partial J}{\partial \theta_1}(\boldsymbol{\theta}) = 0$$

$$\dots$$

$$\frac{\partial J}{\partial \theta_n}(\boldsymbol{\theta}) = 0$$

Normal Equation for Multivariate Linear Regression

Vectorised partial derivative

$$\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{\theta}} = \boldsymbol{X}^T \left(\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y} \right) = 0$$

Normal Equation for Multivariate Linear Regression

$$\frac{\partial J}{\partial \boldsymbol{\theta}} = \boldsymbol{X}^T \left(\boldsymbol{X\theta} - \boldsymbol{y}\right) = 0$$

$$\boldsymbol{X}^T\boldsymbol{X\theta} - \boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{0}$$

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

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

Normal Equation for Multivariate Linear Regression

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

feature values	labels

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_n^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_n^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(m)} & x_2^{(m)} & \cdots & x_n^{(m)} \end{bmatrix} \qquad \boldsymbol{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

Normal Equation vs Gradient Descent

Normal Equation vs Gradient Descent

Normal Equation	Gradient Descent
Non-iterative	Multiple iterations
No learning rate	Experimental learning rate
Doesn’t need feature scaling	Needs feature scaling
Slow if training set is very large $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ $O(n^3)$	Works well with very large training set
Doesn’t work for many cases	Works with other types of tasks too.