Linear Regression with Ordinary Least Square Method

Zengchang Qin (PhD)

2018年4月15日
0:00

LEAST SQUARE

1. Given a training data $D = \{(x^{(1)}, y^{(1)}), \cdots (x^{(N)}, y^{(N)})\}$
   If the model we hope to learn is a linear model:

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

The function we need to learn can map $x^{(i)} \to y^{(i)}$, we need to find the best parameter θ to satisfy the following equations:

$$\begin{cases} \theta_0 + \theta_1 x_1^{(1)} + \theta_2 x_2^{(1)} + \cdots + \theta_n x_n^{(1)} = y^{(1)} \\ \theta_0 + \theta_1 x_1^{(2)} + \theta_2 x_2^{(2)} + \cdots + \theta_n x_n^{(2)} = y^{(2)} \\ \vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \\ \theta_0 + \theta_1 x_1^{(N)} + \theta_2 x_2^{(N)} + \cdots + \theta_n x_n^{(N)} = y^{(N)} \end{cases}$$

They can be written into the form of matrices:

$$N \left\{ \underbrace{\begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_n^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_n^{(2)} \\ \vdots & \vdots & & & \\ 1 & x_1^{(N)} & x_2^{(N)} & \cdots & x_n^{(N)} \end{bmatrix}}_{n+1} \right. \cdot \left. \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix} \right\} n+1 \approx \left. \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{bmatrix} \right\} N$$

In order to find the best approximation, we can solve the following optimization problem:

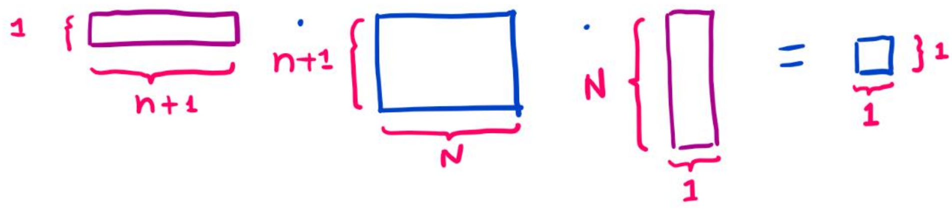
$$X\theta \approx Y \quad \Rightarrow \quad \min \|X\theta - Y\|_2^2 \text{ (square of } L_2 \text{ norm)}$$

To solove that, we set:

$$S = \|X\theta - Y\|_2^2 = 0$$

$$\begin{aligned}
\|X\theta - Y\|_2^2 &= (X\theta - Y)^T (X\theta - Y) \\
&= (\theta^T X^T - Y^T)(X\theta - Y) \\
&= \theta^T X^T X\theta - \theta^T X^T Y - Y^T X\theta + Y^T Y \\
&= \theta^T X^T X\theta - 2\theta^T X^T Y + Y^T Y
\end{aligned}$$

Because $\theta^T X^T Y$ is a scalar, so does $Y^T X\theta$



$$\frac{\partial S}{\partial \theta} = 0 \Rightarrow \frac{\partial(\theta^T X^T X\theta)}{\partial \theta} - 2X^T Y = 0$$

Since $\dfrac{d(u^T v)}{dx} = \dfrac{du^T}{dx} v + \dfrac{dv^T}{dx} u$, then:

$$\frac{d(\theta^T \theta)}{d\theta} = \frac{d\theta^T}{d\theta}\theta + \frac{d\theta^T}{d\theta}\theta = 2\theta$$

Given A is a square matrix

$$\begin{aligned}
\frac{d(\theta^T A\theta)}{d\theta} &= \frac{d\theta^T}{d\theta} A\theta + \frac{d(\theta^T A^T)}{d\theta}\theta \\
&= A\theta + A^T\theta = (A + A^T)\theta
\end{aligned}$$

$X^T X$ is $(n+1)\times(n+1)$ square matrix

$$\frac{d(\theta^T X^T X\theta)}{d\theta} = X^T X\theta + (X^T X)^T\theta = 2X^T X\theta$$

$$\begin{aligned}
\frac{\partial S}{\partial \theta} = 0 &\Rightarrow 2X^T X\theta - 2X^T Y = 0 \\
&\Rightarrow X^T X\theta = X^T Y \\
&\Rightarrow \theta = (X^T X)^{-1} X^T Y
\end{aligned}$$