

Linear Regression 01

Chen Gong

12 October 2019

数据集 $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, 其中 $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$, $i = 1, 2, \dots, N$ 。
数据矩阵为: (这样可以保证每一行为一个数据点)

$$X = (x_1, x_2, \dots, x_N)^T = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{Np} \end{pmatrix}_{N \times P} \quad (1)$$

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}_{N \times 1} \quad (2)$$

设拟合的函数为: $f(w) = W^T x$

1 最小二乘估计: 矩阵表示

很简单可以得到损失函数 (Loss function) 为:

$$L(w) = \sum_{i=1}^N \|w^T x_i - y_i\|^2 \quad (3)$$

$$= (w^T x_1 - y_1, w^T x_2 - y_2, \dots, w^T x_N - y_N) \begin{pmatrix} w^T x_1 - y_1 \\ w^T x_2 - y_2 \\ \vdots \\ w^T x_N - y_N \end{pmatrix} \quad (4)$$

其中:

$$\begin{aligned} (w^T x_1 - y_1, w^T x_2 - y_2, \dots, w^T x_N - y_N) &= [(w^T x_1, w^T x_2, \dots, w^T x_N) - (y_1, y_2, \dots, y_N)] \\ &= W^T X^T - Y^T \end{aligned} \quad (5)$$

所以:

$$L(w) = (W^T X^T - Y^T)(W^T X^T - Y^T)^T \quad (6)$$

$$= (W^T X^T - Y^T)(XW - Y)$$

$$= W^T X^T X - W^T X^T Y - Y^T XW + Y^T Y$$

$$= W^T X^T X - 2W^T X^T Y + Y^T Y$$

(7)

那么我要求的 w , 可记为 $\hat{w} = \operatorname{argmin}_w L(w)$ 。求得这个函数的方法可以使用对 w 求偏导的方法, 那么有:

$$\frac{\partial L(w)}{\partial w} = 2X^T XW - 2X^T Y = 0 \quad (8)$$

解得:

$$W = (X^T X)^{-1} X^T Y \quad (9)$$

2 最小二乘估计: 几何意义

将 X 矩阵从列向量的角度来看, 可以看成是一个 p 维的向量空间 S , 为了简便计算, 令 $W^T X = X\beta$ 。可以看成 Y 向量到 S 的距离最短, 那么将有约束条件:

$$X^T(Y - X\beta) = 0 \quad (10)$$

$$X^T Y - X^T X\beta = 0 \quad (11)$$

$$\beta = (X^T X)^{-1} X^T Y \quad (12)$$

3 最小二乘估计: 概率角度

假设一个分布 $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, 那么所有的观测值可看为 $y = w^T x + \varepsilon$ 。因为 $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, 那么 $p(y|x; w) \sim \mathcal{N}(w^T x, \sigma^2)$ 。我们的目的是求 w 使, y 出现的概率最大, 在这里可以使用极大似然估计 (MLE) 求解。首先写出 $p(y|x; w)$ 的概率密度函数为:

$$p(y|x; w) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y - w^T x)^2}{2\sigma^2}\right) \quad (13)$$

似然函数为 $\ln p(y|x; w)$, 使似然函数最大化的过程求解如下:

$$L(w) = \ln p(y|x; w) = \ln \prod_{i=1}^N p(y_i|x_i; w) \quad (14)$$

$$= \sum_{i=1}^N \ln p(y_i|x_i; w) \quad (15)$$

$$= \sum_{i=1}^N \left(\ln \frac{1}{\sqrt{2\pi}\sigma} + \ln \exp\left(-\frac{(y_i - w^T x)^2}{2\sigma^2}\right) \right) \quad (16)$$

求解目标为 $\hat{w} = \operatorname{argmax}_w L(w)$ ，因为第一项其中并没有包含 w ，于是可以直接省略，那么有：

$$\hat{w} = \operatorname{argmax}_w L(w) \quad (17)$$

$$\begin{aligned} &= \operatorname{argmax}_w \sum_{i=1}^N -\frac{(y_i - w^T x_i)^2}{2\sigma^2} \\ &= \operatorname{argmin}_w \sum_{i=1}^N (y_i - w^T x_i)^2 \end{aligned} \quad (18)$$

然后惊奇的发现后面的求解过程，和最小二乘法的矩阵表示方法的求解过程是一模一样的。那么我可以得到一个结论：最小二乘估计 \iff 极大似然估计 (噪声符合高斯分布)。那么我们的最小二乘估计中隐藏了一个假设条件，那就是噪声符合高斯分布。