Linear Regression 01

Chen Gong

12 October 2019

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

$$X=(x_1,x_2,\cdots,x_N)^T=\begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{pmatrix}=\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{32} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{Np} \end{pmatrix}_{N\times p} \tag{0.1}$$

$$Y=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}_{N\times 1} \tag{0.2}$$

设拟合的函数为：$f(w)=w^Tx$。

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：

$$L(w)=\sum_{i=1}^{N}||w^Tx_i-y_i||^2 \tag{1.1}$$

$$=(w^Tx_1-y_1,w^Tx_2-y_2,\dots,w^Tx_N-y_N)\begin{pmatrix} w^Tx_1-y_1 \\ w^Tx_2-y_2 \\ \vdots \\ w^Tx_N-y_N \end{pmatrix} \tag{1.2}$$

其中:

$$\begin{aligned}(w^Tx_1-y_1,w^Tx_2-y_2,\dots,w^Tx_N-y_N)=&[(w^Tx_1,w^Tx_2,\cdots,w^Tx_N)-(y_1,y_2,\cdots,y_N)] \\ =&w^TX^T-Y^T\end{aligned} \tag{1.3}$$

所以:

$$\begin{aligned}L(w)=&(Xw-Y)^T(Xw-Y) \\ =&w^TX^TXw-2w^TX^TY+Y^TY\end{aligned} \tag{1.4}$$

由于 X^TX 是一个半正定矩阵,$L(w)$ 是一个凸函数,那么我需要求的 w,可记为 $\hat{w} = \arg\min_w \ L(w)$。这是一个无约束优化问题，可以通过求偏导解决。那么有:

$$\frac{\partial L(w)}{\partial w} = 2X^TXw - 2X^TY = 0 \tag{1.5}$$

解得:

$$\hat{w} = (X^TX)^{-1}X^TY \tag{1.6}$$

2 最小二乘估计：几何意义

将 X 矩阵从列向量的角度来看,可以看成一个 p 维的向量空间 S,为了简便计算,令 $w^TX = X\beta$。可以看成 Y 向量到 S 的距离最短，那么将有约束条件:

$$X^T(Y - X\beta) = 0 \tag{2.1}$$

$$X^TY - X^TX\beta = 0 \tag{2.2}$$

$$\beta = (X^TX)^{-1}X^TY \tag{2.3}$$

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 $y = w^Tx + \varepsilon$。因为 $\varepsilon \sim \mathcal{N}(0, \sigma^2)$，那么 $p(y|x;w) \sim \mathcal{N}(w^Tx, \sigma^2)$。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:

$$p(y|x;w) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(y - w^Tx)^2}{2\sigma^2}\right) \tag{3.1}$$

对数似然函数为 $\log \ p(Y|X;w)$，使似然函数最大化的过程求解如下:

$$L(w) = \log \ p(Y|X;w) = \log \prod_{i=1}^{N} p(y_i|x_i;w) \tag{3.2}$$

$$= \sum_{i=1}^{N} \log \ p(y_i|x_i;w) \tag{3.3}$$

$$= \sum_{i=1}^{N} \left(\log \frac{1}{\sqrt{2\pi}\sigma} + \log \ exp\left(-\frac{(y_i - w^Tx)^2}{2\sigma^2}\right)\right) \tag{3.4}$$

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

$$\hat{w} = \arg\max_w \ L(w) \tag{3.5}$$

$$\begin{aligned} &= \arg\max_w \ \sum_{i=1}^{N} -\frac{(y_i - w^Tx_i)^2}{2\sigma^2} \\ &= \arg\min_w \ \sum_{i=1}^{N} (y_i - w^Tx_i)^2 \end{aligned} \tag{3.6}$$

那么我可以可以得到一个结论：最小二乘估计 $\Longleftrightarrow$ 极大似然估计 (噪声符合高斯分布)。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。