

Linear Regression 02

Chen Gong

14 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: $f(w) = W^T x$, 根据上一节我们推导出的结果, 损失函数定义为:

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

解出, $\hat{w} = (X^T X)^{-1} X^T Y$

1 正则化概述

过拟合问题 (over-fitting) 问题是深度学习中一个很重要的问题, 往往是由少量的数据拟合高维的向量所造成的。解决 over-fitting 的方法有很多, 通常是使用这几种思路: 1. 增加数据量; 2. 特征选择/特征提取 (PCA); 3. 增加正则项的方法。

正则项通常可以描述为 Loss Function + Penalty, 也就是 $L(w) + \lambda P(w)$ 。正则化的方法通常有以下两种:

1. Lasso, 其中 $P(w) = \|w\|_1 = \sum_{i=1}^N w_i$
2. Ridge, 岭回归, 也就是 $P(w) = \|w\|_2^2 = \sum_{i=1}^N w_i^2$

2 岭回归频率派角度

Loss function 可写为 $L(w) = \sum_{i=1}^N \|w^T x_i - y_i\|^2 + \lambda W^T W$

$$\begin{aligned}
 J(w) &= \sum_{i=1}^N \|w^T x_i - y_i\|^2 + \lambda W^T W \\
 &= (W^T X^T - Y^T)(XW - Y) + \lambda W^T W \\
 &= W^T X^T XW - W^T X^T Y - Y^T XW - Y^T Y + \lambda W^T W \\
 &= W^T X^T XW - 2W^T X^T Y - Y^T Y + \lambda W^T W \\
 &= W^T (X^T X + \lambda I)W - 2W^T X^T Y - Y^T Y
 \end{aligned} \tag{4}$$

我们的求解目标是 $\hat{w} = \operatorname{argmin}_w J(w)$, 求解过程为:

$$\frac{\partial J(w)}{\partial w} = 2(X^T X + \lambda I)W - 2X^T Y = 0 \tag{5}$$

解得:

$$W = (X^T X + \lambda I)^{-1} X^T Y \tag{6}$$

根据以上的推导我们可以得出, 首先 $(X^T X + \lambda I)$ 一定是可逆的。因为, 半正定矩阵 + 单位矩阵 = 正定矩阵。这里不需要再求伪逆了。

3 岭回归贝叶斯派估计角度

类似于前文提到的贝叶斯回归的角度, 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 $p(y|x; w) \sim \mathcal{N}(w^T x, \sigma^2)$ 。假设 w 符合一个先验分布 $\mathcal{N}(0, \sigma_0^2)$ 。于是, 我们可以得到 $p(w)$ 和 $p(y|w)$ 的解析表达式:

$$p(y|w) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y - w^T x)^2}{2\sigma^2}\right) \tag{7}$$

$$p(w) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{\|w\|^2}{2\sigma_0^2}\right) \tag{8}$$

我们的目标是求 w 的最大后验估计 (MAP), 也就是定义为求 $\hat{w} = \operatorname{argmax}_w p(w|y)$ 。由于

$$p(w|y) = \frac{p(y|w)p(w)}{p(y)} \tag{9}$$

但是 y 是我们的观察量, 所以 $p(y)$ 是一个常量, 在求解优化问题的时候可以不考虑进来。而且, 可以加入 \log 函数来简化运算, 而且与计算结果无关, 于是

$$\operatorname{argmax}_w p(w|y) = \log p(y|w)p(w) \tag{10}$$

代入可得：

$$\operatorname{argmax}_w p(w|y) = \sum_{i=1}^N \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{\|w\|^2}{2\sigma_0^2}\right) \quad (11)$$

$$= \sum_{i=1}^N \log \frac{1}{2\pi\sigma\sigma_0} \exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2} - \frac{\|w\|^2}{2\sigma_0^2}\right) \quad (12)$$

$$= \sum_{i=1}^N \log \frac{1}{2\pi\sigma\sigma_0} + \log \exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2} - \frac{\|w\|^2}{2\sigma_0^2}\right) \quad (13)$$

由于 $\log \frac{1}{2\pi\sigma\sigma_0}$ 与求解无关，所以

$$\operatorname{argmax}_w p(w|y) = \sum_{i=1}^N \log \exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2} - \frac{\|w\|^2}{2\sigma_0^2}\right) \quad (14)$$

$$= \sum_{i=1}^N -\frac{(y_i - w^T x_i)^2}{2\sigma^2} - \frac{\|w\|^2}{2\sigma_0^2} \quad (15)$$

$$(16)$$

公式可以转化为：

$$\operatorname{argmin}_w p(w|y) = \sum_{i=1}^N (y_i - w^T x_i)^2 + \frac{\sigma^2}{\sigma_0^2} \|w\|^2 \quad (17)$$

然后我们惊奇的发现将 $\frac{\sigma^2}{\sigma_0^2}$ 换成 λ 就又变成了和之前从频率角度看岭回归一样的结果。所以，对于上节我们得出的结论：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。我们进一步补充可得，Regularized LSE 可以等价于最大后验估计（MAP）其中噪声为 Gaussian Distribution，并且 w 的先验也为 Gaussian Distribution。