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：

$$\begin{aligned} L(w) &= \sum_{i=1}^N ||w^Tx_i-y_i||^2 && (1.1) \\ &=(w^Tx_1-y_1,w^Tx_2-y_2,\dots,w^Tx_N-y_N)\begin{pmatrix} w^Tx_1-y_1 \\ w^Tx_2-y_2 \\ \vdots \\ w^Tx_N-y_N \end{pmatrix} && (1.2) \end{aligned}$$

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。

Linear Regression 02

Chen Gong

14 October 2019

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 $P(w) = ||w||_1 = \sum_{i=1}^{N} |w_i|$，LASSO 回归等价于最小二乘回归加上 $||w||_1 < \varepsilon$ 条件，也就是将其中的每个维度都尽量压缩到 0，使得系数稀疏化。
2. Redge，岭回归，也就是 $P(w) = ||w||_2^2 = \sum_{i=1}^{N} w_i^2$，等价于最小二乘回归加上了 $w^T w < \varepsilon$ 条件，也就是让系数之间相差不会太大。

2 岭回归频率派角度

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

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

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

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

解得：

$$
W = (X^T X + \lambda I)^{-1} X^T Y \tag{3}
$$

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:

$$p(y|w) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(y-w^T x)^2}{2\sigma^2}\right) \tag{4}$$

$$p(w) = \frac{1}{\sqrt{2\pi}\sigma_0} exp\left(-\frac{w^T w}{2\sigma_0^2}\right) \tag{5}$$

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

$$p(w|y) = \frac{p(y|w)p(w)}{p(y)} \tag{6}$$

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，于是问题变成了求解如下的无约束优化问题:

$$\arg\max_w p(w|y) = \log p(y|w)p(w) \tag{7}$$

代入可得:

$$\begin{aligned}
argmax_w p(w|y) &= \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{||w||^2}{2\sigma_0^2}\right) && (8)\\
&= \sum_{i=1}^{N} \log \frac{1}{2\pi\sigma\sigma_0} \exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2} - \frac{||w||^2}{2\sigma_0^2}\right) && (9)\\
&= \sum_{i=1}^{N} \log \frac{1}{2\pi\sigma\sigma_0} + \log\exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2} - \frac{||w||^2}{2\sigma_0^2}\right) && (10)
\end{aligned}$$

由于 $\log \frac{1}{2\pi\sigma\sigma_0}$ 与求解无关，所以优化问题等价于:

$$\begin{aligned}
\arg\max_w p(w|y) &= \sum_{i=1}^{N} \log\exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2} - \frac{||w||^2}{2\sigma_0^2}\right) && (11)\\
&= \sum_{i=1}^{N} -\frac{(y_i - w^T x_i)^2}{2\sigma^2} - \frac{||w||^2}{2\sigma_0^2} && (12)\\
& && (13)
\end{aligned}$$

公式可以转化为:

$$\arg\min_w p(w|y) = \sum_{i=1}^{N} (y_i - w^T x_i)^2 + \frac{\sigma^2}{\sigma_0^2}||w||^2 \tag{14}$$

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 $\Longleftrightarrow$ 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(L_2 范数正则化）可以等价为最大后验估计 (MAP) 其中噪声为 Guassian Distribution，并且 w 的先验也为 Guassian Distribution。**