Linear Regression 推导

Soul Walker

2020 年 7 月 22 日

1 线性回归模型

输入：训练数据集

$$T = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N))\} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$$

其中, $\mathbf{x}_i \in \chi \in \mathbb{R}^N$ 为实例的特征向量, $\mathbf{x}_i = (\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, ..., \mathbf{x}_i^{(n)})^T$, $y_i \in \mathbb{R}$ 为实例预测值, $i = 1, 2, \cdots, N$ ，实例特征向量 $\mathbf{x}$，$\mathbf{X}$ 为训练样本集，形状为 (N, p)，$\mathbf{Y}$ 是训练标签集，形状为 $(N, 1)$。

输出：实例 $\mathbf{x}$ 的预测值 $\hat{y}$。

那我们首先给出的线性回归的假设就是

$$f(w) = w^T\mathbf{x} \tag{1}$$

2 最小二乘估计

首先我们使用二范数定义的平方误差来定义线性回归的损失函数

$$\begin{aligned}
L(w) &= \sum_{i=1}^{N} ||w^T\mathbf{x}_i - y_i||^2 \\
&= \sum_{i=1}^{N} (w^T\mathbf{x}_i - y_i)^2 \\
&= \begin{pmatrix} w^T\mathbf{x}_1 - y_1 & w^T\mathbf{x}_2 - y_2 & \cdots & w^T\mathbf{x}_N - y_N \end{pmatrix} \begin{bmatrix} w^T\mathbf{x}_1 - y_1 \\ w^T\mathbf{x}_2 - y_2 \\ \vdots \\ w^T\mathbf{x}_N - y_N \end{bmatrix} \\
&= (w^T\mathbf{X}^T - \mathbf{Y}^T)(\mathbf{X}w - \mathbf{Y}) \\
&= w^T\mathbf{X}^T\mathbf{X}w - w^T\mathbf{X}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}w + \mathbf{Y}^T\mathbf{Y} \\
&= w^T\mathbf{X}^T\mathbf{X}w - 2w^T\mathbf{X}^T\mathbf{Y} + \mathbf{Y}^T\mathbf{Y}
\end{aligned} \tag{2}$$

此时就可以得到最小化问题

$$\hat{w} = \arg\min_w L(w) \tag{3}$$

然后我们求偏导

$$\begin{aligned}\frac{\partial L(w)}{\partial w} &= 2\mathbf{X}^T\mathbf{X}w - 2\mathbf{X}^T\mathbf{Y} \overset{令}{=} 0 \\ \mathbf{X}^T\mathbf{X}w &= \mathbf{X}^T\mathbf{Y} \\ w &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\end{aligned} \tag{4}$$

所以这里我们得到了使用最小二乘法对参数的估计

$$\hat{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{5}$$

3 噪声为高斯分布的最大似然估计

对于同样的数据假设，我们引入高斯分布的噪声 ϵ，那么我们现在的模型假设就是

$$y = w^T\mathbf{x} + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{6}$$

那么 $y|\mathbf{x}, w \sim \mathcal{N}(w^T\mathbf{x}, \sigma^2)$，此时的最大似然估计就可以转换为如下

$$\hat{w} = \arg\max_w L(w) \tag{7}$$

其中

$$\begin{aligned}L(w) &= logp(\mathbf{Y}|\mathbf{x}_i, w) \\ &= log\prod_{i=1}^{N} p(y_i|\mathbf{x}_i, w) \\ &= \sum_{i=1}^{N} logp(y_i|\mathbf{x}_i, w) \\ &= \sum_{i=1}^{N}[log\frac{1}{\sqrt{2\pi}\sigma} + logexp\{-\frac{(y - w^T\mathbf{x}_i)^2}{2\sigma^2}\}] \\ &= \sum_{i=1}^{N}[log\frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2}(y_i - w^T\mathbf{x}_i)^2]\end{aligned} \tag{8}$$

这里，$log\frac{1}{\sqrt{2\pi}\sigma}$ 与 w 无关所以式 (7) 可以做如下转换

$$\begin{aligned}\hat{w} &= \arg\max_w L(w) \\ &= \arg\max_w \sum_{i=1}^{N}[-\frac{1}{2\sigma^2}(y_i - w^T\mathbf{x}_i)^2] \\ &= \arg\min_w \sum_{i=1}^{N}(y_i - w^T\mathbf{x}_i)^2\end{aligned} \tag{9}$$

4 正则化

我们有介绍，当造成过拟合时我们有以下三种解决方式

1. 增加数据

2. 降维

3. 正则化

那我们这里的正则化一般是加在损失函数后面来表示对模型的惩罚，这里有两种正则化框架

$$L_1 : \arg\min_w L(w) + \lambda ||w||_1, \lambda > 0 \tag{10}$$

$$L_2 : \arg\min_w L(w) + \lambda ||w||_2^2 = \arg\min_w L(w) + \lambda w^T w, \lambda > 0 \tag{11}$$

L_1(Lasso) 正则化会引起稀疏解。一方面来看，L_1 正则化相当于

$$\begin{cases} \arg\min\limits_w L(w) \\ s.t. ||w||_1 < C \end{cases} \tag{12}$$

平方损失函数在 w 空间是一个椭球，那么上式的求解就是椭球和 $||w||_1 = C$ 的切点。

L_2(Ridge) 正则化可使用式 (2) 中的结果

$$\begin{aligned} J(w) &= \sum_{i=1}^{N} ||w^T \mathbf{x}_i - y_i||^2 + \lambda ||w||_2^2 \\ &= w^T \mathbf{X}^T \mathbf{X} w - 2 w^T \mathbf{X}^T \mathbf{Y} + \mathbf{Y}^T \mathbf{Y} + \lambda w^T w \\ &= w^T (\mathbf{X}^T \mathbf{X} + \lambda I) w - 2 w^T \mathbf{X}^T \mathbf{Y} + \mathbf{Y}^T \mathbf{Y} \end{aligned} \tag{13}$$

那么我们可以得出

$$\hat{w} = \arg\min_w J(w) \tag{14}$$

求解的话我们求 $J(w)$ 对 w 的偏导

$$\frac{\partial J(w)}{\partial w} = 2(\mathbf{X}^T \mathbf{X} + \lambda I) w - 2 \mathbf{X}^T \mathbf{Y} \tag{15}$$

我们令式 (15)=0 即可求得

$$\hat{w} = (\mathbf{X}^T \mathbf{X} + \lambda I)^{-1} \mathbf{X}^T \mathbf{Y} \tag{16}$$