# Linear Regression

ChangYu

2023 年 2 月 17 日

## 1 广义线性回归

对于广义线性回归，我们的目标是对设计矩阵 $X$ 与样本标签 $Y$ 构建模型:

$$f(X) = X\theta$$

目标参数 $\theta$ 是我们的求解目标。为此，我们构建损失函数如下:

$$J(\theta) = \frac{1}{2N}\|Y - X\theta\| = \frac{1}{2N}\sum_{i=1}^{N}(y_i - \sum_{j=1}^{p}\theta_j x_{ij})^2$$

我们的目的是寻求损失函数的最小，即满足求解优化问题:

$$\arg\min_{\theta} J(\theta). = \arg\min_{\theta} \|Y - X\theta\|^2$$

## 2 参数求解

关于广义线性回归的参数求解普遍有两种方法，一种是投影法，一种是求导法，此外还有通用的梯度下降法。

### 2.1 投影法

考虑到 $X\theta$ 所得到的预测向量在设计矩阵 $X$ 的列空间内，因此为了使得向量 $Y$ 和向量 $X\theta$ 的距离最短，则必然有 $Y - X\theta$ 正交于设计矩阵 $X$ 的列空间。

由此立得:

$$X^{\mathrm{T}}(Y - X\theta) = 0$$

即:

$$X^{\mathrm{T}}X\theta = X^{\mathrm{T}}Y$$

暂且考虑设计矩阵为满秩矩阵，不考虑数据集中的多重共线性问题，我们可以解出:

$$\theta = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$$

## 2.2 逐项求导法

考虑损失函数：

$$J(\theta) = \frac{1}{2N}\sum_{i=1}^{N}(y_i - \sum_{j=1}^{p}\theta_j x_{ij})^2$$

去除标准化系数 $\frac{1}{2N}$ 我们不妨对所有的参数进行分别求导：

$$\begin{aligned}
\frac{\partial J}{\partial \theta_1} &= \sum_{i=1}^{N} x_{i1}(y_i - \sum_{j=1}^{p}\theta_j x_{ij}) \\
\frac{\partial J}{\partial \theta_2} &= \sum_{i=1}^{N} x_{i2}(y_i - \sum_{j=1}^{p}\theta_j x_{ij}) \\
&\vdots \\
\frac{\partial J}{\partial \theta_p} &= \sum_{i=1}^{N} x_{ip}(y_i - \sum_{j=1}^{p}\theta_j x_{ij})
\end{aligned}$$

通过使用克莱默法则，可以同样得出

$$\theta = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$$

## 2.3 梯度下降法

梯度下降法是在求解凸函数最值中的一个适用于计算机求解的暴力算法，其核心是寻找任意点后通过计算梯度不断的对点进行调整，从而通过迭代寻找到最小值点的办法。

考虑在任意特殊点，该点处的梯度如下

$$\nabla = (\frac{\partial J}{\partial \theta_1}, \frac{\partial J}{\partial \theta_2}, \cdots, \frac{\partial J}{\partial \theta_p})$$

考虑步长参数 $\alpha$ 做迭代：

$$\begin{aligned}
\theta_1 &= \theta_1 - \alpha\frac{\partial J}{\partial \theta_1} \\
\theta_2 &= \theta_2 - \alpha\frac{\partial J}{\partial \theta_2} \\
&\vdots \\
\theta_p &= \theta_p - \alpha\frac{\partial J}{\partial \theta_p}
\end{aligned}$$

将 $J$ 带入得：

$$\theta_j = \theta_j - \alpha\frac{\partial J}{\partial \theta_1} = \theta_j - \alpha\sum_{i=1}^{N} x_{i1}(y_i - \sum_{j=1}^{p}\theta_j x_{ij})$$

一般而言，梯度下降法在面对凸函数时一定会达到最优解，在面对凸函数时有可能陷入局部最优解。

# 3 线性模型正则化

考虑到线性模型的参数求解过程中，设计矩阵 $X$ 可能是非满秩矩阵，就会导致求解的参数数值过大，会导致模型的稳定性差，因此，我们需要考虑通过对损失函数引入新的惩罚机制，使模型不至于让参数过大。以及，通过引入正则化的方法来解决模型的过拟合问题。

## 3.1 L2 正则化

L2 正则化的方法是在最小二乘法的基础上引入一个惩罚项，得到新的损失函数：

$$J(\theta) = \frac{1}{2N}\sum_{i=1}^{N}(y_i - \sum_{j=1}^{p}\theta_j x_{ij})^2 + \frac{\lambda}{2}\|\theta\|^2$$

通过引入惩罚项 $\|\theta\|$ 目的是为了不让参数的取值过大。下面我们通过梯度下降法求解正则化后的参数 $\theta$

为了方便计算书写，令：

$$H_\theta(x^{(i)}) = \sum_{j=1}^{p}\theta_j x_{ij}$$

于是我们得到：

$$\frac{\partial J}{\partial \theta_j} = \frac{1}{N}\sum_{i=1}^{N}(H_\theta(x^{(i)}) - y^{(i)})x_j + \lambda\theta$$

因此，我们做迭代如下：

$$\theta_j = \theta_j - \alpha\frac{\partial J}{\partial \theta_j} = \theta_j - \alpha(\frac{1}{N}\sum_{i=1}^{N}(H_\theta(x^{(i)}) - y^{(i)})x_j + \lambda\theta)$$

化简后得：

$$\theta_j = (1-\alpha\lambda)\theta_j - \alpha\frac{1}{N}\sum_{i=1}^{N}(H_\theta(x^{(i)}) - y^{(i)})x_j$$

此外，关于 L2 正则化，也可以通过求解解析解的方法得到目标损失函数的解析解：考虑到

$$J(\theta) = \|Y - X\theta\|^2 + \lambda\|\theta\|^2$$

直接计算，不难得到：

$$J(\theta) = \theta^{\mathrm{T}}X^{\mathrm{T}}X\theta - 2\theta^{\mathrm{T}}X^{\mathrm{T}}Y + Y^{\mathrm{T}}Y + \lambda\theta^{\mathrm{T}}\theta$$

对向量 $\theta$ 求导得：

$$\frac{\partial J}{\partial \theta} = 2X^{\mathrm{T}}X\theta - 2X^{\mathrm{T}}Y + \lambda\theta = 0$$

不难得到：

$$\theta = (X^{\mathrm{T}}X + \frac{\lambda}{2}I)^{-1}X^{\mathrm{T}}Y$$

其中 $\lambda$ 为正则化参数。在统计学中，L2 正则化的方法也被称为岭回归。

### 3.2 L1 正则化

L1 正则化与 L2 正则化类似，都是通过设计惩罚项来实现模型的稳定性。

损失函数为:

$$J(\theta) = J_0(\theta) + J_1(\theta) = \frac{1}{2N}\sum_{i=1}^{N}(y_i - \sum_{j=1}^{p}\theta_j x_{ij})^2 + \lambda\|\theta\|_1$$

解决方法与 L2 类似，都是通过梯度下降法寻求最优解。

对 $\theta$ 进行求导:

$$\frac{\partial J}{\partial \theta} = \lambda sgn(\theta) + \frac{\partial J_0}{\partial \theta}$$

因此，通过梯度下降法即得到迭代过程:

$$\theta_j = \theta_j - \alpha\lambda sgn(\theta_j) - \alpha\frac{\partial J_0}{\partial \theta}$$