Regression shrinkage and selection via the lasso 论文总结

Zheng Xin

本文是对阅读论文《Regression shrinkage and selection via the lasso》总结，将对论文原文的关键部分进行梳理，主要内容包括 4 个部分：

第一部分背景，对论文中省略的背景作出解释。

第二部分 lasso 的介绍，对论文中 lasso 给出自己的理解。

第三部分 lasso 重要过程推导，对论文中 lasso 相关公式中，较难理解的部分进行推导。

第四部分结论，看完论文后对 lasso 的总体理解。

一、背景

对数据集 D：$(X^i, y_i), i = 1,2,\cdots,N$，其中$X^i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$和$y_i$分别是第 i 次观测的预测变量和响应变量。线性回归的基本表示为：

$$\mathrm{f}(X^i) = W^T X^i + b$$

最终目的是是得$\mathrm{f}(X_i) \simeq y_i$。

1.普通最小二乘回归 OLS

为了确定 W 和 b，即要最小化$\mathrm{f}(X_i)$与y_i的差值，若使用均方误差表示，那么：

$$(W^*, b^*) = arg\min_{(W,b)} \sum_{i=1}^{N} \left(f(X^i) - y_i\right)^2 = arg\min_{(W,b)} \sum_{i=1}^{N} \left(WX^i + b - y_i\right)^2$$

均方方误差在几几何上表现为欧氏氏距离，其目的是找到一条直线，使样本到直线的距离最小。

将 W 和 b 写成向量形式$\vec{w} = (W;\ b)$，数据集 D 表示为$\mathrm{N} \times (\mathrm{p}+1)$的矩阵，

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{N1} & \cdots & x_{Np} & 1 \end{pmatrix}$$

$$\vec{y} = (y_1; \cdots; y_N)$$

那么，

$$(W^*, b^*) = \vec{w}^* = arg\min_{\vec{w}^*} (\vec{y} - X)^T (\vec{y} - X) = E_{\vec{w}}$$

对$\vec{w}$求偏导得：

$$\frac{\partial E_{\vec{w}}}{\partial \vec{w}} = 2X^T(X_{\vec{w}})$$

当X^TX为满秩矩阵或正定矩阵时，令上式为 0，可求得：

$$X^TX\vec{w} = X^T\vec{y}$$

$$\vec{w}^* = (X^TX)^{-1}X^T\vec{y}$$

令$\overrightarrow{X_\imath}$表示 X 当第 i 行向量，则最终学到的线性回归模型为：

$$\mathrm{f}(\overrightarrow{X_\imath}) = \overrightarrow{X_\imath}(X^TX)^{-1}X^T\vec{y}$$

但是，在实际问题中，X^TX往往并不是满秩矩阵：自行列向量之间存在高度多重共线性，或列向量数大于行向量数。这会导致偏回归系数无解或结果无效，为了能够克服这个问题，可以使用子集选择将高自相关变量删除，或者选用岭回归也能够避免X^TX不可逆的情况。

2. 岭回归

岭回归在X^TX的基础上加上一个较小的λ扰动，从而使得行列式不再为 0：

$$\vec{w}^* = (X^TX + \lambda I)^{-1}X^T\vec{y}$$

在主对角线元素上都加上了λ ，是得矩阵非奇异，随着λ的不断增大，$\vec{w}^*(\lambda)$的各元素$\vec{w}^*{}_i(\lambda)$的绝对值不断减小，它们相对于正确值$\vec{w}_i$的偏差不断变大。

设 OLS 的解为$\vec{w}_i$，岭回归的解为$\vec{w}_i{}'$：

$$\begin{aligned}
\vec{w}_i{}' &= (X^TX + \lambda I)^{-1}X^T\vec{y} \\
&= (X^TX + \lambda I)^{-1}(X^TX)(X^TX)^{-1}X^T\vec{y} \\
&= (X^TX + \lambda I)^{-1}(X^TX)\vec{w} \\
&= (X^TX + \lambda I)^{-1}(X^TX + \lambda I - \lambda I)\vec{w} \\
&= (I - \lambda(X^TX + \lambda I)^{-1})\vec{w} < \vec{w}
\end{aligned}$$

可以看出，$\vec{w}_i{}'$是对$\vec{w}$向原点的压缩，并不会出现某一系数为 0 的稀疏解情况。但是，在实际问题中，特征存在冗余，稀疏解有利于找到有用的维度并减少冗余，提预测高鲁棒性和准确性。

二、lasso 的介绍

lasso 的回归表达式：

$$\vec{w}^* = arg\min_{\vec{w}^*}\left[\sum_{1}^{N}(W^T\overrightarrow{X_\imath} - y_i)^2 + \lambda\sum_{j=1}^{p+1}|w_j|\right] = arg\min_{\vec{w}^*}\sum_{i=1}^{N}(W^T\overrightarrow{X_\imath} - y_i)^2, \sum_{j=1}^{p+1}|w_j| \le t$$

用几何形式表示 lasso 和岭回归：

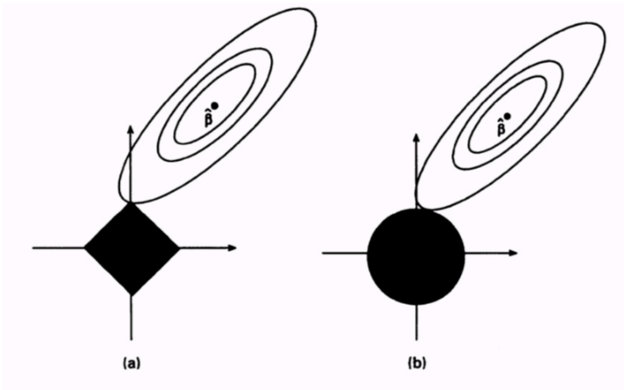

图 2: (a)the lasso(b) 岭回归的估计

如上图所示，lasso 取得的最优解使轴上对应的特征系数为 0，因此产生的稀疏解，这起到了特征选择的作用，删除了无用特征，留下了强特征。当λ逐渐增大时，图中棱形范围不断增大，解范围变小，因此起到了压缩变量的作用。

三、lasso 重要过程推导

对论文中公式：

$$\hat{\beta}_j = \mathrm{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)^+ \quad (3)$$

进行推导。

设$\hat{\beta}$为 OLS 的解，则当正交时：

$$\hat{\beta} = (X^TX)^{-1}X^T\vec{y}$$

lasso 估计为：

$$L = \arg\min_{\beta} \frac{1}{2}(\beta X - y)^2 + \lambda|\beta|$$

$$= arg\min_{\beta} \frac{1}{2}(y^Ty - 2y^TX\beta + \beta^T\beta) + \lambda|\beta|$$

$$= arg\min_{\beta} \frac{1}{2}y^Ty - y^TX\beta + \frac{1}{2}\beta^T\beta + \lambda|\beta|$$

其中，y^Ty为常数，则上式可转化为：

$$L = \arg\min_{\beta}\left(y^TX\beta + \frac{1}{2}\beta^2\right) + \lambda|\beta|$$

$$= arg\min_{\beta}\left(-\hat{\beta}\beta + \frac{1}{2}\beta^2\right) + \lambda|\beta|$$

$$= arg\min_{\beta}\sum_{i=1}^{p+1}\left(-\hat{\beta}_i\beta_i + \frac{1}{2}{\beta_i}^2 + \lambda|\beta_i|\right)$$

对每一个 i，我们希望最小化：

$$L_i = -\hat{\beta}_i\beta_i + \frac{1}{2}{\beta_i}^2 + \lambda|\beta_i|$$

式子右边最后两项为正，因此，要最小化L_i就要保证：

$$当\hat{\beta}_i > 0时，\beta_i \geq 0$$
$$当\hat{\beta}_i < 0时，\beta_i \leq 0$$

（1）$\hat{\beta}_i > 0$，$\beta_i \geq 0$：

L_i对β_i求偏导：

$$-\hat{\beta}_i + \beta_i + \lambda = 0$$
$$\beta_i = \hat{\beta}_i - \lambda$$

由于$\beta_i \geq 0$，因此当且仅当$\hat{\beta}_i - \lambda$非负时成立。

（2）$\hat{\beta}_i < 0$，$\beta_i \leq 0$：

L_i对β_i求偏导：

$$-\hat{\beta}_i + \beta_i - \lambda = 0$$
$$\beta_i = \hat{\beta}_i + \lambda$$

由于$\beta_i \leq 0$，因此当且仅当$\hat{\beta}_i + \lambda$非正时成立。

故两种情况结合，可得所证式子：

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)^+$$

γ即为推导过程中的λ。当λ增大时，$|\beta|$减小；当$\lambda = 0$时，为 OLS 解；当$\lambda > \max_i \beta_i$时，$\beta = 0$。

四、总结

对于回归问题，lasso 有两个重要特点：

1. 对系数进行压缩；
2. 能特征选择。

这两个特点解决了 OLS 的无解情况，相比岭回归多了特征选择的作用。

从另一角度，lasso 可以看作对 OLS 加上 l1 正则，有控制模型复杂度的作用。