带约束的线性回归 Constrained Linear Regression

概述

在线性回归模型$Y = X\theta + \varepsilon$ 的求解当中，求 $\hat{\theta} = \text{argmin}_\theta ||X\theta - Y||_2^2$ ，是一个无约束优化问题，而这则导致模型求解时容易受到异常值的影响从而出现过拟合的情况。为了处理这种情况，一个行之有效的做法是对模型添加约束，降低模型的复杂度。单独使用l1,l2约束为最常用的约束条件，分别对应于LASSO回归和岭回归，还有组合使用l1,l2约束的ElasticNet等。

还有一种特殊的子空间约束，利用线性投影映射(满足 $P^2 = P, P^T = P$ 的 b×b 的矩阵) P 将 θ 映射到 $P\theta$, 以 $P\theta$ 为模型参数，即 $Y = XP\theta + \varepsilon$ ，这样就将参数空间限制在了 P 的投影空间里。然而实际上构造合适的投影 P 是一件很困难的事情，所以这种方法实践上应用不多。

以下简记$J(\theta) = \frac{1}{2}||X\theta - Y||_2^2$ ，对于无约束优化问题，得到的解是完全一致的，但因为此处添加了约束项，此处添加一个常数系数是为了让后面解的形式简洁，实际上不添加这一常数系数得到的解本质上也是相同的。

岭回归 l2-constrained linear regression

往线性回归模型中添加l2限制，求解则变为：

$$\min_\theta J(\theta) \quad \text{subject to } ||\theta||_2^2 \le r^2$$

这个限制实际上是将 θ 限制在一个半径为 r 的超球体内，转化这个约束问题的拉格朗日对偶来求解：

$$\max_\lambda \min_\theta \left[J(\theta) + \frac{\lambda}{2}(||\theta||_2^2 - r^2) \right]$$

尽管我们看到在这里 λ 原则上是与 r 有关的，但实际操作上，我们往往不是给定 r，而是把 λ 看作一个模型参数事先给定，而我们始终最关注的是 θ 即：

$$\hat{\theta} = \text{argmin}_\theta \left[J(\theta) + \frac{\lambda}{2}||\theta||_2^2 \right]$$

式子很容易求解，最小二乘解为：

$$(X^TX + \lambda I)^{-1}X^TY$$

对比无约束的最小二乘解：

$$(X^TX)^{-1}X^TY$$

可以看到，在求逆之前，加了一个λI，$X^TX+\lambda I$ 的主对角元比 X^TX 大，在求逆时能获得更稳定的数值，当λ 充分大时，矩阵逆趋于 $\lambda^{-1}I$，θ 则趋于 0 .

LASSO回归 Least Absolute Shrinkage and Selection Operator

往线性回归模型中添加l1限制，求解则变为：

$$\min_\theta J(\theta) \quad \text{subject to } ||\theta||_1 \le r^2$$

其中 $||\theta||_1$ 为 $\theta=(\theta_0,\theta_1,...,\theta_p)^T$ 的一范数 $||\theta||_1=\sum_i|\theta_i|$

将模型描述为：

$$\min_{\theta,z}\ [J(\theta)+\lambda||z||_1] \quad \text{subject to } \theta=z$$

函数没有连续一阶导数，不适合像上面那样直接求解，这里使用ADMM算法求解此问题，这个优化问题的增广拉格朗日函数为：

$$L(\theta,z,u)=J(\theta)+\lambda||z||_1+u^T(\theta-z)+\frac{1}{2}||\theta-z||^2$$

对 θ 求导

$$\frac{\partial L}{\partial\theta}=-X^T(Y-X\theta)+u+\theta-z$$

令上式等于0,得

$$\begin{aligned}\theta^{(k+1)} &= \text{argmin}_\theta L(\theta,z^{(k)},u^{(k)}) \\ &= (X^TX+I)^{-1}(X^TY+z^{(k)}-u^{(k)})\end{aligned}$$

又由

$$\min_z\left[\lambda|z|+u(\theta-z)+\frac{1}{2}(\theta-z)^2\right]=\max(0,\theta+u-\lambda)+\min(0,\theta+u+\lambda)$$

得

$$\begin{aligned}z^{(k+1)} &= \text{argmin}L(\theta^{(k+1)},z,u^{(k)}) \\ &= \max(0,\theta^{(k+1)}+u^{(k)}-\lambda\mathbf{1})+\min(\mathbf{0},\theta^{(\mathbf{k+1})}+\mathbf{u^{(k)}}+\lambda\mathbf{1})\end{aligned}$$

这里的 $\mathbf{1}$ 为p维全为1的向量，再有 u 的更新式子

$$u^{(k+1)}=u^{(k)}+\theta^{(k+1)}-z^{(k+1)}$$

LASSO回归的其他做法：坐标下降法(Coordinate descent)、最小角回归（LAR）。

附：ADMM算法 Alternating Direction Method of Multipliers

设有如下优化问题：

$$\text{argmin}_{\theta,z}\ f(\theta)+g(z)\ \text{ subject to } A\theta+Bz=c$$

则有增广拉格朗日函数：

$$L(\theta,z,u)=f(\theta)+g(z)+u^T(A\theta+Bz-c)+\frac{1}{2}||A\theta+Bz-c||^2$$

参数更新式为

$$\begin{aligned}\theta^{(k+1)} &= \text{argmin}_\theta L(\theta,z^{(k)},u^{(k)})\\ z^{(k+1)} &= \text{argmin}_z L(\theta^{(k+1)},z,u^{(k)})\\ u^{(k+1)} &= u^{(k)}+A\theta^{(k+1)}+Bz^{(k+1)}-c\end{aligned}$$

附：坐标下降法

设有如下优化问题：

$$\hat{\theta}=\text{argmin}_\theta\ L(\theta)$$

那么p维的 θ 可以用以下方法收敛到极小值点

1. θ 随机取一个初值 $\theta^{(0)}$
2. 在第 k+1 轮迭代，从 $\theta_1^{(k)}$ 到 $\theta_p^{(k)}$ 逐个优化

$$\begin{aligned}\theta_1^{(k+1)} &= \text{argmin}_{\theta_1} L(\theta_1,\theta_2^{(k)},...,\theta_p^{(k)})\\ &.....\\ \theta_i^{(k+1)} &= \text{argmin}_{\theta_i} L(\theta_1^{(k)},...,\theta_i,...,\theta_p^{(k)})\\ &.....\\ \theta_p^{(k+1)} &= \text{argmin}_{\theta_p} L(\theta_1^{(k)},...,\theta_{p-1}^{(k)},\theta_p)\end{aligned}$$

3. 检查 $||\theta^{(k+1)}-\theta^{(k)}||$ ，如果充分小，那么停止迭代，否则重复(2)

附：最小角回归

最小角回归的路径与LASSO相似，但最小角回归并不是从LASSO的模型入手考虑的，也就是说，从结果来看，可以认为最小角回归是LASSO的一种高效解法，但从推导上看则是从不同的起点出发的。要完全获得与LASSO相同的求解路径，应用的是LARS的一种修改版本，在Efron的论文中有非常细致的解说。

详见Efron的论文：LEAST ANGLE REGRESSION

> The main point of this paper is that both Lasso and Stagewise are variants of a basic procedure called Least Angle Regression

推广的 l_p 回归

上面的两种情况一个用的是 l_2 限制,一个用的是 l_1 限制，实际上，可以很容易地推广到任意的 l_p 限制：
$p \in (1, +\infty)$ 时，均有连续一阶导数，因此所有基于梯度的方法均可直接套用;
$p = \infty$ 时，相当于限制了参数的最大值
$p = 0$ 时，相当与限制了非零参数的个数
$p \in (0, 1)$ 时，优化函数是非凸的，无法确保找到全局最优，因此为了要保持凸性且让解有稀疏性的话，最合适的选择是p=1，即LASSO.

弹性网回归 **ElasticNet**

往线性回归模型中同时添加l1限制和l2限制，求解则变为：

$$\hat{\theta} = \text{argmin}_\theta J(\theta) \quad \text{subject to } \ (1-\alpha)||\theta||_1 + \alpha||\theta||_2^2 \leq t \ \text{ for some t}$$

$(1-\alpha)||\theta||_1 + \alpha||\theta||_2^2$ 称为 elastic net penalty , Zhou的论文中巧妙地将l2惩罚融入到数据集中，详见论文中的Lemma 1，将问题转化为一个lasso like 的问题，最后用LAR将求解路径解得。

详见Hui Zhou的论文：Regularization and variable selection via the elastic net