University of Chinese Academy of Sciences

CS091M4042H Assignment 4

Elastic Net Regression*论文笔记

熊兴旺†
Sno: 2018E8013261007
中国科学院计算技术研究所

November 12, 2018

*Regression Shrinkage and Selection via the Elastic Net, with Applications to Microarrays

†email: xingw.xiong@gmail.com; xiongxingwang@ict.ac.cn

1 Introduction

评价模型质量的标准:

1. 对未知数据的**预测精确度**
2. 模型的**可解释性**：即模型应该尽可能简单。简单的模型更能说明预测结果和协变量之间的关系，尤其是当预测变量的数量很大时。

已有回归模型的比较:

- **最小二乘法 (OLS)**: 最小化残差平方和 (RSS)。容易引起过拟合，在模型的预测精确度和可解释性上表现很差。
- **最优子集选择 (Best Subset Selection)**: 从所有预测变量中选择一个非空子集，让他们的系数非零，其他元素系数为 0，然后最小化 RSS。最优子集选择会产生一个稀疏的模型，但是由于它是离散的方法，所以模型的方差很高。
- **岭回归 (Ridge Regression)**: 在满足其 L_2 范数小于某个数时，使 RSS 最小。是一个连续的收缩方法，通过"bias-variance trade-off" 使得预测性能较好。但是产生的模型会比较复杂，因为所有的预测变量都保持在模型中。
- **套索回归 (Lasso Regression)**: 对 L_1 范数进行惩罚，在满足 L_1 范数小于某一个数时，使 RSS 最小。Lasso Regression 可以连续收缩，也可以进行变量选择，在很多情况下，Lasso 都是一个有效的回归模型。但是，在下面三种情形下表现很差:
  1. 当预测变量的数目 p 大于观测变量的数目 n 时，Lasso 是意义不明确的方法，除非 L_1 范数小于某一个常数。由于凸优化的性质，Lasso 最多只能选择 n 个变量，这似乎限制了变量选择方法。
  2. 如果有一组变量，两两相关性比较高，那么 Lasso 往往只会在这组变量中选其中一个，而且不关心具体选了的哪一个变量。
  3. 对于通常的 $n > p$ 时，如果预测变量之间有高度的相关性，经验表明，Ridge Regression 的预测准确性明显由于 Lasso Regression。

一个理想的**基因选择**的方法应该有下面的能力:

- 消除不重要的基因;
- 一旦群组中一个基因进入模型，整个群组的所有基因自动进去 (即**群组选择 (Grouped Selection)**)。

文章提出了新的正则化方法: **弹性网回归 (Elastic Net Regression)**，它在自动选择变量的同时也能实现连续收缩，并具有群组选择的能力，而且在场景三中的预测效果比 Lasso 好。

2 Native Elastic Net

2.1 Definition

假定数据集有 n 个观测值 p 个预测变量，观测值记为 $Y=(y_1,y_2,\ldots,y_n)^T$，预测变量记为 $X=[X_1|\ldots|X_p]$，其中列向量 $X_j=(x_{1j},\ldots,x_{nj})^T, i=1,\ldots,p$ 表示的第 j 个变量的 n 个预测值。假设数据是**中心化**和**标准化**的，即：

$$\sum_{i=1}^{n} y_i=0,\quad \sum_{i=1}^{n} x_{ij}=0,\quad \sum_{i=1}^{n} x_{ij}^2=1,\quad \text{for } j=1,2,\ldots,p. \tag{2.1}$$

对于任意非负的 λ_1，λ_2，朴素弹性网准则定义如下：

$$L(\lambda_1,\lambda_2,\beta)=|Y-X\beta|^2+\lambda_2|\beta|^2+\lambda_1|\beta|_1 \tag{2.2}$$

其中，$|\beta|^2=\sum_{j=1}^{p}\beta_j^2, |\beta|_1=\sum_{j=1}^{p}\beta_j$。通过求上式最小值，可以回归系数 $\hat{\beta}$ 的朴素弹性网估计:

$$\hat{\beta}=\arg\min_{\beta} L(\lambda_1,\lambda_2,\beta) \tag{2.3}$$

通过下面步骤，可以把**朴素弹性网转化为带罚函数的最小二乘法**。令 $\alpha=\frac{\lambda_2}{\lambda_1+\lambda_2}$，那么:

$$\hat{\beta}=\arg\min_{\beta}|Y-X\beta|^2,\quad \text{subject to } (1-\alpha)|\beta|_1+\alpha|\beta|^2\le t \text{ for some } t \tag{2.4}$$

$(1-\alpha)|\beta|_1+\alpha|\beta|^2$ 即为弹性网惩罚，是 Lasso 惩罚和 Ridge 惩罚的凸组合。

2.2 Solution

朴素弹性网求解方法的主要思想是讲其规约为等价的 Lasso 问题。

Lemma 1 *Given data set* $(Y,\mathbf{X})$ *and* (λ_1,λ_2)*, define an artificial data set* $(Y^*,\mathbf{X}^*)$ *by*

$$\mathbf{X}^*_{(n+p)\times p}=(1+\lambda_2)^{-\frac{1}{2}}\begin{pmatrix}\mathbf{X}\\ \sqrt{\lambda_2}\mathbf{I}\end{pmatrix},\quad Y^*_{(n+p)}=\begin{pmatrix}Y\\ 0\end{pmatrix}.$$

Let $\gamma=\frac{\lambda_1}{\sqrt{1+\lambda_2}}$ *and* $\boldsymbol{\beta}^*=\sqrt{1+\lambda_2}\boldsymbol{\beta}$*. Then the naive elastic net criterion can be written as*

$$L(\gamma,\boldsymbol{\beta})=L(\gamma,\boldsymbol{\beta}^*)=\left|Y^*-\mathbf{X}^*\boldsymbol{\beta}^*\right|^2+\gamma\left|\boldsymbol{\beta}^*\right|_1.$$

Let

$$\hat{\boldsymbol{\beta}}^*=\arg\min_{\boldsymbol{\beta}^*} L(\gamma,\boldsymbol{\beta}^*),$$

then

$$\hat{\boldsymbol{\beta}}=\frac{1}{\sqrt{1+\lambda_2}}\hat{\boldsymbol{\beta}}^*.$$

Lemma 1 表明通过**把原始数据集拓展为增广数据，弹性朴素网问题就规约为了 Lasso 问题**。增广数据有 $n+p$ 条观测，矩阵 X^* 的秩为 p，也就是说，在任何情况，朴素弹性网的会选择所有 p 个预测变量。下面分别给出正交设计中，OLS, Ridge, Lasso, native elastic net 的超参数估计值:

$$\hat{\boldsymbol{\beta}}^{ols} = X^T Y \tag{2.5a}$$

$$\hat{\boldsymbol{\beta}}(\text{ridge}) = \frac{\hat{\boldsymbol{\beta}}^{ols}}{1+\lambda_2} \tag{2.5b}$$

$$\hat{\beta}_i(\text{lasso}) = (|\hat{\beta}^{ols}| - \frac{\lambda_1}{2})_+ \text{sgn}(\hat{\beta}_i^{ols}) \tag{2.5c}$$

$$\hat{\beta}_i(\text{native elastic net}) = \frac{(|\hat{\beta}^{ols}| - \frac{\lambda_1}{2})_+}{1+\lambda_2} \text{sgn}(\hat{\beta}_i^{ols}) \tag{2.5d}$$

$$\tag{2.5e}$$

2.3 Results on the grouping effect

Theorem 1 *Given data $(Y, \mathbf{X})$ and parameters (λ_1, λ_2), the response Y is centered and the predictors $\mathbf{X}$ are standardized. Let $\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2)$ be the naive elastic net estimates. Define $D_{\lambda_1,\lambda_2}(i,j) = \frac{1}{|Y|}\left|\left|\hat{\beta}_i(\lambda_1,\lambda_2)\right| - \left|\hat{\beta}_j(\lambda_1,\lambda_2)\right|\right|$.*

1. *If $\mathbf{X}_i = \mathbf{X}_j$, then given any fixed $\lambda_2 > 0$, $\hat{\beta}_i(\lambda_1,\lambda_2) = \hat{\beta}_j(\lambda_1,\lambda_2)$ for all $\lambda_1 \geq 0$, i.e., the whole solution paths of $\hat{\beta}_i$ and $\hat{\boldsymbol{\beta}}_j$ are the same.*

2. *$\forall \lambda_2 > 0$, suppose $\hat{\boldsymbol{\beta}}_i(\lambda_1,\lambda_2)\,\hat{\boldsymbol{\beta}}_j(\lambda_1,\lambda_2) > 0$, then $D_{\lambda_1,\lambda_2}(i,j) \leq \frac{1}{\lambda_2}\sqrt{2(1-\rho)}$, where $\rho = cor(\mathbf{X}_i, \mathbf{X}_j)$.*

该定理说明了严格凸性保证了群组效应。

3 Elastic Net

3.1 Deficiency of the naive elastic net

经验表明，朴素弹性网的表现往往没有想象中那么好，除非它接近 Ridge 或 Lasso 回归。朴素弹性网估计可以看成**两阶段过程**:

1. 先固定 λ_2 寻找岭回归系数;
2. 然后沿着 Lasso 系数的路径收缩。

然而相比于 Lasso 和 Ridge，这两次收缩不会明显减少方差，相反还带了不必要的模型偏差。

3.2 The elastic net estimates

We outline a remedy for the poor performance of the naive elastic net. Let us follow the notation of Section 2.2. Given data $(Y, \mathbf{X})$ and penalty parameter (λ_1, λ_2), after introducing the equivalent artificial data set $(Y^*, \mathbf{X}^*)$, we solve a lasso type problem

$$\hat{\boldsymbol{\beta}}^* = \arg\min_{\boldsymbol{\beta}^*} |Y^* - \mathbf{X}^*\boldsymbol{\beta}^*|^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}} |\boldsymbol{\beta}^*|_1 . \tag{7}$$

The elastic net estimates $\hat{\boldsymbol{\beta}}$ are defined as

$$\hat{\boldsymbol{\beta}}\,(\text{elastic net}) = \sqrt{1+\lambda_2}\hat{\boldsymbol{\beta}}^* . \tag{8}$$

Recall that $\hat{\boldsymbol{\beta}}(\text{naive elastic net}) = \frac{1}{\sqrt{1+\lambda_2}}\hat{\boldsymbol{\beta}}^*$, thus

$$\hat{\boldsymbol{\beta}}(\text{elastic net}) = (1+\lambda_2)\hat{\boldsymbol{\beta}}(\text{naive elastic net}). \tag{9}$$

可以发现，弹性网估计值就是对朴素弹性网估计值的一种放缩。选择 $1+\lambda_2$ 作为调节因子的原因之一是：在正交设计中，预测变量是正交的，此时 Lasso 估计值是最优值 (minimax optimal)，经过 $1+\lambda_2$ 调节之后，弹性网就变成了最优值。
还有一个原因是考虑对岭回归作用矩阵对分解。

of the ridge operator. Since $\mathbf{X}$ are standardized beforehand, we have

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1 & \rho_{12} & \cdot & \rho_{1p} \\ & 1 & \cdot & \cdot \\ & & 1 & \rho_{p-1,p} \\ & & & 1 \end{bmatrix}_{p\times p} .$$

Then for ridge regression with parameter λ_2,

$$\hat{\beta}(\text{ridge}) = \mathbf{R}Y,$$

where $\mathbf{R}$ is the ridge operator

$$\mathbf{R} = (\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{I})^{-1}\mathbf{X}^T.$$

We can rewrite $\mathbf{R}$ as

$$\mathbf{R} \;= \frac{1}{1+\lambda_2}\mathbf{R}^* = \frac{1}{1+\lambda_2}\begin{bmatrix} 1 & \frac{\rho_{12}}{(1+\lambda_2)} & \cdot & \frac{\rho_{1p}}{(1+\lambda_2)} \\ & 1 & \cdot & \cdot \\ & & 1 & \frac{\rho_{p-1,p}}{(1+\lambda_2)} \\ & & & 1 \end{bmatrix}^{-1} \mathbf{X}^T. \tag{10}$$

R^* 矩阵除了被 $\frac{1}{1+\lambda_2}$ 收缩之外，与 OLS 乘子非常类似，这个就是**去相关 (即降低相关性)**。所以，岭回归

能够解决线性回归问题的原因在于：岭回归是用了 $\frac{1}{1+\lambda_2}$ 实现了去相关。
弹性网估计结合了岭回归和 Lasso 回归，岭回归的去相关步骤是多余的。因为尽管岭收缩可以有效控制方差，但是 Lasso 收缩既可以控制方差也可以得到稀疏解。为了避免双重收缩，因此对朴素弹性网乘上一个因子 $1+\lambda_2$。

3.3 Computations: the LARS-EN algorithm

前面已经提到了可以讲弹性网估计转化为等价的 Lasso 回归问题。LARS 是 Lasso 问题的求解算法。LARS-EN 算法是基于 LARS 的，利用了 X^* 矩阵的稀疏性对 LARS 进行了一些调整。

4 Discussion

1. 弹性网是一种收缩和变量选择回归方法，它产生的是一个稀疏模型，并具有群组效应。
2. 真实数据和数值模拟表明了弹性网预测性能比 Lasso 良好。
3. 选取合适的罚函数，弹性网惩罚也可以用于分类问题。
4. 弹性网和极大边际解释有密切的关系。
5. 可以把弹性网看成 Lasso 的一个拓展。