More about LASSO

王 成

上海交通大学数学科学学院

最小二乘法和LASSO

给定样本

$$(\mathbf{x}_1, y_1), \;\cdots, (\mathbf{x}_n, y_n),$$

其中$\mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^p$为解释变量， $y_1, \cdots, y_n \in \mathbb{R}$ 为响应变量.

$$\text{最小二乘法:} \quad \underset{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p}{\arg\min} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - \mathbf{x}_i^\top \beta)^2,$$

$$\text{LASSO:} \quad \underset{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p}{\arg\min} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - \mathbf{x}_i^\top \beta)^2 + \lambda |\beta|_1.$$

Section 1

glmnet

glmnet

Glmnet is a package that fits generalized linear and similar models via penalized maximum likelihood. The regularization path is computed for the LASSO or elastic net penalty at a grid of values (on the log scale) for the regularization parameter lambda.

glmnet

Glmnet is a package that fits generalized linear and similar models via penalized maximum likelihood. The regularization path is computed for the LASSO or elastic net penalty at a grid of values (on the log scale) for the regularization parameter lambda.

glmnet solves the problem:

$$\arg\min_{\beta_0,\beta} \frac{1}{n}\sum_{i=1}^{n} w_i \; l(y_i, \beta_0 + \beta^\top \mathbf{x}_i) + \lambda[(1-\alpha)\frac{\|\beta\|_2^2}{2} + \alpha\|\beta\|_1],$$

这里$w_i \geq 0$是权重, $l(\cdot,\cdot)$是损失函数(negative log-likelihood), $\lambda \geq 0$是调节参数, $\alpha \in [0,1]$是一个控制ℓ_1惩罚和ℓ_2惩罚的调节参数.

glmnet中的模型

- 线性模型(Linear Regression: family = "gaussian")

$$l(y_i, \beta_0 + \beta^\top \mathbf{x}_i) = \frac{1}{2}\{y_i - (\beta_0 + \beta^\top \mathbf{x}_i)\}^2;$$

glmnet中的模型

- 线性模型(Linear Regression: family = "gaussian")

$$l(y_i, \beta_0 + \beta^\top \mathbf{x}_i) = \frac{1}{2}\{y_i - (\beta_0 + \beta^\top \mathbf{x}_i)\}^2;$$

- 逻辑回归(Logistic Regression: family = "binomial")

$$l(y_i, \beta_0 + \beta^\top \mathbf{x}_i) = \log(1 + e^{\beta_0 + \mathbf{x}_i^\top \beta}) - y_i(\beta_0 + \mathbf{x}_i^\top \beta);$$

glmnet中的模型

- Poisson Regression: 假定数据生成机制服从Poisson分布

$$P(y_i = k) = \frac{\lambda_i^k}{k!} e^{-\lambda_i},\ \text{假定}\ \lambda_i = e^{\beta_0 + \mathbf{x}_i^\top \beta},$$

对应损失函数为:

$$l(y_i, \beta_0 + \beta^\top \mathbf{x}_i) = e^{\beta_0 + \mathbf{x}_i^\top \beta} - y_i(\beta_0 + \mathbf{x}_i^\top \beta);$$

glmnet中的模型

- Poisson Regression: 假定数据生成机制服从Poisson分布

$$P(y_i = k) = \frac{\lambda_i^k}{k!} e^{-\lambda_i}, \text{ 假定 } \lambda_i = e^{\beta_0 + \mathbf{x}_i^\top \beta},$$

对应损失函数为:

$$l(y_i, \beta_0 + \beta^\top \mathbf{x}_i) = e^{\beta_0 + \mathbf{x}_i^\top \beta} - y_i(\beta_0 + \mathbf{x}_i^\top \beta);$$

- Cox Regression: family = "cox":

$$l(y_i, \beta^\top \mathbf{x}_i) = \log\{\sum_{j: y_j \geq y_i} e^{\mathbf{x}_j^\top \beta}\} - \mathbf{x}_i^\top \beta.$$

glmnet中的模型

- Poisson Regression: 假定数据生成机制服从Poisson分布

$$P(y_i = k) = \frac{\lambda_i^k}{k!} e^{-\lambda_i}, \text{ 假定 } \lambda_i = e^{\beta_0 + \mathbf{x}_i^\top \beta},$$

对应损失函数为:

$$l(y_i, \beta_0 + \beta^\top \mathbf{x}_i) = e^{\beta_0 + \mathbf{x}_i^\top \beta} - y_i(\beta_0 + \mathbf{x}_i^\top \beta);$$

- Cox Regression: family = "cox":

$$l(y_i, \beta^\top \mathbf{x}_i) = \log\{\sum_{j: y_j \geq y_i} e^{\mathbf{x}_j^\top \beta}\} - \mathbf{x}_i^\top \beta.$$

- GLM families: family = family().

Section 2

其他惩罚函数

SCAD

Fan and Li (2001) 提出smoothly clipped absolute deviation (SCAD) penalty:

$$p_\lambda(\beta) = \begin{cases} \lambda|\beta| & |\beta| \leq \lambda, \\ \frac{2\alpha\lambda|\beta| - \beta^2 - \lambda^2}{2(\alpha-1)} & \lambda < |\beta| < \alpha\lambda, \\ \frac{\lambda^2(\alpha+1)}{2} & |\beta| \geq \alpha\lambda. \end{cases}$$

其中$\alpha > 2$是一个给定的常数,一般$\alpha = 3.7$.

SCAD

Fan and Li (2001) 提出smoothly clipped absolute deviation (SCAD) penalty:

$$p_\lambda(\beta) = \begin{cases} \lambda|\beta| & |\beta| \le \lambda, \\ \frac{2\alpha\lambda|\beta| - \beta^2 - \lambda^2}{2(\alpha-1)} & \lambda < |\beta| < \alpha\lambda, \\ \frac{\lambda^2(\alpha+1)}{2} & |\beta| \ge \alpha\lambda. \end{cases}$$

其中$\alpha > 2$是一个给定的常数,一般$\alpha = 3.7$.

对应$\arg\min_\beta \frac{1}{2}(x-\beta)^2 + p_\lambda(\beta)$的解为:

$$\hat{\beta} = \begin{cases} sgn(x)(|x| - \lambda)_+ & |x| \le 2\lambda, \\ \frac{(\alpha-1)x - sgn(x)a\lambda}{\alpha-2} & 2\lambda < |x| < \alpha\lambda, \\ x & |x| \ge \alpha\lambda. \end{cases}$$

MCP

Zhang (2010)提出minimax concave penalty (MCP):

$$p_\lambda(\beta) = \begin{cases} \lambda|\beta| - \frac{\beta^2}{2\gamma} & |\beta| \le \gamma\lambda, \\ \frac{1}{2}\gamma\lambda^2 & |\beta| > \gamma\lambda, \end{cases}$$

其中$\gamma > 1$是一个给定的常数.

MCP

Zhang (2010)提出minimax concave penalty (MCP):

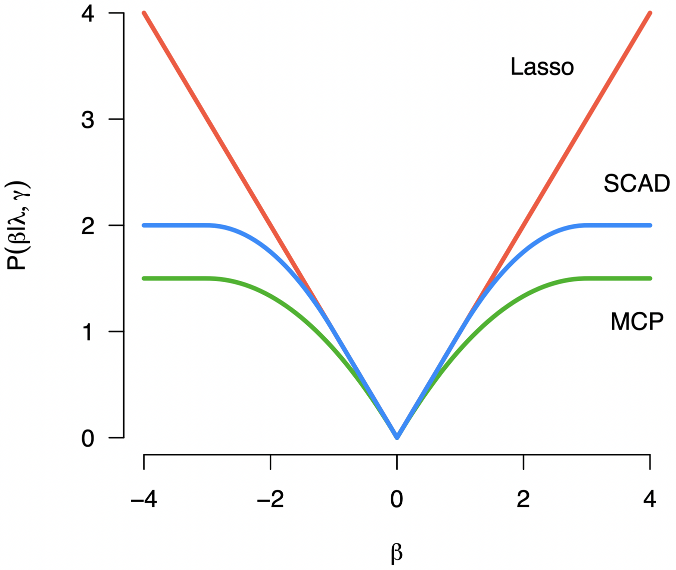
$$p_\lambda(\beta) = \begin{cases} \lambda|\beta| - \frac{\beta^2}{2\gamma} & |\beta| \le \gamma\lambda, \\ \frac{1}{2}\gamma\lambda^2 & |\beta| > \gamma\lambda, \end{cases}$$

其中$\gamma > 1$是一个给定的常数.

对应$\arg\min_\beta \frac{1}{2}(x-\beta)^2 + p_\lambda(\beta)$的解为:

$$\hat{\beta} = \begin{cases} sgn(x)\frac{(|x|-\lambda)_+}{\gamma-1} & |x| \le \gamma\lambda, \\ x & |x| \ge \gamma\lambda. \end{cases}$$

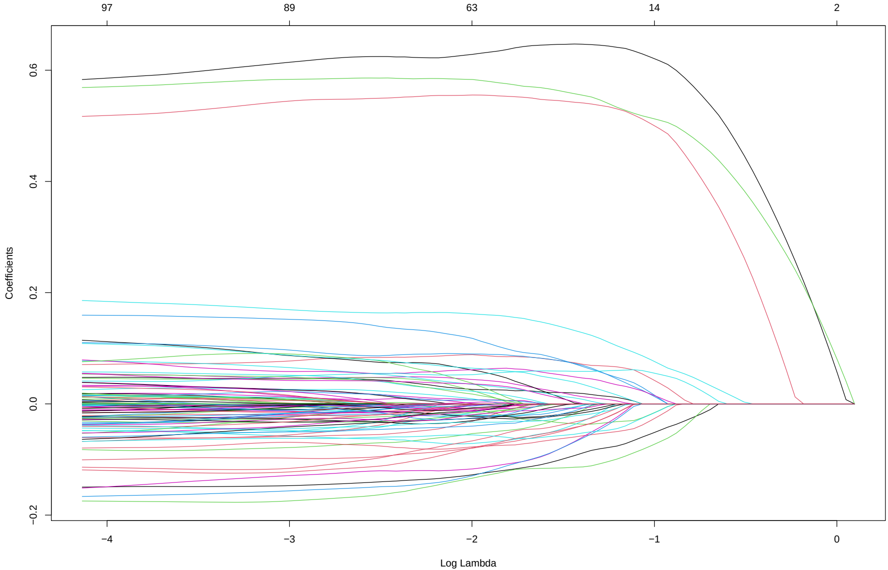
惩罚函数对比



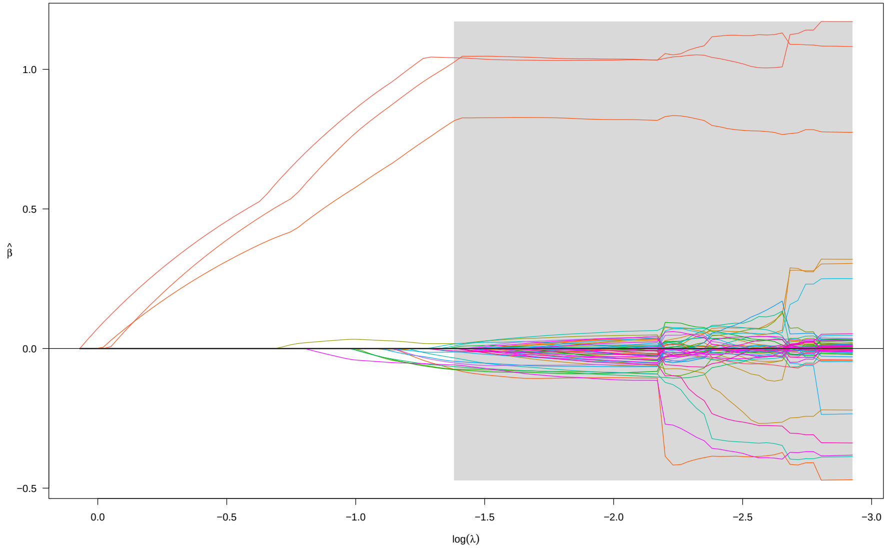
高维数据的线性回归

```
rm(list=ls())
set.seed(123)
library(glmnet)

## Loading required package:  Matrix
## Loaded glmnet 4.1-4

library('ncvreg')
n=100
p=10000
beta=c(1,1,1,rep(0,p-3))
X<-matrix(rnorm(n*p),nrow=p)
epsilon<-rnorm(n)
Y<-t(X)%*%beta+epsilon
```

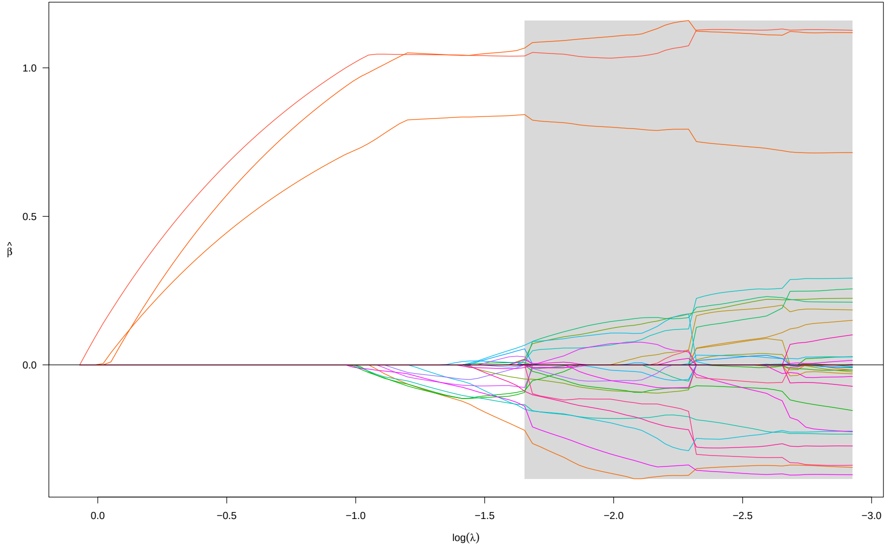
Solution path for LASSO



Solution path for SCAD



Solution path for MCP



Section 3

其他LASSO模型

Elastic net

Zou and Hastie (2005)提出Elastic Net方法:

$$\arg\min_{\beta_0,\beta} \frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta^\top \mathbf{x}_i)^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2,$$

其中$\lambda_1, \lambda_2 \geq 0$是调节参数.

Adaptive LASSO

Zou (2006)提出Adaptive LASSO方法:

$$\arg\min_{\beta_0,\beta} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta^\top \mathbf{x}_i)^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j|,$$

这里$w_j \geq 0$是权重.

Adaptive LASSO

Zou (2006)提出Adaptive LASSO方法:

$$\arg\min_{\beta_0,\beta} \frac{1}{2n}\sum_{i=1}^{n}(y_i-\beta_0-\beta^\top \mathbf{x}_i)^2+\lambda\sum_{j=1}^{p} w_j|\beta_j|,$$

这里$w_j \geq 0$是权重. 直观上，如果真正的$\beta_j \approx 0$, 我们可以设置w_j比较大，反之比较小. 例如:

$$w_j = 1/|\hat{\beta}_j|^\gamma,\ \gamma > 0.$$

Group LASSO

把$\{1, 2, \ldots, p\}$分成J个组

$$I_1 + \cdots + I_J = \{1, 2, \ldots, p\}$$

对于每个指标集I, 定义

$$\|\beta_I\|_2 = \sqrt{\sum_{i \in I} \beta_i^2},$$

Yuan and Lin (2006)提出Group LASSO模型:

$$\arg\min_{\beta_0, \beta} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta^\top \mathbf{x}_i)^2 + \lambda \sum_{j=1}^{J} \|\beta_{I_j}\|_2.$$

Group LASSO

把$\{1,2,\ldots,p\}$分成J个组

$$I_1+\cdots+I_J=\{1,2,\ldots,p\}$$

对于每个指标集I, 定义

$$\|\beta_I\|_2=\sqrt{\sum_{i\in I}\beta_i^2},$$

Yuan and Lin (2006)提出Group LASSO模型:

$$\arg\min_{\beta_0,\beta}\frac{1}{2n}\sum_{i=1}^{n}(y_i-\beta_0-\beta^\top\mathbf{x}_i)^2+\lambda\sum_{j=1}^{J}\|\beta_{I_j}\|_2.$$

Group LASSO分到同一个组的系数会出现同时为零或者非零的特点.

Fused LASSO

Tibshirani et al. (2005)提出Fused LASSO方法:

$$\arg\min_{\beta_0,\beta} \frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta^\top \mathbf{x}_i)^2 + \lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=2}^{p}|\beta_j - \beta_{j-1}|.$$

Fused LASSO

Tibshirani et al. (2005)提出Fused LASSO方法:

$$\arg\min_{\beta_0,\beta} \frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta^\top \mathbf{x}_i)^2 + \lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=2}^{p}|\beta_j - \beta_{j-1}|.$$

Fused LASSO的特点是回归系数非零部分会聚焦到同一段.

Section 4

LASSO的算法

LASSO的算法

对于LASSO问题，优化中针对ℓ_1有大量的研究，其中有代表性的算法

- ADMM: Alternating Direction Method of Multiplier
- FISTA: Beck and Teboulle(2009): A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems
- Cordinent Decent: glmnet

ADMM for LASSO I

$$\text{LASSO} \quad \underset{\beta\in\mathbb{R}^p}{\arg\min}\ \frac{1}{2n}\|(\mathbb{Y}-\mathbb{X}\beta)\|_2^2+\lambda|\beta|_1$$

等价的引入额外的变量

$$\min_{x,z\in\mathbb{R}^p}\frac{1}{2n}\|(\mathbb{Y}-\mathbb{X}x)\|_2^2+\lambda|z|_1,\ \text{s.t. } x-z=0.$$

ADMM for LASSO II

写出Augmented Lagrangian形式

$$L(x,z,u)=\frac{1}{2n}\|\mathbb{Y}-\mathbb{X}x\|_2^2+\lambda|z|_1+\frac{\rho}{2}\|x-z+u\|_2^2.$$

这里的$\rho>0$是ADMM中的常数.为了方便理解，这里写出unscaled form

$$L(x,z,u)=\frac{1}{2n}\|\mathbb{Y}-\mathbb{X}x\|_2^2+\lambda|z|_1+\rho u^\top(x-z)+\frac{\rho}{2}\|x-z\|_2^2.$$

ADMM for LASSO III

给定上一步值(x^k, z^k, u^k), 迭代过程

- $x^{k+1} = \arg\min_x L(x, z^k, u^k) = (\frac{1}{n}\mathbb{X}^\top\mathbb{X} + \rho I)^{-1}(\frac{1}{n}\mathbb{X}^\top\mathbb{Y} + \rho(z^k - u^k)).$
- $z^{k+1} = \arg\min_z L(x^{k+1}, z, u^k) = \mathrm{soft}_{\lambda/\rho}(x^{k+1} + u^k)$
- $u^{k+1} = u^k + x^{k+1} - z^{k+1}$

当两次迭代变化很小的时候可以停止迭代过程。这里的$\mathrm{soft}_\lambda(x)$是软阈值函数：

$$\mathrm{soft}_\lambda(x) = (x - \lambda)I(x > \lambda) + (x + \lambda)I(x < -\lambda).$$

LASSO算法

设置初始值$x^0 = z^0 = u^0 = (0, \dots, 0)$,

- $x^{k+1} = (\frac{1}{n}\mathbb{X}^\top\mathbb{X} + \rho I)^{-1}(\frac{1}{n}\mathbb{X}^\top\mathbb{Y} + \rho(z^k - u^k))$。
- $z^{k+1} = \mathrm{soft}_{\lambda/\rho}(x^{k+1} + u^k)$
- $u^{k+1} = u^k + x^{k+1} - z^{k+1}$,

可以设置迭代停止条件例如$\|z^{k+1} - z^k\|_2 \leq 1e-3$.

Thank you !

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society, Series B*, 67(1):91–108, 2005.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68 (1):49–67, 2006.

C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical Association*, 101(476):1418–1429, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67(2):301–320, 2005.