第七讲：线性模型选择及正则化

软件工程《机器学习》课程

李济洪，王瑞波

山西大学 · 软件学院
wangruibo@sxu.edu.cn

2019 年 10 月 22 日

提纲

上讲回顾：贝叶斯概念学习

基本概念

- 似然分布（极大似然估计）
- 先验概率
- 后验分布（极大后验估计）
- 后验预测分布
- 贝叶斯模型平均
- plug-in 模型

推断过程

$$p(x|D_n) = \sum_h p(x|h)p(h|D_n) \tag{1}$$

$$p(h|D_n) \propto p(h)p(D_n|h) \tag{2}$$

新课导入

多元线性回归算法

$$\mathrm{y} = \beta_0 + \beta_1 \mathrm{x}_1 + \ldots + \beta_\mathrm{p} \mathrm{x}_\mathrm{p} + \epsilon \tag{3}$$

其中，$\epsilon \sim \mathrm{N}(0, \sigma^2)$。

- 假定 $\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{p}$ 与 y 均为线性关系，且对 y 的预测均有作用。

实际情况

- x_i 与 y 的关系为非线性关系；
- x_i 对 y 没有预测作用。

因此，如何从 p 个候选预测变量 $\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{p}$ 中选择重要的预测变量，构建具有高预测性能、高可解释性的线性回归模型？

Credit 数据集

```
> library(ISLR)
Warning message:
package 'ISLR' was built under R version 3.4.4
> Credit
     ID  Income Limit Rating Cards Age Education Gender Student Married        Ethnicity Balance
1     1  14.891  3606    283     2  34        11   Male      No     Yes        Caucasian     333
2     2 106.025  6645    483     3  82        15 Female     Yes     Yes            Asian     903
3     3 104.593  7075    514     4  71        11   Male      No      No            Asian     580
4     4 148.924  9504    681     3  36        11 Female      No      No            Asian     964
5     5  55.882  4897    357     2  68        16   Male      No     Yes        Caucasian     331
6     6  80.180  8047    569     4  77        10   Male      No      No        Caucasian    1151
7     7  20.996  3388    259     2  37        12 Female      No      No African American     203
8     8  71.408  7114    512     2  87         9   Male      No      No            Asian     872
9     9  15.125  3300    266     5  66        13 Female      No      No        Caucasian     279
10   10  71.061  6819    491     3  41        19 Female     Yes     Yes African American    1350
```

问题描述

- **预测变量**：Income、Limit 等 10 个。
- **响应变量**：balance。
- **采用算法**：线性回归算法。

变量选择要解决的几个问题

1. p 个变量可以产生出多少种不同的模型？
2. 变量选择问题可归结为模型比较问题。
   - 需要比较多少组不同的模型？
   - 对于任意两组模型，如何进行客观的比较？
     - 如何合理地估计单个模型的预测性能？
     - 如何比较两个模型的预测性能估计值？

二模型预测性能比较方法

- ⊛ 直接比较点估计；
- ⊛ 基于置信区间进行比较；
- ⊛ 基于统计假设检验进行比较；

算法预测性能的几种估计

1. 交叉验证估计
   - 优点：使用验证集可有效地估计出算法预测性能。
   - 缺点：需要预留出验证集，缩小了训练集规模，可能造成算法欠拟合。
2. 使用训练集估计算法预测性能，例如：RSS、R^2 等。
   - 优点：使用整个数据集拟合模型；
   - 缺点：过优地估计了算法的预测性能（过拟合），易选取较为复杂（如，预测变量较多）的算法。

问题：是否还有其它合理的算法预测性能估计方法？

构造估计时另一个基本思路

在使用训练集估计算法预测性能时，兼顾算法的复杂度。

其它常用的预测性能估计

- C_p 估计；
- AIC: Akaike information criterion;
- BIC: Bayes information criterion;
- 调整 R^2: Adjusted R^2;

C_p 估计量

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2) \tag{4}$$

- RSS: **残差平方和**。
- d: **模型的大小，模型中预测变量的个数**；
- $\hat{\sigma}^2$: σ^2 **的估计值**；

AIC 准则（AKaike Information Criterion）

$$\mathrm{AIC} = \frac{1}{\mathrm{n}\hat{\sigma}^2}(\mathrm{RSS} + 2\mathrm{d}\hat{\sigma}^2) \tag{5}$$

特点

- 适用于许多使用极大似然法估计的模型；
- 与 $\mathrm{C_p}$ 估计量成比例；

BIC 准则（Bayes Information Criterion）

$$\mathrm{BIC} = \frac{1}{\mathrm{n}}(\mathrm{RSS} + \log(\mathrm{n})\mathrm{d}\hat{\sigma}^2) \tag{6}$$

BIC 的特点

- 相对于 AIC，BIC 对模型复杂度有更重的惩罚。$\log(\mathrm{n})>2$，当 $\mathrm{n} > 7$；
- 与 $\mathrm{C_p}$ 及 AIC 相同，BIC 加入了对模型复杂度的惩罚；

调整 R^2

$$\mathrm{R}^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \tag{7}$$

$$\text{Adjusted } \mathrm{R}^2 = 1 - \frac{\mathrm{RSS}(\mathrm{n}-\mathrm{d}-1)}{\mathrm{TSS}/(\mathrm{n}-1)} \tag{8}$$

调整 R^2 的特点

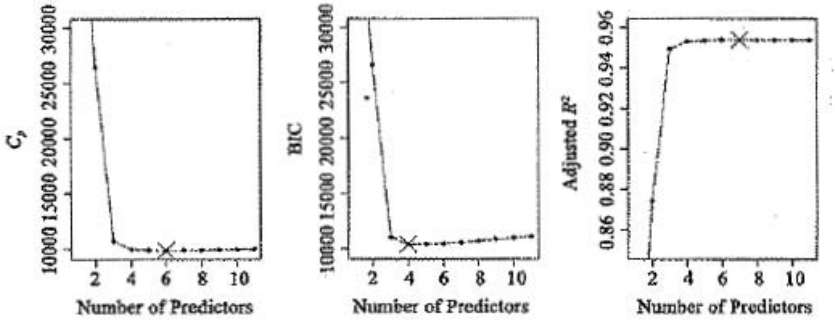
- 调整 R^2 将模型大小 d 融合到估计中。
- 越大越好，最大值为 1；
- 理论上讲，拥有最大调整 R^2 的值只包含了正确的预测变量，没有冗余变量；

验证方法与交叉验证

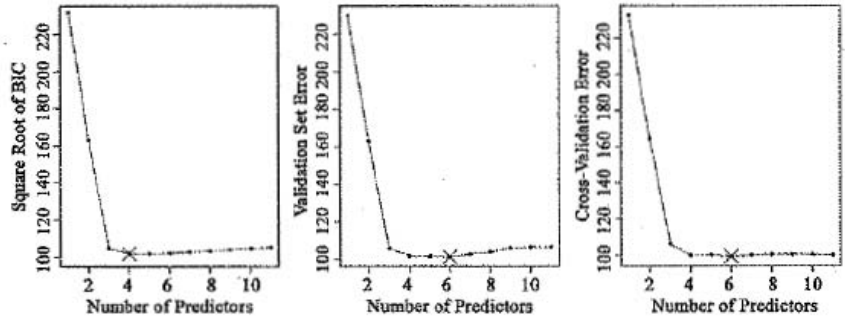
验证方法和交叉验证方法的特点

1. 使用验证集直接估计模型的预测性能；
2. 对潜在的模型有较少的假设；
3. 适用范围广
   - 模型复杂度难以确定时；
   - 误差的方差 σ^2 难以估计时；

C_p、BIC、调整 R^2 在 Credit 数据集上的比较



BIC、验证方法、交叉验证方法在 Credit 数据集上的比较



最优子集选择

基本思想

遍历所有可能的模型，基于预测性能的估计，找到预测性能最好的一组模型。

算法步骤

1. 记不含预测变量的零模型为 $\mathcal{M}_0$，用于估计样本均值；
2. 对于 $\mathrm{k} = 1, 2, \ldots, \mathrm{p}$:
   - 拟合 $\binom{\mathrm{p}}{\mathrm{k}}$ 个包含 k 个预测变量的模型；
   - 在 $\binom{\mathrm{p}}{\mathrm{k}}$ 个 m 模型中选择 RSS 最小或 R^2 最大的模型最为最优模型，记为 $\mathcal{M}_\mathrm{k}$；
3. 根据交叉验证估计、$\mathrm{C_p}$、AIC、BIC 或调整 R^2 从 $\mathcal{M}_0, \ldots, \mathcal{M}_\mathrm{p}$ 中选取一个最优模型；

最优子集选择的特点

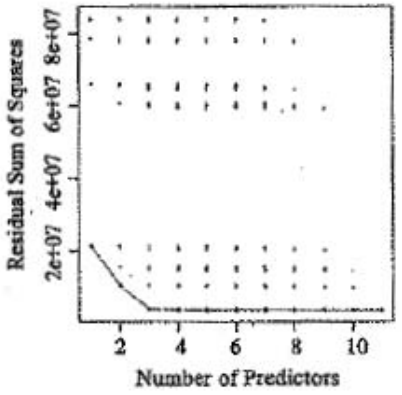
优点

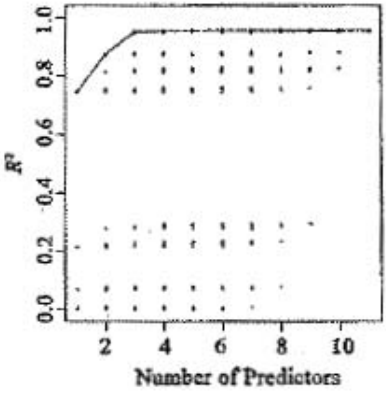
- 方法简单直观；
- 遍历了所有模型，更容易找到最优模型；

缺点

- 计算效率不高。
- $p = 10$ 对应 1000 多个候选模型；
- $p = 20$ 对应 100 多万种候选模型；

最优子集选择在 Credit 数据集上的应用





前向选择方法

基本思想

逐个添加最重要的变量，生成一个模型序列，从中甄别出最优模型。

算法步骤

1. 记不含预测变量的零模型为 $\mathcal{M}_0$；
2. 对于 $\mathrm{k} = 0, 1, \ldots, \mathrm{p} - 1$:
   - 从 $\mathrm{p} - \mathrm{k}$ 个模型中进行选择，每个模型都在模型 $\mathcal{M}_\mathrm{k}$ 的基础上增加一个变量；
   - 在 $\mathrm{p} - \mathrm{k}$ 个模型中选择 RSS 最小或 R^2 最高的模型作为最优模型，记为 $\mathcal{M}_{\mathrm{k}+1}$；
3. 根据交叉验证估计、$\mathrm{C_p}$、AIC、BIC 或调整 R^2 从 $\mathcal{M}_0, \ldots, \mathcal{M}_\mathrm{p}$ 中选取一个最优模型；

后向选择方法

基本思想

逐个剔除不重要的变量，生成一个模型序列，从中甄别出最优模型。

算法步骤

1. 记包含全部 p 个预测变量的全模型为 $\mathcal{M}_{\mathrm{p}}$；
2. 对于 $\mathrm{k} = \mathrm{p}, \mathrm{p}-1, \ldots, 1$:
   - 从 k 个模型中进行选择，在模型 $\mathcal{M}_{\mathrm{k}}$ 的基础上减少一个变量，在模型只含有 $\mathrm{k}-1$ 个变量；
   - 在 k 个模型中选择 RSS 最小或 R^2 最高的模型作为最优模型，记为 $\mathcal{M}_{\mathrm{k}-1}$；
3. 根据交叉验证估计、$\mathrm{C_p}$、AIC、BIC 或调整 R^2 从 $\mathcal{M}_0, \ldots, \mathcal{M}_{\mathrm{p}}$ 中选取一个最优模型；

前向选择和后向选择的特点

- 前向选择和后项选择的计算效率高。
  - 最优子集选择的计算效率为 2^p；
  - 前向选择和后向选择的计算效率为 $p(p+1)/2$；
- 前向选择适用于高维问题（$n < p$）；
- 后向选择必须要求 $p < n$，否则模型无法拟合。
- 相比最优子集选择，前向选择和后向选择为贪心算法，找到全局最优模型的概率较小。

逐步选择方法

基本思想

从空模型开始，逐个加入重要的变量，与此同时，判断模型中是否含有不重要的变量，若含有，则剔除该变量。多次执行该过程，生成一个模型序列，从中甄别出最优模型。

注意

该算法为前向和后向算法的结合使用，本课程不做深入探讨。

最优子集选择、前向选择在 Credit 数据集上的应用

变量个数	最优子集选择	前向选择
1	rating	rating
2	rating income	rating income
3	rating income student	rating income student
4	cards income student limit	rating income student limit

压缩估计方法的基本思想

最小二乘估计

$$\hat{\beta}_{\mathrm{LS}} = \arg\min_{\beta} \mathrm{RSS} = \arg\min_{\beta} \sum_{\mathrm{i}=1}^{\mathrm{n}} [\mathrm{y_i} - (\beta_0 + \beta_1 \mathrm{x_1} + \ldots + \beta_{\mathrm{p}} \mathrm{x_p})]^2 \tag{9}$$

压缩估计的基本思想

将多元线性回归算法的参数估计向 $\bigcirc$ 的方向压缩，提升参数估计的稳定性，减少估计的方差，进而改善模型的拟合效果。

- ⊛ 岭回归（Ridge Regression）
- ⊛ Lasso 回归（Lasso Regression）

岭回归 (Ridge Regression)

$$\hat{\beta}_{\text{ridge}} = \arg\min_{\beta} \text{RSS} = \arg\min_{\beta} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)]^2$$

$$\text{s.t.} \quad \sum_{j=1}^{p} \beta_j^2 \leq t \tag{10}$$

等价形式

$$\hat{\beta}_{\text{ridge}} = \arg\min_{\beta} \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

- RSS: **残差平方和**;
- λ: **拉格朗日乘子（超参数）**;

岭回归的特点

$$\hat{\beta}_{\text{ridge}} = \arg\min_{\beta} \text{RSS} + \lambda \sum_{\text{j}=1}^{\text{p}} \beta_{\text{j}}^2 \tag{11}$$

- 优化问题为权衡问题，λ 为调节参数；$\lambda \to \infty$ 时，参数估计趋于 0。
- 岭回归不对截距项参数 β_0 进行惩罚；
- 岭回归参数估计与预测变量的量纲有关，因此需对数据进行标准化。

$$\tilde{\text{x}}_{\text{ij}} = \frac{\text{x}_{\text{ij}}}{\sqrt{\frac{1}{\text{n}} \sum_{\text{i}=1}^{\text{n}} (\text{x}_{\text{ij}} - \bar{\text{x}}_{\text{j}})^2}} \tag{12}$$

岭估计

$$\hat{\beta}_{\mathrm{LS}} = (\mathrm{X}^\top \mathrm{X})^{-1}\mathrm{X}^\top \mathrm{y} \qquad (13)$$

$$\hat{\beta}_{\mathrm{ridge}} = (\mathrm{X}^\top \mathrm{X} + \lambda \mathrm{I})^{-1}\mathrm{X}^\top \mathrm{y} \qquad (14)$$

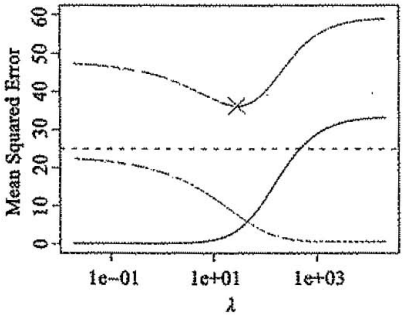
特点

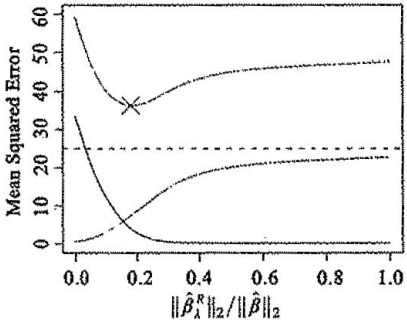
- 方差最小的线性无偏估计。
- 方差较大，特别是 $\mathrm{X}^\top \mathrm{X}$ 不可逆时，该估计失效。

特点

- 有偏估计，但方差较小。
- $\lambda \mathrm{I}$ 的作用是使得矩阵可逆，故增强了估计的稳定性。

岭回归的优势


Mean Squared Error
0
10
20
30
40
50
60
1e−01
1e+01
1e+03
λ


Mean Squared Error
0
10
20
30
40
50
60
0.0
0.2
0.4
0.6
0.8
1.0
$\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$

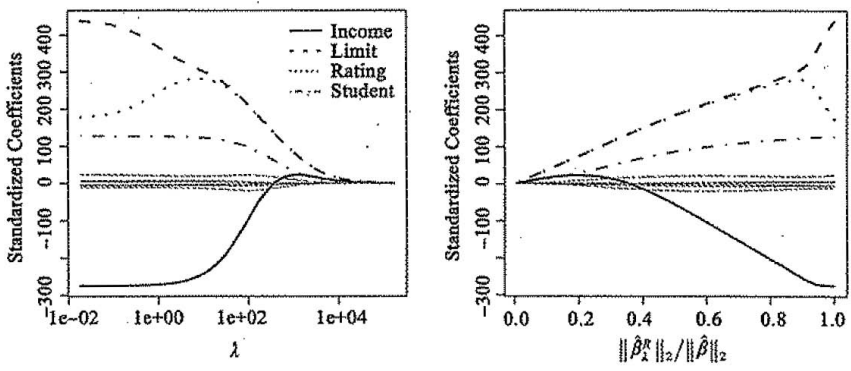
岭回归 v.s. 最优子集选择

对比

项目	岭回归	最优子集选择
计算效率	高	低
真模型的辨识度	低	高
优势领域	最小二乘估计方差大	p 较小

- 岭回归无法精准辨识真实变量，但可以压低预测变量的影响。

岭回归在 Credit 数据集上的应用



Lasso 回归 (Lasso Regression)

$$\hat{\beta}_{\text{ridge}} = \arg\min_{\beta} \text{RSS} = \arg\min_{\beta} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)]^2$$

$$\text{s.t.} \quad \sum_{j=1}^{p} |\beta_j| \leq t \tag{15}$$

等价形式

$$\hat{\beta}_{\text{ridge}} = \arg\min_{\beta} \text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|$$

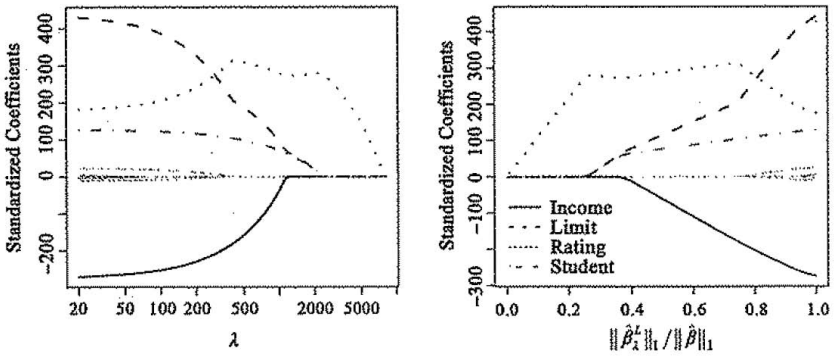
- RSS: **残差平方和**;
- λ: **拉格朗日乘子（超参数）**;

Lasso 估计的特点

Lasso 估计

- 向 0 的方向压缩
- 当 λ 足够大时，Lasso 估计精确为 0。
- Lasso 估计具有稀疏性。
- Lasso 回归有很好的可解释性。

Lasso 回归在 Credit 数据集上的应用



最优子集回归 v.s. 岭回归 v.s. Lasso 回归

最优子集选择问题

$$\min_\beta \mathrm{RSS} \quad \sum_{\mathrm{j}=1}^{\mathrm{p}} \mathrm{I}(\beta_\mathrm{j} \neq 0) \leq \mathrm{t}$$

Lasso 回归问题

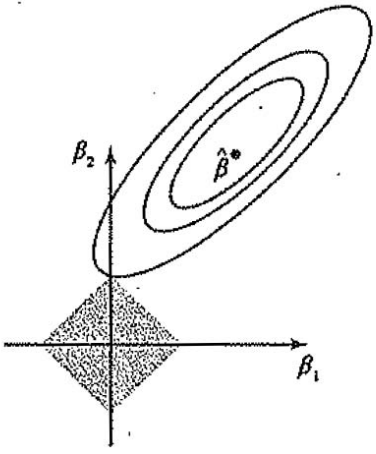
$$\min_\beta \mathrm{RSS} \quad \sum_{\mathrm{j}=1}^{\mathrm{p}} |\beta_\mathrm{j}| \leq \mathrm{t}$$

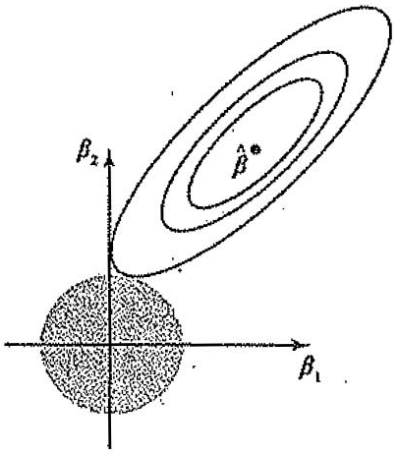
岭回归问题

$$\min_\beta \mathrm{RSS} \quad \sum_{\mathrm{j}=1}^{\mathrm{p}} \beta_\mathrm{j}^2 \leq \mathrm{t}$$

岭回归与 Lasso 回归的比较（续）

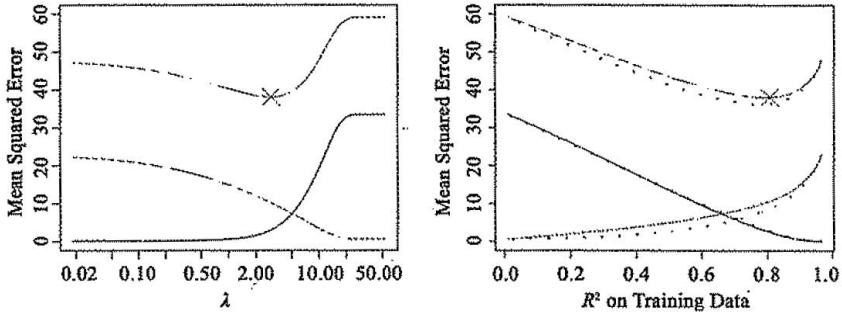
以 2 个预测变量为例





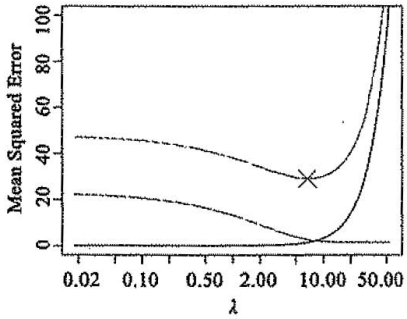
岭回归与 Lasso 回归的比较（续）

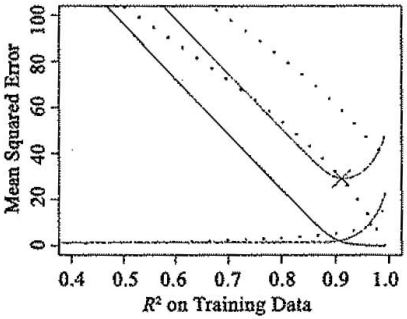
预测变量总数:45，**真实预测变量个数**：45



岭回归与 Lasso 回归的比较（续）

预测变量总数:45，**真实预测变量个数**：2





岭回归与 Lasso 回归的比较（续）

考虑特殊线性回归问题：$\mathrm{RSS} = \sum_{j=1}^{p}(y_j - \beta_j)^2$。

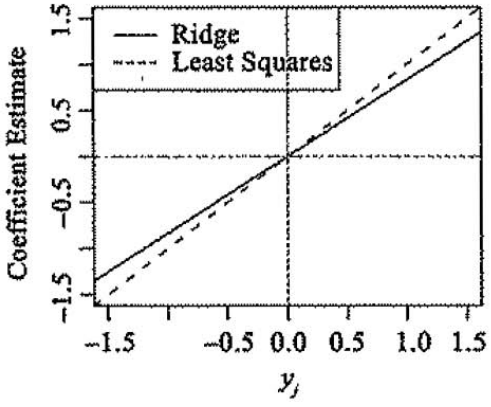
岭估计

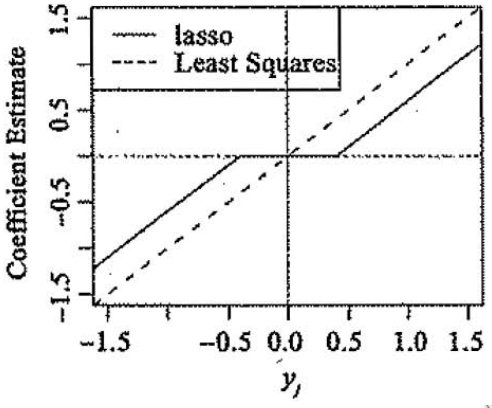
$$\hat{\beta}_{j,\mathrm{ridge}} = \frac{y_j}{1+\lambda} \tag{16}$$

Lasso 估计

$$\hat{\beta}_{j,\mathrm{lasso}} = \begin{cases} y_j - \frac{\lambda}{2}, & \text{if } y_j > \frac{\lambda}{2} \\ y_j + \frac{\lambda}{2}, & \text{if } y_j < \frac{\lambda}{2} \\ 0, & \text{if } |y_j| \le \frac{\lambda}{2} \end{cases} \tag{17}$$

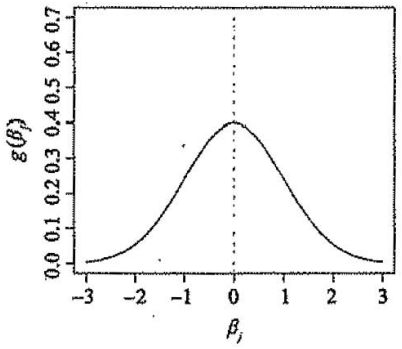
岭回归与 Lasso 回归的比较（续）

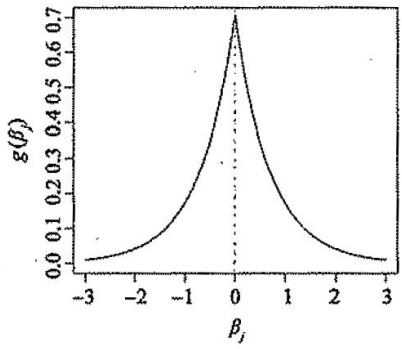




岭回归和 lasso 的贝叶斯解释

$$\mathrm{p}(\beta|\mathrm{X},\mathrm{Y}) \propto \mathrm{f}(\mathrm{Y}|\mathrm{X},\beta)\mathrm{p}(\beta|\mathrm{X}) = \mathrm{f}(\mathrm{Y}|\mathrm{X},\beta)\mathrm{p}(\beta)$$

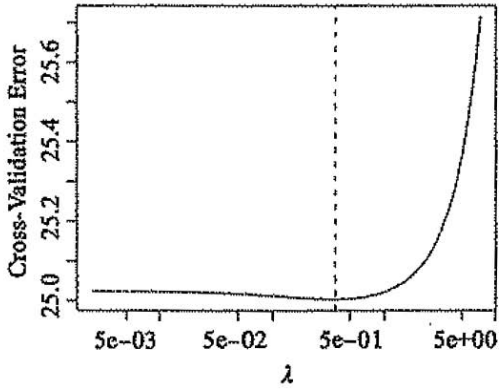


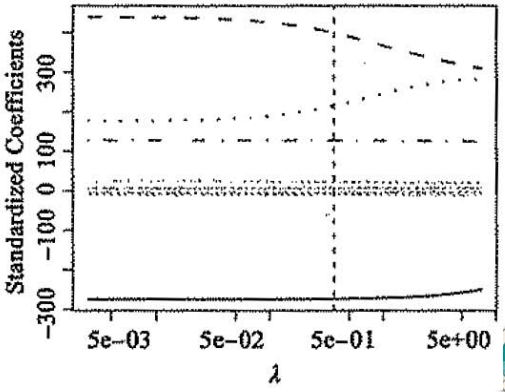


- 岭回归：高斯分布
- Lasso 回归：拉普拉斯分布

岭回归中 λ 的选择

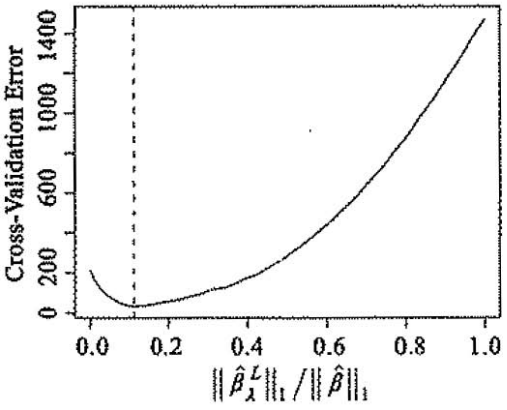
留一交叉验证

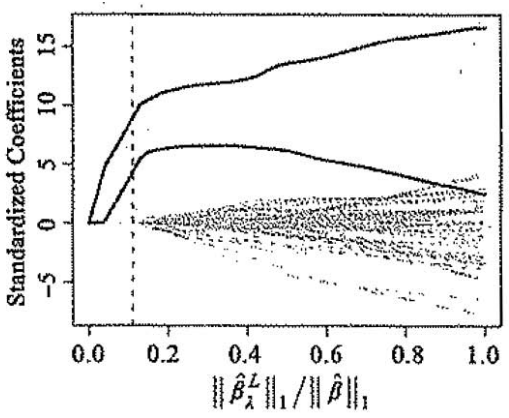




岭回归中 λ 的选择

10 折交叉验证





作业

1. 【推导】岭回归的参数估计形式。
2. 【思考】岭回归估计与极大后验估计的关系。
3. 【编程】岭回归方法和 Lasso 回归方法在 Credit 数据集上的应用实现。

谢谢！

Questions & Answering!