

逻辑回归模型的 Smooth LASSO 及 Spline LASSO 变量选择 *

戴 微* 金百锁

(中国科学技术大学管理学院统计与金融系, 合肥, 230026)

摘要: 对于逻辑回归模型中的参数估计和变量选择问题, 提出了 Smooth LASSO 以及 Spline LASSO. 当变量具有连续性, 使用 Smooth LASSO, 可以获得局部恒定的系数. 但是在有些情况下, 系数可能不同并且缓慢变化, 可以使用 Spline LASSO 来估计参数. 本文通过理论证明模型的可靠性, 利用坐标下降法对模型进行求解, 最后通过模拟验证了模型在变量选择中的准确性以及较好的预测性.

关键词: 逻辑回归; 变量选择; Smooth LASSO; Spline LASSO; 坐标下降法

中图分类号: O212.1

英文引用格式: DAI W, JIN B S. Variable selection for logistic regression via Smooth LASSO and Spline LASSO [J]. Chinese J Appl Probab Statist, 2019, 35(3): 292–304. (in Chinese)

§1. 引 言

作为广义线性模型^[1] 的特例, 逻辑回归具有良好的统计特性, 被广泛地应用到分类问题中^[2-4], 并取得了不错的效果. 然而, 传统的逻辑回归模型选出的变量不具有稀疏性, 而且和其他模型一样, 逻辑回归同样也会出现过拟合的情况. 在线性模型中, 为解决变量选择以及过拟合问题, 我们可以通过添加惩罚项来解决. 因此, 我们可以借鉴线性模型中的方法来解决广义线性模型中的问题.

在线性模型中, 为解决实际问题, 许多学者做了大量的研究. 为了选取稀疏的模型以及提高模型预测的准确率, Tibshirani^[5] 提出 LASSO, 取得了较好的效果, 并被广泛使用. Zou 和 Hastie^[6] 提出了弹性网络, 当多个特征有较强相关性时, 弹性网络表现得更好. 在某些特定领域, 比如基因表达数据上, 其特征按照一定的顺序进行排列, 也就是位置相邻的特征具有较强的相关性. 为解决这类问题, Tibshirani 等^[7] 提出了 Fused LASSO, 能够控制变量选择的稀疏性, 同时也能获取变量的组间效应. Hebiri 和 Van de Geer^[8] 提出了 Smooth LASSO, 该模型同样能获取组间的特征, 但是得到的特征更加光滑. Fused LASSO 和 Smooth LASSO 在每个组间更趋向于选取稳定的系数, 然而在一些情况下, 相邻

*国家自然科学基金面上项目(批准号: 11571337、71873128)和国家自然科学基金重点项目(批准号: 71631006)资助.

*通讯作者, E-mail: dw1992@mail.ustc.edu.cn.

本文 2018 年 3 月 27 日收到, 2018 年 6 月 30 日收到修改稿.

的特征虽然有较强的相关性, 但是系数并不稳定, 而是缓慢变化的, 这时候 Fused LASSO 和 Smooth LASSO 估出的系数不是很准确. 为解决这个问题, Guo 等^[9] 提出了 Spline LASSO, 在保证模型的稀疏性的同时, 也能获取光滑的系数.

由于正则化方法在解决变量选择以及参数估计问题上的有效性, 同样在广义线性模型中也可以引入正则化方法来解决传统广义线性模型的不足. Zhu 和 Hastie^[10] 提出了 L_2 正则化逻辑回归用来解决基因分类问题, 取得了不错的效果. Park 和 Hastie^[11] 提出了 L_1 正则化逻辑回归, 并给出了 path 算法. Bunea^[12] 证明了 L_1 和 $L_1 + L_2$ 正则化逻辑回归模型变量选择的有效性, 并给出了一定的条件, 在该条件下模型具有 Oracle 性质. 这些正则化方法被广泛地应用于各个领域, 并取得了很好的效果. 当变量具有较强的组间效应, 也就是相邻的变量间具有较强的相关性, 这些模型就达不到较好的结果. 因此受线性模型问题的启发, 本文将 Smooth LASSO 以及 Spline LASSO 中的惩罚项引入到逻辑回归模型中, 来克服变量选择问题中的组间效应. 理论上分析了模型的有效性, 并通过坐标下降法对模型进行求解, 通过仿真实验进行验证.

§2. 惩罚逻辑回归模型及理论特性

假设 n 组数据 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, 其中 $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^\top$, y_i 取值 $\{0, 1\}$ 来自总体 Y . 则逻辑回归模型为

$$\mathbb{E}(Y | X = \boldsymbol{x}_i) = \mathbb{P}(Y = 1) = 1/[1 + \exp(-\boldsymbol{x}_i^\top \boldsymbol{\beta})],$$

其中 $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^\top$ 为未知参数. 逻辑回归的求解问题可以表述为如下优化问题:

$$\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^n \{-y_i \boldsymbol{x}_i^\top \boldsymbol{\beta} + \ln[1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})]\}.$$

2.1 Smooth LASSO

将惩罚项 $2r \sum_{i=1}^p |\beta_j| + c \sum_{j=2}^p (\beta_j - \beta_{j-1})^2$ 引入到逻辑回归损失函数中, 其中 r 和 c 为调优参数, 可以得到 $\boldsymbol{\beta}$ 的估计值为

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^n \{-y_i \boldsymbol{x}_i^\top \boldsymbol{\beta} + \ln[1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})]\} + 2r \sum_{i=1}^p |\beta_j| + c \sum_{j=2}^p (\beta_j - \beta_{j-1})^2,$$

第一项惩罚为 LASSO 惩罚, 能够保证变量的稀疏性. 第二项惩罚系数一阶差分的 L_2 范数, 以获取局部恒定的系数.

令 p 为 $\boldsymbol{\beta}$ 的维数, $\hat{I} = \{j \in (1, 2, \dots, p) : \hat{\beta}_j \neq 0\}$, $I^* = \{j \in (1, 2, \dots, p) : \beta_j \neq 0\}$, k^* 为 $\boldsymbol{\beta}$ 中非零的个数, 对于设计矩阵 $X_{n \times p}$, 定义 $\Sigma = X^\top X/n$, 可以给出以下假设:

假设 1 对任意给定的 $\alpha, \epsilon > 0$, 定义一个集合:

$$V_{\alpha,\epsilon} = \left\{ v \in R^p : \sum_{j \notin I^*} |v_j| \leq \alpha \sum_{j \in I^*} |v_j| + \epsilon \right\},$$

对任意 $v \in V_{\alpha,\epsilon}$, 存在 $0 < b < 1$, 使得

$$\mathsf{P}\left(v^\top \Sigma v \geq b \sum_{j \in I^*} v_j^2 - \epsilon\right) = 1.$$

假设 2 存在 $B, D > 0$, 使得

$$\max_{j \in I^*} |\beta_j| \leq B, \quad \|\beta\|_1 \leq D.$$

假设 3 存在常数 $0 < d \leq 1$, 使得

$$\mathsf{P}\left(\max_{j \in I^*, k \neq j} |\rho_{kj}| \leq \frac{d}{k^*}\right) = 1.$$

假设 1 在许多文献中都被广泛使用, 通过选取合适的 b , 控制 X 的相关矩阵 Σ 的最小特征值, 从而保证其正定性. 从集合 $V_{\alpha,\epsilon}$ 的定义出发, 在定理 4 的证明中, 通过选取参数 α, ϵ , 可以使 $\beta - \hat{\beta} \in V_{\alpha,\epsilon}$, 假设 2 给出了系数的一些条件, 控制系数的界以及稀疏性. 假设 3 比假设 1 更加严格, 能够保证真实变量和无关变量的区分度, 用在定理 5 变量选择的相合性证明中.

设计矩阵 X 中的元素上界为 L , 存在 $\delta \in (0, 1)$, 令

$$\epsilon = \frac{\ln 2}{2(p \vee n) + 1} \times \frac{1}{r}, \quad \alpha = 7, \quad c = \frac{r}{16B}, \quad s = \frac{e^{2LD}}{(1 + e^{LD})^4},$$

下面定理给出 Smooth LASSO 方法的一些理论性质:

定理 4 假设 1 和 2 成立的情况下, 对于 $0 < b < 1$, 如果

$$r \geq 6L \sqrt{\frac{2 \ln 2(p \vee n)}{n}} + \frac{1}{2(p \vee n)} + 2L \sqrt{\frac{2 \ln \delta^{-1}}{n}},$$

则

$$\mathsf{P}\left[\|\hat{\beta} - \beta\|_1 \leq \frac{8rk^*}{bs} + \left(\frac{4}{r} + 2s\right)\epsilon\right] \geq 1 - \delta.$$

定理 5 令 $0 < \delta < 1$, 在假设 1 和 2 成立的条件下, 如果

$$\begin{aligned} \min_{j \in I^*} |\beta_j| &\geq \frac{8rk^*}{bs} + \left(\frac{4}{r} + 2s\right)\epsilon, \\ r &\geq 6L \sqrt{\frac{2 \ln 2(p \vee n)}{n}} + \frac{1}{2(p \vee n)} + 2L \sqrt{\frac{2 \ln \delta^{-1}}{n}}, \end{aligned}$$

则

$$\mathsf{P}(I^* \subset \hat{I}) \geq 1 - \delta.$$

注记 6 定理 4 表明, 当真实模型具有稀疏性, 选择较小的调优参数 r , 此时 $rk^* \rightarrow 0$, 当 $\delta \rightarrow 0$ 时, 与普通的 Lasso 方法比较, $\|\hat{\beta}_j - \beta\|_1$ 存在一个较小的界. 定理 5 表明变量选择的相合性, Smooth LASSO 方法能够以概率 1 挑选出所以的真实变量. 2.2 节中的定理 7 和 8 是上面定理的 Spline Lasso 形式, 与上述定理有相同的性质.

2.2 Spline LASSO

将惩罚项 $2r \sum_{i=1}^p |\beta_j| + c \sum_{j=2}^{p-1} [(\beta_{j+1} - \beta_j) - (\beta_j - \beta_{j-1})]^2$ 引入到逻辑回归损失函数中, 可以得到 β 的估计值为

$$\begin{aligned}\tilde{\beta} = & \arg \min_{\beta} \frac{1}{n} \sum_{i=1}^n \{-y_i \mathbf{x}_i^\top \beta + \ln[1 + \exp(\mathbf{x}_i^\top \beta)]\} \\ & + 2r \sum_{i=1}^p |\beta_j| + c \sum_{j=2}^{p-1} [(\beta_{j+1} - \beta_j) - (\beta_j - \beta_{j-1})]^2.\end{aligned}$$

第一项惩罚为 LASSO 惩罚, 能够保证变量的稀疏性. 第二项惩罚系数二阶差分的 L_2 范数, 能够得到缓慢变化的系数.

任意 $\delta \in (0, 1)$, 令

$$\epsilon = \frac{\ln 2}{2(p \vee n) + 1} \times \frac{1}{r}, \quad \alpha = 7, \quad c = \frac{r}{32B}, \quad s = \frac{e^{2LD}}{(1 + e^{LD})^4}.$$

下面给出 Spline LASSO 方法的一些理论特性:

定理 7 假设 1 成立的情况下, 对于 $0 < b < 1$, 如果

$$r \geq 6L \sqrt{\frac{2 \ln 2(p \vee n)}{n}} + \frac{1}{2(p \vee n)} + 2L \sqrt{\frac{2 \ln \delta^{-1}}{n}},$$

则

$$\mathbb{P} \left[\|\tilde{\beta} - \beta\|_1 \leq \frac{8rk^*}{bs} + \left(\frac{4}{r} + 2s \right) \epsilon \right] \geq 1 - \delta.$$

定理 8 令 $0 < \delta < 1$, 在假设 1 和 2 成立的条件下, 如果

$$\begin{aligned}\min_{j \in I^*} |\beta_j| &\geq \frac{8rk^*}{bs} + \left(\frac{4}{r} + 2s \right) \epsilon, \\ r &\geq 6L \sqrt{\frac{2 \ln 2(p \vee n)}{n}} + \frac{1}{2(p \vee n)} + 2L \sqrt{\frac{2 \ln \delta^{-1}}{n}},\end{aligned}$$

则

$$\mathbb{P}(I^* \subset \widehat{I}) \geq 1 - \delta.$$

§3. 仿真分析

通过仿真实验来分析 Smooth LASSO 和 Spline LASSO 方法的表现，并将其与 L_1 和 $L_1 + L_2$ 惩罚的逻辑回归模型进行比较。

自变量 x_1, x_2, \dots, x_{p-1} 服从多元正态分布 $N(0, \Sigma)$ ，响应变量来自于二项分布 $B(1, P)$ ，其中 $P = 1/[1+\exp(-\mathbf{x}\beta)]$, $\mathbf{x} = (1, x_1, x_2, \dots, x_{p-1})$, $\beta = (\beta_1, \beta_2, \dots, \beta_p)^\top$, 选取 $n = 1000$, $p = 50$ 以及两组不同的 β ,

$$\begin{aligned}\beta_1 &= I_{(0 < j \leq 15)} - 0.2(j - 20)I_{(15 < j \leq 20)} + 0 * I_{(20 < j \leq 50)}, \\ \beta_2 &= [-0.02(j - 10)^2 + 2]I_{(0 < j \leq 20)} + 0 * I_{(20 < j \leq 50)},\end{aligned}$$

图 1、图 3 为两组 β 的分布图。对于方差结构 Σ 的选取，也考虑了两种不同的情形。情形 1: $\Sigma = I$ (相邻变量没有相关性); 情形 2: $\Sigma_{ij} = 0.5^{|i-j|}$ (相邻变量有较强的相关性)。所有的调优参数都是通过 10 折交叉验证选取最小偏差 (Deviance) 得到。

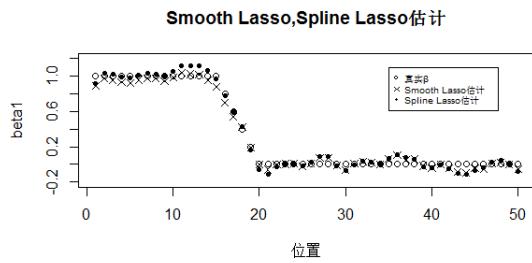


图 1 $\Sigma = I$ 时 β_1 以及 β_1 的估计值

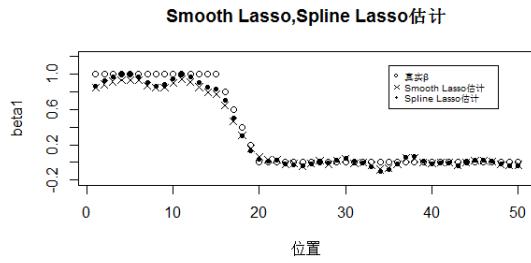


图 2 $\Sigma_{ij} = 0.5^{|i-j|}$ 时 β_1 以及 β_1 的估计值

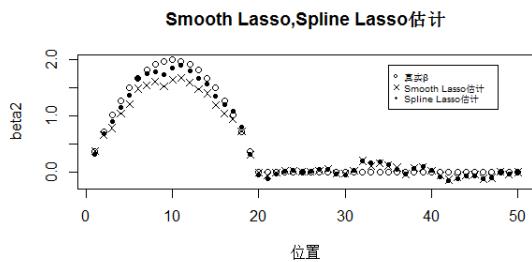


图 3 $\Sigma = I$ 时 β_2 以及 β_2 的估计值

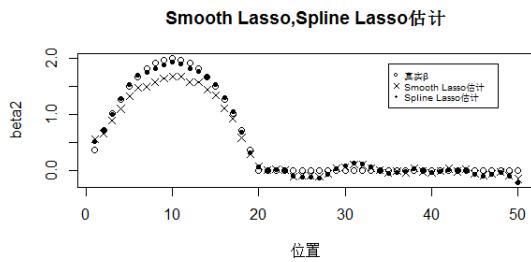


图 4 $\Sigma_{ij} = 0.5^{|i-j|}$ 时 β_2 以及 β_2 的估计值

对于不同的模型表现，我们给出以下几个评价标准：

- 1) 预测准确率：我们生成 1000 组测试样本，用训练集训练出的模型对测试样本的响应变量 Y 进行预测，从而得到每个模型的预测准确率。
- 2) 变量选择：对于不同的模型，可以通过模型系数的敏感度和特异度来判断模型的变量选择能力。定义集合 $A_0 = \{j \in \{1, 2, \dots, p\} : \beta_j \neq 0\}$, $B_0 = \{j \in \{1, 2, \dots, p\} : \beta_j = 0\}$,

$A_1 = \{j \in A_0 : \hat{\beta}_j \neq 0\}$, $B_1 = \{j \in B_0 : \hat{\beta}_j = 0\}$, 敏感度 = $|A_1|/|A_0|$ 为实际非零的系数被该模型挑选出来的部分, 特异度 = $|B_1|/|B_0|$ 为实际为零的系数被该模型剔除的部分, 其中 $|A_0|$ 表示集合 A_0 中元素的个数.

3) 模型估计误差: 用模型系数的估计值和真实值的 L_2 范数 $\|\hat{\beta} - \beta\|_2$ 作为估计误差, 模型估计误差越小, 越接近真实的模型.

从仿真的结果可以看到 Smooth LASSO 以及 Spline LASSO 都能得到平滑的系数, 也有很好的预测准确率. 但是不能获得较好的稀疏性, 对于许多真实值为零的系数, 会得到许多较小的非零估计. 这是因为调优参数是由交叉验证选取出来的, 交叉验证的选取指标是模型的偏差, 保证系数的平滑性时, 使得 L_1 惩罚的调优参数较小, 而得不到较好的稀疏性. 为解决这个问题, 可以设置一个阈值, 在用 Smooth LASSO 和 Spline LASSO 估出系数后, 把绝对值小于阈值的系数用零代替, 从而提高变量选择能力. 如何选择阈值 T , 一种可行的方法就是用交叉验证来选取, 但交叉验证计算量偏大. 另一种是用 Donoho 和 Johnstone^[13] 提出的 $\pm\hat{\sigma}\sqrt{2\ln p}$ 作为阈值, 其中 $\hat{\sigma}$ 可以通过系数的标准差估计得到. 因此这里我们选用 $T = \pm\hat{\sigma}\sqrt{2\ln p}$ 作为阈值来估计系数, 首先将估计出来的系数 $\hat{\beta}$ 的绝对值从小到大排序, 用前面的一半系数的标准差来估计 $\hat{\sigma}$.

结果分析:

从 β 的估计值可以看出, 用 Smooth LASSO 以及 Spline LASSO 方法均能取得不错的效果, 对系数的估计都很接近真实值. 从图 1、图 2 的对比可以看出, 当相邻特征间没有相关性时, Smooth LASSO 要好于 Spline LASSO; 当相邻特征间有较强的相关性时, 两种方法相差不大. 从图 3、图 4 可以看出, 在两种不同的方差结构下 Spline LASSO 均优于 Smooth LASSO.

从下表中各个模型的表现来看, 在两组不同 β 以及不同的方差结构下, 应用 Smooth LASSO 以及 Spline LASSO 方法进行模型估计以及预测准确率均优于传统惩罚逻辑回归. 从特异度可以看出 Smooth LASSO 和 Spline LASSO 虽然稀疏性不够, 但通过阈值控制可以克服这个问题, 同时还能提高模型的预测准确率. 通过对比不同的情况, 也能发现当 β 变化趋势较大时 Spline LASSO 在各方面均优于 Smooth LASSO.

§4. 结论与展望

为解决逻辑回归中的变量选择问题, 当变量具有连续性时, 提出了 Smooth LASSO 以及 Spline LASSO 方法, 该模型均优于传统惩罚逻辑回归. 并且通过阈值控制, 在提高模型的稀疏性同时, 也能提高模型的预测能力. Smooth LASSO 趋向于选取局部恒定的系数, 而 Spline LASSO 能够选取缓慢变化的系数. 本文只考虑了逻辑回归的问题, 从理论分析可以看出, 将 Smooth LASSO 和 Spline LASSO 惩罚项应用于一般的广义线性模型同样具有良好的特性. 为进一步提高模型的变量选择效果以及预测准确率, 除了发展新的数据分

表 1 $\beta = \beta_1, \Sigma = I$ 时各模型的表现

模型	Lasso	EN	Smooth LASSO	Spline LASSO	EN+阈值	Smooth+阈值	Spline+阈值
$\ \hat{\beta} - \beta\ _2$	0.98	0.70	0.36	0.42	0.65	0.32	0.39
敏感度	1	1	1	1	1	1	1
特异度	0.645	0.484	0.452	0.516	0.871	0.903	0.871
准确率	0.869	0.859	0.866	0.863	0.858	0.869	0.869

表 2 $\beta = \beta_1, \Sigma_{ij} = 0.5^{|i-j|}$ 时各模型的表现

模型	Lasso	EN	Smooth LASSO	Spline LASSO	EN+阈值	Smooth+阈值	Spline+阈值
$\ \hat{\beta} - \beta\ _2$	1.43	1.16	0.58	0.44	1.11	0.57	0.44
敏感度	1	1	1	1	0.947	1	1
特异度	0.742	0.581	0.548	0.581	0.871	0.871	0.871
准确率	0.910	0.909	0.916	0.916	0.914	0.917	0.917

表 3 $\beta = \beta_2, \Sigma = I$ 时各模型的表现

模型	Lasso	EN	Smooth LASSO	Spline LASSO	EN+阈值	Smooth+阈值	Spline+阈值
$\ \hat{\beta} - \beta\ _2$	2.29	2.00	1.20	0.69	1.97	1.20	0.66
敏感度	1	1	1	1	1	1	1
特异度	0.677	0.613	0.452	0.516	0.935	0.774	0.839
准确率	0.908	0.907	0.913	0.917	0.909	0.913	0.919

表 4 $\beta = \beta_2, \Sigma_{ij} = 0.5^{|i-j|}$ 时各模型的表现

模型	Lasso	EN	Smooth LASSO	Spline LASSO	EN+阈值	Smooth+阈值	Spline+阈值
$\ \hat{\beta} - \beta\ _2$	2.82	1.75	1.04	0.50	1.68	1.03	0.45
敏感度	1	1	1	1	1	1	1
特异度	0.806	0.581	0.581	0.613	0.871	0.903	0.903
准确率	0.947	0.945	0.950	0.949	0.945	0.950	0.950

析方法, 还可以从试验设计的角度进行考虑. 因此当自变量可控时, 可以通过安排切片正交试验设计^[17,18] 来进一步提高 Smooth LASSO 和 Spline LASSO 的效果.

§5. 定理证明

5.1 定理 4

逻辑回归损失函数:

$$l(\beta) = l(\beta; x, y) = -y\beta'x + \ln(1 + \exp \beta'x).$$

定义风险函数为 $\mathbb{P} l(\beta) = \mathbb{E} l(\beta; Y, X)$, 经验风险为

$$\mathbb{P}_n l(\beta) = \frac{1}{n} \sum_{i=1}^n \left[-Y_i \sum_{j=1}^p \beta_j X_{ij} + \ln \left(1 + \exp \sum_{j=1}^p \beta_j X_{ij} \right) \right].$$

使用 Smooth LASSO 方法得到的 β 估计值为

$$\hat{\beta} = \arg \min_{\beta} \mathbb{P}_n l(\beta) + 2r \|\beta\|_1 + c \sum_{j=2}^p (\beta_j - \beta_{j-1})^2.$$

β 的估计值满足下式

$$\mathbb{P}_n l(\hat{\beta}) + 2r \sum_{j=1}^p |\hat{\beta}_j| + c \sum_{j=2}^p (\hat{\beta}_j - \hat{\beta}_{j-1})^2 \leq \mathbb{P}_n l(\beta) + 2r \sum_{j=1}^p |\beta_j| + c \sum_{j=2}^p (\beta_j - \beta_{j-1})^2,$$

两边同时加上 $\mathbb{P}[l(\hat{\beta}) - l(\beta)] + r \sum_{j=1}^p |\beta_j|$, 移项可得

$$\begin{aligned} & r \|\hat{\beta} - \beta\|_1 + \mathbb{P}[l(\hat{\beta}) - l(\beta)] + c \sum_{j=2}^p (\hat{\beta}_j - \hat{\beta}_{j-1})^2 - c \sum_{j=2}^p (\beta_j - \beta_{j-1})^2 \\ & \leq (\mathbb{P}_n - \mathbb{P})[l(\beta) - l(\hat{\beta})] + r \|\hat{\beta} - \beta\|_1 + 2r \sum_{j=1}^p |\beta_j| - 2r \sum_{j=1}^p |\hat{\beta}_j|. \end{aligned}$$

令

$$L_n = \sup_{\beta' \in \mathbb{R}^p} \frac{(\mathbb{P}_n - \mathbb{P})[l(\beta) - l(\beta')]}{|\beta' - \beta|_1 + \epsilon},$$

利用文献 [15] 中 bounded difference inequality 可得

$$\mathbb{P}(L_n - EL_n \geq u) \leq \exp \left(-\frac{nu^2}{8L^2} \right).$$

令 $u = 2L\sqrt{2 \ln \delta^{-1}/n}$, 则

$$\mathbb{P}(L_n - EL_n \geq u) \leq \delta.$$

由文献 [16] 中的引理 3 可得

$$EL_n \leq 6L \sqrt{\frac{2 \ln 2(p \vee n)}{n}} + \frac{1}{2(p \vee n)}.$$

定义事件 $E = \{L_n \leq r\}$, 如果

$$r \geq 6L \sqrt{\frac{2 \ln 2(p \vee n)}{n}} + \frac{1}{2(p \vee n)} + 2L \sqrt{\frac{2 \ln \delta^{-1}}{n}},$$

$$\begin{aligned} \mathbb{P}(E) &= \mathbb{P}(L_n \leq r) = \mathbb{P}(L_n - EL_n \leq r - EL_n) \geq \mathbb{P}\left(L_n - EL_n \leq 2L \sqrt{\frac{2 \ln \delta^{-1}}{n}}\right) \\ &= 1 - \mathbb{P}(L_n - EL_n > u) \geq 1 - \delta. \end{aligned}$$

令

$$J = \begin{bmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix}_{(p-1) \times p},$$

$K = J^\top J \geq 0$, 在假设 $\max_{j \in I^*} |\beta_j| \leq B$ 下可得

$$\begin{aligned} \sum_{j=2}^p (\hat{\beta}_j - \hat{\beta}_{j-1})^2 - \sum_{j=2}^p (\beta_j - \beta_{j-1})^2 &= \hat{\beta}^\top K \hat{\beta} - \beta^\top K \beta = (\hat{\beta} - \beta)^\top K (\hat{\beta} - \beta) + 2(\hat{\beta} - \beta)^\top K \beta \\ &\geq 2(\hat{\beta} - \beta)^\top K \beta \geq -8B\|\hat{\beta} - \beta\|_1. \end{aligned}$$

由文献 [17] 中的例 4.5 可知 $\mathbb{P} l(\hat{\beta}) - \mathbb{P} l(\beta) \geq \|g_{\hat{\beta}} - g_\beta\|^2$, 其中 $g_\beta = \exp(x\beta)/[1 + \exp(x\beta)]$, $\|\cdot\|$ 为 L_2 范数. 记 $f_\beta(x) = \beta^\top x$, 将 g_β 进行一阶泰勒展开

$$g_{\hat{\beta}} - g_\beta = \frac{\exp(x\beta^*)}{[1 + \exp(x\beta^*)]^2} (f_{\hat{\beta}} - f_\beta),$$

令 $s = e^{2LD}/(1 + e^{LD})^4$, 可以得到

$$\mathbb{P} l(\hat{\beta}) - \mathbb{P} l(\beta) \geq s\|f_{\hat{\beta}} - f_\beta\|_2^2.$$

因此, 当事件 E 成立时, 可以得出

$$\begin{aligned} &r\|\hat{\beta} - \beta\|_1 + s\|f_{\hat{\beta}} - f_\beta\|^2 + c \sum_{j=2}^p (\hat{\beta}_j - \hat{\beta}_{j-1})^2 - c \sum_{j=2}^p (\beta_j - \beta_{j-1})^2 \\ &\leq 2r\|\hat{\beta} - \beta\|_1 + 2r \sum_{j=1}^p |\beta_j| - 2r \sum_{j=1}^p |\hat{\beta}_j| + r\epsilon \leq 4r \sum_{j \in I^*} |\hat{\beta}_j - \beta_j| + r\epsilon. \end{aligned}$$

令 $\gamma_{kj} = EX_k X_j$, 其中 $k, j \in \{1, 2, \dots, p\}$, 令 Γ 为 γ_{kj} 组成的 $p \times p$ 阶矩阵, 所以 $\|f_{\hat{\beta}} - f_\beta\|^2 = (\hat{\beta} - \beta)^\top \Gamma (\hat{\beta} - \beta)$, 进一步可知

$$r\|\hat{\beta} - \beta\|_1 + s(\hat{\beta} - \beta)^\top \Gamma (\hat{\beta} - \beta) - 8cB\|\hat{\beta} - \beta\|_1 \leq 4r \sum_{j \in I^*} |\hat{\beta}_j - \beta_j| + r\epsilon.$$

通过选取 c , 使得 $16cB = r$, 可得

$$\frac{r}{2}\|\hat{\beta} - \beta\|_1 + s(\hat{\beta} - \beta)^\top \Gamma (\hat{\beta} - \beta) \leq 4r \sum_{j \in I^*} |\hat{\beta}_j - \beta_j| + r\epsilon,$$

由上式可以得出

$$\sum_{j \notin I^*} |\hat{\beta}_j - \beta_j| \leq 7 \sum_{j \in I^*} |\hat{\beta}_j - \beta_j| + 2\epsilon.$$

所以 $(\hat{\beta} - \beta) \in V$, 其中 $\alpha = 7$, 由假设 1 可知:

$$(\hat{\beta} - \beta)^\top \Gamma (\hat{\beta} - \beta) \geq b \sum_{j \in I^*} (\hat{\beta}_j - \beta_j)^2 - 2\epsilon,$$

代入上式可得

$$\begin{aligned} \frac{r}{2} \|\hat{\beta} - \beta\|_1 + bs \sum_{j \in I^*} (\hat{\beta}_j - \beta_j)^2 &\leq 4r \sum_{j \in I^*} |\hat{\beta}_j - \beta_j| + (r + 2s)\epsilon \\ &\leq 4ar^2 k^* + \frac{1}{a} \sum_{j \in I^*} (\hat{\beta}_j - \beta_j)^2 + (r + 2s)\epsilon. \end{aligned}$$

上式右边由 Cauchy-Schwarz 不等式 $2x^\top y \leq 2\sqrt{x^2 y^2} \leq ax^2 + y^2/a$ 得到, 其中 x, y 为列向量. 取 $a = 1/b s$, 在事件 E 成立时可得 $\|\hat{\beta} - \beta\|_1 \leq 8rk^*/(bs) + (4/r + 2s)\epsilon$, 所以

$$\mathbb{P}\left[\|\hat{\beta} - \beta\|_1 \leq \frac{8rk^*}{bs} + \left(\frac{4}{r} + 2s\right)\epsilon\right] \geq 1 - \delta.$$

5.2 定理 5

由 I^* 以及 \hat{I} 的定义可知

$$\mathbb{P}(I^* \not\subset \hat{I}) \leq \mathbb{P}(k \in I^* : k \notin \hat{I}) \leq k^* \max_{k \in I^*} \mathbb{P}(\hat{\beta}_k = 0, \beta_k \neq 0).$$

由文献 [12] 中的命题 3.4 可得

$$\begin{aligned} \mathbb{P}(\hat{\beta}_k = 0, \beta_k \neq 0) &\leq \mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n W_i X_{ik}\right| \geq \frac{r}{2}\right) \\ &+ \mathbb{P}\left[\sum_{j=1}^p |\hat{\beta}_j - \beta_j| \left|\frac{1}{n} \sum_{i=1}^n X_{ij} X_{ik}\right| > \frac{r}{2} + \left(1 + \frac{1}{r}\right)\epsilon\right] \\ &+ \mathbb{P}\left[|S_n - B_n| \geq \frac{r}{2} + \left(1 + \frac{1}{r}\right)\epsilon\right], \end{aligned}$$

其中 $W_i = Y_i - p(X_i)$, $B_n = \sum_{j=1}^p (\hat{\beta}_j - \beta_j)(n^{-1} \sum_{i=1}^n X_{ij} X_{jk})$, 并且满足 $n^{-1} \sum X_{ik}^2 = 1$,

$$S_n = \frac{1}{n} \sum_{i=1}^n X_{ik} \left(\frac{\exp \sum_{j=1}^p \hat{\beta}_j X_{ij}}{1 + \exp \sum_{j=1}^p \hat{\beta}_j X_{ij}} - \frac{\exp \sum_{j=1}^p \beta_j X_{ij}}{1 + \exp \sum_{j=1}^p \beta_j X_{ij}} \right).$$

当

$$\min_{j \in I^*} |\beta_j| \geq \frac{8rk^*}{bs} + \left(\frac{4}{r} + 2s\right)\epsilon, \quad r \geq 2L\sqrt{\frac{2\ln(2p/\delta)}{n}},$$

对第一项, 用 Hoeffding 不等式可得

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n W_i X_{ik}\right| \geq \frac{r}{2}\right) \leq \frac{\delta}{p}.$$

在假设 3 成立时, 对第二项, 可以得出以下不等式

$$\mathbb{P}\left[\sum_{j=1}^p |\hat{\beta}_j - \beta_j| \left|\frac{1}{n} \sum_{i=1}^n X_{ij} X_{ik}\right| > \frac{r}{2} + \left(1 + \frac{1}{r}\right)\epsilon\right]$$

$$\leq \mathbb{P} \left[\sum_{j=1}^p |\hat{\beta}_j - \beta_j| > \frac{rk^*}{2d} + \left(1 + \frac{1}{r}\right)\epsilon \right] \leq \frac{\delta}{p}.$$

当

$$d \leq \left[\frac{r}{2} + \left(1 + \frac{1}{r}\right)\epsilon k^* \right] / \left[\frac{8rk^*}{bs} + \left(\frac{4}{r} + 2s\right)\epsilon \right],$$

第三项可以得出以下不等式, 具体过程可以参考文献 [12] 中的证明,

$$\mathbb{P} \left[|S_n - B_n| \geq \frac{r}{2} + \left(1 + \frac{1}{r}\right)\epsilon \right] \leq \frac{\delta}{p}.$$

由上面三项的界可以得出

$$\mathbb{P}(I^* \not\subset \hat{I}) \leq k^* \frac{3}{p} \delta \leq 3\delta.$$

定理 8 和定理 5 类似, 只是在一些常数的选取会有不同, 证明过程可以参照定理 5 的证明.

5.3 定理 7

使用 Spline LASSO 方法得到的 β 估计值

$$\tilde{\beta} = \arg \min_{\beta} \mathbb{P}_n l(\beta) + 2r\|\beta\|_1 + c \sum_{j=2}^{M-1} [(\beta_{j+1} - \beta_j) - (\beta_j - \beta_{j-1})]^2.$$

由定理 4 的证明可以得到

$$\begin{aligned} & r\|\tilde{\beta} - \beta\|_1 + s(\tilde{\beta} - \beta)^T \Gamma(\tilde{\beta} - \beta) + c \sum_{j=2}^{p-1} (\tilde{\beta}_{j+1} + \tilde{\beta}_{j-1} - 2\tilde{\beta}_j)^2 - c \sum_{j=2}^{p-1} (\beta_{j+1} + \beta_{j-1} - 2\beta_j)^2 \\ & \leq 2r\|\tilde{\beta} - \beta\|_1 + 2r \sum_{j=1}^p |\beta_j| - 2r \sum_{j=1}^p |\tilde{\beta}_j| + r\epsilon \leq 4r \sum_{j \in I^*} |\tilde{\beta}_j - \beta_j| + r\epsilon. \end{aligned}$$

同理, 令

$$L = \begin{bmatrix} 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & & \\ & & & 1 & -2 & 1 \end{bmatrix}_{(p-2)*p},$$

则 $Q = L^T L \geq 0$, 因此在假设 $\max_{j \in I^*} |\beta_j| \leq B$ 下可得到

$$\begin{aligned} & \sum_{j=2}^{p-1} (\tilde{\beta}_{j+1} + \tilde{\beta}_{j-1} - 2\tilde{\beta}_j)^2 - \sum_{j=2}^{p-1} (\beta_{j+1} + \beta_{j-1} - 2\beta_j)^2 = \tilde{\beta}^T Q \tilde{\beta} - \beta^T Q \beta \\ & = (\tilde{\beta} - \beta)^T Q (\tilde{\beta} - \beta) + 2(\tilde{\beta} - \beta)^T Q \beta \geq 2(\tilde{\beta} - \beta)^T Q \beta \geq -16B\|\tilde{\beta} - \beta\|_1. \end{aligned}$$

进一步可知

$$r\|\tilde{\beta} - \beta\|_1 + s(\tilde{\beta} - \beta)^\top \Gamma(\tilde{\beta} - \beta) - 16cB\|\tilde{\beta} - \beta\|_1 \leq 4r \sum_{j \in I^*} |\tilde{\beta}_j - \beta_j| + r\epsilon,$$

通过选取 c , 使得 $32cB = r$, 可得

$$\frac{r}{2}\|\tilde{\beta} - \beta\|_1 + s(\tilde{\beta} - \beta)^\top \Gamma(\tilde{\beta} - \beta) \leq 4r \sum_{j \in I^*} |\tilde{\beta}_j - \beta_j| + r\epsilon,$$

同理取 $a = 1/b$, 在事件 E 成立时可得

$$\begin{aligned} \|\tilde{\beta} - \beta\|_1 &\leq \frac{8rk^*}{bs} + \left(\frac{4}{r} + 2s\right)\epsilon, \\ \mathbb{P}\left[\|\tilde{\beta} - \beta\|_1 \leq \frac{8rk^*}{bs} + \left(\frac{4}{r} + 2s\right)\epsilon\right] &\geq 1 - \delta. \end{aligned}$$

参 考 文 献

- [1] MCCULLAGH P, NELDER J A. *Generalized Linear Models* [M]. 2nd ed. New York: Chapman and Hall/CRC, 1989.
- [2] AYALEW L, YAMAGISHI H. The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Mountains, Central Japan [J]. *Geomorphology*, 2005, **65(1-2)**: 15–31.
- [3] KING G, ZENG L C. Logistic regression in rare events data [J]. *Polit Anal*, 2001, **9(2)**: 137–163.
- [4] KEATING K A, CHERRY S. Use and interpretation of logistic regression in habitat-selection studies [J]. *J Wildlife Manage*, 2004, **68(4)**: 774–789.
- [5] TIBSHIRANI R. Regression shrinkage and selection via the lasso [J]. *J Roy Statist Soc Ser B*, 1996, **58(1)**: 267–288.
- [6] ZOU H, HASTIE T. Regularization and variable selection via the elastic net [J]. *J R Stat Soc Ser B Stat Methodol*, 2005, **67(2)**: 301–320.
- [7] TIBSHIRANI R, SAUNDERS M, ROSSET S, et al. Sparsity and smoothness via the fused lasso [J]. *J R Stat Soc Ser B Stat Methodol*, 2005, **67(1)**: 91–108.
- [8] HEBIRI M, VAN DE GEER S. The Smooth-Lasso and other $\ell_1 + \ell_2$ -penalized methods [J]. *Electron J Statist*, 2011, **5**: 1184–1226.
- [9] GUO J H, HU J C, JING B Y, et al. Spline-lasso in high-dimensional linear regression [J]. *J Amer Statist Assoc*, 2016, **111(513)**: 288–297.
- [10] ZHU J, HASTIE T. Classification of gene microarrays by penalized logistic regression [J]. *Biostatistics*, 2004, **5(3)**: 427–443.
- [11] PARK M Y, HASTIE T. L_1 -regularization path algorithm for generalized linear models [J]. *J R Stat Soc Ser B Stat Methodol*, 2007, **69(4)**: 659–677.
- [12] BUNEA F. Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization [J]. *Electron J Statist*, 2008, **2**: 1153–1194.

- [13] DONOHO D L, JOHNSTONE I M. Minimax estimation via wavelet shrinkage [J]. *Ann Statist*, 1998, **26**(3): 879–921.
- [14] DEVROYE L, LUGOSI G. *Combinatorial Methods in Density Estimation* [M]. New York: Springer-Verlag, 2001.
- [15] WEGKAMP M. Lasso type classifiers with a reject option [J]. *Electron J Statist*, 2007, **1**: 155–168.
- [16] STEINWART I. How to compare different loss functions and their risks [J]. *Constr Approx*, 2007, **26**(2): 225–287.
- [17] YANG J F, LIN C D, QIAN P Z G, et al. Construction of sliced orthogonal Latin hypercube designs [J]. *Statist Sinica*, 2013, **23**(3): 1117–1130.
- [18] HUANG H Z, YANG J F, LIU M Q. Construction of sliced (nearly) orthogonal Latin hypercube designs [J]. *J Complexity*, 2014, **30**(3): 355–365.

Variable Selection for Logistic Regression via Smooth LASSO and Spline LASSO

DAI Wei JIN Baisuo

(Department of Statistics and Finance, School of Management, University of Science and Technology
of China, Hefei, 230026, China)

Abstract: Considering a parameter estimation and variable selection problem in logistic regression, we propose Smooth LASSO and Spline LASSO. When the variables are continuous, using Smooth LASSO can select local constant coefficient in each group. However, in some case, the coefficient might be different and change smoothly. Using Spline Lasso to estimate parameter is more appropriate. In this article, we prove the reliability of the model by theory. Finally using coordinate descent algorithm to solve the model.

Keywords: logistic regression; variable selection; Smooth LASSO; Spline LASSO; coordinate descent algorithm

2010 Mathematics Subject Classification: 62F10; 62F12