

第9章 Logistic回归

李高荣

北京师范大学统计学院

E-mail: ligaorong@bnu.edu.cn



本章纲要

1 多元logistic回归

- 多元logistic回归模型
- 极大似然估计
- 预测

2 二分类模型的评估

3 惩罚似然变量选择方法

4 非参数logistic回归

5 多项logistic回归

6 分类方法比较

7 作业



- 扫二维码获取在线课件和相关教学电子资源
- 请遵守电子资源使用协议

本章纲要

1 多元logistic回归

- 多元logistic回归模型
- 极大似然估计
- 预测

2 二分类模型的评估

3 惩罚似然变量选择方法

4 非参数logistic回归

5 多项logistic回归

6 分类方法比较

7 作业

本章纲要

1 多元logistic回归

- 多元logistic回归模型
- 极大似然估计
- 预测

2 二分类模型的评估

3 惩罚似然变量选择方法

4 非参数logistic回归

5 多项logistic回归

6 分类方法比较

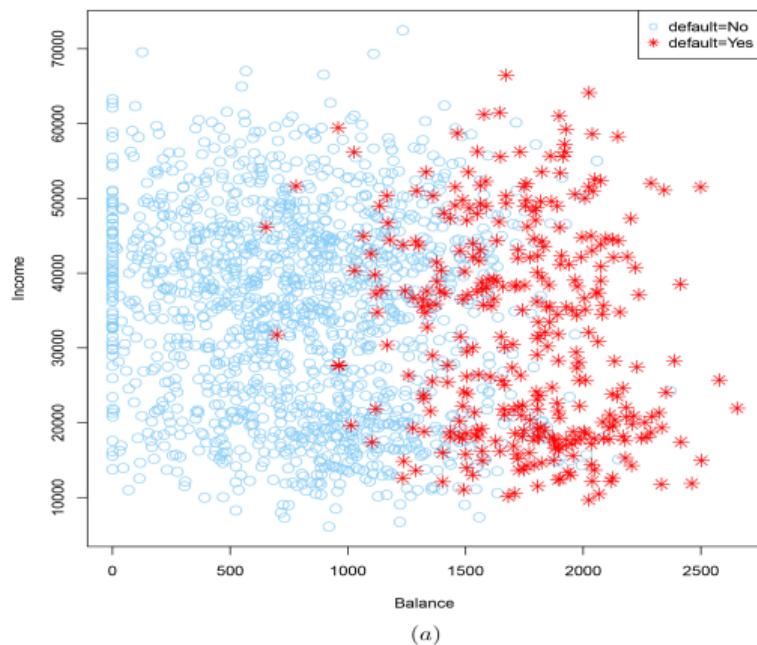
7 作业

■ 信用卡违约数据包含10000个样本和4个变量:

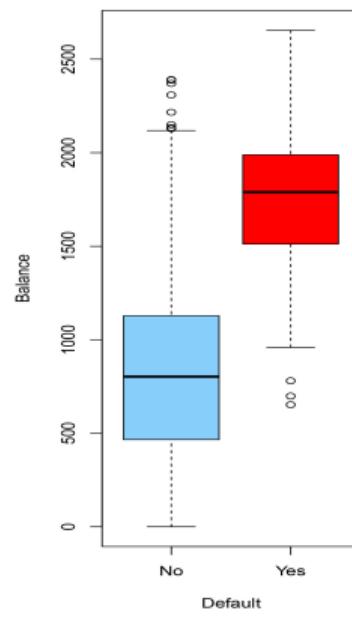
- ① default (记为 Y) — 表示客户是否违约, 为定性变量或因子变量, 如果为“Yes”表示客户违约, 如果为“No”表示客户不违约;
- ② student (记为 X_1) — 表示客户是否为学生, 如果取“Yes”表示客户为学生, 如果取“No”表示客户不是学生;
- ③ balance (记为 X_2) — 表示客户每月信用卡的平均余额;
- ④ income (记为 X_3) — 表示客户的年收入.

■ 主要目的: 通过student(X_1)、balance(X_2)和income(X_3)三个变量预测客户是否具有违约行为.

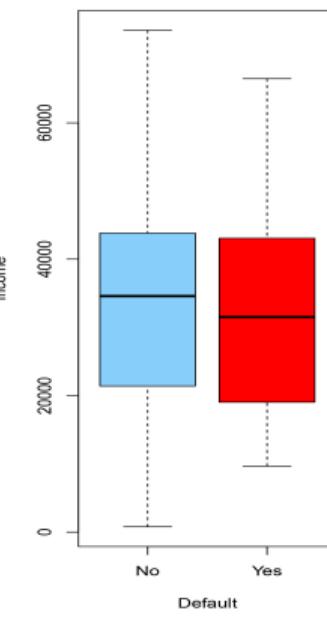
多元logistic回归模型



(a)



(b)



(c)

随机抽取1500个未违约样本和全部333个违约样本绘制客户年收入(income)与每月信用卡平均余额(balance)的散点图, 以及箱线图

多元logistic回归模型

■ 对响应变量 Y , 采用0/1编码; 对协变量 X_1 , 考虑哑变量建模, 即

$$Y = \begin{cases} 1, & \text{default=Yes (违约),} \\ 0, & \text{default=No (未违约),} \end{cases} \quad X_1 = \begin{cases} 1, & \text{student=Yes (客户为学生),} \\ 0, & \text{student=No (客户不是学生).} \end{cases}$$

■ 建立多元logistic回归模型, 核心思想是建立违约概率模型, 即给定协变量向量 $X_1 = x_1, X_2 = x_2$ 和 $X_3 = x_3$ 条件下, 考虑违约($Y = 1$)的条件概率, 即

$$\mathbb{P}(Y = 1 | X_1 = x_1, X_2 = x_2, X_3 = x_3) =: \pi(\mathbf{x}), \quad \text{其中 } \mathbf{x} = (x_1, x_2, x_3)^T.$$

- 违约的条件概率必须满足: $\pi(x) \in [0, 1]$.
- 可根据条件概率 $\pi(x)$ 的大小对客户是否违约作出预测: 如果 $\pi(x) > 0.5$ 时, 则可预测该客户违约, 否则预测其未违约.

♠ **问题:** 如何建立概率模型?

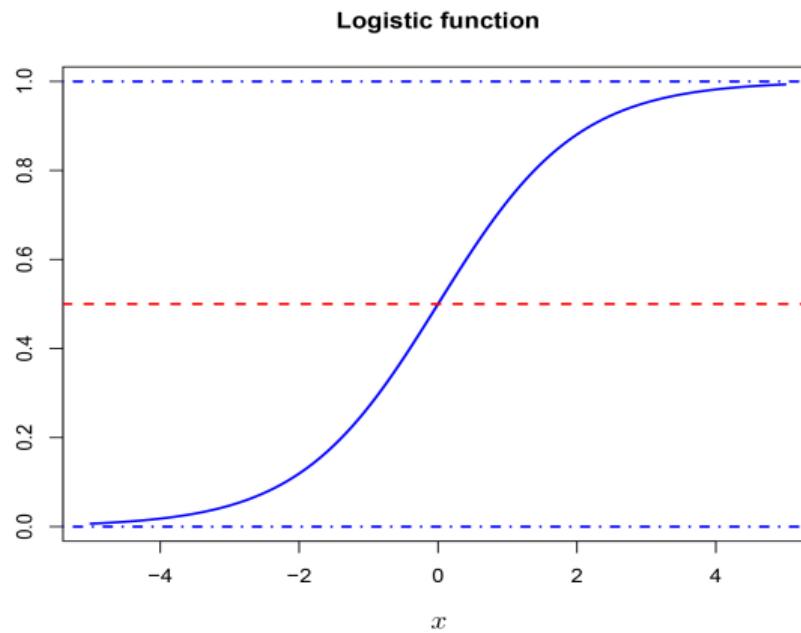
- 违约的条件概率必须满足: $\pi(x) \in [0, 1]$.
- 可根据条件概率 $\pi(x)$ 的大小对客户是否违约作出预测: 如果 $\pi(x) > 0.5$ 时, 则可预测该客户违约, 否则预测其未违约.

♠ **问题:** 如何建立概率模型?

- 最经典的为logistic函数:

$$\pi(x) = \frac{\exp(x)}{1 + \exp(x)} = \frac{1}{1 + \exp(-x)}.$$

多元logistic回归模型



Logistic函数曲线为单调递增的连续S型曲线，取值在区间[0,1]之间

多元logistic回归模型

■ 考虑更一般的情形, 即对 $p+1$ 维协变量向量 $\boldsymbol{X} = (1, X_1, \dots, X_p)^T$ 和取值为0或1的二元响应变量 Y , 建立多元logistic回归模型:

$$\begin{aligned}\mathbb{P}(Y=1|\boldsymbol{X}=\boldsymbol{x}) = \pi(\boldsymbol{x}) &= \frac{\exp(\beta_0 + \beta_1x_1 + \cdots + \beta_px_p)}{1 + \exp(\beta_0 + \beta_1x_1 + \cdots + \beta_px_p)} \\ &= \frac{1}{1 + \exp(-\beta_0 - \beta_1x_1 - \cdots - \beta_px_p)},\end{aligned}$$

其中 $\boldsymbol{x} = (1, x_1, \dots, x_p)^T$, β_0 为未知的截距项参数, β_1, \dots, β_p 为未知的回归系数.

■ 考虑下面的发生比或几率(odds)

$$\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p) \in (0, \infty).$$

■ 给定协变量向量 $\mathbf{X} = \mathbf{x}$, 则有

- ① 如果发生比(odds)接近于0, 表示事件 $\{Y = 1\}$ 发生的概率非常低;
 - ② 如果发生比(odds)接近于 ∞ , 则表示事件 $\{Y = 1\}$ 发生的概率非常高.
- 例如, 对信用卡违约数据, 发生比(odds)取值接近于0表示违约概率非常低, 接近于 ∞ 表示违约概率非常高.

■ 对发生比两边取对数, 可得**对数发生比或对数几率(log-odds)**为

$$\log \left(\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} \right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p =: g(\mathbf{x}).$$

- 当固定其他 $p - 1$ 个协变量 $x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_p$ 的取值, 而第 k 个协变量 x_k 每增加一个单位, **对数发生比(log-odds)**的变化为 β_k , 或者**发生比(odds)**要乘以 $\exp(\beta_k)$, 其中 $k = 1, \dots, p$.
- 记概率 π 的新值为 π^* , 则根据新发生比 $\pi^*/(1 - \pi^*)$ 与原发生比 $\pi/(1 - \pi)$ 的比率定义**优势比或几率比(odds ratio, OR)**.

■ 优势比或几率比(odds ratio, OR)定义为

$$\begin{aligned} \text{OR}_k &= \frac{\frac{\pi^*}{1 - \pi^*}}{\frac{\pi}{1 - \pi}} = \frac{\exp[\beta_0 + \beta_1 x_1 + \cdots + \beta_k(x_k + 1) + \cdots + \beta_p x_p]}{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \cdots + \beta_p x_p)} \\ &= \exp(\beta_k), \quad k = 1, \dots, p. \end{aligned}$$

■ 当事件 $\{Y = 1\}$ 出现的概率极小时, 优势比近似于相对危险度,
且 $\text{OR}_k = \exp(\beta_k)$ 也称为相对危险度.

- 例如, 当 $\beta_k = 0.15$, 意味着 x_k 增加一个单位可引起对数发生比(log-odds) 增加 15%, 新的发生比(odds)为原发生比(odds)的 $\exp(0.15) \approx 1.16$ 倍, 即第 k 个变量的优势比(OR _{k})为 1.16.
 - 也可以讲, 优势比(OR)增加 16%, 因为 $\exp(\beta_k) - 1 = 1.16 - 1 = 0.16$.
-
- ① 若 $\beta_k > 0$, 且 OR _{k} > 1, 说明 X_k 可能导致事件 { $Y = 1$ } 发生的概率上升;
 - ② 若 $\beta_k < 0$, 且 OR _{k} < 1, 说明 X_k 可能导致事件 { $Y = 1$ } 发生的概率下降;
 - ③ 若 $\beta_k = 0$, 且 OR _{k} = 1, 说明 X_k 对事件 { $Y = 1$ } 发生与否没有影响.

本章纲要

1 多元logistic回归

- 多元logistic回归模型
- 极大似然估计
- 预测

2 二分类模型的评估

3 惩罚似然变量选择方法

4 非参数logistic回归

5 多项logistic回归

6 分类方法比较

7 作业

♠ **问题:** 如何估计截距项 β_0 和回归系数 β_1, \dots, β_p ?

♠ **问题:** 如何估计截距项 β_0 和回归系数 β_1, \dots, β_p ?

■ **估计方法:** 极大似然法

■ 假设存在观测训练样本集 $D = \{(\mathbf{x}_i, y_i), i = 1, \dots, n\}$, 其中 $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})^T \in \mathbb{R}^{p+1}$ 为观测的协变量向量, $y_i \in \{0, 1\}$ 为二元响应变量.

■ 对第*i*个个体, y_i 服从两点分布, 即给定观测协变量 \mathbf{x}_i 时, 有

$$\mathbb{P}(y_i = 1 | \mathbf{x}_i) = \pi(\mathbf{x}_i) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip})} = \frac{\exp(\boldsymbol{\beta}^T \mathbf{x}_i)}{1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i)}.$$

极大似然估计

■ 令 $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$ 为 $p + 1$ 维的参数向量, 可得 β 的似然函数为

$$L(\beta) = \prod_{i=1}^n \pi(\mathbf{x}_i)^{y_i} (1 - \pi(\mathbf{x}_i))^{1-y_i}.$$

■ 两边取对数, 可得 β 的对数似然函数为

$$\ell(\beta) = \log(L(\beta)) = \sum_{i=1}^n \left\{ y_i(\beta^T \mathbf{x}_i) - \log[1 + \exp(\beta^T \mathbf{x}_i)] \right\}.$$

■ 极大化对数似然函数 $\ell(\beta)$, 可得 β 的极大似然估计为

$$\hat{\beta} = \arg \max_{\beta} \ell(\beta).$$

- 由于 $\ell(\beta)$ 为非线性的目标函数, 不存在极大似然估计 $\hat{\beta}$ 的解析解.
- **Newton-Raphson 迭代算法:** 假设 $\ell(\beta)$ 满足二阶连续可微, 将 $\ell(\beta)$ 在 $\beta^{(k)}$ 处进行Taylor展开, 可得

$$\begin{aligned}\ell(\beta) \approx & \ell(\beta^{(k)}) + \left(\frac{\partial \ell(\beta)}{\partial \beta} \Big|_{\beta=\beta^{(k)}} \right)^T (\beta - \beta^{(k)}) \\ & + \frac{1}{2} (\beta - \beta^{(k)})^T \left(\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} \Big|_{\beta=\beta^{(k)}} \right) (\beta - \beta^{(k)}).\end{aligned}$$

- $\partial\ell(\beta)/\partial\beta$ 为 $(p+1) \times 1$ 的 **梯度向量**, $\partial^2\ell(\beta)/\partial\beta\partial\beta^T$ 为 $(p+1) \times (p+1)$ 的 **Hessian矩阵**, 分别定义为

$$\frac{\partial\ell(\beta)}{\partial\beta} = \sum_{i=1}^n [y_i - \pi(\mathbf{x}_i)] \mathbf{x}_i,$$

$$\frac{\partial^2\ell(\beta)}{\partial\beta\partial\beta^T} = - \sum_{i=1}^n \pi(\mathbf{x}_i)(1 - \pi(\mathbf{x}_i)) \mathbf{x}_i \mathbf{x}_i^T.$$

- 由 $\partial\ell(\beta)/\partial\beta = 0$, 可得Newton-Raphson迭代公式为

$$\beta^{(k+1)} = \beta^{(k)} - \left(\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} \Big|_{\beta=\beta^{(k)}} \right)^{-1} \frac{\partial \ell(\beta)}{\partial \beta} \Big|_{\beta=\beta^{(k)}}.$$

- 对于给定的 ϵ , 当 $\|\beta^{(k+1)} - \beta^{(k)}\|_2^2 < \epsilon$ 时, 停止迭代, 可得参数向量 β 最终的极大似然估计数值解 $\hat{\beta}$.
- 对logistic回归模型, 把Newton-Raphson 迭代算法也称为Fisher得分迭代算法(Fisher scoring iteration algorithm).

- 在Newton-Raphson迭代公式中, 需要计算Hessian矩阵的逆. 对 $p > n$ 或 $p \gg n$ 的高维情形, Hessian矩阵不可逆, 这时Newton-Raphson迭代算法将失效.
- 为了解决该问题, 可以使用如下的梯度上升算法

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} - \lambda_k \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(k)}},$$

其中 λ_k 称为步长(step size)或学习率(learning rate).

- 梯度上升算法对初始估计很敏感, 只能找到局部最优解.

- 当样本量 $n \rightarrow \infty$ 时, 可证明极大似然估计 $\hat{\beta}$ 具有渐近正态分布, 均值向量为 β , 协方差矩阵的估计为

$$\widehat{\text{Cov}}(\hat{\beta}) \approx \left[\sum_{i=1}^n \hat{\pi}(\mathbf{x}_i) (1 - \hat{\pi}(\mathbf{x}_i)) \mathbf{x}_i \mathbf{x}_i^T \right]^{-1},$$

其中 $\hat{\pi}(\mathbf{x}_i) = \frac{\exp(\hat{\beta}^T \mathbf{x}_i)}{1 + \exp(\hat{\beta}^T \mathbf{x}_i)}$.

- 令 $\text{SE}(\hat{\beta}_k)$ 表示 $\hat{\beta}_k$ 的标准误差, 即为矩阵 $\widehat{\text{Cov}}(\hat{\beta})$ 的第 k 个对角线元素的平方根, 其中 $k = 0, 1, \dots, p$.
- 这时, 可构造 β_k 置信水平为 $1 - \alpha$ 的置信区间, 即

$$\hat{\beta}_k \pm z_{1-\alpha/2} \times \text{SE}(\hat{\beta}_k), \quad k = 0, 1, \dots, p.$$

- 在实际问题中, 需要检验第 k 个协变量 X_k 对事件 $\{Y = 1\}$ 发生的概率是否有显著影响, 即

$$H_{k0} : \beta_k = 0 \longleftrightarrow H_{k1} : \beta_k \neq 0, \quad k = 1, \dots, p.$$

- 可以证明, 当样本量 $n \rightarrow \infty$, 且在原假设 H_{k0} 成立时, Z 统计量渐近服从标准正态分布, 即

$$Z_k = \frac{\hat{\beta}_k}{\text{SE}(\hat{\beta}_k)} \xrightarrow{d} N(0, 1), \quad k = 1, \dots, p.$$

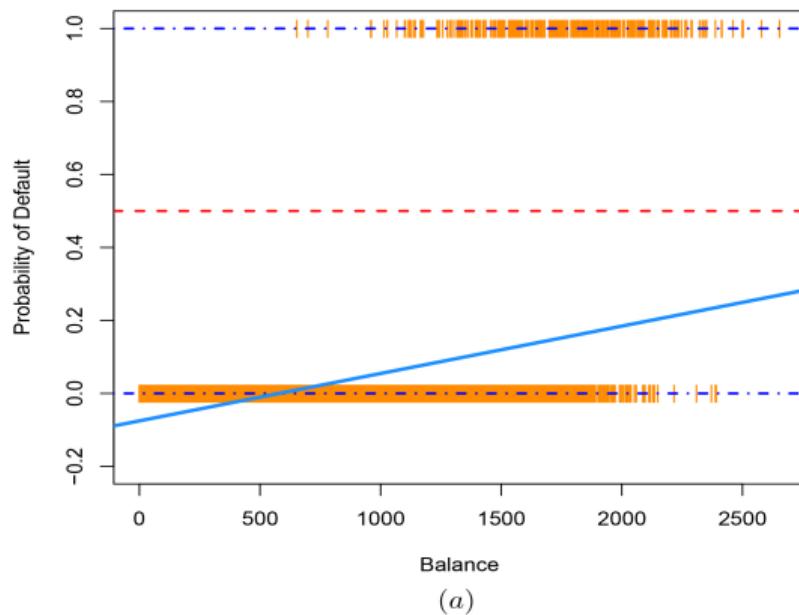
- 如果 $|Z_k| \geq z_{1-\alpha/2}$ 或者计算 Z 统计量的 p 值进行判断, 即 $p_k < \alpha/2$ 时, 则拒绝原假设 H_{k0} , 认为 $\beta_k \neq 0$.
- 在 R 语言中, 函数 `glm()` 中参数 `family=binomial` 时可进行 logistic 回归模型拟合.

- 首先, 建立仅用变量balance预测违约概率的logistic回归模型, 并和线性回归模型的拟合进行比较.
- 可用函数confint()计算参数的95%置信区间, 并用程序包GGally 中的函数ggcoef()进行可视化.

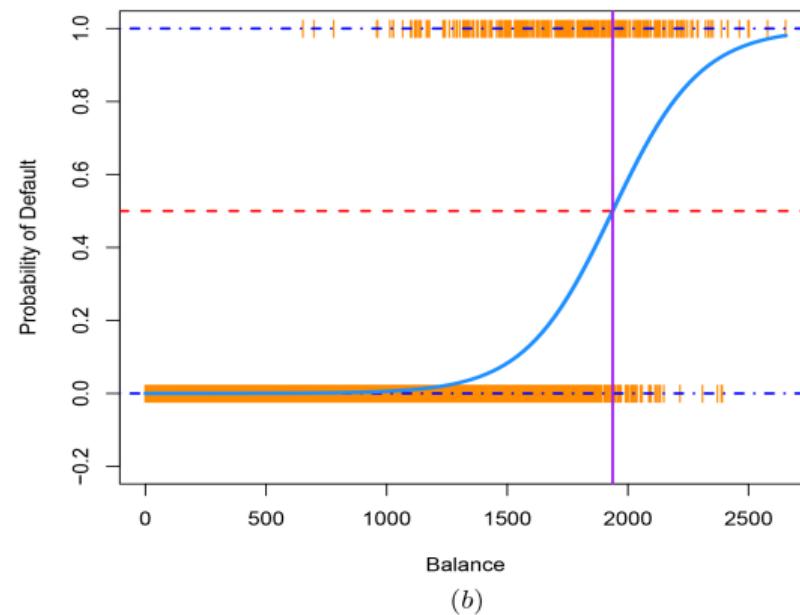
```
library(GGally); library(ISLR2); attach(Default)
balance.glm=glm(default~balance, data = Default, family = "binomial")
> confint(balance.glm)      ## 计算95\%置信区间
                               2.5 %           97.5 %
(Intercept)      -11.383288936      -9.966565064
balance          0.005078926       0.005943365
ggcoef(balance.glm, exclude_intercept = T, vline_color = "red",
        errorbar_color = "blue", errorbar_height = 0.1) + theme_bw()
```

信用卡违约数据分析

Linear Regression for Classification



Logistic Regression for Classification



(a) 线性回归模型的违约概率拟合; (b) logistic回归模型的违约概率拟合, 其中紫色竖线表示决策边界

信用卡违约数据分析

信用卡违约数据中，变量balance预测违约概率的logistic回归模型的系数估计结果

变量	系数估计	标准误差	Z统计量	p值	95%置信区间
截距项	-10.6513	0.3612	-29.49	< 0.0001	[-11.3833, -9.9666]
balance	0.0055	0.0002	24.95	< 0.0001	[0.0051, 0.0059]

- 变量balance每增加一个单位，违约的对数发生比(log-odds)增加0.0055个单位，新的发生比(odds)为原发生比(odds)的 $\exp(0.0055) \approx 1.0055$ 倍，即优势比(OR)为1.0055.
- 所得logistic回归模型为

$$\mathbb{P}(\text{default} = \text{Yes} | \text{balance}) = \pi(\text{balance}) = \frac{\exp(-10.6513 + 0.0055 \times \text{balance})}{1 + \exp(-10.6513 + 0.0055 \times \text{balance})}.$$

■ 其次, 用三个协变量student(X_1), balance(X_2)和income(X_3)对违约概率建立多元logistic 回归模型, 并对模型中参数进行估计和统计推断.

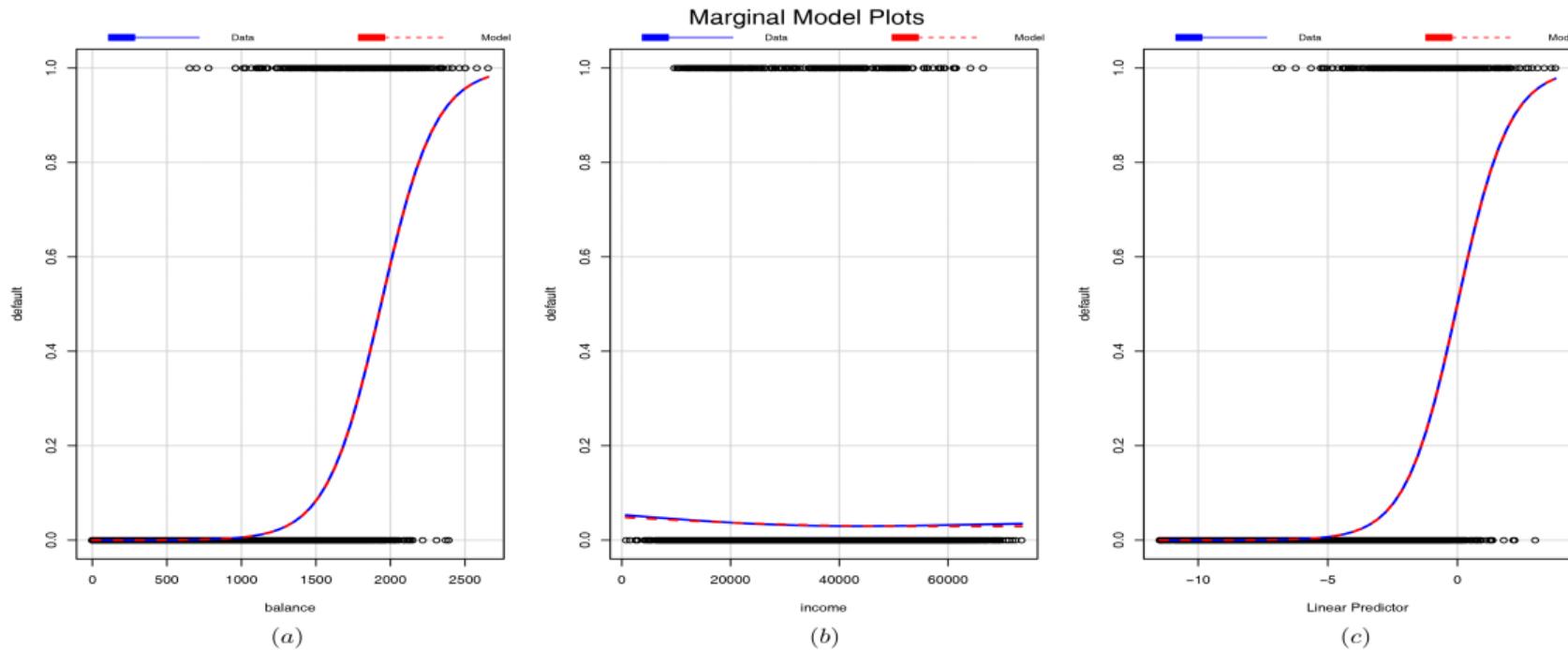
变量	系数估计	标准误差	Z统计量	p值	95%置信区间
截距项	-10.8690	0.4923	-22.080	< 0.0001	[-11.8590, -9.9281]
student[Yes]	-0.6468	0.2363	-2.738	0.0062	[-1.1090, -0.1822]
balance	0.0057	0.0002	24.738	< 0.0001	[0.0053, 0.0062]
income	0.0030	0.0082	0.370	0.7115	[-0.0130, 0.0191]

■ 所得多元logistic回归模型为

$$\mathbb{P}(Y = 1|X_1, X_2, X_3) = \frac{\exp(-10.8690 - 0.6468X_1 + 0.0057X_2 + 0.0030X_3)}{1 + \exp(-10.8690 - 0.6468X_1 + 0.0057X_2 + 0.0030X_3)}.$$

■ 利用程序包alr4中的函数mmps()绘制每个协变量对违约概率拟合的边际logistic 回归模型.

信用卡违约数据分析



信用卡违约数据的边际logistic回归模型拟合图

■ 最后, 可利用AIC准则和BIC准则进行模型选择, 其中AIC和BIC统计量分别定义为

$$\text{AIC} = -2\ell(\hat{\beta}) + 2p = -2 \log(L(\hat{\beta})) + 2p,$$

$$\text{BIC} = -2\ell(\hat{\beta}) + p \log(n) = -2 \log(L(\hat{\beta})) + p \log(n),$$

其中 $\ell(\hat{\beta})$ 表示用 p 个协变量拟合的对数似然函数, $L(\hat{\beta})$ 表示似然函数.

信用卡违约数据分析

```
> glm.aic = step(fit.glm)
Start: AIC=1579.54
default ~ student + balance + income
              Df    Deviance      AIC
- income     1      1571.7    1577.7
<none>          1571.5    1579.5
- student    1      1579.0    1585.0
- balance    1      2907.5    2913.5
Step: AIC=1577.68
default ~ student + balance
              Df    Deviance      AIC
<none>          1571.7    1577.7
- student    1      1596.5    1600.5
- balance    1      2908.7    2912.7
```

信用卡违约数据的逐步回归分析结果

变量	系数估计	标准误差	Z统计量	p值	95%置信区间
截距项	-10.7500	0.3692	-29.116	< 0.0001	[-11.4981, -10.0498]
student[Yes]	-0.7149	0.1475	-4.846	< 0.0001	[-1.0078, -0.4291]
balance	0.0057	0.0002	24.750	< 0.0001	[0.0053, 0.0062]

■ 所得最优的多元logistic回归模型为

$$\mathbb{P}(Y = 1 | X_1, X_2) = \frac{\exp(-10.7500 - 0.7149X_1 + 0.0057X_2)}{1 + \exp(-10.7500 - 0.7149X_1 + 0.0057X_2)}.$$

本章纲要

1 多元logistic回归

- 多元logistic回归模型
- 极大似然估计
- 预测

2 二分类模型的评估

3 惩罚似然变量选择方法

4 非参数logistic回归

5 多项logistic回归

6 分类方法比较

7 作业

- 得到多元logistic回归模型的系数估计 $\hat{\beta}$ 后, 则可预测 $y_0 = 1$ 的条件概率, 即

$$\mathbb{P}(y_0 = 1 | \mathbf{x}_0) = \hat{\pi}(\mathbf{x}_0) = \frac{\exp(\hat{\beta}^T \mathbf{x}_0)}{1 + \exp(\hat{\beta}^T \mathbf{x}_0)},$$

其中 $\mathbf{x}_0 = (1, x_{01}, \dots, x_{0p})^T$ 为给定的测试样本.

- 如果预测概率 $\mathbb{P}(y_0 = 1 | \mathbf{x}_0) > 0.5$ 时, 则可预测 $y_0 = 1$, 否则预测为 $y_0 = 0$.
- 决策边界为: $\{\mathbf{x}_i \mid \hat{\mathbb{P}}(y_i = 1 | \mathbf{x}_i) = 0.5\}$.

■ 对logistic回归模型，也可使用对数发生比(log-odds)来预测响应变量的类别：

- ① 如果对数发生比 $\log\left(\frac{\hat{\pi}(\mathbf{x}_0)}{1 - \hat{\pi}(\mathbf{x}_0)}\right) > 0$, 则可预测 $y_0 = 1$;
- ② 如果对数发生比 $\log\left(\frac{\hat{\pi}(\mathbf{x}_0)}{1 - \hat{\pi}(\mathbf{x}_0)}\right) < 0$, 则可预测 $y_0 = 0$;
- ③ 如果对数发生比 $\log\left(\frac{\hat{\pi}(\mathbf{x}_0)}{1 - \hat{\pi}(\mathbf{x}_0)}\right) = 0$, 则可预测 y_0 落在决策边界上,
可预测为 $y_0 = 0$ 或 1 .

预测—信用卡违约数据分析

- 一个信用卡余额为1500美元, 同时收入为40000美元**学生的违约概率**为

$$\hat{\pi}(x_0) = \frac{\exp(-10.8690 - 0.6468 \times 1 + 0.0057 \times 1500 + 0.0030 \times 40)}{1 + \exp(-10.8690 - 0.6468 \times 1 + 0.0057 \times 1500 + 0.0030 \times 40)} \approx 0.058.$$

- 一个信用卡余额为1500美元, 同时收入为40000美元**非学生的违约概率**为

$$\hat{\pi}(x_0) = \frac{\exp(-10.8690 + 0.0057 \times 1500 + 0.0030 \times 40)}{1 + \exp(-10.8690 + 0.0057 \times 1500 + 0.0030 \times 40)} \approx 0.105.$$

本章纲要

1 多元logistic回归

- 多元logistic回归模型
- 极大似然估计
- 预测

2 二分类模型的评估

3 惩罚似然变量选择方法

4 非参数logistic回归

5 多项logistic回归

6 分类方法比较

7 作业

- 分类模型的评价准则: 精确率(accRate) 和 错误率 (errRate).
- 但是准确率和错误率并不适合“类别不平衡”的数据. 例如, 某种罕见病的发病率仅为1%, 则样本中两个类别的数据高度不平衡. 这时, 采用任何的分类方法, 只要一直预测不发病, 准确率也能高达99%(或错误率为1%).
- 因此, 对于类别不平衡的训练样本, 会导致训练分类模型存在偏差.
- 在实际应用中, 更希望准确预测那些发病的患者, 即所谓的正例(positive cases).

■ 混淆矩阵 (confusion matrix):

分类结果的混淆矩阵(confusion matrix)

		真实分类		
		正例($y = 1$)	反例($y = 0$)	总计
预测分 类	正例($\hat{y} = 1$)	真阳性值(TP)	假阳性值(FP)	$P^* = TP + FP$
	反例($\hat{y} = 0$)	假阴性值(FN)	真阴性值(TN)	$N^* = FN + TN$
总计		$P = TP + FN$	$N = FP + TN$	n

P^* 表示预测为正例的总个数, N^* 表示预测为反例的总个数.

- ① 真阳性值(true positive, TP): 表示真实类别为正例而被预测为正例的个数, 即 $\#\{\hat{y}_i = 1, y_i = 1\}$;
- ② 假阳性值(false positive, FP): 表示真实类别为反例而被预测为正例的个数, 即 $\#\{\hat{y}_i = 1, y_i = 0\}$;
- ③ 假阴性值(false negative, FN): 表示真实类别为正例而被预测为反例的个数, 即 $\#\{\hat{y}_i = 0, y_i = 1\}$;
- ④ 真阴性值(true negative, TN): 表示真实类别为反例而被预测为反例的个数, 即 $\#\{\hat{y}_i = 0, y_i = 0\}$.

由混淆矩阵，可计算二分类模型的评价指标：

- 准确率: $\text{accRate} = (\text{TP} + \text{TN})/n$;
- 错误率: $\text{errRate} = 1 - \text{accRate}$;
- 灵敏度: 在真实为正例的子样本中, 被正确预测为正例的比例, 也称为**真阳性率**(true positive rate, TPR), 定义为

$$\text{sensitivity} = \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{TP}}{\text{P}}.$$

■ **特异度** 定义为在真实为反例的子样本中，被正确预测为反例的比例，也称为**真阴性率**(true negative rate, TNR)，定义为

$$\text{specificity} = \text{TNR} = \frac{\text{TN}}{\text{FP} + \text{TN}} = \frac{\text{TN}}{\text{N}}.$$

■ **1-特异度**(1-specificity) 定义为在真实为反例的子样本中，被错误预测为正例的比例，也称为**假阳性率**(false positive rate, FPR)，刻画了犯第一类错误的大小，定义为

$$1 - \text{specificity} = \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} = \frac{\text{FP}}{\text{N}}.$$

- 在logistic 回归二分类问题中, 默认分类的阈值或决策边界为 $\hat{P}(y = 1|x) = \hat{\pi}(x) = 0.5$, 该阈值或决策边界对于不平衡数据并不是最佳选择.
- 在作分类预测时, 可能犯“**假阳性**”或“**假阴性**”两种不同的错误, 在具体的问题中, 犯这两类错误的成本可能差别很大.
- 例如, 在医学诊断中,
 - ① “假阳性”表示将健康者误判为患者, 其成本可能只是医疗检查等损失;
 - ② “假阴性”将患者误判为健康者, 则会耽误病情并产生严重的后果.

二分类模型的评估

- 为了提高二分类模型预测的准确性，需要选择合适的阈值或决策边界 $\hat{\pi}(\mathbf{x}) = c$ 进行分类，即

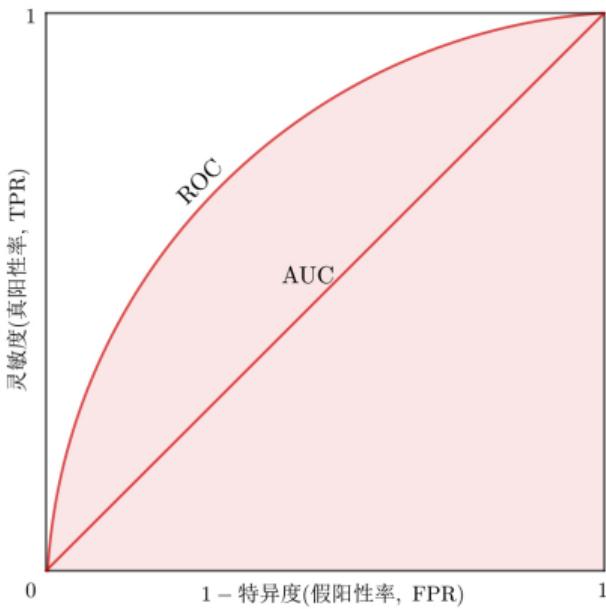
$$\hat{y}_i = \begin{cases} 1, & \text{如果 } \hat{\pi}(\mathbf{x}_i) > c, \\ 0, & \text{如果 } \hat{\pi}(\mathbf{x}_i) \leq c, \end{cases} \quad i = 1, \dots, n.$$

- 如果使用更低的阈值 c ，将预测更多的正例，而预测更少的反例：

- 在真实为正例的子样本中，预测准确率将上升，即灵敏度(真阳性率，TPR)将上升；
- 在真实为反例的子样本中，预测准确率将下降，即特异度(真阴性率，TNR)下降，故1-特异度(假阳性率，FPR)上升。

二分类模型的评估

- 灵敏度(真阳性率, TPR)与1-特异度(假阳性率, FPR)均为阈值 $\hat{\pi}(x) = c$ 的函数.
- 受试者工作特征曲线(ROC曲线): 1-特异度(假阳性率, FPR)作为横坐标, 灵敏度(真阳性率, TPR)作为纵坐标, 让阈值 $\hat{\pi}(x) = c$ 从0连续地变为1, 所得到的一条曲线.
- ROC曲线下面积(AUC): 度量ROC曲线的优良性.



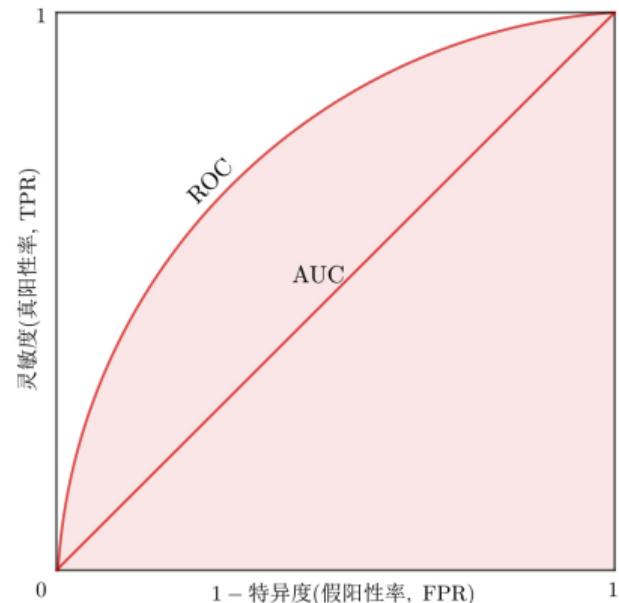
二分类模型的评估

- 当阈值 $c = 0$ 时, 所有样本都被预测为正例, 即 $\#\{\hat{y}_i = 1\} = n$, 则
- $FN = 0$ 和 $TN = 0$, 意味着所有真实正例都被正确预测, 而所有反例都被错误预测.
- 这时, 灵敏度(或真阳性率)为

$$\text{sensitivity} = \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} = 1;$$

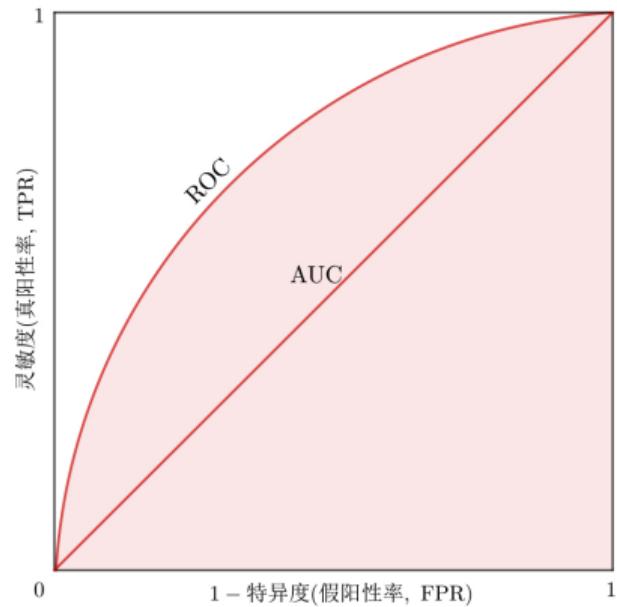
- 1-特异度(或假阳性率)为

$$1 - \text{specificity} = \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} = 1. \quad \text{当 } c = 0 \text{ 时, 坐标为 } (1, 1), \text{ 位于图的最右上角.}$$



二分类模型的评估

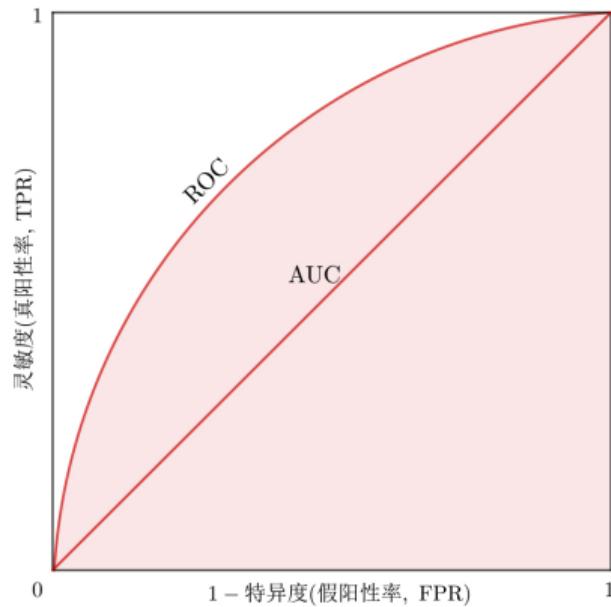
- 当阈值 c 从0变大时, 灵敏度(或真阳性率)将变小;
- 当 $c = 1$ 时, 所有样本都被预测为反例, 即 $\#\{\hat{y}_i = 0\} = n$, 则 $TP = 0$ 和 $FP = 0$, 意味着所有真实反例都被正确预测, 而所有正例都被错误预测.
- 这时, 灵敏度(或真阳性率)为0, 且1-特异度(或假阳性率)也为0.



当 $c = 1$ 时, 坐标为 $(0, 0)$, 位于图的原点.

二分类模型的评估

- AUC值一般介于0.5与1之间:
- 如果AUC=1, 表示模型对正例和反例都正确预测, 这是无法达到的理想情况;
- 如果AUC=0.5, 表示模型的预测结果类似于随机猜测;
- 如果AUC<0.5, 表示模型的预测结果还不如随机猜测.
- 在R语言中, 可用程序包ROCR和pROC绘制ROC曲线.



二分类模型的评估

- 对类别不平衡的数据, 也可利用Cohen(1960)提出的kappa指标进行度量;
- 程序包vcd中的函数Kappa()可计算kappa指标, 其中kappa指标也可用于多分类问题的评价.

Kappa指标含义的解释

kappa指标的取值	kappa指标的解释
$\text{kappa} \leq 0.2$	一致性很差 (poor agreement)
$0.2 < \text{kappa} \leq 0.4$	一致性较差 (fair agreement)
$0.4 < \text{kappa} \leq 0.6$	一致性中等 (moderate agreement)
$0.6 < \text{kappa} \leq 0.8$	一致性较好 (good agreement)
$0.8 < \text{kappa} \leq 1$	一致性很好 (great agreement)

- 固定种子`set.seed(2023)`, 把数据随机分成70%的训练样本和30%的测试样本.
- 首先, 在训练集上拟合多元logistic回归模型.

```
library(ISLR2); library(ROCR); library(knitr)
attach(Default); set.seed(2023)
tr.id = sample(nrow(Default), 0.7*nrow(Default))
train = Default[tr.id, ]; test = Default[-tr.id, ]
train.m = glm(default ~ student + balance + income,
              family = "binomial", data = train)
```

信用卡违约数据分析

■ 然后, 用函数predict()进行预测, 计算测试集上的混淆矩阵.

- ① 函数predict()中的参数type="response", 表示预测事件发生的条件概率;
- ② 默认取type="link", 表示预测 $\hat{\beta}^T x_i$, 即对数发生比(log-odds).

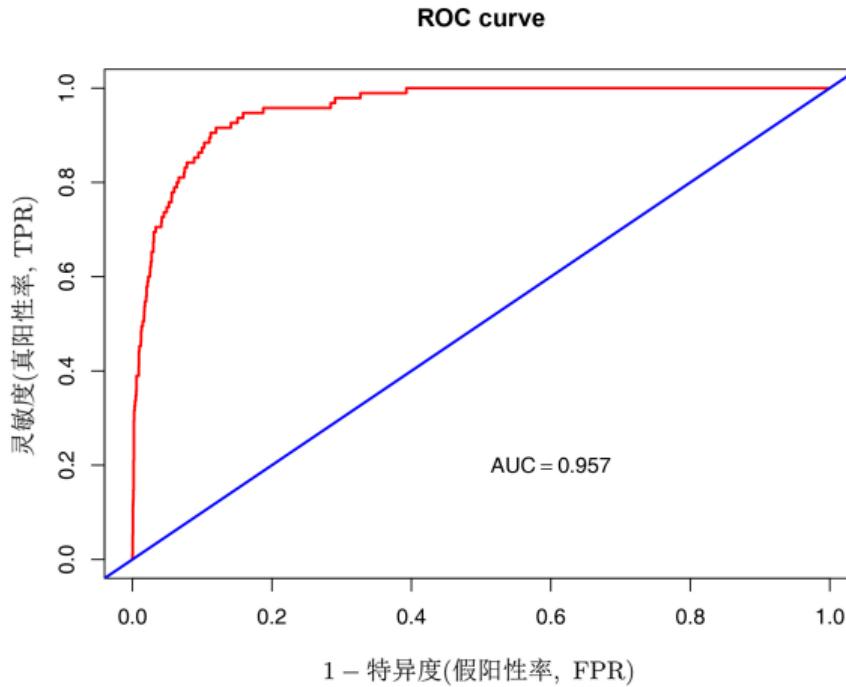
```
prob.test = predict(train.m, type = "response", newdata = test)
pred.test = prob.test > 0.5
pred.test[which(pred.test == FALSE)] = "No"
pred.test[which(pred.test == TRUE)] = "Yes"
con.mat = table(Predicted = pred.test, Actual = test$default)
> con.mat    ## 混淆矩阵
```

		Actual
Predicted	No	Yes
No	2891	62
Yes	14	33

■ 根据混淆矩阵计算准确率、错误率、灵敏度、特异度和召回率.

```
(accRate = (con.mat[1,1]+con.mat[2,2])/sum(con.mat))
(errRate = (con.mat[2,1]+con.mat[1,2])/sum(con.mat))
(sensitivity = con.mat[2,2]/(con.mat[1,2]+con.mat[2,2]))
(specificity = con.mat[1,1]/(con.mat[1,1]+con.mat[2,1]))
(recRate = con.mat[2,2]/(con.mat[2,1]+con.mat[2,2]))
res=data.frame(accRate, errRate, sensitivity, specificity, recRate)
colnames(res)=c("准确率", "错误率", "灵敏度", "特异度", "召回率")
> kable(res, digits = 4)
| 准确率 | 错误率 | 灵敏度 | 特异度 | 召回率 |
| -----: | -----: | -----: | -----: | -----: |
| 0.9747 | 0.0253 | 0.3474 | 0.9952 | 0.7021 |
```

■ 使用程序包ROCR在30%测试集上绘制ROC曲线, AUC=0.957.



- 用程序包vcd中的函数Kappa()计算kappa指标.
- 在测试集中, kappa指标的值为0.4533, 表明预测值与真实值之间具有一致性中等(moderate agreement)表现.

```
library(vcd)
```

```
> Kappa(con.mat)
```

	value	ASE	z	Pr (> z)
Unweighted	0.4533	0.0523	8.669	4.374e-18
Weighted	0.4533	0.0523	8.669	4.374e-18

本章纲要

1 多元logistic回归

- 多元logistic回归模型
- 极大似然估计
- 预测

2 二分类模型的评估

3 惩罚似然变量选择方法

4 非参数logistic回归

5 多项logistic回归

6 分类方法比较

7 作业

- 对于多元logistic回归模型，也需要讨论模型中哪些协变量对事件 $\{Y = 1\}$ 发生的概率有显著性影响，而哪些协变量对事件 $\{Y = 1\}$ 发生的概率无显著性影响。
- 解决办法：**采用惩罚的变量选择方法，即

$$\begin{aligned} Q(\boldsymbol{\beta}) &= \ell(\boldsymbol{\beta}) - n \sum_{j=1}^p p_\lambda(|\beta_j|) \\ &= \sum_{i=1}^n \left\{ y_i (\boldsymbol{\beta}^\top \mathbf{x}_i) - \log \left[1 + \exp(\boldsymbol{\beta}^\top \mathbf{x}_i) \right] \right\} - n \sum_{j=1}^p p_\lambda(|\beta_j|). \end{aligned}$$

惩罚似然变量选择方法

■ $p_\lambda(\cdot)$ 是惩罚函数，可以取

- ① 岭回归
- ② Lasso
- ③ SCAD
- ④ 自适应Lasso
- ⑤ 弹性网

■ $\lambda \geq 0$ 是调节参数或截断参数，是用来控制模型的复杂度，采用CV、GCV 或BIC等数据驱动的方法进行选取。

■ 极大化惩罚似然目标函数 $Q(\beta)$, 可得 β 的 惩罚极大似然估计, 即

$$\hat{\beta} = \arg \max_{\beta} Q(\beta).$$

■ 在R语言中, 可利用下面程序包对多元logistic回归模型进行惩罚似然变量选择:

- ① [glmnet](#)
- ② [gcdnet](#)
- ③ [ncvreg](#)

Heart数据集分析

- 对程序包`ncvreg`中的Heart数据集进行Lasso和SCAD分析.
- Heart数据集包含462个观测样本, 二元响应变量y (1表示患有心脏病, 0 表示未患有心脏病), 和9个协变量为

- ① `sbp` (收缩压)
- ② `tobacco` (累计烟草消费量, 单位: kg)
- ③ `ldl` (低密度脂蛋白胆固醇)
- ④ `adiposity` (脂肪组织浓度)
- ⑤ `famhist` (1表示家族存在心脏病, 0表示家族不存在心脏病)
- ⑥ `typea` (A型行为的测试度量得分)
- ⑦ `obesity` (肥胖)
- ⑧ `alcohol` (当前饮酒量)
- ⑨ `age` (个体的年龄)

- 目的: 考察9个协变量中哪些协变量对患有心脏病概率有显著影响.
- 首先, 利用Lasso方法对Heart数据集进行分析.

```
library(glmnet); library(ncvreg)
data(Heart);      set.seed(2023)
lasso.cv = cv.glmnet(Heart$X, Heart$y, family="binomial", alpha=1)
plot(lasso.cv)          ## 绘制交叉验证误差图
lasso.fit = glmnet(Heart$X, Heart$y, family = "binomial")
plot(lasso.fit, xvar="lambda", label=TRUE, lwd=2) ## Lasso估计路径图
> lasso.cv$lambda.min           > lasso.cv$lambda.1se
[1] 0.009921424                [1] 0.03649477
```

Heart数据集分析

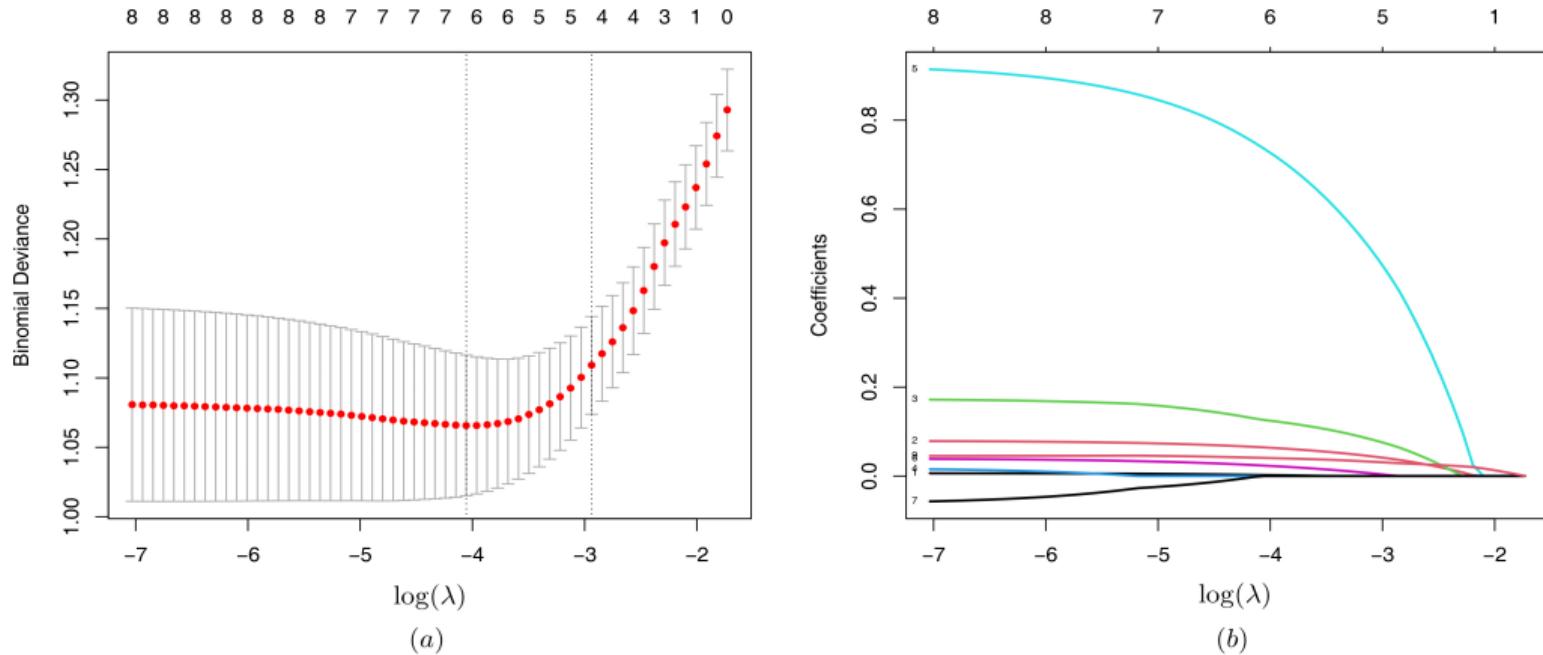


Figure 1: (a) Lasso 的交叉验证误差图; (b) Lasso估计随着 λ 变化的路径图. CV误差最小的 λ 为 $\hat{\lambda} \approx 0.0099$, 利用“一个标准差”准则选取的 λ 为 $\tilde{\lambda} \approx 0.0365$

■ 使用函数coef()分别提取 $\hat{\lambda}$ 和 $\tilde{\lambda}$ 对应回归系数的Lasso估计.

```
> coef(lasso.cv, s="lambda.min")    > coef(lasso.cv, s="lambda.1se")
10 x 1 sparse Matrix of class "dgCMatrix"
                                         s1
(Intercept) -5.73717333
sbp          0.00416794
tobacco      0.07056290
ldl          0.14789200
adiposity     .
famhist      0.81077289
typea         0.02967640
obesity      -0.01618876
alcohol       .
age           0.04396417
                                         s1
(Intercept) -3.70344526
sbp          .
tobacco      0.05029280
ldl          0.09549068
adiposity     .
famhist      0.57306288
typea         0.01108731
obesity      .
alcohol       .
age           0.03496002
```

■ 其次, 利用SCAD方法对Heart数据集进行分析.

```
scad.cv=cv.ncvreg(Heart$X,Heart$y,family="binomial",penalty="SCAD")
> summary(scad.cv)
SCAD-penalized logistic regression with n=462, p=9
At minimum cross-validation error (lambda=0.0144):
-----
```

Nonzero coefficients: 7

Cross-validation error (deviance): 1.07

R-squared: 0.20

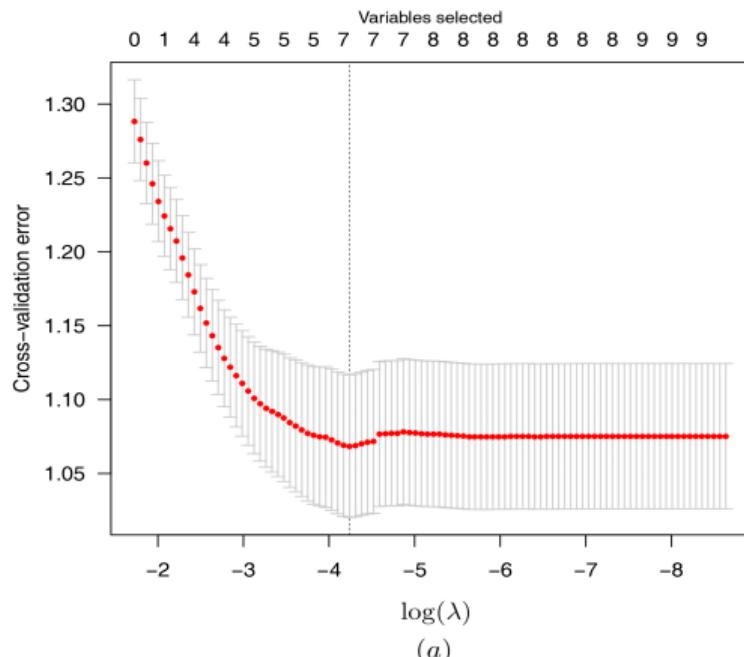
Signal-to-noise ratio: 0.25

Prediction error: 0.268

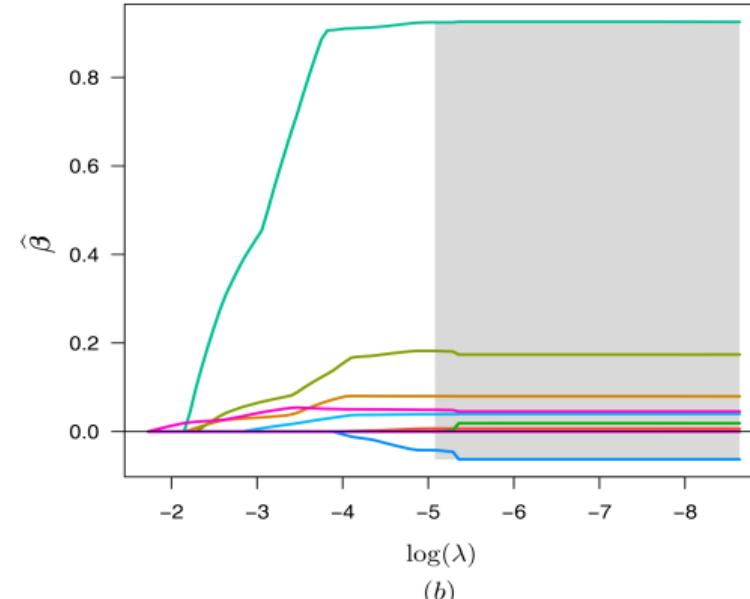
绘制交叉验证图和SCAD估计路径图

```
scad.fit = scad.cv$fit; plot(scad.cv); plot(scad.fit, log=TRUE)
```

Heart数据集分析



(a)



(b)

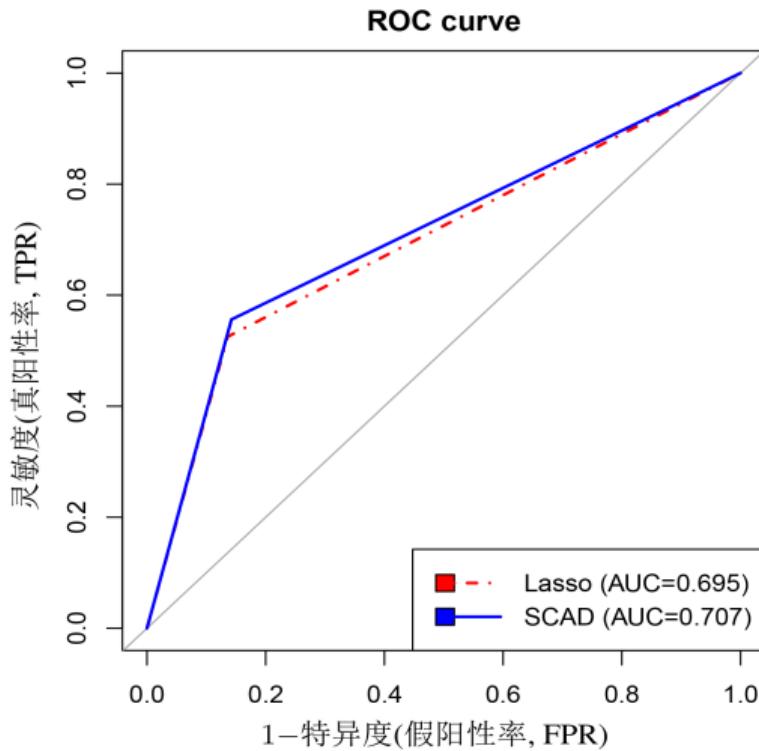
(a) SCAD 的交叉验证误差图; (b) SCAD估计随着 λ 变化的路径图

Heart数据集分析

```
> scad.cv$lambda.min  
[1] 0.01439429  
> scad.fit$beta[, scad.cv$min] ## lambda=0.0144对应的系数估计  
(Intercept) sbp tobacco ldl adiposity  
-6.346175650 0.001815207 0.080107355 0.169622961 0.000000000  
famhist typea obesity alcohol age  
0.912011700 0.037695525 -0.015560248 0.000000000 0.050179621  
> coef(scad.fit, lambda = 0.05) ## lambda=0.05对应的系数估计  
(Intercept) sbp tobacco ldl adiposity  
-3.397134970 0.000000000 0.031306612 0.064116832 0.000000000  
famhist typea obesity alcohol age  
0.436932032 0.005588174 0.000000000 0.000000000 0.041152826
```

Heart数据集分析

- ① 最后，使用函数predict()对Lasso方法和SCAD方法所得结果进行预测，并对两种方法的预测结果进行比较。
- ② 调节参数都取使用10折CV方法选取的最优调节参数。
- ③ 利用程序包pROC 绘制ROC曲线和计算AUC值



本章纲要

1 多元logistic回归

- 多元logistic回归模型
- 极大似然估计
- 预测

2 二分类模型的评估

3 惩罚似然变量选择方法

4 非参数logistic回归

5 多项logistic回归

6 分类方法比较

7 作业

- 当响应变量 Y 的取值为0或1的二元变量时, 考虑单个协变量 X 的非参数logistic回归模型:

$$\mathbb{P}(Y = 1|X = x) = \frac{\exp(g(x))}{1 + \exp(g(x))},$$

其中 $g(x)$ 为未知的连续光滑函数.

- 进一步, 可得到对数发生比(log-odds)为

$$\log \left(\frac{\mathbb{P}(Y = 1|X = x)}{\mathbb{P}(Y = 0|X = x)} \right) = g(x).$$

■ 对于非参数函数 $g(x)$ 的拟合, 同样可采用如下方法进行估计:

- ① 多项式回归
- ② 回归样条
- ③ 自然样条
- ④ 光滑样条
- ⑤ N-W核光滑方法
- ⑥ 局部线性光滑方法

非参数logistic回归

- 本节以光滑样条为例对非参数函数 $g(x)$ 进行拟合, 进而得到条件概率 $\mathbb{P}(Y = 1|X = x)$ 的光滑估计, 并用来分类或者风险评分.
- 假设 $\{(x_i, y_i), i = 1, \dots, n\}$ 是来自非参数logistic回归模型的一组i.i.d.的随机样本, 其中 $y_i \in \{0, 1\}$ 为二元响应变量.
- 构造如下的惩罚对数似然目标函数

$$\ell(g; \lambda) = \sum_{i=1}^n \left\{ y_i g(x_i) - \log \left[1 + \exp(g(x_i)) \right] \right\} - \frac{1}{2} \lambda \int \{g''(x)\}^2 dx,$$

其中 $\lambda \geq 0$ 称为光滑参数, $g''(x)$ 是 $g(x)$ 的二阶导数.

■ 令非参数函数 $g(x)$ 可被自然三次样条基函数逼近, 即

$$g(x) \approx \sum_{j=1}^n \beta_j b_j(x),$$

其中 $b_j(x)$ 是自然样条的基函数, 且 $j = 1, \dots, n$.

■ 首先, 介绍一些符号

- ▶ $\boldsymbol{\beta} = (\beta_1, \dots, \beta_n)^T$ 为 $n \times 1$ 的未知参数向量;
- ▶ $\mathbf{B}(x_i) = (b_1(x_i), \dots, b_n(x_i))^T$ 为 $n \times 1$ 的基函数向量;
- ▶ Ω_n 为 $n \times n$ 的矩阵, 其第 (j, k) 个元素为 $\{\Omega_n\}_{jk} = \int b_j''(x) b_k''(x) dx$.

■ 这时, 惩罚对数似然目标函数 $\ell(g; \lambda)$ 可以写为

$$\ell(\boldsymbol{\beta}; \lambda) = \sum_{i=1}^n \left\{ y_i \boldsymbol{\beta}^T \mathbf{B}(x_i) - \log \left[1 + \exp (\boldsymbol{\beta}^T \mathbf{B}(x_i)) \right] \right\} - \frac{1}{2} \lambda \boldsymbol{\beta}^T \boldsymbol{\Omega}_n \boldsymbol{\beta}.$$

■ **Newton-Raphson迭代算法:** 假设 $\ell(\boldsymbol{\beta}; \lambda)$ 满足二阶连续可微, 分别计算梯度向量和Hessian矩阵为

$$\frac{\partial \ell(\boldsymbol{\beta}; \lambda)}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n (y_i - \pi_i) \mathbf{B}(x_i) - \lambda \boldsymbol{\Omega}_n \boldsymbol{\beta},$$

$$\frac{\partial^2 \ell(\boldsymbol{\beta}; \lambda)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = - \sum_{i=1}^n \pi_i (1 - \pi_i) \mathbf{B}(x_i) \mathbf{B}^T(x_i) - \lambda \boldsymbol{\Omega}_n.$$

■ 其中 $\pi_i = \frac{\exp(\boldsymbol{\beta}^T \mathbf{B}(x_i))}{1 + \exp(\boldsymbol{\beta}^T \mathbf{B}(x_i))}$, 且 $i = 1, \dots, n$.

■ 对给定的光滑参数 λ , 可得Newton-Raphson迭代公式为

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} - \left(\frac{\partial^2 \ell(\boldsymbol{\beta}; \lambda)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(k)}} \right)^{-1} \frac{\partial \ell(\boldsymbol{\beta}; \lambda)}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(k)}}.$$

■ 对于给定的 ϵ , 当 $\|\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^{(k)}\|_2^2 < \epsilon$ 时, 停止迭代, 可得参数向量 $\boldsymbol{\beta}$ 最终的极大似然估计数值解 $\hat{\boldsymbol{\beta}}$.

■ 因此, 最终可得 $g(x)$ 和条件概率的估计分别为

$$\hat{g}(x) = \hat{\beta}^T \mathbf{B}(x) = \sum_{j=1}^n \hat{\beta}_j b_j(x), \quad \hat{\mathbb{P}}(Y = 1 | X = x) = \frac{\exp(\hat{g}(x))}{1 + \exp(\hat{g}(x))}.$$

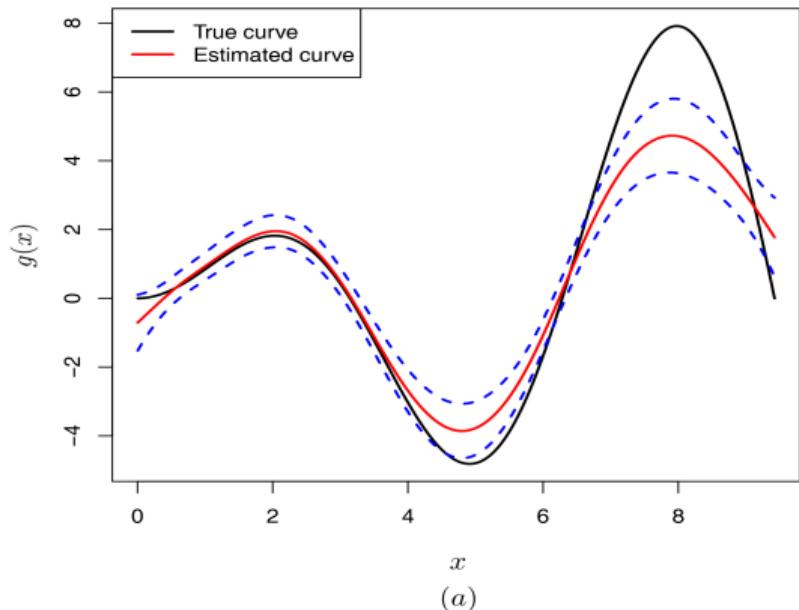
■ 拟合非参数logistic 回归模型的程序包和函数:

- ① 程序包mgcv中的函数gam()
- ② 程序包SemiPar中的函数spm()
- ③ 程序包gam中的函数gam()
- ④ 程序包sm中的函数sm.binomial()
- ⑤ 程序包gss中的函数gssanova()

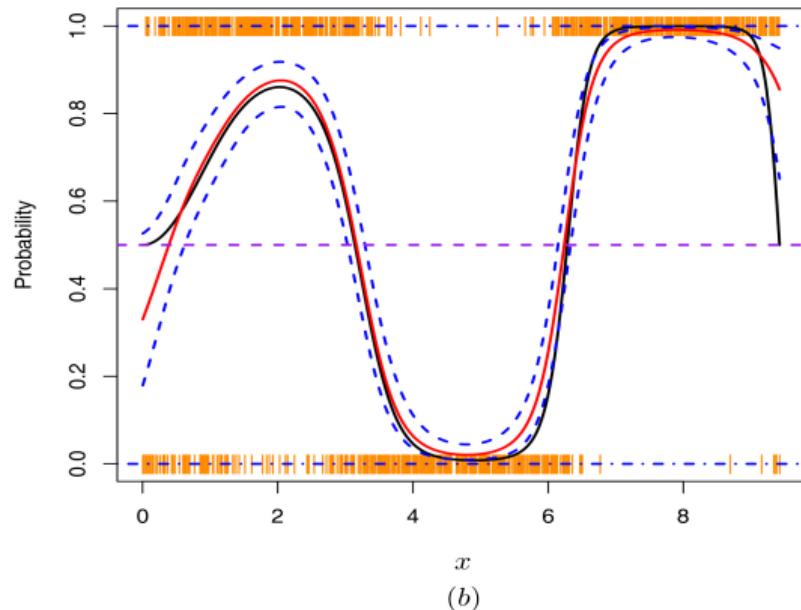
- 从非参数logistic回归模型中产生 $n = 1000$ 个随机样本 $\{(x_i, y_i), i = 1, \dots, n\}$, 其中
 - ▶ x_i 从区间为 $[0, 3\pi]$ 的均匀分布中产生;
 - ▶ 非参数函数取 $g(x_i) = x_i \sin(x_i)$;
 - ▶ 给定 x_i 和 $g(x_i)$ 后, 响应变量 y_i 从 $\text{Bernoulli}(\pi(x_i))$ 中生成0或1的二元变量;
 - ▶ $\pi(x_i) = \frac{\exp(g(x_i))}{1 + \exp(g(x_i))}$, 且 $i = 1, \dots, 1000$.

- 使用程序包gss中的函数gssanova() 拟合非参数logistic回归模型;
- 非参数函数 $g(x)$ 的估计默认为三次光滑样条;
- 绘制非参数函数 $g(x)$ 的真实曲线、拟合曲线和95% 置信带;
- 绘制事件 $\{y_i = 1\}$ 条件概率的真实曲线、拟合曲线和95% 置信带.
- 结果显示: 非参数函数 $g(x)$ 的拟合曲线在边界点和曲率大的位置拟合效果较差, 比较平滑的位置有很好的拟合效果.
- 除了边界点位置, 所提方法对条件概率的拟合都有非常好的效果.

非参数logistic回归的模拟研究



(a)



(b)

(a) 非参数函数 $g(x)$ 的真实曲线、拟合曲线和95%置信带; (b) 条件概率的真实曲线、拟合曲线和95%置信带

- 如果考虑 $\mathbf{X} = (X_1, \dots, X_p)^T$, 则建立如下的非参数广义可加logistic回归模型

$$\log \left(\frac{\mathbb{P}(Y = 1 | \mathbf{X} = \mathbf{x})}{\mathbb{P}(Y = 0 | \mathbf{X} = \mathbf{x})} \right) = \beta_0 + g_1(x_1) + \dots + g_p(x_p),$$

其中 $\mathbf{x} = (x_1, \dots, x_p)^T$, β_0 是截距项, $g_1(\cdot), \dots, g_p(\cdot)$ 是 p 个未知的一元连续光滑函数.

- 同样利用后移算法获得 β_0 和 $g_1(\cdot), \dots, g_p(\cdot)$ 的估计.

- 对程序包ISLR2中的Wage数据集, 用非参数广义可加logistic回归模型来预测Wage数据中个人年收入超过250千美元的可能性.
- 令 $\mathbf{x} = (\text{year}, \text{age}, \text{education})^T$, 非参数广义可加logistic回归模型为

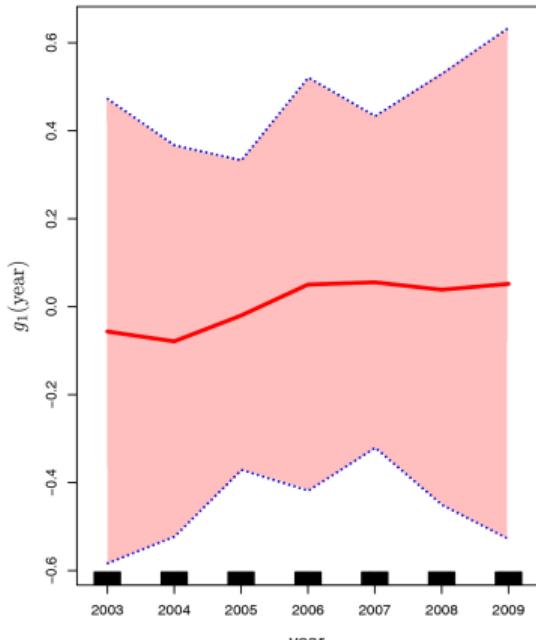
$$\log \left(\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} \right) = \beta_0 + g_1(\text{year}) + g_2(\text{age}) + g_3(\text{education}).$$

- $\pi(\mathbf{x}) = \mathbb{P}(I(\text{wage} > 250) | \text{year}, \text{age}, \text{education})$.
- 利用程序包gam中的函数gam()进行拟合.

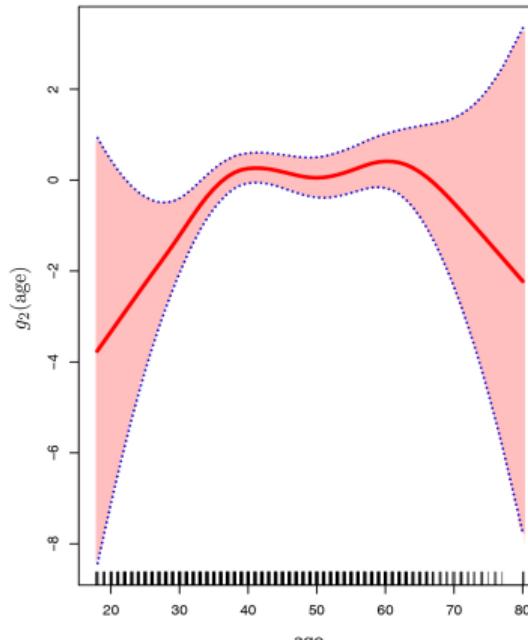
非参数广义可加logistic回归模型— Wage数据分析

```
library(ISLR2)
library(gam)
attach(Wage)
mod = gam(I(wage > 250) ~ ns(year, 4) + s(age, 5) + education,
          family = binomial, data = Wage,
          subset=(education != "1. < HS Grad"))
par(mfrow = c(1, 3))
plot(mod, se = T, col = "red", lwd = 3)
```

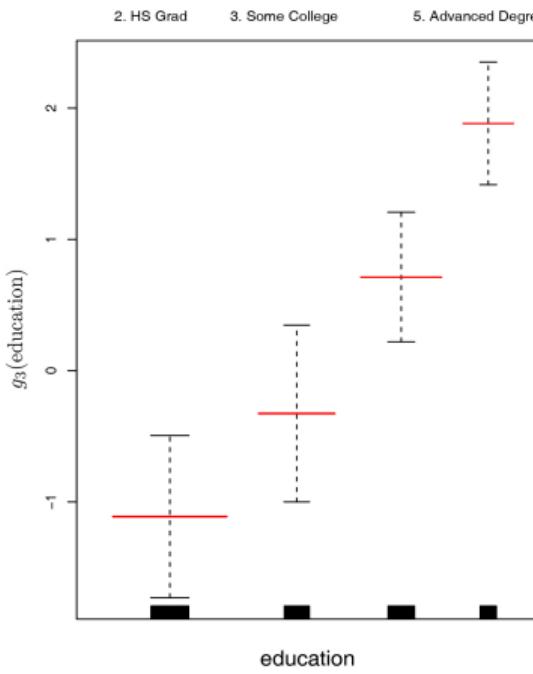
非参数广义可加logistic回归模型—Wage数据分析



(a)



(b)



(c)

本章纲要

1 多元logistic回归

- 多元logistic回归模型
- 极大似然估计
- 预测

2 二分类模型的评估

3 惩罚似然变量选择方法

4 非参数logistic回归

5 多项logistic回归

6 分类方法比较

7 作业

■ 多项**logistic**回归：多分类问题的**logistic**回归.

■ 设 $D = \{(x_i, y_i), i = 1, \dots, n\}$ 为训练样本集，其中

- ▶ $x_i = (1, x_{i1}, \dots, x_{ip})^T \in \mathbb{R}^{p+1}$ 为观测的协变量向量；
- ▶ $y_i \in \{1, 2, \dots, J\}$ 为类别变量，其中 $J \geq 2$.

■ 对于多分类问题，给定第*i*个个体和协变量向量 x_i ，条件概率满

$$\sum_{j=1}^J \mathbb{P}(y_i = j | x_i) = 1.$$

■ 考虑如下的多项logistic回归模型

$$\mathbb{P}(y_i = j | \mathbf{x}_i) = \pi_j(\mathbf{x}_i) = \frac{\exp(\boldsymbol{\beta}_j^T \mathbf{x}_i)}{1 + \sum_{s=2}^J \exp(\boldsymbol{\beta}_s^T \mathbf{x}_i)}, \quad j = 2, \dots, J,$$

其中 $\boldsymbol{\beta}_j = (\beta_{0j}, \beta_{1j}, \dots, \beta_{pj})^T$ 为第 j 类对应的 $p + 1$ 维参数向量.

■ 把第1类($j = 1$)作为参照类别, 且令 $\boldsymbol{\beta}_1 = \mathbf{0}$.

多项logistic回归

■ 根据 $\sum_{j=1}^J \mathbb{P}(y_i = j | \mathbf{x}_i) = 1$, 可得第1类的条件概率为

$$\mathbb{P}(y_i = 1 | \mathbf{x}_i) = \pi_1(\mathbf{x}_i) = 1 - \sum_{j=2}^J \pi_j(\mathbf{x}_i) = \frac{1}{1 + \sum_{s=2}^J \exp(\boldsymbol{\beta}_s^T \mathbf{x}_i)}.$$

■ 响应变量 y 归属于第 j 类 ($j = 2, \dots, J$) 的条件概率与 y 归属于第1类的条件概率之比为

$$\frac{\mathbb{P}(y_i = j | \mathbf{x}_i)}{\mathbb{P}(y_i = 1 | \mathbf{x}_i)} = \exp(\boldsymbol{\beta}_j^T \mathbf{x}_i), \quad j = 2, \dots, J.$$

- 条件概率比称为事件 $\{y_i = j\}$ 与事件 $\{y_i = 1\}$ 发生的发生比 (odds),也称为相对风险(relative risk).
- 根据模型的特点, 可以定义第*i*个观测样本的似然函数为

$$L_i(\beta_2, \dots, \beta_J) = \prod_{j=1}^J \mathbb{P}(y_i = j | \mathbf{x}_i)^{I(y_i=j)}.$$

- 进一步, 第*i*个观测样本的对数似然函数为

$$\ell_i(\beta_2, \dots, \beta_J) = \log(L_i(\beta_2, \dots, \beta_J)) = \sum_{j=1}^J I(y_i=j) \log(\mathbb{P}(y_i=j | \mathbf{x}_i)).$$

多项logistic回归

- 对所有观测样本的对数似然函数进行求和，并关于 β_2, \dots, β_J 极大化求和的对数似然函数，可得 β_2, \dots, β_J 的极大似然估计为

$$\begin{aligned}(\hat{\beta}_2, \dots, \hat{\beta}_J) &= \arg \max_{\beta_2, \dots, \beta_J} \sum_{i=1}^n \ell_i(\beta_2, \dots, \beta_J) \\&= \arg \max_{\beta_2, \dots, \beta_J} \sum_{i=1}^n \sum_{j=1}^J I(y_i = j) \log (\mathbb{P}(y_i = j | \mathbf{x}_i)).\end{aligned}$$

- 可用Newton-Raphson迭代算法求解，在R语言中，可用程序包[nnet](#)中的函数[multinom\(\)](#)进行多项logistic回归分析.

Fisher Iris数据分析

■ Fisher Iris数据集有四个属性:

- ① 萼片长度
- ② 萼片宽度
- ③ 花瓣长度
- ④ 花瓣宽度



Iris Versicolor

■ 数据共有150个样本, 分为三类:

- ① 前50个样本是属于第1类—Setosa
- ② 中间的50个样本属于第2类—Versicolor
- ③ 最后50个样本属于第3类—Virginica



Iris Setosa



Iris Virginica

- 固定种子`set.seed(2023)`, 将该数据集随机分成两部分: 75 个样本作为训练集, 剩余75个样本作为测试集.
- 首先, 在训练集上用程序包nnet中的函数`multinom()`拟合多项logistic回归模型.

```
library(nnet); set.seed(2023)
n = nrow(iris)
index = sample(n, size = trunc(0.50 * n))
iris.train = iris[index, ]; iris.test = iris[-index, ]
fit.multi = multinom(Species~., data=iris.train, trace=FALSE)
```

Fisher Iris数据分析

```
> summary(fit.multi)      ## 输出结果
Call:
multinom(formula = Species ~ ., data = iris.train, trace = FALSE)
Coefficients:
              (Intercept)    S.Length    S.Width     P.Length    P.Width
versicolor     99.223      -19.719    -32.931      42.163    -6.2775
virginica    -112.607      -11.967    -54.004      64.728     61.0027
Std. Errors:
              (Intercept)    S.Length    S.Width     P.Length    P.Width
versicolor     132.48       79.266     38.748      26.457     34.272
virginica     138.42       76.462     42.167      22.916     32.408
Residual Deviance: 3.5194
AIC: 23.519
```

Fisher Iris数据分析

■ 利用函数confint()计算回归系数和优势比(OR)的95%置信区间.

```
> confint(fit.multi)
, , versicolor
              2.5 %   97.5 %
(Intercept) -160.4332 358.879
Sepal.Length -175.0781 135.640
Sepal.Width  -108.8759  43.014
Petal.Length  -9.6907  94.017
Petal.Width   -73.4488  60.894
, , virginica
              2.5 %   97.5 %
(Intercept) -383.9091 158.696
Sepal.Length -161.8295 137.896
Sepal.Width  -136.6499  28.642
Petal.Length   19.8138 109.643
Petal.Width   -2.5166 124.522
> exp(confint(fit.multi))
, , versicolor
              2.5 %   97.5 %
2.1123e-70 7.2289e+155
9.2158e-77 8.0873e+58
5.1974e-48 4.7924e+18
6.1858e-05 6.7796e+40
1.2635e-32 2.7916e+26
, , virginica
              2.5 %   97.5 %
1.8638e-167 8.3335e+68
5.2278e-71 7.7171e+59
4.5050e-60 2.7479e+12
4.0274e+08 4.1413e+47
8.0733e-02 1.2002e+54
```

■ 利用函数predict()在测试集iris.test上进行预测，并计算测试集上的混淆矩阵和测试错误率。

```
pred.test = predict(fit.multi, newdata=iris.test, type="class")
(test.table=table(Predicted=pred.test, Actual=iris.test$Species))
```

		Actual		
Predicted	setosa	versicolor	virginica	
setosa	25	0	0	
versicolor	0	22	2	
virginica	0	1	25	

```
> (accRate = sum(diag(test.table))/sum(test.table))
[1] 0.96
```

- 最后, 利用程序包vcd中的函数Kappa()计算kappa指标.
- 结果表明, 测试集iris.test上的kappa指标为0.940, 表示预测值与真实值之间具有一致性很好(great agreement)的拟合效果.

```
library(vcd)
```

```
> Kappa(test.table)
```

	value	ASE	z	Pr (> z)
Unweighted	0.940	0.0340	27.7	1.5e-168
Weighted	0.956	0.0253	37.8	0.0e+00

本章纲要

1 多元logistic回归

- 多元logistic回归模型
- 极大似然估计
- 预测

2 二分类模型的评估

3 惩罚似然变量选择方法

4 非参数logistic回归

5 多项logistic回归

6 分类方法比较

7 作业

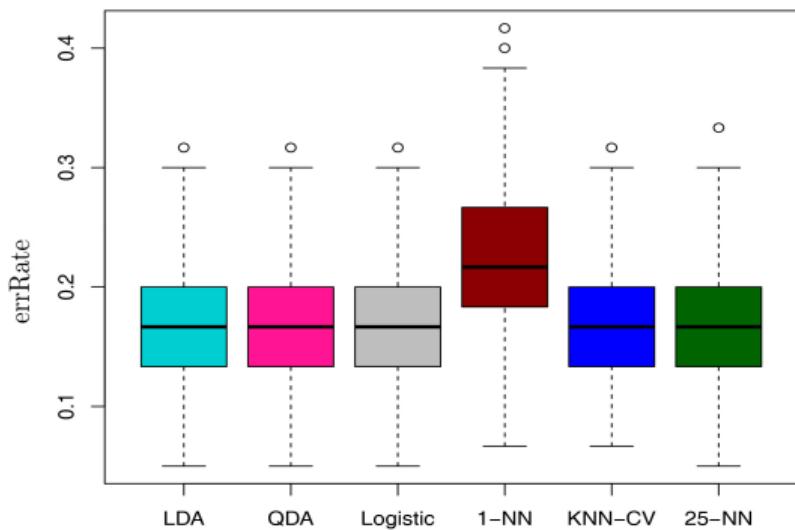
- 在四种情形下，对四种方法进行比较：logistic回归、LDA、QDA和KNN.
- 生成 $n = 200$ 个独立同分布的简单随机样本 $\{(y_i, \mathbf{x}_i), i = 1, \dots, 200\}$ ，其中 $y_i \in \{0, 1\}$ 的二元类别变量， $\mathbf{x}_i = (x_{i1}, x_{i2})^T$ 为二维协变量向量.
- 把 $n = 200$ 个样本随机分为70%的训练集和30%的测试集.
- 在训练集上对四种方法进行拟合，在测试集上计算测试错误率和混淆矩阵，重复模拟500次试验，绘制基于500次试验所得测试错误率的箱线图.
- 绘制ROC曲线和计算AUC值，对不同方法进行比较.

■ **情形 1:** 设置两个二元正态总体, 分别为 $N_2(\mu_1, \Sigma_1)$ 和 $N_2(\mu_2, \Sigma_2)$. 情形1考虑两个正态总体的均值向量不同, 而协方差矩阵相同, 即 $\mu_1 \neq \mu_2, \Sigma_1 = \Sigma_2 = \Sigma$, 其中

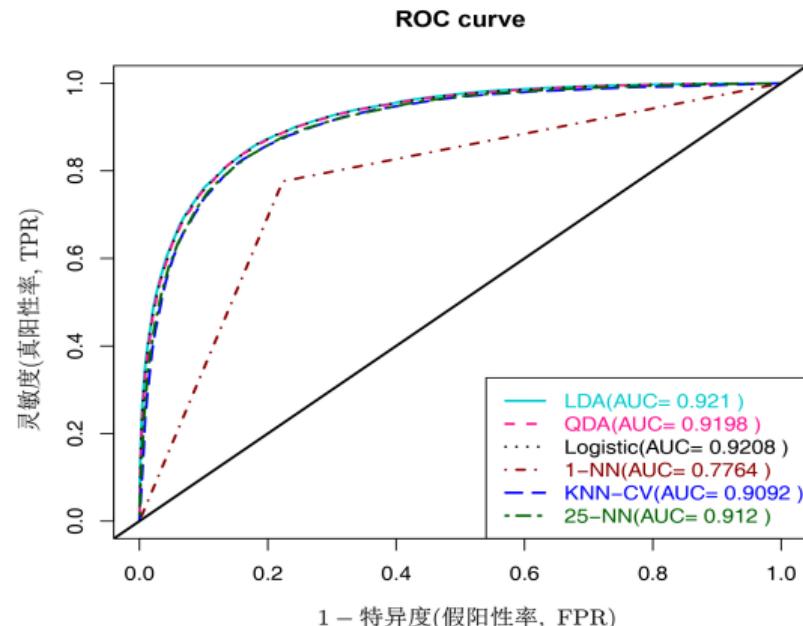
$$\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}.$$

- 当 $\Sigma_1 = \Sigma_2 = \Sigma$ 时, 决策边界为线性的.
- 因此, 情形1对线性分类方法更有利, 如LDA方法和logistic回归方法.

分类方法比较



(a)



(b)

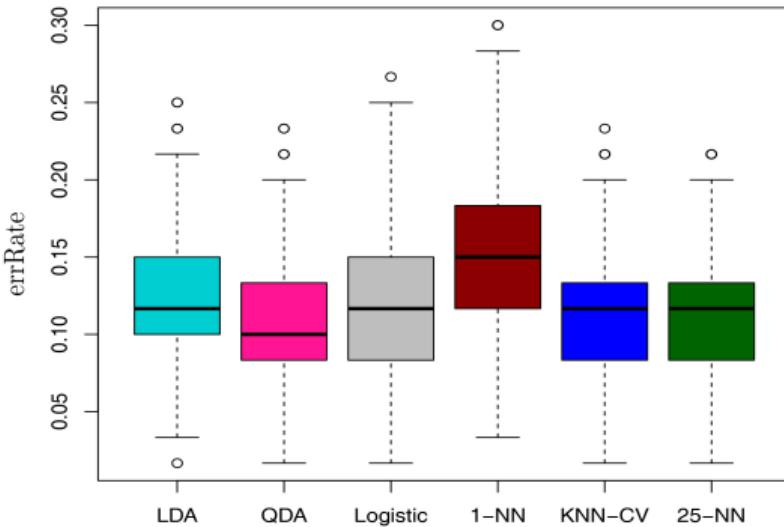
情形1：六种分类方法的测试错误率箱线图和ROC曲线图. (a) 六种分类方法的测试错误率箱线图; (b) 六种分类方法的ROC曲线和AUC值

■ **情形 2:** 数据产生过程与情形1类似, 但情形2考虑两个正态总体的协方差矩阵不同, 即 $\Sigma_1 \neq \Sigma_2$, 其中

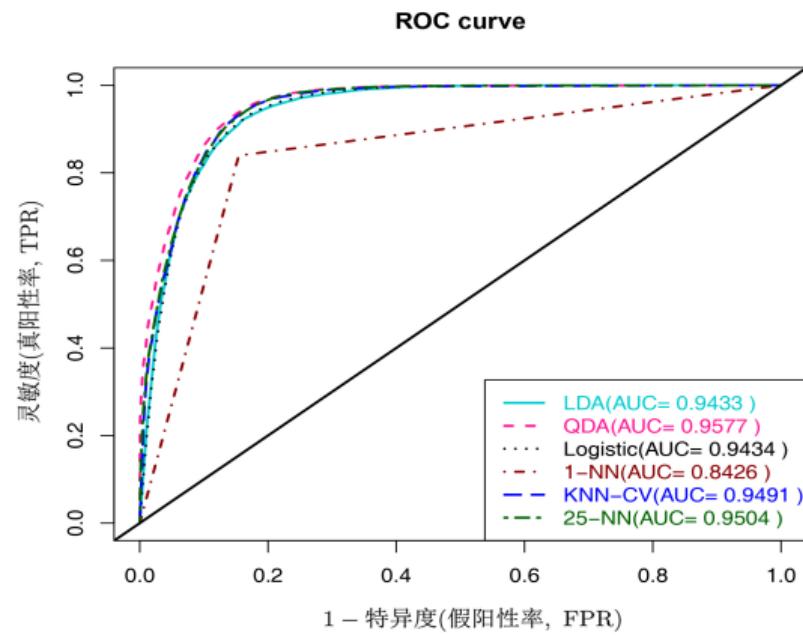
$$\Sigma_1 = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 1 & -0.6 \\ -0.6 & 1 \end{pmatrix}.$$

- 当 $\Sigma_1 \neq \Sigma_2$ 时, 决策边界为非线性的.
- 因此, 情形2对非线性分类方法更有利, 如QDA和KNN分类方法.

分类方法比较



(a)



(b)

情形2: 六种分类方法的测试错误率箱线图和ROC曲线图. (a) 六种分类方法的测试错误率箱线图; (b) 六种分类方法的ROC曲线和AUC值

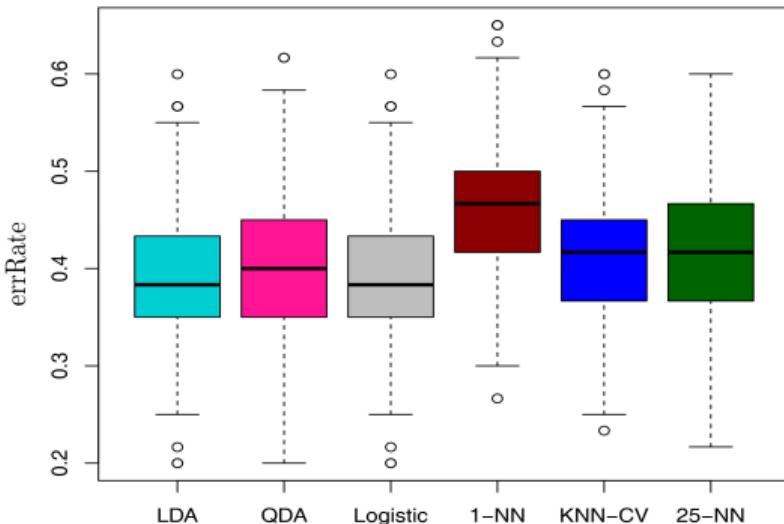
■ **情形3:** 假设从logistic回归模型中产生 $n = 200$ 个随机样本 $\{(\mathbf{x}_i, y_i), i = 1, \dots, 200\}$, 其中 $\mathbf{x}_i = (x_{i1}, x_{i2})^T \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, 其中 $\boldsymbol{\mu} = (0, 2)^T$, 且协方差矩阵 $\boldsymbol{\Sigma}$ 相同于情形1和情形2中的 $\boldsymbol{\Sigma}_1$.

■ 给定 \mathbf{x}_i 后, 响应变量 y_i 从 $Bernoulli(\pi(\mathbf{x}_i))$ 中生成0或1的二元变量, 其中

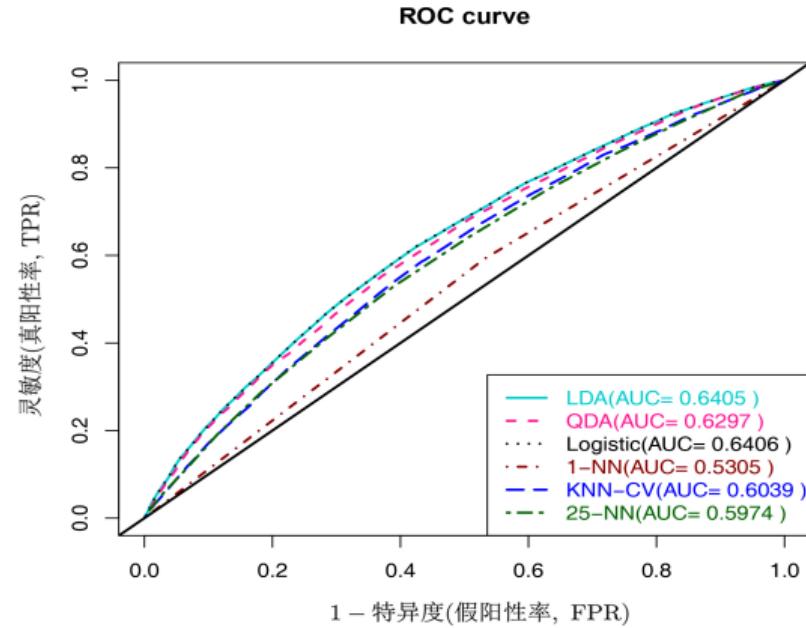
$$\pi(\mathbf{x}_i) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})},$$

其中 $\beta_0 = 0.3, \beta_1 = 0.7, \beta_2 = -0.4$.

分类方法比较



(a)



(b)

情形3: 六种分类方法的测试错误率箱线图和ROC曲线图. (a) 六种分类方法的测试错误率箱线图; (b) 六种分类方法的ROC曲线和AUC值

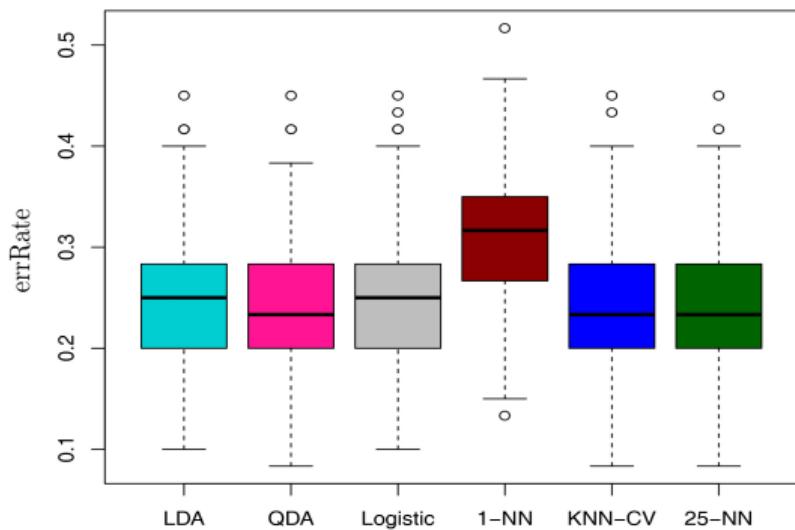
■ **情形 4:** 数据产生类似于情形3, 仅在logistic回归模型中增加了平方项 x_{i1}^2 和 x_{i2}^2 , 以及交互项 $x_{i1}x_{i2}$. 这时, $\pi(\mathbf{x}_i)$ 为

$$\pi(\mathbf{x}_i) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}^2 + \beta_4 x_{i2}^2 + \beta_5 x_{i1}x_{i2})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}^2 + \beta_4 x_{i2}^2 + \beta_5 x_{i1}x_{i2})},$$

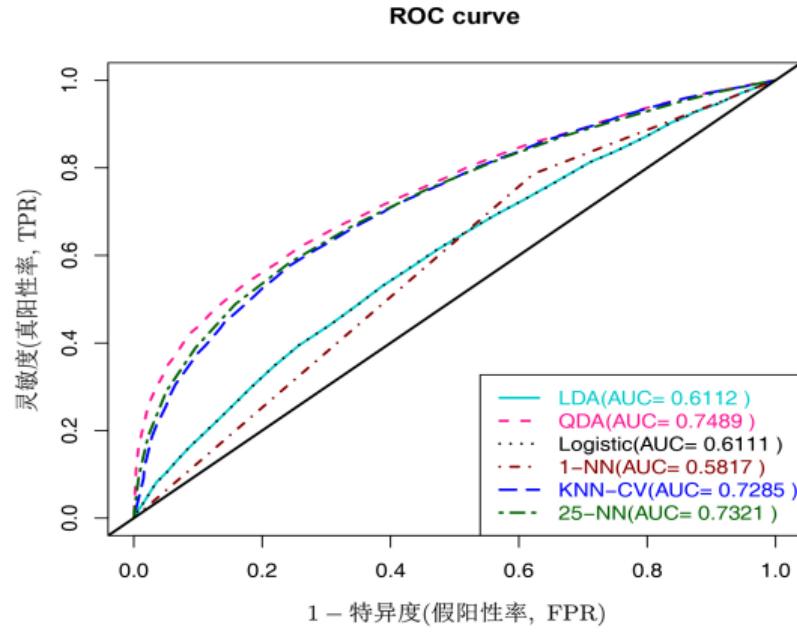
其中 $\beta_0 = 0.3, \beta_1 = 0.7, \beta_2 = -0.4, \beta_3 = 0.4, \beta_4 = 0.5, \beta_5 = 1.$

■ 当模型中增加二次项和交互项后, 更有利于非线性的分类方法.

分类方法比较



(a)



(b)

情形4: 六种分类方法的测试错误率箱线图和ROC曲线图. (a) 六种分类方法的测试错误率箱线图; (b) 六种分类方法的ROC曲线和AUC值

本章纲要

1 多元logistic回归

- 多元logistic回归模型
- 极大似然估计
- 预测

2 二分类模型的评估

3 惩罚似然变量选择方法

4 非参数logistic回归

5 多项logistic回归

6 分类方法比较

7 作业

作业

[习题见教材: 统计学习(R语言版) — 习题9]

- **课后思考题:** 第5题、第6题、第15题
- **需要完成的课后作业:** 第2题、第3题、第14题
- **应用:** 第4题、第8题、第10题. 具体要求:
 - ① 能使用R语言把数据读入, 并对数据中的每个变量进行了解;
 - ② 能用学过的一些统计方法, 按照题目要求, 利用R语言对数据进行一些简单的分析, 并思考数据分析的结果.



谢谢，请多提宝贵意见！