第 5 章 逻辑回归

在统计学中，二分类问题的常见估计方法为“逻辑回归”(logistic regression，简记 Logit)；尽管这是机器学习的分类问题。

5.1 逻辑回归

在监督学习中，存在大量关于“是与否”的二分类问题，在各行业有着广泛的应用。

比如，人脸识别(是否为人脸、是否为某人的人脸)、自动驾驶(是否应刹车)、银行贷款(是否批准贷款申请)、邮件过滤(是否为垃圾邮件)等。

以过滤垃圾邮件为例，假设响应变量只有两种可能取值，即$y=1$(垃圾邮件)或$y=0$(正常邮件)。这种 0-1 变量称为**虚拟变量**(dummy variable)或“哑变量”。

记特征向量为$\boldsymbol{x}_i \equiv (x_{i1}\ x_{i2}\ \cdots\ x_{ip})'$，比如不同词汇出现的频率。

最简单的建模方法为“线性概率模型”(Linear Probability Model)：

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i = \boldsymbol{x}_i' \boldsymbol{\beta} + \varepsilon_i \quad (i = 1, \cdots, n) \tag{5.1}$$

其中，参数向量$\boldsymbol{\beta} \equiv (\beta_1\ \beta_2\ \cdots\ \beta_p)'$。

线性概率模型的优点在于，计算方便(就是 OLS 估计)，且容易得到边际效应(即回归系数)。

但线性概率模型一般并不适合作预测。

明知y的取值非 0 即 1，但根据线性概率模型所作的预测值却可能出现$\hat{y} > 1$或$\hat{y} < 0$的不现实情形，参见图 5.1。

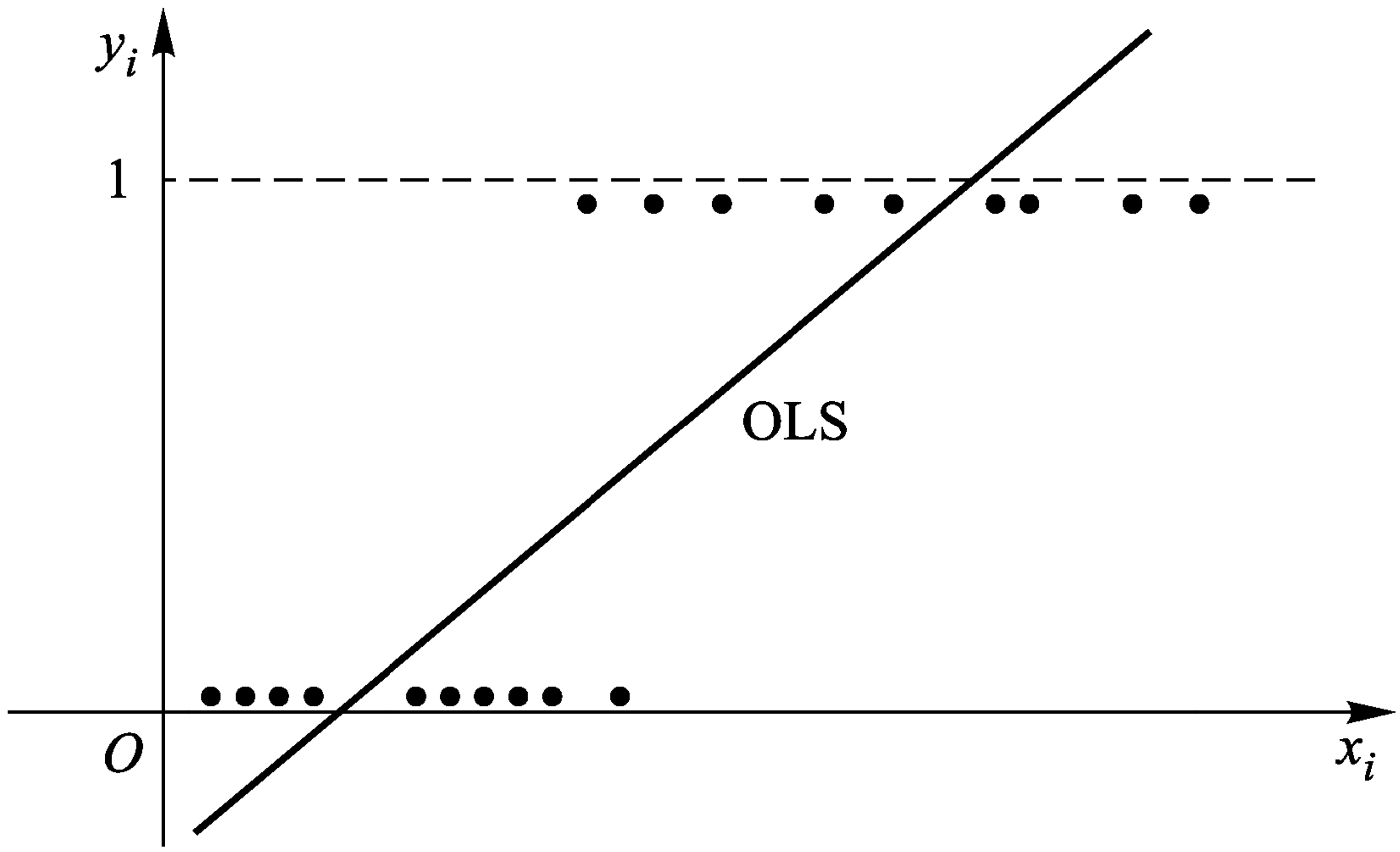

图 5.1　线性概率模型

为使y的预测值总是介于$[0,1]$之间，在给定$\boldsymbol{x}$的情况下，考虑y的两点分布概率：

$$\begin{cases} \mathrm{P}(y=1\,|\,\boldsymbol{x}) = F(\boldsymbol{x},\boldsymbol{\beta}) \\ \mathrm{P}(y=0\,|\,\boldsymbol{x}) = 1-F(\boldsymbol{x},\boldsymbol{\beta}) \end{cases} \tag{5.2}$$

其中，函数$F(\mathbf{x},\boldsymbol{\beta})$称为**连接函数**(link function) ，因为它将特征向量$\mathbf{x}$与响应变量y连接起来。

连接函数的选择具有一定的灵活性。通过选择合适的连接函数$F(\boldsymbol{x},\boldsymbol{\beta})$(比如，某随机变量的累积分布函数)，可以保证$0 \le \hat{y} \le 1$。

在给定$\boldsymbol{x}$的情况下，y的条件期望为

$$\mathrm{E}(y\,|\,\boldsymbol{x}) = 1\cdot \mathrm{P}(y=1\,|\,\boldsymbol{x}) + 0\cdot \mathrm{P}(y=0\,|\,\boldsymbol{x}) = \mathrm{P}(y=1\,|\,\boldsymbol{x}) \tag{5.3}$$

可将模型的拟合值(预测值)理解为事件“$y=1$”的发生概率。如果$F(\boldsymbol{x},\boldsymbol{\beta})$为标准正态的累积分布函数(cumulative distribution function)，则

$$\mathrm{P}(y=1\,|\,\boldsymbol{x}) = F(\boldsymbol{x},\boldsymbol{\beta}) = \Phi(\boldsymbol{x}'\boldsymbol{\beta}) \equiv \int_{-\infty}^{\boldsymbol{x}'\boldsymbol{\beta}} \phi(t)\,dt \tag{5.4}$$

其中，$\phi(\cdot)$与$\Phi(\cdot)$分别为标准正态的密度函数与累积分布函数。此模型称为**概率单位模型**(Probit)。

由于标准正态的密度函数之积分并无解析表达式，须进行数值积分，故计算不便。

由于回归参数$\boldsymbol{\beta}$出现于积分上限，故无法解释其含义。

如果连接函数$F(\boldsymbol{x}, \boldsymbol{\beta})$为“逻辑分布”(logistic distribution)的累积分布函数，则

$$\mathrm{P}(y=1 \mid \boldsymbol{x}) = F(\boldsymbol{x}, \boldsymbol{\beta}) = \Lambda(\boldsymbol{x}'\boldsymbol{\beta}) \equiv \frac{\exp(\boldsymbol{x}'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}'\boldsymbol{\beta})} = \frac{1}{1+\exp(-\boldsymbol{x}'\boldsymbol{\beta})} \tag{5.5}$$

其中，函数$\Lambda(\cdot)$的定义为$\Lambda(z) \equiv \dfrac{e^z}{1+e^z} = \dfrac{1}{1+e^{-z}}$。

此模型称为**逻辑回归**(Logistic Regression)或“逻辑斯蒂回归”，简记 Logit。

对逻辑函数$\Lambda(\cdot)$求导数，即可得到逻辑分布的密度函数：

$$\begin{aligned}\frac{d\Lambda(z)}{dz} &= \frac{d(1+e^{-z})^{-1}}{dz} = (-1)(1+e^{-z})^{-2}e^{-z}(-1) \\ &= \frac{1}{1+e^{-z}}\cdot\frac{e^{-z}}{1+e^{-z}} = \Lambda(z)\left[1-\Lambda(z)\right] = \frac{e^{z}}{(1+e^{z})^{2}}\end{aligned} \tag{5.6}$$

在 Python 中，画逻辑分布的密度函数与累积分布函数，可输入命令：

```
In [1]: import numpy as np
   ...: from scipy.stats import logistic
   ...: import matplotlib.pyplot as plt
   ...: fig, ax = plt.subplots(1, 2, figsize=(8, 4))
   ...: x = np.linspace(-5, 5, 100)
   ...: ax[0].plot(x, logistic.pdf(x), linewidth=2)
   ...: ax[0].vlines(0, 0, .255, linewidth=1)
   ...: ax[0].hlines(0, -5, 5, linewidth=1)
   ...: ax[0].set_title('Logistic Density')
   ...: ax[1].plot(x, logistic.cdf(x), linewidth=2)
   ...: ax[1].vlines(0, 0, 1, linewidth=1)
   ...: ax[1].hlines(0, -5, 5, linewidth=1)
   ...: ax[1].set_title('Logistic CDF')
```

其中，第 2 个命令从 SciPy 的 stats 子模块导入 `logistic` 类(class)；`logistic.pdf()`与 `logistic.cdf()`分别为逻辑分布的密度函数与

累积分布函数。第 4 个命令将画布分为1×2的画轴，而整个画布尺寸为 8 英寸宽与 4 英寸高。结果参见图 5.2。

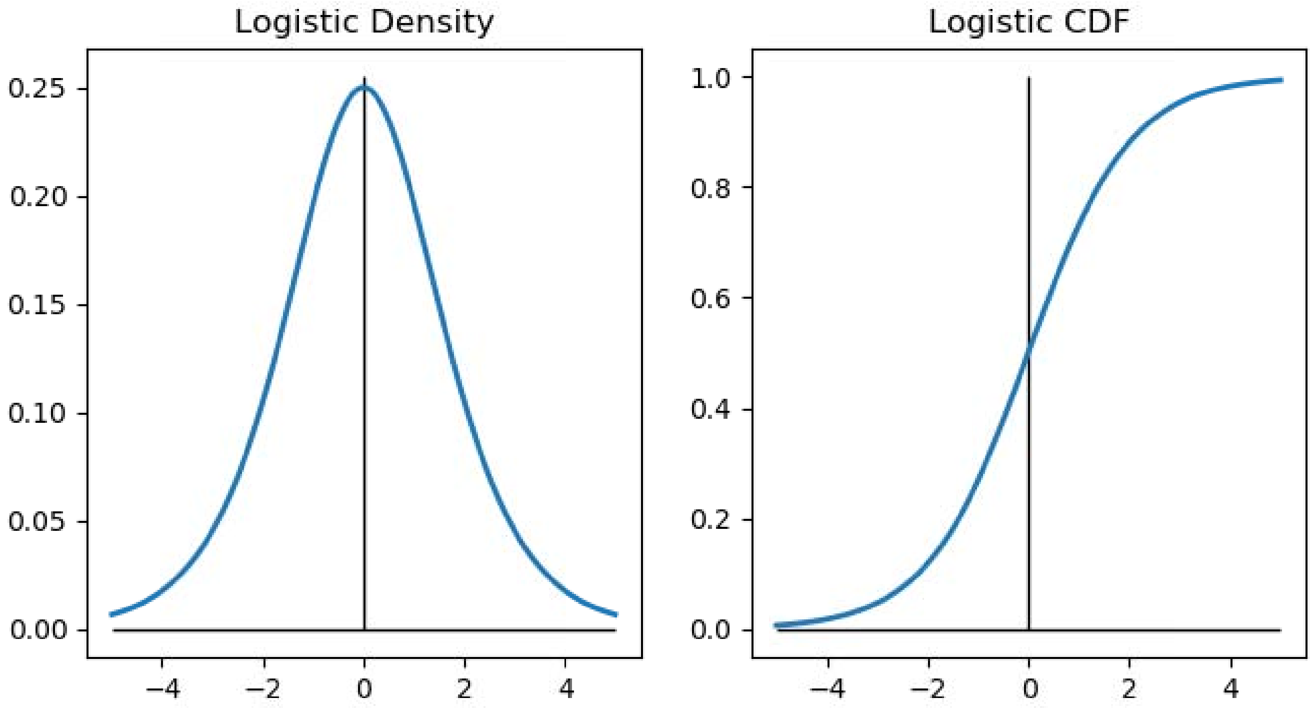

图 5.2　逻辑分布的密度函数与累积分布函数

逻辑分布的密度函数关于原点对称，期望为 0。

由于逻辑分布的累积分布函数之形状类似于(拉长的)大写英文字母 S，故在机器学习中常称为**S 型函数**(sigmoid function)，记为$\sigma(z)$，广泛用于神经网络模型(参见第 15 章)。

有时，S 型函数(sigmoid function)也泛指所有形如 S 的函数。

逻辑分布的累积分布函数$\Lambda(\cdot)$有解析表达式，故计算 Logit 更为方便。

由于逻辑函数$\Lambda(z)=\dfrac{e^z}{1+e^z}$的形式很好，故可通过“几率比”(odds ratio)解释 Logit 回归系数$\boldsymbol{\beta}$的意义。

在统计学中，将 Probit 与 Logit 模型统称为**广义线性模型**(Generalized

Linear Model，简记 GLM)，因为二者的模型均可写为如下形式：

$$\mathrm{P}(y=1\,|\,\boldsymbol{x})=F(\boldsymbol{x}'\boldsymbol{\beta}) \qquad (5.7)$$

在上式中，非线性的连接函数$F(\cdot)$作用于线性函数$\boldsymbol{x}'\boldsymbol{\beta}$，由此决定“$y=1$”的条件概率$\mathrm{P}(y=1\,|\,\boldsymbol{x})$。

如果使用某连续变量的累积分布函数作为连接函数$F(\cdot)$，则$F(\cdot)$为严格单调函数，故其逆函数$F^{-1}(\cdot)$存在。

以此逆函数$F^{-1}(\cdot)$作用于方程的两边，则可得到一个线性模型：

$$F^{-1}\left[\mathrm{P}(y=1\,|\,\boldsymbol{x})\right]=\boldsymbol{x}'\boldsymbol{\beta} \tag{5.8}$$

对于 Logit 模型，此逆函数$F^{-1}(\cdot)$为“对数几率”(log odds)，也称为“逻辑变换”(logit transformation)，它将概率变换为相应的对数几率(详见下文)。

5.2 最大似然估计

Logit 模型本质上为非线性模型，故一般使用**最大似然估计**(Maximum Likelihood Estimation，简记 MLE)，而不使用最小二乘法。

尽管可以用“非线性最小二乘法”(Nonlinear Least Square，简记 NLS)估计 Logit 模型(依然最小化残差平方和)，但 NLS 的估计效率不及 MLE。

考虑第i个观测值，由于$y_i = 0$或 1，故y_i服从“两点分布”(Bernoulli distribution)，这是“二项分布”（binomial distribution）的特例。

第i个观测数据的条件概率为

$$P(y_i \mid \boldsymbol{x}_i) = \begin{cases} \Lambda(\boldsymbol{x}_i'\boldsymbol{\beta}), & if \;\; y_i = 1 \\ 1 - \Lambda(\boldsymbol{x}_i'\boldsymbol{\beta}), & if \;\; y_i = 0 \end{cases} \tag{5.9}$$

其中，$\Lambda(z) = \dfrac{e^z}{1+e^z}$为逻辑分布的累积分布函数。将其更紧凑地写为

$$P(y_i \mid \boldsymbol{x}_i) = \left[\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})\right]^{y_i} \left[1 - \Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})\right]^{1-y_i} \tag{5.10}$$

方程(5.10)与(5.9)等价(分别代入y_i = 0 或 1 即可验证)。

给定训练样本$\left\{\mathbf{x}_i, y_i\right\}_{i=1}^n$，假设样本中的个体相互独立，则整个样本的联合概率为

$$P(\mathbf{y} \mid \mathbf{X}) = \prod_{i=1}^n \left[\Lambda(\boldsymbol{x}_i' \boldsymbol{\beta})\right]^{y_i} \left[1 - \Lambda(\boldsymbol{x}_i' \boldsymbol{\beta})\right]^{1-y_i} \qquad (5.11)$$

该样本的**似然函数**(likelihood function)：

$$L(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) = \prod_{i=1}^n \left[\Lambda(\boldsymbol{x}_i' \boldsymbol{\beta})\right]^{y_i} \left[1 - \Lambda(\boldsymbol{x}_i' \boldsymbol{\beta})\right]^{1-y_i} \qquad (5.12)$$

将上式取对数，可得**对数似然函数**(loglikelihood function)，并选择参数$\boldsymbol{\beta}$使其最大化：

$$\max_{\boldsymbol{\beta}} \ \ln L(\boldsymbol{\beta} \mid \boldsymbol{y}, \mathbf{X}) = \sum_{i=1}^{n} y_i \ln\left[\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})\right] + \sum_{i=1}^{n} (1 - y_i) \ln\left[1 - \Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})\right] \tag{5.13}$$

所得最优解$\hat{\boldsymbol{\beta}}$即为最大似然估计(MLE)。

由于目标函数为非线性函数，故不存在解析解。

一般使用数值计算的方法，比如牛顿法，求解此非线性最大化问题。

具体来说，利用逻辑函数$\Lambda(\cdot)$的导数公式，可得对数似然函数(5.13)的梯度向量：

$$
\begin{aligned}
\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial \sum_{i=1}^{n} y_i \ln\left[\Lambda(\mathbf{x}_i'\boldsymbol{\beta})\right]}{\partial \boldsymbol{\beta}} + \frac{\partial \sum_{i=1}^{n} (1-y_i) \ln\left[1-\Lambda(\mathbf{x}_i'\boldsymbol{\beta})\right]}{\partial \boldsymbol{\beta}} \\
&= \sum_{i=1}^{n} y_i \frac{1}{\Lambda_i} \Lambda_i (1-\Lambda_i)\mathbf{x}_i - \sum_{i=1}^{n} (1-y_i) \frac{1}{1-\Lambda_i} \Lambda_i (1-\Lambda_i)\mathbf{x}_i \\
&= \sum_{i=1}^{n} y_i (1-\Lambda_i)\mathbf{x}_i - \sum_{i=1}^{n} (1-y_i)\Lambda_i \mathbf{x}_i \\
&= \sum_{i=1}^{n} (y_i - \Lambda_i)\mathbf{x}_i
\end{aligned} \tag{5.14}
$$

其中，记$\Lambda_i \equiv \Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})$。

对梯度向量(5.14)再次求偏导数，可得黑塞矩阵：

$$\frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}'} = \frac{\partial\sum_{i=1}^{n}\big(y_i - \Lambda(\mathbf{x}_i'\boldsymbol{\beta})\big)\mathbf{x}_i}{\partial\boldsymbol{\beta}'} = \underbrace{\frac{\partial\sum_{i=1}^{n} y_i\mathbf{x}_i}{\partial\boldsymbol{\beta}'}}_{=\mathbf{0}} - \frac{\partial\sum_{i=1}^{n}\Lambda(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i}{\partial\boldsymbol{\beta}'}$$

$$= -\sum_{i=1}^{n}\Lambda_i(1-\Lambda_i)\mathbf{x}_i\mathbf{x}_i' \tag{5.15}$$

在上式中，由于$0<\Lambda_i<1$，故黑塞矩阵为负定(negative definite)。

故 Logit 的对数似然函数为“凹函数”(concave function)，一定存在唯一的最大值。

将梯度向量(5.14)与黑塞矩阵(5.15)代入牛顿法的迭代公式，即可得到迭代算法。

可将此迭代公式视为“加权最小二乘法”(Weighted Least Squares)的解，而此权重在每步迭代时需更新，故称为“迭代重加权最小二乘法”(Iterative Reweighted Least Squares，简记 IRLS)。

由于黑塞矩阵$\dfrac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\,\partial \boldsymbol{\beta}'}$的表达式(5.15)并不包含响应变量$y$，故它等价于黑塞矩阵的期望$\mathrm{E}\left(\dfrac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\,\partial \boldsymbol{\beta}'}\right)$(给定$\mathbf{x}$，对$y$求条件期望)。

在统计学中，黑塞矩阵的期望之负数称为“费雪信息矩阵”(Fisher information matrix)。

对于 Logit 模型，IRLS 也称为“费雪得分迭代”(Fisher scoring iteration)。

5.3 Logit 模型的解释

对于非线性模型，其估计量$\hat{\boldsymbol{\beta}}$一般并非边际效应(marginal effects)。

对于 Logit 模型，可使用微积分的链式法则(chain rule)，计算第k个特征变量x_k的边际效应：

$$\frac{\partial \mathrm{P}(y=1|\boldsymbol{x})}{\partial x_k} = \frac{\partial \Lambda(\boldsymbol{x}'\boldsymbol{\beta})}{\partial x_k} = \frac{\partial \Lambda(\boldsymbol{x}'\boldsymbol{\beta})}{\partial(\boldsymbol{x}'\boldsymbol{\beta})} \cdot \frac{\partial(\boldsymbol{x}'\boldsymbol{\beta})}{\partial x_k} = \lambda(\boldsymbol{x}'\boldsymbol{\beta}) \cdot \beta_k$$

(5.16)

其中，$\lambda(z) \equiv \dfrac{e^z}{(1+e^z)^2}$为逻辑分布的密度函数。由表达式(5.16)可知，

非线性模型的边际效应通常不是常数，随着特征向量$\mathbf{x}$而变。

可根据样本数据，计算**平均边际效应**(Average Marginal Effects，简记 AME)，即分别计算在每个样本点$\mathbf{x}_i$上的边际效应，然后针对整个样本进行平均。

记变量x_k对于个体i的边际效应为$\widehat{AME}_{ik}$（$\widehat{\cdot}$ 表示为样本估计值)，则变量x_k的平均边际效应为

$$\widehat{AME}_k = \frac{1}{n}\sum_{i=1}^{n}\widehat{AME}_{ik} \quad (k=1,\cdots,p) \quad (5.17)$$

其中，$\widehat{AME}_{ik} = \lambda(\mathbf{x}_i'\hat{\boldsymbol{\beta}})\cdot\hat{\beta}_k$，参见式(5.16)。

既然$\hat{\boldsymbol{\beta}}$并非边际效应，那么它究竟有什么含义？记事件“$y=1$”发生的条件概率为$p \equiv \mathrm{P}(y=1 \mid \boldsymbol{x})$，则该事件不发生的条件概率为$1-p=\mathrm{P}(y=0 \mid \boldsymbol{x})$。

对于 Logit 模型，由于$p=\dfrac{\exp(\boldsymbol{x}'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}'\boldsymbol{\beta})}$，而$1-p=\dfrac{1}{1+\exp(\boldsymbol{x}'\boldsymbol{\beta})}$，故事件发生与不发生的“几率”为

$$\text{几率} \equiv \frac{p}{1-p}=\exp(\boldsymbol{x}'\boldsymbol{\beta}) \quad (5.18)$$

其中，$\dfrac{p}{1-p}$称为**几率**(odds)或**相对风险**(relative risk)。

例如，在一个检验药物疗效的随机实验中，“$y=1$”表示“生”，而“$y=0$”表示“死”。如果几率为 2，则意味着存活的概率是死亡概率的两倍，故存活概率为 2/3，而死亡概率为 1/3。

对方程(5.18)两边取对数可得：

$$\ln\left(\frac{p}{1-p}\right)=\boldsymbol{x}'\boldsymbol{\beta}=\beta_1 x_1+\cdots+\beta_p x_p \tag{5.19}$$

其中，$\ln\left(\frac{p}{1-p}\right)$称为**对数几率**(log-odds)，而上式右边为线性函数，这正是上文提及的“逻辑变换”(logit transformation)。

根据方程(5.19)，回归系数β_k表示当变量x_k增加一个微小量时，引起对数几率的边际变化：

$$\beta_k = \frac{\partial \ln\left(\frac{p}{1-p}\right)}{\partial x_k} \approx \frac{\Delta\left(\frac{p}{1-p}\right)\Big/\frac{p}{1-p}}{\Delta x_k} \qquad (5.20)$$

这意味着，可将β_k解释为**半弹性**(semi-elasticity)，即当x_k增加 1 单位，可引起几率$\left(\frac{p}{1-p}\right)$变化的百分比：

$$\frac{\Delta odds}{odds} = \frac{\Delta\left(\frac{p}{1-p}\right)}{\frac{p}{1-p}} \approx \beta_k \cdot \underbrace{\Delta x_k}_{=1} = \beta_k \qquad (5.21)$$

例如，$\beta_k = 0.12$，意味着x_k增加 1 单位可引起几率增加 12%。

以上解释隐含假设x_k为连续变量，且可求导数。

如果x_k为离散变量(比如，性别、子女数)，则无法微分，可使用如下解释方法。

假设x_k增加 1 单位，从x_k变为x_k+1，记概率p的新值为p^*，则可根据新几率$\frac{p^*}{1-p^*}$与原几率$\frac{p}{1-p}$的比率定义**几率比**(odds ratio)：

$$几率比 \equiv \frac{\dfrac{p^*}{1-p^*}}{\dfrac{p}{1-p}} = \frac{\exp\left[\beta_1 x_1 + \cdots + \beta_k (x_k + 1) + \cdots + \beta_p x_p\right]}{\exp(\beta_1 x_1 + \cdots + \beta_k x_k + \cdots + \beta_p x_p)} = \exp(\beta_k) \tag{5.22}$$

若$\beta_k = 0.12$，则几率比$\exp(\beta_k) = e^{0.12} = 1.13$。

这意味着，当x_k增加 1 单位时，新几率变为原几率的 1.13 倍，或增加 13%，因为$\exp(\beta_k) - 1 = 1.13 - 1 = 0.13$。

如果β_k较小，则$\exp(\beta_k) - 1 \approx \beta_k$(将$\exp(\beta_k)$泰勒展开)，则以上两种方法基本等价。

如果x_k至少必须变化 1 个单位(比如性别、婚否等虚拟变量，以及子女个数等)，则应使用$\exp(\beta_k)$。

若使用 Probit 模型，由于其连接函数较为复杂，故无法使用几率比对其系数$\hat{\beta}$进行类似的解释，这是 Probit 模型的劣势。

5.4 非线性模型的拟合优度

非线性模型不存在平方和分解公式，一般无法使用R^2度量拟合优度。

对于使用 MLE 进行估计的非线性模型，可使用**准**R^2(Pseudo R^2)或**伪**R^2度量模型的拟合优度。准R^2由 McFadden (1974)提出，其定义为

$$\text{准}R^2 \equiv \frac{\ln L_0 - \ln L_1}{\ln L_0} \qquad (5.23)$$

其中，$\ln L_1$为原模型的对数似然函数之最大值，而$\ln L_0$为以常数项为唯一变量的对数似然函数之最大值。

由于y为离散的两点分布，似然函数的最大可能值为 1(即取值概率为 1)，

故对数似然函数的最大可能值为 0，记为$\ln L_{\max}$。

显然，$0 \geq \ln L_1 \geq \ln L_0$，而$0 \leq 准R^2 \leq 1$，参见图 5.3。

由于 $\ln L_{\max} = 0$ ，故可将“准R^2”写为

$$准R^2 = \frac{\ln L_1 - \ln L_0}{\ln L_{\max} - \ln L_0} \tag{5.24}$$

其中，分子为加入除常数项外的变量后，对数似然函数的实际增加值$(\ln L_1 - \ln L_0)$；而分母为对数似然函数的最大可能增加值$(\ln L_{\max} - \ln L_0)$。

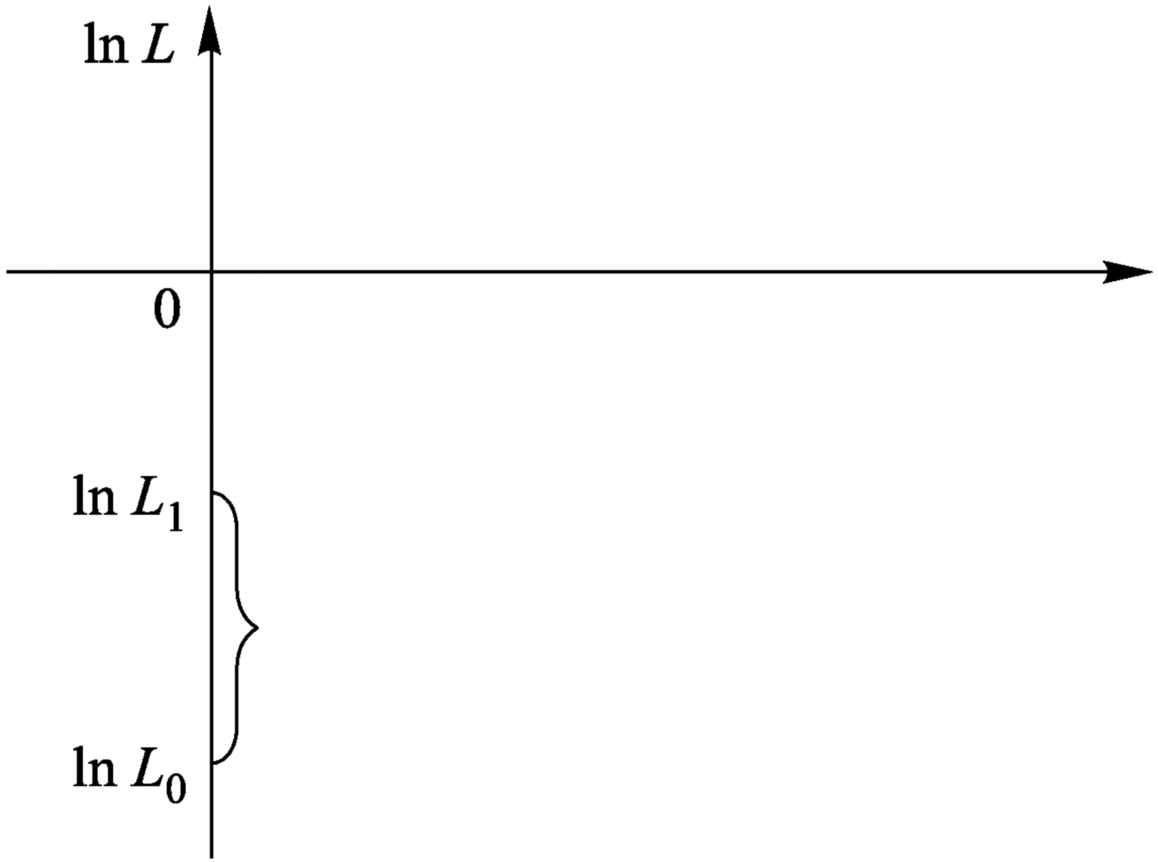

图 5.3　准R^2的计算

在统计学中，还常使用**偏离度**(deviance)的概念。偏离度也称为**残差偏离**

度(residual deviance)，其定义为

$$\text{(残差) 偏离度} \equiv -2\ln L_1 \qquad (5.25)$$

其中，$\ln L_1$为原模型的对数似然函数之最大值。

偏离度表达式中的 2，只是为了凑成统计学中“似然比检验”(likelihood ratio test)统计量而设。

偏离度表达式中的负号，则使得最大化$2\ln L_1$的问题，变为最小化$-2\ln L_1$，参见图 5.4。

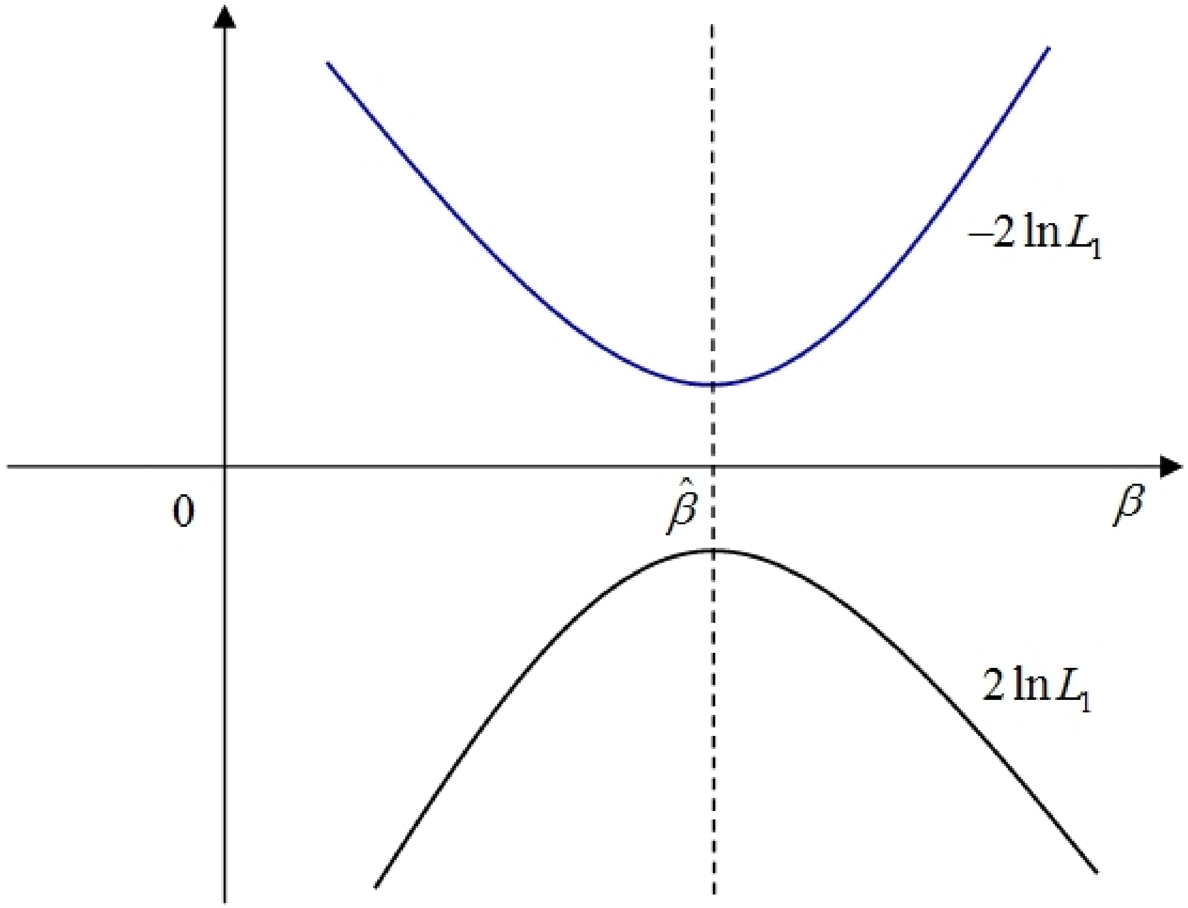

图 5.4　偏离度的示意图

在线性模型中，最小化问题的目标函数为“残差平方和”(Residual Sum of Squares，简记 RSS)。

从图 5.4 可知，对非线性模型进行 MLE 估计，残差偏离度$-2\ln L_1$所起的作用与线性模型的残差平方和类似，其可能的最小值也是 0。

偏离度可视为线性模型的残差平方和概念之推广，故有时也译为“偏差平方和”。

进一步，对于仅包含常数项的**零模型**(null model)，可定义其相应的**零偏离度**(null deviance)：

$$\text{零偏离度} \equiv -2\ln L_0 \qquad (5.26)$$

其中，$\ln L_0$为以常数项为唯一变量的对数似然函数之最大值。在零模型中加入其他变量之后，其偏离度的改进程度，则反映了这些变量对于y的解释能力：

$$\text{零偏离度} - \text{偏离度} = 2(\ln L_1 - \ln L_0) \tag{5.27}$$

$2(\ln L_1 - \ln L_0)$正是检验原假设“除常数项外所有回归系数均为 0”的似然比检验之统计量。根据残差偏离度与零偏离度，容易计算准R^2：

$$\text{准}R^2 = \frac{\text{零偏离度} - \text{偏离度}}{\text{零偏离度}} = \frac{\ln L_0 - \ln L_1}{\ln L_0} \tag{5.28}$$

5.5 Logit 模型的预测

得到 Logit 模型的估计系数后，即可预测“$y_i = 1$”的条件概率：

$$\hat{p}_i \equiv \widehat{\mathrm{P}(y_i = 1 \mid \boldsymbol{x}_i)} = \Lambda(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}) \equiv \frac{\exp(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})}{1+\exp(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})} \tag{5.29}$$

如果预测概率 $\hat{p}_i > 0.5$，则可预测 $\hat{y}_i = 1$。反之，如果 $\hat{p}_i < 0.5$，则可预测 $\hat{y}_i = 0$。

如果 $\hat{p}_i = 1 - \hat{p}_i = 0.5$，则可预测 $\hat{y}_i = 0$或1。

对于二分类问题，在特征空间(feature space)中，所有满足“$\hat{p}_i = 0.5$”的样本点之集合称为**决策边界**(decision boundary)：

$$\text{决策边界} \equiv \left\{ \mathbf{x}_i : \hat{p}_i(\mathbf{x}_i) = 0.5 \right\} \tag{5.30}$$

在决策边界，可以无差别地预测$\hat{y}_i = 0$或 1。

对于 Logit 模型，也可使用对数几率来预测其响应变量的类别：

$$\ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}} \tag{5.31}$$

如果对数几率$\ln\left(\dfrac{\hat{p}_i}{1-\hat{p}_i}\right)>0$，则可预测$\hat{y}_i=1$。

反之，如果$\ln\left(\dfrac{\hat{p}_i}{1-\hat{p}_i}\right)<0$，则可预测$\hat{y}_i=0$。

若$\ln\left(\dfrac{\hat{p}_i}{1-\hat{p}_i}\right)=0$，则观测值落在决策边界上，可预测$\hat{y}_i=0$或 1。

从方程(5.31)可知，对于 Logit 模型，其决策边界为线性函数，因为

$$\ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}} = 0 \qquad (5.32)$$

此时，决策边界可写为$\left\{\mathbf{x}_i : \mathbf{x}_i'\hat{\boldsymbol{\beta}} = 0\right\}$。

线性的决策边界将特征空间分割为两部分。其中，在决策边界的一侧$\left\{\mathbf{x}_i : \mathbf{x}_i'\hat{\boldsymbol{\beta}} > 0\right\}$，可预测 $y=1$。反之，在决策边界的另一侧$\left\{\mathbf{x}_i : \mathbf{x}_i'\hat{\boldsymbol{\beta}} < 0\right\}$，则预测$y = 0$。

5.6 二分类模型的评估

对于监督学习问题，一般用预测效果来评估其模型的性能。

具体到分类问题，一个常用指标为**准确率**(accuracy)，也称为“正确预测的百分比”(percent correctly predicted)。

只要将样本数据的预测值$\hat{y}_i$与实际值y_i进行比较，即可计算正确预测的百分比：

$$\text{准确率} \equiv \frac{\sum_{i=1}^{n} I(\hat{y}_i = y_i)}{n} \quad (5.33)$$

其中，$I(\cdot)$为**示性函数**(indicator function)。

有时我们更专注于错误预测的百分比，即**错误率**(error rate)或**错分率**(misclassification rate)：

$$
\text{错分率} \equiv \frac{\sum_{i=1}^{n} I(\hat{y}_i \neq y_i)}{n} \tag{5.34}
$$

其中，求和式$\sum_{i=1}^{n} I(\hat{y}_i \neq y_i)$为样本中错误预测的样例数。

如果所考虑样本为训练集，则为“训练误差”(training error)。如果所考虑样本为测试集，则为“测试误差”(test error)。准确率与错误率之和为 1。

准确率或错分率并不适用于“类别不平衡”(class imbalance)的数据。

例如，假设某种罕见病的发病率仅为百分之一。此时，样本中的两个类别高度不平衡。即使不用任何机器学习的算法，只要一直预测不发病，也能达到 99%的准确率(或 1%的错分率)。

我们更希望算法能够准确地预测那些发病的个体，即所谓“正例”(positive cases，简称 positives)。

为此，根据模型预测的正例(也称“阳性”)与反例(也称“阴性”)，以及实际观测的正例与反例，可将样本数据分为以下四类，并用一个矩阵来表示，即所谓**混淆矩阵**(confusion matrix)，参见表 5.1：

表 5.1 分类结果的混淆矩阵

		实 际 观 测 值	
		正例(positives)	反例(negatives)
预测值	正例 (positives)	**真阳性** (True Positive, **TP**) $(\hat{y}=1, y=1)$	**假阳性** (False Positive, **FP**) $(\hat{y}=1, y=0)$
	反例 (negatives)	**假阴性** (False Negative, **FN**) $(\hat{y}=0, y=1)$	**真阴性** (True Negative, **TN**) $(\hat{y}=0, y=0)$

混淆矩阵的左上角为**真阳性**或“真正例”(True Positive)，简记 TP，即预测正例($\hat{y}=1$)，而实际也是正例($y=1$)的情形。

右上角为**假阳性**或“假正例”(False Positive)，简记 FP，类似于假警报(false alarm)，即预测正例($\hat{y}=1$)，但实际为反例($y=0$)的情形。

混淆矩阵的左下角为**假阴性**或“假反例”(False Negative)，简记 FN，即预测反例($\hat{y}=0$)，而实际为正例($y=1$)的情形。

右下角为**真阴性**或“真反例”(True Negative)，简记 TN，即预测反例($\hat{y}=0$)，而实际也是反例($y=0$)的情形。

根据混淆矩阵的信息，可设计更为精细的模型评估指标。

比如，从纵向角度考察混淆矩阵的第 1 列，在实际为正例的子样本中，定义其预测正确的比例为**灵敏度**(sensitivity) 或**真阳率**(true positive rate)：

$$\text{灵敏度}=\text{真阳率} \equiv \frac{\text{TP}}{\text{TP}+\text{FN}} \qquad (5.35)$$

灵敏度也称为“查准率”(precision)，它反映了在实际为正例的子样本中，正确预测的比例；尤其适用于上文关于罕见病的案例。

类似地，考虑混淆矩阵的第 2 列，在实际为反例的子样本中，定义其预测正确的比例为**特异度**(specificity)，也称为**真阴率**(true negative rate)：

$$\text{特异度}=\text{真阴率}\equiv\frac{\text{TN}}{\text{FP}+\text{TN}} \tag{5.36}$$

进一步，“1−特异度”则为在实际为反例的子样本中，错误预测的比例，也称为**假阳率**(false positive rate)：

$$\text{假阳率}=1-\text{特异度}\equiv\frac{\text{FP}}{\text{FP}+\text{TN}} \tag{5.37}$$

也可以从横向角度考察混淆矩阵。

比如，考虑混淆矩阵的第 1 行，在预测为正例的子样本中，定义其预测正确的比例为**查全率或召回率**(recall)：

$$\text{查全率} \equiv \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{5.38}$$

5.7 ROC 与 AUC

迄今为止，我们默认用于分类的“门槛值”(threshold)为$\hat{p} = 0.5$。

事实上，从决策理论的角度，这未必是最佳选择。

从混淆矩阵可知，在作预测时，可能犯两类不同的错误，即“假阳性”与“假阴性”。在具体的业务中，这两类错误的成本可能差别很大。

例 在诊断疾病时(病人=正例)，“假阳性”将健康者误判为病人，其成本可能只是多做些医疗检查；而“假阴性”将病人视为健康者，则会耽误病情，后果更为严重。

例 银行在审批贷款申请时(断供=正例),“假阳性”将正常客户视为劣质客户而拒绝贷款，其成本只是少赚些利润；而“假阴性”将劣质客户视为正常客户而放贷，则会面临因断供而损失本金的巨大成本。

作预测的两类错误，其成本可能并不对称。此时，应根据具体的业务需要，考虑使用合适的门槛值 $\hat{p}=c$ 进行分类。

比如，为了降低错误放贷的损失，银行可将分类为劣质客户的门槛值降低到 $\hat{p}=0.2$。这意味着，如有 20%或以上的概率客户会断供，则判断为断供，并拒绝贷款。

使用更低的门槛值，将预测更多的正例，而预测更少的反例。

此时，在实际为正例的子样本中，预测准确率将上升，即灵敏度上升。而在实际为反例的子样本中，预测准确率将下降，即特异度下降，故“1−特异度”上升。

灵敏度与“1−特异度”均为门槛值 $\hat{p}=c$ 的函数，可记为“sensitivity(c)”与“1−specificity(c)”。

如果将“1−specificity(c)”放于坐标横轴，而把“sensitivity(c)”放于纵轴，然后让门槛值 $\hat{p}=c$ 的取值从 0 连续地变为 1，则可得到一条曲线，即所谓**接收器工作特征曲线**(Receiver Operating Characteristic Curve，简记 ROC 曲线)，参见图 5.5。

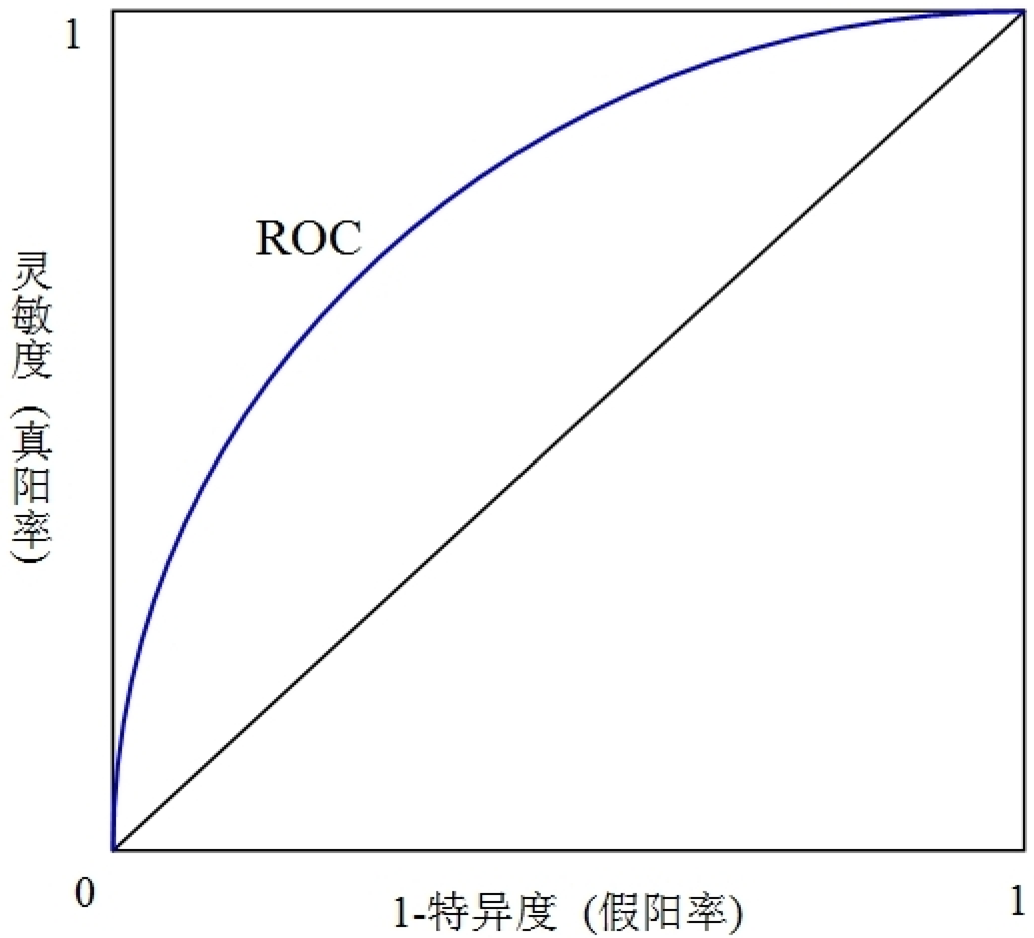

图 5.5　ROC 曲线的示意图

当门槛值低至$\hat{p} = c = 0$时，则“草木皆兵”(宁可错杀一万，不可放过一人)，所有样例都被预测为正例，即$\hat{y} = 1$。此时，混淆矩阵变为表 5.2：

表 5.2 分类门槛值为 0 的混淆矩阵

<table>
<tr><td></td><td colspan="2">实际值</td></tr>
<tr><td rowspan="2">预测值</td><td>TP</td><td>FP</td></tr>
<tr><td>FN = 0</td><td>TN = 0</td></tr>
</table>

因此，当门槛值为 0 时，

$$\text{灵敏度}=\text{真阳率}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}=\frac{\mathrm{TP}}{\mathrm{TP}+0}=1$$

所有实际正例都被预测对，而

$$\text{假阳率}=1-\text{特异度}=\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}=\frac{\mathrm{FP}}{\mathrm{FP}+0}=1$$

所有实际反例都被预测错)，此时 ROC 曲线位于图 5.5 的右上角，坐标为(1, 1)。

反之，如果走向另一极端，将门槛值提高至 $\hat{p}=c=1$，则所有样例都被预测为反例，即 $\hat{y}=0$。此时，混淆矩阵变为表 5.3：

表 5.3 分类门槛值为 1 的混淆矩阵

	实际值	
预测值	TP = 0	FP = 0
	FN	TN

因此，当门槛值为 1 时，

$$\text{灵敏度}=\text{真阳率}=\frac{\text{TP}}{\text{TP}+\text{FN}}=\frac{0}{0+\text{FN}}=0$$

所有实际正例都被预测错)，而

$$\text{假阳率}=1-\text{特异度}=\frac{\text{FP}}{\text{FP}+\text{TN}}=\frac{0}{0+\text{TN}}=0$$

所有实际反例都被预测对，此时 ROC 曲线位于图 5.5 的左下角(即原点)，坐标为(0, 0)。这是一种“老好人”的做法，但无法捕捉真正的正例。

当门槛值取值为$0 \le c \le 1$时，则可得到整条 ROC 曲线。

由于纵轴为实际正例中的准确率(灵敏度)，而横轴为实际反例中的错误率(1-特异度)，故我们希望模型的 ROC 曲线越靠近左上角越好。

因此，为衡量 ROC 曲线的优良程度，可使用 ROC **曲线下面积**(Area Under the Curve，简记 AUC)来度量，参见图 5.6。

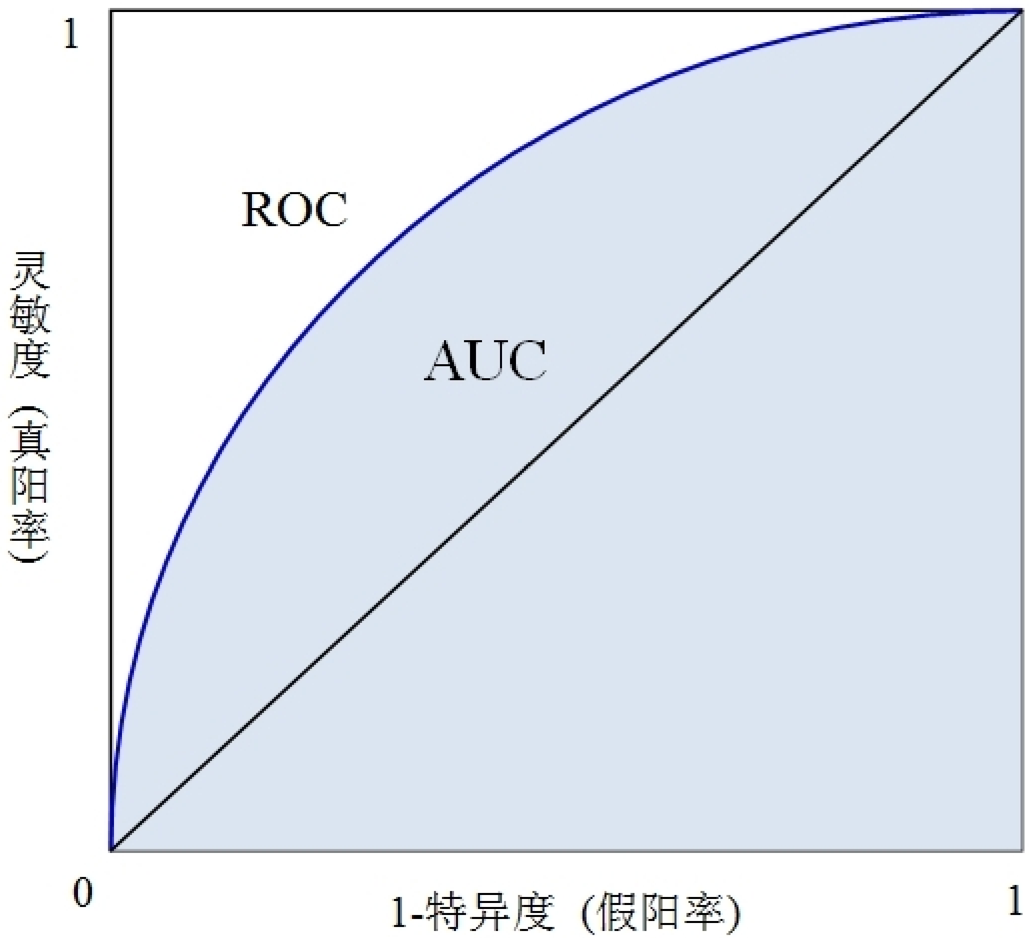

图 5.6　AUC 的示意图

如果 AUC 为 1，则意味着模型对于所有正例与反例的预测都是正确的，这一般是无法达到的理想状态。

如果 ROC 曲线与 45 度线(从原点到(1, 1)的对角线)重合，则意味着该模型的预测结果无异于随机猜测。比如，样本中正例与负例各占一半，而通过从$[0,1]$区间的均匀分布随机抽样来预测概率。此时，AUC 为 0.5。

AUC 小于 0.5 的情形十分罕见，这意味着模型的预测结果还不如随机猜测。

对于二分类问题，在比较不同模型的预测效果时，常使用 AUC。由于 AUC 为衡量预测效果的综合性指标，可使用此单一指标比较不同的算法。

5.8 科恩的 kappa

对于类别不平衡的样本，还可使用“kappa”或“科恩 kappa”(Cohen's kappa)(Cohen, 1960)，以去掉准确率指标中可能的虚高“水分”。

回到表 5.1 的混淆矩阵，并将表 5.1 中每个单元的频数转化为频率(频数占全样本的比重)，即所谓“列联表”(contingency table)，参见表 5.4。

表 5.4 分类结果的列联表

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="2">实 际 观 测 值</td><td></td></tr>
<tr><td>正例(positives)</td><td>反例(negatives)</td><td>合计</td></tr>
<tr><td rowspan="2">预测值</td><td>正例(positives)</td><td>p_{11}
真阳性(TP)的比重</td><td>p_{12}
假阳性(FP)的比重</td><td>$p_{1\cdot}$</td></tr>
<tr><td>反例(negatives)</td><td>p_{21}
假阴性(FN)的比重</td><td>p_{22}
真阴性(TN)的比重</td><td>$p_{2\cdot}$</td></tr>
<tr><td></td><td>合计</td><td>$p_{\cdot 1}$</td><td>$p_{\cdot 2}$</td><td>1</td></tr>
</table>

显然，$p_{11}+p_{12}+p_{21}+p_{22}=1$。

定义 $p_{\cdot 1} \equiv p_{11}+p_{21}$，即样本中实际正例的比重；而 $p_{\cdot 2} \equiv p_{12}+p_{22}$，即样本中实际负例的比重。

定义 $p_{1\cdot} \equiv p_{11}+p_{12}$，即样本中预测正例的比重；而 $p_{2\cdot} \equiv p_{21}+p_{22}$，即样本中预测负例的比重。

假设预测值与实际值分别来自两个不同的“评分者”(raters)。我们希望评估这两个评分者之评分结果(rating)的一致性(agreement)。、

首先，记“观测到的一致性”(Observed Agreement)，也就是准确率为

$$\mathrm{P}_O = p_{11} + p_{22} \qquad (5.39)$$

当然，两个评分者之间的观测一致性P_O可能只是因为随机因素所致。

另一方面，如果两个评分者的打分完全相互独立，则二者“期望的一致性”(Expected Agreement)为

$$\mathrm{P}_E = p_{\cdot 1}p_{1\cdot} + p_{\cdot 2}p_{2\cdot} \qquad (5.40)$$

期望一致性P_E度量在评分者相互独立的情形，二者打分碰巧一致的概率；即一致打分为正例($p_{\cdot 1}p_{1\cdot}$)与一致打分为反例($p_{\cdot 2}p_{2\cdot}$)的概率之和。

考察观测一致性P_O相对于期望一致性P_E的改进,可定义科恩的 kappa：

$$\text{kappa} \equiv \frac{\mathrm{P}_O - \mathrm{P}_E}{1 - \mathrm{P}_E} \qquad (5.41)$$

其中，分子$(\mathrm{P}_O - \mathrm{P}_E)$为从随机一致性到观测一致性的实际改进，而分母$(1 - \mathrm{P}_E)$为从随机一致性到完全一致的最大可能改进。

在上式中，分母起着一种标准化的作用。

由于$\mathrm{P}_O \le 1$，故$\text{kappa} \le 1$。

一般来说，$\text{kappa} \geq 0$。

但$\text{kappa} < 0$的情形依然可能出现，这意味着观测一致性还不如随机猜测的一致性，在实践中比较罕见。

究竟kappa多大才表示模型的预测性能好，这是一个主观判断。

关于kappa含义的解释，表 5.5 提供一个较为常用的指南(guideline)：

表 5.5 kappa 含义的解释

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

与仅适用于二分类的 AUC 指标相比，科恩 kappa 可用于多分类问题。

5.9 逻辑回归的 Python 案例

本节以 `titanic` 数据为例，演示逻辑回归的 Python 操作。

该数据包括泰坦尼克号乘客的存活数据。

泰坦尼克号邮轮是当时最大的客运轮船，在从英国南安普顿开往美国纽约的处女航中于 1912 年 4 月 14 日撞冰山沉没。泰坦尼克号海难是和平时期死亡人数最多的海难之一，也是最广为人知的海难之一(1997 年同名好莱坞电影热映)。

* 详见教材，以及配套 Python 程序（现场演示）。