数据科学导论第 9 讲——分类方法

王小宁

中国传媒大学数据科学与智能媒体学院

2025 年 05 月 08 日

中国传媒大学
COMMUNICATION UNIVERSITY OF CHINA

目录

问题的提出

分类

- 当因变量取离散值时，我们称之为分类。比如我们在信用卡违约预测的时候，我们的因变量 y 取值是，这是一个二元（binary）的取值。模仿回归的表达式，我们可以将分类问题写成

$$y = f(x)$$

- 其中，f 是关于 x 的函数，往往被称为分类器（classifier），比如 Logistic 模型、决策树、随机森林和支持向量机等都是经典的分类器。

分类问题

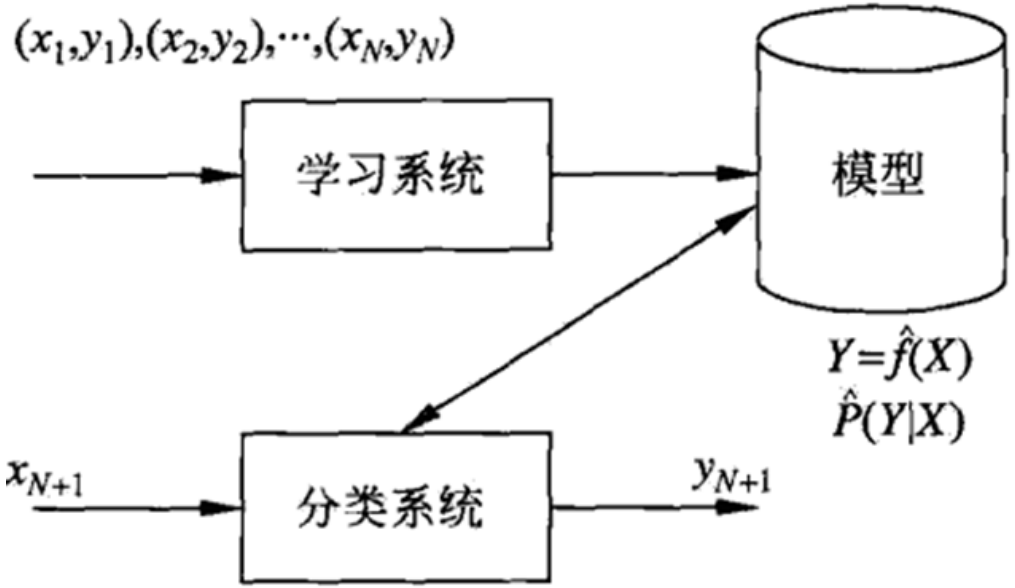


图 1: 分类问题的结构

Probit 与 Logistic 模型

线性概率模型（Linear Probability Model, LPM）

- 考虑二元选择模型：

$$Y_i = X_i\beta + \varepsilon_i, i = 1, 2, \cdots, N$$

- 其中 X_i 包含常数项的 k 元设计矩阵,Y_i 是二元取值的因变量：$Y_i = 1$ 表示某一事件发生，$Y_i = 0$ 表示某一事件不发生。
- 如果假定 $E(\varepsilon_i|X_i) = 0$, 则总体的回归方程为：$E(Y_i|X_i) = X_i\beta$

- 假定在给定 X_i 的时候，某一事件发生的概率为 p, 不发生的概率为 $1-p$, 即 $Prob(Y_i=1|X_i)=p, Prob(Y_i=0|X_i)=1-p$, 又因为 Y_i 只取 1 和 0 两个值，所以其条件期望为：
  $E(Y_i|X_i)=1*Prob(Y_i=1|X_i)+0*Prob(Y_i=0|X_i)=p$
- 因此可得：$E(Y_i|X_i)=X_i\beta=Prob(Y_i|X_i)=p$

- 当给定自变量 X_i 的时候，某一事件发生（即取值为 1）的平均概率。在上式中，这一概率体现为线性的形式 $X_i\beta$，因此:

$$Y_i = X_i\beta + \varepsilon_i$$

- 称为线性概率模型（Linear Probability Model, LPM）。这实际上就是用普通的线性回归方法对二元取值的因变量直接建模。

线性概率模型

- 对于线性概率模型，我们也可以采用普通的最小二乘法进行估计，但是会存在如下三个问题：

1. 我们对线性概率模型进行的拟合，实际上是对某一事件发生的平均概率的预测，即

$\hat{Y}_i = \hat{Prob}(Y_i|X_i) = X_i\hat{\beta}$。但是，这里的 $X_i\beta$ 值并不能保证在 0 和 1 之间，完全有可能出现大于 1 或小于 0 的情形。

2. 由于 Y 是二元变量，因此扰动项 $\varepsilon_i = Y_i - X_i\beta$ 也应该是二元变量，它应该服从二项分布，而不是我们通常假定的正态分布。但是当样本足够多时，二项分布收敛于正态分布。

线性概率模型

3. 在 LPM 中，扰动项的方差是异方差的（与 X_i 有关，非常数）

- 由于存在着上述的诸多问题，因此对于二元定性因变量，一般不推荐使用 LPM，而是需要其他更为科学的方法。

Probit 与 Logistic 模型

- 在 LPM 中，通过适当的假设可以使得 $Y_i = 1$ 的概率 $Prob(Y_i|X_i)$ 与 X_i 是线性关系，即：

$$p(X_i) = Prob(Y_i = 1|X_i) = F(X_i\beta) = X_i\beta$$

- 同时为了保证估计的概率的取值范围能在 [0,1] 区间上，一个直接的想法就是在外面套上一个分布函数 $F(\cdot)$

常用的分布函数

1. 如果分布函数在 $F(X_i\beta)$ 用标准正态分布函数 $\Phi(\cdot)$, 即:

$$p(X_i) = Prob(Y_i = 1|X_i) = \Phi(X_i\beta) = \int_{-\infty}^{X_i\beta} \frac{1}{\sqrt{2\pi}} exp(-\frac{z^2}{2})dz$$

其中 $\Phi(X_i\beta)$ 是正态分布的分布函数，取值范围 [0,1] 这时的概率模型为 Probit 模型。

2. 如果分布函数在 $F(X_i\beta)$ 用 Logistic 分布函数 $\Lambda(\cdot)$, 即:

$$p(X_i) = Prob(Y_i = 1|X_i) = \Lambda(X_i\beta) = \frac{exp(X_i\beta)}{1 + exp(X_i\beta)}$$

特别地对于 Logit 模型来说还可以写成：

$$log(\frac{p(X_i)}{1-p(X_i)}) = X_i\beta = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}$$

- 其中 $log(\frac{p(X_i)}{1-p(X_i)})$ 称为链接函数（Link Function）。
- $\frac{p(X_i)}{1-p(X_i)}$ 称为赔率、优势比 (Odds Ratio)，即 Y_i 取 1 的概率与取 0 的概率比值。

基于潜变量模型的理解

- 二元选择模型也可以从潜变量回归模型的角度去解释，首先考察以下模型：

$$Y_i^* = X_i\beta + \varepsilon_i, i = 1, 2, \cdots, T$$

$$Y_i = \begin{cases} 1, if & Y_i^* > 0 \\ 0, if & Y_i^* \leq 0 \end{cases}$$

- 其中，Y_i^* 是潜变量或隐变量（Latent Variable），它无法获得实际观测值，但是却可以观测到 $Y_i^* > 0$ 还是 $Y_i^* \leq 0$。因此，我们实际上观测到的变量是 Y_i 而不是 Y_i^*。上式称为潜变量响应函数 (Latent Response Function) 或指示函数 (Index Function)

Probit 模型

如果我们假设：

1. $E(\varepsilon|X_i) = 0$
2. ε 是 i.i.d. 的正态分布
3. $rank(X_i) = k$

根据潜变量响应函数可得 Y_i 的概率特征：
$Prob(Y_i = 1|X_i) = \int_{-X_i\beta}^{\infty} f(\varepsilon_i)d\varepsilon_i$

则当 $f(\varepsilon_i)$ 为标准正态分布的概率密度函数：$\phi(\varepsilon_i) = \frac{1}{\sqrt{2\pi}} exp(-\frac{\varepsilon_i^2}{2})$
$Prob(Y_i = 1|X_i)$ 可以写成：

$$Prob(Y_i = 1|X_i) = 1 - \Phi(-X_i\beta) = \Phi(X_i\beta)$$

Logit 模型

如果我们假设：

1. $E(\varepsilon|X_i) = 0$
2. ε 是 i.i.d. 的 Logistic 分布
3. $rank(X_i) = k$

根据潜变量响应函数可得 Y_i 的概率特征：

$$Prob(Y_i = 1|X_i) = Prob(Y_i^* > 0|X_i) = 1 - \int_{-\infty}^{-X_i\beta} f(\varepsilon_i)d\varepsilon_i$$

又因为 $\int_{-\infty}^{X_i\beta} f(\varepsilon_i)d\varepsilon_i = \frac{exp(-X_i\beta)}{1+exp(-X_i\beta)}$，因此：

$Prob(Y_i = 1|X_i)$ 可以写成：

$$Prob(Y_i = 1|X_i) = \frac{exp(-X_i\beta)}{1+exp(-X_i\beta)} = \Lambda(X_i\beta)$$

最大似然估计（MLE）

Probit 和 Logit 模型的参数估计常用最大似然法。对于 Probit 或 Logit 模型：

$$Prob(Y_i = 1|X_i) = F(X_i\beta)$$

$$Prob(Y_i = 0|X_i) = 1 - F(X_i\beta)$$

所以似然函数为：

$$L = \prod_{i=1}^{N} F(X_i\beta)^{Y_i}(1 - F(X_i\beta))^{1-Y_i}$$

对数似然函数为：

$$log(L) = \sum_{i=1}^{N} Y_i * F(X_i\beta) + (1 - Y_i)(1 - F(X_i\beta))^{1-Y_i}$$

最大化这个对数似然函数，对 β 求导令其为 0 不存在显示解或封闭解，所以要用非线性方程的迭代的方法进行求解。常用的有 Newton-Raphson 法或二次爬坡法（Quadratic hill climbing）。

边际效应分析

对于 Probit 模型来说，其边际效应为：

$$\frac{\partial Prob(Y_i = 1|X_i)}{\partial X_i} = \phi(X_i\beta)\beta$$

对于 Logit 模型，其边际效应为：

$$\frac{\partial Prob(Y_i = 1|X_i)}{\partial X_i} = \Lambda(X_i\beta)(1 - \Lambda(X_i\beta))\beta$$

- 在 Probit 和 Logit 模型中，自变量对 Y_i 取值为 1 的概率的边际影响不是常数，会随着自变量取值的变化而变化。所以对于 Probit 和 Logit 模型来说，它们的边际影响不能像线性回归模型直接等于系数。
- 这两个模型的边际效应分析常用的方法，是计算其平均边际效应，即对于非虚拟的自变量，一般是用其样本均值代入到两式，估计平均边际影响。

似然比检验

- 似然比检验类似于检验模型整体显著性的 F 检验，原假设为全部自变量的系数都为 0，检验的统计量 LR 为：

$$LR = 2(lnL - lnL_0)$$

其中 lnL 为对概率模型进行 MLE 估计的对数似然函数值，lnL_0 为只有截距项的模型的对数似然函数值，往往也称为空模型，即模型中不包含任何自变量。当原假设成立时，LR 的渐近分布是自由度为 $k-1$（即除截距项外的自变量的个数）的 χ^2 分布。

预测

- 如果我们得到了系数的估计值 $\hat{\beta}$, 我们就可以预测出在给定 X_0 下，$P(X_0) = Prob(Y_i = 1|X_0)$ 的概率预测值，即：

Probit:

$$\hat{P}(X_0) = \hat{Prob}(Y_i = 1|X_0) = \Phi(X_0\beta) = \int_{-\infty}^{X_0\hat{\beta}} \frac{1}{\sqrt{2\pi}} exp(-\frac{z^2}{2})dz$$

Logit:

$$\hat{P}(X_0) = \hat{Prob}(Y_i = 1|X_0) = \Lambda(X_0\beta) = \frac{exp(X_0\hat{\beta})}{1 + exp(X_0\hat{\beta})}$$

也可以直接求链接函数的值：

$$\hat{Link} = log(\frac{\hat{p}(X_0)}{1 - \hat{p}(X_0)}) = X_0\hat{\beta} = \beta_0 + \hat{\beta}_1 X_{10} + \cdots + \hat{\beta}_p X_{p0}$$

判别分析

概述

- 对于二元因变量，Logistic 模型直接对 $Pr(Y = k|X = x)$ 进行建模，即在给定自变量 X 下建立因变量 Y 的条件分布模型。相反，判别分析采取的方法是先对每一个给定的 Y 建立自变量 X 的分布，然后使用贝叶斯定理反过来再去估计 $pr(Y = k|X = x)$
- 何时使用判别分析？一方面，当类别的区分度较高的时候，或者当样本量 n 较小且自变量 X 近似服从正态分布时，Logistic 模型的参数估计会相对不够稳定，而判别分析就不存在这样的问题；另一方面，在现实生活中，有很多因变量取值超过两类的情形，虽然我们可以把二元 Logistic 模型推广到多元的情况，但这在实际应用中并不常用。实际中对于因变量取多类别的问题，我们更常使用的是判别分析法。
- 主要介绍三种常用方法，包括朴素贝叶斯判别分析、线性判别分析和二次判别分析。

Naive Bayes 判别分析

对于分类模型，我们的目的是构建从输入空间（自变量空间）到输出空间（因变量空间）的映射（函数）：

$$f(X) \longrightarrow Y$$

- 它将输入空间划分成几个区域，每个区域对应一个类别。区域的边界可以是各种函数形式，其中最重要且最常用的一类就是线性的。
- 为了确定边界函数，在构造分类器时，我们最关注的便是一组测试观测值上的测试错误率，在一组测试观测值上的误差计算具有以下形式：

$$Ave(I(y_0 \neq \hat{y}_0))$$

一个非常简单的分类器是将每个观测值分到它最大可能所在的类别中，即给定 $X = x_0$ 的情况下，将它分到条件概率最大的 j 类中是比较合理的：

$$max_j Pr(Y = j|X = x_0)$$

这类方法称为贝叶斯分类器，这种分类器将产生最低的测试错误率，称为贝叶斯错误率，在 $X = x_0$ 这一点的错误率为 $1 - max_j Pr(Y = j|X = x_0)$ 整个分类器的贝叶斯错误率为：

$$1 - E(max_j Pr(Y = j|X))$$

线性判别分析（linear discriminant analysis, LDA）

- 我们在进行分类时，首先要获取 $f_k(x)$ 的估计，然后代入第 k 类的后验分布估计 $p_k(x)$, 并根据 $p_k(x)$ 的值，将观测分到值最大的一类中。为了获取 $f_k(x)$ 的估计，首先对其做一些假设。
- 通常假设 $f_k(x)$ 的分布是正态的。当 $p = 1$ 时，密度函数为一维正态密度函数，其次是各类方差相同，均为 σ^2
- 线性分类器是将观测值 $X = x$ 分到上式中 $p_k(x)$ 最大的一类。
- 判别函数 $\hat{\delta}_k$ 是关于 x 的线性函数，所以称该方法为线性判别分析。

多元线性判别分析

若自变量维度 $P > 1$, 假设 $X = (X_1, X_2, \cdots, X_p)$ 服从一个均值不同、协方差矩阵相同的多元正态分布，即假设第 k 类观测服从一个多元正态分布 $N(\mu_k, \Sigma)$, 其中 μ_k 是一个均值向量，Σ 为所有 K 类共同的协方差矩阵，其密度函数形式为:

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\sum|^{1/2}} exp(-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k))$$

通过类似于一维自变量的方法，我们可以知道贝叶斯分类器将 $X = x$ 分入 $\delta_k(x)$ 最大的一类。

二次判别分析 (quadratic discriminant analysis, QDA)

- LDA 假设每一类观测服从协方差矩阵相同的多元正态分布，但现实中可能很难满足这样的假设。
- 二次判别分析放松了这一假设，虽然 QDA 分类器也假设每一类观测服从一个正态分布，并把参数估计代入贝叶斯定理进行预测，但 QDA 假设每一类观测有自己的协方差矩阵，即假设第 k 类的观测服从分布为 $X \sim N(\mu_k, \Sigma_k)$, 此时，二次判别函数为：

QDA

$$\delta_k(x) = -\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)+\log\pi_k$$
$$= -\frac{1}{2}x^T\Sigma_k^{-1}x + x^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k + \log\pi_k$$

QDA分类器把μ_k、Σ_k、π_k的估计值代入上式然后将观测分入使$\hat{\delta}_k(x)$值最大的一类。我们发现判别函数$\hat{\delta}_k(x)$是关于x的二次函数，类别k和l的决策边界也是一条曲线边界，这也是二次判别分析名字的由来。

图 2: QDA

分类问题评判准则

混淆矩阵

混淆矩阵

		预测分类		
		− 或零	+ 或非零	总计
真实分类	− 或零	真阴性值（TN）	假阳性值（FP）	**N**
	+ 或非零	假阴性值（FN）	真阳性值（TP）	**P**
	总计	**N***	**P***	

于是，模型整体的正确率可表示为 $\text{accuracy} = (\text{TN} + \text{TP})/(\text{N} + \text{P})$

相应的，整体错误率即为 $1 - \text{accuracy}$ 。

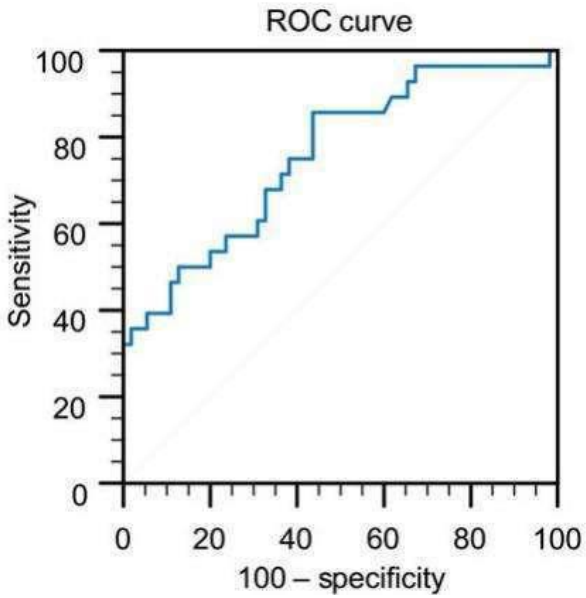
其他指标

分类和诊断测试中重要的评价指标

名称	定义	相同含义名称
假阳性率	FP/N	第**I**类错误，1−特异度（1-specificity）
真阳性率	TP/P	1−第**II**类错误，灵敏度（sensitivity），召回率（recall）
预测阳性率	TP/P*	精确度，1−假阳性率
预测阴性率	TN/N*	

ROC 曲线

- ROC(接受者操作特性曲线) 主要是用于 X 对 Y 的预测准确率情况。
- 曲线下方部分的面积被称为 AUC（Area Under Curve），用来表示预测准确性，AUC 值越高，也就是曲线下方面积越大，说明预测准确率越高。曲线越接近左上角（X 越小，Y 越大），预测准确率越高。



R 实现

Logistic

- R 中可以用 glm() 函数拟合广义线性模型，包含 Logistic 回归中的 probit 模型和 logit 模型。
- 函数的基本形式为：

glm (formula , family = family (link = function) , data =)

- 其中，formula 是模型表达式，与 lm() 的表达式一致。family 参数用于设置模型的连接函数对应的分布族，比如 gaussian 分布，Poisson 分布等.

LDA 和 QDA

- R 中的 MASS 包提供了 lda() 和 qda() 函数分别做线性判别分析和二次判别分析，e1071 包提供了 naiveBayes() 函数做朴素贝叶斯分类。
- R 中的 ROCR 包可以用于生成 ROC 曲线。

分类

- 当因变量取离散值时，我们称之为分类。模仿回归的表达式，我们可以将分类问题写成

$$y = f(x)$$

- 其中，f 是关于 x 的函数，往往被称为分类器（classifier），比如 Logistic 模型、决策树、随机森林和支持向量机等都是经典的分类器。

本周推荐

1. 一本书：《传染-塑造消费、心智、决策的隐秘力量》，乔纳·伯杰，电子工业出版社，2017.8《疯传》姊妹篇；《数字化生存》，NIcholas Negroponte，电子工业出版社，2017（写于 1996,MIT Media Lab）
2. 2 部电影：《统计的乐趣（The joy of stats）》，BBC；《大空头（The Big Short）》，2015；利益风暴（商海通牒）
3. 练习：《R 语言实战（第 2 版）》，《R 语言实战（第 2 版）》，第 13 章代码和第 17 章代码实现 (11 或 12)，2025.5.28

谢　谢!