中级应用统计学

判别分析

张晨峰

华东理工大学商学院

2016年11月22日

5 判别分析

主要内容

- 简介
- 总体的分离和分类
- 费希尔判别

5.1 简介

判别分析

判别分析是在类（组）别先验已知的情况下的应用。判别分析的目的在于把一个或几个观测值分配到这些已知的类（组）别中。

判别与分类

判别分析本质上是一种探索性的分割方法，分类方法导出一些明确定义的法则，可用于分配新的对象。

判别与回归

当被解释变量是非度量变量时，一般的多元回归不适合解决此类问题，而判别分析适用于此类情景。

5.2 总体的分离和分类

两个总体的分类

全部可能的样本结果的集合被分成R_1和R_2两个区域，要是某个新观测值落入R_1，就将它分配到总体π_1，若落入R_2，就将它分配到总体π_2。

5.2 总体的分离和分类

Table 11.1

π_1: Riding-mower owners		π_2: Nonowners	
x_1 (Income in \$1000s)	x_2 (Lot size in 1000 ft^2)	x_1 (Income in \$1000s)	x_2 (Lot size in 1000 ft^2)
90.0	18.4	105.0	19.6
115.5	16.8	82.8	20.8
94.8	21.6	94.8	17.2
91.5	20.8	73.2	20.4
117.0	23.6	114.0	17.6
140.1	19.2	79.2	17.6
138.0	17.6	89.4	16.0
112.8	22.4	96.0	18.4
99.0	20.0	77.4	16.4
123.0	20.8	63.0	18.8
81.0	22.0	81.0	14.0
111.0	20.0	93.0	14.8

5.2 总体的分离和分类

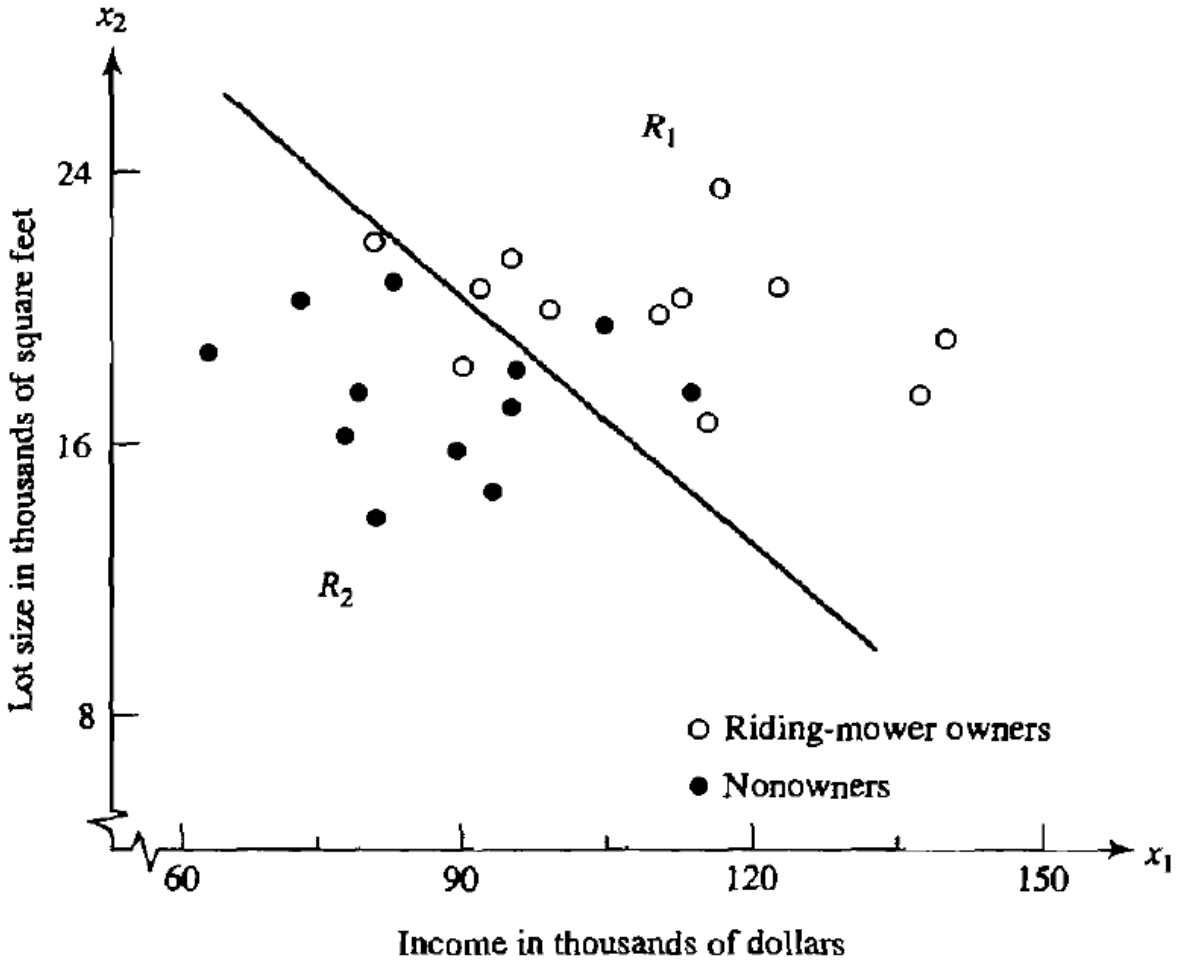

Figure 11.1 Income and lot size for riding-mower owners and nonowners.

5.2 总体的分离和分类

两个总体的期望（平均）错分代价

假设两个总体π_1和π_2，p_1和p_2分别为π_1和π_2的先验概率，将π_1中的对象错分到π_2的条件概率为$P(2|1)$，即

$$P(2|1) = P(X \in R_2|\pi_1) = \int_{R_2=\Omega-R_2} f_1(x)dx$$

类似地，将π_2中的对象错分到π_1的条件概率为$P(1|2)$，即，

$$P(1|2) = P(X \in R_1|\pi_2) = \int_{R_1} f_2(x)dx$$

则$P_{11} = P(1|1)p_1$，$P_{12} = P(1|2)p_2$，$P_{21} = P(2|1)p_1$，$P_{22} = P(2|2)p_2$。

5.2 总体的分离和分类

		Classify as:	
		π_1	π_2
True population:	π_1	0	$c(2\|1)$
	π_2	$c(1\|2)$	0

两个总体的期望（平均）错分代价

$$ECM = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2$$

5.2 总体的分离和分类

最小ECM法则

使ECM达到最小的区域R_1和$_2$由满足以下不等式的x值所定义：

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq (\frac{c(1|2)}{c(2|1)})(\frac{p_2}{p_1})$$

$$R_2 : \frac{f_1(x)}{f_2(x)} < (\frac{c(1|2)}{c(2|1)})(\frac{p_2}{p_1})$$

5.2 总体的分离和分类

Example 11.2 (Classifying a new observation into one of the two populations) A researcher has enough data available to estimate the density functions $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ associated with populations π_1 and π_2, respectively. Suppose $c(2|1) = 5$ units and $c(1|2) = 10$ units. In addition, it is known that about 20% of *all* objects (for which the measurements $\mathbf{x}$ can be recorded) belong to π_2. Thus, the prior probabilities are $p_1 = .8$ and $p_2 = .2$.

Suppose the density functions evaluated at a new observation $\mathbf{x}_0$ give $f_1(\mathbf{x}_0) = .3$ and $f_2(\mathbf{x}_0) = .4$. Do we classify the new observation as π_1 or π_2? To answer the

5.2 总体的分离和分类

两个多元正态总体的分离

假设协方差矩阵相等的情况，则X对总体π_1和π_2的联合密度为

$$f_i(x) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} exp[-\frac{1}{2}(x-\mu_i)'\Sigma^{-1}(x-\mu_i)]$$

最小ECM区域变成

$$R_1 : exp[-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2)] \geq (\frac{c(1|2)}{c(2|1)})(\frac{p_2}{p_1})$$

$$R_2 : exp[-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2)] < (\frac{c(1|2)}{c(2|1)})(\frac{p_2}{p_1})$$

5.2 总体的分离和分类

两个正态总体的估计的最小ECM法则

若

$$(\bar{x}_1 - \bar{x}_2)' S_p^{-1} x_0 - \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S_p^{-1}(\bar{x}_1 + \bar{x}_2) \geq ln[(\frac{c(1|2)}{c(2|1)})(\frac{p_2}{p_1})]$$

则将x_0分配给π_1，否则将x_0分配给π_2。

5.2 总体的分离和分类

The investigators (see [4]) provide the information

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} -.0065 \\ -.0390 \end{bmatrix}, \qquad \bar{\mathbf{x}}_2 = \begin{bmatrix} -.2483 \\ .0262 \end{bmatrix}$$

and

$$\mathbf{S}_{\text{pooled}}^{-1} = \begin{bmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix}$$

Therefore, the equal costs and equal priors discriminant function

Measurements of AHF activity and AHF-like antigen on a woman who may be a hemophilia A carrier give $x_1 = -.210$ and $x_2 = -.044$. Should this woman be classified as π_1 (normal) or π_2 (obligatory carrier)?

5.2 总体的分离和分类

多个总体的期望（平均）错分代价

$$ECM = \sum_{i=1}^{g} p_i \sum_{k=1, k\neq i}^{g} P(k|i)c(k|i)$$

5.2 总体的分离和分类

多个总体的最小ECM法则

使ECM达到极小的分类域，可通过将x分配给使

$$\sum_{i=1,i\neq k}^{g} p_i f_i(x) c(k|i)$$

最小的总体$\pi_k(k = 1, ..., g)$来定义。当最小的总体不止一个时，则将x指派给其中的任一总体即可。

最小ECM法则的三要素

- 先验概率
- 错分代价
- 密度函数

5.2 总体的分离和分类

多个总体错分代价相同时的最小ECM法则

若

$$p_k f_k(x) > p_i f_i(x), i \neq k$$

则将x分配到π_k，或等价的，若

$$lnp_k f_k(x) > lnp_i f_i(x), i \neq k$$

则将x分配到π_k。

5.2 总体的分离和分类

		True population		
		π_1	π_2	π_3
Classify as:	π_1	$c(1\|1) = 0$	$c(1\|2) = 500$	$c(1\|3) = 100$
	π_2	$c(2\|1) = 10$	$c(2\|2) = 0$	$c(2\|3) = 50$
	π_3	$c(3\|1) = 50$	$c(3\|2) = 200$	$c(3\|3) = 0$
Prior probabilities:		$p_1 = .05$	$p_2 = .60$	$p_3 = .35$
Densities at $\mathbf{x}_0$:		$f_1(\mathbf{x}_0) = .01$	$f_2(\mathbf{x}_0) = .85$	$f_3(\mathbf{x}_0) = 2$

5.2 总体的分离和分类

多个正态总体情况下的一般判别分类

若

$$lnp_k f_k(x) = lnp_k - (\frac{p}{2})ln(2\pi) - \frac{1}{2}ln|\Sigma_k| - \frac{1}{2}(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k) = max_i lnp_i f_i(x)$$

则将x分到π_k

二次判别得分

$$d_i^Q(x) = \frac{1}{2}ln|\Sigma_i| - \frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i) + lnp_i$$

5.2 总体的分离和分类

多个正态总体情况下的一般判别分类

若二次判别得分

$$d_k^Q(x) = max\{d_1^Q(x), d_2^Q(x), ..., d_g^Q(x)\}$$

则将x分到π_k。

协方差阵相同时的线性判别得分

$$d_i(x) = \mu_i'\Sigma^{-1}x - \frac{1}{2}\mu_i'\Sigma^{-1}\mu_i + lnp_i$$

样本线性判别得分

$$d_i(x) = \bar{x}_i' S_p^{-1}x - \frac{1}{2}\bar{x}_i' S_p^{-1}\bar{x}_i + lnp_i$$

5.2 总体的分离和分类

协方差相同时等价的判别函数

定义x到样本均值向量$\bar{x}_i$的平方距离

$$D_i^2(x) = (x - \bar{x}_i)' S_p^{-1} (x - \bar{x}_i)$$

于是判别法则为，若

$$-\frac{1}{2} D_i^2(x) + lnp_i$$

最大，则将x分到总体π_i

5.4 费希尔判别

费希尔的思想

费希尔的想法是将多元观测值x变换成一元观测值y，使得由总体π_1和π_2导出的y尽可能地分离。

投影

假设g个类（组），每类（组）π_i有N_i个p维的样本点，令α为R^p中的任一向量，则$y = \alpha' x$为x向以α为法线方向的投影。

5.4 费希尔判别

组内平方和

组内平方和为$\alpha' W \alpha$，其中

$$W = \sum_{i=1}^{g} (N_i - 1) S_i = \sum_{i=1}^{g} \sum_{j=1}^{N_i} (x_{i,j} - \bar{x}_i)(x_{i,j} - \bar{x}_i)'$$

组间平方和

组内平方和为$\alpha' B \alpha$，其中

$$B = \sum_{i=1}^{g} (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})'$$

5.4 费希尔判别

求解

最大化下式

$$\frac{\alpha' B \alpha}{\alpha' W \alpha}$$

即得到α的解

定理

若A和B都是对称阵且$B > 0$，则有

$$max \quad \frac{x'Ax}{x'Bx} = \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p = min \quad \frac{x'Ax}{x'Bx}$$

其中$\lambda_1, ..., \lambda_p$为$B^{-1}A$的特征值，使$\frac{x'Ax}{x'Bx}$最大（最小）的向量$B^{-1}A$是对应最大（最小）特征值的特征向量。

5.4 费希尔判别

费希尔线性判别函数的解

费希尔准则下的线性判别函数$y = \alpha' x$的解α是矩阵$W^{-1}B$的最大特征根所对应的特征向量。

用费希尔判别量将对象进行分类

若

$$\sum_{j=1}^{r}(\hat{y}_i - \bar{y}_{ki})^2 = \sum_{j=1}^{r}[\hat{\alpha}_j'(x - \bar{x}_k)]^2 \leq \sum_{j=1}^{r}[\hat{\alpha}_j'(x - \bar{x}_i)]^2$$

则将x分入π_k。