第六章 判别分析

6.3 Bayes判别

Bayes 判别和 Bayes 估计的思想方法是一样的，即假定对研究的对象已经有一定的认识，这种认识常用先验概率来描述，当我们取得一个样本后，就可以用样本来修正已有的先验概率分布，得出后验概率分布，再通过后验概率分布进行各种统计推断。

第六章 判别分析

1.误判概率与误判损失

设有两个总体X_1和X_2，根据某一个判别规则，将实际上为X_1的个体判为X_2或者将实际上为X_2的个体判为X_1的概率就是误判概率，一个好的判别规则应该使误判概率最小。

除此之外还有一个误判损失问题或者说误判产生的花费（cost）问题。如，把X_1的个体误判到X_2的损失比X_2的个体误判到X_1严重得多，则人们在作前一种判断时就要特别谨慎。

譬如在药品检验中把有毒的样品判为无毒后果比无毒样品判为有毒严重得多，因此一个好的判别规则还必须使误判损失最小。

第六章 判别分析

仍考虑两个总体的情况，两个总体X_1与X_2分别具有密度函数$f_1(x)$与$f_2(x)$，其中x为p维向量。

记Ω为x的所有待判定样本的全体，称样本空间；R_1为根据我们的规则要判为X_1的那些x的全体；$R_2=\Omega-R_1$是要判为X_2的那些x的全体；

显然R_1与R_2互斥完备。某样本x实际是来自X_1，但被判为X_2的概率为

$$P(2|1)=P(x\in R_2\,|\,X_1),$$

来自X_2，但被判为X_1的概率为

$$P(1|2)=P(x\in R_1|X_2).$$

类似地，来自X_1被判为X_1的概率，来自X_2被判为X_2的概率分别为

$$P(1|1)=P(x\in R_1|X_1),$$

$$P(2|2)=P(x\in R_2|X_2).$$

又设p_1, p_2分别表示总体X_1和X_2的先验概率，且

$p_1 + p_2 = 1$，于是

$$P(\text{正确地判为}X_1) = P(\text{来自}X_1, \text{被判为}X_1) =$$
$$P(x \in R_1 \mid X_1) \times P(X_1) = P(1|1) \times p_1,$$
$$P(\text{误判到}X_1) = P(\text{来自}X_2, \text{被判为}X_1) =$$
$$P(x \in R_1 \mid X_2) \cdot P(X_2) = P(1|2) \cdot p_2.$$

类似地有

$$P(\text{正确地判为}X_2) = P(2|2) \cdot p_2,$$

$$P(\text{误判到}X_2) = P(2|1) \cdot p_1.$$

设$L(1|2)$表示来自X_2误判为X_1引起的损失，$L(2|1)$表示来自X_1误判为X_2引起的损失，并规定$L(1|1)=L(2|2)=0$。

将上述的误判概率与误判损失结合起来，定义平均误判损失（Expected Cost of Misclassification，简记为ECM）如下

$$\mathrm{ECM}(R_1,R_2)=L(2|1)P(2|1)p_1+L(1|2)P(1|2)p_2$$

一个合理的判别规则应使ECM达到极小。

2.两总体的Bayes判别

由上面叙述，要选择样本空间Ω的一个划分R_1和$R_2=\Omega-R_1$使得平均损失ECM达到极小。

定理 6.3 极小化平均误判损失ECM的R_1和R_2为

$$R_1=\left\{x:\frac{f_1(x)}{f_2(x)}\geq\frac{L(1|2)}{L(2|1)}\cdot\frac{p_2}{p_1}\right\},$$

$$R_2=\left\{x:\frac{f_1(x)}{f_2(x)}<\frac{L(1|2)}{L(2|1)}\cdot\frac{p_2}{p_1}\right\},$$

（当$\frac{f_1(x)}{f_2(x)}=\frac{L(1|2)}{L(2|1)}\cdot\frac{p_2}{p_1}$时，即$x$为边界点，它可归入$R_1$，$R_2$的任何一个，为了方便就将它归入$R_1$）。

由上述定理，得到两总体的 Bayes 判别准则

$$\begin{cases} x \in X_1, \quad \text{当}x\text{使得} \dfrac{f_1(x)}{f_2(x)} \geq \dfrac{L(1|2)}{L(2|1)} \cdot \dfrac{p_2}{p_1}, \\ x \in X_2, \quad \text{当}x\text{使得} \dfrac{f_1(x)}{f_2(x)} < \dfrac{L(1|2)}{L(2|1)} \cdot \dfrac{p_2}{p_1}. \end{cases}$$

第六章 判别分析

两总体 Bayes 判别的步骤：

（1）新样本点 $x_0=[x_{01},x_{02},\cdots,x_{0p}]^T$ 的密度函数比 $f_1(x_0)/f_2(x_0)$；

（2）损失比 $L(1|2)/L(2|1)$；

（3）先验概率比 p_2/p_1。

（4）
$$\begin{cases} x\in X_1, & \text{当}x\text{使得}\dfrac{f_1(x)}{f_2(x)}\geq\dfrac{L(1|2)}{L(2|1)}\cdot\dfrac{p_2}{p_1}, \\ x\in X_2, & \text{当}x\text{使得}\dfrac{f_1(x)}{f_2(x)}<\dfrac{L(1|2)}{L(2|1)}\cdot\dfrac{p_2}{p_1}. \end{cases}$$

第六章 判别分析

损失和先验概率以比值的形式出现是很重要的，因为确定两种损失的比值（或两总体的先验概率的比值往往比确定损失本身（或先验概率本身）来得容易。

第六章 判别分析

下面列举三种特殊情况：

（1） 当 $p_2 / p_1 = 1$

$$\begin{cases} x \in X_1, \quad \text{当}x\text{使得} \dfrac{f_1(x)}{f_2(x)} \geq \dfrac{L(1|2)}{L(2|1)}, \\ x \in X_2, \quad \text{当}x\text{使得} \dfrac{f_1(x)}{f_2(x)} < \dfrac{L(1|2)}{L(2|1)}. \end{cases}$$

（2） 当 $L(1|2)/L(2|1)=1$ 时

$$\begin{cases} x\in X_1, & \text{当}x\text{使得}\dfrac{f_1(x)}{f_2(x)}\geq\dfrac{p_2}{p_1}, \\ x\in X_2, & \text{当}x\text{使得}\dfrac{f_1(x)}{f_2(x)}<\dfrac{p_2}{p_1}. \end{cases}$$

（3） $p_1/p_2=L(1|2)/L(2|1)=1$ 时

$$\begin{cases} x\in X_1, & \text{当}x\text{使得}\dfrac{f_1(x)}{f_2(x)}\geq 1, \\ x\in X_2, & \text{当}x\text{使得}\dfrac{f_1(x)}{f_2(x)}<1. \end{cases}$$

第六章 判别分析

对于具体问题:

如果先验概率或者其比值(p2/p1)都难以确定，此时就利用特殊规则(1)　p2/p1=1;

如误判损失或者其比值都是难以确定，此时就利用特殊规则（2）$L(1|2)/L(2|1)=1$;

如果上述两者都难以确定,则利用特殊规则(3),一种无可奈何的办法。

当然判别也变得很简单，若$f_1(x)\geq f_2(x)$，则判$x\in X_1$，否则判$x\in X_2$。

第六章 判别分析

将上述的两总体 Bayes 判别应用于正态总体 $X_i \sim N_p(\mu_i, \Sigma_i)$ $(i=1,2)$，分两种情况讨论。

（1）$\Sigma_1 = \Sigma_2 = \Sigma$，$\Sigma$ 正定，此时 X_i 的密度为

$$f_i(x) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(x-\mu_i)^T \Sigma^{-1}(x-\mu_i)\right].$$

定理 设总体 $X_i \sim N_p(\mu_i, \Sigma)$ $(i=1,2)$，其中 Σ 正定，则使平均误判损失极小的划分为

$$\begin{cases} R_1 = \{x : W(x) \geq \beta\}, \\ R_2 = \{x : W(x) < \beta\}. \end{cases}$$

$$\frac{f_1(x)}{f_2(x)} \geq \frac{L(1|2)}{L(2|1)} \cdot \frac{p_2}{p_1},$$
$$\frac{f_1(x)}{f_2(x)} < \frac{L(1|2)}{L(2|1)} \cdot \frac{p_2}{p_1}.$$

其中

$$W(x) = [x - \frac{1}{2}(\mu_1 + \mu_2)]^T \Sigma^{-1}(\mu_1 - \mu_2),$$

$$\beta = \ln \frac{L(1|2) \cdot p_2}{L(2|1) \cdot p_1}.$$

第六章 判别分析

如果总体的μ_1,μ_2和Σ未知，通过计算样本的$\hat{\mu}_1,\hat{\mu}_2$和$\hat{\Sigma}$来代替μ_1,μ_2和Σ，得到的判别函数

$$W(x)=[x-\frac{1}{2}(\hat{\mu}_1+\hat{\mu}_2)]^T\hat{\Sigma}^{-1}(\hat{\mu}_1-\hat{\mu}_2)$$

称为 Anderson 线性判别函数

判别的规则为

$$\begin{cases} x\in X_1, & \text{当}x\text{使得}W(x)\geq\beta, \\ x\in X_2, & \text{当}x\text{使得}W(x)<\beta, \end{cases}$$

$$\beta=\ln\frac{L(1\mid 2)\cdot p_2}{L(2\mid 1)\cdot p_1}.$$

第六章 判别分析

这里应该指出，总体参数用其估计来代替，所得到的规则，仅仅只是最优（在平均误判损失达到极小的意义下）规则的一个估计，这时对于一个具体问题来讲，我们并没有把握说所得到的规则能够使平均误判损失达到最小，但当样本的容量充分大时，估计$\hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma}$分别和$\mu_1, \mu_2, \Sigma$很接近，因此我们有理由认为“样本”判别规则的性质会很好。

（2）$\Sigma_1 \neq \Sigma_2$（Σ_1, Σ_2正定）

由于误判损失 EMC 极小化的划分规则，依赖于密度函数之比 $f_1(x)/f_2(x)$ 或 对数 $\ln(f_1(x)/f_2(x))$. 把协方差矩阵不等的两个多元正态密度代入这个比后，包含 $|\Sigma_i|^{1/2}$（$i=1,2$）的因子不能消去，而且 $f_i(x)$ 的指数部分也不能组合成简单表达式，因此，对于 $\Sigma_1 \neq \Sigma_2$ 时，可得判别区域

$$\begin{cases} R_1 = \{x : W(x) \geq K\}, \\ R_2 = \{x : W(x) < K\}, \end{cases}$$

其中

$$W(x)=-\frac{1}{2}x^T(\Sigma_1^{-1}-\Sigma_2^{-1})x+(\mu_1^T\Sigma_1^{-1}-\mu_2^T\Sigma_2^{-1})x$$

$$K=\ln\left(\frac{L(1|2)p_2}{L(2|1)p_1}\right)+\frac{1}{2}\ln\frac{|\Sigma_1|}{|\Sigma_2|}+\frac{1}{2}(\mu_1^T\Sigma_1^{-1}\mu_1-\mu_2^T\Sigma_2^{-1}\mu_2)$$

显然，判别函数$W(x)$是关于x的二次函数，它比$\Sigma_1=\Sigma_2$时的情况复杂得多。如果$\mu_i,\Sigma_i(i=1,2)$未知，仍可采用其估计来代替。

例 6.3　下表是某气象站预报有无春旱的实际资料，x_1与x_2都是综合预报因子（气象含义从略），有春旱的是6个年份的资料，无春旱的是8个年份的资料，它们的先验概率分别用6/14和8/14来估计，并设误判损失相等，试建立 Anderson 线性判别函数。

第六章 判别分析

序号		1	2	3	4	5	6	7	8
春早	x_1	24.8	24.1	26.6	23.5	25.5	27.4		
	x_2	-2.0	-2.4	-3.0	-1.9	-2.1	-3.1		
	$W(x_1,x_2)$	3.0156	2.8796	10.0929	-0.0322	4.8098	12.0960		
无春早	x_1	22.1	21.6	22.0	22.8	22.7	21.5	22.1	21.4
	x_2	-0.7	-1.4	-0.8	-1.6	-1.5	-1.0	-1.2	-1.3
	$W(x_1,x_2)$	-6.9371	-5.6602	-6.8144	-2.4897	-3.0303	-7.1958	-5.2789	-6.4097

$$W(x)=[x-\frac{1}{2}(\hat{\mu}_1+\hat{\mu}_2)]^T\hat{\Sigma}^{-1}(\hat{\mu}_1-\hat{\mu}_2)$$

称为 Anderson 线性判别函数

判别的规则为

$$\begin{cases} x\in X_1, & \text{当}x\text{使得}W(x)\geq\beta, \\ x\in X_2, & \text{当}x\text{使得}W(x)<\beta, \end{cases} \qquad \beta=\ln\frac{L(1|2)\cdot p_2}{L(2|1)\cdot p_1}.$$

由表中数据计算得

$$\hat{\mu}_1 = [25.3167, \quad -2.4167]^T ,$$

$$\hat{\mu}_2 = [22.0250, \quad -1.1875]^T ,$$

$$\hat{\Sigma} = \begin{bmatrix} 1.0819 & -0.3109 \\ -0.3109 & 0.1748 \end{bmatrix},$$

$$\beta = \ln \frac{p_2}{p_1} = 0.288.$$

将上述计算结果代入 Anderson 线性判别函数得

$$W(x) = W(x_1, x_2) = 2.0893x_1 - 3.3165x_2 - 55.4331.$$

第六章 判别分析

判别限为 0.288，将表中的数据代入$W(x)$，计算的结果填在表中$W(x_1,x_2)$相应的栏目中，错判的只有一个，即春旱中的第 4 号，与历史资料的拟合率达 93%。

序号		1	2	3	4	5	6	7	8
春旱	x_1	24.8	24.1	26.6	23.5	25.5	27.4		
	x_2	-2.0	-2.4	-3.0	-1.9	-2.1	-3.1		
	$W(x_1,x_2)$	3.0156	2.8796	10.0929	-0.0322	4.8098	12.0960		
无春旱	x_1	22.1	21.6	22.0	22.8	22.7	21.5	22.1	21.4
	x_2	-0.7	-1.4	-0.8	-1.6	-1.5	-1.0	-1.2	-1.3
	$W(x_1,x_2)$	-6.9371	-5.6602	-6.8144	-2.4897	-3.0303	-7.1958	-5.2789	-6.4097

计算的Matlab程序如下：

```
clc,clear
a=[24.8   24.1   26.6  23.5   25.5  27.4
-2.0  -2.4   -3.0   -1.9   -2.1   -3.1]';
b=[22.1  21.6   22.0  22.8   22.7  21.5   22.1   21.4
-0.7  -1.4   -0.8  -1.6   -1.5  -1.0  -1.2  -1.3]';
n1=6; n2=8;
mu1=mean(a), mu2=mean(b) %计算两个总体样本的均值向量,注意得到的是
行向量
sig1=cov(a);sig2=cov(b); %计算两个总体样本的协方差矩阵
sig=((n1-1)*sig1+(n2-1)*sig2)/(n1+n2-2) %计算两总体公共协方差阵的估计
beta=log(8/6)
syms x  [1,2]  %定义符号行向量[x1,x2]
wx=(x-0.5*(mu1+mu2))*inv(sig)*(mu1-mu2)';  %构造判别函数
wx=vpa(wx,6)  %显示判别函数
ahat=subs(wx, x, {a(:,1),a(:,2)})';  %计算总体1样本的判别函数值
bhat=subs(wx, x, {b(:,1),b(:,2)})';  %计算总体2样本的判别函数值
ahat=vpa(ahat,6), bhat=vpa(bhat,6) %显示6位数字的符号数
      sol1=(double(ahat)>beta),  sol2=(double(bhat)<beta)   %回代，计算误判
```