判别分析

类型：判断样品属于已知类型中哪一类。

判别分析模型：

设有 k 个总体 G_1，G_2，…，G_k，它们都是 p 元总体，其数量指标是

$$\boldsymbol{X}=(\boldsymbol{X}_1,\boldsymbol{X}_2,...,\boldsymbol{X}_p)^T$$

设总体 G_i 的分布函数是 $F_i(x)=F_i(x_1, x_2, \cdots, x_p)$，i=1,2,…,k，通常是连续型总体，即 G_i 具有概率密度 $f_i(x)=f_i(x_1, x_2, \cdots, x_p)$。对于任一新样品数据 $x=(x_1, x_2, \cdots, x_p)^T$,要判断它来自哪一个总体 G_i。

通常各个总体 G_i 的分布是未知的，它需要由各总体 G_i 取得的样本数据资料来估计。一般，先要估计各个总体的均值向量与协方差矩阵。从每个总体 G_i 取得的样本叫训练样本。判别分析从各训练样本中的提取各总体的信息，构造一定的判别准则，判断新样品属于哪个总体。

从统计学的角度，要求判别准则在某种准则下是最优的，例如错判的概率最小或错判的损失最小等。

由于判别准则的不同，有各种不同的判别分析方法：距离判别、Bayes 判别和 Fisher 判别等。

一、距离判别

1. 两个总体的距离判别

1.1 距离定义

马氏平方距离：设 x，y 是从均值向量为 μ、协方差矩阵为 Σ 的总体 G 中抽取的两个样品，马氏距离定义为：

$$\begin{aligned}&d^2(x,y)=(x-y)^T\Sigma^{-1}(x-y),\\&d^2(x,G)=(x-\mu)^T\Sigma^{-1}(x-\mu),\\&d^2(G_1,G_2)=(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2)\end{aligned}\qquad(1)$$

1.2 双总体的判别规则

设 G_1，G_2 为两个不同的 p 元已知总体，G_i 的均值向量是 $\mu_i, i=1,2,$ G_i 的协方差矩阵是 $\Sigma_i, i=1,2$。设 $x=(x_1,x_2,...,x_p)^T$ 是一个待判样品，距离判别准则为

$$\begin{cases} x\in G_1, d(x,G_1)\le d(x,G_2) \\ x\in G_2, d(x,G_1)> d(x,G_2) \end{cases} \tag{2}$$

即当 x 到 G_1 的马氏距离不超过到 G_2 的马氏距离时，判 x 来自 G_1；反之，判 x 来自 G_2。

1.3 两个矩阵协方差矩阵相等的情况

1.3.1 已知 Σ,μ_1,μ_2

$$\begin{aligned} d^2(x,G_2)-d^2(x,G_1) &= (x-\mu_2)^T\Sigma^{-1}(x-\mu_2)-(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) \\ &= 2(\mu_1-\mu_2)^T\Sigma^{-1}[x-\tfrac{1}{2}(\mu_1+\mu_2)] \end{aligned} \tag{3}$$

记

$$W(x)=a^T(x-\bar{\mu})$$

其中 $a=\Sigma^{-1}(\mu_1-\mu_2), \bar{\mu}=\frac{1}{2}(\mu_1+\mu_2)$，则

$$d^2(x,G_2)-d^2(x,G_1)=2W(x) \tag{4}$$

距离判别简化为

$$\begin{cases} x\in G_1, 若\ W(x)\ge 0; \\ x\in G_2, 若\ W(x)<0. \end{cases} \tag{5}$$

1.3.1 未知 Σ,μ_1,μ_2

数据资料来自两个总体的训练样本，每个样品皆是 p 元向量。

总体 G_1 的训练样本：$x_1^{(1)},x_2^{(1)},...,x_{n_1}^{(1)}$，容量：$n_1$；

总体 G_2 的训练样本：$x_1^{(2)},x_2^{(2)},...,x_{n_2}^{(2)}$，容量：$n_2$；

要以训练样本估计 μ_1,μ_2及Σ，其估计量分别为

$$\hat{\mu}_1 = \bar{x}^{(1)}\ ,\ \hat{\mu}_2 = \bar{x}^{(2)}\ ,$$
$$\hat{\Sigma} = \frac{(n_1-1)S_1+(n_2-1)S_2}{n_1+n_2-2}. \tag{6}$$

其中 S_1，S_2 为两个训练样本的协方差矩阵。

距离判别规则为：

$$\begin{cases} x\in G_1\text{，若 } \hat{W}(x)\geq 0; \\ x\in G_2\text{，若 } \hat{W}(x)<0. \end{cases} \tag{7}$$

其中 $\hat{W}(x)=\hat{a}^T(x-\bar{x})$，$\hat{a}=\hat{\Sigma}^{-1}(\bar{x}^{(1)}-\bar{x}^{(2)})$，$\bar{x}=\frac{1}{2}(\bar{x}^{(1)}+\bar{x}^{(2)})$.

1.4 两个总体协方差矩阵不相等的情况

1.4.1 已知 $\Sigma_1,\Sigma_2,\mu_1,\mu_2$

令

$$d_1^2(x)=(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1),$$
$$d_2^2(x)=(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2). \tag{8}$$

距离判别规则如下：

$$\begin{cases} x\in G_1\text{，若 } d_1^2(x)\leq d_2^2(x), \\ x\in G_2\text{，若 } d_1^2(x)> d_2^2(x). \end{cases} \tag{9}$$

1.4.2 未知 $\Sigma_1,\Sigma_2,\mu_1,\mu_2$

数据资料来自两个总体的训练样本，每个样品皆是 p 元向量。

要以训练样本估计 μ_1,μ_2及$\Sigma_1,\Sigma_1.$，然后用估计值进行判断。

1.5 判别准则的评价

当一个判别准则提出以后，还要研究其优良性。考察一个判别准则的优良性，要考察误判概率，即考察 x 属于 G_1 而误判为属于 G_2，或 x 属于 G_2 而误判为属于 G_1 的概率。下面介绍一训练样本为基础的用回代方法估计误判率的方法。

1.5.1 误差率回代估计法

将全体训练样本作为新样品，逐个回代已建立的判别准则中判别归属，

这个过程称为回判，回判结果如下：

回判情况 实际归类	G_1	G_2
G_1	n_{11}	n_{12}
G_2	n_{21}	n_{22}

误判率的回代估计为：

$$\hat{a} = \frac{n_{12} + n_{21}}{n_1 + n_2} \tag{10}$$

它常常比真实误判率小，但可以作为真实误判率的一种估计。

1.5.2 误判率的交叉确认估计

误判率的交叉确认估计是每次剔出训练样本中的一个样品，利用其余容量为 n_1+n_2-1 的训练样本建立判别准则，再用建立的判别准则对删除的那个样品作判别。对训练样本中的每个样品作上述分析，以其误判的比例作为误判率的估计。具体步骤如下：

1）从总体 G_1 的容量为 n_1 的训练样本开始，剔除其中的一个样品，用剩余的容量为 n_1-1 的训练样本和总体 G_2 的训练样本建立判别函数；

2）用建立的判别函数对删除的那个样品做判别；

3）重复步骤 1)、2)，直到 G_1 的训练样本中的 n_1 个样品一次被删除，又进行判别。其误判样品个数记为 n_{12}；

4）对总体 G_2 的训练样本重复步骤 1)、2)、3)，并记其误判样品个数为 n_{21}。

误判率的交叉确认估计为：

$$\hat{a} = \frac{n_{12} + n_{21}}{n_1 + n_2} \tag{11}$$

1.6 多总体的距离判别

设有 k 个总体 G_1，G_2，…，G_k，均值向量分别为 $\mu_1,\mu_2,...,\mu_k$,协方差矩阵分别为 $\Sigma_1,\Sigma_2,...,\Sigma_k$,，类似两总体的距离判别方法，计算新样品 x 到各总体的马氏距离，比较这 k 个距离，判定 x 属于其马氏距离最短的总体。若最短距离在不只一个总体达到，则可将 x 判归具有最短距离总体的任一个。

当总体的均值向量和协方差矩阵未知时，使用训练样本作估计。也可以与两总体相同的方式作误判率的回代估计与交叉确认估计。

二、Bayes 判别

2.1 Bayes 判别的基本思想

Bayes 统计是现代统计学的重要分支，其基本思想是：假定对所研究的对象（总体）在抽样前已有一定的认识，常用先验分布来描述这种认识，然后给予抽取的样本再对先验认识作修正，得到后验分布，而各种统计推断均基于后验分布进行。将Bayes 统计的思想用于判别分析，就得到Bayes 判别。

设 G_1，G_2，…，G_k 为 k 个 p 元总体，分别具有概率密度 $f_1(x),f_2(x),...,f_k(x).$在进行判别分析以前，我们已对各总体有一定的了解。一般说来，一个待判样品应该首先考虑判入有较大可能出现的总体之中。在 Bayes 判别中，开应该考虑误判引起的损失。

2.2 两个总体的Bayes 判别

2.2.1 一般讨论

考虑两个 p 元总体 G_1 和 G_2，它们分别具有概率密度 $f_1(x)$，$f_2(x)$，

G_1 和 G_2 出现的先验概率为 p_1 和 p_2，且 $p_1+p_2=1$。

对于 p 元指标 $x=(x_1, x_2, \cdots, x_p)^T$ 来自 R^p。一个判别法则实质上是对 R^p 的一个划分，记为 R_1 和 R_2，并满足下列条件：

$$R_1 \cup R_2 = R^p, R_1 \cap R_2 = \varnothing. \tag{12}$$

一个划分 $R=(R_1, R_2)$ 相当于一个判别准则 R。在判别准则 R 下将来自 G_1 的样品误判为 G_2 的概率是

$$P(2|1,R)=\int_{R_2} f_1(x)dx \tag{13}$$

而将来自 G_2 的样品误判为 G_1 的概率为

$$P(1|2,R)=\int_{R_1} f_2(x)dx \tag{14}$$

设将 G_1 误判为 G_2 造成的损失是 $c(2|1)$，而将 G_2 误判为 G_1 在造成的损失是 $c(1|2)$。Bayes 判别即寻求 $R=(R_1, R_2)$，使平均误判损失达到最小。下面总假定 $c(1|1)=c(2|2)=0$。

情况 1：c(1|2)=c(2|1)

当得到新样品 x 后，由 Bayes 公式得总体 G_1，G_2 的后验概率是

$$\begin{cases} P(G_1|x)=\dfrac{p_1 f_1(x)}{p_1 f_1(x)+p_2 f_2(x)}; \\ P(G_2|x)=\dfrac{p_2 f_2(x)}{p_1 f_1(x)+p_2 f_2(x)}. \end{cases} \tag{15}$$

两总体 Bayes 判别的一个最优划分是

$$\begin{cases} R_1=\{x: P(G_1|x)\geq P(G_2|x)\}; \\ R_2=\{x: P(G_1|x)< P(G_2|x)\}. \end{cases} \tag{16}$$

此时的 **Bayes 判别法则**：

$$\begin{cases} x\in G_1: \text{ 若 } \mathrm{P}(G_1|x)\geq \mathrm{P}(G_2|x); \\ x\in G_2: \text{ 若 } \mathrm{P}(G_1|x)< \mathrm{P}(G_2|x). \end{cases} \tag{17}$$

最优划分 R 使得平均误判概率

$$p^* = p_1 P(2|1,R)+p_2 P(1|2,R) \tag{18}$$

达到最小。

情况2：c(1|2)≠c(2|1)

关于先验分布 p_1、p_2，误判所造成的平均损失为：

$$L = c(2|1)p_1P(2|1,R) + c(1|2)p_2P(1|2,R). \quad (19)$$

Bayes 判别（即使 L 达到最小）的最优划分为：

$$\begin{cases} R_1 = \{x: c(2|1)P(G_1|x) \geq c(1|2)P(G_2|x)\}; \\ R_2 = \{x: c(2|1)P(G_1|x) < c(1|2)P(G_2|x)\}. \end{cases} \quad (20)$$

此时的**Bayes判别法则**:

$$\begin{cases} x \in G_1: \text{ 若 } c(2|1)\mathrm{P}(G_1|x) \geq c(1|2)\mathrm{P}(G_2|x); \\ x \in G_2: \text{ 若 } c(2|1)\mathrm{P}(G_1|x) < c(1|2)\mathrm{P}(G_2|x). \end{cases} \quad (21)$$

2.2.2 两个正态总体的Bayes判别

需分 c(1|2)和 c(2|1)相等与否，两个总体的协方差矩阵相等与否分别讨论。（详细参见：范金城，梅长林编著. 数据分析：P174-177. 北京：科学出版社，2002.）

2.2.3 误判率的计算

（参见：范金城，梅长林编著. 数据分析：P177-182. 北京：科学出版社，2002.）

2.3 多个总体的Bayes判别

（参见：范金城，梅长林编著. 数据分析：P182-187. 北京：科学出版社，2002.）

判别分析课堂例题

例题1：某气象站预报某地区有无春旱的观测资料中，x_1 与 x_2 是与气象有关的综合预报因子。数据包括发生春旱的 6 个年份的 x_1，x_2 观测值和无春

旱的 8 个年份的相应观测值（见下表）。

表 某地区有无春旱的观测数据

G_1:有春旱			G_2:无春旱		
序号	X_1	X_2	序号	X_1	X_2
1	24.6	-2.0	1	22.1	-0.7
2	24.7	-2.4	2	21.6	-1.4
3	26.6	-3.0	3	22.0	-0.8
4	23.5	-1.9	4	22.8	-1.6
5	25.5	-2.1	5	22.7	-1.5
6	27.4	-3.1	6	21.5	-1.0
			7	22.1	-1.2
			8	21.4	-1.3

在假定 $\Sigma_1 = \Sigma_2 = \Sigma$ 条件下，建立距离判别函数并估计误判率；

解：经过计算

$$\bar{x}^{(1)}=\begin{bmatrix}25.3167\\-2.4167\end{bmatrix},\qquad \bar{x}^{(2)}=\begin{bmatrix}22.0250\\-1.1875\end{bmatrix},$$

$$S_1=\begin{bmatrix}2.2137 & -0.6577\\-0.6577 & 0.2697\end{bmatrix},\ S_2=\begin{bmatrix}0.2736 & -0.0632\\-0.0632 & 0.1069\end{bmatrix}$$

$$\bar{x}=\frac{1}{2}(\bar{x}^{(1)}+\bar{x}^{(2)})=\begin{bmatrix}23.6709\\-1.8021\end{bmatrix},$$

$$S=\frac{(n_1-1)S_1+(n_2-1)S_2}{n_1+n_2-2}=\begin{bmatrix}1.0820 & -0.3109\\-0.3109 & 0.1747\end{bmatrix}$$

判别函数为：

$$W(x)=\hat{a}^T(x-\bar{x})=\begin{bmatrix}2.0889 & -3.3179\end{bmatrix}\left(\begin{bmatrix}x_1\\x_2\end{bmatrix}-\begin{bmatrix}23.6709\\-1.8021\end{bmatrix}\right)$$

$$=-55.4255+2.0889x_1-3.3179x_2$$

利用回代法将总体 G_1（春旱）的第 4 号样品误判来自总体 G_2（无春旱）的样品，误判率为

$$\alpha=1/14=0.0714$$

利用交叉确认法，同样将总体 G_1（春旱）的第 4 号样品误判来自总体 G_2（无春旱）的样品，误判率为

$$\alpha=1/14=0.0714$$

例题 2：我国山区某大型化工厂，在厂区及邻近地区挑选有代表性的 15 个大气取样点，每日 4 次同时抽取大气样品，测定其中含有的 6 种气体的浓度，前后共 4 天，每个取样点没种气体实测 16 次。计算每个取样点没种气体的平均浓度，数据见下表所示。气体数据对应得污染地区分类如表中最后一列所示。现有两个取自该地区的 4 个气体样本，气体指标如表中后 4 行所示，试判别这 4 个样品的污染分类。

表　大气样品数据表

气体	氯	硫　化　氢	二 氧 化 硫	碳 4	环 氧 氯 丙 烷	环己烷	污 染 分 类

1	0.056	0.084	0.031	0.038	0.0081	0.022	1
2	0.040	0.055	0.100	0.110	0.0220	0.0073	1
3	0.050	0.074	0.041	0.048	0.0071	0.020	1
4	0.045	0.050	0.110	0.100	0.0250	0.0063	1
5	0.038	0.130	0.079	0.170	0.0580	0.043	2
6	0.030	0.110	0.070	0.160	0.0500	0.046	2
7	0.034	0.095	0.058	0.160	0.200	0.029	1
8	0.030	0.090	0.068	0.180	0.220	0.039	1
9	0.084	0.066	0.029	0.320	0.012	0.041	2
10	0.085	0.076	0.019	0.300	0.010	0.040	2
11	0.064	0.072	0.020	0.250	0.028	0.038	2
12	0.054	0.065	0.022	0.280	0.021	0.040	2
13	0.048	0.089	0.062	0.260	0.038	0.036	2
14	0.045	0.092	0.072	0.200	0.035	0.032	2
15	0.069	0.087	0.027	0.050	0.089	0.021	1
样品1	0.052	0.084	0.021	0.037	0.0071	0.022	
样品2	0.041	0.055	0.110	0.110	0.0210	0.0073	
样品3	0.030	0.112	0.072	0.160	0.056	0.021	
样品4	0.074	0.083	0.105	0.190	0.020	1.000	

Matlab 函数介绍：

函数名称：classify

调用格式：[class,err,…]=classify(sample,training,group,…)

说明：sample：待判样品；

training：训练样本；

group：分类变量。

注意:

sample 与 training 具有相同的列数；

group 与 training 具有相同的行数。

返回：class：样品的分类结果；err：误判率的估计。

Matlab 程序：

```
training=[0.056 0.084 0.031 0.038 0.0081 0.022;
0.040 0.055 0.100 0.110 0.0220 0.0073;
0.050 0.074 0.041 0.048 0.0071 0.020;
0.045 0.050 0.110 0.100 0.0250 0.0063;
0.038 0.130 0.079 0.170 0.0580 0.043;
0.030 0.110 0.070 0.160 0.0500 0.046;
0.034 0.095 0.058 0.160 0.200 0.029;
0.030 0.090 0.068 0.180 0.220 0.039;
0.084 0.066 0.029 0.320 0.012 0.041;
0.085 0.076 0.019 0.300 0.010 0.040;
0.064 0.072 0.020 0.250 0.028 0.038;
0.054 0.065 0.022 0.280 0.021 0.040;
0.048 0.089 0.062 0.260 0.038 0.036;
0.045 0.092 0.072 0.200 0.035 0.032;
0.069 0.087 0.027 0.050 0.089 0.021];
group=[1 1 1 1 2 2 1 1 2 2 2 2 2 2 1]';
```

```
sample=[0.052 0.084 0.021 0.037 0.0071 0.022;
0.041 0.055 0.110 0.110 0.0210 0.0073;
0.030 0.112 0.072 0.160 0.056 0.021;
0.074 0.083 0.105 0.190 0.020 1.000];
[class,err]=classify(sample,training,group)
class=[1 1 2 2]
```