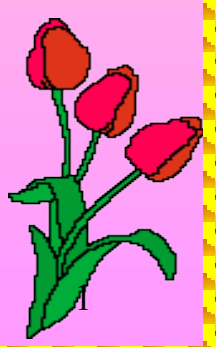
第七章 判别分析

第一节 距离判别法

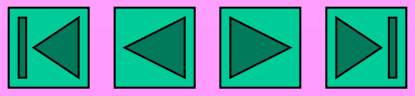
第二节 贝叶斯判别法



判别分析又称“分辨法”，它是在分类确定的条件下，根据某一研究对象的各种特征值判别其类型归属问题的一种多变量统计分析方法。

即一种统计判别和分组技术，通过一定数量样本的一个分组变量和相应的其他多元变量的已知信息，来确定分组与其他多元变量信息所属的样本进行判别分组方法。

其基本原理是按照一定的判别准则，建立一个或多个判别函数，用研究对象的大量资料确定判别函数中的待定系数，并计算判别指标。据此即可确定某一样本归属于何类。



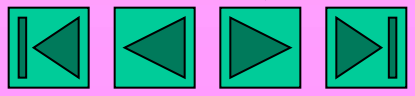
例如：

1. 通过病人的某些检查指标，需要辨识出该病人所患疾病归于与哪一类；

2. 根据气象的多项指标数据，需要判别出未来天气是晴或阴，有雨或无雨等问题；

一般来说，判别问题的数学描述为：

设有m个p维总体$G_1,G_2,\ldots,G_m$,分别服从一定的分布$F_1(x), F_2(x),\ldots, F_m(x)$,现有一个新样品$x=(x_1,x_2,\ldots,x_p)^T$,如何判别此样品归属哪一个总体的问题。



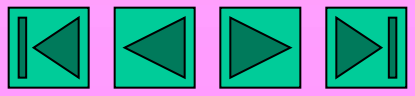
判别问题的解决方法分类：

1. 根据判别中的组数，可以分为两组判别分析和多组判别分析；

2. 根据判别函数的形式，可以分为线性判别和非线性判别；

3.根据判别式处理变量的方法不同，可以分为逐步判别、序贯判别等；

4.根据判别标准不同，可以分为距离判别、Fisher判别、Bayes判别法等。



判别分析方法基本步骤

第一步：明确样品的级别、类别标准。

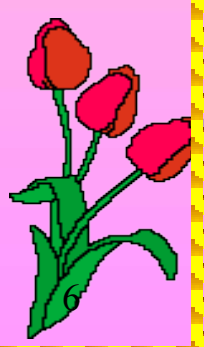
第二步：根据相应的测定值，去构造一个依赖于指标值的判别函数与判别规则 。

第三步：据此规则作出样品归属的判断。

第一节 距离判别法

一 两总体情况

二 多总体情况

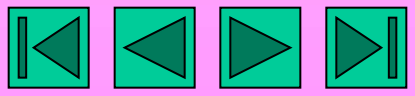


一 两总体情况

1. 两总体判别问题

设有2个p维总体G_1,G_2,分别服从一定的分布$F_1(x)$, $F_2(x)$,现有一个新样品$x=(x_1,x_2,\ldots,x_p)^T$,如何判别此样品归属哪一个总体的问题，即判别x来自G_1或G_2 的问题。

解决方法：先根据已知信息建立样本x到两类G_1或G_2 的距离，再由距离的远近判别此样本x的归属类。



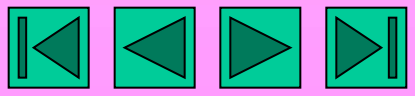
定义样本x到G_1或G_2的距离为$d(x,G_1)$或$d(x,G_2)$，再按下述规则判别x的归属：

如果$d(x,G_1)<d(x,G_2)$,则判别$x\in G_1$;

如果$d(x,G_1)>d(x,G_2)$,则判别$x\in G_2$;

如果$d(x,G_1)=d(x,G_2)$,则不判（待判）。

可见距离函数的选取是关键的要素，若采用欧氏距离，虽然计算相对简单，但若总体各分量的量纲不同，会影响距离值的变化，因此通常采用马氏距离。



2 .马氏距离函数

当两总体均为正态总体，利用其协方差阵选用马氏距离：

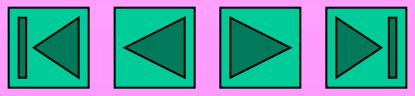
两点间的距离：

$$D(x,y)=\sqrt{(x-y)^T\Sigma_i^{-1}(x-y)}\quad(i=1,2)$$

点到总体的距离：

$$D(x,G_1)=\sqrt{(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)}\quad(i=1,2)$$

其中μ_i和Σ_i分别为总体G_i的均值向量和协方差阵



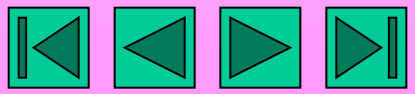
如此定义样本x到G_1或G_2的距离为$D(x,G_1)$或$D(x,G_2)$，再按下述规则判别x的归属：

如果$D(x,G_1) < D(x,G_2)$,则判别$x \in G_1$;

如果$D(x,G_1) > D(x,G_2)$,则判别$x \in G_2$;

如果$D(x,G_1) = D(x,G_2)$,则不判（待判）。

按此规则判别，称为距离判别法。



1）若协方差阵相等 $\Sigma_1 = \Sigma_2 = \Sigma$

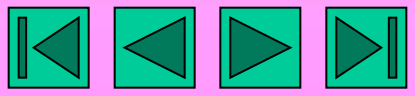
则距离为

$$D(x,y) = \sqrt{(x-y)^T \Sigma^{-1}(x-y)}$$

$$D(x,G_i) = \sqrt{(x-\mu_i)^T \Sigma^{-1}(x-\mu_i)}, \quad i = 1,2$$

此时x到两总体的距离平方的差为

$$D^2(x,G_2) - D^2(x,G_1) = 2(x - \frac{\mu_1 + \mu_2}{2})^T \Sigma^{-1}(\mu_1 - \mu_2)$$



取判别函数：

$$W(x)=\frac{1}{2}[D^2(x,G_2)-D^2(x,G_1)]$$

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

则判别准则为：

若$W(x)>0$,则判别$x\in G_1$;

若$W(x)<0$,则判别$x\in G_2$;

若$W(x)=0$,则不判（待判）。

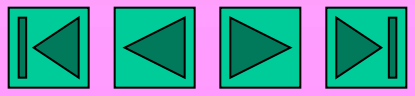
2）若协方差阵不相等　$\Sigma_1 \neq \Sigma_2$

令x到两总体的距离平方差的函数(判别函数）：

$$W(x)=\frac{1}{2}[D^2(x,G_2)-D^2(x,G_1)]$$

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

这样，前述判别准则等价于：



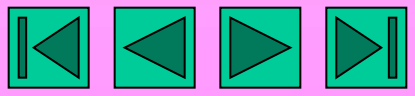
判别准则

若$W(x)>0$,则判别$x\in G_1$;

若$W(x)<0$,则判别$x\in G_2$;

若$W(x)=0$,则不判（待判）。

可见，W(x)与0的差距越大，样品到两总体差异越明显，作出正确判断的把握越大，W(x)越接近0，越造成误判。



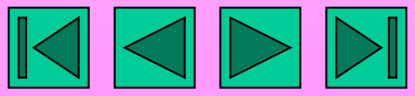
在实际中，当总体均值向量与协方差未知时,可用样本值来估计：

设$x_{(1)}^{(a)}, x_{(2)}^{(a)}, \mathrm{L}\ , x_{(n_a)}^{(a)}$是来自$G_a$的样本$(a = 1, 2)$,

$$x_{(i)}^{(a)} = (x_{i1}^{(a)}, x_{i2}^{(a)}, \mathrm{L}\ , x_{ip}^{(a)})^T, i = 1, 2, \mathrm{L}\ , n_a$$

$$\hat{\mu}_1 = \bar{x}^{(1)} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{(i)}^{(1)}, \hat{\mu}_2 = \bar{x}^{(2)} = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{(i)}^{(2)}$$

$$\hat{\Sigma}_a = S_{(a)}^2 = \frac{1}{n_a - 1}\sum_{k=1}^{n_a}(x_{(k)}^{(a)} - \bar{x}^{(a)})(x_{(k)}^{(a)} - \bar{x}^{(a)})^T$$



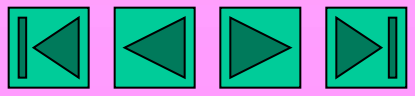
若样本容量均为n时，则样本均值向量与样本协方差为：

设$x_{(1)}^{(a)}, x_{(2)}^{(a)}, \mathrm{L}\ , x_{(n)}^{(a)}$是来自$G_a$的样本$(a=1,2)$,

$$x_{(i)}^{(a)} = (x_{i1}^{(a)}, x_{i2}^{(a)}, \mathrm{L}\ , x_{ip}^{(a)})^T, i = 1,2, \mathrm{L}\ , n$$

$$\hat{\mu}_1 = \bar{x}^{(1)} = \frac{1}{n}\sum_{i=1}^{n} x_{(i)}^{(1)}, \hat{\mu}_2 = \bar{x}^{(2)} = \frac{1}{n}\sum_{i=1}^{n} x_{(i)}^{(2)}$$

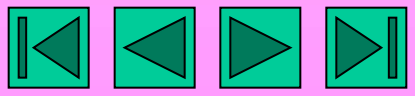
$$\hat{\Sigma}_k = S_{(a)}^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_{(k)}^{(a)} - \bar{x}^{(a)})(x_{(k)}^{(a)} - \bar{x}^{(a)})^T$$



$$\hat{\mu}_1 = \bar{x}^{(a)} = \frac{1}{n}\sum_{i=1}^{n} x_{(i)}^{(a)} = \frac{1}{n}\left[\begin{pmatrix} x_{11}^{(a)} \\ x_{12}^{(a)} \\ \mathbf{M} \\ x_{1p}^{(a)} \end{pmatrix} + \begin{pmatrix} x_{21}^{(a)} \\ x_{22}^{(a)} \\ \mathbf{M} \\ x_{2p}^{(a)} \end{pmatrix} + \mathbf{L} + \begin{pmatrix} x_{n1}^{(a)} \\ x_{n2}^{(a)} \\ \mathbf{M} \\ x_{np}^{(a)} \end{pmatrix}\right] = \begin{pmatrix} \bar{x}_1^{(a)} \\ \bar{x}_2^{(a)} \\ \mathbf{M} \\ \bar{x}_p^{(a)} \end{pmatrix}$$

$$\Sigma_a = S_{(a)}^2 = (s_{ij}^{(a)}) \qquad a = 1, 2$$

$$s_{ij}^{(a)} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ki}^{(a)} - \bar{x}_i^{(a)})(x_{kj}^{(a)} - \bar{x}_j^{(a)})$$



3 两总体距离判别法的步骤

1）先计算x到两总体的距离的平方值：

$$D^2(x,G_i)=(x-\mu_i)^T\Sigma^{-1}(x-\mu_i),\quad i=1,2$$

2) 确定判别函数的值：

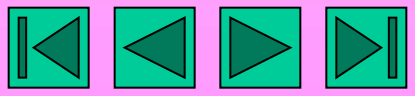
$$W(x)=\frac{1}{2}[D^2(x,G_2)-D^2(x,G_1)]$$

3）确定判别规则进行判断：

若$W(x)>0$,则判别$x\in G_1$;

若$W(x)<0$,则判别$x\in G_2$;

若$W(x)=0$,则不判（待判）。



例7.1.1 设有两个二元正态总体:

$$G_1 \sim N(\mu_1, \Sigma), G_2 \sim N(\mu_2, \Sigma),$$

根据过去收集到的资料已估计出:

$$\hat{\mu}_1 = \begin{pmatrix} 2 \\ 6 \end{pmatrix}, \hat{\mu}_2 = \begin{pmatrix} 4 \\ 2 \end{pmatrix}, \hat{\Sigma} = \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}$$

现有一个样品$x=(3,5)^T$，试计算x到各总体的马氏距离，并用距离判别法判别x的归属。

解： 1) x到两个总体的马氏距离的平方为：

$$D^2(x,G_i)=(x-\mu_i)^T\Sigma^{-1}(x-\mu_i),\quad i=1,2$$

其中 $$\hat{\Sigma}=\begin{pmatrix}1&1\\1&4\end{pmatrix}\Rightarrow\hat{\Sigma}^{-1}=\frac{1}{3}\begin{pmatrix}4&-1\\-1&1\end{pmatrix}$$

$$D^2(x,G_1)=\left[\begin{pmatrix}3\\5\end{pmatrix}-\begin{pmatrix}2\\6\end{pmatrix}\right]^T\frac{1}{3}\begin{pmatrix}4&-1\\-1&1\end{pmatrix}\left[\begin{pmatrix}3\\5\end{pmatrix}-\begin{pmatrix}2\\6\end{pmatrix}\right]$$

$$=\frac{1}{3}(1,-1)\begin{pmatrix}4&-1\\-1&1\end{pmatrix}\begin{pmatrix}1\\-1\end{pmatrix}=\frac{7}{3}$$

$$D(x,G_1)=\sqrt{7/3}$$

$$D^2(x,G_2)=\left[\begin{pmatrix}3\\5\end{pmatrix}-\begin{pmatrix}4\\2\end{pmatrix}\right]^T\frac{1}{3}\begin{pmatrix}4&-1\\-1&1\end{pmatrix}\left[\begin{pmatrix}3\\5\end{pmatrix}-\begin{pmatrix}4\\2\end{pmatrix}\right]$$

$$=\frac{1}{3}(-1,3)\begin{pmatrix}4&-1\\-1&1\end{pmatrix}\begin{pmatrix}-1\\3\end{pmatrix}=\frac{13}{3}$$

$$D(x,G_1)=\sqrt{13/3}$$

2) 确定判别函数的值：

$$W(x)=\frac{1}{2}[D^2(x,G_2)-D^2(x,G_1)]=\frac{1}{2}\times\frac{6}{3}=1>0$$

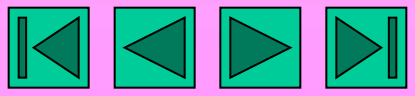
3) 判断：因W(x)>0, 所以此x属于G_1。

二　多总体情况

1. 多总体判别问题

设有m个p维总体$G_1,G_2,\ldots,G_m$分别服从一定的分布$F_1(x), F_2(x),\ldots,F_m(x)$现有一个新样品$x=(x_1,x_2,\ldots,x_p)^T$,如何判别此样品归属哪一个总体的问题，即判别x来自哪个总体的问题。

解决方法：先根据已知信息建立样本x到各类的距离，再由距离的远近判别此样本x的归属类。



设μ_i和Σ_i分别为总体G_i的均值向量和协方差阵

$(i=1,2,\mathrm{L}\ ,m)$　则两点间距离为

$$D(x,y)=\sqrt{(x-y)^T\Sigma_i^{-1}(x-y)}$$

则点到G_i的距离为

$$D(x,G_i)=\sqrt{(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)},\quad i=1,2,\mathrm{L}\ ,m$$

其中各总体的均值与协方差均用矩估计代替，再令判别函数：