# 多元统计与矩阵分析

张锋　*8125345@qq.com*

中国地质大学, 计算机学院, 武汉

# 第4章

# 判别分析

# 内容

(1) 基本思想

(2) 距离判别法

(3) 贝叶斯判别

(4) 费歇判别

(5) 例子

# 内容

(1) 基本思想

(2) 距离判别法

(3) 贝叶斯判别

(4) 费歇判别

(5) 例子

# 判别分析的基本思想

- 基本思想

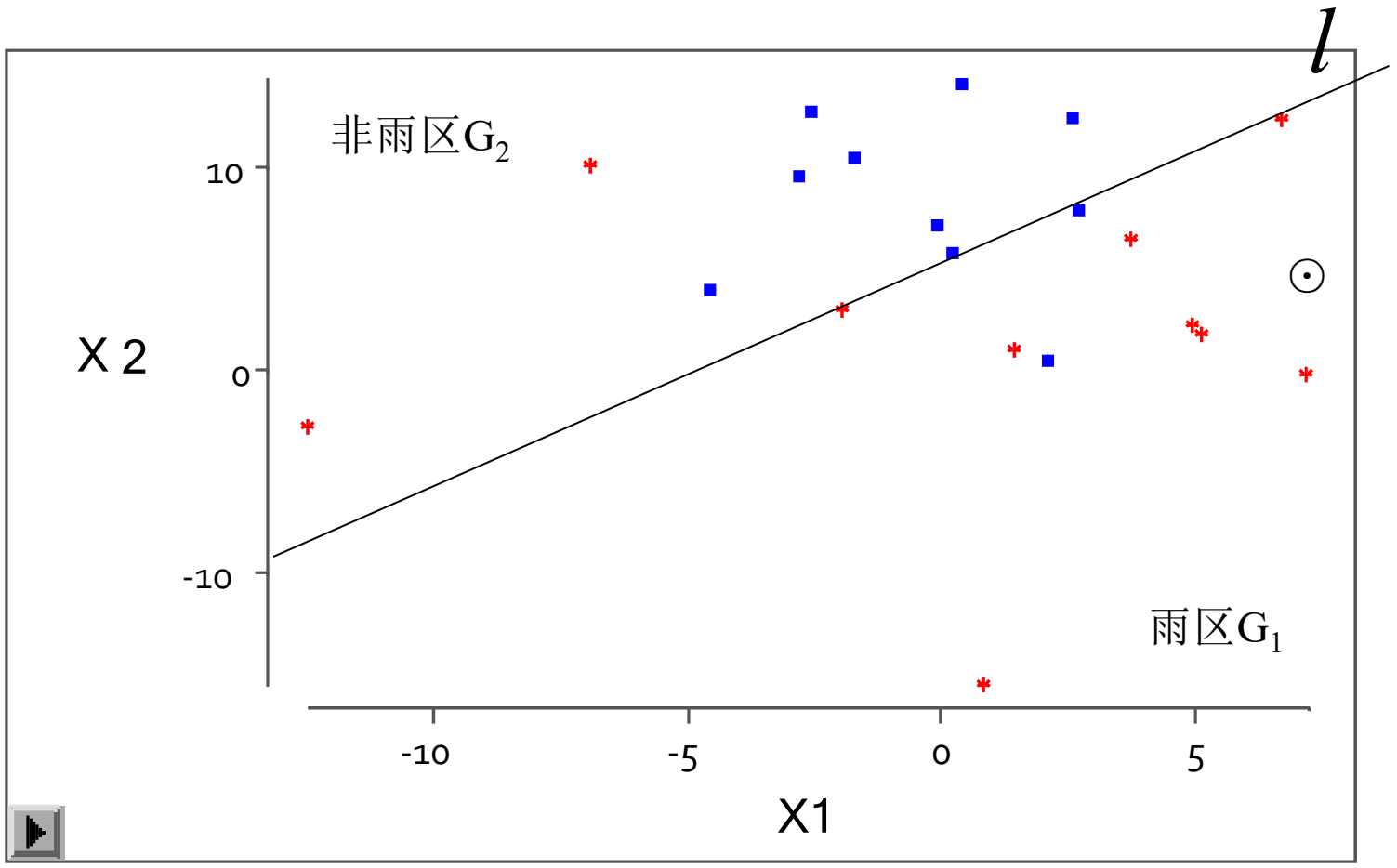
根据已知类别的样本所提供的信息，总结出分类的规律性，建立判别公式和判别准则，判别新的样本点所属类型，是判别个体所属群体的一种统计方法。

根据经验，今天与昨天的湿度差及今天的压差（气压与温度之差）是预报明天下雨或不下雨的两个重要因素。今测得 $x_1$ =8.1，$x_2$ =2.0，试问应预报明天下雨还是不下雨?

这个问题是两类判别问题，总体分为两类，用$G_1$表示下雨，$G_2$表示不下雨。为进行预报，应先收集一批资料，从已有的资料中找出规律，再作预报。

# 收集过去10个雨天和非雨天 $x_1$ 和 $x_2$ 的数值

| 雨天 | | 非雨天 | |
|---|---|---|---|
| $x_1$ | $x_2$ | $x_1$ | $x_2$ |
| -1.9 | 3.2 | 0.2 | 6.2 |
| -6.9 | 10.4 | -0.1 | 7.5 |
| 5.2 | 2.0 | 0.4 | 14.6 |
| 5.0 | 2.5 | 2.7 | 8.3 |
| 7.3 | 0.0 | 2.1 | 0.8 |
| 6.8 | 12.7 | -4.6 | 4.3 |
| 0.9 | -15.4 | -1.7 | 10.9 |
| -12.5 | -2.5 | -2.6 | 13.1 |
| 1.5 | 1.3 | 2.6 | 12.8 |
| 3.8 | 6.8 | -2.8 | 10.0 |


l
非雨区G2
雨区G1
X 2
X1
10
0
-10
-10
-5
0
5

# 内容

(1) 基本思想

(2) 距离判别法

(3) 贝叶斯判别

(4) 费歇判别

(5) 例子

# 距离判别法的基本思想

- 最直观的想法是计算样品到第 i 类总体的距离，哪个距离最小就将它判归哪个总。
- 所以，我们首先考虑的是是否能够构造一个恰当的距离函数，通过样本与某类别之间距离的大小，判别其所属类别。

# 判别分析中常用马氏距离

样品 $\boldsymbol{x}$ 和 $\boldsymbol{G}_i$ 类之间的马氏距离定义为 $\boldsymbol{x}$ 与 $\boldsymbol{G}_i$ 类重心间的距离：

$$d^2(\boldsymbol{x}, \boldsymbol{G}_i) = (\boldsymbol{x} - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i) \quad i = 1,2,\cdots,k$$

马氏距离不受变量间的相关性和量纲的影响

# 两个总体距离判别法

## 1、总体协差阵相等

先考虑两个总体的情况，设有两个协差阵Σ相同的p维正态总体，对给定的样品 $\boldsymbol{x}$ ，判别一个样品 $\boldsymbol{x}$ 到底是来自哪一个总体，一个最直观的想法是计算 $\boldsymbol{x}$ 到两个总体的距离。故我们用马氏距离来给定判别规则，有：

$$\begin{cases} \boldsymbol{x} \in \boldsymbol{G}_1, & \text{如} d^2(\boldsymbol{x}, \boldsymbol{G}_1) < d^2(\boldsymbol{x}, \boldsymbol{G}_2), \\ \boldsymbol{x} \in \boldsymbol{G}_2, & \text{如} d^2(\boldsymbol{x}, \boldsymbol{G}_2) < d^2(\boldsymbol{x}, \boldsymbol{G}_1) \\ \text{待判}, & \text{如} d^2(\boldsymbol{x}, \boldsymbol{G}_1) = d^2(\boldsymbol{x}, \boldsymbol{G}_2) \end{cases}$$

$$d^2(x,G_2)-d^2(x,G_1)$$

$$=(x-\mu_2)'\Sigma^{-1}(x-\mu_2)-(x-\mu_1)'\Sigma^{-1}(x-\mu_1)$$

$$=x'\Sigma^{-1}x-2x'\Sigma^{-1}\mu_2+\mu_2'\Sigma^{-1}\mu_2$$

$$-(x'\Sigma^{-1}x-2x'\Sigma^{-1}\mu_1+\mu_1'\Sigma^{-1}\mu_1)$$

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

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

令 $\overline{\mu}=\frac{\mu_1+\mu_2}{2}$ $\qquad \alpha=\Sigma^{-1}(\mu_1-\mu_2)=(a_1,a_2,\cdots,a_p)'$

$$=2[x-\overline{\mu}]'\alpha$$

$$W(x)=[x-\frac{(\mu_1+\mu_2)}{2}]'\Sigma^{-1}(\mu_1-\mu_2)=(x-\bar{\mu})'\ \alpha$$

$$=\alpha'(x-\bar{\mu}) \qquad \alpha=\Sigma^{-1}(\mu_1-\mu_2)=(a_1,a_2,\cdots,a_p)'$$

$$=a_1(x_1-\bar{\mu}_1)+\cdots+a_p(x_p-\bar{\mu}_p)$$

则前面的判别法则表示为

$$\begin{cases} x\in G_1, & \text{如}W(x)>0, \\ x\in G_2, & \text{如}W(x)<0。 \\ \text{待判}, & \text{如}W(x)=0 \end{cases}$$

显然，W(X)是$x_1,x_2...x_p$的线性函数，称W(X)为线性判别函数

当 $\mu_1$，$\mu_2$，$\boldsymbol{\Sigma}$ 未知时，可通过样本来估计。设 $x_1^{(1)}$，$x_2^{(1)}$，…，$x_{n_1}^{(1)}$ 是从总体 $G_1$ 中取出的样本。$x_1^{(2)}$，$x_2^{(2)}$，…，$x_{n_2}^{(1)}$ 是从总体 $G_2$ 中取出的样本。则

$$\hat{\mu}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i^{(1)} = \overline{x}^{(1)}$$

$$\hat{\mu}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_i^{(2)} = \overline{x}^{(2)}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{(n_1-1)\boldsymbol{S}_1 + (n_2-1)\boldsymbol{S}_2}{n_1+n_2-2}$$

其中，$\boldsymbol{S}_1$，$\boldsymbol{S}_2$ 为总体 $G_1$，$G_2$ 的样本协方差矩阵。

此时，线性判别函数为

$$W(x) = \left(x - \frac{\overline{x}^{(1)} + \overline{x}^{(2)}}{2}\right)' \hat{\boldsymbol{\Sigma}}^{-1}(\overline{x}^{(1)} - \overline{x}^{(2)})$$

$$= (\overline{x}^{(1)} - \overline{x}^{(2)})' \hat{\boldsymbol{\Sigma}}^{-1}\left(x - \frac{\overline{x}^{(1)} + \overline{x}^{(2)}}{2}\right)$$

当 $p=2$ 时，$W(x)$ 为一直线。当 $p>2$ 时，$W(x)$ 为一平面。这条直线或平面把空间的点分为两个部分，一部分属 $G_1$，另一部分属 $G_2$。

特别地，当p=1时，若两个总体分别为 $N(\mu_1,\sigma^2)$ 和 $N(\mu_2,\sigma^2)$ 则判别函数为

$$W(x)=(x-\bar{\mu})\frac{1}{\sigma^2}(\mu_1-\mu_2)$$ ,其中 $\bar{\mu}=\frac{1}{2}(\mu_1+\mu_2)$

不妨设 $\mu_1<\mu_2$

则 $W(x)$ 的符号取决于 $x>\bar{\mu}$ 还是 $x<\bar{\mu}$

因此判别规则可写成：

$$\begin{cases} x\in G_1 & \text{若 } x<\bar{\mu} \\ x\in G_2 & \text{若 } x>\bar{\mu} \\ \text{待判} & \text{若 } x=\bar{\mu} \end{cases}$$

用距离判别所得到的准则看上去是颇为合理的，但用这个判别法有时会错判。如 $x$ 来自 $G_1$，但却落入 $D_2$，被判为属 $G_2$，错判的概率为图中阴影部分的面积，记为 $P(2/1)$，类似地有 $P(1/2)$

显然，$$P(2/1)=1-\Phi\left(\frac{\bar{\mu}-\mu_1}{\sigma}\right)=1-\Phi\left(\frac{\mu_2-\mu_1}{2\sigma}\right)$$

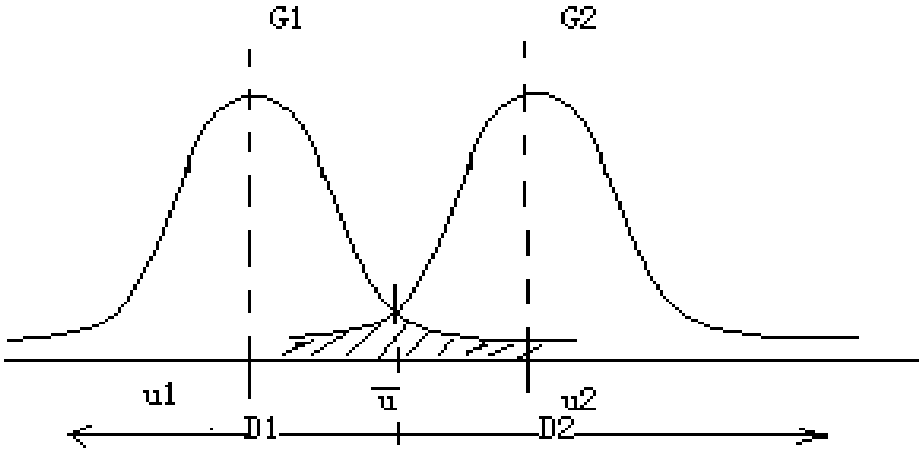
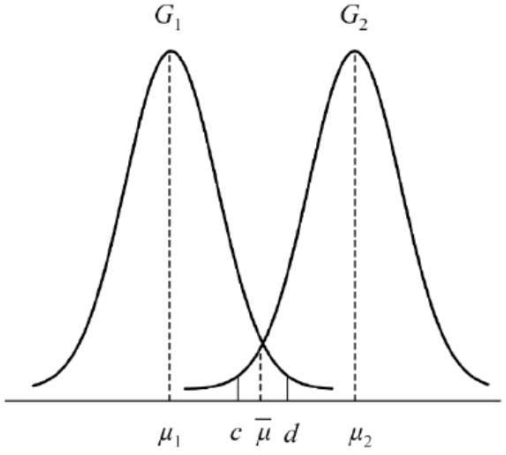


图7-2

$$\begin{cases} x \in G_1 & \text{若 } x \leqslant c \\ x \in G_2 & \text{若 } x \geqslant d \\ \text{待判} & \text{若 } c < x < d \end{cases}$$



## 2、当总体的协差阵已知，且不相等

$$\begin{cases} x \in G_1, & \text{如} d^2(x, G_1) < d^2(x, G_2), \\ x \in G_2, & \text{如} d^2(x, G_2) < d^2(x, G_1) \\ \text{待判}, & \text{如} d^2(x, G_1) = d^2(x, G_2) \end{cases}$$

$$W(x) = d^2(x, G_2) - d^2(x, G_1)$$

$$= (x - \mu_2)' \Sigma_2^{-1} (x - \mu_2) - (x - \mu_1)' \Sigma_1^{-1} (x - \mu_1)$$

判别准则：

$$\begin{cases} x \in G_1, & \text{如} W(x) > 0, \\ x \in G_2, & \text{如} W(x) < 0。 \\ \text{待判}, & \text{如} W(x) = 0 \end{cases}$$

特别地，当p=1时，若两个总体分别为$N\left(\mu_1,\sigma_1^2\right)$和 $N\left(\mu_2,\sigma_2^2\right)$

则判别函数为 $W(x)=\dfrac{(x-\mu_2)^2}{\sigma_2^2}-\dfrac{(x-\mu_1)^2}{\sigma_1^2}$

当 $\mu_1<x<\mu_2$

$$W(x)=\frac{\mu_2-x}{\sigma_2}-\frac{x-\mu_1}{\sigma_1}=\frac{\sigma_1\mu_2+\sigma_2\mu_1-x(\sigma_1+\sigma_2)}{\sigma_1\sigma_2}$$

$$=-\frac{\sigma_1+\sigma_2}{\sigma_1\sigma_2}\left(x-\frac{\sigma_1\mu_2+\sigma_2\mu_1}{\sigma_1+\sigma_2}\right)$$

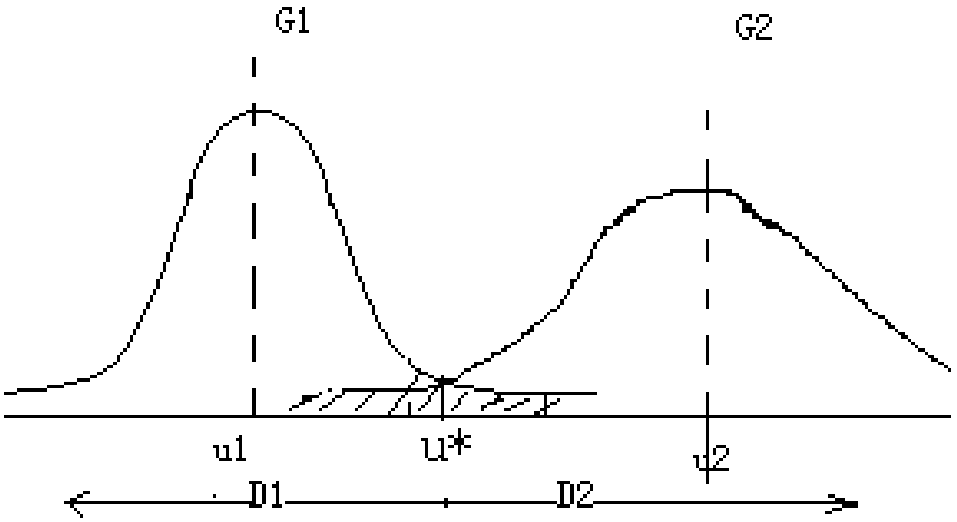
令$\mu^*=\dfrac{\sigma_1\mu_2+\sigma_2\mu_1}{\sigma_1+\sigma_2}$

判别规则为：$\begin{cases}\text{当}x<\mu^*\text{时}\quad，\quad\text{则}x\in G_1\\ \text{当}x>\mu^*\text{时}\quad，\quad\text{则}x\in G_2\end{cases}$

$$\mu^* - \mu_1 = \frac{\frac{\sigma_2\mu_1 + \sigma_1\mu_2}{\sigma_1 + \sigma_2} - \mu_1}{\sigma_1} = \frac{\mu_2 - \mu_1}{\sigma_1 + \sigma_2}$$

$$\mu_2 - \mu^* = \frac{\mu_2 - \frac{\sigma_2\mu_1 + \sigma_1\mu_2}{\sigma_1 + \sigma_2}}{\sigma_2} = \frac{\mu_2 - \mu_1}{\sigma_1 + \sigma_2}$$

$\mu^*$ 到两个总体的马氏距离相等



**例 4.2** 某企业生产新式大衣，将新产品的样品分寄给9个城市百货公司的进货员，并附寄调查意见表征求对新产品的评价，评价分质量、款式、颜色三个方面，以10分制评分。结果5位喜欢，4位不喜欢，具体评价见表4-2。

**表 4-2　产品评价表**

| 组别 | 序号 | 产品特征 | | |
|---|---|---|---|---|
| | | 质量($x_1$) | 款式($x_2$) | 颜色($x_3$) |
| 喜欢组 | 1 | 8 | 9.5 | 7 |
| | 2 | 9 | 8.5 | 6 |
| | 3 | 7 | 8.0 | 9 |
| | 4 | 10 | 7.5 | 8.5 |
| | 5 | 8 | 6.5 | 7 |
| 不喜欢组 | 1 | 6 | 3 | 5.5 |
| | 2 | 3 | 4 | 3.5 |
| | 3 | 4 | 2 | 5 |
| | 4 | 3 | 5 | 4 |

(1) 求两类样本均值。

$$\bar{x}^{(1)}=\begin{bmatrix}8.4\\8.0\\7.5\end{bmatrix}\qquad \bar{x}^{(2)}=\begin{bmatrix}4.0\\3.5\\4.5\end{bmatrix}\qquad \bar{x}^{(1)}-\bar{x}^{(2)}=\begin{bmatrix}4.4\\4.5\\3\end{bmatrix}\qquad \frac{\bar{x}^{(1)}+\bar{x}^{(2)}}{2}=\begin{bmatrix}6.2\\5.75\\6\end{bmatrix}$$

(2) 计算样本协方差矩阵，从而求出$\hat{\boldsymbol{\Sigma}}$及$\hat{\boldsymbol{\Sigma}}^{-1}$。

$$\boldsymbol{S}_1=\frac{1}{4}\begin{bmatrix}5.2&-0.5&-1\\-0.5&5&-1.25\\-1&-1.25&6\end{bmatrix}=\begin{bmatrix}1.3&-0.125&-0.25\\-0.125&1.25&-0.312\,5\\-0.25&-0.312\,5&1.5\end{bmatrix}$$

$$\boldsymbol{S}_2=\frac{1}{3}\begin{bmatrix}6 & -3 & 3.5\\ -3 & 5 & -2.5\\ 3.5 & -2.5 & 2.5\end{bmatrix}=\begin{bmatrix}2 & -1 & 1.167\\ -1 & 1.667 & -0.833\\ 1.167 & -0.833 & 0.833\end{bmatrix}$$

$$\text{故}\hat{\boldsymbol{\Sigma}}=\frac{4\boldsymbol{S}_1+3\boldsymbol{S}_2}{5+4-2}=\frac{1}{7}\begin{bmatrix}11.2 & -3.5 & 2.5\\ -3.5 & 10 & -3.75\\ 2.5 & -3.75 & 8.5\end{bmatrix}$$

$$\hat{\boldsymbol{\Sigma}}^{-1}=\begin{bmatrix}0.716\,0 & 0.205\,6 & -0.119\,8\\ 0.205\,6 & 0.897\,8 & 0.335\,6\\ -0.119\,8 & 0.335\,6 & 1.006\,9\end{bmatrix}$$

(3) 求线性判别函数。

$$W(x)=(\overline{x}^{(1)}-\overline{x}^{(2)})'\hat{\boldsymbol{\Sigma}}^{-1}\left(x-\frac{\overline{x}^{(1)}+\overline{x}^{(2)}}{2}\right)$$

$$=3.716\,6x_1+5.952\,0x_2+4.004\,8x_3-81.295\,6$$

(4) 对已知类别的样品判别分类。

对已知类别的样品(通常成为训练样品)用线性判别函数进行判别归类，结果如表 4-3 所示。回代率为 100%，全部判对。

**表 4-3　回判结果**

| 样品 | 判别函数 $W(x)$ 的值 | 原类号 | 判归类别 |
|---|---|---|---|
| 1 | 33.01 | 1 | 1 |
| 2 | 26.77 | 1 | 1 |
| 3 | 28.38 | 1 | 1 |
| 4 | 34.55 | 1 | 1 |
| 5 | 15.16 | 1 | 1 |
| 6 | −19.11 | 2 | 2 |
| 7 | −21.24 | 2 | 2 |
| 8 | −32.32 | 2 | 2 |
| 9 | −24.37 | 2 | 2 |

(5) 对待判样品判别归类。

如果有一潜在顾客，他对新产品的质量、款式、颜色分别评价 6、8、8，其评价值为 $W(x)=3.716\,6\times6+5.952\,0\times8+4.004\,8\times8-81.295\,6=20.66>0$，故他属喜欢组。

# 多总体的距离判别法

多个总体时，在思想方法上只是两总体距离判别的推广。

设有$k$个总体$G_1$，$G_2$，…，$G_k$，它们的均值和协方差矩阵分别为$\mu_i$，$\boldsymbol{\Sigma}_i$，$i=1$，$2$，…，$k$。

从每个总体$G_i$中抽取$n_i$个样品，$i=1$，$2$，…，$k$，每个样品观测$p$个指标。今对任一样品$x=(x_1, x_2, \cdots, x_p)'$，问$x$属于哪一类？

# 多总体的距离判别法

## 1. 各总体协差阵相等时 ($\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k = \Sigma$)

$$d^2(x, G_i) = (x - \mu_i)'\Sigma^{-1}(x - \mu_i)$$
$$= x'\Sigma^{-1}x - 2x'\Sigma^{-1}\mu_i + \mu_i'\Sigma^{-1}\mu_i$$
$$f_i(x) = -2x'\Sigma^{-1}\mu_i + \mu_i'\Sigma^{-1}\mu_i$$

上式中的第一项 $x'\Sigma^{-1}x$ 与 i 无关，则舍去，得一个等价的函数

# 多总体的距离判别法

将上式中提-2，得

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

则距离判别法的判别函数为：

令 $$f_i(x) = (x'\Sigma^{-1}\mu_i - \frac{1}{2}\mu_i'\Sigma^{-1}\mu_i)$$

判别规则为 $f_l(x) = \max\limits_{1\le i\le k} f_i(x)$，则 $x \in G_l$

# 多总体的距离判别法

当 $\mu_1, \mu_2, \cdots, \mu_k, \boldsymbol{\Sigma}$ 未知时，可通过样本来估计。设从总体 $G_i$ 中抽取的样本为 $x_1^{(i)}, x_2^{(i)}, \cdots, x_{n_i}^{(i)}, i=1, 2, \cdots, k$，则

$$\hat{\mu}_i = \frac{1}{n_i}\sum_{l=1}^{n_i} x_l^{(i)} \hat{=} \overline{x}^{(i)}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{\sum_{i=1}^{k}(n_i-1)\boldsymbol{S}_i}{\sum_{i=1}^{k} n_i - k} = \frac{(n_1-1)\boldsymbol{S}_1 + (n_2-1)\boldsymbol{S}_2 + \cdots + (n_k-1)\boldsymbol{S}_k}{n_1+n_2+\cdots+n_k-k}$$

其中，$\boldsymbol{S}_i = \frac{1}{n_i-1}\sum_{l=1}^{n_i}(x_l^{(i)} - \overline{x}^{(i)})(x_l^{(i)} - \overline{x}^{(i)})'$ 为总体 $G_i$ 的样本协方差矩阵。

# 多总体的距离判别法

**2. 各总体协差阵不等情形($\Sigma_1$, $\Sigma_2$, …, $\Sigma_k$ 不相等时)**

此时，判别函数为：$W_{ij}(x)=(x-\mu_j)'\sum_j^{-1}(x-\mu_j)-(x-\mu_i)'\sum_j^{-1}(x-\mu_i)$。

相应的判别规则为

$$\begin{cases} x\in G_i & \text{当 } W_{ij}(x)>0,\ i\neq j \\ \text{待判} & \text{当 } W_{ij}(x)=0 \end{cases}$$

当 $\mu_i$, $\Sigma_i$ $(i=1, 2, \cdots, k)$ 未知时，用 $\mu_i$, $\Sigma_i$ 的估计量代替，$\hat{\mu}_i=\overline{x}^{(i)}$, $\hat{\Sigma}_i=S_i$，即

$$W_{ij}(x)=(x-\overline{x}^{(j)})'S_j^{-1}(x-\overline{x}^{(j)})-(x-\overline{x}^{(i)})S_i^{-1}(x-\overline{x}^{(i)})$$

# 内容

(1) 基本思想

(2) 距离判别法

(3) 贝叶斯判别

(4) 费歇判别

(5) 例子

距离判别只要求知道总体的数字特征，不涉及总体的分布函数，当总体均值和协方差未知时，就用样本的均值和协方差矩阵来估计。距离判别方法简单实用，但没有考虑到每个总体出现的机会大小，即先验概率，没有考虑到错判的损失。贝叶斯判别法正是为了解决这两个问题提出的判别分析方法。

贝叶斯公式是一个我们熟知的公式

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{\sum_i P(A \mid B_i)P(B_i)}$$

# 朴素贝叶斯分类例子

- 办公室新来了一个雇员小王，大家都在猜测小王是好人还是坏人？
- 假设一个人是好人或坏人的概率均为0.5。
- 好人做好事的概率为0.9，坏人做好事的概率为0.2。
- 一天，小王做了一件好事，那么，小王是好人还是坏人？

# 判别准则1：后验概率最大

设有总体 $G_i(i=1,2,\cdots,k)$ ，$G_i$ 具有概率密度函数 $f_i(x)$ 。并且根据以往的统计分析，知道 $G_i$ 出现的概率为 $q_i$ （先验概率）。即当样本 $x_0$ 发生时，求它属于某类的概率。由贝叶斯公式计算后验概率，有：

$$P(G_i \mid x_0)=\frac{q_i f_i(x_0)}{\sum_j q_j f_j(x_0)}$$

若 $P(G_l \mid x_0)=\frac{q_l f_l(x_0)}{\sum_j q_j f_j(x_0)}=\max_{1\le i\le k}\frac{q_i f_i(x_0)}{\sum_j q_j f_j(x_0)}$

则 $x_0$ 判给 $G_l$。在正态的假定下，$f_i(x)$ 为正态分布的密度函数。

## 特别地，总体服从正态分布的情形

$$q_l f_l(x_0) = \max_{1\le i\le k} q_i f_i(x_0), \quad \text{则} x_0 \text{判给} G_l。$$

$$\text{若} f_i(x) = \frac{1}{(2\pi|\Sigma_i|)^{1/2}} \exp[-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)]$$

$$\text{则}, q_i f_i(x) = q_i \frac{1}{(2\pi|\Sigma_i|)^{1/2}} \exp[-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)]$$

上式两边取对数并去掉与i无关的项，则等价的判别函数为：

$$z_i(x) = \ln(q_i f_i(\mathbf{x}))$$

$$= \ln q_i - \frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(x-\mu^{(i)})'\Sigma_i^{-1}(x-\mu_i)]$$

$$z_i(x) = \ln(q_i f_i(\mathbf{x})) = \ln q_i - \frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(x-\mu^{(i)})'\Sigma_i^{-1}(x-\mu_i)]$$

问题转化为若 $Z_l(x) = \max_{1\le i\le k}[Z_i(x)]$，则判 $x \in G_l$ 。

当协方差阵相等 $\sum_1 = \sum_2 = \ldots = \sum_k$

当先验概率相等，$q_1 = \cdots = q_k = \frac{1}{k}$

完全成为距离判别法 。

# 判别准则2：错判损失最小

设有总体 $G_i(i=1,2,\cdots,k)$ ，$G_i$ 具有概率密度函数 $f_i(x)$ 。并且根据以往的统计分析，知道 $G_i$ 出现的概率为 $q_i$ （先验概率）。即当样本 $x_0$ 发生时，求它是否属于某类，需要考虑其错判损失。

$P(G_i|x_0)=\dfrac{q_i f_i(x_0)}{\sum_j q_j f_j(x_0)}$ 后验概率判别式为

$$P(G_l|x_0)=\frac{q_l f_l(x_0)}{\sum_j q_j f_j(x_0)}=\max_{1\le i\le k}\frac{q_i f_i(x_0)}{\sum_j q_j f_j(x_0)}=\max_{1\le i\le k} q_i f_i(x_0)$$

设$c(i|j)$为样本来自$G_j$但误判为的损失$G_i$（风险）, $i, j = 1,2,\ldots,k$

$$P(G_l|x_0)=\min_{1\le i,j\le k} q_i f_i(x_0)c(i|j)$$

# 判别准则2：错判损失最小

设有总体 $G_i(i=1,2,\cdots,k)$ ，$G_i$ 具有概率密度函数 $f_i(x)$ 。并且根据以往的统计分析，知道 $G_i$ 出现的概率为 $q_i$（先验概率）。

设$D_1, D_2, \ldots, D_k$为样本$D$的一个划分，互不相交。其最优划分满足下述 **ECM(expected cost of misclassification)**最小。

$$ECM(D_1, D_2, \ldots, D_k) = \sum_{i=1}^{k} q_i \sum_{j=1}^{k} q_j f_i(x_0) c(i|j)$$

# 内容

(1) 基本思想

(2) 距离判别法

(3) 贝叶斯判别

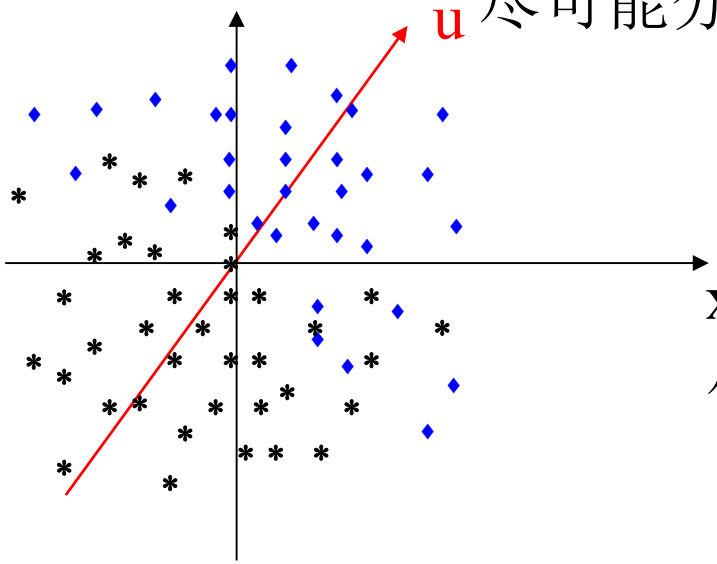
(4) 费歇判别

(5) 例子

# 费歇判别法

## 两个总体的费歇（Fisher）判别法

费歇判别的基本思想是投影，将k组p维数据投影到某一个方向，使其投影的组与组之间尽可能地分开。




能使总体单位
尽可能分开的方向

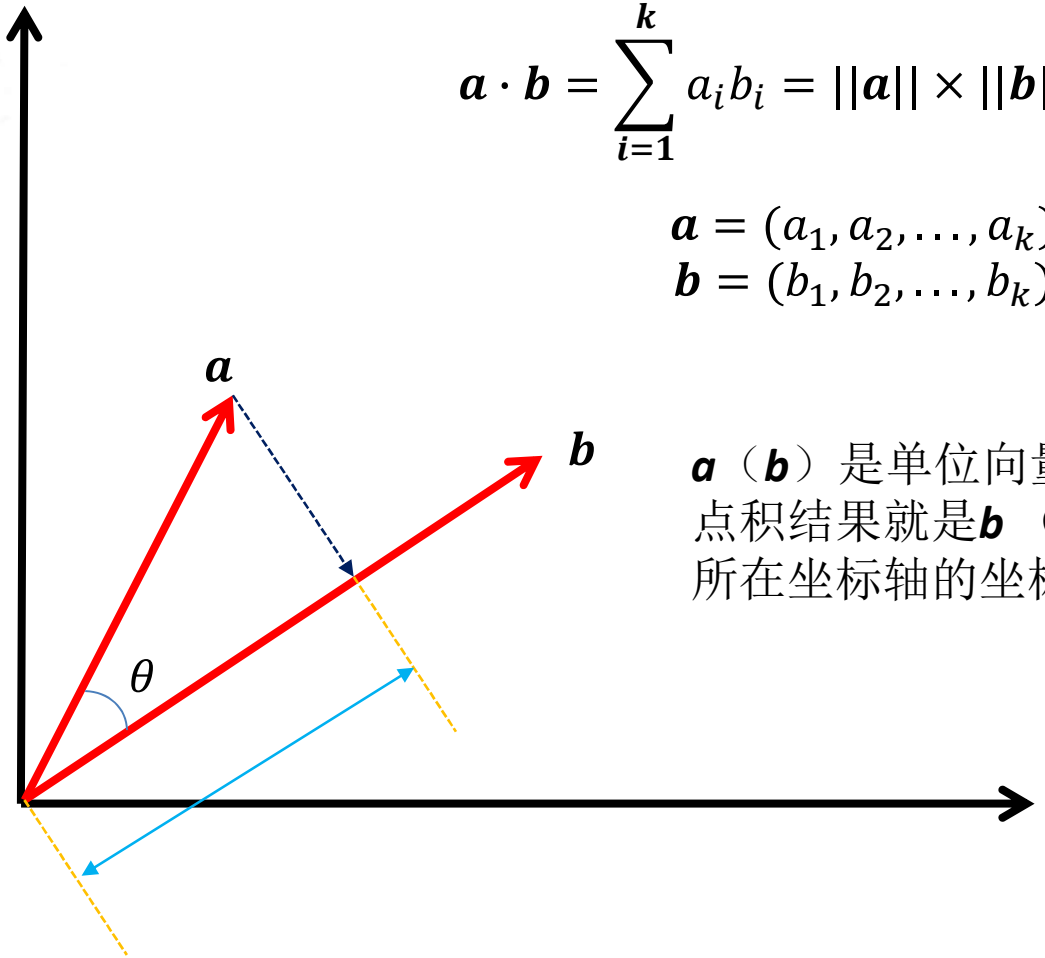
不能使总体
尽可能分开的方

旋转坐标轴至总体单位尽可能分开的方向，此时分类变量被简化为一个

# 向量点积的几何意义

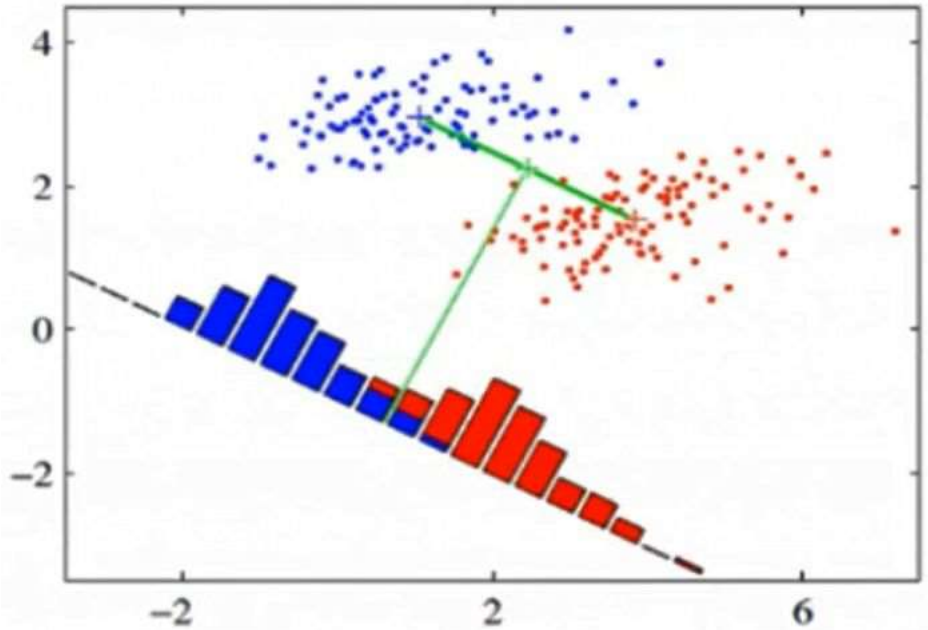
$$\boldsymbol{a}\cdot\boldsymbol{b}=\sum_{i=1}^{k}a_ib_i=||\boldsymbol{a}||\times||\boldsymbol{b}||\times\boldsymbol{cos}(\boldsymbol{\theta})$$

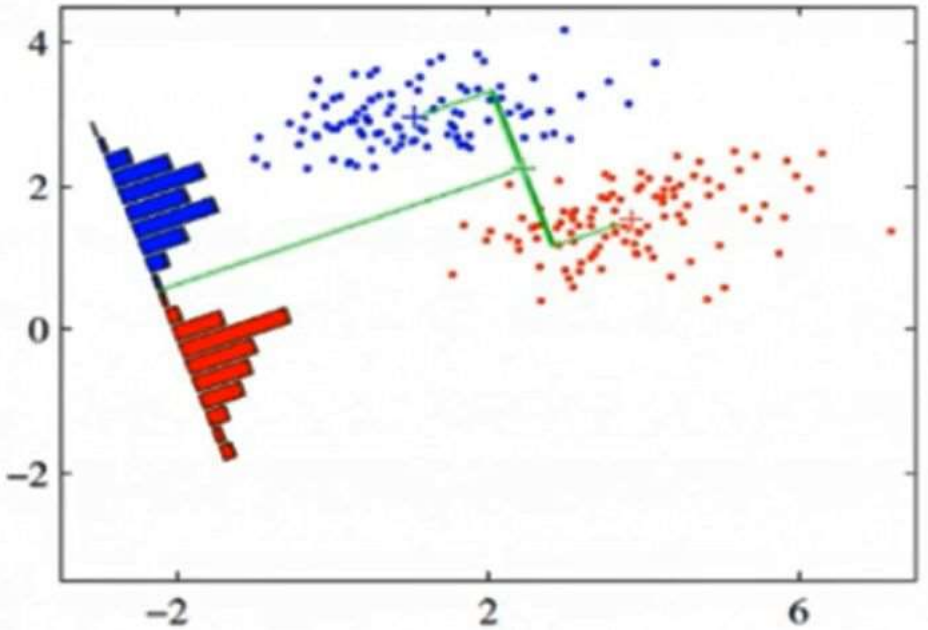
$$\boldsymbol{a}=(a_1,a_2,\ldots,a_k)$$
$$\boldsymbol{b}=(b_1,b_2,\ldots,b_k)$$




$\boldsymbol{a}$（$\boldsymbol{b}$）是单位向量，那么：
点积结果就是$\boldsymbol{b}$（$\boldsymbol{a}$）在$\boldsymbol{a}$（$\boldsymbol{b}$）所在坐标轴的坐标值。

# 费歇尔判别基本原理




$$\mathrm{dis}(C_1, C_2) = ||\tilde{\boldsymbol{\mu}}_1 - \tilde{\boldsymbol{\mu}}_2||^2$$

$$\tilde{\boldsymbol{\mu}}_1 = \boldsymbol{\omega}^T\boldsymbol{\mu}_1 \quad \tilde{\boldsymbol{\mu}}_2 = \boldsymbol{\omega}^T\boldsymbol{\mu}_2$$

$$\Rightarrow \begin{cases} \max\limits_{\boldsymbol{\omega}} ||\boldsymbol{\omega}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)||^2 \\ s.t.\ \boldsymbol{\omega}^T\boldsymbol{\omega} = 1 \end{cases} \Rightarrow \max_{\boldsymbol{\omega}} J(\boldsymbol{\omega}) = \frac{||\boldsymbol{\omega}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)||^2}{D_1 + D_2}$$

$$\boldsymbol{\mu}_1 = \frac{\sum_{x \in C_1} \boldsymbol{x}}{n_1} \quad \boldsymbol{\mu}_2 = \frac{\sum_{x \in C_2} \boldsymbol{x}}{n_2}$$

$$D_1 = \sum_{x \in C_1} (\boldsymbol{\omega}^T\boldsymbol{x} - \boldsymbol{\omega}^T\boldsymbol{\mu}_1)^2 = \sum_{x \in C_1} \boldsymbol{\omega}^T(\boldsymbol{x} - \boldsymbol{\mu}_1)(\boldsymbol{x} - \boldsymbol{\mu}_1)^T\boldsymbol{\omega}$$

$$D_2 = \sum_{x \in C_2} \boldsymbol{\omega}^T(\boldsymbol{x} - \boldsymbol{\mu}_2)(\boldsymbol{x} - \boldsymbol{\mu}_2)^T\boldsymbol{\omega}$$

# 费歇尔判别基本原理

$$J(\boldsymbol{\omega}) = \frac{||\boldsymbol{\omega}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)||^2}{D_1+D_2} = \frac{||\boldsymbol{\omega}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)||^2}{\sum_{x\in C_1}\boldsymbol{\omega}^T(\boldsymbol{x}-\boldsymbol{\mu}_1)(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\omega}+\sum_{x\in C_2}\boldsymbol{\omega}^T(\boldsymbol{x}-\boldsymbol{\mu}_2)(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\omega}} =$$

$$\frac{||\boldsymbol{\omega}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)||^2}{\sum_{x\in C_i}\boldsymbol{\omega}^T(\boldsymbol{x}-\boldsymbol{\mu}_i)(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\omega}} = \frac{||\boldsymbol{\omega}^T\boldsymbol{d}||^2}{\boldsymbol{\omega}^T\boldsymbol{E}\boldsymbol{\omega}} = \frac{\boldsymbol{Q}}{\boldsymbol{R}}$$

$$\ln\big(J(\boldsymbol{\omega})\big) = \ln(Q) - \ln(R)$$

令$\frac{\partial \ln(J(\boldsymbol{\omega}))}{\partial \boldsymbol{\omega}} = \frac{1}{Q}\frac{\partial Q}{\partial \boldsymbol{\omega}} - \frac{1}{R}\frac{\partial R}{\partial \boldsymbol{\omega}}=0$，则$\frac{2}{J(\boldsymbol{\omega})}(\boldsymbol{\omega}^T\boldsymbol{d})\boldsymbol{d} = 2\boldsymbol{E}\boldsymbol{\omega}$

得到$\boldsymbol{\omega} = \boldsymbol{E}^{-1}\boldsymbol{d}$

注意 $\boldsymbol{\omega}^T\boldsymbol{d}$ 是一个值（标量），它的大小并不影响的求解 $\boldsymbol{\omega}$，$\boldsymbol{\omega}$的关键是“方向”

$$\mathrm{dis}(C_1, C_2) = ||\tilde{\boldsymbol{\mu}}_1 - \tilde{\boldsymbol{\mu}}_2||^2$$

$$\tilde{\boldsymbol{\mu}}_1 = \boldsymbol{\omega}^T\boldsymbol{\mu}_1 \quad \tilde{\boldsymbol{\mu}}_2 = \boldsymbol{\omega}^T\boldsymbol{\mu}_2$$

$$\Rightarrow \begin{cases} \max\limits_{\boldsymbol{\omega}} ||\boldsymbol{\omega}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)||^2 \\ s.t.\ \boldsymbol{\omega}^T\boldsymbol{\omega} = 1 \end{cases} \Rightarrow \max_{\boldsymbol{\omega}} J(\boldsymbol{\omega}) = \frac{||\boldsymbol{\omega}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)||^2}{D_1 + D_2}$$

$$\boldsymbol{\mu}_1 = \frac{\sum_{x\in C_1}\boldsymbol{x}}{n_1} \quad \boldsymbol{\mu}_2 = \frac{\sum_{x\in C_2}\boldsymbol{x}}{n_2}$$

$$D_1 = \sum_{x\in C_1}(\boldsymbol{\omega}^T\boldsymbol{x} - \boldsymbol{\omega}^T\boldsymbol{\mu}_1)^2 = \sum_{x\in C_1}\boldsymbol{\omega}^T(\boldsymbol{x}-\boldsymbol{\mu}_1)(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\omega}$$

$$D_2 = \sum_{x\in C_2}\boldsymbol{\omega}^T(\boldsymbol{x}-\boldsymbol{\mu}_2)(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\omega}$$

# 费歇判别的基本思想

**Fisher**判别法由**Fisher**在1936年提出，是根据方差分析的思想建立起来的一种能较好区分各个总体的线性判别法，该判别方法对总体的分布不做任何要求。

从两个总体中抽取具有p个指标的样品观测数据，借助于方差分析的思想构造一个线性判别函数：

$$y = c_1 x_1 + c_2 x_2 + \cdots + c_p x_p = \alpha' x_i$$

这里的$\alpha = (c_1, c_2, c_3 ..., c_p)$称为判别系数

系数 $c_1, c_2, \cdots, c_p$ 确定的原则

使组间离差平方和最大，而组内离差平方和最小。

假设我们可以得到一个线性判别函数：

$$y = c_1 x_1 + c_2 x_2 + \cdots + c_p x_p$$

我们把两个总体的样品数据代入上面的判别式

$$y_i^{(1)} = c_1 x_{i1}^{(1)} + c_2 x_{i2}^{(1)} + \cdots + c_p x_{ip}^{(1)} \qquad i = 1,2,\cdots,n_1$$

$$y_i^{(2)} = c_1 x_{i1}^{(2)} + c_2 x_{i2}^{(2)} + \cdots + c_p x_{ip}^{(2)} \qquad i = 1,2,\cdots,n_2$$

$$\bar{y}^{(1)} = \frac{1}{n}\sum_{i=1}^{n_1} y_i^{(1)} \qquad \bar{y}^{(1)} = c_1 \bar{x}_1^{(1)} + c_2 \bar{x}_2^{(1)} + \cdots + c_p \bar{x}_p^{(1)} = \sum_{k=1}^{p} c_k \bar{x}_k^{(1)}$$

$$\bar{y}^{(2)} = \frac{1}{n}\sum_{i=1}^{n_2} y_i^{(2)} \qquad \bar{y}^{(2)} = c_1 \bar{x}_1^{(2)} + c_2 \bar{x}_2^{(2)} + \cdots + c_p \bar{x}_p^{(2)} = \sum_{k=1}^{p} c_k \bar{x}_k^{(2)}$$

为了使判别函数能够很好地区分来自不同总体$G_1$和$G_2$的样品，自然希望：

(1) $\bar{y}^{(1)}$和 $\bar{y}^{(2)}$ 的差异越大越好

（2）来自同一总体的各个样品之间的差异越小越好。

即 $y_i^{(1)}(i=1,2,\cdots,n_1)$的离差平方和 $\sum_{i=1}^{n_1}\left(y_i^{(1)}-\bar{y}^{(1)}\right)^2$ 越小越好

即 $y_i^{(2)}(i=1,2,\cdots,n_2)$的离差平方和 $\sum_{i=1}^{n_2}\left(y_i^{(2)}-\bar{y}^{(2)}\right)^2$ 越小越好

$$Q=\left(\bar{y}^{(1)}-\bar{y}^{(2)}\right)^2 \rightarrow \max$$

$$R=\sum_{i=1}^{n_1}\left(y_i^{(1)}-\bar{y}^{(1)}\right)^2+\sum_{i=1}^{n_2}\left(y_i^{(2)}-\bar{y}^{(2)}\right)^2 \rightarrow \min$$

$$I=\frac{Q}{R} \rightarrow \max$$

$$I = \frac{Q}{R} \rightarrow \max$$

$$\ell nI = \ell nQ - \ell nR \rightarrow \max$$

令 $\dfrac{\partial \ell nI}{\partial c_k} = 0 \quad (k = 1,2,\cdots,p)$

由于 $\dfrac{\partial \ell nI}{\partial c_k} = \dfrac{1}{Q}\dfrac{\partial Q}{\partial c_k} - \dfrac{1}{R}\dfrac{\partial R}{\partial c_k} = 0$

故 $\dfrac{1}{I}\dfrac{\partial Q}{\partial c_k} = \dfrac{\partial R}{\partial c_k}$

$$Q=\left(\bar{y}^{(1)}-\bar{y}^{(2)}\right)^2=\left[\sum_{k=1}^{p}c_k\left(\bar{x}_k^{(1)}-\bar{x}_k^{(2)}\right)\right]^2 \triangleq\left(\sum_{k=1}^{p}c_k d_k\right)^2$$

其中$d_k=\bar{x}_k^{(1)}-\bar{x}_k^{(2)}\quad k=1,2,\cdots,p$

即 $\begin{pmatrix}d_1\\d_2\\\vdots\\d_p\end{pmatrix}=\begin{pmatrix}\bar{x}_1^{(1)}-\bar{x}_1^{(2)}\\\bar{x}_2^{(1)}-\bar{x}_2^{(2)}\\\vdots\\\bar{x}_p^{(1)}-\bar{x}_p^{(2)}\end{pmatrix}=\left(\bar{x}^{(1)}-\bar{x}^{(2)}\right)$ 为两类总体的样本均值差

$$\frac{\partial Q}{\partial c_k}=2\left(\sum_{l=1}^{p}c_l d_l\right)d_k$$

$$R=\sum_{i=1}^{n_1}\left(y_i^{(1)}-\overline{y}^{(1)}\right)^2+\sum_{i=1}^{n_2}\left(y_i^{(2)}-\overline{y}^{(2)}\right)^2$$

$$=\sum_{i=1}^{n_1}\left[\sum_{k=1}^{p}c_k\left(x_{ik}^{(1)}-\overline{x}_k^{(1)}\right)\right]^2+\sum_{i=1}^{n_2}\left[\sum_{k=1}^{p}c_k\left(x_{ik}^{(2)}-\overline{x}_k^{(2)}\right)\right]^2$$

$$=\sum_{i=1}^{n_1}\left[\sum_{k=1}^{p}c_k\left(x_{ik}^{(1)}-\overline{x}_k^{(1)}\right)\cdot\sum_{l=1}^{p}c_l\left(x_{il}^{(1)}-\overline{x}_l^{(1)}\right)\right]$$

$$+\sum_{i=1}^{n_2}\left[\sum_{k=1}^{p}c_k\left(x_{ik}^{(2)}-\overline{x}_k^{(2)}\right)\cdot\sum_{l=1}^{p}c_l\left(x_{il}^{(2)}-\overline{x}_l^{(2)}\right)\right]$$

$$=\sum_{k=1}^{p}\sum_{l=1}^{p}c_kc_l\left[\sum_{i=1}^{n_1}\left(x_{ik}^{(1)}-\overline{x}_k^{(1)}\right)\left(x_{il}^{(1)}-\overline{x}_l^{(1)}\right)+\sum_{i=1}^{n_2}\left(x_{ik}^{(2)}-\overline{x}_k^{(2)}\right)\left(x_{il}^{(2)}-\overline{x}_l^{(2)}\right)\right]$$

$$\hat{=}\sum_{k=1}^{p}\sum_{l=1}^{p}c_kc_ls_{kl}$$

其中

$$s_{kl}=\sum_{i=1}^{n_1}\left(x_{ik}^{(1)}-\overline{x}_k^{(1)}\right)\left(x_{il}^{(1)}-\overline{x}_l^{(1)}\right)+\sum_{i=1}^{n_2}\left(x_{ik}^{(2)}-\overline{x}_k^{(2)}\right)\left(x_{il}^{(2)}-\overline{x}_l^{(2)}\right)$$

$$\frac{\partial R}{\partial c_k}=2\left(\sum_{l=1}^{p}c_l s_{kl}\right)=2c_1 s_{k1}+2c_2 s_{k2}+\cdots+2c_p s_{kp}$$

$$(k=1,2,\cdots,p)$$

$$\frac{2}{I}\left(\sum_{l=1}^{p}c_l d_l\right)d_k=2\sum_{l=1}^{p}c_l s_{kl}\qquad k=1,2,\cdots,p$$

令 $\beta=\frac{1}{I}\sum_{l=1}^{p}c_l d_l$

$\beta$ 是常数因子，不依赖于 $k$

它对方程组只起共同扩大倍数的作用，不影响判别结果，不妨取 $\beta = 1$

于是得到 $\sum_{l=1}^{p} c_l s_{kl} = d_k \qquad k = 1,2,\cdots,p$

$$s_{k1}c_1 + s_{k2}c_2 + \cdots + s_{kp}c_p = d_k \qquad k = 1,2,\cdots,p$$

$$\begin{cases} s_{11}c_1 + s_{12}c_2 + \cdots + s_{1p}c_p = d_1 \\ s_{21}c_1 + s_{22}c_2 + \cdots + s_{2p}c_p = d_2 \\ \cdots \\ s_{p1}c_1 + s_{p2}c_2 + \cdots + s_{pp}c_p = d_p \end{cases}$$

用矩阵表示：

$$\begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix} \cdot \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_p \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_p \end{pmatrix}$$

因此得到

$$\begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_p \end{pmatrix} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}^{-1} \cdot \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_p \end{pmatrix}$$

$$\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ \vdots \\ c_p \end{bmatrix} = \begin{bmatrix} s_{11} & s_{12} & \cdots & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & \cdots & s_{2p} \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ \vdots & \vdots & & & \vdots \\ s_{p1} & s_{p2} & \cdots & \cdots & s_{pp} \end{bmatrix}^{-1} \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ \vdots \\ d_p \end{bmatrix} = E^{-1} \cdot d$$

$$\begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_p \end{bmatrix} = \begin{bmatrix} \overline{x}_1^{(1)} - \overline{x}_1^{(2)} \\ \overline{x}_2^{(1)} - \overline{x}_2^{(2)} \\ \vdots \\ \overline{x}_p^{(1)} - \overline{x}_p^{(2)} \end{bmatrix}$$

两总体的
积差阵之和

称 $y(x) = c_1 x_1 + \cdots + c_p x_p$ 为判别函数.

# 两总体fisher判别的基本步骤

（1）建立判别函数
（2）计算判别临界值
（3）建立判别准则

## 判别函数

根据前面判别函数的导出，得到了 $y(x)=\alpha' x$

$$\alpha = \begin{bmatrix} c_1 \\ c_2 \\ \cdots \\ c_p \end{bmatrix} E^{-1}(\overline{x_1}-\overline{x_2})$$

这里的E为两总体的样本积差阵之和

# 判别临界值

定义临界点为 $y_c = \begin{cases} \dfrac{\bar{y}^{(1)} + \bar{y}^{(2)}}{2} & \text{两总体方差相等时} \\ \dfrac{\hat{\sigma}_2 \bar{y}^{(1)} + \hat{\sigma}_1 \bar{y}^{(2)}}{\hat{\sigma}_2 + \hat{\sigma}_1} & \text{两总体方差不相等时} \end{cases}$

其中

$$\hat{\sigma}_1 = \sqrt{\frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left( y_i^{(1)} - \bar{y}^{(1)} \right)^2}$$

$$\hat{\sigma}_2 = \sqrt{\frac{1}{n_2 - 1} \sum_{i=1}^{n_2} \left( y_i^{(2)} - \bar{y}^{(2)} \right)^2}$$

# 判别准则

不妨假定 $\bar{y}^{(1)} > \bar{y}^{(2)}$ ，则判别准则为：

$$\begin{cases} 若y(x) > y_c, 则x \in G_1 \\ 若y(x) < y_c, 则x \in G_2 \\ 若y(x) = y_c, 待判 \end{cases}$$

**例 4.3** 某外贸公司为推销某一新产品，为保险起见，在新产品大量上市前将该产品的样品寄往12个国家的进口代理商，并附意见调查表，要求对该产品给予评估，评估的因素有式样、包装及耐久性三项。评分表用10分制，最后要求说明是否愿意购买。12个代理商的调查结果如表4-4所示。若第13个国家的进口代理商评分为(9，5，8)，问该代理商是否愿意购买此产品。

**表4-4　产品调查结果**

| 组别 | 序号 | 产品特征 | | |
|---|---|---|---|---|
| | | 质量($x_1$) | 款式($x_2$) | 颜色($x_3$) |
| 购买组 | 1 | 9 | 8 | 7 |
| | 2 | 7 | 6 | 6 |
| | 3 | 10 | 7 | 8 |
| | 4 | 8 | 4 | 5 |
| | 5 | 9 | 9 | 3 |
| | 6 | 8 | 6 | 7 |
| | 7 | 7 | 5 | 6 |
| 非购买组 | 1 | 6 | 4 | 4 |
| | 2 | 3 | 6 | 6 |
| | 3 | 6 | 3 | 3 |
| | 4 | 2 | 4 | 5 |
| | 5 | 1 | 2 | 2 |

**解**：(1) 求两总体的样本均值$\overline{x}^{(1)}$和$\overline{x}^{(2)}$。

$$\overline{x}^{(1)}=\begin{bmatrix}8.29\\6.43\\6.00\end{bmatrix},\ \overline{x}^{(2)}=\begin{bmatrix}\overline{x}_1^{(2)}\\\overline{x}_2^{(2)}\\\overline{x}_3^{(2)}\end{bmatrix}=\begin{bmatrix}3.20\\3.80\\4.00\end{bmatrix}$$

(2) 求两总体样本均值之差。

$$\boldsymbol{d}=\overline{x}^{(1)}-\overline{x}^{(2)}=\begin{bmatrix}5.09\\2.63\\2.00\end{bmatrix}$$

(3) 求两总体的样本离差平方和矩阵$\boldsymbol{E}$。

先求各$s_{kl}$，如

$$s_{11}=\sum_{i=1}^{7}(x_{1i}^{(1)}-\overline{x}_1^{(1)})^2+\sum_{i=1}^{5}(x_{1i}^{(2)}-\overline{x}_1^{(2)})^2=22.228\ 57$$

$$s_{12}=\sum_{i=1}^{7}(x_{1i}^{(1)}-\overline{x}_1^{(1)})(x_{2i}^{(1)}-\overline{x}_2^{(1)})+\sum_{i=1}^{5}(x_{1i}^{(2)}-\overline{x}_1^{(2)})(x_{2i}^{(2)}-\overline{x}_2^{(2)})=8.342\ 88$$

$$\boldsymbol{E}=\begin{bmatrix}22.228\ 57 & 8.342\ 88 & 2\\ & 26.514\ 27 & 6\\ & & 26\end{bmatrix}$$

求 $\boldsymbol{E}$ 的逆矩阵得

$$\boldsymbol{E}^{-1}=\begin{bmatrix} 0.051\,01 & & \\ -0.016\,00 & 0.044\,81 & \\ -0.000\,23 & -0.009\,11 & 0.040\,58 \end{bmatrix}$$

(4) 求判别系数。

$$\boldsymbol{\alpha}=(c_1,c_2,c_3)'=\boldsymbol{E}^{-1}\boldsymbol{d}$$

$$\boldsymbol{\alpha}=\begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}=\begin{bmatrix} 0.051\,01 & -0.016\,00 & -0.000\,23 \\ -0.016\,00 & 0.044\,81 & -0.009\,11 \\ -0.000\,23 & -0.009\,11 & 0.040\,58 \end{bmatrix}\cdot\begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix}=\begin{bmatrix} 0.216\,92 \\ 0.018\,20 \\ 0.056\,04 \end{bmatrix}$$

(5) 得判别函数。

$$y(x)=\boldsymbol{\alpha}'x=0.216\,92x_1+0.018\,2x_2+0.056\,04x_3$$

(6) 将样本均值$\overline{x}^{(1)}$ 和$\overline{x}^{(2)}$ 代入判别函数，得

$$\begin{aligned}\overline{y}^{(1)}&=\boldsymbol{\alpha}'\overline{x}^{(1)}\\&=0.216\,92\times8.29+0.018\,2\times6.43+0.056\,04\times6\\&=2.251\,533\end{aligned}$$

$$\overline{y}^{(2)}=\boldsymbol{\alpha}'\overline{x}^{(2)}=0.098\,746\,4$$

判别临界值

$$y_c=\frac{\overline{y}^{(1)}+\overline{y}^{(2)}}{2}=1.62$$

则判别准则为

$$\begin{cases} x\in G_1 & \text{若 } y(x)>y_c \\ x\in G_2 & \text{若 } y(x)<y_c \\ \text{待判} & \text{若 } y(x)=y_c \end{cases}$$

(7) 对已知类别的样品判别分类。

对已知类别的样品(训练样品)用线性判别函数进行判别归类，结果如表 4-5 所示。回代率为 100% 判对。

**表 4-5　训练样品的判别结果**

| 组别 | 样品 | 判别函数 $y(x)$ 的值 | 原类号 | 判归类别 |
|---|---|---|---|---|
| 购买组 | 1 | 2.49 | 1 | 1 |
| | 2 | 1.96 | 1 | 1 |
| | 3 | 2.74 | 1 | 1 |
| | 4 | 2.09 | 1 | 1 |
| | 5 | 2.28 | 1 | 1 |
| | 6 | 2.24 | 1 | 1 |
| | 7 | 1.95 | 1 | 1 |
| 非购买组 | 1 | 1.16 | 2 | 2 |
| | 2 | 1.10 | 2 | 2 |
| | 3 | 1.52 | 2 | 2 |
| | 4 | 0.79 | 2 | 2 |
| | 5 | 0.37 | 2 | 2 |

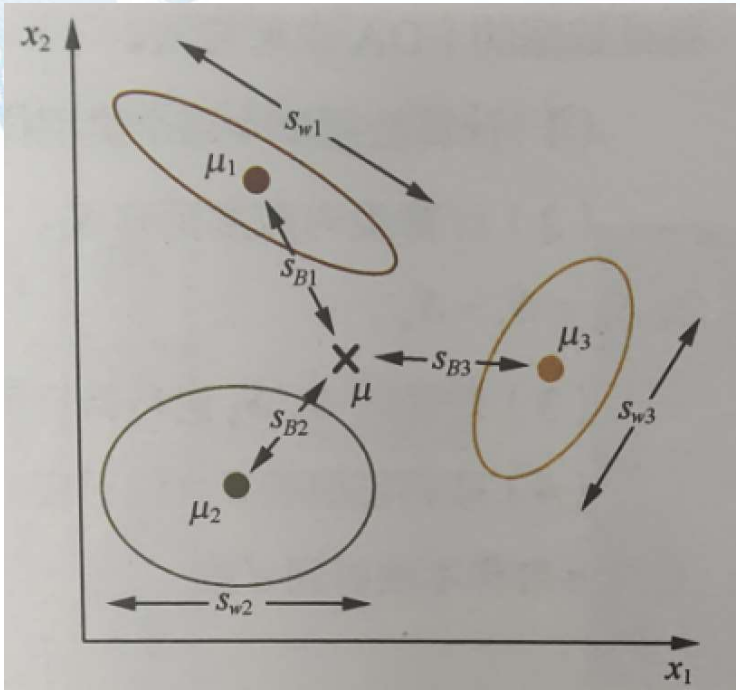
(8) 对待判样品判别归类。

对待判样品 $x=(9, 5, 8)$ 代入判别函数得

$$y(x)=0.216\ 92\times 9+0.018\ 2\times 5+0.056\ 04\times 8=2.491\ 6>y_c$$

故 $x$ 属购买组 $G_1$。

# 多总体费歇尔判别基本原理




$$\max_{\boldsymbol{W}} J(\boldsymbol{W}) = \frac{tr(W^T \boldsymbol{S}_b W)}{tr(W^T \boldsymbol{S}_w W)}$$

$$\max_{\boldsymbol{\omega}} j(\boldsymbol{\omega}) = \frac{\boldsymbol{\omega}^T \boldsymbol{S}_b \boldsymbol{\omega}}{\boldsymbol{\omega}^T \boldsymbol{S}_w \boldsymbol{\omega}}$$

$$(\boldsymbol{\omega}^T \boldsymbol{S}_w \boldsymbol{\omega}) \boldsymbol{S}_b \boldsymbol{\omega} = (\boldsymbol{\omega}^T \boldsymbol{S}_b \boldsymbol{\omega}) \boldsymbol{S}_w \boldsymbol{\omega}$$

$$\boldsymbol{S}_b \boldsymbol{\omega} = \lambda \boldsymbol{S}_w \boldsymbol{\omega}$$

$$\boldsymbol{S}_w{}^{-1} \boldsymbol{S}_b \boldsymbol{\omega} = \lambda \boldsymbol{\omega}$$

$$\boldsymbol{S}_w = \sum_{x \in C_i} (\boldsymbol{x} - \boldsymbol{\mu}_i)^2 = \sum_{x \in C_i} (\boldsymbol{x} - \boldsymbol{\mu}_i)(\boldsymbol{x} - \boldsymbol{\mu}_i)^T$$

$$\boldsymbol{S}_b = \sum_{i=1}^{k} n_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T$$

$$\boldsymbol{\omega} = \boldsymbol{S}_w{}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu})$$

# 多个总体的Fisher判别法

## (一) 判别函数

Fisher判别法实际上是致力于寻找一个最能反映组和组之间差异的投影方向，即寻找线性判别函数 $Y(x)=c_1x_1+\cdots+c_px_p$，设有$k$个总体$G_1,G_2,\cdots,G_k$，分别有均值向量$\bar{\mu}_1$，$\bar{\mu}_2$ ,...,$\bar{\mu}_k$和协方差阵$\Sigma_1,\cdots,\Sigma_k$，分别从各总体中得到样品：

$$X_1^{(1)},\cdots,X_{n_1}^{(1)}$$
$$X_1^{(2)},\cdots,X_{n_2}^{(2)}$$
$$\cdots$$
$$X_1^{(k)},\cdots,X_{n_k}^{(k)}$$

$$n_1 + n_2 + \cdots + n_k = n$$

第i个总体的样本均值向量 $\overline{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_j^{(i)}$

综合的样本均值向量 $\overline{X}_i = \frac{1}{n}\sum_{i=1}^{k} n_i \overline{X}_i$

第i个总体样本组内离差平方和

$$V_i = \sum_{j=1}^{n_i}(X_j^{(i)} - \overline{X}_i)(X_j^{(i)} - \overline{X}_i)'$$

综合的组内离差平方和

$$E = V_1 + V_2 + \cdots + V_k = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_j^{(i)} - \overline{X}_i)(X_j^{(i)} - \overline{X}_i)'$$

组间离差平方和 $B=\sum_{i=1}^{k} n_i(\overline{X}_i-\overline{X})(\overline{X}_i-\overline{X})'$

因为$Y(x)=c_1x_1+\cdots+c_px_p$

$$V_{iy}=\sum_{i=1}^{n_i}(Y_t^{(i)}-\overline{Y}_i)^2=\sum_{t=1}^{n_i}(Y_t^{(i)}-\overline{Y}_i)(Y_t^{(i)}-\overline{Y})'=C'V_iC$$

$$E_0=\sum_{i=1}^{k}V_{iy}=\sum_{i=1}^{k}C'V_iC=C'EC$$

$$B_0=\sum_{i=1}^{k}n_i(\overline{Y}_i-\overline{Y})^2=\sum_{i=1}^{n_i}n_i(\overline{Y}_i-\overline{Y})(\overline{Y}_i-\overline{Y})'=C'BC$$

如果判别分析是有效的，则所有的样品的线性组合$Y(x)=c_1x_1+\cdots+c_px_p$ 满足组内离差平方和小，而组间离差平方和大。则

$$\Delta^2(C)=\frac{B_0}{E_0}=\frac{C'BC}{C'EC}=\max$$

$\Delta^2(C)=\frac{B_0}{E_0}=\frac{C'BC}{C'EC}$的最大值是$B$相对于$E$最大的特征根$\lambda_1$。

而$\lambda_1$ 所对应的特征向量即 $C_1=(c_{11},\cdots,c_{p1})'$。

Fisher样品判别函数是

$$\hat{Y}_1(x)=\hat{c}_{11}x_1+\cdots+\hat{c}_{p1}x_p$$

然而，如果组数k太大，讨论的指标太多，则一个判别函数是不够的，这时需要寻找第二个，甚至第三个线性判别函数

$\Delta^2(C)=\dfrac{B_0}{E_0}=\dfrac{C'BC}{C'EC}$的最大值是$B$相对于$E$第二大的特征根

其特征向量构成第二个判别函数的系数。

$$C_2=(c_{12},\cdots,c_{p2})'$$

$$\hat{Y}_2(x)=\hat{c}_{12}x_1+\cdots+\hat{c}_{p2}x_p$$

类推得到m(m<k)个线性函数。

关于需要几个判别函数得问题，需要累计判别效率达到85%以上，即有

$$\Delta^2(C)=\frac{B_0}{E_0}=\frac{C'BC}{C'EC}$$

设 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p$ 为B相对于E的特征根，则

$$\frac{\lambda_1+\lambda_2+\cdots+\lambda_k}{\lambda_1+\lambda_2+\cdots+\lambda_p}\leq 85\%$$

以m个线性判别函数得到的函数值为新的变量，再进行距离判别。

判别规则：

设$Y_i(X)$为第i个线性判别函数，$(i=1,2,\cdots,m)$，

$$d(x,G_k)=\sum_{i=1}^{r}(y_i(x)-y_i(\bar{x}_k))^2$$

则 $$d(x,G_t)=\min_{1\le j\le k} d(x,G_k)\ \ x\in G_t$$

# 内容

(1) 基本思想

(2) 距离判别法

(3) 贝叶斯判别

(4) 费歇判别

(5) 例子

# 线性判别例题——数字笔迹识别

本次采用的数字笔迹识别例题数据有10992个观测值和17个变量，变量中的前16个变量是识别数字的自变量，第17个变量是识别数字的因变量，即最终识别的数字。

数据大致格式如下：

| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 88 | 92 | 2 | 99 | 16 | 66 | 94 | 37 | 70 | 0 | 0 | 24 | 42 | 65 | 100 | 100 | 8 |
| 80 | 100 | 18 | 98 | 60 | 66 | 100 | 29 | 42 | 0 | 0 | 23 | 42 | 61 | 56 | 98 | 8 |
| 0 | 94 | 9 | 57 | 20 | 19 | 7 | 0 | 20 | 36 | 70 | 68 | 100 | 100 | 18 | 92 | 8 |
| 95 | 82 | 71 | 100 | 27 | 77 | 77 | 73 | 100 | 80 | 93 | 42 | 56 | 13 | 0 | 0 | 9 |
| 68 | 100 | 6 | 88 | 47 | 75 | 87 | 82 | 85 | 56 | 100 | 29 | 75 | 6 | 0 | 0 | 9 |
| 70 | 100 | 100 | 97 | 70 | 81 | 45 | 65 | 30 | 49 | 20 | 33 | 0 | 16 | 0 | 0 | 1 |
| 40 | 100 | 0 | 81 | 15 | 58 | 100 | 57 | 47 | 87 | 50 | 88 | 40 | 42 | 36 | 0 | 4 |
| 3 | 71 | 0 | 95 | 45 | 100 | 100 | 99 | 79 | 78 | 48 | 53 | 31 | 24 | 54 | 0 | 7 |
| 79 | 87 | 98 | 81 | 71 | 100 | 72 | 73 | 100 | 66 | 91 | 21 | 48 | 0 | 0 | 13 | 9 |
| 92 | 95 | 30 | 100 | 34 | 68 | 87 | 89 | 84 | 78 | 100 | 35 | 64 | 0 | 0 | 19 | 9 |
| 58 | 64 | 100 | 96 | 27 | 100 | 0 | 63 | 79 | 65 | 91 | 72 | 48 | 36 | 10 | 0 | 9 |
| 34 | 89 | 3 | 70 | 1 | 25 | 49 | 0 | 100 | 23 | 100 | 67 | 56 | 99 | 0 | 100 | 0 |
| 0 | 90 | 46 | 100 | 88 | 92 | 79 | 69 | 60 | 48 | 39 | 27 | 47 | 6 | 100 | 0 | 2 |
| 20 | 71 | 0 | 29 | 31 | 0 | 78 | 12 | 100 | 51 | 84 | 93 | 37 | 100 | 8 | 66 | 0 |
| 100 | 100 | 67 | 98 | 41 | 80 | 44 | 50 | 78 | 42 | 68 | 16 | 35 | 2 | 0 | 0 | 5 |
| 91 | 69 | 48 | 57 | 9 | 79 | 60 | 100 | 100 | 75 | 95 | 40 | 64 | 8 | 0 | 0 | 9 |

# 进行线性判别的python代码如下：

```
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.discriminant_analysis import QuadraticDiscriminantAnalysis
from sklearn.metrics import confusion_matrix
import numpy as np
import pandas as pd

def Fold(y,x,z,seed=8888):
    id0=np.arange(len(y))
    part=[]
    for i in np.unique(y):
        part.append(id0[y==i])
    zid=[];xn=[];yn=[]
    np.random.seed(seed)
    for k in part:
        np.random.shuffle(k)
        yn.extend(y[k])
        xn.extend(x[k])
        zid.extend((list(range(z))*int(len(k)/z+1))[:len(k)])
    return  zid,yn,xn
```

```
def CCV(clf,zid,xn,yn):
    y_pred=[];yN=[]
    for j in np.unique(zid):
        clf.fit(xn[zid!=j],yn[zid!=j])
        yN.extend(yn[zid==j])
        y_pred.extend(clf.predict(xn[zid==j]))
    y_pred=np.array(y_pred)
    yN=np.array(yN)
    error=np.sum(yN!=y_pred)/len(y)
    cmatrix=confusion_matrix(yN,y_pred)
    return (error,cmatrix)
v=pd.read_csv("pendigits.csv",index_col=False)
x=np.array(v[v.columns[:16]])
y=np.array(v[v.columns[16]])
zid,yn,xn=Fold(y=y,x=x,z=10,seed=1010)
xn=np.array(xn)
yn=np.array(yn)
lda=LinearDiscriminantAnalysis()
er,cm=CCV(lda,zid,xn,yn)
print("Error rate=",er)
print(cm)
```

# 判别的错误率和混淆矩阵如图：

```
Error rate= 0.13440091507006005
[[327   5   0   0   5   0   0   0  25   1]
 [  0 234  69   0   1  39   0   8   0  13]
 [  0  18 343   0   0   0   0   3   0   0]
 [  0   3   0 319   0   0   0  12   0   2]
 [  0   0   0   0 355   4   1   2   0   2]
 [  1   0   0  23   0 236   0   1   6  68]
 [  5   0   0   0   1   2 328   0   0   0]
 [  0  29   5   4  10   3   0 301   0  12]
 [ 23   0   0   0   0  11   0   0 300   1]
 [  0  14   0   2   8  24   0   3   1 284]]
```

错误率为13.44%，混淆矩阵的对角线是判断正确的数目

# 小测试

- Fisher判别分析可以用于以下哪些任务？（两个答案）
  - A. 分类任务
  - B. 回归任务
  - C. 聚类任务
  - D. 特征选择

# 小测试

- Fisher判别分析中，最佳投影方向的计算需要用到什么信息？
  - A. 样本均值
  - B. 样本协方差矩阵
  - C. 类别数量
  - D. 样本数量