# 判别分析

TUTU

# 判别分析基本概念

♣ 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。

♣ 判别分析与聚类分析的区别与联系：

- 判别分析：已知研究对象分为若干个类别，并且已经取得每一类别的一批观测数据，在此基础上寻求出分类的规律性，建立判别准则，然后对未知类别的样品进行判别分类。

  将共性的聚类结果，作为已知类别的样本的信息 (训练样本)，对未知类别的样品 (测试样本) 进行判别分类。

- 聚类分析：一批样品划分为几类事先并不知道，正需要通过聚类分析来给以确定类型。

  用不同的聚类方法可能得到不同的结果，保留共性的聚类结果；对于用不同方法归类不同的少数样品，再结合判别分析加以判断归类。

# 距离判别

♣ 基本思想：

计算样品 $x$ 到第 $i$ 个类的距离 $d^2(x, G_i)$，哪个距离最小，就将它判归哪个总体。

♣ 马氏距离：$d^2(\boldsymbol{x}, G_i) = (\boldsymbol{x} - \boldsymbol{\mu}_i)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i)$，不受变量间的相关性和量纲的影响

♣ 总体协差阵相等时 ($\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$)：

- 判别函数：$f_i(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i - \dfrac{1}{2} \boldsymbol{\mu}_i^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i$
- 判别规则：若 $f_l(\boldsymbol{x}) = \max\limits_{1 \leq i \leq k} f_i(\boldsymbol{x})$，$\boldsymbol{x} \in G_l$
- 两个总体的错判概率：$P(2|1) = 1 - \Phi\left(\dfrac{\mu_2 - \mu_1}{2\sigma}\right)$

# 距离判别

♣ 两个总体的距离判别：设 $G_1(n_1$ 个$)$ 和 $G_2(n_2$ 个$)$ 为两个不同的 $p$ 元总体，$G_1(G_2)$ 的均值向量为 $\boldsymbol{\mu}_1(\boldsymbol{\mu}_2)$，协方差阵为 $\boldsymbol{\Sigma}_1(\boldsymbol{\Sigma}_2)$，$\boldsymbol{x}^{\mathrm{T}} = (x_1, x_2, \cdots, x_p)$ 是待判样品，$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ 且已知

- $\boldsymbol{x}$ 到 $G_1, G_2$ 的距离的平方 (二次判别函数) 分别为

$$d^2(\boldsymbol{x}, G_i) = (\boldsymbol{x} - \boldsymbol{\mu}_i)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i), \quad (i = 1, 2)$$

- 判别准则为

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

# 距离判别

♣ 两个总体的距离判别：(相等协差阵)

- 令 $\bar{\boldsymbol{\mu}} = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$，
  则 $d^2(\boldsymbol{x}, G_2) - d^2(\boldsymbol{x}, G_1) = 2(\boldsymbol{x} - \bar{\boldsymbol{\mu}})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$，
  令 $W(\boldsymbol{x}) = (\boldsymbol{x} - \bar{\boldsymbol{\mu}})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$，则判别准则为

$$\begin{cases} \boldsymbol{x} \in G_1, & W(\boldsymbol{x}) > 0, \\ \boldsymbol{x} \in G_2, & W(\boldsymbol{x}) < 0. \end{cases}$$

- 若 $p = 1$，两个总体是 $N(\mu_i, \sigma^2)$，则 $W(x) = (x - \bar{\mu})\frac{1}{\sigma^2}(\mu_1 - \mu_2)$，
  判别准则为

$$\begin{cases} x \in G_1, & x < \bar{\mu}, \\ x \in G_2, & x > \bar{\mu}. \end{cases}$$

## 距离判别

♣ 总体协差阵不相等时：$d^2(\boldsymbol{x}, G_i) = (\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)$

♣ 两个总体的距离判别：(不等协差阵)

- $W(\boldsymbol{x}) = (\boldsymbol{x}-\boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) - (\boldsymbol{x}-\boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)$
- 判别准则：

$$\begin{cases} \boldsymbol{x} \in G_1, & W(\boldsymbol{x}) > 0, \\ \boldsymbol{x} \in G_2, & W(\boldsymbol{x}) < 0. \end{cases}$$

- 若 $p=1$，两个总体是 $N(\mu_i, \sigma_i^2)$，则 $W(x) = -\dfrac{\sigma_1+\sigma_2}{\sigma_1\sigma_2}\left(x - \dfrac{\sigma_1\mu_2+\sigma_2\mu_1}{\sigma_1+\sigma_2}\right)$，令 $\mu^* = \dfrac{\sigma_1\mu_2+\sigma_2\mu_1}{\sigma_1+\sigma_2}$，判别准则为

$$\begin{cases} x \in G_1, & x < \mu^*, \\ x \in G_2, & x > \mu^*. \end{cases}$$

# Bayes 判别

♣ 距离判别与 Bayes 判别：

- 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，没有考虑错判的损失
- Bayes 判别解决了这两个问题

♣ 总体 $G_i$ 的概率密度为 $f_i(\boldsymbol{x})$，$G_i$ 出现的概率为 $q_i$，

则 $P(G_i|\boldsymbol{x}_0) = \dfrac{q_i f_i(\boldsymbol{x})}{\sum q_i f_i(\boldsymbol{x})}$

♣ 判别准则：若 $q_l f_l(\boldsymbol{x}) = \max\limits_{1\le i\le k} q_i f_i(\boldsymbol{x})$，则 $\boldsymbol{x}_0 \in G_l$

# Bayes 判别

♣ 两个总体的 Bayes 判别：

若 $f_i(\boldsymbol{x}) = \dfrac{1}{[(2\pi)^k|\boldsymbol{\Sigma}|]^{1/2}} \exp\left[-\dfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}^{(i)})^{\mathrm{T}}\boldsymbol{\Sigma}(\boldsymbol{x}-\boldsymbol{\mu}^{(i)})\right]$，则

- 判别函数：$z_i(\boldsymbol{x}) = \ln q_i - \dfrac{1}{2}\ln|\boldsymbol{\Sigma}_i| - \dfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}^{(i)})^{\mathrm{T}}\boldsymbol{\Sigma}(\boldsymbol{x}-\boldsymbol{\mu}^{(i)})$
- 判别准则：若 $Z_l(\boldsymbol{x}) = \max\limits_{1\le i\le k} Z_i(\boldsymbol{x})$，$\boldsymbol{x}\in G_l$
- 协差阵相等时，$z_i(\boldsymbol{x}) = \ln q_i - \dfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}^{(i)})^{\mathrm{T}}\boldsymbol{\Sigma}(\boldsymbol{x}-\boldsymbol{\mu}^{(i)})$
- 当先验概率相等时，完全成为距离判别法

# Fisher 判别

♣ 基本思想：

其基本思想是投影，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。

♣ 两个总体 Fisher 判别的判别函数：$y = c_1 x_1 + c_2 x_2 + \cdots + c_p x_p$

- 组间离差平方和最大，组内离差平方和最小
- $\bar{y}^{(i)} = \frac{1}{n_i} \sum y_i = \sum c_k \bar{x}_k$
- $\max\ Q = (\bar{y}^{(1)} - \bar{y}^{(2)})^2$；$\min\ R = \sum_j (y_i^{(j)} - \bar{y}^{(j)})^2$；$\max\ I = \frac{Q}{R}$
- 令 $d_i = \bar{x}_i^{(1)} - \bar{x}_i^{(2)}$，
  $$s_{kl} = \sum_{i=1}^{n_1} (x_{ik}^{(1)} - \bar{x}_k^{(1)})(x_{il}^{(1)} - \bar{x}_l^{(1)}) + \sum_{i=1}^{n_2} (x_{ik}^{(2)} - \bar{x}_k^{(2)})(x_{il}^{(2)} - \bar{x}_l^{(2)}),$$
  $E = (s_{kl})_{p\times p}$，则$\boldsymbol{c} = E^{-1}\boldsymbol{d}$

# Fisher 判别

♣ 两个总体 Fisher 判别：定义临界判别点为

$$y_c = \begin{cases} \frac{\bar{y}^{(1)}+\bar{y}^{(2)}}{2}, & \text{两总体方差相等} \\ \frac{\hat{\sigma}_2\bar{y}^{(1)}+\hat{\sigma}_1\bar{y}^{(2)}}{\hat{\sigma}_1+\hat{\sigma}_2}, & \text{两总体方差不等} \end{cases}$$

，且 $\bar{y}^{(1)} > \bar{y}^{(2)}$，则判别准则为

$$\begin{cases} x \in G_1, & y(x) > y_c, \\ x \in G_2, & y(x) < y_c. \end{cases}$$

# Fisher 判别

♣ 多个总体 Fisher 判别的判别函数：$\boldsymbol{X}=(x_1,x_2,\cdots,x_p)^{\mathrm{T}}$，从 $G_i$ 中取 $n_i$ 个样本 $\boldsymbol{X}^{(i)}$，样本均值向量为 $\bar{\boldsymbol{X}}_i=\dfrac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{X}_j^{(i)}$，协方差阵为 $\boldsymbol{\Sigma}_i$

- 综合的样本均值向量：$\bar{\boldsymbol{X}}=\dfrac{1}{n}\sum n_i\bar{\boldsymbol{X}}_i$
- 第 $i$ 个总体样本组内离差平方和：
  $$\boldsymbol{V}_i=\sum_{j=1}^{n_i}n_i(\boldsymbol{X}_j^{(i)}-\bar{\boldsymbol{X}}_i)(\boldsymbol{X}_j^{(i)}-\bar{\boldsymbol{X}}_i)^{\mathrm{T}}$$
- 综合的组内离差平方和：$E=\sum V_i$
- 组间离差平方和：$\boldsymbol{B}=\sum_{i=1}^{k}n_i(\bar{\boldsymbol{X}}_i-\bar{\boldsymbol{X}})(\bar{\boldsymbol{X}}_i-\bar{\boldsymbol{X}})^{\mathrm{T}}$
- $\max\ \Delta^2(\boldsymbol{C})=\dfrac{\boldsymbol{C}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{C}}{\boldsymbol{C}^{\mathrm{T}}\boldsymbol{E}\boldsymbol{C}}$，则求 $\boldsymbol{B}$ 相对于 $\boldsymbol{E}$ 的特征根

# Fisher 判别

♣ 多个总体 Fisher 判别的判别函数：设 $\boldsymbol{E}^{-1}\boldsymbol{B}$ 的非零特征值为 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$，相应的特征向量为 $\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_p$

- 取 $\boldsymbol{c} = \boldsymbol{c}_1$ 时，得到一个 Fisher 判别函数：
  $\hat{Y}_1(\boldsymbol{x}) = \hat{c}_{11}x_1 + \cdots + \hat{c}_{p1}x_p$，此时，$\Delta^2(\boldsymbol{C})$ 达到最大值$\lambda_1$
- 如果判别函数不够，可建立更多的判别函数，则前 $k$ 个线性判别函数为 $\hat{Y}_i(\boldsymbol{x}) = \boldsymbol{c}_i^{\mathrm{T}}\boldsymbol{x}, (i = 1, 2, \cdots, k)$
- 前 $k$ 个线性判别函数的累计判别能力为 $\dfrac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$
- 设 $Y_i(\boldsymbol{x})$ 为第 $i$ 个线性判别函数，$d(x, G_k) = \sum_{i=1}^{m}(Y_i(\boldsymbol{x}) - Y_i(\boldsymbol{x}_k))^2$，
  则判别准则：若 $d(x, G_t) = \min_{1\leq j\leq k} d(x, G_k)$，则 $x \in G_t$

# 判别分析 SAS 代码

♣ SAS 代码:

```
/*距离判别, listerr是回判, pool=
    no表示认为协差阵不等, yes是
    相等*/
proc discrim data=unknowdata
    listerr testdata=knowndata
    out=out testout=testout
    outstat=outstat pool=no;
class x5;
var x1-x4;
run;

/*Bayes判别, 多了priors*/
proc discrim data=unknowdata
    listerr testdata=knowndata
    out=out testout=testout
    outstat=outstat pool=no;
class x5;
var x1-x4;
priors '1'=0.2 '2'=0.8;
run;

/*Fisher判别*/
proc candisc data=yourdata out=
    out;
class x6;
var x1-x5;
run;

proc plot data=out;
plot can2*can1=x6;
run;
```