

第七章 地理系统的聚类分析和判别分析



河北师范大学：胡引翠

主要内容



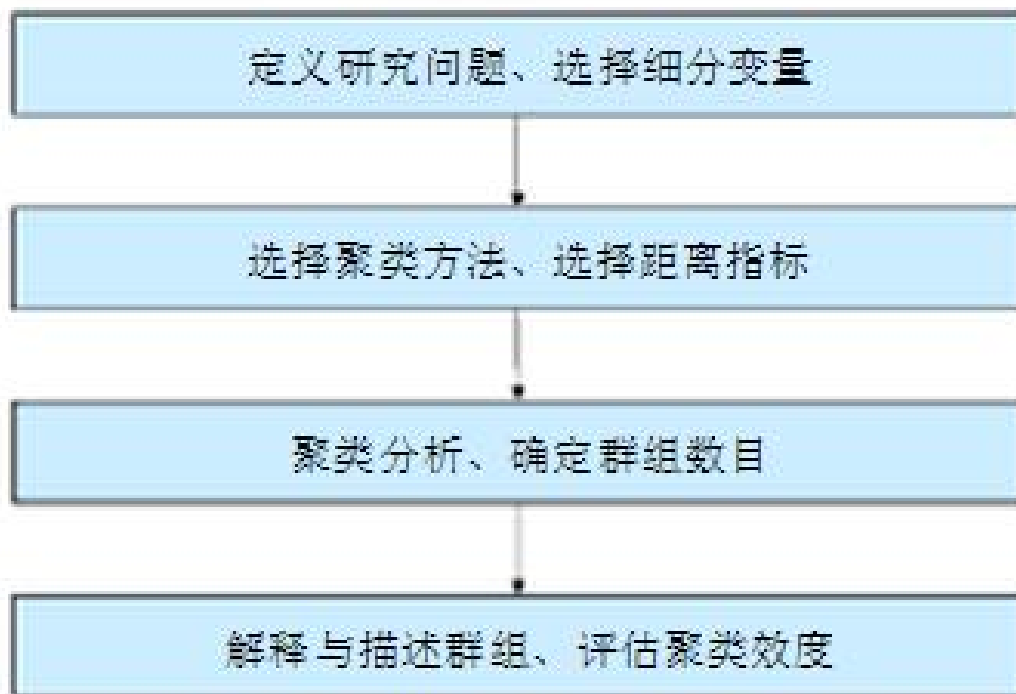
§ 1地理系统的聚类分析

§ 2地理系统的判别分析

§ 1地理系统的聚类分析



地理系统分类的步骤





地理系统分类的分类

“聚类分析” 方法有系统聚类法、动态聚类法和模糊聚类法等。

根据分类对象的不同，分为样品聚类（**Q**型聚类）和变量聚类（**R**型聚类）。

样品聚类（**Q**型聚类）：在**SPSS**中称对事件（**cases**）进行聚类，或是说对观测量进行聚类，根据被观测的对象的各种特征值进行分类。

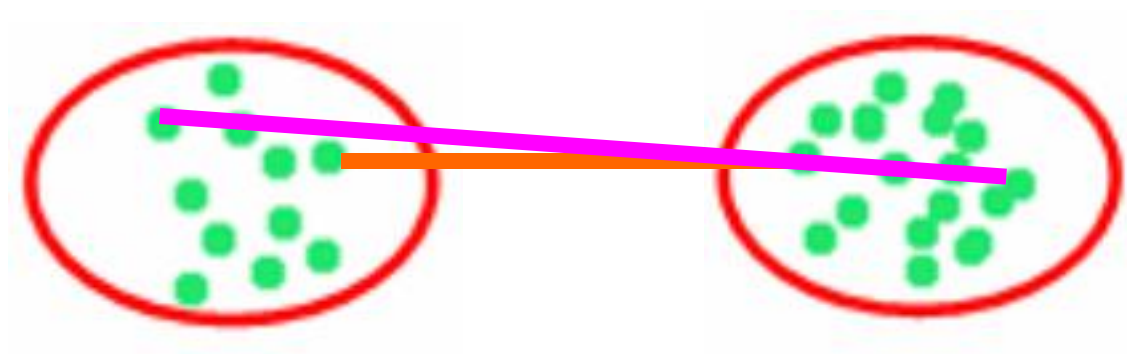
变量聚类（**R**型聚类）：反映同一事物特点的变量有很多，变量之间往往具有一定的相关关系，需要找出彼此独立的有代表性的变量，就需要先进行变量聚类。（批量生产）



■ 地理系统的聚类方法

定义类与类之间距离的方法：

- 最短距离法
- 最长距离法
- 类平均法
- 重心法
- 离差平方和法





§ 1地理系统的聚类分析

§ 2地理系统的判别分析





§ 1地理系统的聚类分析

§ 2地理系统的判别分析

在我们的日常生活和工作实践中，常常会遇到判别分析问题，即根据历史上划分类别的有关资料和某种最优准则，确定一种判别方法，判定一个新的样本归属哪一类。

如，在天气预报中，我们有一段较长时间关于某地区每天气象的记录资料（晴阴雨、气温、气压、湿度等），现在想建立一种用连续五天的气象资料来预报第六天是什么天气的方法。



§ 1地理系统的聚类分析

§ 2地理系统的判别分析

把这类问题用数学语言来表达，可以叙述如下：

设有 n 个样本，对每个样本测得 p 项指标（变量）的数据，已知每个样本属于 k 个类别（或总体） G_1, G_2, \dots, G_k 中的某一类，且它们的分布函数分别为 $F_1(x), F_2(x), \dots, F_k(x)$ 。我们希望利用这些数据，找出一种判别函数，使得这一函数具有某种最优性质，能把属于不同类别的样本点尽可能地区别开来，并对测得同样 p 项指标（变量）数据的一个新样本，能判定这个样本归属于哪一类。

§ 2地理系统的判别分析`



2.1判别分析的基本原理

◆ 判别分析的概念

如何根据两类（或几类）个体的某些属性或特征来分辨或判别两类（或几类）个体。

与聚类分析的**相同**之处：都能确定地理类型。

与聚类分析的**差别**：判别分析兼有**判别**和**分析**两种性质，但以判别为主，判别分析必须事先已知类型的划分，而聚类分析则不必事先知道已知类型，类型的划分是聚类的结果。

作用：

- ①对已分好的类型进行合理性检验。
- ②判别某地理类型的归属问题和确定区域界限。
- ③评价各要素特征值在判别分析中的大小。



判别分析和聚类分析往往联合使用。当总体分类不清楚时，先用聚类分析对一批样本进行分类，再用判别分析构建判别式对新样本进行判别。

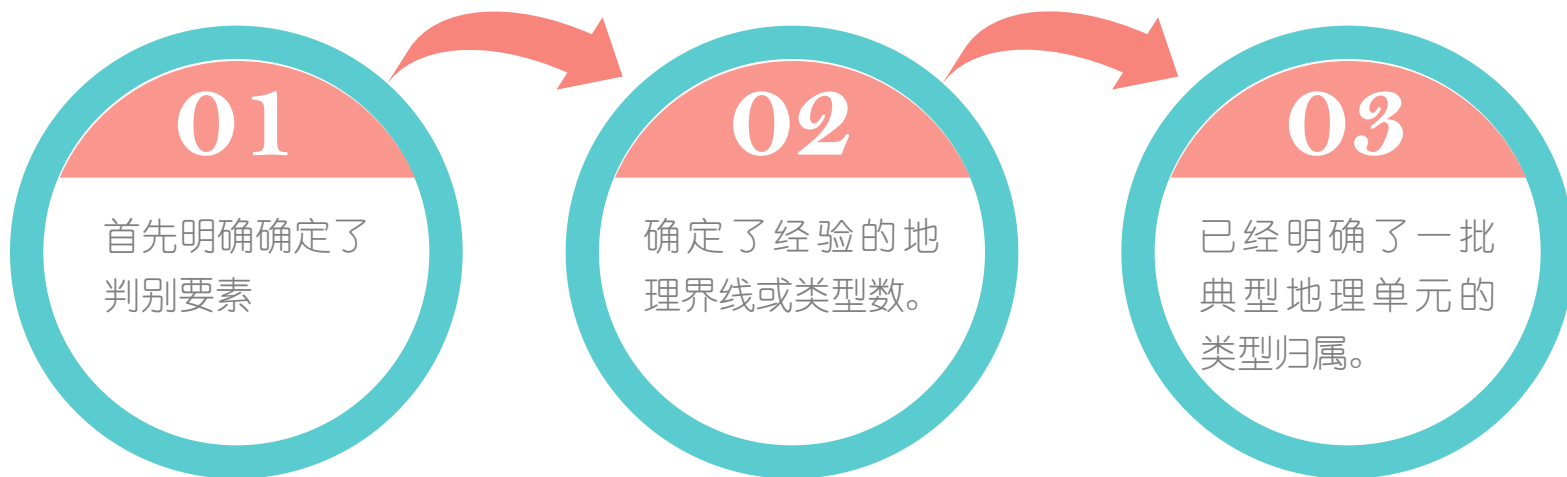
§ 2地理系统的判别分析`



2.1判别分析的基本原理

◆ 判别分析的条件

判别分析：如何根据两类（或几类）个体的某些属性或特征来分辨或判别两类（或几类）个体。



判别分析的条件

§ 2地理系统的判别分析`



■ 2.1判别分析的基本原理

判别分析的准则

费歇准则（Fisher）:对原始数据进行线性组合，使其形成一个新变量，即判别函数，判别函数要求能够使得类间差别最大，而类内（部）差别最小。

贝叶斯（Bayes）准则:在于寻找一种方法把点集合（地理单元），划分为若干个互不相交的子空间，即使得： $R_p \cap R_q = \phi$

并使各点集合经过这样的划分之后，所产生的判断错误（两种不同地理单元混淆）的概率最小。

把点集合向低维空间投影，根据投影后的点集合加以区分。这时是关键在于寻找最佳的能分辨不同性质点集合的投影方向。

在此可以看出，两种判别准则都是把地理单元视为判别要素所组成的指标空间的一个点。

§ 2地理系统的判别分析`



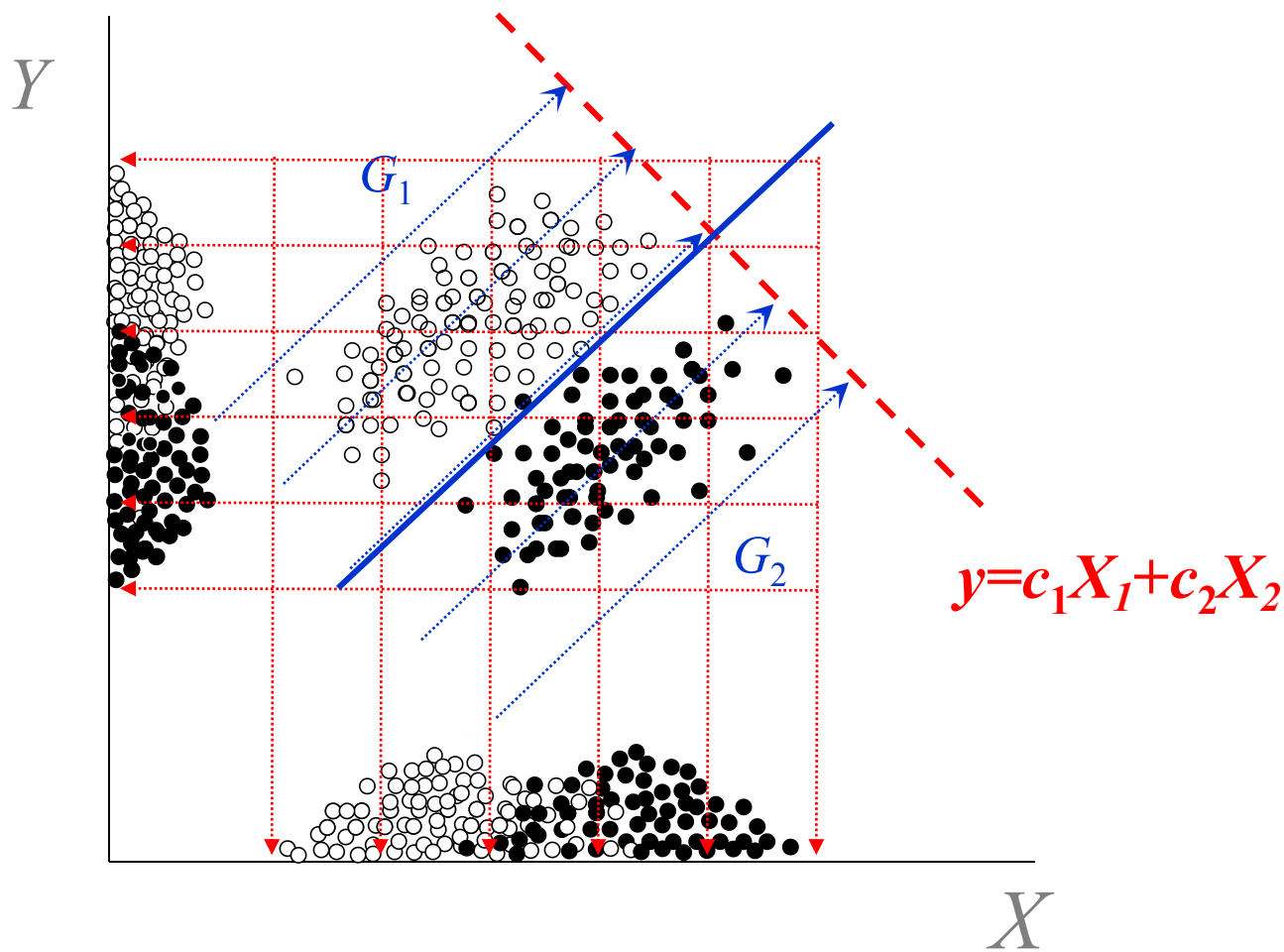
2.1判别分析的基本原理

◆ 判别分析的准则

费歇准则

- 线性组合，形成新变量-判别函数；
- 使各类均值之间的差别最大；
- 各类内部离差平方和最小；
- 即类间均值差与类内方差比最大。

两类Fisher判别示意图



§ 2地理系统的判别分析`



2.1判别分析的基本原理

◆ 判别分析的准则

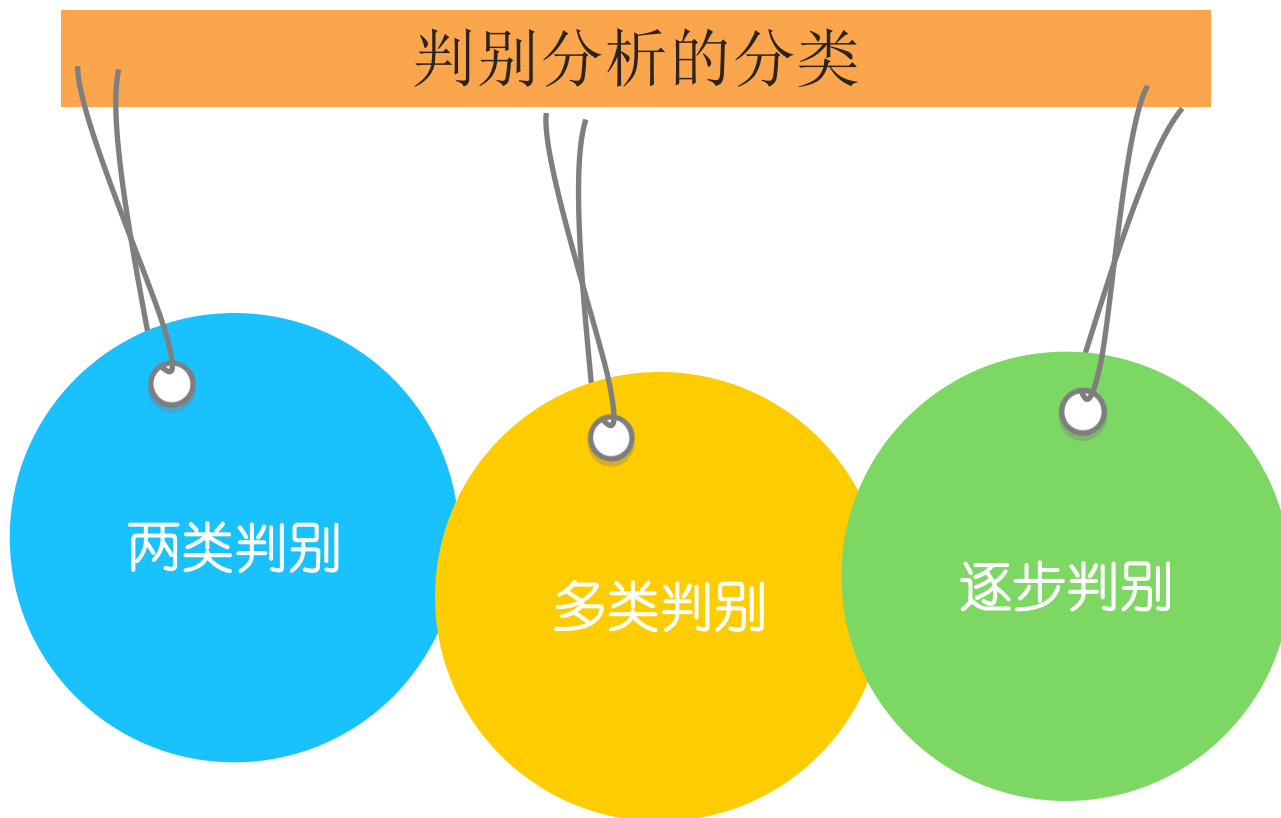
贝叶斯准则

- 把数据分成几类或几组；
- 算出未知类归属于各已知类的概率；
- 把它划归于概率值最大的类中；
- 还可根据错分损失来判定。

§ 2地理系统的判别分析`



■ 2.1判别分析的基本原理



§ 2地理系统的判别分析`



◆ 1、两类地理判别-问题抽象

总体A和B；

分别抽出 n_1 个和 n_2 个样品；

测得每个样品的 m 个指标值 x_1, x_2, \dots, x_m ；

要求依据这 n_1+n_2 个样本指标值，判别一个新样品属于A还是属于B。



◆ 两类地理判别-思考

若A和B中存在一种差异较大的指标，则易解；

例如：非洲较高大和较矮小的两种人，可用身高来分类

实际常常遇到不存在一种差异较大的指标可以作为分类依据的情况。



例如 设地区有“湿润”，“干旱”两种，我们要判别一个地区在一个时期是“湿润”还是“干旱”。

根据过去的资料可知，湿润干旱情况与与降水量有关，也和气温有关，因此可以用降水量和日均气温这两个量去预测。

以 x_1 代表降水量，以 x_2 表示气温。现在假定调查了 n 个时期，得到 n 组数据。这 n 组数据反应的有湿润的也有干旱的，不妨设有 r 组湿润， l 组干旱 ($l=n-r$)，则可将 n 组数据分组如下：

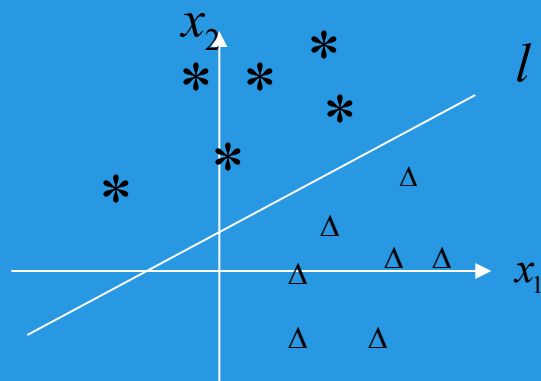


湿润组 $\left\{ \begin{array}{l} (x_{11}^0, x_{12}^0) \\ (x_{21}^0, x_{22}^0) \\ \vdots \\ (x_{r1}^0, x_{r2}^0) \end{array} \right.$

干旱组

$\left\{ \begin{array}{l} (x_{11}^1, x_{12}^1) \\ (x_{21}^1, x_{22}^1) \\ \vdots \\ (x_{l1}^1, x_{l2}^1) \end{array} \right.$

将这n组数据标在平面上，以“*”表示湿润组所对应的点，以“ Δ ”表示干旱数据对应的点，若能得到如图所示的点聚图，即湿润时期的数据和干旱时期的数据有较为明显的区别



我们就可以根据这种趋势直观的做出某些判断。



例如，若某个预测时期的数据对应的点为“*”则我们应判断这一时期为湿润期，若对应点为“ Δ ”则应判断这一时期为干旱期。因此，在预测时，重要的问题是要找出分界线 l ，其方程为

$$c_0 + c_1x_1 + c_2x_2 = 0$$

使得当某个时期的数据 (x_1, x_2) 为已知时代入上式左端，若有

$$c_0 + c_1x_1 + c_2x_2 > 0 \quad \text{即}$$

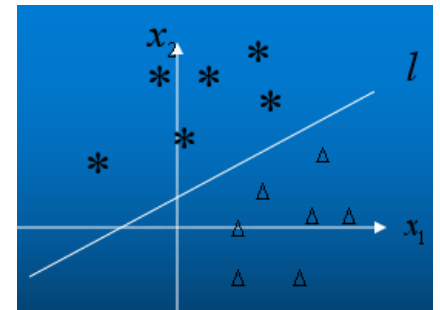
$$c_1x_1 + c_2x_2 > -c_0$$

则预测这时期为湿润期，若有

$$c_0 + c_1x_1 + c_2x_2 < 0 \quad \text{即}$$

$$c_1x_1 + c_2x_2 < -c_0$$

则预测这个时期为干旱期。

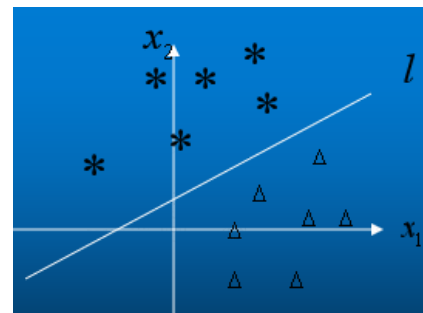




这种预测分析的方法就是判别分析法。在利用这种方法时必须要求湿润期的数据和干旱期的数据之间有一条较明显的分界线，对这一点，我们后面将进一步阐述。

我们令

$$y = c_1 x_1 + c_2 x_2$$

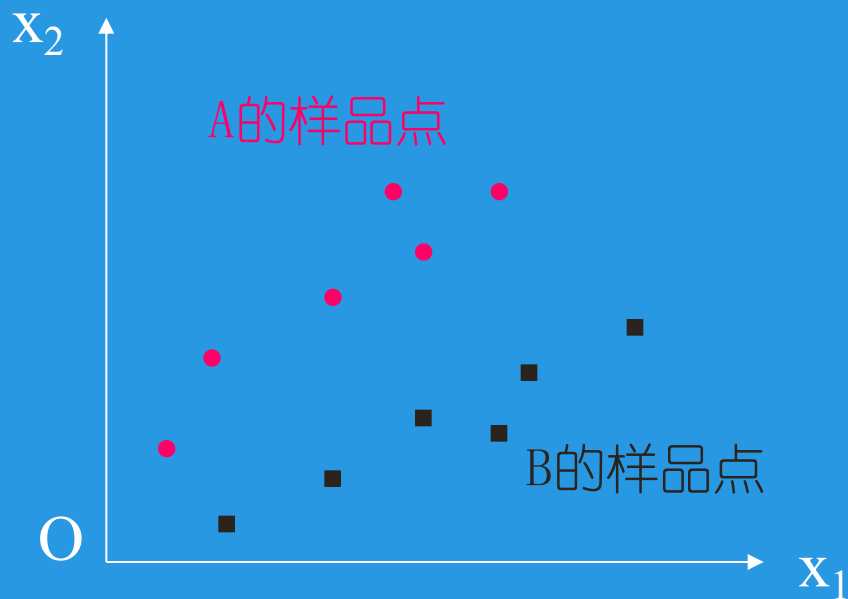


称此函数为线性判别函数，称 $y_0 = -c_0$ 为临界值。

进行判别分析就是要在某种最优准则下，确定线性判别函数的系数 c_1, c_2 以及临界值 $-c_0$ 。



◆ 两类地理判别-思考



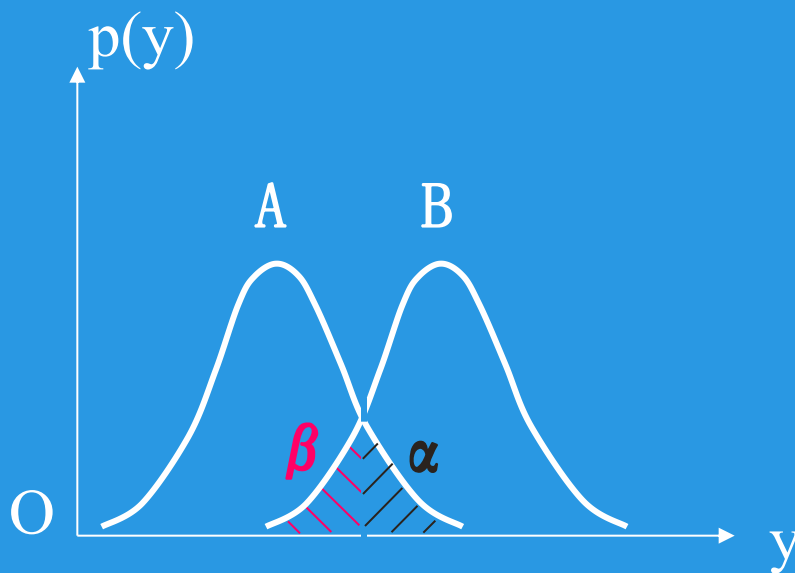
$$c_1x_1 + c_2x_2 - y_0 = 0$$



◆ 两类地理判别-两类错误

第一类错误

第二类错误





◆ 两类地理判别

假设预测因子有 p 个指标，即 x_1, x_2, \dots, x_p ，有 n 组观察或调查得到的数据。判别分析就是要根据这些数据，在适当的判别准则下，确定判别函数：

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

并找出临界值 y_0 。

我们将要判别的两组分别标记为A和B（如A代表湿润，B代表干旱）。对于 p 个判别指标。不妨设组A有 s 组数据，组B有 t 组数据， $n=s+t$ ，现将数据分组如下：



组A的数据

$$\begin{cases} (x_{11}^0, x_{12}^0, \dots, x_{1p}^0) \\ (x_{21}^0, x_{22}^0, \dots, x_{2p}^0) \\ \vdots \\ (x_{s1}^0, x_{s2}^0, \dots, x_{sp}^0) \end{cases}$$

组B的数据

$$\begin{cases} (x_{11}^1, x_{12}^1, \dots, x_{1p}^1) \\ (x_{21}^1, x_{22}^1, \dots, x_{2p}^1) \\ \vdots \\ (x_{t1}^1, x_{t2}^1, \dots, x_{tp}^1) \end{cases}$$

下面反过来思考整个问题，假定用

$$y = l_1 x_1 + l_2 x_2 + \dots + l_p x_p$$

作为判别函数，则组A的数值对应的判别值为

$$y_1^0 = l_1 x_{11}^0 + l_2 x_{12}^0 + \dots + l_p x_{1p}^0$$

$$y_2^0 = l_1 x_{21}^0 + l_2 x_{22}^0 + \dots + l_p x_{2p}^0$$

$$\vdots$$

$$y_s^0 = l_1 x_{s1}^0 + l_2 x_{s2}^0 + \dots + l_p x_{sp}^0$$



组B的数值对应的判别值为

$$y_1^1 = l_1 x_{11}^1 + l_2 x_{12}^1 + \cdots + l_p x_{1p}^1$$

$$y_2^1 = l_1 x_{21}^1 + l_2 x_{22}^1 + \cdots + l_p x_{2p}^1$$

\vdots

$$y_t^1 = l_1 x_{t1}^1 + l_2 x_{t2}^1 + \cdots + l_p x_{tp}^1$$

又作

$$\bar{y}^0 = \frac{1}{s} \sum_{i=1}^s y_i^0$$

$$\bar{y}^1 = \frac{1}{t} \sum_{i=1}^t y_i^1$$

即 \bar{y}^0 为组A的代表, \bar{y}^1 为组B的代表。



◆ 判别函数

综合判断标准 y

$$y = f(x_1, x_2, \dots, x_m)$$

$$y = c_1 x_1 + c_2 x_2 + \dots + c_m x_m = \sum_{k=1}^m c_k x_k$$

式中： $c_k(k=1,2,\dots,m)$ 为判别系数，可以反映各要素或特征值的作用方向、分辨能力和贡献率的大小。

综上 (1) , (2) Fisher最优判别准则为函数

$$L(l_1, l_2, \dots, l_p) = \frac{(\bar{y}^0 - \bar{y}^1)^2}{\sum_{i=1}^s (y_i^0 - \bar{y}^0)^2 + \sum_{i=1}^t (y_i^1 - \bar{y}^1)^2}$$

越大越好。从而最优判别函数的系数 c_1, c_2, \dots, c_p 为函数 $L(l_1, l_2, \dots, l_p)$ 的极大值点。由微分学可知, c_1, c_2, \dots, c_p 为方程组

$$\frac{\partial L(l_1, l_2, \dots, l_p)}{\partial l_i} = 0, j = 1, 2, \dots, p$$


的解。






◆ 判别函数

我们通过判别值 y 来进行判别，为使组A同组B之间有明显的区别，自然希望它们的代表值之间的差距越大越好。即为充分反映出两种地理类型的差别，须

$$[\bar{y}(A) - \bar{y}(B)]^2$$


$$\sum_{i=1}^{n_1} [y_i(A) - \bar{y}(A)]^2 + \sum_{i=1}^{n_2} [y_i(B) - \bar{y}(B)]^2$$




◆ 判别函数

$$\frac{\partial I}{\partial C_k} = 0$$

$$\begin{cases} l_{11}C_1 + l_{12}C_2 + \cdots + l_{1m}C_m = \bar{x}_{1A} - \bar{x}_{1B} = d_1 \\ l_{21}C_1 + l_{22}C_2 + \cdots + l_{2m}C_m = \bar{x}_{2A} - \bar{x}_{2B} = d_2 \\ \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \quad \quad \quad \vdots \\ l_{m1}C_1 + l_{m2}C_2 + \cdots + l_{mm}C_m = \bar{x}_{mA} - \bar{x}_{mB} = d_m \end{cases}$$



◆ 判别函数-判定未知类的归属

各类的均值

$$\begin{cases} \bar{y}(A) = \sum_{k=1}^m C_k \bar{x}_k(A) \\ \bar{y}(B) = \sum_{k=1}^m C_k \bar{x}_k(B) \end{cases}$$

判别指标（判别临界值）

$$y_c = \frac{n_A \bar{y}(A) + n_B \bar{y}(B)}{n_A + n_B}$$



◆ 判别函数-判定未知类的归属

若 $\bar{y}(A) > \bar{y}(B)$

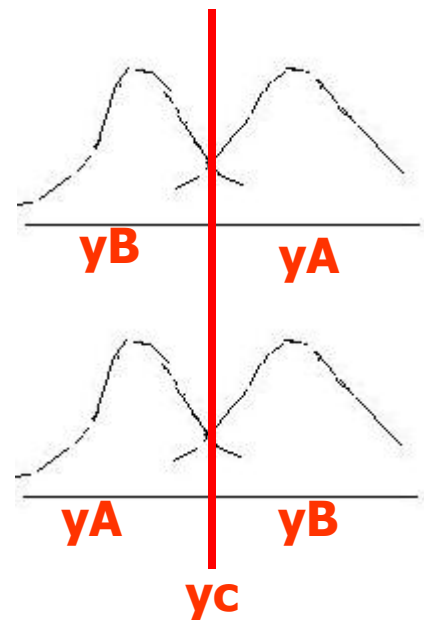
未知类 $y > y_c$, 则 y 归为 A 类。

未知类 $y < y_c$, 则 y 归为 B 类。

若 $\bar{y}(A) < \bar{y}(B)$

未知类 $y > y_c$, 则 y 归为 B 类。

未知类 $y < y_c$, 则 y 归为 A 类。





◆ 提高判别效果-判别变量的贡献率

(1) 计算综合距离系数 D^2

$$D^2 = (n_A + n_B - 2) \sum_{k=1}^m C_k d_k$$

(2) 分别计算各判别变量 x_k 的贡献率 $x_k(\%)$

$$x_k(\%) = \frac{c_k \cdot d_k}{D^2} (n_A + n_B - 2) \times 100\%$$



◆ 判别分析的显著性检验

$$F = \left[\frac{n_A n_B}{(n_A + n_B)(n_A + n_B - 2)} \right] \left[\frac{n_A + n_B - p - 1}{p} \right] \times D^2$$

p 为要素（变量）的个数。

F 服从自由度为 p 和 $(n_A + n_B - p - 1)$ 的 F 分布。



◆ 1.推求判别函数式

第一步，为求解判别系数 c_k 计算中间数据

- 计算各变量在各类（组）内的累加和
- 计算各变量的均值
- 计算各类（组）相应变量均值之差
- 计算各类（组）离均差平方和、离均差积和及两类（组）之和

第二步，计算判别系数 c_k 和判别函数式



- ◆ 2. 计算判别指标（临界值） y_c
- ◆ 3. 判别标准与对研究对象作判别
 - 作出判别标准
 - 对已知类作判别验证
 - 判别未知类点的归属



◆ 4. 判别能力的显著性检验

计算综合距离系数 D^2

计算F值

查F分布临界值表并比较

◆ 5. 计算各判别变量 (x_k) 的贡献率



Fisher判别基本思想是投影，即将原来在 R 维空间的自变量组合投影到维度较低的 D 维空间去，投影的原则是使得每一类内的离差尽可能小，而不同类间投影的离差尽可能大。然后使用典型变量计算出各类别在低维空间中的重心坐标，给出的判别式也是用于计算各样品的坐标值，最后用各观测点离各类别重心距离的远近来做出所属类别的判断。



Fisher判别的优势在于对分布、方差等都没有什么限制，应用范围较广。另外，用该判别方法建立的判别方程可以直接用手工计算的方法进行新观察对象的判别，这在许多时候是非常方便的。

§ 2地理系统的判别分析`



◆ 2、多类地理判别-问题抽象

总体A、B、C等多个类别；

分别抽出 n_1 、 n_2 、 n_3 、 n_4 、个样品；

测得每个样品的 m 个指标值 x_1, x_2, \dots, x_m ；

要求依据这些样本指标值，判别一个新样品属于哪一类。

§ 2地理系统的判别分析`



◆ 2、多类地理判别-Bayes判别

基本思想是认为所有 P 个类别都是空间中互斥的子域，每个观测都是空间中的一个点。

考虑先验概率的前提下，利用Bayes公式按照一定准则构造一个判别函数，分别计算该样品落入各个子域的概率，所有概率中最大的一类就被认为是该样品所属的类别。

Bayes判别强项是进行多类判别，但要求总体呈多元正态分布，应用范围窄。



最大后验准则

办公室新来了一个雇员小王，小王是好人还是坏人大家都在猜测。按人们主观意识，一个人是好人或坏人的概率均为0.5。坏人总是要做坏事，好人总是做好事，偶尔也会做一件坏事，一般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，小王是好人的概率有多大，你现在把小王判为何种人。

$$\begin{aligned} & P(\text{好人} / \text{做好事}) \\ &= \frac{P(\text{好人})P(\text{做好事} / \text{好人})}{P(\text{好人})P(\text{做好事} / \text{好人}) + P(\text{坏人})P(\text{做好事} / \text{坏人})} \\ &= \frac{0.5 \times 0.9}{0.5 \times 0.9 + 0.5 \times 0.2} = 0.82 \end{aligned}$$



$P(\text{坏人} / \text{做好事})$

$$= \frac{P(\text{坏人})P(\text{做好事} / \text{坏人})}{P(\text{好人})P(\text{做好事} / \text{好人}) + P(\text{坏人})P(\text{做好事} / \text{坏人})}$$

$$= \frac{0.5 \times 0.2}{0.5 \times 0.9 + 0.5 \times 0.2} = 0.18$$



- ◆ Bayes判别既考虑到各个总体出现的先验概率，又要考虑到错判造成的损失，
- ◆ 贝叶斯公式是一个我们熟知的公式

$$P(B_i | A) = \frac{P(A | B_i)P(B_i)}{\sum P(A | B_i)P(B_i)}$$



贝叶斯的统计思想总是假定对研究的对象已有一定的认识,常用先验概率分布来描述这种认识;然后抽取一个样本,用样本来修正已有的认识(先验概率分布),得到后验概率分布.各种统计推断都通过后验概率分布来进行.将贝叶斯思想用于判别分析就得到贝叶斯判别法.



设有 k 个总体 G_1, G_2, \dots, G_k . 假设事先对所研究的问题有一定的认识, 这种认识常用先验概率来描述, 即已知这 k 个总体各自出现的概率(验前概率)为 q_1, q_2, \dots, q_k (显然 $q_i > 0, q_1 + q_2 + \dots + q_k = 1$). 比如研究人群中得癌(G_1)和没有得癌(G_2)两类群体的问题, 由长期经验知: $q_1 = 0.001, q_2 = 0.999$. 这组验前概率 q_1, \dots, q_k 称为**先验概率**.



标准的贝叶斯判别法应该计算后验概率分布. 即计算当样品 X 已知时, 它属于 G_t 的概率, 记为 $P(G_t|X)$ (或 $P(t|X)$), 这个概率作为判别归类的准则, 其概率意义更为直观. 假定总体 G_t 的概率密度函数 $f_t(x)$ ($t=1, \dots, k$) 给定, 由条件概率的定义可以导出:

$$P(t|X) = P\{X \in G_t | X \text{ 已知}\} = \frac{q_t f_t(x)}{\sum_{i=1}^k q_i f_i(x)}.$$



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 。即当样本 x_0 发生时, 求他属于某类的概率。由贝叶斯公式计算后验概率, 有:

$$P(G_i | x_0) = \frac{q_i f_i(x_0)}{\sum q_j f_j(x_0)}$$

判别规则

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

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。



当抽取了一个未知总体的样品值 \mathbf{x} ，要判别它属于那个总体，
只要先计算出 k 个按先验概率加权的误判平均损失

$$h_j(\mathbf{x}) = \sum_{i=1}^k q_i C(j/i) f_i(\mathbf{x})$$

然后比较其大小，选取其中最小的，则判定样品属于该总体。

§ 2地理系统的判别分析`



◆ 3、逐步判别

变量选择和逐步判别：

变量的选择是判别分析中的一个重要的问题，变量选择是否恰当，是判别分析效果优劣的关键。

如果在某个判别问题中，将起最重要的变量忽略了，相应的判别函数的效果一定不好。

另一方面，如果判别变量个数太多，计算量必然大，会影响估计的精度。特别当引入了一些判别能力不强的变量时，还会严重地影响判别的效果。



逐步判别法采用有进有出的算法，即每一步都进行检验。首先，将判别能力最强的变量引进判别函数，而对较早进入判别函数的变量，随着其他变量的进入，其显著性可能发生变化，如果其判别能力不强了，则删除。

向前选入 开始时模型中没有变量。每一步，Wilks的统计量最小者，进入模型。当不再有未被选入的变量小于选入的临界值时，向前选入过程停止。



提问答疑