第14章 聚类分析与判别分析

介绍： 1、聚类分析

2、判别分析

分类学是人类认识世界的基础科学。聚类分析和判别分析是研究事物分类的基本方法，广泛地应用于自然科学、社会科学、工农业生产的各个领域。

14.1.1 聚类分析

- 根据事物本身的特性研究个体分类的方法，原则是同一类中的个体有较大的相似性，不同类中的个体差异很大。
- 根据分类对象的不同，分为样品（观测量）

聚类和变量聚类两种：

- 样品聚类：对观测量(Case)进行聚类（不同的目的选用不同的指标作为分类的依据，如选拔运动员与分课外活动小组）
- 变量聚类：找出彼此独立且有代表性的自变量，而又不丢失大部分信息。在生产活动中不乏有变量聚类的实例，如：衣服号码（身长、胸围、裤长、腰围）、鞋的号码。变量聚类使批量生产成为可能。

14.1.2 判别分析

- 判别分析是根据表明事物特点的变量值和它们所属的类，求出判别函数。根据判别函数对未知所属类别的事物进行分类的一种分析方法。
- 在自然科学和社会科学的各个领域经常遇到需要对某个个体属于哪一类进行判断。如动物学家对动物如何分类的研究和某个动物属于哪一类、目、纲的判断。
- 不同：判别分析和聚类分析不同的在于判别分析要求已知一系列反映事物特征的数值变量的值，并且已知各个体的分类（训练样本）。

14.1.3 聚类分析与判别分析的SPSS过程

- 在**Analyze✍Classify**下：
  - **K-Means Cluster**：　观测量快速聚类分析过程
  - **Hierarchical Cluster**：分层聚类（进行观测量聚类和变量聚类的过程
  - **Discriminant**：进行判别分析的过程

14.2 快速样本聚类过程(Quick Cluster)

- 使用 **k** 均值分类法对观测量进行聚类
- 可使用系统的默认选项或自己设置选项，如分为几类、指定初始类中心、是否将聚类结果或中间数据数据存入数据文件等。
- 快速聚类实例**(P342，data14-01a):**使用系统的默认值进行：对运动员的分类（分为**4**类）
  - **Analyze✍Classify✍K-Means Cluster**
    - **Variables: x1,x2,x3**
    - **Label Case By: no**

- **Number of Cluster: 4**
- 比较有用的结果：聚类结果形成的最后四类中心点**(Final Cluster Centers)**和每类的观测量数目（**Number of Cases in each Cluster**）
- 但不知每个运动员究竟属于哪一类？这就要用到**Save**选项

14.2 快速样本聚类过程(Quick Cluster)中的选项

.使用快速聚类的选择项：

- 类中心数据的输入与输出：**Centers**选项
- 输出数据选择项：**Save**选项
- 聚类方法选择项：**Method**选项
- 聚类何时停止选择项：**Iterate**选项
- 输出统计量选择项：**Option**选项

14.2 指定初始类中心的聚类方

法例题P343

- 数据同上（**data14-01a**）：以四个四类成绩突出者的数据为初始聚类中心(种子)进行聚类。类中心数据文件**data14-01b**（但缺一列**Cluster_**，不能直接使用，要修改）。对运动员的分类（还是分为**4**类）
- **Analyze✍Classify✍K-Means Cluster**
  - **Variables: x1,x2,x3**
  - **Label Case By: no**
  - **Number of Cluster: 4**
  - **Center: Read initial from: data14-01b**
  - **Save: Cluster membership**和Distance from Cluster Center
  - 比较有用的结果（可将结果与前面没有初始类中心比较）：
    - 聚类结果形成的最后四类中心点**(Final Cluster Centers)**
    - 每类的观测量数目（**Number of Cases in each Cluster**）
    - 在数据文件中的两个新变量**qc1_1**（每个观测量最终被分配到哪一类）和**qc1_2**（观测量与所属类中心点的距离）

14.3 分层聚类(Hierarchical Cluster)

- 分层聚类方法：
  - 分解法:先视为一大类，再分成几类
  - 凝聚法:先视每个为一类,再合并为几大类
- 可用于观测量(样本)聚类(Q型)和变量聚类(R型)
- 一般分为两步（自动,可从**Paste**的语句知道,**P359**）：
  - **Proximities**：先对数据进行的预处理(标准化和计算距离等)
  - **Cluster**：然后进行聚类分析
- 两种统计图：树形图(**Dendrogram**)和冰柱图(**Icicle**)
- 各类型数据的标准化、距离和相似性计算 **P348-354**
  - 定距变量、分类变量、二值变量
  - 标准化方法**p353**：**Z Scores**、**Range -1 to 1**、**Range 0 to 1**等

14.3.4 用分层聚类法进行观测量聚类实例P358

- 对**20**种啤酒进行分类(**data14-02**)，变量包括：**Beername**(啤酒名称)、**calorie**(热量)、**sodium**(钠含量)、**alcohol**(酒精含量)、 **cost**(价格)

- **Analyze→Classify →Hierarchical Cluster:**
  - **Variables: calorie,sodium,alcohol, cost** 成分和价格
  - **Label Case By: Beername**
  - **Cluster：Case, Q聚类**
  - **Display:** 选中**Statistics**，单击Statistics
    - Agglomeration Schedule 凝聚状态表
    - Proximity matrix：距离矩阵
    - Cluster membership：Single solution：4 显示分为4类时，各观测量所属的类
  - **Method: Cluster (Furthest Neighbor), Measure-Interval (Squared Euclidean distance), Transform Value (Range 0-1/By variable (值-最小值)/极差)**
  - **Plots: (Dendrogram) Icicle(Specified range of cluster, Start-1,Stop-4, by-1), Orientation (Vertical纵向作图)**
  - **Save: Cluster Membership(Single solution [4])**
  - 比较有用的结果：根据需要进行分类，在数据文件中的分类新变量**clu4_1**等

14.3.5 用分层聚类法进行变量聚类

- 变量聚类，是一种降维的方法，用于在变量众多时寻找有

代表性的变量，以便在用少量、有代表性的变量代替大变量集时，损失信息很少。

- 与进行观测量聚类雷同，不同点在于：
  - 选择**Variable**而非**Case**
  - **Save**选项失效，不建立的新变量

14.3.6 变量聚类实例1 P366

- 上面啤酒分类问题**data14-02**。
- **Analyze→Classify →Hierarchical Cluster:**
  - **Variables: calorie,sodium,alcohol, cost** 成分和价格
  - **Cluster：Variable, R**聚类
  - **Method：**
    - **Cluster Method ： Furthest Neighbor**
    - **Measure-Interval：Pearson Correlation**
    - **Transform Values：Z Score （By Variable）**
  - **Plots: Dendrogram** 树型图
  - **Statistics：** Proximity matrix：相关矩阵

- 比较有用的结果：根据相关矩阵和树型图，可知**calorie**(热量)和alcohol(酒精含量)的相关系数最大，首先聚为一类。从整体上看，聚为三类是比较好的结果。至于热量和酒精含量选择哪个作为典型指标代替原来的两个变量，可以根据专业知识或测度的难易程度决定。

14.3.6 变量聚类实例2 P368

- 有**10**个测试项目，分别用变量X_1-X_{10}表示，**50**名学生参加测试。想从**10**个变量中选择几个典型指标。**data14-03**
- **Analyze→Classify →Hierarchical Cluster:**
  - **Variables: X1-X10**
  - **Cluster：Variable, R**聚类
  - **Method：**
    - **Cluster Method ： Furthest Neighbor**
    - **Measure-Interval：Pearson Correlation**
  - **Plots: Dendrogram** 树型图
  - **Statistics：** Proximity matrix相关矩阵
  - 比较有用的结果：可以从树型图中看出聚类过程。具体聚为几类最为合理，根据专业知识来定。而每类中的典型指标的选择，可用**p370**的相关指数公式的计算，然后比较类中各个变量间的相关指数，哪个大，就选哪个变量作为此类的代表变量。

14.4 判别分析P374

- 判别分析的概念：是根据观测到的若干变量值，判断研究对象如何分类的方法。
- 要先建立判别函数

  $Y=a_1x_1+a_2x_2+...a_nx_n$，其中:Y为判别分数(判别值)，$x_1\ x_{2...}x_n$为反映研究对象特征的变量，$a_1\ a_2...a_n$为系数
- SPSS对于分为m类的研究对象，建立m个线性判别函数。对于每个个体进行判别时，把观测量的各变量值代入判别函数，得出判别分数，从而确定该个体属于哪一类，或计算属于各类的概率，从而判别该个体属于哪一类。还建立标准化和未标准化的典则判别函数。
- 具体见下面吴喜之教授有关判别分析的讲义

补充：聚类分析与判别分析

- 以下的讲义是吴喜之教授有关聚类分析与判别分析的讲义，我觉得比书上讲得清楚。
- 先是聚类分析一章
- 再是判别分析一章

聚类分析

分类

- 俗语说，物以类聚、人以群分。
- 但什么是分类的根据呢？
- 比如，要想把中国的县分成若干类，就有很多种分类法；
- 可以按照自然条件来分，
- 比如考虑降水、土地、日照、湿度等各方面；
- 也可以考虑收入、教育水准、医疗条件、基础设施等指标；
- 既可以用某一项来分类，也可以同时考虑多项指标来分类。

聚类分析

- 对于一个数据，人们既可以对变量（指标）进行分类(相当于对数据中的列分类)，也可以对观测值（事件，样品）来分类（相当于对数据中的行分类）。
- 比如学生成绩数据就可以对学生按照理科或文科成绩（或者综合考虑各科成绩）分类，
- 当然，并不一定事先假定有多少类，完全可以按照数据本身的规律来分类。

- 本章要介绍的分类的方法称为聚类分析（**cluster analysis**）。对变量的聚类称为**R**型聚类，而对观测值聚类称为**Q**型聚类。这两种聚类在数学上是对称的，没有什么不同。

饮料数据（**drink.sav** ）

- 16种饮料的热量、咖啡因、钠及价格四种变量

如何度量远近？

- 如果想要对100个学生进行分类，如果仅仅知道他们的数学成绩，则只好按照数学成绩来分类；这些成绩在直线上形成100个点。这样就可以把接近的点放到一类。
- 如果还知道他们的物理成绩，这样数学和物理成绩就形成二维平面上的100个点，也可以按照距离远近来分类。
- 三维或者更高维的情况也是类似；只不过三维以上的图形无法直观地画出来而已。在饮料数据中，每种饮料都有四个变量值。这就是四维空间点的问题了。

两个距离概念

- 按照远近程度来聚类需要明确两个概念：一个是点和点之间的距离，一个是类和类之间的距离。
- 点间距离有很多定义方式。最简单的是欧氏距离，还有其他的距离。
- 当然还有一些和距离相反但起同样作用的概念，比如相似性等，两点越相似度越大，就相当于距离越短。
- 由一个点组成的类是最基本的类；如果每一类都由一个点组成，那么点间的距离就是类间距离。但是如果某一类包含不止一个点，那么就要确定类间距离，
- 类间距离是基于点间距离定义的：比如两类之间最近点之间的距离可以作为这两类之间的距离，也可以用两类中最远点之间的距离作为这两类之间的距离；当然也可以用各类的中心之间的距离来作为类间距离。在计算时，各种点间距离和类间距离的选择是通过统计软件的选项实现的。不同的选择的结果会不同，但一般不会差太多。

向量$x=(x_1,\cdots, x_p)$与$y=(y_1,\cdots, y_p)$之间的距离或相似系数：

欧氏距离：

Euclidean

平方欧氏距离：

Squared Euclidean

夹角余弦(相似系数1)：

cosine

Pearson correlation

(相似系数2):

Chebychev: $Max_i|x_i-y_i|$

Block(绝对距离): $\Sigma_i|x_i-y_i|$

Minkowski:

当变量的测量值相差悬殊时,要先进行标准化. 如R为极差, s 为标准差, 则标准化的数据为每个观测值减去均值后再除以R或s. 当观测值大于0时, 有人采用Lance和Williams的距离

类G_p与类G_q之间的距离D_{pq}
($d(x_i,x_j)$表示点$x_i \in G_p$和$x_j \in G_q$之间的距离)

最短距离法:

最长距离法:

重心法:

离差平方和:

(Wald)

类平均法:

(中间距离, 可变平均法,可变法等可参考各书).

在用欧氏距离时, 有统一的递推公式

(假设G_r是从G_p和G_q合并而来):

Lance和Williams给出(对欧氏距离)统一递推公式:

$$D^2(k,r)=\alpha_p D^2(k,p)+\alpha_q D^2(k,q)$$
$$+\beta D^2(p,q) \quad +\gamma|D^2(k,p)-D^2(k,q)|$$

前面方法的递推公式可选择参数而得:

方法	α_i (i=p,q)	β	γ

最短距离 ½ 0 -1/2

最长距离 ½ 0 1/2

重心 n_i/n_r $-\alpha_p\alpha_q$ 0

类平均 n_i/n_r 0 0

离差平方和 $(n_i+n_k)/(n_r+n_k)$ $-n_k/(n_r+n_k)$ 0

中间距离 1/2 -1/4 0

可变法 (1-β)/2 β(<1) 0

可变平均 (1- β) n_i/n_r β(<1) 0

有了上面的点间距离和类间距离的概念，就可以介绍聚类的方法了。这里介绍两个简单的方法。

事先要确定分多少类：k-均值聚类

- 前面说过，聚类可以走着瞧，不一定事先确定有多少类；但是这里的**k-**均值聚类（**k-means cluster**，也叫快速聚类，**quick cluster**）却要求你先说好要分多少类。看起来有些主观，是吧！
- 假定你说分**3**类，这个方法还进一步要求你事先确定**3**个点为“聚类种子”**(SPSS**软件自动为你选种子**)**；也就是说，把这**3**个点作为三类中每一类的基石。
- 然后，根据和这三个点的距离远近，把所有点分成三类。再把这三类的中心（均值）作为新的基石或种子（原来的“种子”就没用了），重新按照距离分类。
- 如此叠代下去，直到达到停止叠代的要求（比如，各类最后变化不大了，或者叠代次数太多了）。显然，前面的聚类种子的选择并不必太认真，它们很可能最后还会分到同一类中呢。下面用饮料例的数据来做**k-**均值聚类。
-
- 假定要把这16种饮料分成3类。利用SPSS，只叠代了三次就达到目标了（计算机选的种

子还可以）。这样就可以得到最后的三类的中心以及每类有多少点

根据需要，可以输出哪些点分在一起。结果是：第一类为饮料1、10；第二类为饮料2、4、8、11、12、13、14；第三类为剩下的饮料3、5、6、7、9、15、16。

SPSS实现（聚类分析）

- **K-**均值聚类
- 以数据**drink.sav**为例，在**SPSS**中选择**Analyze－Classify－K-Menas Cluster**，
- 然后把**calorie**（热量）、**caffeine**（咖啡因）、**sodium**（钠）、**price**（价格）选入**Variables**，
- 在**Number of Clusters**处选择**3**（想要分的类数），
- 如果想要知道哪种饮料分到哪类，则选**Save**，再选**Cluster Membership**等。

- 注意**k-**均值聚类只能做**Q**型聚类，如要做**R**型聚类，需要把数据阵进行转置。

事先不用确定分多少类：分层聚类

- 另一种聚类称为分层聚类或系统聚类（**hierarchical cluster**）。开始时，有多少点就是多少类。
- 它第一步先把最近的两类（点）合并成一类，然后再把剩下的最近的两类合并成一类；
- 这样下去，每次都少一类，直到最后只有一大类为止。显然，越是后来合并的类，距离就越远。再对饮料例子来实施分层聚类。

对于我们的数据，SPSS输出的树型图为

聚类要注意的问题

- 聚类结果主要受所选择的变量影响。如果去掉一些变量，或者增加一些变量，结果会很不同。
- 相比之下，聚类方法的选择则不那么重要了。因此，聚类之前一定要目标明确。
- 另外就分成多少类来说，也要有道理。只要你高兴，从分层聚类的计算机结果可以得到任何可能数量的类。但是，聚类的目的是要使各类距离尽可能的远，而类中点的距离尽可能的近，而且分类结果还要有令人信服的解释。这一点就不是数学可以解决的了。

SPSS实现(聚类分析)

- 分层聚类
- 对drink.sav数据在SPSS中选择Analyze—Classify—Hierarchical Cluster,
- 然后把calorie（热量）、caffeine（咖啡因）、sodium（钠）、price（价格）选入Variables,
- 在Cluster选Cases（这是Q型聚类：对观测值聚类），如果要对变量聚类（R型聚类）则选

Variables,

- 为了画出树状图，选Plots，再点Dendrogram等。

啤酒成分和价格数据（data14-02)

啤酒名	热量	钠含量	酒精	价格
Budweiser	144.00	19.00	4.70	.43
Schlitz	181.00	19.00	4.90	.43
Ionenbrau	157.00	15.00	4.90	.48
Kronensourc	170.00	7.00	5.20	.73
Heineken	152.00	11.00	5.00	.77
Old-milnaukee	145.00	23.00	4.60	.26
Aucsberger	175.00	24.00	5.50	.40
Strchs-bohemi	149.00	27.00	4.70	.42
Miller-lite	99.00	10.00	4.30	.43
Sudeiser-lich	113.00	6.00	3.70	.44
Coors	140.00	16.00	4.60	.44
Coorslicht	102.00	15.00	4.10	.46
Michelos-lich	135.00	11.00	4.20	.50
Secrs	150.00	19.00	4.70	.76
Kkirin	149.00	6.00	5.00	.79

Pabst-extra-l 68.00 15.00 2.30 .36

Hamms 136.00 19.00 4.40 .43

Heilemans-old 144.00 24.00 4.90 .43

Olympia-gold- 72.00 6.00 2.90 .46

Schlite-light 97.00 7.00 4.20 .47

Statistics→Classify →Hierarchical Cluster:

Variables:啤酒名和成分价格等

Cluster(Case, Q型聚类)

Display: (Statistics)(Agglomeration Schedule凝聚状态表), (Proximity matrix), Cluster membership(Single solution, [4])

Method: Cluster (Furthest Neighbor), Measure-Interval (Squared Euclidean distance), Transform Value (Range 0-1/By variable (值-最小值)/极差)

Plots: (Dendrogram) Icicle(Specified range of cluster, Start-1,Stop-4, by-1), Orientation (Vertical)

Save: Cluster Membership(Single solution [4])

啤酒例子

下表(Proximity matrix)中行列交叉点为两种啤酒之间各变量的欧氏距离平方和

凝聚过程:Coefficients为不相似系数,由于是欧氏距离,小的先合并.

分为四类的聚类结果

冰柱图(icicle)

聚类树型图

学生测验数据（data14-03)

50个学生，X1-X10个测验项目

要对这10个变量进行变量聚类

（ **R** 型聚类），过程和**Q**型聚类（观测量聚类，对**cases)**一样

Statistics→Classify →Hierarchical Cluster:

Variables:x1-x10

Cluster(Variable, R型聚类)

Display: (Statistics) (Proximity matrix), Cluster membership(Single solution, [2])

Method: Cluster (Furthest Neighbor), Measure-Interval (Pearson correlation，用Pearson相关系数),

Plots: Icicle(All Cluster)

学生测验例子

下表**(Proximity matrix)**中行列

交叉点为两个变量之间变量的欧氏距离平方和

分为两类的聚类结果

冰柱图**(icicle)**

判别分析

判别

- 有一些昆虫的性别很难看出，只有通过解剖才能够判别；
- 但是雄性和雌性昆虫在若干体表度量上有些综合的差异。于是统计学家就根据已知雌雄的昆虫体表度量（这些用作度量的变量亦称为预测变量）得到一个标准，并且利用这个标准来判别其他未知性别的昆虫。

- 这样的判别虽然不能保证百分之百准确，但至少大部分判别都是对的，而且用不着杀死昆虫来进行判别了。

判别分析(discriminant analysis)

- 这就是本章要讲的是判别分析。
- 判别分析和前面的聚类分析有什么不同呢？
- 主要不同点就是，在聚类分析中一般人们事先并不知道或一定要明确应该分成几类，完全根据数据来确定。
- 而在判别分析中，至少有一个已经明确知道类别的“训练样本”，利用这个数据，就可以建立判别准则，并通过预测变

量来为未知类别的观测值进行判别了。

判别分析例子

- 数据disc.sav:企图用一套打分体系来描绘企业的状况。该体系对每个企业的一些指标（变量）进行评分。
- 这些指标包括：企业规模(is)、服务(se)、雇员工资比例(sa)、利润增长(prr)、市场份额(ms)、市场份额增长(msr)、流动资金比例(cp)、资金周转速度(cs)等等。
- 另外，有一些企业已经被某杂志划分为上升企业、稳定企业和下降企业。
- 我们希望根据这些企业的上述变量的打分和它们已知的类别（三个类别之一：group-1代表上升，group-2代表稳定，group-3代表下降）找出一个分类标准，以对没有被该刊物分类的企业进行分类。
- 该数据有90个企业（90个观测值），其中30个属于上升型，30个属于稳定型，30个属于下降型。这个数据就是一个“训练样本”。

Disc.sav数据

根据距离的判别（不用投影）

- Disc.sav数据有8个用来建立判别标准(或判别函数)的（预测）变量，另一个（group）是类别。
- 因此每一个企业的打分在这8个变量所构成的8维空间中是一个点。这个数据有90个点，
- 由于已经知道所有点的类别了，所以可以求得每个类型的中心。这样只要定义了如何计算距离，就可以得到任何给定的点（企业）到这三个中心的三个距离。
- 显然，最简单的办法就是离哪个中心距离最近，就属于哪一类。通常使用的距离是所谓的Mahalanobis距离。用来比较到各个中心距离的数学函数称为判别函数(discriminant function).这种根据远近判别的方法，原理简单，直观易懂。

Fisher判别法(先进行投影)

- 所谓**Fisher**判别法，就是一种先投影的方法。
- 考虑只有两个（预测）变量的判别分析问题。
- 假定这里只有两类。数据中的每个观测值是二维空间的一个点。见图（下一张幻灯片）。
- 这里只有两种已知类型的训练样本。其中一

类有**38**个点（用“**o**”表示），另一类有**44**个点（用“*”表示）。按照原来的变量（横坐标和纵坐标），很难将这两种点分开。

- 于是就寻找一个方向，也就是图上的虚线方向，沿着这个方向朝和这个虚线垂直的一条直线进行投影会使得这两类分得最清楚。可以看出，如果向其他方向投影，判别效果不会比这个好。
- 有了投影之后，再用前面讲到的距离远近的方法来得到判别准则。这种首先进行投影的判别方法就是**Fisher**判别法。

逐步判别法(仅仅是在前面的方法中加入变量选择的功能)

- 有时，一些变量对于判别并没有什么作用，为了得到对判别最合适的变量，可以使用逐步判别。也就是，一边判别，一边引进判别能力最强的变量，
- 这个过程可以有进有出。一个变量的判别能力的判断方法有很多种，主要利用各种检验，例如Wilks’ Lambda、Rao’s V、The Squared Mahalanobis Distance、Smallest F ratio或The Sum of Unexplained Variations等检验。其细节这里就不赘述了；这些不同方法可由统计软件的各种选项来实现。逐步判别的其他方面和前面的无异。

Disc.sav例子

- 利用SPSS软件的逐步判别法淘汰了不显著的流动资金比例(cp)，还剩下七个变量is，se，sa，prr，ms，msr，cs，得到两个典则判别函数（Canonical Discriminant Function Coefficients）：
  - 0.035IS+3.283SE+0.037SA-0.007PRR+0.068MS-0.023MSR-0.385CS-3.166
  - 0.005IS+0.567SE+0.041SA+0.012PRR+0.048MS+0.044MSR-0.159CS-4.384

这两个函数实际上是由**Fisher**判别法得到的向两个方向的投影。这两个典则判别函数的系数是下面的**SPSS**输出得到的：

Disc.sav例子

- 根据这两个函数，从任何一个观测值（每个观测值都有7个变量值）都可以算出两个数。把这两个数目当成该观测值的坐标，这样数据中的150个观测值就是二维平面上的150个点。它们的点图在下面图中。

Disc.sav例子

- 从上图可以看出，第一个投影（相应于来自

于第一个典则判别函数横坐标值）已经能够很好地分辨出三个企业类型了。这两个典则判别函数并不是平等的。其实一个函数就已经能够把这三类分清楚了。SPSS的一个输出就给出了这些判别函数（投影）的重要程度：

前面说过，投影的重要性是和特征值的贡献率有关。该表说明第一个函数的贡献率已经是**99%**了，而第二个只有**1%**。当然，二维图要容易看一些。投影之后，再根据各点的位置远近算出具体的判别公式（**SPSS**输出）：

Disc.sav例子

- 具体的判别公式（SPSS输出），由一张分类函数表给出：

该表给出了三个线性分类函数的系数。把每个观测点带入三个函数，就可以得到分别代表三类的三个值，哪个值最大，该点就属于相应的那一类。当然，用不着自己去算，计算机软件的选项可以把这些训练数据的每一个点按照这里的分类法分到某一类。当然，我们一开始就知道这些训练数据的各个观测值的归属，但即使是这些训练样本的观测值（企业）按照这里推导出的分类函数来分类，也不一定全都能够

正确划分。

Disc.sav例子

- 下面就是对我们的训练样本的分类结果（SPSS）：

误判和正确判别率

- 从这个表来看，我们的分类能够**100%**地把训练数据的每一个观测值分到其本来的类。
- 该表分成两部分；上面一半（**Original**）是用从全部数据得到的判别函数来判断每一个点的结果（前面三行为判断结果的数目，而后三行为相应的百分比）。
- 下面一半（**Cross validated**）是对每一个观测值，都用缺少该观测的全部数据得到的判别函数来判断的结果。
- 这里的判别结果是**100%**判别正确，但一般并不一定。

Disc.sav例子

- 如果就用这个数据，但不用所有的变量，而只用4个变量进行判别：企业规模（is）、服务(se)、雇员工资比例(sa)、资金周转速度

(cs)。结果的图形和判别的正确与否就不一样了。下图为两个典则判别函数导出的150个企业的二维点图。它不如前面的图那么容易分清楚了

原先的图

Disc.sav例子

- 下面是基于4个变量时分类结果表：

- 这个表的结果是有87个点（96.7%）得到正确划分，有3个点被错误判别；其中第二类有两个被误判为第一类，有一个被误判为第三类。

判别分析要注意什么？

- 训练样本中必须有所有要判别的类型，分类必须清楚，不能有混杂。
- 要选择好可能由于判别的预测变量。这是最重要的一步。当然，在应用中，选择的余地不见得有多大。
- 要注意数据是否有不寻常的点或者模式存在。还要看预测变量中是否有些不适宜的；这可以用单变量方差分析（ANOVA）和相关

分析来验证。

- 判别分析是为了正确地分类，但同时也要注意使用尽可能少的预测变量来达到这个目的。使用较少的变量意味着节省资源和易于对结果进行解释。

判别分析要注意什么？

- 在计算中需要看关于各个类的有关变量的均值是否显著不同的检验结果（在**SPSS**选项中选择**Wilks' Lambda**、**Rao's V**、**The Squared Mahalanobis Distance**或**The Sum of Unexplained Variations**等检验的计算机输出），以确定是否分类结果是仅仅由于随机因素。
- 此外成员的权数（**SPSS**用**prior probability**，即“先验概率”，和贝叶斯统计的先验概率有区别）需要考虑；一般来说，加权要按照各类观测值的多少，观测值少的就要按照比例多加权。
- 对于多个判别函数，要弄清各自的重要性。
- 注意训练样本的正确和错误分类率。研究被误分类的观测值，看是否可以找出原因。

SPSS选项

- 打开**disc.sav**数据。然后点击**Analyze**－

Classify－Discriminant，

- 把**group**放入**Grouping Variable**，再定义范围，即在**Define Range**输入**1－3**的范围。然后在**Independents**输入所有想用的变量；但如果要用逐步判别，则不选**Enter independents together**，而选择**Use stepwise method**，
- 在方法（**Method**）中选挑选变量的准则（检验方法；默认值为**Wilks' Lambda**）。
- 为了输出**Fisher**分类函数的结果可以在**Statistics**中的**Function Coefficient**选**Fisher**和**UnStandardized**（点则判别函数系数），在**Matrices**中选择输出所需要的相关阵；
- 还可以在**Classify**中的**Display**选**summary table, Leave-one-out classification**；注意在**Classify**选项中默认的**Prior Probability**为**All groups equal**表示所有的类都平等对待，而另一个选项为**Compute from group sizes**，即按照类的大小加权。
- 在**Plots**可选 **Combined-groups, Territorial map**等。

14.4.3 判别分析实例P379

- 鸢尾花数据(花瓣,花萼的长宽) **5**个变量:花瓣长**(slen)**,花瓣宽**(swid)**, 花萼长**(plen)**, 花萼宽**(pwid)**, 分类号**(1:Setosa, 2:Versicolor, 3:Virginica)(data14-04)**

Statistics→Classify →Discriminant:

Variables: independent (slen,swid,plen,pwid) Grouping(spno) Define range(min-1,max-3)

Classify: prior probability(All group equal) use covariance matrix (Within-groups) Plots (Combined-groups, Separate-groups, Territorial map) Display (Summary table)

Statistics: Descriptive (Means) Function Coefficients (Fisher's, Unstandardized) Matrix (Within-groups correlation, Within-groups covariance, Separate-groups covariance, Total covariance)

Save: (Predicted group membership,

Discriminant Scores, Probability of group membership)

鸢尾花数据(数据分析过程简明表)

鸢尾花数据(原始数据的描述)

鸢尾花数据(合并类内相关阵和协方差阵)

鸢尾花数据(总协方差阵)

鸢尾花数据(特征值表)

Eigenvalue:用于分析的前两个典则判别函数的特征值, 是组间平方和与组内平方和之比值. 最

大特征值与组均值最大的向量对应, 第二大特征值对应着次大的组均值向量

典则相关系数(**canonical correlation**):是组间平方和与总平方和之比的平方根.被平方的是由组间差异解释的变异总和的比.

鸢尾花数据(**Wilks' Lambda**统计量)

检验的零假设是各组变量均值相等. **Lambda**接近**0**表示组均值不同,接近**1**表示组均值没有不同. **Chi-square**是**lambda**的卡方转换, 用于确定其显著性.

鸢尾花数据(有关判别函数的输出)

标准化的典则判别函数系数(使用时必须用标准化的自变量)

鸢尾花数据(有关判别函数的输出)

典则判别函数系数

鸢尾花数据(有关判别函数的输出)

这是类均值(重心)处的典则判别函数值

这是典则判别函数(前面两个函数)在类均值(重心)处的值

鸢尾花数据(用判别函数对观测量分类结果)

先验概率(没有给)

费歇判别函数系数

把自变量代入三个式子,哪个大归谁.

Territorial Map

Canonical Discriminant

Function 2

-12.0 -8.0 -4.0 .0 4.0 8.0 12.0

超蝌蝌蝌蝌超蝌蝌蝌蝌超蝌蝌蝌蝌超蝌蝌蝌蝌超蝌蝌蝌蝌超蝌蝌蝌蝌

12.0 12 23

12 23

12 23

12 23

12 23

12 23

8.0 12 23

12 23

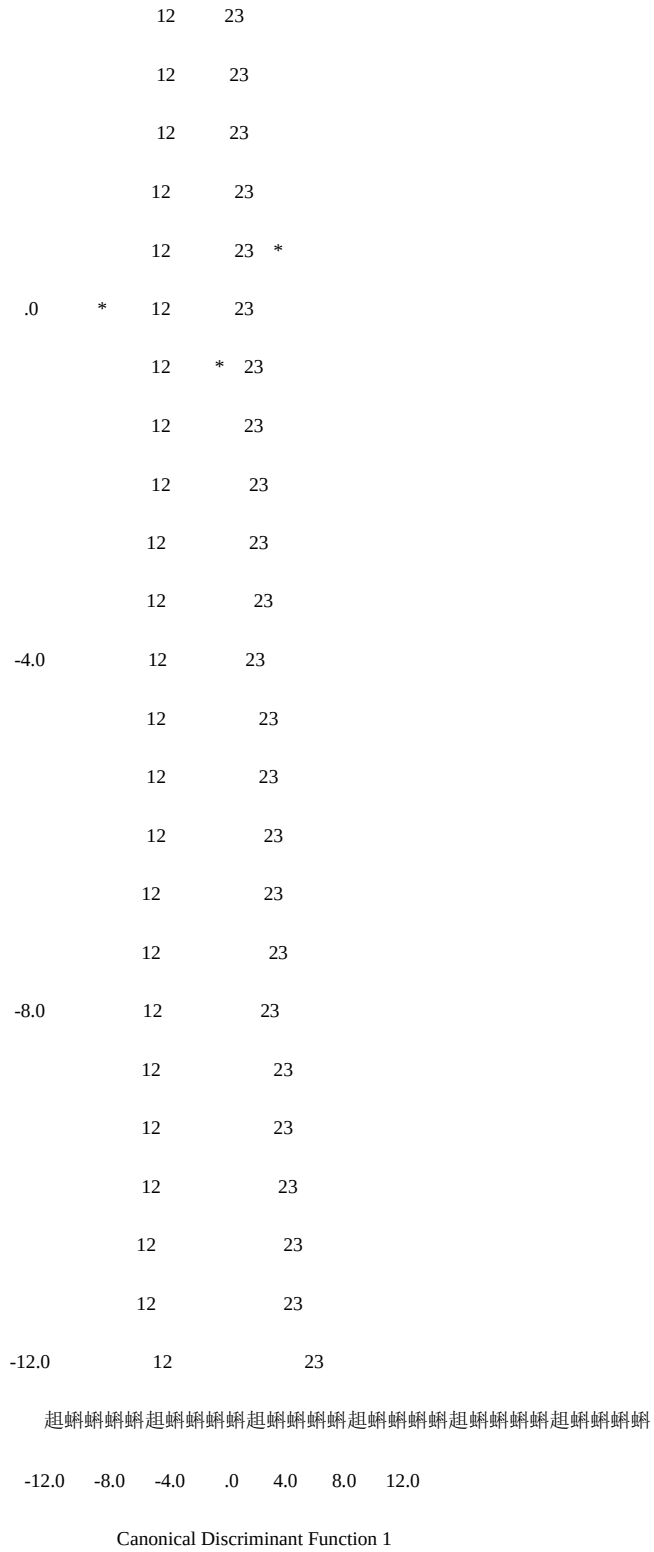
12 23

12 23

12 23

12 23

4.0 12 23



Symbols used in territorial map

Symbol	Group	Label
1	1	刚毛鸢尾花
2	2	变色鸢尾花
3	3	佛吉尼亚鸢尾花

*　Indicates a group centroid

鸢尾花数据

Territory Map(区域图)

Canonical Discriminate Function 1

Versus

Canonical Discriminate Function 2

三种鸢尾花的典则变量值把一个典则变量组成的坐标平面分成三个区域. *为中心坐标.

鸢尾花数据(预测分类结果小结)

可以看出分错率