第五章、判别分析

November 20, 2018

判别分析是用于判断样品所属类型的一种统计方法.

例如

1. 在气象学中,根据已有气象资料来判断明天是阴天还是晴天,是有雨还是无雨.
2. 在经济学中,根据人均国民收入、人均工农业产值、人均消费水平等多种指标来判定一个国家的经济发展程度所属类型
3. 在市场预测中，根据以往调查所得的种种指标，判断下季度的产品是畅销、平常还是滞销等等

判别分析例一

Fisher于1936年发表的鸢尾花(Iris)数据被广泛地作为判别分析的例子。数据是对3种鸢尾花：刚毛鸢尾花(setosa第1组)、变色鸢尾花(versicolor第2组) 和佛吉尼亚鸢尾花(virginica第3组) 各抽取一个容量为50的样本，测量(单位为mm):

- 花萼长（sepallen）$x1$
- 花萼宽（sepalwid）$x2$
- 花瓣长（petallen）$x3$
- 花瓣宽（petalwid）$x4$
- 分组标记为S.

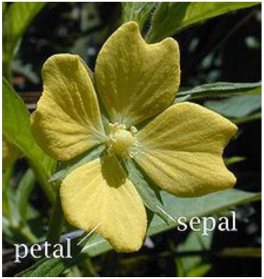
判别分析的目标：根据收集到的资料建立判别函数，然后对新来的鸢尾花，根据测量花萼长、花萼宽、花瓣长和花瓣宽来判别其所属的类型。

Example: Classification of Iris flowers

Iris setosa

Iris versicolor

Iris virginica



Classify according to sepal/petal length/width

判别分析例二

从已有的研究来看，四种不同的检测指标对胃癌均有一定的预报作用

- VacA, CagA 和UreaA, UreaB抗体的水平
- 血清肿瘤标志物CEA,CA199,CA724
- 血清蛋白质
- 血清PGⅠ，PGⅡ和PGⅠ/PGⅡ比值

但单个指标的预报功能都非常弱，因此，不能引起病人足够的重视。

判别分析的目标：通过上面的四个方面建立胃癌风险预警综合模型(判别函数)，对病人胃癌可能进行分级预报。

无论在哪个领域,判别分析问题都可以这样描述:

- 所谓判别方法,就是给出空间$\mathbb{R}^m$ 的一种划分: $D = D_1, \cdots,$ D_k . 不同一划分分别对应不同的判别方法.
- 设有k个m维总体$G_1, G_2, \cdots, G_k$, 其分布特征已知(如已知分布函数分别为$F_1(x)$, $F_2(x)$, $\cdots$, $F_k(x)$,或知道来自各个总体的训练样本). 对给定的一个新样品X, 我们要判断它来自哪个总体.
- 几个常用的判别方法有:距离判别、Bayes判别、Fisher判别、逐步判别、序贯判别等.

5.1 距离判别

距离判别的基本思想:样品和哪个总体距离最近,就判断它属于哪个总体.

距离判别法也称直观判别法.

一、马氏距离

已知有两个类G_1和G_2,

- G_1是设备A生产的产品,设备A 的产品质量高,其平均耐磨度$\mu^{(1)}$=80,反映设备精度的方差σ_1^2=0.25;
- G_2设备B生产的同类产品,其平均耐磨度$\mu^{(2)}$=75,设备精度的方差σ_2^2=4.
- 今有一产品X_0,测得耐磨度$\mu^{(0)}$=78,判断该产品来自哪个总体?

直观: $|78-80|=2<|75-78|=3$, 所以直观上X_0更像是设备A生产的.但没有考虑设备精度的方差.

相对性的距离 考虑产品的方差

记X_0与G_1和G_2的相对平方距离为$d_1^2(x)$和$d_2^2(x)$,则

$$
\begin{aligned}
d_1^2(x_0) &= \frac{(x_0-\mu^{(1)})^2}{\sigma_1^2} = \frac{(78-80)^2}{0.25} = 16. \\
d_2^2(x_0) &= \frac{(x_0-\mu^{(2)})^2}{\sigma_2^2} = \frac{(78-75)^2}{4.00} = 2.25.
\end{aligned}
$$

由于$d_2^2(x_0) < d_1^2(x_0)$, 所以判断X_0为设备B生产的更为合理.

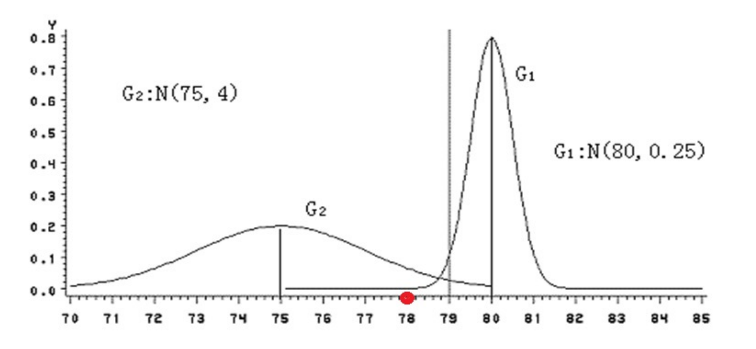


Figure: 两个正态总体距离判别法的示意图

一般情形：

假设总体G_1的分布为$N(\mu^{(1)},\sigma_1^2)$, 总体G_2的分布为$N(\mu^{(2)},\sigma_2^2)$, 则利用相对距离可以找出分界点(假设$\mu^{(2)}<\mu^{(1)}$), 令

$$\frac{(x-\mu^{(1)})^2}{\sigma_1^2}=\frac{(x-\mu^{(2)})^2}{\sigma_2^2}\rightarrow x=\frac{\mu^{(1)}\sigma_2+\mu^{(2)}\sigma_1}{\sigma_1+\sigma_2}\overset{def}{=}\mu^*$$

$$\mu_*\overset{def}{=}\frac{\mu^{(1)}\sigma_2-\mu^{(2)}\sigma_1}{\sigma_1-\sigma_2}$$

而按这种距离最近的判别准则为:

$$\begin{cases} X\in G_1, & \mu^*<x<\mu_* \\ X\in G_2, & x\le\mu^* \ \text{ or } \ x\ge\mu_* \end{cases}$$

给出m元总体的马氏距离的定义

定义5.1.1

设总体G为m元总体,均值向量为$\mu = (\mu_1, \mu_2, \cdots, \mu_m)'$, 协方差矩阵为$\Sigma = (\sigma_{ij})_{m\times m}$, 则样品X与总体$G$ 的马氏距离定义为

$$d^2(X, G) = (X - \mu)'\Sigma^{-1}(X - \mu)$$

二、两总体的距离判别

设有两个总体G_1和G_2,来自G_i 的训练样本为

$$X_{(t)}^{(i)} = (x_{(t1)}^{(i)}, x_{(t2)}^{(i)}, \cdots, x_{(tm)}^{(i)})' \qquad t = (1, 2, \cdots, n_i),\ i = (1, 2),$$

则均值向量$\mu^{(i)}$, $i = 1, 2$的估计量为

$$\bar{X}^{(i)} = \left(\frac{1}{n_i} \sum_{t=1}^{n_i} x_{t1}^{(i)}, \cdots, \frac{1}{n_i} \sum_{t=1}^{n_i} x_{tm}^{(i)} \right)' = \left(\bar{x}_1^{(i)}, \bar{x}_2^{(i)}, \cdots, \bar{x}_m^{(i)} \right)'$$

总体G_i的协方差阵Σ_i的估计S_i(称为**组内协方差阵**)为

$$S_i = \frac{1}{n_i - 1} A_i = (s_{lj}^{(i)})_{m \times m}, \ \ i = 1, 2.$$

$$\text{其中}A_i=\sum_{t=1}^{n_i}(X_{(t)}^{(i)}-\bar{X}^{(i)})(X_{(t)}^{(i)}-\bar{X}^{(i)})'\text{称为组内离差阵}$$

$$s_{lj}^{(i)}=\frac{1}{n_i-1}\sum_{t=1}^{n_i}(x_{tl}^{(i)}-\bar{x}^{(i)})(x_{tj}^{(i)}-\bar{x}^{(i)}),$$

$$i=1,2,\ \ l,j=1,\cdots,m.$$

当假定$\Sigma_1=\Sigma_2=\Sigma$时,协方差矩阵的估计为

$$S=\frac{1}{n-2}\sum_{i=1}^{2}A_i=(s_{lj})_{m\times m}$$

称S为**合并样本协方差阵**，其中

$$s_{lj}=\frac{1}{n-2}\sum_{i=1}^{2}\sum_{t=1}^{n_i}(x_{tl}^{(i)}-\bar{x}^{(i)})(x_{tj}^{(i)}-\bar{x}^{(i)})$$

$$(l,j)=(1,\cdots,m)$$

1.$\Sigma_1 = \Sigma_2$时的判别方法(线性判别)

计算样品X到两个总体的距离$d^2(X,G_1)$和$d^2(X,G_2)$,判别准则为:

$$\begin{cases} X \in G_1, & d^2(X,G_1) < d^2(X,G_2) \\ X \in G_2, & d^2(X,G_1) \geq d^2(X,G_2) \end{cases}$$

这里的距离是马氏距离,利用马氏距离的定义及相同的协方差阵可简化马氏距离的计算公式:

$$\begin{aligned} d^2(X,G_i) &= (X-\bar{X}^{(i)})'S^{-1}(X-\bar{X}^{(i)}) \\ &= X'S^{-1}X - 2[(S^{-1}\bar{X}^{(i)})'X - \frac{1}{2}(\bar{X}^{(i)})'S^{-1}\bar{X}^{(i)}] \\ &= X'S^{-1}X - 2Y_i(X) \end{aligned}$$

所以计算马氏距离可改为对X的线性函数$Y_i(X)$的计算:

$$Y_i(X) = (S^{-1}\bar{X}^{(i)})'X - \frac{1}{2}(\bar{X}^{(i)})'S^{-1}\bar{X}^{(i)}$$

- $Y_i(X)$称为**线性判别函数**,
- $a_i = S^{-1}\bar{X}^{(i)}$称为**线性系数向量**,
- $c_i = -\frac{1}{2}(\bar{X}^{(i)})'S^{-1}\bar{X}^{(i)}$ 称为**常数项**

若考察两个马氏距离之差,

$$\begin{aligned} & d^2(X,G_2)-d^2(X,G_1) \\ = \quad & 2\left(X-\frac{1}{2}(\bar{X}^{(1)}+\bar{X}^{(2)})\right)' S^{-1}\left(\bar{X}^{(1)}-\bar{X}^{(2)}\right) \\ \overset{def}{=} \quad & 2W(X). \end{aligned}$$

其中

$$W(X)=(X-X^*)'S^{-1}(\bar{X}^{(1)}-\bar{X}^{(2)})$$

$$X^*=\frac{1}{2}(\bar{X}^{(1)}+\bar{X}^{(2)})$$

则判别准则还可改为:

$$\begin{cases} X\in G_1, & W(X)>0 \\ X\in G_2, & W(X)\le 0 \end{cases}$$

$W(X)$ 也称为线性判别函数.

考察$m=1$的特殊情况,设两总体为正态总体,分布为$N(\mu^{(1)},\sigma^2)$和$N(\mu^{(2)},\sigma^2)$,这是判别函数为

$$W(x)=(x-\frac{1}{2}(\mu^{(1)}+\mu^{(2)}))\sigma^{-2}(\mu^{(1)}-\mu^{(2)})$$

错判问题

记$P(2|1)$表示为属于G_1的样品被误判为G_2 的概率.

$$\begin{aligned} P(2|1) &= \Pr(X\text{判给}G_2|X\in G_1)=\Pr\left(X\geq\frac{\mu^{(1)}+\mu^{(2)}}{2}\right) \\ &= 1-\Phi(\frac{\mu^{(1)}-\mu^{(2)}}{2\sigma}) \\ P(1|2) &= P(2|1)=1-\Phi(\frac{\mu^{(1)}-\mu^{(2)}}{2\sigma}). \end{aligned}$$

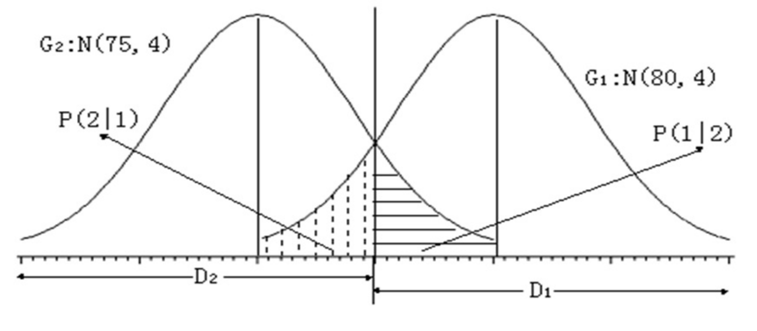


Figure: 错判概率示意图

所以当两个总体的均值靠的很近时,错判概率会很大,这时作判别分析没意义,只有当两总体均值有显著差异时,作判别分析才有意义.

注意：在进行判别分析前，通常要进行总体均值检验，只有在均值检验有差异的情形下，判别分析才有意义。

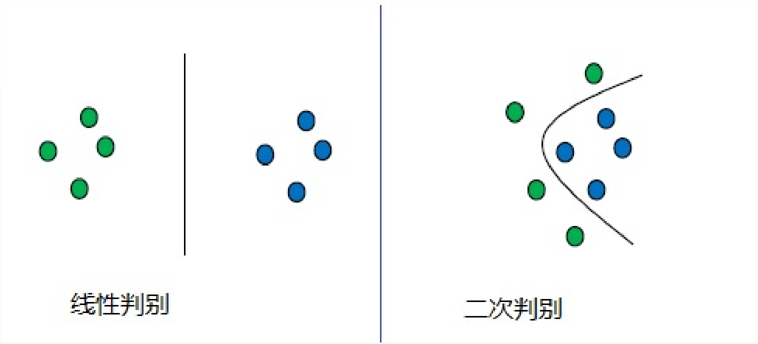
2.$\Sigma_1 \neq \Sigma_2$时的判别方法(二次判别)

计算判别函数,令

$$
\begin{aligned}
W(X) &= d^2(X, G_2) - d^2(X, G_1) \\
&= (X - \mu^{(2)})' S_2^{-1} (X - \mu^{(2)}) - (X - \mu^{(1)})' S_1^{-1} (X - \mu^{(1)}) \\
&= X'(S_2^{-1} - S_1^{-1})X - 2(\mu^{(2)} - \mu^{(1)})' X \\
&\quad + (\mu^{(2)'} S_2^{-1} \mu^{(2)} - \mu^{(1)'} S_1^{-1} \mu^{(1)}) \\
&\triangleq Z(X) + Z_0.
\end{aligned}
$$

其中$Z(X)$为X的二次函数(因$\Sigma_1 \neq \Sigma_2$), Z_0 是一常数.判别准则可以写为:

$$
\begin{cases}
X \in G_1, & W(X) > 0, \\
X \in G_2, & W(X) \leq 0.
\end{cases}
$$



注意：

- 经检验不能拒绝协方差矩阵相等，一般可采用线性判别. 但仍可采用二次判别，二次判别的正确率高于线性判别.
- 如果数据量很大，二次判别所需要的时间远大于线性差别.

当$m=1$时,设G_i的分布为$N(\mu^{(i)},\sigma_i^2) \quad (i=1,2)$. 不妨设$\mu^{(2)}<\mu^{(1)}$ 这时马氏距离的平方根是

$$d_i(x)=\frac{|x-\mu^{(i)}|}{\sigma_i}.$$

当观测值满足$\mu^{(2)}<x<\mu^{(1)}$ 时,

$$d_2(x)-d_1(x) \quad = \quad \frac{\sigma_1+\sigma_2}{\sigma_1\sigma_2}(x-\mu^*),$$

其中

$$\mu^*=\frac{\mu^{(1)}\sigma_2+\mu^{(2)}\sigma_1}{\sigma_1+\sigma_2}.$$

当$m=1$时,设G_i的分布为$N(\mu^{(i)},\sigma_i^2)\quad(i=1,2)$. 不妨设$\mu^{(2)}<\mu^{(1)}$ 这时马氏距离的平方根是

$$d_i(x)=\frac{|x-\mu^{(i)}|}{\sigma_i}.$$

当观测值满足$\mu^{(2)}<x<\mu^{(1)}$ 时,

$$d_2(x)-d_1(x) = \frac{\sigma_1+\sigma_2}{\sigma_1\sigma_2}(x-\mu^*),$$

其中

$$\mu^*=\frac{\mu^{(1)}\sigma_2+\mu^{(2)}\sigma_1}{\sigma_1+\sigma_2}.$$

这时判别准则为

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

R中线性判别：加载MASS包

模型语句：

$lda(x, grouping, prior = proportions, tol = 1.0e-4, method, CV = FALSE, nu, ...)$

- prior=先验概率向量，缺省为先验概率相等
- tol=容忍度，回归分析中关于共线性诊断所提到的一样
- method="moment"均值方差均为标准估计，"mle"为极大似然估计

判别语句：

$predict(object, newdata, prior = object\$prior, dimen, method = c("plug-in", "predictive", "debiased"), ...)$

- object为lda输出的结果，
- newdata即为要分析的新的数据库

例5.1.1(盐泉含钾性判别)

表 5.1 盐泉的特征数值

盐泉类别	序号	X1	X2	X3	X4	类别号
第一类含钾盐泉（A 盆地）	1	13.85	2.79	7.80	49.60	A
	2	22.31	4.67	12.31	47.80	A
	3	28.82	4.63	16.18	62.15	A
	4	15.29	3.54	7.50	43.20	A
	5	28.79	4.90	16.12	58.10	A
第二类含钠盐泉（B 盆地）	6	2.18	1.06	1.22	20.60	B
	7	3.85	0.80	4.06	47.10	B
	8	11.40	0.00	3.50	0.00	B
	9	3.66	2.42	2.14	15.10	B
	10	12.10	0.00	5.68	0.00	B
待判盐泉	11	8.85	3.38	5.17	26.10	
	12	28.60	2.40	1.20	127.0	
	13	20.70	6.70	7.60	30.20	
	14	7.90	2.40	4.30	33.20	
	15	3.19	3.20	1.43	9.90	
	16	12.40	5.10	4.43	24.60	
	17	16.80	3.40	2.31	31.30	
	18	15.00	2.70	5.02	64.00	

- 第一步：均值差异性检验：$H_0 : \mu_1 = \mu_2$ 的$HotellingT2$ 统计量为14.46436, $P = 0.0059 < 0.01$, 说明两类盐泉有显著差异. 具体为：读入数据
  - $D511 <-read.table(''ex511.txt'', header = F)$
  - $D511A <-D511[1 : 5, 1 : 4]$
  - $D511B <-D511[6 : 10, 1 : 4]$
  - $HotellingsT2(D511A, D511B)$
  - 输出结果：T.2 = 14.4644, df1 = 4, df2 = 5, p-value = 0.005891.

- 第二步：判别分析：用R中加载的MASS包，
  - $results <-lda(V5 \sim V1+V2+V3+V4, D511)$
  - 输出结果：

```
Call:
lda(V5 ~ V1 + V2 + V3 + V4, data = D511)

Prior probabilities of groups:
  A   B
0.5 0.5

Group means:
     V1    V2     V3    V4
A 21.812 4.106 11.982 52.17
B  6.638 0.856  3.320 16.56

Coefficients of linear discriminants:
          LD1
V1 -0.7794490
V2 -0.6888651
V3  1.4115135
V4 -0.1192217
```

- 第三步：样本回判
  - $predict(results, D511)\$class$
  - 输出结果：

$$[1]\ A\ A\ A\ A\ A\ B\ B\ B\ B\ B$$

$$Levels: A\ B$$

全部判别正确

- 第四步：新样本判别
  - $D511P < -read.table("ex511A.txt", header = F)$
  - $predict(results, D511P)\$class$
  - 输出结果：

$$[1]B\ A\ A\ B\ B\ A\ A\ A$$

$$Levels : A\ B$$

三、多总体的距离判别

设有k个m元总体:$G_i \quad i=1,\cdots,k$,均值向量和协方差矩阵分别为$\mu^{(i)},\Sigma_i$,判断样品X来自哪个总体.

计算X到各个总体的马氏距离$d_i^2(X) \quad i=1,\cdots,k$

三、多总体的距离判别

设有k个m元总体:$G_i \quad i=1,\cdots,k$,均值向量和协方差矩阵分别为$\mu^{(i)},\Sigma_i$,判断样品X来自哪个总体.

计算X到各个总体的马氏距离$d_i^2(X) \quad i=1,\cdots,k$

判别准则:

$$X \in G_l, \quad d_l^2 = \min_{i=1,\cdots,k}\{d_i^2(X)\}$$

上面讨论的差别方法仅考虑样本之间的距离, **距离判别的缺点:**

- 该判别法与总体出现的机会大小完全无关.
- 没有考虑错判损失.
- 其它

如果进一步考虑上述问题，如何给出判别?

5.2 Bayes判别法及广义平方距离判别法

Bayes的统计思想:

假定对研究的对象已经有一定的认识,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(先验分布),得到后验概率分布.

- 将Bayes思想用于判别分析就得到Bayes判别方法.
- Bayes判别方法也是给出空间$\mathbb{R}^m$的一种划分, 由此构成的判别方法称为Bayes判别方法.

一、先验概率

设有k个总体,$G_1, \cdots, G_k$.假设事先对所研究的有一定的认识,即已知k个总体各自出现的概率为$q_1, \cdots, q_k$(验前概率).这组验前概率$q_1, \cdots, q_k$称为先验概率.

比如研究人群中得癌(G_1)和没有癌症(G_2)两类群体的问题,由长期经验知:$q_1 = 0.001$,$q_2 = 0.999$.

一、先验概率

设有k个总体,$G_1, \cdots, G_k$.假设事先对所研究的有一定的认识,即已知k个总体各自出现的概率为$q_1, \cdots, q_k$(验前概率).这组验前概率$q_1, \cdots, q_k$称为先验概率.

比如研究人群中得癌(G_1)和没有癌症(G_2)两类群体的问题,由长期经验知:$q_1 = 0.001$,$q_2 = 0.999$.

先验概率的给出方法:

1. 利用历史资料及经验进行估计
2. 利用训练样本中给类样本占总体的比例n_i/n作为q_i的值
3. 假定$q_1 = \cdots = q_k = 1/k$

二、广义平方距离

定义样品X到总体$G_t \quad (t=1,\cdots,k)$的广义平方距离为:

$$D_t^2(X) = D^2(X, G_t) = d_t^2(X) + g_1(t) + g_2(t),$$

其中

$$g_1(t) = \begin{cases} \ln|S_t|, & \text{若各组协方差阵}\Sigma_i\text{不全相等} \\ 0 & \text{若各组协方差阵}\Sigma_i\text{全相等} \end{cases}$$

$$g_2(t) = \begin{cases} -2\ln|q_t|, & \text{若先验概率不全相等} \\ 0 & \text{若先验概率全相等} \end{cases}$$

广义平方距离判别法为:

$$\text{判}X \in G_t, \qquad \text{若}D_t^2(X) < D_i^2(X) \quad (i=1,\cdots,k).$$

三、后验概率(条件概率)

把当样品X已知时,它属于G_t的概率记为$P(G_t|X)$或$P(t|X)$,假定总体G_t的概率密度为$f_t(x) \quad (t=1,\cdots,k)$给定,由条件概率的定义可以导出:

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

若假设$G_t(t=1,\cdots,k)$为正态总体,其密度函数为

$$f_t(x)=(2\pi)^{-m/2}|\Sigma_t|^{-1/2}\exp(-0.5d_t^2(x)),$$

其中$d_t^2(x)=(x-\mu^{(t)})'\Sigma_t^{-1}(x-\mu^{(i)})$.

则X属于第t组的概率为:

$$P(t|X=x)=\frac{\exp(-0.5D_t^2(x))}{\sum_{i=1}^{k}\exp(-0.5D_i^2(x))}$$

采用后验概率的判别标准为:

判$X\in G_t$, 当$P(t|X=x)>P(i|X=x)\quad(i\neq t, i=1,\cdots,k)$

在正态假设下按后验概率最大进行归类的原则,等价于按广义平方距离最小准则进行归类.

错判概率的估计方法:

1. 利用训练样本作为检验集,即用判别方法对已知类别的样本进行回判,统计判错的个数几占样品总数的比率.
2. 当训练样本较大,留出部分已知类别的样本作为检验集,统计判错的比率.
3. 舍一法(交叉检验法),每次留出一个已知类别的样品,而用其余n-1个样品建立判别准则,然后对留出的这个已知类别的样品进行判别归类.统计错判个数及比率.

四、贝叶斯判别准则

贝叶斯判别准则： 给出空间$\mathbb{R}^m$ 的一种划分: $D = (D_1, \cdots, D_k)$, 使得所带来的平均损失达到最小（即错判损失最小）.

1.错判概率和错判损失

我们把属于G_i的样品X,用判别法D 判别时却判给$G_j (j \neq i)$(即错判)的概率记为$P(j|i;D)$(或简记为$P(j|i)$).则有

$$P(j|i;D) = \int \cdots \int_{D_j} f_i(x_1, \cdots, x_m) dx_1 \cdots dx_m$$

$$= \int_{D_j} f_i(X) dX, \quad (j \neq i). \qquad \text{错判概率}$$

用这种判别法会发生错判,如X来自G_1,但却落入D_2, 被判为属G_2. 错判的概率为下图中阴影左半部分的面积,并记为$P(2|1)$. 类似有$P(1|2)$.

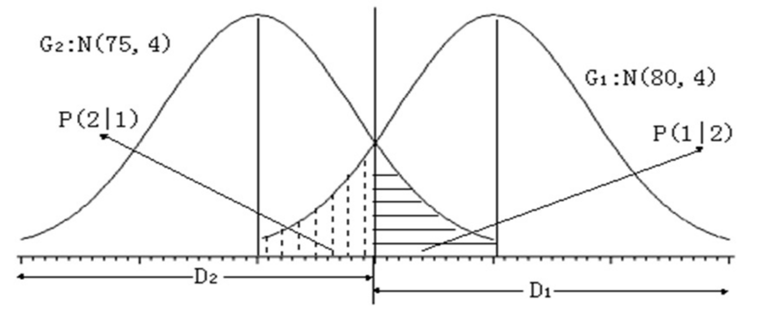


Figure: 错判概率示意图

我们把属于G_i的样品X,用判别法D 判别时却判给$G_j(j \neq i)$(即错判)的损失用$L(j|i;D)$表示,不会引起混淆情况下简计$L(j|i)$.
$L(j|i)$的赋值方法:

1. 由经验人为赋值
2. 假定错判损失相等

$$L(j|i;D) = \begin{cases} 1, & i \neq j, \\ 0, & i = j. \end{cases} = 1 - \delta_{ij}.$$

2.关于先验概率的平均损失

引入先验概率,判别法D关于先验概率的平均损失定义为:

$$g(D) = \sum_{t=1}^{k} q_t \sum_{j=1}^{k} P(j|t)L(j|t) \overset{def}{=} \sum_{t=1}^{k} q_t r_t(D)$$

其中$r_t(D) = q_t \sum_{j=1}^{k} P(j|t)L(j|t)$（错判损失），表示实属$G_t$的样品被错判其他总体的损失。

3.贝叶斯(Bayes)判别准则

定义5.2.1 设有k个总体: $G_1, \cdots, G_k$,相应的先验概率为$q_1, q_2, \cdots, q_k$. 如果有判别法D^*, 使得D^* 带来的平均损失达到最小,即

$$g(D^*) = \min_{\text{一切}D} g(D),$$

则称判别法D^* 符合贝叶斯判别准则,或称D^* 是贝叶斯判别的解.

4.符合贝叶斯判别准则(贝叶斯判别的解)

定理5.2.1 设有k个总体:$G_1,\cdots,G_k$,已知G_i的联合密度为$f_i(X)$,先验概率为$q_1,\cdots,q_k$,错判损失为$L(j|i)$,则贝叶斯判别的解$D^*=\{D_1^*,\cdots,D_k^*\}$ 为

$$D_t^*=\{X|h_t(X)<h_j(X),j\neq t,j=1,\cdots,k\}\quad(t=1,\cdots,k),$$

其中

$$h_j(X)=\sum_{i=1}^{k}q_iL(j|i)f_i(X)$$

它表示把样品X判归G_j的平均损失.

注:定理可以看出,贝叶斯判别准则不仅考虑了后验概率而且还考虑了错判损失。

证明：

$$\begin{aligned}
g(D^*) &= \sum_{i=1}^{k} q_i \sum_{t=1}^{k} L(t|i) P(t|i) \\
&= \sum_{i=1}^{k} q_i \sum_{t=1}^{k} L(t|i) \Pr(X \in D_t^* | X \in G_i) \\
&= \sum_{i=1}^{k} q_i \sum_{t=1}^{k} L(t|i) \int_{D_t^*} f_i(x) dx \\
&= \sum_{t=1}^{k} \int_{D_t^*} \sum_{i=1}^{k} q_i L(t|i) f_i(x) dx \\
&= \sum_{t=1}^{k} \int_{D_t^*} h_t(x) dx
\end{aligned}$$

类似有

$$g(D) = \sum_{t=1}^{k} \int_{D_t} h_t(x)dx$$

所以

$$\begin{aligned} g(D^*) - g(D) &= \sum_{t=1}^{k} \int_{D_t} h_t(x)dx - \sum_{j=1}^{k} \int_{D_j^*} h_t(x)dx \\ &= \sum_{t=1}^{k} \sum_{j=1}^{k} \int_{D_t^* \cap D_t} \left(h_t(x) - h_j(x)\right) dx \end{aligned}$$

由D^*的定义可知，$h_t(x) < h_j(x)$，$j = 1, 2, \cdot k, j \neq t.$

推论 当$L(j|i) = 1 - \delta_{ij}$ 时(即错判损失相等), 其中$\delta_{ij} = 1$, $i = j$; $\delta_{ij} = 0$, $i \neq j$. 贝叶斯的解$D^* = \{D_1^*, \cdots, D_k^*\}$ 为

$$D_t^* = \{X|q_t f_t(X) > q_j f_j(X), j \neq t, j = 1, \cdots, k\} \quad (t = 1, \cdots, k),$$

例5.2.1 试导出$k=2$ 时的贝叶斯判别的解.

解

$$h_1(X)=q_2f_2(X)L(1|2) \quad h_2(X)=q_1f_1(X)L(2|1)$$

从而

$$D_1=\{X|q_2f_2(X)L(1|2)<q_1f_1(X)L(2|1)\},$$

$$D_2=\{X|q_2f_2(X)L(1|2)\geq q_1f_1(X)L(2|1)\},$$

若令判别函数为

$$W(X)=\frac{f_1(X)}{f_2(X)} \quad d=\frac{q_2L(1|2)}{q_1L(2|1)}$$

则贝叶斯判别准则为:

$$\begin{cases} X\in G_1, & W(X)>d \\ X\in G_2, & W(X)\leq d \end{cases}$$

5.正态总体的贝叶斯判别法

设$G_i \sim N_m(\mu^{(i)}, \Sigma_i) \quad (i = 1, \cdots, k)$,并假定错判损失相等,先验概率为$q_1, \cdots, q_k$

(1)当$\Sigma_1 = \Sigma_2 = \cdots = \Sigma_m \overset{def}{=} \Sigma$ 时,总体G_i的概率密度为$f_i(X)$,则

$$q_i f_i(X) = \frac{q_i}{(2\pi)^{m/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(X - \mu^{(i)})'\Sigma^{-1}(X - \mu^{(i)})\right\}$$

$$\begin{aligned}\ln q_i f_i(X) &= -\frac{1}{2}[\ln|\Sigma| + m\ln(2\pi) + X'\Sigma^{-1}X] \\ &\quad +\ln q_i - \frac{1}{2}(\mu^{(i)})'\Sigma^{-1}\mu^{(i)} + X'\Sigma^{-1}\mu^{(i)} \\ &= C_0 + C_{i0} + X'C_i \overset{def}{=} C_0 + Y_i(X)\end{aligned}$$

其中:

- $C_0 = -\frac{1}{2}[\ln|\Sigma| + m\ln(2\pi) + X'\Sigma^{-1}X]$是与$G_i$无关的依赖与X的常数.
- $C_{i0} = \ln q_i - \frac{1}{2}(\mu^{(i)})'\Sigma^{-1}\mu^{(i)}$;
- $C_i = \Sigma^{-1}\mu^{(i)} \overset{def}{=} (C_{i1}, \cdots, C_{im})'$.

当$\Sigma_1 = \cdots = \Sigma_k = \Sigma$ 未知时,

- 由样本可计算第i个总体的样本均值向量为$\bar{X}^{(i)}$为$\mu^{(i)}$的估计.
- 各总体的协方差阵假设为相等，样本协方差阵的估计

$$S = \frac{1}{n-k}(A_1 + \cdots + A_k).$$

- 贝叶斯的解$D^* = \{D_1^*, \cdots, D_k^*\}$为

$$D_t^* = \{X | Y_t(X) > Y_j(X), j \neq t, j = 1, \cdots, k\} \quad (t = 1, \cdots, k),$$

其中$Y_t(X) = C_{j0} + C_j' X$,称为线性判别函数,$C_j = S^{-1}\bar{X}^{(j)}$ 称为判别系数,$C_{j0} = \ln q_i - \frac{1}{2}(\bar{X}^{(j)})' S^{-1} \bar{X}^{(j)}$ 为常数项.

(2)当$\Sigma_i(i=1,\cdots,k)$不全相等

$$\ln q_i f_i(X) = d_0 - \frac{1}{2}[-2\ln q_i + \ln|\Sigma_i| + (X-\mu^{(i)})'\Sigma_i^{-1}(X-\mu^{(i)})]$$

$$\overset{def}{=} d_0 + Z_i(X)$$

其中$d_0 = -\frac{m}{2}\ln(2\pi)$是常数.

当$\mu^{(i)}, \Sigma_i$未知时,样本可计算第i个总体的样本均值向量为$\bar{X}^{(i)}$,样本协方差阵为S_i

贝叶斯判别的解$D^* = \{D_1^*, \cdots, D_k^*\}$为

$$D_t^* = \{X | Z_t(X) > Z_j(X), j \neq t, j = 1, \cdots, k\} \quad (t = 1, \cdots, k),$$

其中

$$Z_j(X) = \ln q_j f_j(X) - d_0 = -\frac{1}{2} D_j^2(X)$$

称$Z_j(X)$为二次判别函数.

可见在正态条件下,贝叶斯判别方法与广义平方距离判别法是一致的.

例5.2.2 (胃癌的鉴别)

表5.2是从病例中随机抽取的部分资料。这里有三个总体：胃癌、萎缩性胃炎和非胃炎患者. 从每个总体抽5个病人，每人化验4项生化指标：血清铜蛋白(X_1)、蓝色反应(X_2)、尿吲哚乙酸(X_3)和中性硫化物(X_4).试用广义平方距离判别法建立判别准则，并对这15个样品进行判别归类.

表 5.2 胃癌检验的生化指标值

类别		序号	血清铜蛋白 X_1	蓝色反应 X_2	吲哚乙酸 X_3	中性硫化物 X_4
胃癌患者	胃癌患者	1	228	134	20	11
		2	245	134	10	40
		3	200	167	12	27
		4	170	150	7	8
		5	100	167	20	14
非胃癌患者	萎缩性胃癌患者	6	225	125	7	14
		7	130	100	6	12
		8	150	117	7	6
		9	120	133	10	26
		10	160	100	5	10
	非胃炎患者	11	185	115	5	19
		12	170	125	6	4
		13	165	142	5	3
		14	135	108	2	12
		15	100	117	7	2

注：X_3，X_4 是原始数据的 100 倍.

第一步：检验三组的均值是否有差异，

$$H_0: \mu_1 = \mu_2 = \mu_3, \ H_1: \text{不全相等}$$

并进一步检验各两组之间的均值是否有差异。

$$H_0: \mu_1 = \mu_2 \ \ H_1: \mu_1 \neq \mu_2$$
$$H_0: \mu_2 = \mu_3 \ \ H_1: \mu_2 \neq \mu_3$$
$$H_0: \mu_1 = \mu_3 \ \ H_1: \mu_1 \neq \mu_3$$

第二步：检验协方差矩阵是否相等

$$H_0 : \Sigma_1 = \Sigma_2 = \Sigma_3,\ \ H_1 : \text{不全相等}$$

从下面的输出结果可以看出，不能拒绝协方差矩阵相等的假设。采用varcomp协方差矩阵检验是否相等的函数.

Equality of Covariances Matrices Test

data: covmat
corrected lambda* = 23.524, df = 20, p-value = 0.4396

第三步：采用广义平方距离给出判别

- 加载MASS包，调用qda函数
- $D522 < -read.table("ex522.txt", header = F)$
- $results < -qda(V1\ V2 + V3 + V4 + V5, D522)$

```
$class
 [1] 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3
Levels: 1 2 3
Call:
qda(V1 ~ V2 + V3 + V4 + V5, data = D522)

Prior probabilities of groups:
        1         2         3
0.3333333 0.3333333 0.3333333

Group means:
     V2    V3   V4   V5
1 188.6 150.4 13.8 20.0
2 157.0 115.0  7.0 13.6
3 151.0 121.4  5.0  8.0
```

根据平方距离给出的判别函数对样本进行回判。

- $predict(results, D522)$

```
$class
 [1] 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3
Levels: 1 2 3

$posterior
             1            2            3
1  1.000000e+00  0.000000e+00 4.424058e-19
2  1.000000e+00  1.378099e-22 1.013387e-18
3  9.524673e-01  4.753271e-02 2.378453e-31
4  9.744963e-01 6.770577e-102 2.550371e-02
5  1.000000e+00  0.000000e+00 5.112352e-53
6  1.570255e-12  8.937141e-01 1.062859e-01
7  4.720564e-03  9.824368e-01 1.284262e-02
8  3.717880e-06  9.999963e-01 6.334871e-24
9  1.385891e-06  7.798879e-01 2.201108e-01
10 3.983102e-10  9.968484e-01 3.151628e-03
11 6.582191e-04  2.403927e-20 9.993418e-01
12 3.687193e-10 2.344737e-196 1.000000e+00
13 3.746839e-05  3.510061e-29 9.999625e-01
14 9.215652e-03 1.217308e-176 9.907843e-01
15 6.656535e-09  1.630444e-06 9.999984e-01
```

R中有不同功能用于判别分析的包，例如，加载WMDB 包，可作加权的距离判别分析

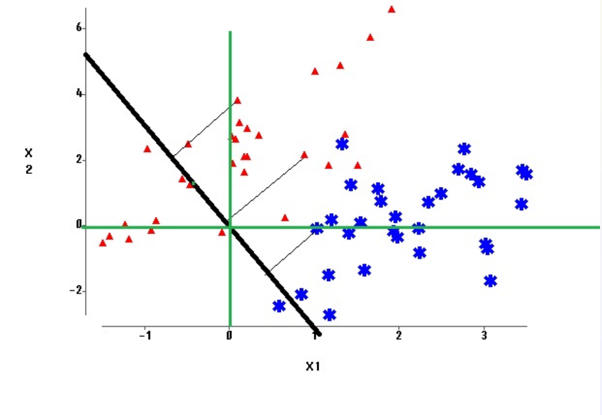
- $X < -D522[,2:5]$
- $Y < -as.factor(D522[,1])$
- $wmd(X, Y, diag(rep(0.25, 4)))$

```
      1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
blong 1 1 1 1 1 2 2 2 3  2  3  3  3  3  3
[1] "num of wrong judgement"
[1] 9
[1] "samples divided to"
[1] 3
[1] "samples actually belongs to"
[1] 2
Levels: 1 2 3
[1] "percent of right judgement"
[1] 0.9333333
```

5.2 Fisher判别法

一、Fisher判别的基本思想

通过投影后，使组与组之间尽可能分开. 如下图当$m = 2$, $k = 2$时,寻找方向a, 使两组数据投影后在一维直线上尽可能区分开



利用线性投影和方差分析的思想.

设从m元总体$G_t(t=1,\cdots,k)$分别抽取样本如下:

$$X_{(i)}^{(t)}=(x_{i1}^{(t)},\cdots,x_{im}^{(t)})\quad (t=1,\cdots,k;i=1,\cdots,n_t)$$

投影后为:

$$G_1:\quad a'X_{(1)}^{(1)},\cdots,a'X_{(n_1)}^{(1)},\qquad \bar{X}^{(1)}=\frac{1}{n_1}\sum_{j=1}^{n_1}X_{(j)}^{(1)}$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$G_k:\quad a'X_{(1)}^{(k)},\cdots,a'X_{(n_k)}^{(k)},\qquad \bar{X}^{(k)}=\frac{1}{n_k}\sum_{j=1}^{n_k}X_{(j)}^{(k)}$$

注：投影后可以看成是一维的分组样本，考虑样本总的平方和、组内平方和以及组间平方和。

投影后为一元数据,其组间平方和为

$$\begin{aligned} B_0 &= \sum_{t=1}^{k} n_t(a'\bar{X}^{(t)} - a'\bar{X})^2 \\ &= a'[\sum_{t=1}^{k} n_t(\bar{X}^{(t)} - \bar{X})(\bar{X}^{(t)} - \bar{X})']a \\ &= a'Ba \end{aligned}$$

其中B为组间离差阵

$$B = \sum_{t=1}^{k} n_t(\bar{X}^{(t)} - \bar{X})(\bar{X}^{(t)} - \bar{X})'$$

合并的组内平方和:

$$
\begin{aligned}
A_0 &= \sum_{t=1}^{k}\sum_{j=1}^{n_t}(a'\bar{X}_{(j)}^{(t)} - a'\bar{X}^{(t)})^2 \\
&= a'Aa
\end{aligned}
$$

其中:

$$
A = \sum_{t=1}^{k}\sum_{j=1}^{n_t}(\bar{X}_{(j)}^{(t)} - \bar{X}^{(t)})(\bar{X}_{(j)}^{(t)} - \bar{X}^{(t)})'
$$

要使得投影后，k个总体（类）的均值差异尽可能的大. 即使得比值

$$
\frac{a'Ba}{a'Aa} \stackrel{def}{=} \Delta(a)
$$

尽可能的大.

归纳为如下的规划问题：求$a \in R^m$使$a'Ba$ 在$a'Aa = 1$条件下达极大. 即

$$\begin{cases} \max_a \frac{a'Ba}{a'Aa}, \\ s.t. a'Aa = 1. \end{cases}$$

二、线性判别函数的求法

利用Lagranre方法，令$\varphi(a) = a'Ba - \lambda(a'Aa - 1)$

解方程组

$$\begin{cases} \frac{\partial \varphi}{\partial a} = 2(B - \lambda A)a = 0 \\ \frac{\partial \varphi}{\partial \lambda} = 1 - a'Aa = 0 \end{cases}$$

结论： 设$A^{-1}B$的非零特征值为$\lambda_1 \geq \cdots \geq \lambda_r > 0$，相应的满足约束条件的特征向量为$l_1, \cdots, l_r$,取$a = l_1$ 可使$\Delta(a)$达最大,且最大值为λ_1。

一般$\Delta(a)$称为判别效率. 由以上结论知，此时判别效率为λ_1。

附录：定理7.2

设B 是p 阶对称矩阵, λ_i 是B的第i大的特征值, l_i 是相应于λ_i的B 的标准化特征向量$(i = 1, \ldots, p)$ 为任一非零p 维向量, 那么有

$$\lambda_p \leq \frac{x'Bx}{x'x} \leq \lambda_1,$$

上式右边等号当$x = cl_1$时成立, 左边等号当$x = cl_p$时成立.

令$x = A^{\frac{1}{2}}a$, 则

$$\Delta(a) = \frac{x'Cx}{x'x}$$

其中$C = A^{-\frac{1}{2}}BA^{-\frac{1}{2}}$, 与$A^{-1}B$有相同的特征根.

定义5.3.1 设$A^{-1}B$的非零特征值为$0 < \lambda_1 \le \cdots \le \lambda_r$，相应的满足约束条件的特征向量为$l_1, \cdots, l_r$,称

$$P_1 = \lambda_1 / \sum_{i=1}^{r} \lambda_i$$

为线性判别函数$u_1(X) = l_1' X$的判别能力;称

$$P_{(l)} = (\lambda_1 + \cdots + \lambda_l) / \sum_{i=1}^{r} \lambda_i$$

为前l个$(l \le r)$线性判别函数$u_1(X) = l_1' X, \cdots, u_l(X) = l_l' X$ 的累计判别能力.

三、Fisher判别准则

判别准则I(基本Fisher判别)

- 取最大特征根对应的特征向量l_1, 构作Fisher判别函数$l_1'X$, 将原来m维的总体简单化为一维总体的形式;
- 采用距离判别（或其它一维总体的判别方法）进行判别.

例5.3.1 若$k=2$,试求Fisher线性判别函数及其相应的判别效率.

解 两总体的组间离差阵B为

$$B = n_1(\bar{X}^{(1)} - \bar{X})(\bar{X}^{(1)} - \bar{X})' + n_2(\bar{X}^{(2)} - \bar{X})(\bar{X}^{(2)} - \bar{X})'$$

利用$\bar{X} = \frac{1}{n_1+n_2}(n_1\bar{X}^{(1)} + n_2\bar{X}^{(2)})$得

$$B = \frac{n_1 n_2}{n_1 + n_2}(\bar{X}^{(1)} - \bar{X}^{(2)})(\bar{X}^{(1)} - \bar{X}^{(2)})'.$$

$$B_0 = a'Ba$$

合并组内离方差阵$A = A_1 + A_2$,其中

$$A_t = \sum_{i=1}^{n_t}(x_{(i)}^{(t)} - \bar{X}^{(t)})(x_{(i)}^{(t)} - \bar{X}^{(t)})' \quad i = 1, 2$$

$$A_0 = a'Aa.$$

and

$$\Delta(a) = \frac{B_0}{A_0}.$$

由于B的秩等于1,故特征方程$|A^{-1}B - \lambda I| = 0$的非零特征根只有一个.

因为

$$A^{-1}B = \frac{n_1 n_2}{n_1+n_2} A^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)})(\bar{X}^{(1)} - \bar{X}^{(2)})'$$

有线性代数知:AB和BA的非零特征根相同.所以$A^{-1}B$的非零特征根等同于:

$$\frac{n_1 n_2}{n_1+n_2}(\bar{X}^{(1)} - \bar{X}^{(2)})' A^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)}) = \frac{n_1 n_2}{n_1+n_2} d^2,$$

其中

$$d^2 = (\bar{X}^{(1)} - \bar{X}^{(2)})' A^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)})$$

因为$\frac{n_1 n_2}{n_1+n_2} d^2$是个数值,所以它就是所求的非零特征根$\lambda$.

记l为对应于λ的在条件$l'Al=1$的特征向量,满足$Bl=\lambda Al$,所以取$l=\frac{1}{d}A^{-1}(\bar{X}^{(1)}-\bar{X}^{(2)})$可满足上述条件.于是得Fisher线性判别函数为

$$u(X)=l'X=\frac{1}{d}X'A^{-1}(\bar{X}^{(1)}-\bar{X}^{(2)})$$

相应的判别效率

$$\Delta(l)=\frac{n_1n_2}{n_1+n_2}(\bar{X}^{(1)}-\bar{X}^{(2)})'A^{-1}(\bar{X}^{(1)}-\bar{X}^{(2)})$$

序贯分析（sequential analysis）

- 序贯分析（sequential analysis）数理统计学的一个分支。其名称源出于美国统计学家瓦尔德在1947年发表的一本同名著作。
- 序贯最开始是指抽样时采用的一种方法。序贯抽样方案是指在抽样时，不事先规定总的抽样个数（观测或实验次数），而是先抽少量样本, 根据其结果，再决定停止抽样或继续抽样、抽多少，这样下去，直至决定停止抽样为止。
- 将序贯思想应用于判别分析，即得出Fisher判别的第二类判别方法。

判别准则II

设

$$u_{(ij)}^{(t)} = l_j' X_{(i)}^{(t)}, t = 1, \cdots, k, j = 1, \cdots, r, i = 1, \cdots, n_t$$

则k个m元样本在$l_1, l_2, \cdots, l_r$上投影后的均值分别为:

$$\bar{u}_j^{(t)} = \frac{1}{n} \sum_{i=1}^{n_t} u_{(ij)}^{(t)} = l_j' \bar{X}^{(t)}, j = 1, \cdots r, t = 1, \cdots, k.$$

投影后的方差分别为:

$$\hat{\sigma}_j^{(t)} = \frac{1}{n-1} \sum_{i=1}^{n_t} (u_{(ij)}^{(t)} - \bar{u}_j^{(t)})^2 = l_j' S_t l_j, j = 1, \cdots r, t = 1, \cdots, k.$$

序贯判别思想:

- 先取判别效率最大的线性判别函数$u_1(X) = l_1'X$,对样品X,计算它在l_1上投影:若存在唯一的i_1,使

$$\frac{|u_1(X) - \bar{u}_1^{(i_1)}|}{\hat{\sigma}_1^{(i_1)}} = \min_{t=1,\cdots,k} \frac{|u_1(X) - \bar{u}_1^{(t)}|}{\hat{\sigma}_1^{(t)}}$$

  时,判$X \in G_{i_1}$.

- 如果存在j个总体$G_{k_1}, \cdots, G_{k_j}$,使其与$u_1(X)$距离相等且为最小,记序号集$L = \{k_1, \cdots, k_j\}$, 则再取判别效率为λ_2(次大)的判别函数$u_2(X) = l_2'X$.

当存在唯一的i_2,使

$$\frac{|u_2(X)-\bar{u}_2^{(i_2)}|}{\hat{\sigma}_2^{(i_2)}}=\min_{t\in L}\frac{|u_2(X)-\bar{u}_2^{(t)}|}{\hat{\sigma}_2^{(t)}}$$

时,判$X\in G_{i_2}$.

- 如果第二个判别函数仍不能判别样品X 所属总体,则还可以取第三个线性判别函数,依此类推.

有了Fisher判别函数(即将多空间上的点投影到一维)后，可采用一维的判别方法给出判别准则. 下面以$k=2$为例推导出按距离准则判别样品归类的判别法. 设两总体均值为$\bar{X}^{(1)}$和$\bar{X}^{(2)}$,线性判别函数的数值为

$$\bar{\mu}^{(1)} = l'\bar{X}^{(1)}, \bar{\mu}^{(2)} = l'\bar{X}^{(2)}$$

若投影后两总体方差相等,则阈值点

$$\bar{\mu} = \frac{1}{2}(l'\bar{X}^{(1)} + l'\bar{X}^{(2)})$$

若投影后两总体方差不等,则阈值点

$$\mu^* = \frac{\hat{\sigma}_2\bar{\mu}^{(1)} + \hat{\sigma}_1\bar{\mu}^{(2)}}{\hat{\sigma}_1 + \hat{\sigma}_2}$$

其中$\hat{\sigma}_t$是投影后总体G_t的样本方差,且$\hat{\sigma}_t = l'S_tl$

判别准则　(不妨设$l'\bar{X}^{(1)} > l'\bar{X}^{(2)}$)

- 投影后方差相等

$$\begin{cases} X \in G_1 & u(X) > \bar{\mu} \\ X \in G_2 & u(X) < \bar{\mu} \\ \text{待判} & u(X) = \bar{\mu} \end{cases}$$

- 投影后方差不等

$$\begin{cases} X \in G_1 & u(X) > \mu^* \\ X \in G_2 & u(X) < \mu^* \\ \text{待判} & u(X) = \mu^* \end{cases}$$

判别准则III(主成分分析思想)

如果有r个非零特征根,相应的有r个线性判别函数$u_1(X), \cdots, u_r(X)$, 这时, 相当于把原来m 个变量综合成r 个新变量.

在实用中常取$l \leq r$,且满足

$$\sum_{i=1}^{l} \lambda_i / \sum_{i=1}^{r} \lambda_i \geq P_0 \quad (p_0\text{一般取}0.7)$$

这样m元总体的判别问题即化为l元总体的判别问题(降维), 而且l个新变量互不相关,可按距离判别准则来归类.

例5.3.2 对表5.2中胃癌检验的生化指标值用主成分判别法进行判别归类.

- 第一步：均值检验和协方差检验
- 第二步：加载adegenet(MASS, ade4)
- 第三步：读入数据
  - $D522 < -read.table("ex522.txt", header = F)$
  - $X < -D522[, 2 : 5]$
  - $Y < -D522[, 1]$

- 第四步：进行主成分判别
  - $re1 < -dapc(X, Y)$,# 会出现累积贡献率图，
  - Choose the number PCs to retain (>= 1): 1 #根据图选择主成分的个数
  - Choose the number discriminant functions to retain (>= 1): 1 # 选择用于主成分判别的主成分个数

- 第五步：回判
  - $predict(re1, X)$

```
$assign
 [1] 1 1 1 2 3 3 1 3 3 2 2 3 1 2 3
Levels: 1 2 3

$posterior
          1         2         3
 [1,] 0.5891914 0.2285853 0.1822233
 [2,] 0.6750288 0.1856064 0.1393648
 [3,] 0.4794356 0.2804507 0.2401138
 [4,] 0.3300522 0.3441481 0.3257997
 [5,] 0.1210917 0.4068540 0.4720543
 [6,] 0.1785776 0.3951101 0.4263123
 [7,] 0.5728086 0.2365553 0.1906362
 [8,] 0.1631301 0.3989081 0.4379618
 [9,] 0.2429939 0.3758935 0.3811126
[10,] 0.2755332 0.3646495 0.3598173
[11,] 0.3200217 0.3480557 0.3319226
[12,] 0.1944713 0.3908279 0.4147007
[13,] 0.3874317 0.3208019 0.2917665
[14,] 0.3048609 0.3538515 0.3412876
[15,] 0.1104645 0.4081241 0.4814114
```

- 采用一个主成分的判别效果并不好。
- 如果选用二个主成分去判别，出输出结果:

```
$assign
 [1] 1 1 1 1 1 2 3 2 3 2 3 2 2 3 3
Levels: 1 2 3
```

Figure: 主成分判别回判结果

- 采用二个主成分后，第一类判别正确，但第二第三类仍不理想。

R中还有一些用于判别分析的包：

例如：Mixture Discriminant Analysis, 包的名称：mda, 函数

- mda(formula, data, subclasses, sub.df, tot.df, dimension, eps, iter, weights, method, keep.fitted, trace, ...)
- fda(formula, data, weights, theta, dimension, eps, method, keep.fitted, ...)

R中还有一些用于判别分析的包:

- rda: rda provides classification for high dimensional data by means of shrunken centroid regularized discriminant analysis;
- class: it provides k-nearest neighbours by knn().
- knncat: it provides k-nearest neighbours for categorical variable.
- SensoMineR: it provides FDA() for factorial discriminant analysis.
- A number of packages provide for dimension reduction with the classification, such as klaR, superpc, gpls, hddplot, ROCR, predbayescor, etc.

判别效果的检验及各变量判别能力的检验

判别能力依赖:

- (1)样本是否来自不同的总体; 即: 进行总体均值是否相等的假设检验.
- (2)m 个判别指标区组能力; 即: 进行各变量判别能力的假设检验.

假设总体G_i 的分布为$N_m(\mu^{(t)}, \Sigma_t)$, $t = 1, 2, \cdots, k$, $X_{(i)}^{(t)}$, $t = 1, 2, \cdots, k$; $i = 1, 2, \cdots, n_t$, 为来自G_t 的m 元样本.

一、两总体判别效果的检验

$$H_0 : \mu^{(1)} = \mu^{(2)}$$

根据第三章的结论，计算两总体之间的马氏距离

$$d^2(1,2) = (\overline{X}^{(1)} - \overline{X}^{(2)})' S^{-1} (\overline{X}^{(1)} - \overline{X}^{(2)})$$

构造F统计量

$$F = \frac{(n_1+n_2-m-1)n_1n_2}{m(n_1+n_2)(n_1+n_2-2)} d^2(1,2) \sim F(m, n_1+n_2-m-1)$$

两个总体均值有显著性时，对两总体讨论判别问题是有意义的.
但当两个总体均值没有显著性差异时，如果盲目地应用判别分析的方法进行判别分析，则错判的机会很大，判别分析没有意义.

二、k个总体判别效果的检验

第一步：检验各总体均值是否有显著差异

$$H_0 : \mu^{(1)} = \mu^{(2)} = \cdots = \mu^{(k)}$$

根据第三章的结论，利用似然比原则导出Λ 统计量

$$\Lambda = \frac{|A|}{|A+B|} = \frac{|A|}{|T|} \sim \Lambda(m, n-k, k-1)$$

第二步：如果上述检验显著，则进行两两配对检验，说明各总体之间的差异

$$H_0^{(ij)} : \mu^{(i)} = \mu^{(j)}$$

如果假设各总体的协方差阵相等，则采用马氏距离构造的F统计量

$$F_{ij} = \frac{(n-k-m+1)}{m(n-k)} \frac{n_i n_j}{n_i + n_j} d^2(i,j) \sim F(m, n-k-m+1)$$

如果检验发现$H_0^{(ij)}$ 为真，则将第i 类和第j 类合并.

第三步：分类检验结束后，检验各变量判别能力

- 变量判别能力的度量
- 变量判别能力的检验

变量判别能力的度量

- 消去法求行列式的值

$$|A| = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{vmatrix} = a_{11} \begin{vmatrix} 1 & \frac{a_{12}}{a_{11}} & \cdots & \frac{a_{1m}}{a_{11}} \\ 0 & a_{22}^{(1)} & \cdots & a_{2m}^{(1)} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & a_{m2}^{(1)} & \cdots & a_{mm}^{(1)} \end{vmatrix}$$

$$= a_{11} \begin{vmatrix} a_{22}^{(1)} & \cdots & a_{2m}^{(1)} \\ \vdots & \vdots & \vdots \\ a_{m2}^{(1)} & \cdots & a_{mm}^{(1)} \end{vmatrix} = a_{11}a_{22}^{(1)}\cdots a_{mm}^{(m-1)}.$$

- 上述的消去法并不一定要从a_{11}开始，可以从$a_{i_1i_1}$ 开始

$$|A| \quad = \quad = a_{i_1i_1}a_{i_2i_2}^{(1)}\cdots a_{i_mi_m}^{(m-1)}.$$

- 对$|T|$也可以进行类似的计算

$$|T| \quad = \quad = t_{i_1i_1}t_{i_2i_2}^{(1)}\cdots t_{i_mi_m}^{(m-1)}.$$

仍从统计量

$$\Lambda = \frac{|A|}{|A+B|} = \frac{|A|}{|T|}$$

开始

- 构造单个变量的判别能力统计量

(1) m个指标的判别能力统计量

$$\begin{aligned}\Lambda_{(m)} &= \frac{|A|}{|T|} = \frac{a_{11}a_{22}^{(1)}\cdots a_{mm}^{(m-1)}}{t_{11}t_{22}^{(1)}\cdots t_{mm}^{(m-1)}} \\ &= U_{(1,2,\cdots,m)}\end{aligned}$$

度量m个指标$X_1,\cdots,X_m$ 对k 个总体的判别能力，$\Lambda_{(m)}$越小，说明判别效果越好.

(2) 现考虑前$m-1$个变量的判别能力统计量

$$\begin{aligned}\Lambda_{(m-1)} &= \frac{|A_{m-1}|}{|T_{m-1}|} = \frac{a_{11}a_{22}^{(1)}\cdots a_{(m-1)(m-1)}^{(m-2)}}{t_{11}t_{22}^{(1)}\cdots t_{(m-1)(m-1)}^{(m-2)}} \\ &= U_{(1,2,\cdots,m-1)}\end{aligned}$$

(3)定义第m个指标的判别能力统计量记：

$$U_{m|1,2,\cdots,m-1}=\frac{U_{(1,2,\cdots,m)}}{U_{(1,2,\cdots,m-1)}}=\frac{a_{mm}^{(m-1)}}{t_{mm}^{(m-1)}}$$

显然，$U_{m|1,2,\cdots,m-1}$ 表示第m个指标X_m 对k 个总体的判别能力，$U_{m|1,2,\cdots,m-1}$ 越小，说明判别效果越好.

(4)定义第j个指标的判别能力统计量对于第j个变量的判别能力可以用类似的统计量来

$$U_{j|1,2,\cdots,j-1,j+1,\cdots,m-1}=\frac{U_{(1,2,\cdots,m)}}{U_{(1,2,\cdots,j-1,j+1,\cdots,m-1)}}$$

- 构造单个变量的判别能力检验

(1) 若已经r个变量$X_{i_1},\cdots,X_{i_r}, r<m$ 的k 个总体的判别效果显著. 其判别能力的统计量为

$$U_{(i_1,i_2,\cdots,i_r)} = \frac{a_{i_1i_1}a^{(1)}_{i_2i_2}\cdots a^{(r-1)}_{(i_r)(i_r)}}{t_{i_1i_1}t^{(1)}_{i_2i_2}\cdots t^{(r-1)}_{i_ri_r}}$$

(2) 考虑增加新的变量$X_{i_{r+1}}$，相应的对k 个总体的判别能力会提高（至少保持原来的判别能力）. 统计量为

$$U_{(i_1,i_2,\cdots,i_r,i_{r+1})} = U_{(i_1,i_2,\cdots,i_r)}\cdot U_{i_{r+1}|(i_1,i_2,\cdots,i_r)}.$$

其中：

$$U_{i_{r+1}|(i_1,i_2,\cdots,i_r)} = \frac{a^r_{i_{r+1}i_{r+1}}}{t^{(r)}_{i_{r+1}i_{r+1}}}.$$

注意：统计量$U_{i_{r+1}|(i_1,i_2,\cdots,i_r)}$检验的是除去了$X_{i_1},\cdots,X_{i_r}$个变量后，新增加的变量$X_{i_{r+1}}$的判别能力. 即检验假设

$$H_0 : \mu^{(1)}_{i_{r+1}|(i_1,\cdots,i_r)} = \mu^{(2)}_{i_{r+1}|(i_1,\cdots,i_r)} = \cdots = \mu^{(k)}_{i_{r+1}|(i_1,\cdots,i_r)}$$

- 判别能力统计量的分布

在假设总体为正态的前提下，可以证明

(1) $U_{(i_1,i_2,\cdots,i_r)} \sim \Lambda(r, n-k, k-1)$

(2) $U_{(i_1,i_2,\cdots,i_r,i_{r+1})} \sim \Lambda(r+1, n-k, k-1)$

(3) $U_{i_{r+1}|(i_1,i_2,\cdots,i_r)} \sim \Lambda(1, n-k-r, k-1)$.

利用Λ 统计量与F统计量的关系，对H_0 给出检验.

1. 向前法
2. 向后法
3. 逐步法

逐步法的主要思想：

逐个引入变量，每次把一个判别能力最强的变量引入判别式，每引入一个新变量，对判别式中的老变量逐个进行检验，如其判别能力因新变量的引入而变得不显著，应把它从判别式中剔除。

基本步骤

- 引入所有变量中判别能力最强的一个变量.
  (一元方差分析，分析每个变量的差别能力，用F大的为最有差别能力)
- 检验变量对k类判别是否显著; 如果显著则引入，如果不显著则无法建立判别式，即所有的变量对总体无法判别, 变量选择过程停止, 需要引入新的变量.
- 第一个变量引入后，再从其余的变量中选择一个判别能力再强的变量，检验第二个变量的判别能力，如果显著则引入，如果不显著，判别变量选择过程结束.

- 第二个变量引入后，对于引入的变量X_{i_1}, X_{i_2} 重新进行检验，从中先取一个判别能力最弱的变量进行检验，如果显著，则进入第三个变量的选择，如果不显著，则把入选的变量重新删除，然后进入第三个变量的选择。
- $\cdots\cdots$
- 如果现在有r个变量已经入选，在引入$r+1$个变量时，要先对r 个变量中判别能力最弱的一个作检验，显著则进一下个变量的选择，如果不显著则要删除变量，并对剩余的$r-1$个变量中最弱的一个作检验，
- 如果没有变量可以入选，也没有变量可以删除，则变量选择过程最后结束.

Stepwise Diagonal Discriminant Analysis: SDDA

Usage: sdda(X, y, priors, start = rep(FALSE, ncol(X)), never = rep(FALSE, ncol(X)),method="lda",...)

- **X**: Training data matrix - rows are observations, columns are variables.
- **y**: A factor of true class labels, or a numeric vector with values 1, 2, 3, ... G, where G is the number of classes.
- **priors**: Prior probabilities for the different classes, if left unspecified these default to equal probability to belong to each group
- ....

- $D511 < -read.table(''ex511.txt'', header = F)$
- $attach(D511)$
- $A < as.factor(V5)$
- $MD511 < -as.matrix(D511[, 1 : 4])$
- $s1 < -sdda(MD511, A)$
- $s2 < -predict(s1, MD511)$
- $table(s2, A)$