多元统计分析 (4)

钟瑜　222018314210044

2020 年 12 月 3 日

5-2　设三个总体 G_1, G_2 和 G_3 的分布分别为：$N(2, 0.5^2)$，$N(0, 2^2)$和 $N(3, 1^2)$. 试问样品 $x=2.5$ 应判归哪一类？

(1) 按距离判别准则；

(2) 按贝叶斯判别准则（取 $q_1=q_2=q_3=\frac{1}{3}, L(j|i)=\begin{cases}1, i\neq j\\ 0, i=j\end{cases}$）.

解. *(1)*

$$\begin{aligned} d(x, G_1) &= \frac{(x-\mu_1)^2}{\sigma_1^2} = 1 \\ d(x, G_2) &= \frac{(x-\mu_2)^2}{\sigma_2^2} = 1.5625 \\ d(x, G_3) &= \frac{(x-\mu_3)^2}{\sigma_3^2} = 0.25 \end{aligned} \tag{1}$$

由于 $d(x, G_3)$ 最小，故应判归第三类.

(2) 由于

```
> dnorm(2.5,2,0.5)
[1] 0.4839414
> dnorm(2.5,0,2)
[1] 0.09132454
> dnorm(2.5,3,1)
[1] 0.3520653
```

而 $q_1 = q_2 = q_3 = 1$,

$$q_2 f_2 = 0.09132454 < q_3 f_3 = 0.3520653 < q_1 f_1 = 0.4839414$$

由推论

推论　当 $L(j|i)=1-\delta_{ij}$ 时(即错判损失都相等)，则贝叶斯判别的解 $D^*=\{D_1^*, \cdots, D_k^*\}$ 为

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

故判归第一类.

5-8　用逐步判别法选择判别变量的过程中(已知训练样本总容量 $n=30, k=3$,考察的变量个数 $m=4$). 已知在第一步引入变量 X_3 后合并组内离差阵 A 和总离差阵 T 分别化为

$$A^{(1)}=T_3(A)=\begin{bmatrix} 28571.5 & 683.4 & -1.123 & 9464.3 \\ 683.4 & 114.9 & -0.519 & 1230.0 \\ 1.123 & 0.519 & 0.0027 & 3.845 \\ 9464.3 & 1230.0 & -3.845 & 15375.8 \end{bmatrix},$$

$$T^{(1)}=T_3(T)=\begin{bmatrix} 28884.9 & 671.2 & -1.172 & 9233.8 \\ 671.2 & 148.3 & -0.347 & 1877.6 \\ 1.172 & 0.347 & 0.0018 & 0.508 \\ 9233.8 & 1877.6 & -0.508 & 27925.9 \end{bmatrix}.$$

试问下一步可否引入变量？引哪一个？

解.

```
> #n=30,k=3
> A<-matrix(c(28571.5,683.4,-1.123,9464.3,
+             683.4,1149,-0.519,1230.0,
+             1.123,0.519,0.0027,3.845,
+             9464.3,1230.0,-3.845,15375.8),byrow = TRUE,ncol=4)
> T<-matrix(c(28884.9,671.2,-1-172,9233.8,
+             671.2,148.3,-0.347,1877.6,
+             1.172,0.347,0.0018,0.508,
+             9233.8,1877.6,-0.508,27925.9),byrow=TRUE,ncol=4)
> lambda<-det(A)/det(T)
> lambda
[1] 10.68135
```

5-10　已知某研究对象分为三类,每个样品考察 4 项指标,各类的观测样品数分别为 7,4,6;另外还有 3 个待判样品(所有观测数据见表 5.4). 假定样本均来自正态总体.

（1）试用马氏距离判别法进行判别分析，并对 3 个待判样品进行判别归类.

（2）使用其他的判别法进行判别分析，并对 3 个待判样品进行判别归类，然后比较之.

表 5.4　判别分类的数据

样品号	X_1	X_2	X_3	X_4	类别号
1	6.0	−11.5	19.0	90.0	1
2	−11.0	−18.5	25.0	−36.0	3
3	90.2	−17.0	17.0	3.0	2
4	−4.0	−15.0	13.0	54.0	1
5	0.0	−14.0	20.0	35.0	2
6	0.5	−11.5	19.0	37.0	3
7	−10.0	−19.0	21.0	−42.0	3
8	0.0	−23.0	5.0	−35.0	1
9	20.0	−22.0	8.0	−20.0	3
10	−100.0	−21.4	7.0	−15.0	1
11	−100.0	−21.5	15.0	−40.0	2
12	13.0	−17.2	18.0	2.0	2
13	−5.0	−18.5	15.0	18.0	1
14	10.0	−18.0	14.0	50.0	1
15	−8.0	−14.0	16.0	56.0	1
16	0.6	−13.0	26.0	21.0	3
17	−40.0	−20.0	22.0	−50.0	3
1	−8.0	−14.0	16.0	56.0	
2	92.2	−17.0	18.0	3.0	
3	−14.0	−18.5	25.0	−36.0	

解.

(1)

```
> data<-read.csv("E:/4.多元统计分析/zuoye/4/10.csv")
> data$type<-as.factor(data$type)
> str(data)
'data.frame':   20 obs. of  5 variables:
$ x1  : num  6 -11 90.2 -4 0 0.5 -10 0 20 -100 ...
$ x2  : num  -11.5 -18.5 -17 -15 -14 -11.5 -19 -23 -22 -21.4 ...
$ x3  : int  19 25 17 13 20 19 21 5 8 7 ...
$ x4  : int  90 -36 3 54 35 37 -42 -35 -20 -15 ...
$ type: Factor w/ 3 levels "1","2","3": 1 3 2 1 2 3 3 1 3 1 ...

DDAM<-function (TrnX, TrnG, TstX = NULL, var.equal = FALSE){ #多元距离判别函数
        if ( is.factor(TrnG) == FALSE){
                mx<-nrow(TrnX); mg<-nrow(TrnG)
                TrnX<-rbind(TrnX, TrnG)
                TrnG<-factor(rep(1:2, c(mx, mg)))
        }
```

```
        if (is.null(TstX) == TRUE) TstX<-TrnX
        if (is.vector(TstX) == TRUE)  TstX<-t(as.matrix(TstX))
        else if (is.matrix(TstX) != TRUE)
        TstX<-as.matrix(TstX)
        if (is.matrix(TrnX) != TRUE) TrnX<-as.matrix(TrnX)
        nx<-nrow(TstX)
        blong<-matrix(rep(0, nx), nrow=1, dimnames=list("blong", 1:nx))
        g<-length(levels(TrnG))
        mu<-matrix(0, nrow=g, ncol=ncol(TrnX))
        for (i in 1:g)
        mu[i,]<-colMeans(TrnX[TrnG==i,])
        D<-matrix(0, nrow=g, ncol=nx)
        if (var.equal == TRUE  || var.equal == T){
                for (i in 1:g)
                D[i,]<- mahalanobis(TstX, mu[i,], var(TrnX))
        }
        else{
                for (i in 1:g)
                D[i,]<- mahalanobis(TstX, mu[i,], var(TrnX[TrnG==i,]))
        }
        for (j in 1:nx){
                dmin<-Inf
                for (i in 1:g)
                if (D[i,j]<dmin){
                        dmin<-D[i,j]; blong[j]<-i
                }
        }
        blong
}

> G<-gl(3,20)
> X<-data[,-5]
> DDAM(X,G)
Error in TrnX[TrnG == i, ] : (下标)逻辑下标太长
```

(2) 线性判别分析:

```
> library(MASS)
> View(A)
> ld<-lda(data$type~data$x1+data$x2+data$x3+data$x4)
> ld
Call:
```

```
lda(data$type ~ data$x1 + data$x2 + data$x3 + data$x4)  #公式

Prior probabilities of groups: #先验概率
        1         2         3
0.4117647 0.2352941 0.3529412

Group means: #各组均值向量
    data$x1   data$x2  data$x3   data$x4
1 -14.42857 -17.34286 12.71429  31.14286
2   0.80000 -17.42500 17.50000   0.00000
3  -6.65000 -17.33333 20.16667 -15.00000

Coefficients of linear discriminants: #第一和第二线性判别函数系数
                  LD1          LD2
data$x1 -0.009870482 -0.022839978
data$x2 -0.542919566  0.106647088
data$x3 -0.047312575  0.024128295
data$x4  0.068388163 -0.001001915

Proportion of trace: #两个判别式对判别的贡献大小
  LD1   LD2
0.996 0.004
> Z<-predict(ld)
> Z
$class
[1] 1 3 3 1 2 3 3 1 1 1 2 2 1 1 1 3 3
Levels: 1 2 3

$posterior
1           2            3
1  9.890173e-01 0.010600406 3.823032e-04
2  4.776285e-05 0.097968167 9.019841e-01
3  2.740321e-03 0.446093207 5.511665e-01
4  9.820309e-01 0.017255788 7.132755e-04
5  1.432291e-01 0.560319463 2.964514e-01
6  4.703544e-03 0.318882368 6.764141e-01
7  5.519341e-05 0.106111183 8.938336e-01
8  5.551665e-01 0.385618402 5.921511e-02
9  6.323055e-01 0.327991369 3.970312e-02
10 9.859693e-01 0.013226287 8.044569e-04
11 2.413737e-01 0.463482728 2.951436e-01
```

```
12 3.019693e-02 0.528897636 4.409054e-01
13 9.102735e-01 0.083079446 6.647017e-03
14 9.988689e-01 0.001120492 1.056153e-05
15 9.410302e-01 0.054669487 4.300268e-03
16 6.245518e-04 0.190138106 8.092373e-01
17 1.271870e-04 0.117787436 8.820854e-01

$x
LD1         LD2
1   2.20029533  0.28118465
2  -2.73225351  0.19394572
3  -1.49988665 -2.18959050
4   2.02112021  0.02761879
5  -0.19184441  0.23084042
6  -1.36998966  0.45990602
7  -2.69174289  0.02728051
8   0.61694890 -1.02077375
9   0.76050442 -1.31357007
10  2.00846388  1.46207772
11 -0.02544884  1.66948725
12 -0.74500229 -0.42254338
13  1.37461015 -0.23848051
14  3.19082694 -0.58394621
15  1.51252117  0.29600685
16 -1.98199600  0.48258009
17 -2.44712675  0.63797639

> X1<-X[18:20,]
> C<-predict(ld,X1)
> C
$class
[1] 1 3 3 1 2 3 3 1 1 1 2 2 1 1 1 3 3 1 3 3
Levels: 1 2 3

$posterior
1           2           3
1  9.890173e-01 0.010600406 3.823032e-04
2  4.776285e-05 0.097968167 9.019841e-01
3  2.740321e-03 0.446093207 5.511665e-01
4  9.820309e-01 0.017255788 7.132755e-04
5  1.432291e-01 0.560319463 2.964514e-01
```

```
6  4.703544e-03 0.318882368 6.764141e-01
7  5.519341e-05 0.106111183 8.938336e-01
8  5.551665e-01 0.385618402 5.921511e-02
9  6.323055e-01 0.327991369 3.970312e-02
10 9.859693e-01 0.013226287 8.044569e-04
11 2.413737e-01 0.463482728 2.951436e-01
12 3.019693e-02 0.528897636 4.409054e-01
13 9.102735e-01 0.083079446 6.647017e-03
14 9.988689e-01 0.001120492 1.056153e-05
15 9.410302e-01 0.054669487 4.300268e-03
16 6.245518e-04 0.190138106 8.092373e-01
17 1.271870e-04 0.117787436 8.820854e-01
18 9.410302e-01 0.054669487 4.300268e-03
19 2.224781e-03 0.429111664 5.686636e-01
20 5.285885e-05 0.099358550 9.005886e-01

$x
LD1         LD2
1   2.20029533  0.28118465
2  -2.73225351  0.19394572
3  -1.49988665 -2.18959050
4   2.02112021  0.02761879
5  -0.19184441  0.23084042
6  -1.36998966  0.45990602
7  -2.69174289  0.02728051
8   0.61694890 -1.02077375
9   0.76050442 -1.31357007
10  2.00846388  1.46207772
11 -0.02544884  1.66948725
12 -0.74500229 -0.42254338
13  1.37461015 -0.23848051
14  3.19082694 -0.58394621
15  1.51252117  0.29600685
16 -1.98199600  0.48258009
17 -2.44712675  0.63797639
18  1.51252117  0.29600685
19 -1.56694019 -2.21114217
20 -2.70264207  0.26246566
```

5-11　某城市的环保监测站于1982年在全市均匀地布置了14个监测点，每日三次定时抽取大气样品，测量大气中二氧化硫、氮氧化物和飘尘的含量. 前后5天，每个取样点(监测点)每种污染元素实测15次，取15次实测值的平均作为该取样点大气污染元素的含量(数据见表5.5). 表中最后一列给出的类号是使用第六章将介绍的聚类分析方法分析得到的结果(第1类为严重污染地区，第2类为一般污染地区，第3类为基本没有污染地区).

(1) 试用广义平方距离判别法建立判别准则(假设三个总体为多元正态总体，其协方差阵相等，先验概率取为各类样本的比例)，并列出回判结果.

(2) 该城市另有两个单位在同一期间测定了所在单位大气中这三种污染元素的含量(见表5.5中最后两行)，试用马氏距离判别方法判断这两个单位的污染情况属哪一类.

表5.5　大气污染数据

样品 \ 污染元素号	二氧化硫 (X_1)	氮氧化物 (X_2)	飘　尘 (X_3)	类别
1	0.045	0.043	0.265	2
2	0.066	0.039	0.264	2
3	0.094	0.061	0.194	2
4	0.003	0.003	0.102	3
5	0.048	0.015	0.106	3
6	0.210	0.066	0.263	1
7	0.086	0.072	0.274	2
8	0.196	0.072	0.211	1
9	0.187	0.082	0.301	1
10	0.053	0.060	0.209	2
11	0.020	0.008	0.112	3
12	0.035	0.015	0.170	3
13	0.205	0.068	0.284	1
14	0.088	0.058	0.215	2
15	0.101	0.052	0.181	
16	0.045	0.005	0.122	

解.

(1) 若三个总体为多元正太总体, 协方差阵相等, 先验概率取为各类样本的比例. 则

$$D_t^2(X) = d_t^2(X) + g_2(t)$$

三个总体的先验概率为

$$\begin{aligned} q_1 &= \frac{3}{7} \\ q_2 &= \frac{2}{7} \\ q_3 &= \frac{2}{7} \end{aligned} \tag{2}$$

那么

$$\begin{aligned} g_2(1) &= -2ln|q_1| = 1.694596 \\ g_2(2) &= -2ln|q_2| = 2.505526 \\ g_2(3) &= -2ln|q_3| = 2.505526 \end{aligned} \tag{3}$$

```
> data1<-read.csv("E:/4.多元统计分析/zuoye/4/11.csv")
> data11<-data1[1:14,]
> data12<-data1[15:16,]
> d1<-data11[which(data11$type==1), ]
> d2<-data11[which(data11$type==2), ]
> d3<-data11[which(data11$type==3), ]
> s1<-cov(d1[,-4])
> s2<-cov(d2[,-4])
> s3<-cov(d3[,-4])
> mu1<-colMeans(d1[,-4])
> mu2<-colMeans(d2[,-4])
> mu3<-colMeans(d3[,-4])
> D1<-mahalanobis(d1[,-4],mu1,s1)
> D2<-mahalanobis(d2[,-4],mu2,s2)
> D3<-mahalanobis(d3[,-4],mu3,s3)
> D1;D2;D3
6    8    9   13
2.25 2.25 2.25 2.25
1         2         3         7        10        14
1.9796692 2.5083563 2.1627834 4.1247155 3.3089299 0.9155458
4    5   11   12
2.25 2.25 2.25 2.25
```

(2)

```
> D11<-mahalanobis(data12[,-4],mu1,s1)
> D22<-mahalanobis(data12[,-4],mu2,s2)
> D33<-mahalanobis(data12[,-4],mu3,s3)
> D11;D22;D33
15       16
8368.24 33328.59
15        16
5.914895 38.756728
15      16
7020274 1944463
```

15 判为 2,16 判为 2.