《多元统计分析》第五次上机作业
判别分析

3180103000 许乐乐

- 实验目的与要求：

  通过本试验项目，能够理解并掌握如下内容：

  1. 处理判别分析的基本步骤；
  2. 熟悉各类判别方法；

习题一

利用第五章的数据和上机指导五，熟悉R在判别分析中的应用（请动手操作）。

一、线性判别——协方差阵相等

例5.1.1 P182

把A盆地和B盆地看作两个不同的总体，并假定两总体协方差阵相等，本例中变量个数$m=4$，两类总体各有5个训练样本($n_1=n_2=5$)，另有8个待判样品，接下来我们进行判别归类。

- $Hotelling\ T^2$检验是一种常用多变量检验方法，是单变量检验的自然推广，常用于两组均向量的比较。

  设两个样本含量分析为n,m的样本来自具有公共协方差阵的q维正态分布$N(\mu_1,\Sigma)$,$N(\mu_2,\Sigma)$,欲检验

  $$H_0:\mu_1=\mu_2,\qquad H_1:\mu_1\neq\mu_2$$

  分别计算出两样本每个变量的均值构成的均向量$\bar{X}$,$\bar{Y}$及合并的组内协方差阵S，则统计量T^2为

  $$T^2=\frac{nm}{n+m}(\bar{X}-\bar{Y})'S^{-1}(\bar{X}-\bar{Y})$$

  其中，$S=(A_1+A_2)/(n+m-2)$为合并协方差矩阵，A_1,A_2分别为两样本的离差阵。

  求得T^2后，可查相应界值表得到p值，从而作出结论。但通常将其转换为统计量F再由F分布得到P值。

```
library(ICSNP) # 加载ICSNP包(包括mvtnorm, ICS, survey)
D511<-read.table("ex511.txt", header=F)
D511A<-D511[1:5,1:4]
D511B<-D511[6:10,1:4]
HotellingsT2(D511A, D511B)
```

```
    Hotelling's two sample T2-test

data:  D511A and D511B
T.2 = 14.464, df1 = 4, df2 = 5, p-value =
0.005891
alternative hypothesis: true location difference is not equal to
c(0,0,0,0)
```

- 判别分析，判别函数建立

`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`为容忍度，回归分析中关于共线性诊断所提到的一样
- `method="moment"`均值方差均为矩估计，`method="mle"`为极大似然估计
- `CV`参数是输出leave-one-out的后验概率

```
library(MASS)
attach(D511)
results<-lda(V5~V1+V2+V3+V4, D511)
```

```
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
```

- 利用判别函数做判别

`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

```
D511P<-read.table("ex511A.txt", header=F)
predict(results, D511P)
```

```
$class
[1] B A A B B A A A
Levels: A B

$posterior
                A
1 0.001639701083
2 1.000000000000
3 1.000000000000
4 0.083024235408
5 0.000001190922
6 0.999999999887
7 1.000000000000
```

```
8 1.000000000000

                    B
1
0.998360298917242094951518538437085226178169250488281250000000000000000000000000000000000000000000000000000000000000000000
2
0.000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000001932625
3
0.000000000000000000012696186439811967499972922279738440920482389628887176513671875000000000
4
0.916975764592251385565191412752028554677963256835937500000000000000000000000000000000000000000000000000000000000000000000
5
0.999998809077888961560631742031546309590339660644531250000000000000000000000000000000000000000000000000000000000000000000
6
0.000000000112961054163488245748875327123528222728054970502853393554687500000000000000000000
7
0.000000000000000000000000001161894355979006759004533433365935479741892777383327484130859375
8
0.000000000000000000000713590329521449026313820174216573377634631469845771789550781250000000

$x
          LD1
1   1.0536512
2 -31.2985593
3  -7.5286829
4   0.3947245
5   2.2416596
6  -3.7639282
7  -9.8136273
8  -8.0017623

> View(D511P)
> summary(results)
        Length Class  Mode
prior   2      -none- numeric
counts  2      -none- numeric
means   8      -none- numeric
scaling 4      -none- numeric
lev     2      -none- character
svd     1      -none- numeric
N       1      -none- numeric
call    3      -none- call
terms   3      terms  call
xlevels 0      -none- list
> results
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
```

二、二次判别函数——协方差阵不等

例5.2.2 P191

此例中总体个数$k=3$，变量个数$m=4$，各组样品个数为：$n_1=n_2=n_3=5(n=15)$。这是多总体的判别归类问题。假设三个总体的协方差阵不相等，采用广义平方距离进行判别归类。我们在实际应用中把胃癌发病率的先验概率取为相等。MASS扩展包的`qda()`采用的是广义平方距离。

```
D522<-read.table("ex522.txt", header=F)
library(MASS)
results<-qda(V1~V2+V3+V4+V5, D522)
predict(results, D522)
```

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

三、主成分判别法

例5.2.2 P191

```
library(adegenet)
X<-D522[, 2:5]
Y<-as.factor(D522[,1])
# 开始有误！！！！
D<-dapc(X, Y)
Summary(D)
predict(D, X)
```

四、逐步判别法

例5.1.1 P182

```
D511<-read.table("ex511.txt", header=F)
attach(D511)
A<-as.factor(V5)
MD511<-as.matrix(D511[,1:4])
library(caret)
slda<-train(V5 ~ .,data=D511,method="stepLDA",
            trControl=trainControl(method = "cv"))
B<-predict(slda,D511)
table(B,V5) # 查看错判
```

五、多个总体协方差矩阵是否相等的检验

- 函数`varcomp(list, df)`
    - `list`是协方差阵的列表
    - `df=length(covmat)`计算协方差阵的样本量

  这个函数原是R的一个包，但后被从R 包中删除了。具体代码如下，请大家用之前先运行这个函数。

```
varcomp <- function(covmat,n) {
   if (is.list(covmat)) {
    if (length(covmat) < 2)
        stop("covmat must be a list with at least 2 elements")
    ps <- as.vector(sapply(covmat,dim))
    if (sum(ps[1] == ps) != length(ps))
        stop("all covariance matrices must have the same dimension")
    p <- ps[1]
        q <- length(covmat)
        if (length(n) == 1)
        Ng <- rep(n,q)
    else if (length(n) == q)
        Ng <- n
    else
        stop("n must be equal length(covmat) or 1")

    DNAME <- deparse(substitute(covmat))
   }

   else
    stop("covmat must be a list")

   ng <- Ng - 1
   Ag <- lapply(1:length(covmat),function(i,mat,n) { n[i] * mat[[i]]
},mat=covmat,n=ng)
   A <- matrix(colSums(matrix(unlist(Ag),ncol=p^2,byrow=T)),ncol=p)
   detAg <- sapply(Ag,det)
   detA <- det(A)
   V1 <- prod(detAg^(ng/2))/(detA^(sum(ng)/2))
   kg <- ng/sum(ng)
   l1 <- prod((1/kg)^kg)^(p*sum(ng)/2) * V1
   rho <- 1 - (sum(1/ng) - 1/sum(ng))*(2*p^2+3*p-1)/(6*(p+1)*(q-1))
   w2 <- p*(p+1) * ((p-1)*(p+2) * (sum(1/ng^2) - 1/(sum(ng)^2)) - 6*(q-
1)*(1-rho)^2) / (48*rho^2)
   f <- 0.5 * (q-1)*p*(p+1)
   STATISTIC <- -2*rho*log(l1)
```

```
   PVAL <- 1 - (pchisq(STATISTIC,f) + w2*(pchisq(STATISTIC,f+4) -
pchisq(STATISTIC,f)))
   names(STATISTIC) <- "corrected lambda*"
   names(f) <- "df"
   RVAL <- structure(list(statistic = STATISTIC, parameter = f,p.value =
PVAL, data.name = DNAME, method = "Equality of Covariances Matrices
Test"),class="htest")
   return(RVAL)
}
```

```
data1<-D511[1:5,1:4]
data2<-D511[6:10,1:4]
s1<-cov(data1)
s2<-cov(data2)
covmat<-list(s1,s2)
varcomp(covmat,n=c(5,5))
```

```
    Equality of Covariances Matrices Test

data:  covmat
corrected lambda* = 19.055, df = 10, p-value =
0.09607
```

p-value = 0.09607，这说明如果显著水平为5%,不能否协方差阵相等的假设。

习题二

采用“肝胆病患者检查数据”（见附件）。这是一组医院病人的资料，基本包括了四变量分别为：总胆红素(umol/L)，白蛋白（g/L），碱性磷酸酶，谷丙转氨酶和医生诊断结果，希望通过这组数据（学习样本）建立判别肝、胆疾病的判别函数，并应用于“体检数据”中，根据体检资料（见附件）分析是否有得肝胆疾病的可能性；

- 检验组间均值显著性差异（多元方差分析）

  多元方差分析(multivariate analysis of variance, MANOVA)是单变量方差分析和$Hotelling\ T^2$检验的推广，用于多组均向量间的比较。

  设含量为$n_1, n_2, \ldots, n_g$的g个样本分别来自q维多元正态分布，$Nq(\mu_1, \Sigma), Nq(\mu_2, \Sigma), \ldots, Nq(\mu_g, \Sigma)$，则可根据交并原则或极大似然比原则推出多个统计量用于判断多个均向量是否来自同一总体。常用的统计量有$Wilks\lambda$，$Pillai$迹，$Hotelling-Lawley$迹和Roy的最大特征根。他们都以组内和组间离差阵有关，且其推断结论一般都一致，故这里仅介绍最常用的$Wilks$统计量：

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

  其中A为合并的组内离差阵，B为组间离差阵，T为总离差阵。Λ满足于自由度分别为$q, n-g, g-1$的$Wilks$分布，可查其相应界值表得到p值，但通常也是转换为F分布后再得到p值。

  用`manova()`进行多元方差分析检验。

```
d<-read_csv("hw5_2_1.csv")
fit<-manova(cbind(BIL,Alb,ALP,ALT)~group,data=d)
summary(fit, test="Wilks")
```

```
           Df  Wilks approx F num Df den Df
group       1 0.3661   146.74      4    339
Residuals 342
                          Pr(>F)
group     < 0.00000000000000022 ***
Residuals
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

由p值可知，否定H_0，这表明这五个组的指标之间有显著的差异。

- 分析组间方差显著性差异
- 线性判别

```
# 线性判别
library(MASS)
results = lda(group~BIL+Alb+ALP+ALT, d)
print(results)
```

```
Call:
lda(group ~ BIL + Alb + ALP + ALT, data = d)

Prior probabilities of groups: # 先验概率
         1          2          3          4          5
0.31395349 0.04360465 0.17151163 0.09883721 0.37209302

Group means: # 各组向量均值
        BIL      Alb       ALP        ALT
1  88.96111 33.05463 139.00926  310.74074
2 155.84667 34.84000 192.13333 1303.60000
3  19.15593 42.70339  77.64407   31.77966
4  23.54412 39.04706 109.41176   47.67647
5  14.05234 47.69375  60.10938   23.28906

Coefficients of linear discriminants: # 线性判别函数的系数
             LD1           LD2           LD3          LD4
BIL -0.002624856 -0.0014151382  0.0169912889 -0.006453130
Alb  0.158422305 -0.1181964481  0.0033508644 -0.099499017
ALP -0.005435544  0.0001076113 -0.0121008443 -0.017931771
ALT -0.002135831 -0.0028019445 -0.0007891008  0.001607685

Proportion of trace:
   LD1    LD2    LD3    LD4
0.8617 0.1300 0.0071 0.0011
```

从均值数据上来看，组3和组5并没有明显差异，而组2的ALT值显著高，组4的ALP较高，组1的ALP和ALT较高。

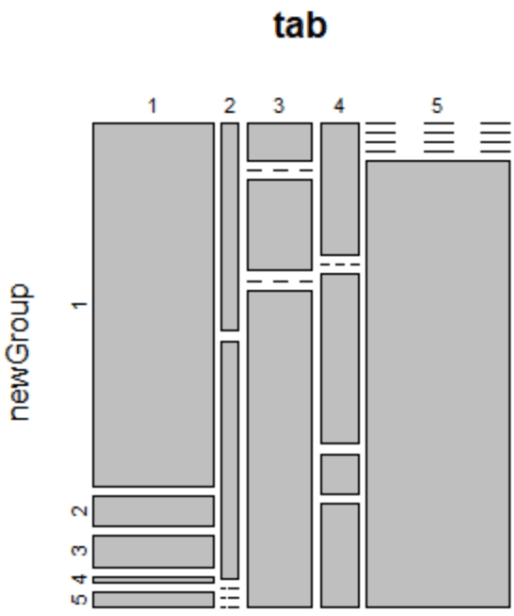
```
# 回判
pre<-predict(results,d[,1:4])
# 或直接写为predict(results)
newGroup<-pre$class # 预测的所属类的结果
cbind(d$group,newGroup) # 显示预测前后分组结果
# 对模型进行评价
tab<-table(d$group,newGroup) # 绘制混淆矩阵
```

```
   newGroup
      1   2   3   4   5
  1  88   7   8   1   4
  2   7   8   0   0   0
  3   5   0  12   0  42
  4  10   0  13   3   8
  5   0   0   0   0 128
```

```
erro<-1-sum(diag(prop.table(tab))) # 计算误判率
```

```
[1] 0.3052326
```

```
plot(tab) # 可视化
```



由上可知：

- 一共错判了105组数据，错判比率30.52%。
- 组2、组3和组4被误判的比较严重。

- 二次判别

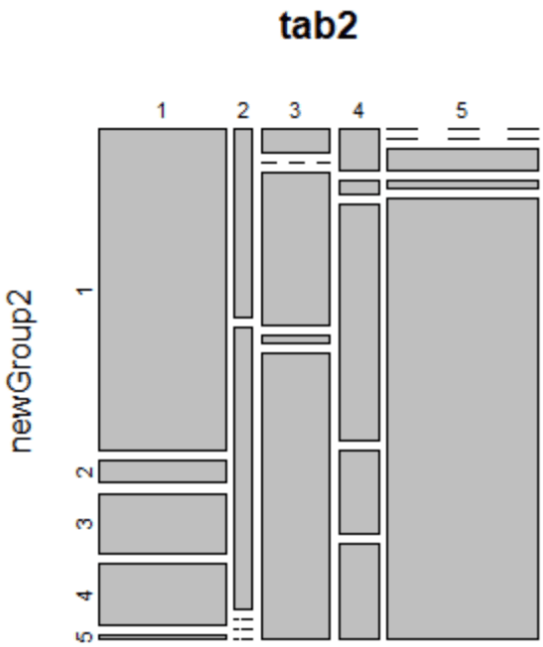
```
results2 = qda(group~BIL+Alb+ALP+ALT, d)
print(results)
# 回判
pre2<-predict(results2,d[,1:4])
# 或直接写为predict(results)
newGroup2<-pre2$class # 预测的所属类的结果
cbind(d$group,newGroup2) # 显示预测前后分组结果
# 对模型进行评价
tab2<-table(d$group,newGroup2) # 绘制混淆矩阵
```

```
   newGroup2
      1   2   3   4   5
  1  74   5  14  14   1
  2   6   9   0   0   0
  3   3   0  19   1  36
  4   3   1  17   6   7
  5   0   0   6   2 120
```

```
erro2<-1-sum(diag(prop.table(tab2))) # 计算误判率
```

```
[1] 0.3372093
```

```
plot(tab2) # 可视化
```



由上可知：

- 错判了116个，错判比率33.72%，比线性判别法错判比率更高。
- 组3和组4误判几率高，但没有线性判别法那么严重，但组1的错判率更高了。

- 对新数据进行判断

```
# 对新数据进行判断
new<-read_csv("hw5_2_2.csv")
new.na<-na.omit(new)
```

- 线性判别法

```
pre_new<-predict(results,new.na)
pre_new$class
```

```
  [1] 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
 [25] 5 5 5 5 5 5 3 5 5 5 5 5 5 5 5 5 5 5 5 5 5 3 5
 [49] 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
 [73] 5 5 5 5 5 5 5 3 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
 [97] 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
[121] 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
[145] 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
[169] 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
[193] 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
[217] 5 5 5 5 5 5
Levels: 1 2 3 4 5
```

```
table(newGroup_new)
```

```
newGroup_new
  1   2   3   4   5
  0   0   3   0 219
```

线性判别法预测出来只有3个人患慢胆，其他人正常。

- 二次判别法

```
pre_new2<-predict(results2,new.na)
newGroup_new2<-pre_new2$class # 预测的所属类的结果
```

```
  [1] 5 5 5 5 5 5 5 5 3 3 5 5 5 5 5 5 5 3 5 5 3 5 5 5
 [25] 5 5 5 5 5 5 5 5 5 5 5 3 5 5 5 5 3 5 5 5 5 5 3 5
 [49] 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
 [73] 5 5 5 5 5 5 5 5 5 5 5 3 5 5 5 5 5 5 5 5 5 3 5
 [97] 5 5 5 5 5 5 5 5 5 3 5 5 5 5 5 5 5 5 1 5 5 5 5
[121] 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 3 5 3 5 5 5
[145] 5 5 5 5 5 5 3 5 5 5 5 5 5 5 5 5 3 5 3 5 5 5 5 5
[169] 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
[193] 5 5 5 5 5 5 3 5 5 5 5 5 5 5 5 5 5 5 1 5 5 3 5
[217] 5 1 3 5 5 5
Levels: 1 2 3 4 5
```

```
table(newGroup_new2)
```

```
newGroup_new
  1   2   3   4   5
  3   0  18   0 201
```

二次判别法预测出来有3个人患肝炎，18个人患慢胆，其他人正常。