第二次作业

辛柏嬴 2020111753

2023-04-12

#### （1） 习题4.11

path1 <- "/Users/xinby/Desktop/Sufe/Multivariate-Stat-Analysis/hw2/exec4.11.xlsx"  
data1 <- readxl::read\_excel(path1)  
g = factor(data1$g)  
y = cbind(data1$x1, data1$x2)  
fit = manova(y ~ g)  
summary(fit, test = "Wilks")

## Df Wilks approx F num Df den Df Pr(>F)  
## g 2 0.92539 1.1069 4 112 0.3569  
## Residuals 57

从上述输出结果可见，Wilk统计量的p值为0.3569，大于0.05，因此无法拒绝原假设。故在统计学上认为没有明显差距。

#### （2） 例4.4.2

**a. 检验轮廓的平行性**

path2 <- "/Users/xinby/Desktop/Sufe/Multivariate-Stat-Analysis/hw2/examp4.4.2.xlsx"  
data2 <- readxl::read\_excel(path2)  
  
# separate the data into grps:  
x <- filter(data2, g == 1) %>%  
 select(-g)  
y <- filter(data2, g == 2) %>%  
 select(-g)  
dat2 <- select(data2, -g)  
  
# calculate average and variance of data:  
x.bar <- apply(x, 2, mean)  
y.bar <- apply(y, 2, mean)  
S1 <- cov(x)  
S2 <- cov(y)  
sp <- (29 \* S1 + 29 \* S2)/(58)  
C <- matrix(c(-1, 0, 0, 1, -1, 0, 0, 1, -1, 0, 0, 1), ncol = 4)  
  
# to test if parallel:  
c1 <- C %\*% (x.bar - y.bar)  
c2 <- C %\*% sp %\*% t(C)  
T2 <- 30 \* 30/(30 + 30) \* t(c1) %\*% solve(c2, c1)  
T.05 <- 3 \* 58/56 \* qf(0.95, 3, 56)  
  
print(T2)  
print(T.05)

## [,1]  
## [1,] 8.016171  
## [1] 8.605018

由上述输出结果可知，T2结果小于临界值，因此不能拒绝原平行假设。

**b. 检验两轮廓是否重合**

s1 <- sum(x.bar - y.bar)  
s2 <- sum(sp)  
# calculate T\_square  
T2\_b <- 30 \* 30/(30 + 30) \* s1^2/s2  
print(T2\_b)  
# compare with F quantile  
qf(0.95, 1, 58)

## [1] 1.53277  
## [1] 4.006873

由上述输出结果可知，计算结果小于临界值，无法拒绝原假设，量轮廓重合。

**c. 检验两轮廓水平**

# calculate T\_square  
Cz <- 0.5 \* C %\*% (x.bar + y.bar)  
S <- cov(dat2)  
csc <- C %\*% S %\*% t(C)  
T2\_c <- (30 + 30) \* t(Cz) %\*% solve(csc, Cz)  
T.05\_c <- 3 \* 59/57 \* qf(0.95, 3, 57)  
print(T2\_c)  
print(T.05\_c)

## [,1]  
## [1,] 24.82071  
## [1] 8.590518

由上述输出结果可知，计算结果大于临界值，拒绝原假设。 因此认为轮廓不是水平的，四个问题的回答有区别。

#### （3） 习题5.5

**a. 试给出判别规则，并预报明天是否会下雨，用回代法估计误判概率**

# import data  
path3.train <- "/Users/xinby/Desktop/Sufe/Multivariate-Stat-Analysis/hw2/exec5.5.xlsx"  
path3.test <- "/Users/xinby/Desktop/Sufe/Multivariate-Stat-Analysis/hw2/exec5.5a.xlsx"  
dat3 <- readxl::read\_excel(path3.train)  
dat3.test <- readxl::read\_excel(path3.test)  
  
# separate data by grps  
rain <- filter(dat3, g == 1) %>%  
 select(-g)  
rain <- as.matrix(rain)  
nrain <- filter(dat3, g == 2) %>%  
 select(-g)  
nrain <- as.matrix(nrain)  
dat3.train <- dat3 %>%  
 select(-g)  
dat3.train <- as.matrix(dat3.train)  
dat3.test <- t(as.matrix(dat3.test))  
  
# test if same variance (the result shows not equal)  
var1 <- cov(rain)  
var2 <- cov(nrain)  
  
# calculate mean and var for each grp  
mu1 <- apply(rain, 2, mean)  
mu2 <- apply(nrain, 2, mean)  
  
# calculate discriminator W(x) with new data  
d1 <- t(dat3.test - mu1) %\*% solve(var1, dat3.test - mu1)  
d2 <- t(dat3.test - mu2) %\*% solve(var2, dat3.test - mu2)  
Wx <- d1 - d2  
print(Wx)  
  
# calculate misjudge rate  
mis1.2 = 0  
mis2.1 = 0  
all = 0  
  
# go through the train set to see if it is correctly judged  
for (line in 1:20) {  
 # calculate each obs' distance  
 d1.test <- t(dat3.train[line, ] - mu1) %\*% solve(var1, dat3.train[line, ] - mu1)  
 d2.test <- t(dat3.train[line, ] - mu2) %\*% solve(var2, dat3.train[line, ] - mu2)  
 # judge which group it belongs to  
 Wx.test <- 0  
 if (d1.test - d2.test <= 0) {  
 Wx.test[line] = 1  
 } else {  
 Wx.test[line] = 2  
 }  
 # count the correct rate  
 if (Wx.test[line] == 1 && dat3$g[line] == 2) {  
 mis1.2 = mis1.2 + 1  
 }  
 if (Wx.test[line] == 2 && dat3$g[line] == 1) {  
 mis2.1 = mis2.1 + 1  
 }  
 all = all + 1  
}  
  
misjudge\_rate\_2.1 = mis2.1/all  
misjudge\_rate\_1.2 = mis1.2/all  
print(misjudge\_rate\_2.1)  
print(misjudge\_rate\_1.2)

## [,1]  
## [1,] -1.950508  
## [1] 0  
## [1] 0.25

由于方差计算认为两组相差较大，故采用二次型判别规则。 令

则有：

由上述输出结果可知，W(x)<0，因此认为明天会下雨。误判概率如上输出所示。

**b. 给定先验概率，p1=0.3, p2=0.7，预报每天是否会下雨**

由Bayes准则的最大后验法，在正态分布条件下有：

p1 <- 0.3  
p2 <- 0.7  
Pr1 <- exp(-0.5 \* (d1 + log(det(var1)) - 2 \* log(p1)))  
Pr2 <- exp(-0.5 \* (d2 + log(det(var2)) - 2 \* log(p2)))  
Prob.pi1\_x <- Pr1/(Pr1 + Pr2)  
Prob.pi2\_x <- Pr2/(Pr1 + Pr2)

由上述输出内容可知，由Bayes判别，认为明天不下雨。

**c. 在b的条件下考虑误判期望c(2|1)=3c(1|2)，判断是否要在明天举行活动？**

此时加入考虑误判期望时，由于二次判别的判别效果严重依赖于数据的正态分布， 而在本题中数据数量较少，故采用线性判别规则：

criteria <- log(3 \* (p1)/(p2))  
print(criteria)  
  
a <- solve(cov(dat3.train), mu1 - mu2)  
mu = 0.5 \* (mu1 + mu2)  
print(t(a) %\*% (dat3.test - mu))

## [1] 0.2513144  
## [,1]  
## [1,] 0.3622237

由上述输出结果，根据判别规则，当考虑误判代价时，明天不应该举行活动。

#### (4) 习题5.6

**a. 对于14名运动员，分别在方差相等和不等的假设下进行Bayes判别**

在正态分布假设下，若考虑误判概率相等，先验概率相等，当同方差时有：

在异方差时有：

其中

# import the data  
path4.train <- "/Users/xinby/Desktop/Sufe/Multivariate-Stat-Analysis/hw2/exec5.6.xlsx"  
path4.test <- "/Users/xinby/Desktop/Sufe/Multivariate-Stat-Analysis/hw2/exec5.6a.xlsx"  
dat4.train <- readxl::read\_excel(path4.train)  
dat4.test <- readxl::read\_excel(path4.test)  
  
# Same Variance Assumption: USING `lda`  
prd\_1 <- lda(g ~ x1 + x2 + x3 + x4 + x5 + x6, prior = c(0.5, 0.5), data = dat4.train)  
  
# Different Variance Assumption: USING `qda`  
prd\_2 <- qda(g ~ x1 + x2 + x3 + x4 + x5 + x6, prior = c(0.5, 0.5), data = dat4.train)  
  
# Output the classification result  
id <- c(1:14)  
class\_1 <- predict(prd\_1, dat4.test)$class  
class\_2 <- predict(prd\_2, dat4.test)$class  
result <- cbind(id, class\_1, class\_2)  
print("classification :")  
print(result)

## [1] "classification :"  
## id class\_1 class\_2  
## [1,] 1 1 1  
## [2,] 2 1 1  
## [3,] 3 1 1  
## [4,] 4 1 1  
## [5,] 5 1 1  
## [6,] 6 1 1  
## [7,] 7 1 1  
## [8,] 8 2 2  
## [9,] 9 2 2  
## [10,] 10 1 1  
## [11,] 11 2 2  
## [12,] 12 2 2  
## [13,] 13 1 2  
## [14,] 14 2 2

由上述输出结果可见，同方差假设下，一级运动员有1～7、10、13，健将级运动员有：8、9、11、12、14； 异方差假设下，一级运动员有1～7、10，健将级运动员有：8、9、11～14.

**b. 试按照回代法和交叉验证法分别对（1）的误判概率进行估计**

# Same Var in-sample validation  
class\_train.eqvar <- predict(prd\_1, dat4.train)$class  
real\_train <- dat4.train$g  
print("same var - in\_sample vald.")  
print(prop.table(table(g = real\_train, class\_train.eqvar), 1))  
  
# cross validation  
prd\_1.cv <- lda(g ~ x1 + x2 + x3 + x4 + x5 + x6, prior = c(0.5, 0.5), CV = TRUE,  
 data = dat4.train)  
print("same var - cross vald.")  
print(prop.table(table(g = real\_train, prd\_1.cv$class), 1))  
  
# Diff Var in-sample validation  
class\_train.dfvar <- predict(prd\_2, dat4.train)$class  
print("diff. var - in\_sample vald.")  
print(prop.table(table(g = real\_train, class\_train.dfvar), 1))  
  
# cross validation  
prd\_2.cv <- qda(g ~ x1 + x2 + x3 + x4 + x5 + x6, prior = c(0.5, 0.5), CV = TRUE,  
 data = dat4.train)  
print("diff var - cross vald.")  
print(prop.table(table(g = real\_train, prd\_2.cv$class), 1))

## [1] "same var - in\_sample vald."  
## class\_train.eqvar  
## g 1 2  
## 1 1 0  
## 2 0 1  
## [1] "same var - cross vald."  
##   
## g 1 2  
## 1 1.00 0.00  
## 2 0.08 0.92  
## [1] "diff. var - in\_sample vald."  
## class\_train.dfvar  
## g 1 2  
## 1 1 0  
## 2 0 1  
## [1] "diff var - cross vald."  
##   
## g 1 2  
## 1 1 0  
## 2 0 1

检验结果如上输出所示。 由此可知，对于同方差假设，回代法检验的误判概率都为0，交叉验证的误判概率为P1|2=0.08, P2|1=0； 对于异方差假设，回代法和交叉验证法得到的误判概率均为0.

**c. 假设同方差，p1=0.8, p2=0.2，试着对这14名运动员进行Bayes判别**

prd\_3 <- lda(g ~ x1 + x2 + x3 + x4 + x5 + x6, prior = c(0.8, 0.2), data = dat4.train)  
result <- predict(prd\_3, dat4.test)$class  
cbind(id, result)

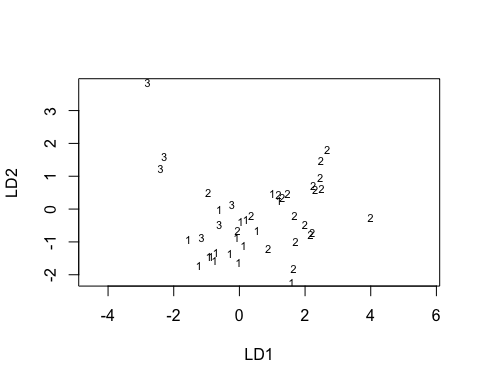
## id result  
## [1,] 1 1  
## [2,] 2 1  
## [3,] 3 1  
## [4,] 4 1  
## [5,] 5 1  
## [6,] 6 1  
## [7,] 7 1  
## [8,] 8 1  
## [9,] 9 2  
## [10,] 10 1  
## [11,] 11 2  
## [12,] 12 2  
## [13,] 13 1  
## [14,] 14 2

由上述输出可知，在考虑先验概率的条件下，一级运动员有1～8、10、13；健将级运动员有9、11、12、14.

#### (5) 习题5.8

**试给出Fisher判别函数，将所有品牌的两个判别函数得分画成散点图，用不同符号表示不同厂商**

path5 <- "/Users/xinby/Desktop/Sufe/Multivariate-Stat-Analysis/hw2/exec5.8.xlsx"  
dat5 <- readxl::read\_excel(path5)  
fisher <- lda(dat5$g ~ ., dat5)  
print(fisher$scaling)  
plot(fisher)



## LD1 LD2  
## x1 0.022344104 0.045417606  
## x2 0.369109646 -0.332405063  
## x3 -0.837675738 -0.386499597  
## x4 -0.000763493 -0.006017311  
## x5 1.420281838 1.039957871  
## x6 0.202200109 -0.203863959  
## x7 0.195246015 -0.235306430  
## x8 -0.030687468 -0.026966644

散点图如上图所示。第i判别函数的形式为：

其系数如上输出所示。在本题中，由于，故可以参考第一、二判别函数。