REPORT

수강과목	:	다변량통계학(II)
담당교수	:	최용석
학 과	:	통계학과
학 번	:	201611531
이 름	:	정호재
제출일자	:	2020.09.24

HW1. for Multivariate Statistics II

September 17, 2020

Chapter 6. Discriminant and Classification analysis(DCA)

1. [Data 6.12.1](salmon.txt) shows the growth ring diameter of salmon caught in both two clusters of Alaska and Canada. Diameter of rings for the first-year freshwater growth and for the first-year marine growth were measured in hundredths of an inch. In addition females and males are coded as 1 and 2 respectively.

(1) Test the bivariate normality of two clusters.

```
> #install.packages("MASS")
> #install.packages("MVN")
> #install.packages("haven")
> library(MASS)
> library(MVN)
> setwd("D:/학교/2020 2학기 정호재/다변량통계학2/20200917/Rdata(all)")
> salmon<-read.table("salmon.txt", header=T)
> #attach(salmon)
> head(salmon) # 변수는 총 4가지/class가 타겟변수
  class sex  x1  x2
1     1   2 108 368
2     1   1 131 355
3     1   1 105 469
4     1   2  86 506
5     1   1  99 402
6     1   2  87 423
> str(salmon)
'data.frame':   100 obs. of  4 variables:
 $ class: int  1 1 1 1 1 1 1 1 1 1 ...
 $ sex  : int  2 1 1 2 1 2 1 2 2 1 ...
 $ x1   : int  108 131 105 86 99 87 94 117 79 99 ...
```

```
 $ x2   : int  368 355 469 506 402 423 440 489 432 403 ...

> unique(salmon[,1]) #타겟 값 보기/unique 중복 없이 유일한 값만 추출
[1] 1 2

> # MVN tests based on the Skewness and Kurtosis Statistics
> salmon_1=salmon[1:50, 3:4]
> salmon_2=salmon[51:100, 3:4] # 각각 종류별로 데이터 분리
> salmon_1
    x1  x2
1  108 368
2  131 355
3  105 469
4   86 506
5   99 402
6   87 423
7   94 440
8  117 489
9   79 432
10  99 403
11 114 428
12 123 372
13 123 372
14 109 420
15 112 394
16 104 407
17 111 422
18 126 423
19 105 434
20 119 474
21 114 396
22 100 470
23  84 399
24 102 429
25 101 469
26  85 444
27 109 397
28 106 442
29  82 431
30 118 381
31 105 388
32 121 403
33  85 451
34  83 453
35  53 427
36  95 411
37  76 442
38  95 426
39  87 402
40  70 397
41  84 511
42  91 469
43  74 451
44 101 474
45  80 398
46  95 433
47  92 404
48  99 481
49  94 491
50  87 480
############################################################
```

```
> salmon_2
     x1  x2
51  129 420
52  148 371
53  179 407
54  152 381
55  166 377
56  124 389
57  156 419
58  131 345
59  140 362
60  144 345
61  149 393
62  108 330
63  135 355
64  170 386
65  152 301
66  153 397
67  152 301
68  136 438
69  122 306
70  148 383
71   90 385
72  145 337
73  123 364
74  145 376
75  115 354
76  134 383
77  117 355
78  126 345
79  118 379
80  120 369
81  153 403
82  150 354
83  154 390
84  155 349
85  109 325
86  117 344
87  128 400
88  144 403
89  163 370
90  145 355
91  133 375
92  128 383
93  123 349
94  144 373
95  140 388
96  150 339
97  124 341
98  125 346
99  153 352
100 108 339
```

```
> result_salmon_1 = mvn(salmon_1)
> result_salmon_2 = mvn(salmon_2) # 다변량 정규성 검정
```

```
> result_salmon_1
$multivariateNormality
              Test            Statistic             p value Result
1 Mardia Skewness    6.67860033109113 0.153879248834123     YES
2 Mardia Kurtosis -0.543809625889093 0.586572484145819     YES
3             MVN                <NA>                <NA>     YES

$univariateNormality
           Test  Variable Statistic    p value Normality
1 Shapiro-Wilk     x1          0.9874     0.8664     YES
2 Shapiro-Wilk     x2          0.9778     0.4639     YES

$Descriptives
    n    Mean   Std.Dev Median Min Max    25th   75th         Skew    Kurtosis
x1 50   98.38 16.14335    99.0   53 131   86.25 109.0 -0.2117158 -0.2055946
x2 50 429.66 37.40436   427.5 355 511 402.00 452.5   0.2277552 -0.7279458

> result_salmon_2
$multivariateNormality
              Test            Statistic             p value Result
1 Mardia Skewness    0.782204925537997 0.940817697109154     YES
2 Mardia Kurtosis -0.0288686753407277 0.976969328661072     YES
3             MVN                <NA>                <NA>     YES

$univariateNormality
           Test  Variable Statistic    p value Normality
1 Shapiro-Wilk     x1          0.9816     0.6197     YES
2 Shapiro-Wilk     x2          0.9847     0.7574     YES

$Descriptives
    n    Mean   Std.Dev Median Min Max    25th    75th         Skew    Kurtosis
x1 50  137.46 18.05797   140.0   90 179  124.00  151.50 -0.17056928 -0.3209330
x2 50 366.62 29.88747   369.5 301 438 345.25 385.75 -0.07051701 -0.2212072
```

mvn test결과 두 데이터 모두 YES이므로 다변량 정규성 만족한다.

(2) Test the homogeneity of the covariance matrices of the two clusters.

```
> salmon<-read.table("salmon.txt", header=T)
> #attach(salmon)
> group=as.factor(class) #섹터 별로 저장
> group
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1
 [50] 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2
 [99] 2 2
Levels: 1 2
> #공분산행렬 동질성 검정
> cov.Mtest=function(x,ina,a=0.05){
+   ## x is the data set
+   ## ina is a numeric vector indicating the groups of the data set
+   ## a is the significance level, set to 0.05 by default
+   x=as.matrix(x)
+   p=ncol(x) ## dimension of the data set
+   n=nrow(x) ## total sample size
+   k=max(ina) ## number of groups
+   nu=rep(0,k) ## the sample size of each group will be stored here later
+   pame=rep(0,k) ## the determinant of each covariance will be stored here
+   ## the next "for" function calculates the covariance matrix of each group
+   nu=as.vector(table(ina))
+   mat=mat1=array(dim=c(p,p,k))
+   for (i in 1:k) {
+     mat[,,i]=cov(x[ina==i,])
+     pame[i]=det(mat[,,i]) ## the detemirnant of each covariance matrix
+     mat1[,,i]=(nu[i]-1)*cov(x[ina==i,]) }
+   ## the next 2 lines calculate the pooled covariance matrix
+   Sp=apply(mat1,1:2,sum)
+   Sp=Sp/(n-k)
+   for (i in 1:k)
+     pamela=det(Sp) ## determinant of the pooled covariance matrix
+   test1=sum((nu-1)*log(pamela/pame))
+   gama1=(2*(p^2)+3*p-1)/(6*(p+1)*(k-1))
+   gama2=(sum(1/(nu-1))-1/(n-k))
+   gama=1-gama1*gama2
+   test=gama*test1 ## this is the M (test statistic)
+   df=0.5*p*(p+1)*(k-1) ## degrees of freedom of the chi-square distribution
+   pvalue=1-pchisq(test,df) ## p-value of the test statistic
+   crit=qchisq(1-a,df) ## critical value of the chi-square distribution
```

```
+     list(M.test=test,degrees=df,critical=crit,p.value=pvalue) }
> ina=as.numeric(as.factor(salmon[, 1]))
> x=salmon[, 3:4]
> cov.Mtest(x, ina)
$M.test
[1] 10.69615

$degrees
[1] 3

$critical
[1] 7.814728

$p.value
[1] 0.01348769
```

p.value가 0.05보다 작으므로 귀무가설 기각 분산이 동질하지 않음

(3) Select LDA and QDA according to the results of (1) and (2).

다변량 정규성 만족, 공분산행렬 동질성 성립하지 않으므로 QDA 실시

```
> QDA=qda(class~x1+x2, data=salmon, prior=c(1,1)/2)
> QDA
Call:
qda(class ~ x1 + x2, data = salmon, prior = c(1, 1)/2)

Prior probabilities of groups:
  1   2
0.5 0.5

Group means:
       x1       x2
1   98.38 429.66
2 137.46 366.62
> qcluster=predict(QDA, salmon)$class # predict(모델, 예측할 데이터)/ #qda로 예측한 결과 그룹 저장
> qct=table(class, qcluster) #실제 데이터의 타겟변수와 예측한 데이터를 table형태로 비교
> qct
      qcluster
class   1   2
    1 45   5
    2   2 48
> # Total percent correct
> mean(class==qcluster)
[1] 0.93
```

93%로 높은 예측률 보임

훈련할 데이터에 전체데이터를 넣었기에 즉 test데이터가 없어서 높은 예측률을 보인 것 일수도 있다.

(4) Divide this data into two clusters of gender 1 or 2 and apply discriminant analysis.

```
> unique(salmon[,2]) #타겟값 보기/unique 중복없이 유일한 값만 추출
[1] 2 1
> # MVN tests based on the Skewness and Kurtosis Statistics
> salmon_1=salmon[salmon$sex==1,3:4]
> salmon_2=salmon[salmon$sex==2,3:4] # 각각 종류별로 데이터 분리
> salmon_1
     x1  x2
2   131 355
3   105 469
5    99 402
7    94 440
10   99 403
11  114 428
13  123 372
16  104 407
20  119 474
21  114 396
27  109 397
29   82 431
31  105 388
32  121 403
33   85 451
34   83 453
35   53 427
36   95 411
37   76 442
38   95 426
40   70 397
43   74 451
45   80 398
46   95 433
48   99 481
50   87 480
51  129 420
52  148 371
53  179 407
57  156 419
59  140 362
62  108 330
63  135 355
65  152 301
66  153 397
67  152 301
70  148 383
72  145 337
73  123 364
79  118 379
81  153 403
83  154 390
84  155 349
87  128 400
88  144 403
91  133 375
92  128 383
94  144 373
98  125 346
99  153 352
100 108 339
```

```
##########################################################
> salmon_2
    x1  x2
1  108 368
4   86 506
6   87 423
8  117 489
9   79 432
12 123 372
14 109 420
15 112 394
17 111 422
18 126 423
19 105 434
22 100 470
23  84 399
24 102 429
25 101 469
26  85 444
28 106 442
30 118 381
39  87 402
41  84 511
42  91 469
44 101 474
47  92 404
49  94 491
54 152 381
55 166 377
56 124 389
58 131 345
60 144 345
61 149 393
64 170 386
68 136 438
69 122 306
71  90 385
74 145 376
75 115 354
76 134 383
77 117 355
78 126 345
80 120 369
82 150 354
85 109 325
86 117 344
89 163 370
90 145 355
93 123 349
95 140 388
96 150 339
97 124 341
> result_salmon_1 = mvn(salmon_1)
> result_salmon_2 = mvn(salmon_2) # 다변량 정규성 검정
```

```
> result_salmon_1
$multivariateNormality
             Test            Statistic             p value Result
1 Mardia Skewness    1.12480142375522 0.890317484538312     YES
2 Mardia Kurtosis -0.788312168481732 0.430514131680041     YES
3             MVN                <NA>               <NA>     YES

$univariateNormality
          Test   Variable Statistic    p value Normality
1 Shapiro-Wilk     x1          0.9755     0.3694     YES
2 Shapiro-Wilk     x2          0.9849     0.7575     YES

$Descriptives
    n       Mean   Std.Dev Median Min Max   25th   75th         Skew    Kurtosis
x1 51 118.0784 28.15375     119   53 179   97.0 144.0 -0.10941704 -0.8379360
x2 51 397.1373 42.54551     398 301 481 371.5 426.5 -0.07609466 -0.3620709

> result_salmon_2
$multivariateNormality
             Test            Statistic             p value Result
1 Mardia Skewness   4.97385160669907 0.289991022331544     YES
2 Mardia Kurtosis -1.59401489134059 0.110932699618718     YES
3             MVN                <NA>               <NA>     YES

$univariateNormality
          Test   Variable Statistic    p value Normality
1 Shapiro-Wilk     x1          0.9649     0.1510     YES
2 Shapiro-Wilk     x2          0.9574     0.0737     YES

$Descriptives
    n       Mean   Std.Dev Median Min Max 25th 75th       Skew    Kurtosis
x1 49 117.7551 23.84632     117   79 170   101   134 0.3198573 -0.8126463
x2 49 399.1837 50.22146     388 306 511   355   432 0.4894025 -0.6500867
```

mvn test결과 두 데이터 모두 YES이므로 다변량 정규성 만족한다.

```
> group=as.factor(sex) #섹터 별로 저장
> group
  [1] 2 1 1 2 1 2 1 2 2 1 1 2 1 2 2 1 2 2 2 1 1 2 2 2 2 2 1 2 1 2 1 1 1 1 1 1 1 2 1 2 2 1 2 1 1
2 1 2
 [50] 1 1 1 1 2 2 2 1 2 1 2 2 1 1 2 1 1 1 2 2 1 2 1 1 2 2 2 2 2 1 2 1 2 1 1 2 2 1 1 2 2 1 1 2 1 2
2 2 1
 [99] 1 1
Levels: 1 2

앞에서 사용한 cov.Mtest 사용
> ina=as.numeric(as.factor(salmon[, 2]))
> x=salmon[, 3:4]
> cov.Mtest(x, ina)
$M.test
[1] 3.728298

$degrees
[1] 3

$critical
[1] 7.814728

$p.value
[1] 0.2923372

p.value가 0.05보다 크므로 귀무가설 채택 분산이 동질함

다변량 정규성 만족, 공분산행렬 동질성 성립하므로 LDA 실시 #
> LDA=lda(sex~x1+x2, data=salmon, prior=c(1,1)/2)
> LDA
Call:
lda(sex ~ x1 + x2, data = salmon, prior = c(1, 1)/2)

Prior probabilities of groups:
  1   2
0.5 0.5

Group means:
         x1       x2
1 118.0784 397.1373
2 117.7551 399.1837
```

```
Coefficients of linear discriminants:
           LD1
x1 0.01337064
x2 0.02463053
> lcluster=predict(LDA, salmon)$class # predict(모델, 예측할 데이터)/ #qda로 예측한 결과 그룹 저장
> lct=table(sex, lcluster) #실제 데이터의 타겟변수와 예측한 데이터를 table형태로 비교
> lct
    lcluster
sex  1  2
  1 26 25
  2 26 23
> # Total percent correct
> mean(sex==lcluster)
[1] 0.49
```

49%로 낮은 예측률 보인다.