데이터과학을 위한 R프로그래밍

9주차. k-인접기법과 판별분석

이혜선 교수
포항공과대학교 산업경영공학과

목차

9주차. k-인접기법과 판별분석

9주차

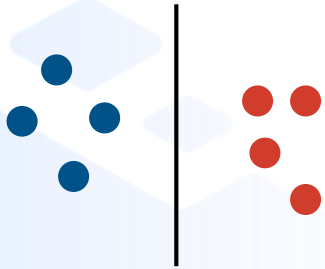
3차시

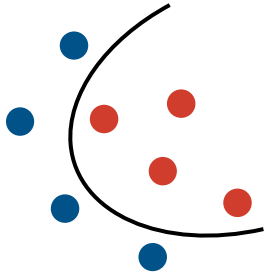
판별분석 II

이차판별분석

선형판별분석 vs. 이차판별분석

LDA	QDA
두범주를 분류하는 선형판별함수	판별함수가 변수들에 대한 이차함수로 표현
범주간 동일한 분산-공분산행렬을 가정	공분산 행렬이 범주별로 다른 경우





POSTECH

이차판별분석(QDA)

모집단 등분산 검정

- 분산-공분산 행렬이 범주별로 다른 경우, 이차판별분석(QDA)을 실시

Box's M-test

귀무가설 : 모집단의 분산-공분산 행렬이 동일
대립가설 : 모집단의 분산-공분산 행렬이 동일 X

등분산검정을 위한 패키지 : biotools

```
install.packages("biotools")
library(biotools)
boxM(iris[1:4], iris$Species)
```

```
> boxM(iris[1:4], iris$Species)

        Box's M-test for Homogeneity of Covariance Matrices

data:  iris[1:4]
Chi-Sq (approx.) = 140.94, df = 20, p-value < 2.2e-16
```

p-value~0

귀무가설(등분산 가정)이 기각 → QDA 실시!

POSTECH

이차판별분석(QDA)

☑ QDA 함수 : qda(종속변수~독립변수 , data=학습 데이터 이름, prior=사전 확률)

```
# Quadratic Discriminant Analysis (QDA)
iris.qda <- qda(Species ~ ., data=train, prior=c(1/3,1/3,1/3))
iris.qda
```

```
> iris.qda
Call:
qda(Species ~ ., data = train, prior = c(1/3, 1/3, 1/3))

Prior probabilities of groups:
    setosa versicolor  virginica
 0.3333333  0.3333333  0.3333333

Group means:
           Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa         5.051613    3.461290     1.480645   0.2387097
versicolor     5.935484    2.745161     4.267742   1.3129032
virginica      6.634211    2.965789     5.597368   2.0289474
```

경우에 따라 다르게 줄 수 있음

독립변수에 대한 그룹별 평균값

POSTECH

이차판별분석(QDA)

- 검증 데이터에 QDA 결과를 적용하여 범주 추정

```
# predict test data set n=50
testpredq <- predict(iris.qda, test)
```

```
> testpredq
$class
 [1] setosa     setosa     setosa     setosa     setosa
 [6] setosa     setosa     setosa     setosa     setosa
[11] setosa     setosa     setosa     setosa     setosa
[16] setosa     setosa     setosa     setosa     versicolor
[21] versicolor versicolor versicolor versicolor virginica
[26] versicolor versicolor versicolor versicolor versicolor
[31] versicolor versicolor versicolor versicolor versicolor
[36] versicolor versicolor versicolor virginica  virginica
[41] virginica  virginica  virginica  virginica  virginica
[46] virginica  virginica  virginica  virginica  virginica
Levels: setosa versicolor virginica

$posterior
        setosa   versicolor    virginica
2   1.000000e+00 2.889417e-18 6.455011e-33
8   1.000000e+00 2.178894e-22 4.689120e-36
14  1.000000e+00 6.827538e-18 7.188376e-31
16  1.000000e+00 1.663584e-32 1.131649e-48
19  1.000000e+00 2.144558e-24 4.509102e-40
```

추정 범주

세 개 범주의 사후 확률(posterior probability)을 구한 후 max값의 범주로 할당

판별분석 II - 이차판별분석

POSTECH

	class	posterior.setosa	posterior.versicolo	posterior.virginica
2	setosa	1.0000	0.0000	0.0000
3	setosa	1.0000	0.0000	0.0000
4	setosa	1.0000	0.0000	0.0000
5	setosa	1.0000	0.0000	0.0000
6	setosa	1.0000	0.0000	0.0000
7	setosa	1.0000	0.0000	0.0000
8	setosa	1.0000	0.0000	0.0000
9	setosa	1.0000	0.0000	0.0000
10	setosa	1.0000	0.0000	0.0000
11	setosa	1.0000	0.0000	0.0000
12	setosa	1.0000	0.0000	0.0000
13	setosa	1.0000	0.0000	0.0000
14	setosa	1.0000	0.0000	0.0000
15	setosa	1.0000	0.0000	0.0000
16	setosa	1.0000	0.0000	0.0000
17	setosa	1.0000	0.0000	0.0000
18	setosa	1.0000	0.0000	0.0000
19	setosa	1.0000	0.0000	0.0000
20	setosa	1.0000	0.0000	0.0000
21	versicolor	0.0000	0.9983	0.0017
22	versicolor	0.0000	0.9978	0.0022
23	versicolor	0.0000	0.9970	0.0030
24	versicolor	0.0000	0.9998	0.0002
25	versicolor	0.0000	0.9977	0.0023
26	virginica	0.0000	0.2947	0.7053
27	versicolor	0.0000	1.0000	0.0000
28	versicolor	0.0000	0.8348	0.1652
29	versicolor	0.0000	0.9908	0.0092
30	versicolor	0.0000	1.0000	0.0000
31	versicolor	0.0000	1.0000	0.0000
32	versicolor	0.0000	0.9831	0.0169
33	versicolor	0.0000	0.9967	0.0033
34	versicolor	0.0000	1.0000	0.0000
35	versicolor	0.0000	0.9986	0.0014
36	versicolor	0.0000	0.9997	0.0003
37	versicolor	0.0000	0.9995	0.0005
38	versicolor	0.0000	1.0000	0.0000
39	versicolor	0.0000	0.9997	0.0003

class	posterior.setosa	posterior.versicolor	posterior.virginica
virginica	0.0000	0.0000	1.0000
virginica	0.0000	0.0032	0.9968
virginica	0.0000	0.0000	1.0000
virginica	0.0000	0.0000	1.0000
virginica	0.0000	0.1410	0.8590
virginica	0.0000	0.0057	0.9943
virginica	0.0000	0.0393	0.9607
virginica	0.0000	0.0000	1.0000
virginica	0.0000	0.0000	1.0000
virginica	0.0000	0.0434	0.9566
virginica	0.0000	0.0000	1.0000
virginica	0.0000	0.0565	0.9435

실제는 versicolor인데 → virginica로 분류

POSTECH

이차판별분석(QDA)

정확도 산정 : 오분류율 (검증데이터)

```
# accuracy of QDA
confusionMatrix(testpredq$class,testLabels)
```

```
> confusionMatrix(testpred$class,testLabels)
Confusion Matrix and Statistics

            Reference
Prediction   setosa versicolor virginica
  setosa         19          0         0
  versicolor      0         18         0
  virginica       0          1        12

Overall Statistics

               Accuracy : 0.98
                 95% CI : (0.8935, 0.9995)
    No Information Rate : 0.38
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9695

 Mcnemar's Test P-Value : NA
```

- versicolor를 virginica로 잘못 예측
- 정확도 : 49/50 → 98%
- 오분류율 : 1/50 → 2%