데이터과학을 위한 R프로그래밍

9주차. k-인접기법과 판별분석

이혜선 교수
포항공과대학교 산업경영공학과

목차

9주차. k-인접기법과 판별분석

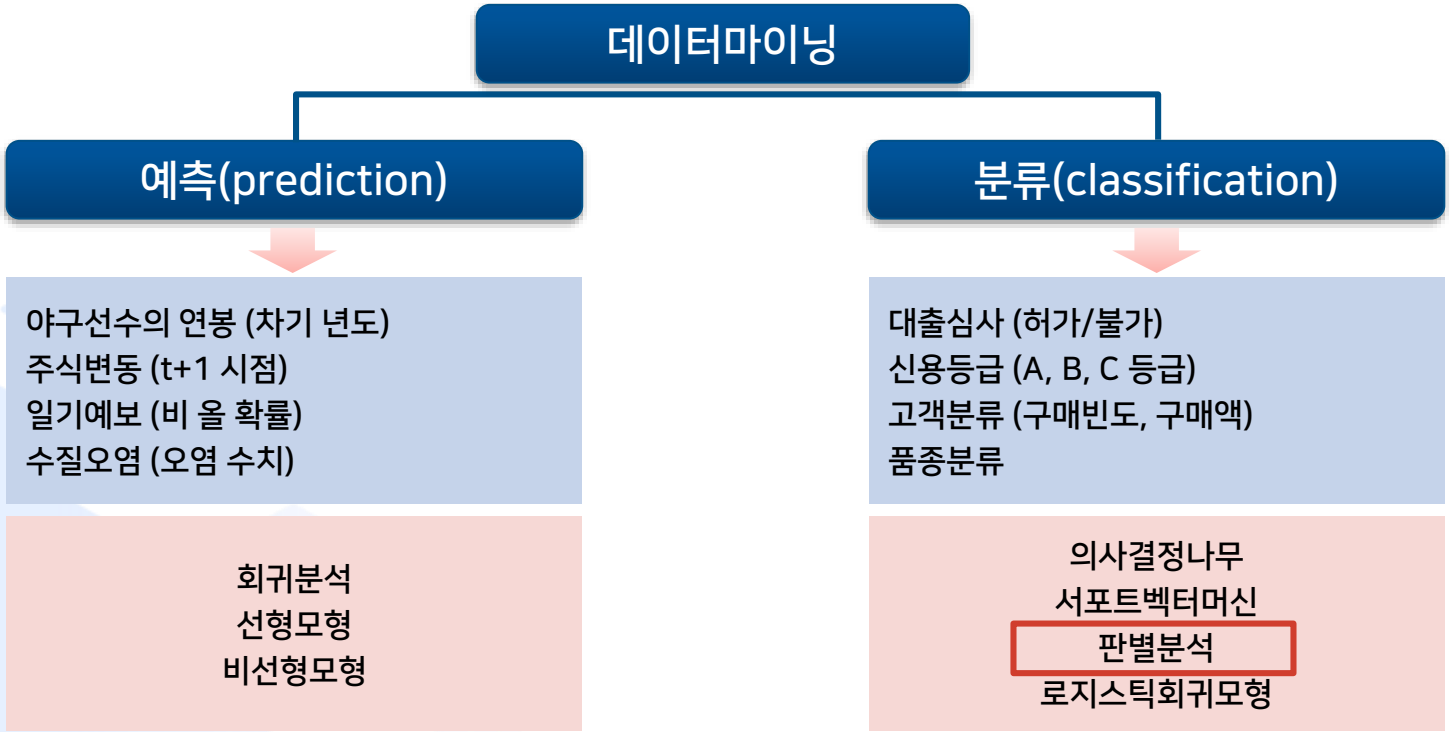
9주차

2차시

판별분석 I
선형판별분석

POSTECH

판별분석

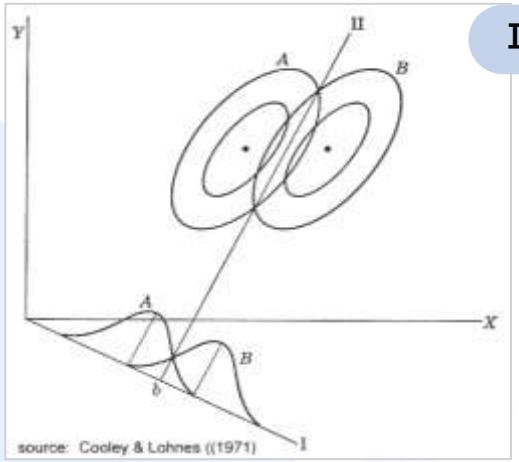


POSTECH

판별분석

판별분석(Discriminant Analysis)

- 객체를 몇 개의 범주로 분류
- 범주들을 가장 잘 구분하는 변수 파악 및 범주간 차이를 가장 잘 표현하는 함수 도출



피셔(Fisher) 방법

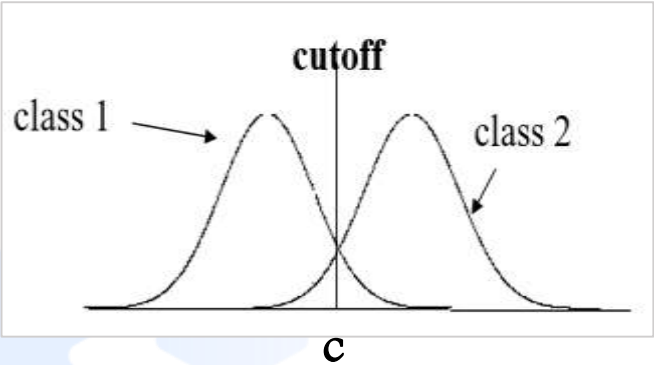
의사 결정 이론

- 선형판별분석 (LDA; Linear DA)
  - 정규분포의 분산 - 공분산 행렬이 범주에 관계없이 동일한 경우
- 이차판별분석 (QDA; Quadratic DA)
  - 정규분포의 분산 - 공분산 행렬이 범주별로 다른 경우

POSTECH

판별분석

의사결정이론



범주 1, 2에 대한 확률밀도함수를 $f_1(x), f_2(x)$

범주 1, 2에 속할 사전확률을 π_1, π_2

오분류 총 확률 = $P\{\text{범주 1로 오분류}\} + P\{\text{범주 2로 오분류}\}$

$$= \pi_2 \int_{-\infty}^{c} f_2(x)dx + \pi_1 \int_{c}^{\infty} f_1(x)dx$$

$\pi_2 \int_{-\infty}^{c} f_2(x)dx + \pi_1 \int_{c}^{\infty} f_1(x)dx$ 를 최소로 하는 c를 c^*라 하면 다음 식이 성립

$$\pi_2 f_2(c^*) = \pi_1 f_1(c^*) \Leftrightarrow \frac{\pi_2}{\pi_1} = \frac{f_1(c^*)}{f_2(c^*)}$$

POSTECH

예제 데이터

Iris 데이터 train/test 분할

```
# read csv file# read csv file
iris<-read.csv("iris.csv", stringsAsFactors = TRUE)
attach(iris)
```

데이터 불러들이기

```
# iris data n=150
set.seed(1000)
n <- nrow(iris)
# split : train set 100, test set 50
tr.idx <- sample.int(n, size = round(2/3* n))
```

데이터분할(학습데이터 80%, 검증데이터 20%)

```
# attributes in training and test
iris.train<-iris[tr.idx,-5]
iris.test<-iris[-tr.idx,-5]
```

iris.train(독립변수4개를 포함한 100개의 데이터)
iris.test(독립변수4개를 포함한 50개의 데이터)

trainLabels(학습데이터의 타겟변수)
testLabels(검증데이터의 타겟변수)

POSTECH

선형판별분석(LDA)

- 패키지(MASS) 설치
- LDA 함수 : lda(종속변수 ~ 독립변수 , data=학습 데이터 이름, prior= 사전 확률)

```
# install the MASS package for LDA
install.packages("MASS")
library(MASS)

# Linear Discriminant Analysis (LDA) with training data n=100
iris.lda <- lda(Species ~ ., data=train, prior=c(1/3,1/3,1/3))
iris.lda
```

동일한 의미

```
Species ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width
```

사전 확률(prior probability)
: 원인 A가 발생할 확률인 P(A)와 같이 결과가 나타나기 전에 결정되어 있는 확률

POSTECH

선형판별분석(LDA)

학습 데이터 LDA 결과

```
> iris.lda
Call:
lda(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

Coefficients of linear discriminants:
                    LD1        LD2
Sepal.Length  0.8907558 -0.1072740
Sepal.Width   1.7077575 -2.2338358
Petal.Length -2.1513701  0.7355423
Petal.Width  -2.9073216 -2.3919728

Proportion of trace:
   LD1    LD2
0.9905 0.0095
```

- 첫 번째 범주 판별 함수

LD1 = 0.89 Sepal.Length + 1.71 Sepal.Width
- 2.15 Petal.Length - 2.91 Petal.Width

- 두 번째 범주 판별 함수

LD2 = - 0.11 Sepal.Length - 2.23 Sepal.Width
+ 0.74 Petal.Length - 2.39 Petal.Width

LD1이 between-group variance의 99%를 설명
LD2가 between-group variance의 1%를 설명

POSTECH

선형판별분석(LDA)

☑ 검증 데이터에 LDA 결과를 적용하여 범주 추정

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

```
> testpred <- predict(iris.lda, test)
> testpred
$class
 [1] setosa     setosa     setosa     setosa     setosa     setosa
 [7] setosa     setosa     setosa     setosa     setosa     setosa
[13] setosa     setosa     setosa     setosa     setosa     setosa
[19] setosa     versicolor versicolor versicolor versicolor versicolor
[25] virginica  versicolor versicolor versicolor versicolor versicolor
[31] versicolor versicolor versicolor versicolor versicolor versicolor
[37] versicolor versicolor virginica  virginica  virginica  virginica
[43] virginica  virginica  virginica  virginica  virginica  virginica
[49] virginica  virginica
Levels: setosa versicolor virginica

$posterior
       setosa   versicolor    virginica
2   1.000000e+00 1.173765e-17 9.991990e-37
8   1.000000e+00 2.526721e-20 7.057281e-40
14  1.000000e+00 2.742945e-19 4.300404e-39
16  1.000000e+00 8.883431e-29 9.438773e-50
19  1.000000e+00 3.299737e-23 3.645303e-43
```

추정 범주

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

POSTECH

선형판별분석(LDA)

산정된 사후확률결과

	class	posterior.s	posterior.v	posterior.v	x.LD1	x.LD2
2	setosa	1	1.17E-17	9.99E-37	7.047432	0.857184
3	setosa	1	2.53E-20	7.06E-40	7.604473	0.026477
4	setosa	1	2.74E-19	4.30E-39	7.449121	0.940083
5	setosa	1	8.88E-29	9.44E-50	9.354295	-2.76085
6	setosa	1	3.30E-23	3.65E-43	8.190099	-1.03424
7	setosa	1	8.23E-23	1.53E-42	8.08592	-1.11698
8	setosa	1	4.68E-25	5.44E-46	8.665407	-0.74515
9	setosa	1	9.84E-15	1.03E-31	6.220302	-0.33135
10	setosa	1	7.04E-16	7.84E-34	6.565774	0.342148
11	setosa	1	5.37E-22	3.60E-42	7.997761	-0.06853
12	setosa	1	1.14E-16	4.00E-35	6.780558	0.57898
13	setosa	1	1.30E-19	1.74E-38	7.379311	-0.49483
14	setosa	1	1.49E-29	1.70E-51	9.631194	-1.88778
15	setosa	1	1.04E-16	2.25E-35	6.817191	0.837267
16	setosa	1	7.14E-22	1.30E-41	7.914791	-0.58321
17	setosa	1	2.38E-10	2.30E-27	5.420104	2.151027
18	setosa	1	2.56E-18	2.58E-37	7.158742	0.3905
19	setosa	1	3.53E-16	1.81E-34	6.667624	0.628714
20	setosa	1	3.83E-18	4.18E-37	7.121756	0.442599
21	versicolor	3.29E-22	0.996817	0.003183	-2.30959	-0.11591
22	versicolor	2.35E-23	0.999526	0.000474	-2.40516	1.637745
23	versicolor	2.40E-23	0.995962	0.004038	-2.53281	0.376484
24	versicolor	5.87E-20	0.999893	0.000107	-1.6915	0.620767
25	versicolor	2.77E-20	0.999297	0.000703	-1.86517	-0.30014
26	virginica	4.62E-28	0.296337	0.703663	-3.68663	-1.02317
27	versicolor	4.35E-17	0.999991	9.26E-06	-1.01683	0.456462
28	versicolor	5.02E-27	0.735249	0.264751	-3.45512	-0.27592
29	versicolor	1.72E-23	0.993012	0.006988	-2.59228	0.133183
30	versicolor	2.93E-12	1	2.24E-08	0.233199	1.29596

- 실제는 versicolor인데 → virginica로 분류됨

POSTECH

선형판별분석(LDA)

정확도 산정 : 오분류율(검증데이터)

```
# accuracy of LDA
library(caret)
confusionMatrix(testpred$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
```

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