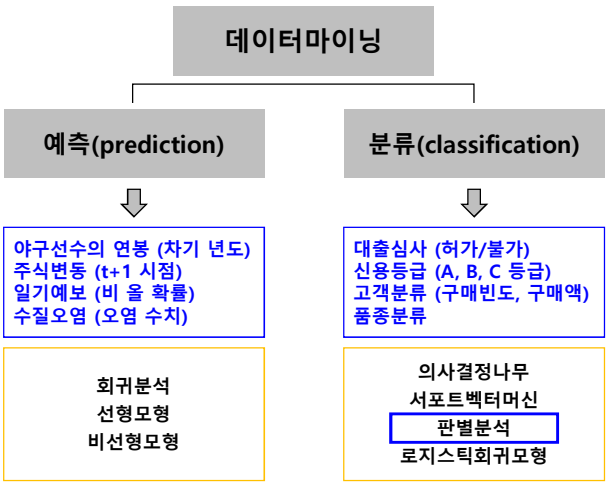
10.3 판별분석(Discriminant Analysis) I

- 선형판별분석 -

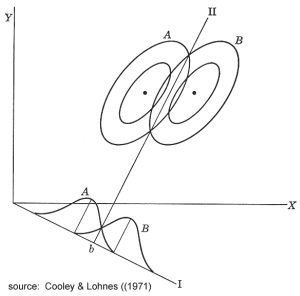
1. 판별분석



1. 판별분석

• 판별분석(Discriminant Analysis)

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

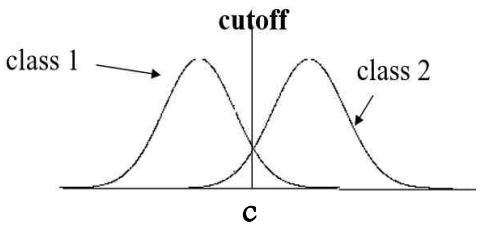


source: Cooley & Lohnes ((1971)

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

1. 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 = $P\{$범주 1로 오분류$\}$ + $P\{$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

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

2. 예제 데이터

- Iris 데이터 train/test 분할

```
# read csv file
iris<-read.csv("iris.csv")
# head(iris)
# str(iris)
attach(iris)
```

데이터 불러들이기

```
# training/ test data : n=150
set.seed(1000)
N=nrow(iris)
tr.idx=sample(1:N, size=N*2/3, replace=FALSE)
```

데이터분할 (학습데이터 2/3, 검증데이터 1/3)

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

train<-iris[tr.idx,]
test<-iris[-tr.idx,]
```

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

trainLabels (학습데이터의 타겟변수)
testLabels (검증데이터의 타겟변수)

3. 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

3. 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

3. 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

3. 선형판별분석(LDA)

• 산정된 사후확률결과

	ID	setosa	versicolor	virginica	True_class
2	2	1	1.17E-17	9.99E-37	setosa
3	8	1	2.53E-20	7.06E-40	setosa
4	14	1	2.74E-19	4.30E-39	setosa
5	16	1	8.88E-29	9.44E-50	setosa
6	19	1	3.30E-23	3.65E-43	setosa
7	20	1	8.23E-23	1.53E-42	setosa
8	23	1	4.68E-25	5.44E-46	setosa
9	24	1	9.84E-15	1.03E-31	setosa
10	25	1	7.04E-16	7.84E-34	setosa
11	29	1	5.37E-22	3.60E-42	setosa
12	30	1	1.14E-16	4.00E-35	setosa
13	32	1	1.30E-19	1.74E-38	setosa
14	34	1	1.49E-29	1.70E-51	setosa
15	39	1	1.04E-16	2.25E-35	setosa
16	41	1	7.14E-22	1.30E-41	setosa
17	42	1	2.38E-10	2.30E-27	setosa
18	43	1	2.56E-18	2.58E-37	setosa
19	46	1	3.53E-16	1.81E-34	setosa
20	48	1	3.83E-18	4.18E-37	setosa
21	53	3.29E-22	0.99682	0.00318	versicolor
22	54	2.35E-23	0.99953	0.00047	versicolor
23	55	2.40E-23	0.99596	0.00404	versicolor
24	59	5.87E-20	0.99989	0.00011	versicolor
25	62	2.77E-20	0.9993	0.0007	versicolor
26	71	4.62E-28	0.29634	0.70366	versicolor

ID	setosa	versicolor	virginica	True_class
71	4.62E-28	0.29634	0.70366	versicolor
72	4.35E-17	0.99999	9.26E-06	versicolor
78	5.02E-27	0.73525	0.26475	versicolor
79	1.72E-23	0.99301	0.00699	versicolor
80	2.93E-12	1	2.24E-08	versicolor
83	5.49E-17	1	4.09E-06	versicolor
91	1.55E-23	0.99929	0.00071	versicolor
92	3.95E-22	0.99831	0.00169	versicolor
93	1.03E-18	0.99999	1.27E-05	versicolor
95	1.30E-21	0.99967	0.00033	versicolor
96	1.01E-17	0.99998	1.74E-05	versicolor
97	1.25E-19	0.99989	0.00011	versicolor
99	8.59E-12	1	2.48E-08	versicolor
100	1.65E-19	0.99992	7.57E-05	versicolor
105	3.48E-46	2.35E-06	1	virginica
111	5.09E-32	0.01773	0.98227	virginica
113	3.93E-39	0.00026	0.99974	virginica
116	3.02E-40	3.60E-05	0.99996	virginica
120	1.80E-34	0.185	0.815	virginica
126	3.74E-36	0.00403	0.99597	virginica
127	2.52E-30	0.20058	0.79942	virginica
129	2.09E-44	1.51E-05	0.99998	virginica
137	2.57E-44	1.35E-06	1	virginica
138	4.90E-35	0.00794	0.99206	virginica
141	3.81E-45	1.65E-06	1	virginica
150	1.30E-33	0.02009	0.97991	virginica

- test set의 ID=71번에 오분류
- 실제는 versicolor인데-> virginica로 분류됨

3. 선형판별분석(LDA)

• 정확도 산정 : 오분류율 (검증데이터)

```
# accuracy of LDA
CrossTable(x=testLabels,y=testpred$class, prop.chisq=FALSE)
```

```
 Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  50

                 | testpred$class
      testLabels |     setosa | versicolor |  virginica | Row Total |
-----------------|------------|------------|------------|-----------|
          setosa |         19 |          0 |          0 |        19 |
                 |      1.000 |      0.000 |      0.000 |     0.380 |
                 |      1.000 |      0.000 |      0.000 |           |
                 |      0.380 |      0.000 |      0.000 |           |
-----------------|------------|------------|------------|-----------|
      versicolor |          0 |         18 |          1 |        19 |
                 |      0.000 |      0.947 |      0.053 |     0.380 |
                 |      0.000 |      1.000 |      0.077 |           |
                 |      0.000 |      0.360 |      0.020 |           |
-----------------|------------|------------|------------|-----------|
       virginica |          0 |          0 |         12 |        12 |
                 |      0.000 |      0.000 |      1.000 |     0.240 |
                 |      0.000 |      0.000 |      0.923 |           |
                 |      0.000 |      0.000 |      0.240 |           |
-----------------|------------|------------|------------|-----------|
    Column Total |         19 |         18 |         13 |        50 |
                 |      0.380 |      0.360 |      0.260 |           |
-----------------|------------|------------|------------|-----------|
```

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