10강. 판별분석(1)

- 판별분석 개요
- Fisher 판별분석 모형
- 선형판별분석
- 판별함수 모형 평가
- R 선형판별분석

1. 판별분석의 개요

◆ 판별분석

측정된 변수들을 이용하여 각 개체들이 2개 이상의 그룹 중 어느 그룹에 속하는 지를 판별하는 분석방법을 말함

◆ 판별분석이 이용되기 위한 조건

각 개체는 여러 개의 그룹 중에서 어느 그룹에 속해있는 지 알려져 있어야 함

◆ 판별분석 과정

소속그룹이 이미 알려진 케이스에 대하여 변수들을 측정하고
이들 변수들을 이용하여 각 그룹들을 가장 잘 구분할 수 있는 판별식을 만들어
분별하는 과정

1. 판별분석의 개요

◆ 판별분석이 이용되는 예

예1) 어느 발굴현장에서 새로운 유물이 발견된 경우에 이 유물이 가능한 두 종족 중에 어느 종족의 유물인지를 판별하고자 하는 경우

- 두 종족의 유물들의 특징, 예를 들어 유물들의 크기, 모양 등을 이용하여 어느 종족의 유물인지를 판별할 수 있는 판별식을 만든 다음 이를 이용하여 새로운 유물이 어느 종족의 유산인지를 판별

예2) 은행에서 기업에 대출을 해주는 경우, 은행에서 돈을 대출해주기 전에 대상기업의 도산가능성에 대한 판단을 하는 경우

- 과거 도산한 기업들과 도산하지 않은 기업들에 대한 각종 자료(기업들의 자산, 부채, 매출액, 당기순이익 등)를 관측한 후에 기업의 도산여부를 판별할 수 있는 판별함수를 만든 다음 이를 이용하여 대상 기업들을 판별

2. Fisher 판별분석 모형

◆ 그룹의 수가 2개인 경우의 판별함수

- 두 그룹을 G1, G2라고 하고, 각각의 그룹에서 p개의 설명변수가 관측된 경우

< 판별분석 자료 형태 >

집단	1				2			
설명변수	X_1	X_2	$\cdots$	X_p	X_1	X_2	$\cdots$	X_p
관측값	X_{11}	X_{12}	$\cdots$	X_{1p}	X_{11}	X_{12}	$\cdots$	X_{1p}
	X_{21}	X_{22}	$\cdots$	X_{2p}	X_{21}	X_{22}	$\cdots$	X_{2p}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$X_{n_1 1}$	$X_{n_1 2}$	$\cdots$	$X_{n_1 p}$	$X_{n_2 1}$	$X_{n_2 2}$	$\cdots$	$X_{n_2 p}$

- 집단의 수가 2개이고 p개의 설명변수가 있는 경우에 판별함수는 하나
- 판별함수는 다음과 같이 표현됨

$$Y = b_1 X_1 + b_2 X_2 + \cdots + b_n X_n = \boldsymbol{b}' X$$

$$\boldsymbol{b}' = (b_1, b_2, \cdots, b_p), \quad \boldsymbol{X} = (X_1, X_2, \cdots, X_p)$$

⇒ Fisher의 판별함수(Fisher Discriminant Function)라고 함

2. Fisher 판별분석 모형

◆ 판별함수 계수벡터 b를 구하는 방법

- 판별함수 작성에 있어서 기본 개념 : 사전에 분류된 그룹들을 판별오류가 최소가 되도록 선형함수를 작성. 즉, 그룹 내 분산에 비해 상대적으로 그룹간 분산이 최대가 되도록 판별함수를 작성

-벡터 b 는 $\lambda = \frac{\text{그룹간분산}}{\text{그룹내분산}}$ 이 최대가 되도록 하는, 즉 두 집단 사이의 평균사이의 거리가 최대가 되도록 하는 b를 구함.

- 이를 만족하는 계수벡터 $b = \Sigma^{-1}(\mu_2 - \mu_1)$ 이며,
  판별함수는 $Y = b'X = (\mu_2 - \mu_1)'\Sigma^{-1}X$

이를 Fisher의 선형판별함수(Fisher's Linear Discriminant Function)라고 함.

- 표본이 관찰된 경우 $Y = b'X = (\overline{X_2} - \overline{X_1})'S^{-1}X$

2. Fisher 판별분석 모형

◆ 두 집단에서 판별함수를 유도하는 과정

< 두 집단에서 관측된 세 변수 자료>

제 1 집단(G_1)			제 2 집단(G_2)		
X_1	X_2	X_3	X_1	X_2	X_3
98	81	38	93	74	37
103	84	38	94	78	35
103	86	42	96	80	35
105	86	42	101	84	39
109	88	44	102	85	38
123	92	50	104	83	39
133	99	51			

평균벡터 :

그룹 G_1 : $\overline{X}_1 = \begin{bmatrix} 110.571 \\ 88.000 \\ 43.571 \end{bmatrix}$

그룹 G_2 : $\overline{X}_2 = \begin{bmatrix} 98.333 \\ 80.667 \\ 37.167 \end{bmatrix}$

분산-공분산 행렬 :

그룹 G_1 : $S_1 = \begin{bmatrix} 160.6190 & 73.5000 & 63.1191 \\ & 35.0000 & 29.1667 \\ & & 27.2857 \end{bmatrix}$ 그룹 G_2 : $S_2 = \begin{bmatrix} 21.0667 & 17.3333 & 6.7333 \\ & 17.4667 & 4.4667 \\ & & 3.3667 \end{bmatrix}$

$$S = \frac{6S_1 + 5S_2}{11} = \begin{bmatrix} 97.1862 & 47.9697 & 37.4892 \\ & 27.0303 & 17.9394 \\ & & 16.4134 \end{bmatrix} \Rightarrow b = S^{-1}(\overline{X_2} - \overline{X_1}) = (0.545, -0.558, -1.024)'$$

선형판별함수 $Y = 0.545X_1 - 0.558X_2 - 1.024X_3$

2. Fisher 판별분석 모형

◆ R 계산

```
> data7 = read.csv("c:/data/mva/data7-1.csv")
> head(data7, 3)
  group  x1 x2 x3
1    g1  98 81 38
2    g1 103 84 38
3    g1 103 86 42
> data7_g1 = data7[data7$group=='g1', -1]
> data7_g2 = data7[data7$group=='g2', -1]
> g1_mean = sapply(data7_g1, mean)
> g2_mean = sapply(data7_g2, mean)
> g1_mean
      x1       x2       x3
110.57143  88.00000  43.57143
> g2_mean
      x1       x2       x3
98.33333 80.66667 37.16667
```

group	x1	x2	x3
g1	98	81	38
g1	103	84	38
g1	103	86	42
g1	105	86	42
g1	109	88	44
g1	123	92	50
g1	133	99	51
g2	93	74	37
g2	94	78	35
g2	96	80	35
g2	101	84	39
g2	102	85	38
g2	104	83	39

평균벡터 :

그룹 G_1 : $\overline{X}_1 = \begin{bmatrix} 110.571 \\ 88.000 \\ 43.571 \end{bmatrix}$

그룹 G_2 : $\overline{X}_2 = \begin{bmatrix} 98.333 \\ 80.667 \\ 37.167 \end{bmatrix}$

2. Fisher 판별분석 모형

```
> n1 = nrow(data7_g1)
> n2 = nrow(data7_g2)
> cov_g1 = cov(data7_g1)
> cov_g2 = cov(data7_g2)
> cov_g = ((n1-1)*cov_g1 + (n2-1)*cov_g2) / (n1+n2-2)
> cov_g
          x1       x2       x3
x1 97.18615 47.96970 37.48918
x2 47.96970 27.03030 17.93939
x3 37.48918 17.93939 16.41342

> b = solve(cov_g) %*% (g2_mean - g1_mean)
> round(b, 3)
     [,1]
x1  0.545
x2 -0.558
x3 -1.024
```

$$S = \frac{6S_1 + 5S_2}{11} = \begin{bmatrix} 97.1862 & 47.9697 & 37.4892 \\ & 27.0303 & 17.9394 \\ & & 16.4134 \end{bmatrix}$$

평균벡터 :

그룹 G_1 : $\overline{X}_1 = \begin{bmatrix} 110.571 \\ 88.000 \\ 43.571 \end{bmatrix}$

그룹 G_2 : $\overline{X}_2 = \begin{bmatrix} 98.333 \\ 80.667 \\ 37.167 \end{bmatrix}$

$$b = S^{-1}(\overline{X_2} - \overline{X_1}) = (0.545, -0.558, -1.024)'$$

$$Y = 0.545X_1 - 0.558X_2 - 1.024X_3$$

3. Fisher 판별함수를 이용한 분류

◆ 분류절차

(1) 판별함수를 이용하여 두 집단에 있어서의 판별함수값의 평균을 구함.

$$\overline{Y}_1 = (\overline{X}_2 - \overline{X}_1)' S^{-1} \overline{X}_1$$
$$\overline{Y}_2 = (\overline{X}_2 - \overline{X}_1)' S^{-1} \overline{X}_2$$

(2) 두 집단의 분류점으로서 두 중심 $\overline{Y}_1$와 $\overline{Y}_2$의 중앙위치를 구함.

$$Y_c = \frac{\overline{Y}_1 + \overline{Y}_2}{2} = \frac{1}{2}(\overline{X}_2 - \overline{X}_1)' S^{-1} (\overline{X}_1 + \overline{X}_2)$$

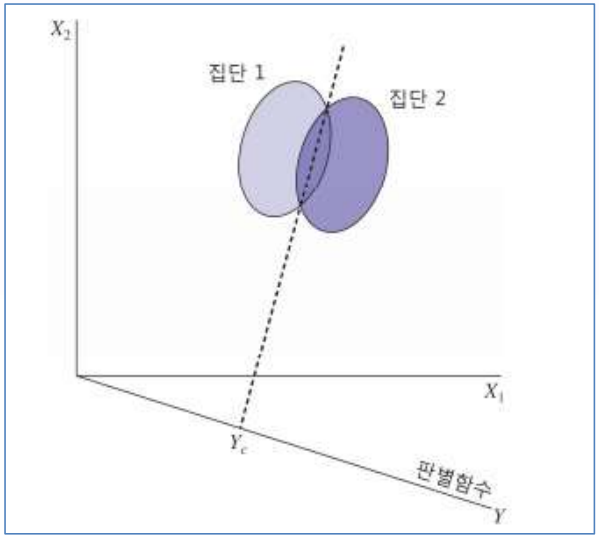
(3) 새로운 관측값의 판별 함수값을 구함.

$$Y_i = (\overline{X}_2 - \overline{X}_1)' S^{-1} X_i$$

(4) $\overline{Y}_1 < \overline{Y}_2$ 인 경우, i 번째 판별함수값 Y_i가

$Y_i \leq Y_c$ 이면 집단 1 로 분류,

$Y_i > Y_c$ 이면 집단 2 로 분류.



3. Fisher 판별함수를 이용한 분류

첫번째 케이스의 분류

```
> y1_mean = g1_mean %*% b
> y2_mean = g2_mean %*% b
> y1_mean
          [,1]
[1,] -33.51548
> y2_mean
          [,1]
[1,] -29.52853
> yc = (y1_mean + y2_mean)/2
> yc
          [,1]
[1,] -31.52201
> case1 = data7_g1[1,]
> case1 = as.matrix(case1)
> y1 = case1 %*% b
> y1
       [,1]
1 -30.74943
>
```

- y1 = -30.75
- yc = -31.52
- => yc < y1 이므로 그룹 2로 분류됨.

4. 선형판별분석 (LDA)

선형판별분석 lda(linear discriminant analysis) : Fisher 판별분석의 일반화된 방법

두 그룹 G_1, G_2

제1집단의 설명변수 : $X_{G1} = (X_1, X_2, \cdots, X_p)'$

제2집단의 설명변수 : $X_{G2} = (X_1, X_2, \cdots, X_p)'$

여기서 $p \times 1$ 변수벡터 X_{G1}과 X_{G2}는 각각 다변량 정규분포를 따른다고 가정,

$$X_{G1} \sim N(\boldsymbol{\mu_1}, \Sigma_1)$$
$$X_{G2} \sim N(\boldsymbol{\mu_2}, \Sigma_2)$$

베이지안 규칙(Bayesian rule)을 이용한 분류방법 :

만약, $P(G_2|X = \boldsymbol{x}) > P(G_1|X = \boldsymbol{x})$ 이면 그룹 G_2로 분류.

4. 선형판별분석 (LDA)

로그-우도비(log of likelihood ratios)를 이용한 베이지안 최적해 :

"어떤 기준 T에 대해서

$$(\boldsymbol{x}-\boldsymbol{\mu_1})^T\Sigma_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu_1})+\ln|\Sigma_1|-(\boldsymbol{x}-\boldsymbol{\mu_2})^T\Sigma_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu_2})-\ln|\Sigma_2| > T$$

이면 그룹 G_2 로 분류."

여기서 두 그룹의 분산이 같다고 가정하면, 즉 $\Sigma_1 = \Sigma_2 = \Sigma$ 이면,

분류규칙은

$$(\boldsymbol{x}-\boldsymbol{\mu_1})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu_1})-(\boldsymbol{x}-\boldsymbol{\mu_2})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu_2}) > T$$

이 되고, 이를 정리하면 다음과 같이 선형판별함수를 이용한 분류규칙이 됨.

"어떤 기준 c 에 대해서 $\boldsymbol{b}'\boldsymbol{x} > c$ 이면 그룹 G_2 로 분류."

5. LDA 를 이용한 분류

```
> library(MASS)
> data7 = read.csv("c:/data/mva/data7-1.csv")
> data7_lda = lda(group ~ ., data=data7)
> data7_lda
Call:
lda(group ~ ., data = data7)

Prior probabilities of groups:
       g1        g2
0.5384615 0.4615385

Group means:
          x1          x2          x3
g1 110.57143 88.00000 43.57143
g2  98.33333 80.66667 37.16667

Coefficients of linear discriminants:
          LD1
x1  0.2727114
x2 -0.2794634
x3 -0.5128692
```

group	x1	x2	x3
g1	98	81	38
g1	103	84	38
g1	103	86	42
g1	105	86	42
g1	109	88	44
g1	123	92	50
g1	133	99	51
g2	93	74	37
g2	94	78	35
g2	96	80	35
g2	101	84	39
g2	102	85	38
g2	104	83	39

분류점수 $D_i = 0.273\,(x_{i1} - \bar{x}_1) - 0.279\,(x_{i2} - \bar{x}_2) - 0.513\,(x_{i3} - \bar{x}_3)$

5. LDA 를 이용한 분류

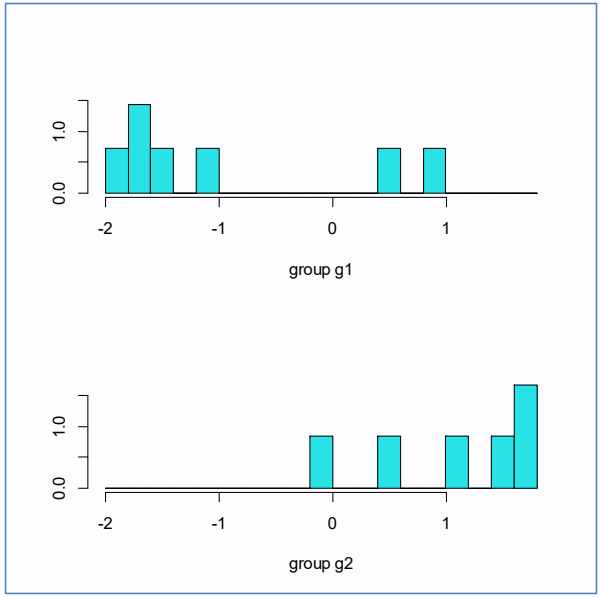
분류점수 $D_i = 0.273(x_{i1} - \bar{x}_1) - 0.279(x_{i2} - \bar{x}_2) - 0.513(x_{i3} - \bar{x}_3)$

```
> pred_lda = predict(data7_lda, newdata=data7)
> names(pred_lda)
[1] "class"     "posterior" "x"
> pred_lda$class
 [1] g2 g2 g1 g1 g1 g1 g1 g2 g2 g2 g1 g2 g2
Levels: g1 g2
> head(pred_lda$posterior)
         g1         g2
1 0.3501401 0.64985989
2 0.1588189 0.84118107
3 0.9719480 0.02805197
4 0.9210121 0.07898791
5 0.9689978 0.03100215
6 0.9851558 0.01484415
> head(pred_lda$x)
         LD1
1  0.4637161
2  0.9888827
3 -1.6215208
4 -1.0760981
5 -1.5699178
6 -1.9470273
```

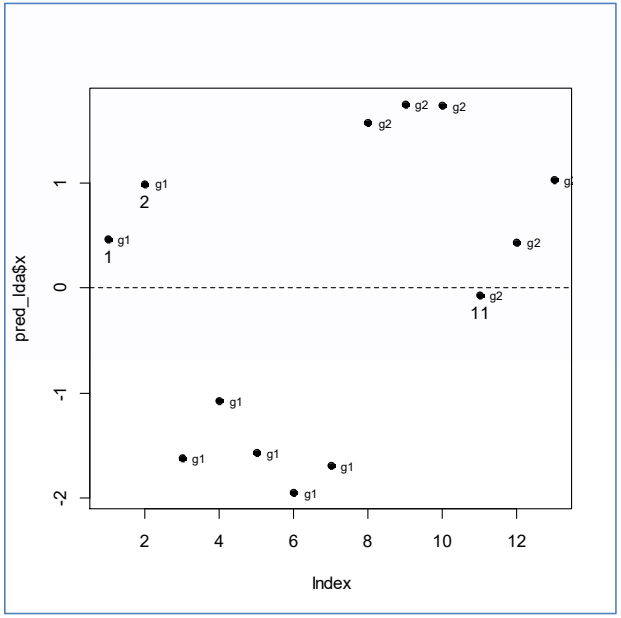
```
> # 참고 : 앞의 결과를 이용하여 분류점수 구하기
> prior = data7_lda$prior
> scaling = data7_lda$scaling
> scaling
          LD1
x1  0.2727114
x2 -0.2794634
x3 -0.5128692
> d_means = data7_lda$means
> d_means
         x1       x2       x3
g1 110.57143 88.00000 43.57143
g2  98.33333 80.66667 37.16667
>  means <- colSums(prior * d_means)
> means
      x1       x2       x3
104.92308  84.61538  40.61538

> x = data7[,-1]
> dscore=scale(x, center=means, scale=FALSE) %*% scaling
> head(dscore, 3)
          LD1
[1,]  0.4637161
[2,]  0.9888827
[3,] -1.6215208
```

5. LDA 를 이용한 분류

```
> library(klaR)
> ldahist(pred_lda$x, g=data7$group)
```




```
> library(klaR)
> ldahist(pred_lda$x, g=data7$group)
> dev.new()
> plot(pred_lda$x, pch=19)
> text(pred_lda$x, data7$group, cex=0.7, pos=4)
> abline(h=0, lty=2)
> identify(pred_lda$x)
[1]  1  2 11
```



6. 판별모형 평가

1) 람다(Wilk's lambda) 통계량

$$\lambda = \frac{|W|}{|T|}$$

W : within-groups sum of square matrix

T : total sum of square matrix

- 값이 0 에 가까울수록 좋고, 1에 가까울수록 나쁨을 의미

```
> x = data7[,-1]
> head(x, 3)
   x1 x2 x3
1  98 81 38
2 103 84 38
3 103 86 42
> data7_man <- manova(as.matrix(x) ~ pred_lda$class)
> wilks_test = summary(data7_man, test="Wilks")
> wilks_test
               Df   Wilks approx F num Df den Df    Pr(>F)
pred_lda$class  1 0.16683   14.983      3      9 0.0007604 ***
Residuals      11
---
```

6. 판별모형 평가

2) 분류표(confusion matrix)

```
> confm_lda = table(data7$group, pred_lda$class)
> confm_lda
     g1 g2
  g1  5  2
  g2  1  5
> prop.table(confm_lda,1)
           g1        g2
  g1 0.7142857 0.2857143
  g2 0.1666667 0.8333333
> error = 1 - sum(diag(prop.table(confm_lda)))
> error
[1] 0.2307692
```

- 그룹1 : 잘못 분류된 케이스는 총 7에서 2 케이스 (2/7 *100 = 28.6%)
- 그룹2 : 잘못 분류된 케이스는 총 6에서 1 케이스 (1/6*100 = 16.7 %)
- 전체적인 분류율 : 10/13*100 = 76.9 %

(오분류율 = 23.1 %)

7. 판별 변수 선택

◆ 변수선택 방법

① 앞으로부터의 선택 (forward discriminant analysis)

② 단계별 선택 (stepwise discriminant analysis)

Wilk's Lambda 통계량을 이용

Selection rule : minimize Wilk's Lambda

예)

Variable	F-to-Enter	Wilk's Lambda	
Gas	69.965	0.9087	
Style	8.349	0.9882	⇒ 선택변수 : Reput
Reput	96.012	0.8788	
Handling	32.695	0.9551	

8. 판별분석 과정

◆ 판별분석 과정

① 각 관찰값으로부터 집단구분과 여러 개의 설명변수들을 측정.

② 관찰값이 어느 집단에 속하는지 판별하는데 도움이 되는 변수 선택.

③ 선택된 변수들의 이용하여 판별함수를 만들어 집단들을 구분하는 기준 마련.

④ 판별함수를 이용하여 집단들이 얼마나 정확하게 구별되는 지를 파악.

⑤ 어느 집단에 속하는 지를 알 수 없는 새로운 관측값이 어느 집단에 속하는지 판별.

다음시간에는

11강 판별분석(2)

수고했습니다