7장. 판별분석

판별분석

1. 판별분석 개요
2. 선형판별식의 기하
3. 판별의 원리
4. 선형판별함수
5.1. 정준상관분석
5.2. 정준판별분석
6. 단계적 판별분석
7. 로지스틱 회귀를 이용한 판별분석
8. 의사결정나무

1. 판별분석 개요

❑ 판별분석(Discriminant Analysis) 정의

- 여러 집단에서부터 나온 개체들을 관찰값의 특성에 따라 분류하는 기준을 마련하고 새로운 개체를 이 기준에 따라 분류하는 분석 기법
- 그룹을 나타내는 하나의 변수(집단변수(분류변수))와 집단을 판별하는데 사용되는 다수의 서로 연관된 변수들(판별변수)로 구성됨

❑ 판별분석(Discriminant Analysis) 목적

- 외적 기준에 의하여 정해진 2개 이상의 그룹을 가장 잘 판별하는 설명변수 조합을 찾음으로써 향후 분류 및 예측 등에 활용
  - 각 대상들의 소속집단을 파악하여 주는 판별식을 찾아냄
  - 대상들을 집단으로 분류하는데 의미있는 독립변수들이 어떠한 것인가를 알아냄
  - 각 집단들간에 의미있는 차이가 있는가를 알려줌
  - 판별식을 이용하여 새로운 한 대상을 어느 집단으로 분류할 것인가를 예측

판별분석 개요

❑ 선형판별분석을 위한 가정

– 독립변수들이 다변량 정규분포(multivariate normality)를 이루며,

– 종속변수에 의해 범주화되는 그룹들의 분산-공분산행렬(variance-covariance matrices)이 동일

– 다중정규성 가정을 충족시키지 못하는 자료를 판별분석을 하는 경우

: 판별함수의 추정에 문제를 야기시키며, 이 경우 logistic regression이 사용

다중 정규성 가정을 요구하지 않음.

– 분산-공분산 행렬이 동일하다는 가정이 충족되지 못하는 경우

: 보다 큰 분산-공분산 행렬을 갖는 그룹에 많은 관측치가 분류되는 문제점 발생

판별분석 개요

❑ 판별함수의 수와 판별분석을 위한 표본의 크기

- "종속변수 집단 수 – 1"과 "독립변수의 수" 중에서 작은 값만큼의 판별함수가 만들어짐
- 판별분석을 위해서는 관측치의 개수(표본의 크기)가 독립변수 수의 20배 이상이 되는 것이 요구되며, 종속변수의 각 범주에 최소한 20개가 요구
- 표본의 크기가 이를 충족시키지 못하면 분석결과는 불안정(unstable : 판별식을 구성하는 각 독립변수와 전체 판별식의 설명력과 예측력을 신뢰할 수 없다는 의미)

판별분석 개요

❑ 판별분석(Discriminant Analysis) 과정

- 판별 과정 (discrimination): 관찰된 변수로부터(판별변수 값) 전체집단을 2개의 집단으로 분류하기 위해 기준이 되는 판별함수의 구축 및 해석
  - 사전에 외적 기준으로 정해진 2개 이상의 그룹간 차이 밝혀내기
  - 그룹간 차이를 가장 잘 구분 짓는 설명변수 조합 찾기
- 분류과정 (classification): 새로운 개체들을 선택된 판별 방법에 따라 가장 적절한 집단으로 분류
  - 소속 그룹이 알려지지 않은 개체의 그룹 알아내기

→ 위 두 과정은 서로 분리되어 단독으로 처리 되기보다는 많은 경우 동시에 서로 복합되어 처리 : 판별 및 분류분석을 통틀어 판별분석이라고도 함

판별분석 개요

❑ 일반화 거리를 이용한 판별식

– 적용방법

어떤 특정한 개체 x로부터 부분집단 w_m의 중심 M_m까지의 마할라노비스 (Mahalanobis) 거리제곱을 부분집단 w_m의 공분산행렬 S_m를 사용하여 구하면,

$$d_m^2(x) = (x - M_m)S_m^{-1}(x - M_m)$$

만일 각 집단의 분산공분산행렬이 차이가 없어 공통공분산행렬을 사용하게 된다면 S_m 대신 공통공분산행렬인 S를 이용

$$d_m^2(x) = (x - M_m)S^{-1}(x - M_m)$$

$d_m^2(x)$이 최소가 되는 집단 m에 개체 x를 분류

판별분석 용어 및 개념

❖ 사후확률 (posterior probability): $\mathrm{P}(y = g|\mathbf{x}), \quad g = 1, \cdots, G$

❖ 사전확률 (prior probability): $\pi_g = \mathrm{P}(y = g), \quad g = 1, \cdots, G$

❖ 확률밀도함수(probability density functior $\mathrm{f}_g(\mathbf{x}) = \mathrm{P}(\mathbf{x}|y = g), \quad g = 1, \cdots, G$

❖ 베이즈 정리(Bayes' rule):

$$\mathrm{P}(y = g|\mathbf{x} = \mathbf{x}) = \frac{\pi_g \, \mathrm{f}_g(\mathbf{x})}{\sum_{k=1}^{G} \pi_k \, \mathrm{f}_k(\mathbf{x})}, \quad g = 1, \cdots, G$$

❖ 확률밀도함수에 대한 가정

- 선형판별분석, 이차판별분석 (선형의 경우 동일한 공분산행렬)

$$\mathrm{f}_g(\mathbf{x};\ \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_g|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_g)'\boldsymbol{\Sigma}_g^{-1}(\mathbf{x} - \boldsymbol{\mu}_g)\right]$$

- 로지스틱 판별분석 (로지스틱 회귀분석)

$$\mathrm{f}_g(\mathbf{x};\ \boldsymbol{\beta}) = \frac{\exp(\alpha + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + \exp(\alpha + \beta_1 x_1 + \cdots + \beta_p x_p)}$$

판별분석 용어 및 개념

1. 선형판별함수 (Linear Discriminant Function, LDA)

❖ 모집단 공분산행렬에 대한 동일성 : $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$

❖ 선형판별함수 :

$$\begin{aligned} \mathrm{f}_g(\mathbf{x}) \propto \mathrm{L}_g(\mathbf{x}) &= -\frac{1}{2}\bar{\mathbf{x}}_g'\mathbf{S}^{-1}\bar{\mathbf{x}}_g + \bar{\mathbf{x}}_g'\mathbf{S}^{-1}\mathbf{x} \\ &= a_g + \mathbf{b}_g'\mathbf{x} \quad (\text{단},\ a_g = -\frac{1}{2}\bar{\mathbf{x}}_g'\mathbf{S}^{-1}\bar{\mathbf{x}}_g,\ \mathbf{b}_g = \mathbf{S}^{-1}\bar{\mathbf{x}}_g) \\ &= a_g + b_{g1}x_1 + b_{g2}x_2 + \cdots + b_{gp}x_p \end{aligned}$$

- 선형판별계수 : $\mathbf{b}_g = (b_{g1},\ b_{g2},\ \cdots,\ b_{gp})'$
- 사후확률을 구한 후

$$\begin{cases} \mathrm{P}_1(\mathbf{x}) \ge \mathrm{P}_2(\mathbf{x})\ [\text{즉},\ \mathrm{L}_1(\mathbf{x}) \ge \mathrm{L}_2(\mathbf{x})] & \longrightarrow \quad \text{집단 } G_1\text{에 분류} \\ \mathrm{P}_1(\mathbf{x}) < \mathrm{P}_2(\mathbf{x})\ [\text{즉},\ \mathrm{L}_1(\mathbf{x}) < \mathrm{L}_2(\mathbf{x})] & \longrightarrow \quad \text{집단 } G_2\text{에 분류} \end{cases}$$

- 모집단 공분산행렬이 다르다면 이차판별함수(Quadratic Discriminant Function, QDA) 이용.

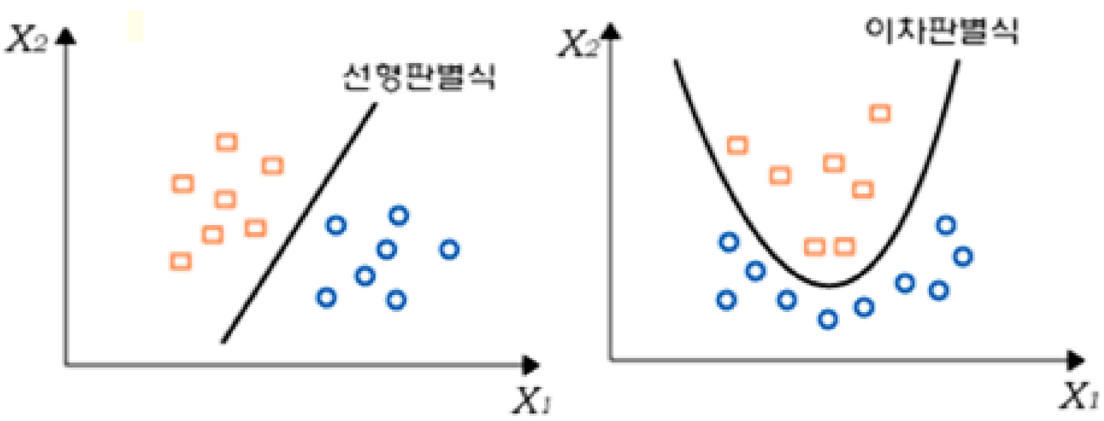
판별분석 용어 및 개념

❑ 일반화 거리를 이용한 판별식

– 적용방법

분산공분산행렬을 어떤 것으로 사용했는가에 따라 다음과 같은 선형판별식 또는

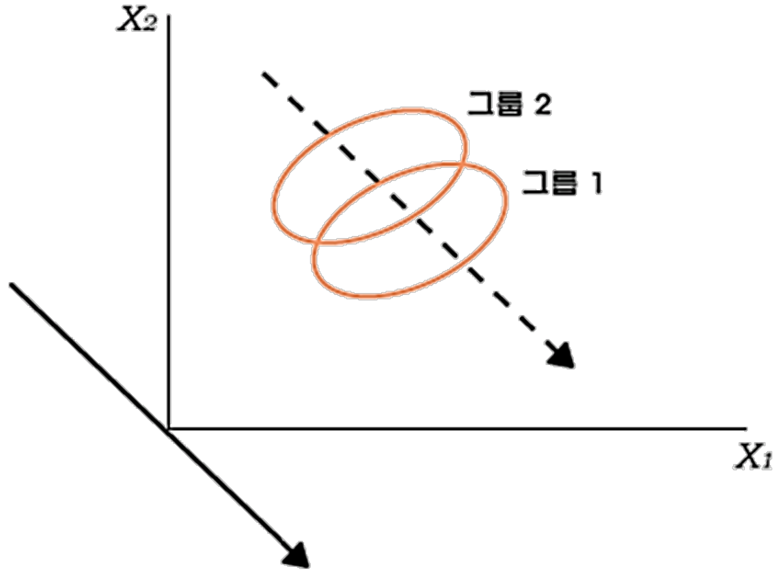
이차판별식을 통해 집단을 구분



2. 선형판별식의 기하

❑ 선형판별식 기하

- 두 개의 그룹이 그림과 같이 분포되어 있다면 두 개의 변수를 동시에 고려한 새로운 축에 의해 구분했을 때 두 그룹의 차이를 명료하게 해 줄 수 있음
- 선형 판별분석은 이와 같은 축을 찾아 집단을 구분하는 것



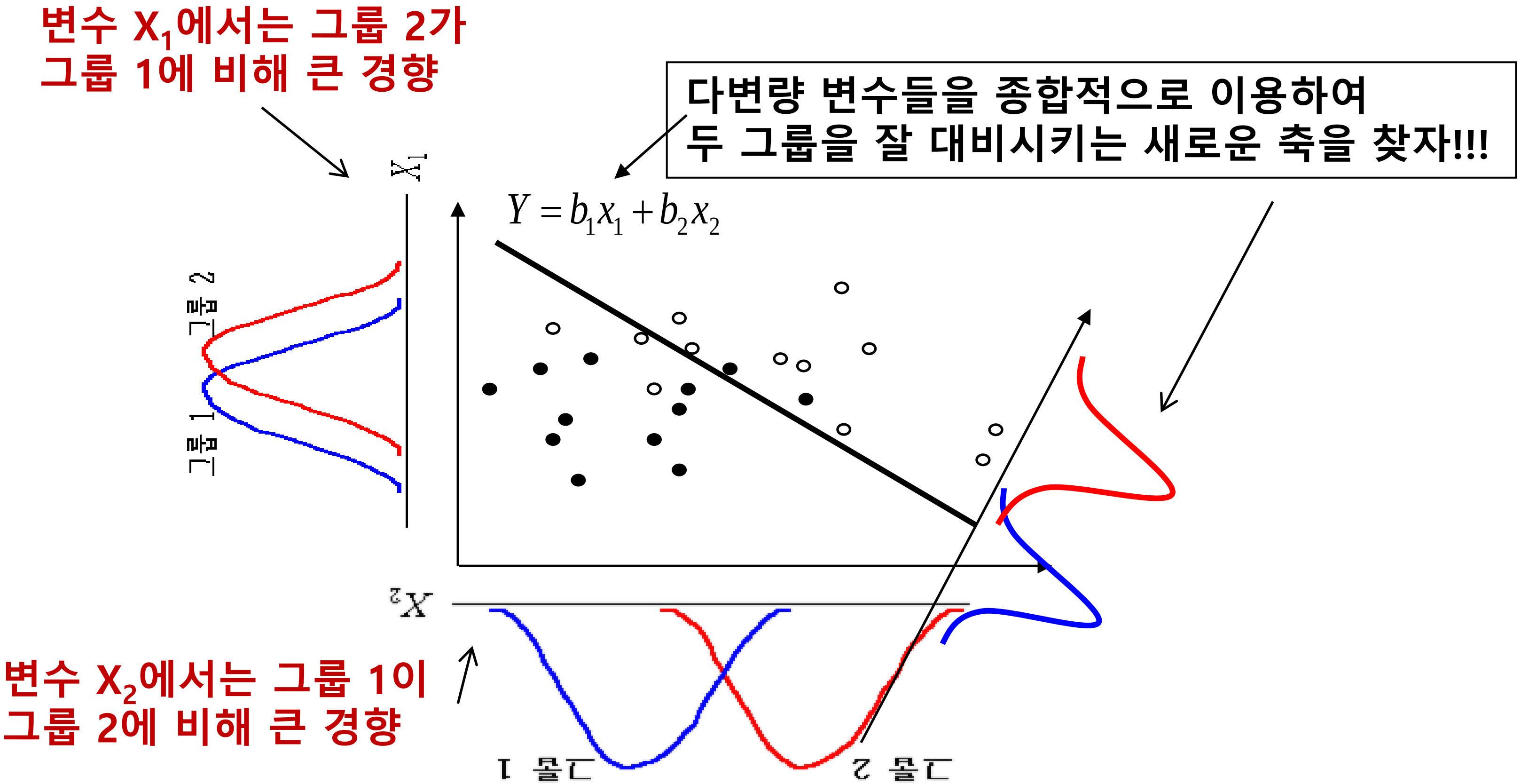
$Y = a_1X_1 + a_2X_2$

- 판별변수 x들의 선형결합을 통해 그룹을 가장 잘 분리하는 새로운 축을 만들어 냄

선형판별식의 기하

❑ 선형판별식 기하

– 다변량 변수의 선형결합에 의하여 전체 데이터를 구분함.



3. 선형판별식의 원리

❑ 원리

- 분석자료 형태 : P개의 변수로 이루어진 다변량 자료를 2개의 그룹으로 분리함.

  그룹 1 (G=1): $(x_{i1}^{(1)}, x_{i2}^{(1)}, \cdots, x_{ip}^{(1)}), \quad i = 1, \cdots, n_1$

  그룹 2 (G=2): $(x_{i1}^{(2)}, x_{i2}^{(2)}, \cdots, x_{ip}^{(2)}), \quad i = 1, \cdots, n_2$

- 각 그룹은 다변량 정규분포를 따른다는 가정하에 표본평균벡터와 표본공분산 행렬로 요약

  그룹 1 (G=1): $\overline{x}^{(1)}, \ S^{(1)}$

  그룹 2 (G=2): $\overline{x}^{(2)}, \ S^{(2)}$

- 만약 두 그룹이 모집단 수준에서 동일한 공분산 행렬을 가질 경우

$$S = \frac{(n_1 - 1)S^{(1)} + (n_2 - 1)S^{(2)}}{(n_1 - 1) + (n_2 - 1)}$$

선형판별식의 원리

❑ 원리

– **임의의 관측벡터:** $x^T = (x_1, x_2, \cdots, x_p)$

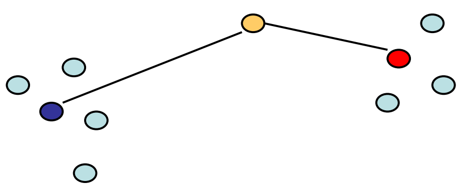
- **그룹 평균(중심점)에서 마할라노비스 거리가 작은 쪽으로 판정**

$$(x-\bar{x}^{(1)})^T S^{-1}(x-\bar{x}^{(1)}) \ge (x-\bar{x}^{(2)})^T S^{-1}(x-\bar{x}^{(2)})$$

임의의 관측벡터가
그룹 2로 판단되는 기준

$$\Leftrightarrow -2\bar{x}^{(1)^T} S^{-1} x + \bar{x}^{(1)^T} S^{-1}\bar{x}^{(1)} \ge -2\bar{x}^{(2)^T} S^{-1} x + \bar{x}^{(2)^T} S^{-1}\bar{x}^{(2)}$$

$$\Leftrightarrow \underbrace{(\bar{x}^{(2)} - \bar{x}^{(1)})^T S^{-1}}_{b^T} x \ge \underbrace{(\bar{x}^{(2)^T} S^{-1}\bar{x}^{(2)} - \bar{x}^{(1)^T} S^{-1}\bar{x}^{(1)})/2}_{c}$$



– $\boldsymbol{x^T}$**에 대한 판별식,** $\boldsymbol{b^T x = b_1 x_1 + \cdots + b_p x_p}$ **과 분류기준**

- $\boldsymbol{x} \in$ 그룹$\mathbf{1}$ $\leftrightarrow$ $\boldsymbol{b^T x < c}$
- $\boldsymbol{x} \in$ 그룹$\mathbf{2}$ $\leftrightarrow$ $\boldsymbol{b^T x \ge c}$

소속 그룹이 알려지지 않은 임의의 관측벡터 $\boldsymbol{x}$ **가**
어느 그룹에서 나온 것일까를 판별하는 기준

선형판별식의 원리

❑ 원리

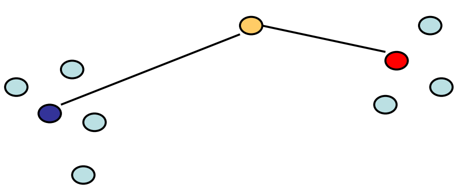
– **임의의 관측벡터:** $x^T = (x_1, x_2, \cdots, x_p)$

- $\boldsymbol{x} \in$ 그룹**1** $\leftrightarrow$ $\boldsymbol{b^T x < c}$
- $\boldsymbol{x} \in$ 그룹**2** $\leftrightarrow$ $\boldsymbol{b^T x \geq c}$

$$(x-\overline{x}^{(1)})^T S^{-1}(x-\overline{x}^{(1)}) \geq (x-\overline{x}^{(2)})^T S^{-1}(x-\overline{x}^{(2)})$$

$$\Leftrightarrow -2\overline{x}^{(1)^T} S^{-1}x + \overline{x}^{(1)^T} S^{-1}\overline{x}^{(1)} \geq -2\overline{x}^{(2)^T} S^{-1}x + \overline{x}^{(2)^T} S^{-1}\overline{x}^{(2)}$$

$$\Leftrightarrow \underbrace{\overline{x}^{(2)^T} S^{-1}x - \overline{x}^{(2)^T} S^{-1}\overline{x}^{(2)}/2}_{L_2(x)} \geq \underbrace{\overline{x}^{(1)^T} S^{-1}x - \overline{x}^{(1)^T} S^{-1}\overline{x}^{(1)}/2}_{L_1(x)}$$



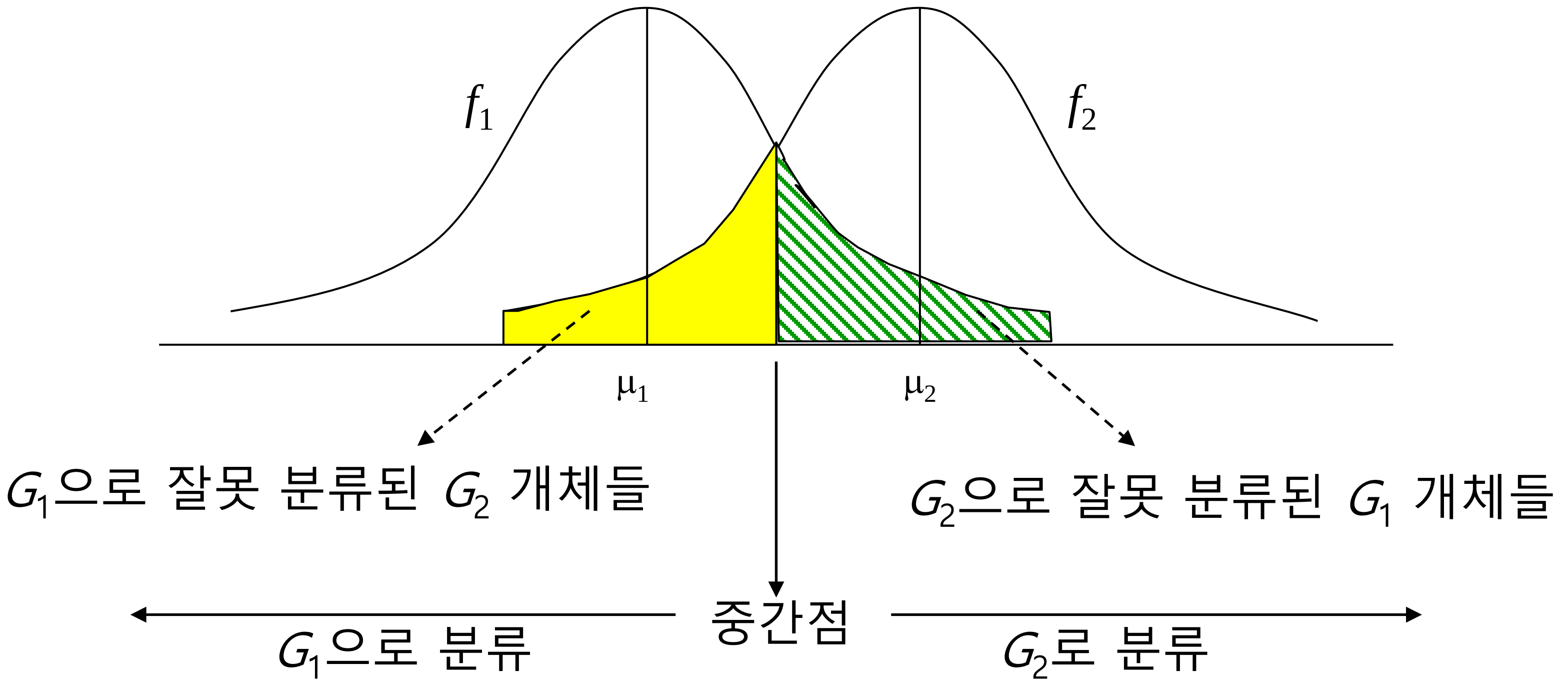
– **판별함수:** $\boldsymbol{L_g(x) = \overline{x}^{(g)^T} S^{-1} x - \overline{x}^{(g)^T} S^{-1}\overline{x}^{(g)}/2}$

✓ $\boldsymbol{x} \in$ 그룹**1** $\leftrightarrow$ $\boldsymbol{L_1(x) \geq L_2(x)}$

✓ $\boldsymbol{x} \in$ 그룹**2** $\leftrightarrow$ $\boldsymbol{L_2(x) \geq L_1(x)}$

선형판별식의 원리

❑ 원리



f_1
f_2
μ_1
μ_2
G_1으로 잘못 분류된 G_2 개체들
G_2으로 잘못 분류된 G_1 개체들
중간점
G_1으로 분류
G_2로 분류

선형판별식의 원리

❑ 원리

- 한 개의 연속형 관찰 변수 X에 근거하여 개체를 두 개의 집단(G_1 또는 G_2)으로 판별
  - 가정
    - 집단 G_1에서 변수 X는 평균이 μ_1이고 분산 Σ인 정규분포
    - 집단 G_2에서 변수 X는 평균이 μ_2이고 분산 Σ 인 정규분포
  - 판별방법
    - 변수 $X=x$ 에 대하여 집단 G_1에 속할 확률 $f_1(x)$과 집단 G_2에 속할 확률 $f_2(x)$에 대하여 $f_1(x) > f_2(x)$ 면 G_1에 할당

선형판별식의 원리

❑ 원리

– 판별에 사용될 변수의 개수가 $X_1, X_2, \ldots, X_p$, 즉, p개의 경우로 확장 가능

– 판별 방법

✓ 변수들 $X_1=x_1$, $X_2=x_2$, …, $X_p=x_p$를 가지는 개체에 대하여 집단 G_1에 속할 확률 $f_1(x)$과 집단 G_2에 속할 확률 $f_2(x)$을 계산하여 $f_1(x) > f_2(x)$면 G_1에 할당

✓ 각 그룹이 다른 사전 확률 π_1, π_2를 가지는 경우 $\pi_1 f_1(x) > \pi_2 f_2(x)$ 면 G_1에 할당

4. 선형판별함수(2그룹)

❑ 가정

– 관찰된 변수 $X_1, X_2, \ldots, X_p$ 가 집단 G_1에 속한다면 평균벡터 μ_1 인 다변량 정규 분포를 따름

– 관찰된 변수 $X_1, X_2, \ldots, X_p$ 가 집단 G_2에 속한다면 평균벡터 μ_2 인 다변량 정규 분포를 따름

– 두 집단의 공분산행렬은 동일

❑ 판별방법

- $\alpha^T(x - \mu) > 0$ 면 개체를 G_1에 할당한다.
- $\alpha^T(x - \mu) = a + b^T x$ (선형판별함수)

여기서 $\alpha^T(x - \mu)$ 의 $\alpha = \Sigma^{-1}(\mu_1 - \mu_2)$ 그리고 $\mu = (\mu_1 + \mu_2)/2$

$$\alpha^T(x - \mu) = (\mu_1 - \mu_2)^T\Sigma^{-1}(x - \mu) = (\mu_1 - \mu_2)^T\Sigma^{-1}x - (\mu_1 - \mu_2)^T\Sigma^{-1}(\mu_1 + \mu_2)/2$$

$$= -(\mu_1 - \mu_2)^T\Sigma^{-1}(\mu_1 + \mu_2)/2 + (\mu_1 - \mu_2)^T\Sigma^{-1}x$$

$$= a + b^T x$$

선형판별함수(2그룹)

❑ Fisher의 idea

- 다변량 자료 x를 두 집단 G1 과 G2를 가능한 한 다르게 나타낼 수 있도록 하는 일변량으로 변환하자!

- $$z = a_1x_1 + a_2x_2 + \cdots + a_px_p$$

- 목적

✓ 두 집단의 표본 평균들을 가장 크게 차이 나도록 만드는 x 의 선형 결합을 선택

✓ 합동 공분산행렬을 가진 다변량 정규분포에서의 선형 판별 함수와 동일

Linear Discriminant Analysis in R

Finance 데이터 (finance.csv)

- 재무지표에 의한 기업분류
- 46개 기업의 재무지표

측정변수

X1: 총부채 대비 현금 유출입

X2: 총자산 대비 순이익

X3: 채무 대비 자산

X4: 순매출 대비 자산

Y: 1='파산기업', 2='건전기업'

Linear Discriminant Analysis in R

Finance 데이터 (finance.csv)

```
> data1<-read.csv("finance.csv",header=T)
> head(data1)
     id     x1     x2    x3    x4 y
1     1 -0.448 -0.410 1.086 0.452 1
2     2 -0.563 -0.311 1.513 0.164 1
3     3  0.064  0.015 1.007 0.397 1
4     4 -0.072 -0.093 1.454 0.258 1
5     5 -0.100 -0.091 1.564 0.668 1
6     6 -0.142 -0.065 0.706 0.279 1
```

❑ 재무지표에 의한 기업분류 (46개 기업의 재무지표)

측정변수

X1: 총부채 대비 현금 유출입

X2: 총자산 대비 순이익

X3: 채무 대비 자산

X4: 순매출 대비 자산

Y: 1=‘파산기업’, 2=‘건전기업’

Linear Discriminant Analysis in R

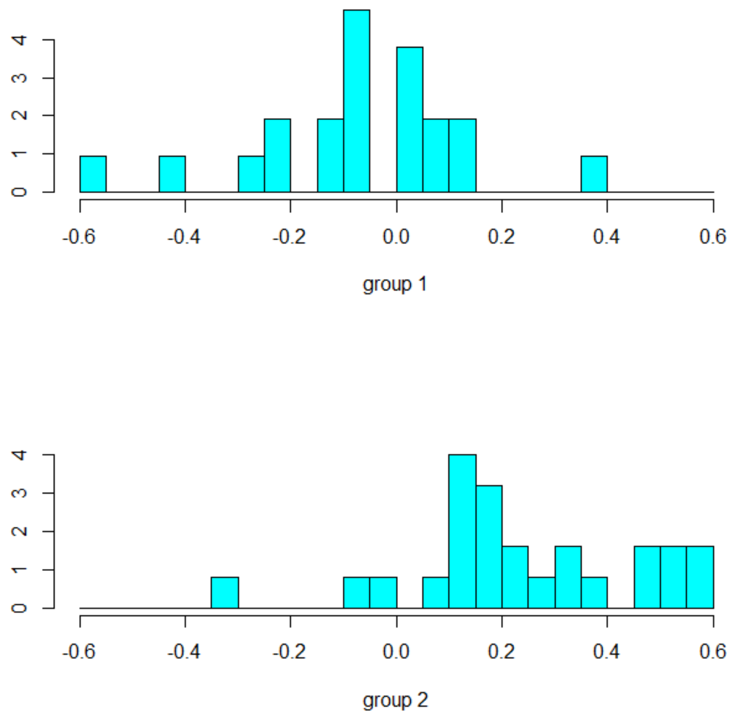
```
> attach(data1)
> library(MASS)
> lda.model1<- lda(y~x1+x2+x3+x4,prior=c(0.5,0.5)) #y 값의 사전확률 1/2씩 할당됨.
> lda.model2<- lda(y~x1+x2+x3+x4) # 사전확률은 y의 1과 2의 비율로 결정됨. 여기서 1은 21개, 2는 25개. 21/46= 0.4565217
> lda.model2
Prior probabilities of groups: # 사전확률
        1         2
0.4565217 0.5434783

Group means:
           x1          x2       x3        x4
1 -0.07319048 -0.08180952 1.367048 0.4378571
2  0.23492000  0.05472000 2.580400 0.4281200

Coefficients of linear discriminants:# 선형판별계수
          LD1
x1  1.0023665
x2  3.9998578
x3  0.8450508
x4 -1.0153181
```

Linear Discriminant Analysis in R

```
> lda.model2$means
           x1          x2       x3        x4
1 -0.07319048 -0.08180952 1.367048 0.4378571
2  0.23492000  0.05472000 2.580400 0.4281200

> lda.model2$counts # y 값 요약.
 1  2
21 25

> lda.model2$prior  # 사전확률
        1         2
0.4565217 0.5434783

> lda.model2$lev  # y의 수준(level)
[1] "1" "2“

> lda.model2$N  # 전체 샘플의 수
[1] 46
```

Linear Discriminant Analysis in R

```
> ldahist(data=x1,g=y,type="histogram") #x1의 그룹별 히스토그램
#(아래 그림 참고)
> ldahist(data=x2, g=y, type="density") #x2의 그룹별 확률밀도함수
> ldahist(data=x3, g=y, type="both")  # x3의 그룹별 히스토그램과 확
률밀도함수
> ldahist(data=x4, g=y) # x4의 그룹별 비교
```



Linear Discriminant Analysis in R

```
> y.fit<-predict(lda.model2,data1[,-1])$class
# 적합한 모델(lda.model2) 에 데이터를 넣어서 y의 분류함.
# 즉 fitted value of y 를 구하는 과정

> table(y.fit,y) # confusion matrix
       y
y.fit   1  2
    1 18  1
    2  3 24
```

오분류율은 (1+3)/46 = 0.08695652

모형을 만들때 쓰인 데이터와 다시 예측에 쓰인 데이터가 동일하기 때문에,과적합이 일반적으로 발생.

즉, 오분류율이 실제 예측보다 낮은 경향이 있다.

선형판별함수(2그룹)

❑ 분류

– 재대입 분류(resubstitution)

실제집단	예측집단	
	G1	G2
G1	N_{11}	N_{12}
G2	N_{21}	N_{22}

- 정확도(Hit ratio) $=(N_{11}+N_{22})/(N_{11}+N_{12}+N_{21}+N_{22})$
- 총 오분류율(Error rate) $=(N_{12}+N_{21})/(N_{11}+N_{12}+N_{21}+N_{22})$
- 분류표의 정확도는 과다하게, 오류율은 과소하게 잡히기 마련

=> 과적합을 방지할 필요가 있다.

선형판별함수(2그룹)

❑ 과적합 방지를 위한 방법

– 교차타당성(Cross-Validation) 기법

- 분류 시 각 관찰값은 자신을 제외한 다른 모든 자료로부터 유도된 함수에 의해 분류
- 즉, (n-1)의 샘플을 이용하여 판별함수를 만든 후, 제외되었던 샘플의 y 값을 예측한다. predicted value of y.
- 위의 과정을 n 번 반복하여 각 샘플에 대한 예측값을 구한다.
- 분석표본(analysis sample, training sample)
  - ✓분류규칙의 산출에 필요한 표본
- 평가표본(validation sample, holdout sample)
  - ✓분류규칙의 수행평가를 위한 표본

교차 타당성(cross-validation)을 이용

```
> lda.model.cv<-lda(y~x1+x2+x3+x4,CV=TRUE)
> y.pred<- lda.model.cv$class
> table(y.pred,y)
      y
y.pred  1  2
     1 17  2
     2  4 23
> lda.model.cv$posterior          # 사후확률
             1           2
1  9.973683e-01 0.002631680
2  9.681017e-01 0.031898254
3  7.322880e-01 0.267711980
4  7.534481e-01 0.246551910
5  8.508852e-01 0.149114823
6  9.052172e-01 0.094782756
7  6.824647e-01 0.317535338
8  7.825799e-01 0.217420132
9 6.646031e-01 0.335396908
…
```

선형판별함수(3그룹)

❑ 가정

– 다변량 정규성, 합동 공분산행렬

❑ 판별방법

– 각 두 개 집단간 선형판별함수

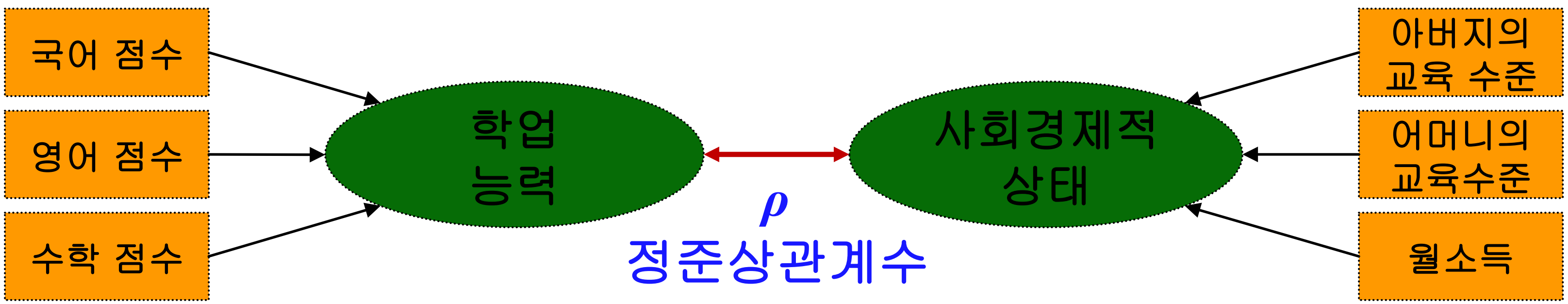
$$h_{12}(x) = (\mu_1 - \mu_2)^T \Sigma^{-1}\left[x - \frac{1}{2}(\mu_1 + \mu_2)\right] \qquad h_{13}(x) = (\mu_1 - \mu_3)^T \Sigma^{-1}\left[x - \frac{1}{2}(\mu_1 + \mu_3)\right]$$

$$h_{23}(x) = (\mu_2 - \mu_3)^T \Sigma^{-1}\left[x - \frac{1}{2}(\mu_2 + \mu_3)\right]$$

- $h_{12}(x) > 0$ 과 $h_{13}(x) > 0$ 이면 집단 G_1 에 할당
- $h_{12}(x) < 0$ 과 $h_{23}(x) > 0$ 이면 집단 G_2 에 할당
- $h_{13}(x) < 0$ 과 $h_{23}(x) < 0$ 이면 집단 G_3 에 할당

5.1 정준상관 분석(Canonical Correlation Analysis)

정준상관 분석의 개념



$$\begin{cases} v & = \boldsymbol{\alpha}'\mathbf{x} & = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p x_p \\ w & = \boldsymbol{\beta}'\mathbf{y} & = \beta_1 y_1 + \beta_2 y_2 + \cdots + \beta_q y_q \end{cases}$$

$$\rho = \max_{\boldsymbol{\alpha},\ \boldsymbol{\beta}} \mathrm{Corr}(v, w)$$

❖ 정준변량 (Canonical Variate)

❖ 정준계수 (Canonical Coefficient)

정준상관 분석(Canonical Correlation Analysis)

정준상관 분석의 개념

- 두 개의 변수 집단 간의 선형성 상관 관계를 파악하고 양으로 표현하고자 할 때
- Hotelling(1935)에 의해 제안된 방법.

-(수학계산속도와 계산능력), (독해속도와 독해능력) 두 개 변수 집단 간의 상관관계 계산

▶ 단순상관계수 : (한 개 변수, 한 개 변수)에 대한 상관성

▶ 다중상관계수 : (한 개 변수, 여러 개 변수)에 대한 상관성

▶ 정준상관계수 : (여러 개 변수, 여러 개 변수) 에 대한 상관성

다차원에 놓인 두 변수 집단간의 관계를 저차원의 정준변수 쌍으로 전환하여 관계를 설명할 수 있음. 정준상관계수가 정준변수간의 상관성을 나타냄.

정준상관 분석(Canonical Correlation Analysis)

정준상관 분석의 개념

$$\rho(v,w) = \frac{\alpha' \Sigma_{xy} \beta}{\sqrt{(\alpha' \Sigma_{xx} \alpha)(\beta' \Sigma_{yy} \beta)}}$$

위의 상관계수를 최대로 하는 정준계수벡터 α 와 β 를 찾는다.

▶ 정준변수를 구하는 과정

1. 첫 번째 정준변수 쌍(first canonical variate pair) (v_1, w_1)은 $|\rho(v,w)|$를 최대로 하며 $Var(v_1) = Var(w_1)$=1 인 변수들의 선형결합식이다.

2. 두 번째 정준변수 쌍(second canonical variate pair) (v_2, w_2)는 (v_1, w_1)과 독립이면서 $|\rho(v,w)|$를 최대로 하며 $Var(v_2) = Var(w_2)$=1 인 변수들의 선형결합식이다.

3. 위와 같은 과정을 반복한다. (정준변수의 개수 = s = min(p,q))

5.2 정준판별분석(canonical discriminant analysis)

❑ 정준판별분석

– 판별변수들에 대한 선형결합 형식의 차원축소를 통해 판별함수 구축

– 선형결합의 표준화 변량 즉 정준판별변수를 이용한 판별분석

– 표준화 정준계수(pooled within-class standardized coefficient)

: 각 판별변수의 상대적 중요도를 의미

정준판별분석(canonical discriminant analysis)

- 선형판별함수($s \le \min(p, k\text{-}1)$)

$$l_1^T x = a_{11}x_1 + a_{21}x_2 + \dots + a_{p1}x_p$$
$$l_2^T x = a_{12}x_1 + a_{22}x_2 + \dots + a_{p2}x_p$$
$$\vdots$$
$$l_s^T x = a_{1s}x_1 + a_{2s}x_2 + \dots + a_{ps}x_p$$

– s개의 선형판별함수를 통해 k개의 집단이 최대한 분리되도록 결정

– 그룹간 변동과 그룹내 변동의 비를 최대화하는 계수를 찾자!

정준판별분석(canonical discriminant analysis)

- 정준판별함수의 계수 (s) 결정 기준
  - s번째 정준판별함수의 공헌도
  - 일반화 우도비 검정 (정규성 가정)
    - 귀무가설: 현재 및 이후 판별함수는 필요 없다.
    - 대립가설: 추가적인 판별함수가 필요하다.

- 개체 분류
  - 정준판별 함수값이 집단 평균 중 가장 가까운 값을 가지는 집단으로 분류

6. 단계적 판별분석

❑ 단계적 판별분석

– 독립변수의 수가 많을 때 어떠한 변수가 좋은 분류에 영향을 미치는가를 알기 어려움

– 이러한 경우에 단계적으로 독립변수를 투입하여 유의한 독립변수를 이용한 판별함수를 찾는 방법

❑ 변수선택(제거)의 방법

– 전방선택법(Forward Selection)
  • 독립변수의 수를 단계별로 증가 시켜나가는 방법

– 후방선택법(Backward Selection)
  • 독립변수의 수를 단계별로 줄여나가는 방법

– 단계별 변수증감법(Stepwise Method)
  • 독립변수의 수를 단계별로 증감하는 방법

단계적 판별분석

❑ 변수선택(제거)의 기준

– 특정변수가 위의 변수선택 판정기준에 의해 도입(또는 제거)을 위해 점검되기에 앞서 일단 partial F-test를 통과하는 것을 전제로 함

– 각 단계에서의 partial F-test(또는 Wilks’ Lambda test)의 유의확률을 산출하고 유의수준과 비교하여 결정

– 각 변수의 기여도를 수정 평균(adjusted mean)의 그룹간 차이로 보고 이에 대한 통계적 유의성(p 값)을 평가하여 변수의 진입과 퇴출을 결정

✓ Wilks’ Lambda : 가장 중요한 변수가 먼저 판별식에 포함되는 방식으로 가장 일반적으로 사용

7. 로지스틱 회귀를 이용한 판별분석

❑ 로지스틱 회귀분석

– 반응값이 연속적이지 않고 범주형일때 사용하는 분석기법
– 연구자는 성공확률을 추정하고, 그 값에 유의한 영향을 미치는 설명변수가
무엇인가를 알아보는 문제에 적용
– 반응변수가 이항(예/아니오, 생존,사망 등)일 때와 다항(명목형, 순서형) 일때로 나눌 수 있음
– 로짓 분석이라고도 함

- 설명변수(연속형) X 반응변수(연속형) : 회귀분석 사용
- 설명변수(연속형) X 반응변수(명목형) : 로짓분석, 로지스틱 회귀분석

예) x : 월수입, y : 외제차 구입여부(1, 0)
x : 독성물질의 양 y : 사망여부(1, 0)

로지스틱 회귀를 이용한 판별분석

❑ 로지스틱 회귀 분석 특징

– 분류 또는 독립변수의 영향력을 알고자 하는 경우에 사용함

– 로짓 분석은 이산형 변수에 대한 분석이며 주로 비선형적 관계를 규명하기 때문에 정규성 가정을 가지고 있지 않음

– 독립변수에 명목/순서형 척도가 포함된 경우 판별분석*보다 주로 로짓 분석을 사용

– 로짓 분석 모형의 적합성 판단은 최우추정법 (Maximum Likelihood Method)으로 수행

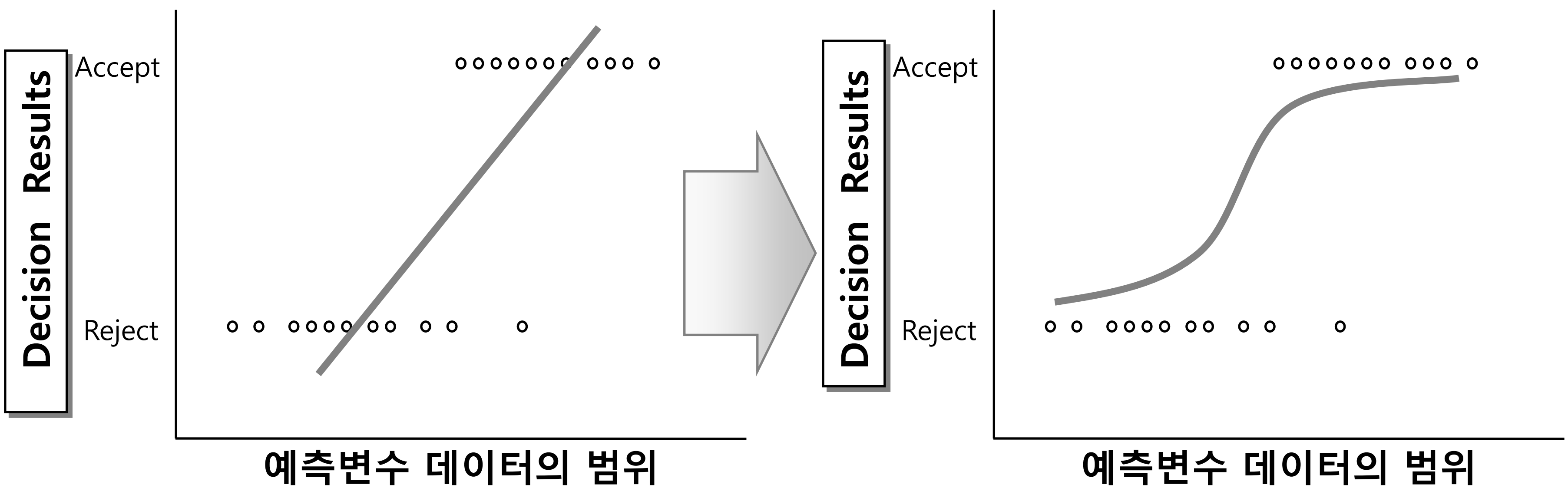
* 선형회귀분석은 최소자승법(Least Squares Method)임

* 판별분석(Discriminant Analysis) : 미리 정해진 그룹간의 차이를 잘 설명하여 줄 수 있는 설명변수들의 선형결합(판별함수)을 찾고, 그에 따라 새로운 개체를 분류하는 분석

로지스틱 회귀를 이용한 판별분석

❑ 로지스틱 회귀 분석의 적용

로지스틱 회귀 모형은 예측 변수에 대한 이산적인 의사결정 과정에서 성공 또는 승인의 확률을 더욱 정확하게 표현하는 회귀 분석 방법임



선형 회귀 분석

로지스틱 회귀 분석

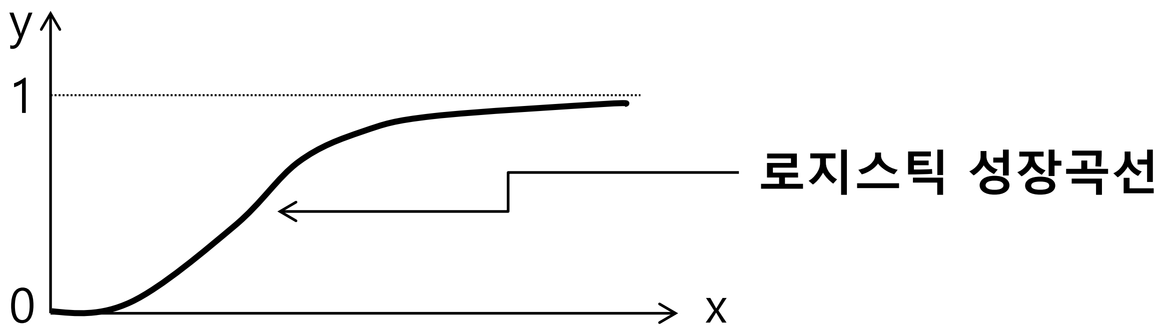
로지스틱 회귀를 이용한 판별분석

❑ 이항 반응변수의 경우

[데이터의 형태]

개체번호	반응 변수	설명 변수			
1	y_1	X_{11}	X_{21}	$\cdots$	X_{k1}
2	y_2	X_{12}	X_{22}	$\cdots$	X_{k2}
⋯	⋯	⋯	⋯	⋯	⋯
N	y_N	X_{1N}	X_{2N}	$\cdots$	X_{kN}

반응변수 y가 두 결과값인 0과 1을 갖고 설명변수 X=x 값을 갖을 때 성공확률 Px를 갖는 베르누이 분포를 따른다고 가정



로지스틱 회귀를 이용한 판별분석

❑ 이항 반응변수의 경우

– 설명변수가 X=x 일 때 종속변수 Y=y 일 확률

$$Y = \begin{cases} 1 & p_x \\ 0 & 1 - p_x \end{cases} \Leftrightarrow$$

– 이때의 성공확률 p_x에 대해 선형 확률 모형을 생각

$$p_x = \alpha + \beta x$$

이때의 모형은 구조적인 결함이 내재

✓ 우선, 확률이기에 0과 1사이의 값을 가져야 하나 선형 확률모형에서는 모든 실수의 값을 가지게 됨

✓ 충분히 크거나 작은 x값에 대해서는 성공확률 p_x 는 0보다 작거나 1보다 큰 값을 갖게 됨

✓ 추정에 있어서 최소제곱추정량(LSE)이 선형 불편추정량에 있어서 더 이상 최소분산을 갖지 않게 됨

로지스틱 회귀를 이용한 판별분석

❑ 이항 반응변수의 경우

– 성공확률 p_x에 대해 선형 확률 모형인

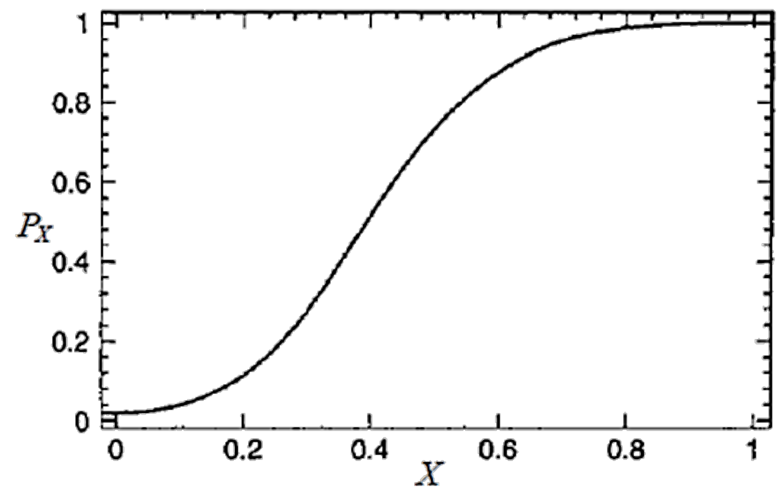
$p_x = \alpha + \beta x$ 의 구조적인 결함의 해결

1, 0 -∞ ~ +∞

$$P(Y = y|X = x) = p_x^y (1 - p_x)^{1-y} \quad (y = 0, 1)$$

일반적으로 위와 같은 구조는 다음과 같은 곡선이 그 성질을 잘 반영한다고 알려져 있음

로지스틱 반응함수
(Logistic Response Function)



로지스틱 회귀를 이용한 판별분석

❑ 이항 반응변수의 경우

로지스틱 회귀 모형에서 종속변수가 2개로 나누어진 경우의 분석

$$\frac{p_x}{1-p_x} = exp(\alpha + \beta x)$$

로짓 (logit : log unit=log odds)

$$\ln\left(\frac{p_x}{1-p_x}\right) = \alpha + \beta x$$

p_x 는 성공 확률 $\quad 0 \le p_x \le 1$

로그를 풀어 p_x 에 대하여 정리하면

Event Probability

$$p_x = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)} \quad \Rightarrow \quad \hat{p}_x = \frac{\exp(\hat{\alpha}+\hat{\beta} x)}{1+\exp(\hat{\alpha}+\hat{\beta} x)}$$

로지스틱 회귀를 이용한 판별분석

❑ 이항 반응변수의 경우

$$p_x = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$$

⬄ β가 0 일 때에는 성공확률 p_x 가 설명변수 X와 무관한 상수값

설명변수와 반응변수간의 관계가 독립

| β|가 커짐에 따라 곡선의 변화율은 커짐

⬄ 성공확률에 대한 오즈(odds)

$$\frac{p_x}{1-p_x} = exp(\alpha + \beta x)$$

⬄ $\ln(\frac{p_x}{1-p_x}) = \alpha + \beta x$: 로짓 모형

▪ 수학에서 그래프 모양을 로지스틱 곡선이라 부르는데 기인하여 모형을 로지스틱 회귀모형 혹은 로짓(logit : Log Unit)모형이라고 함

로지스틱 회귀를 이용한 판별분석

❑ 이항 반응변수의 경우

만약 변수가 여러 개 이면

$$p_x = \frac{\exp(\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}{1+\exp(\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}$$

로짓모형 : $\ln\left(\frac{p_x}{1-p_x}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$

β_i의 추정치 b_i, 이의 표준오차 s_{b_i}

MLE를 이용하여 추정

[가설]

$$H_0 : \beta_i = 0 \qquad vs. \qquad H_1 : \beta_i \neq 0$$

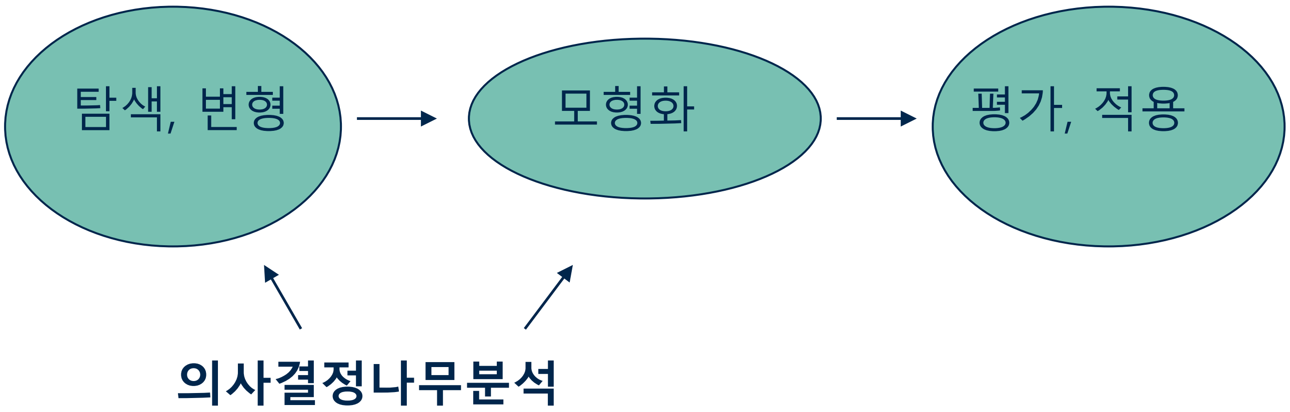
β_i는 근사적으로 정규분포

표준정규분포: $H_0 : \beta_i = 0$ 下

[검정통계량] 왈드통계량 $\chi_w{}^2 = \left(\frac{b_i}{s_{b_i}}\right)^2 \sim \chi^2(1)$

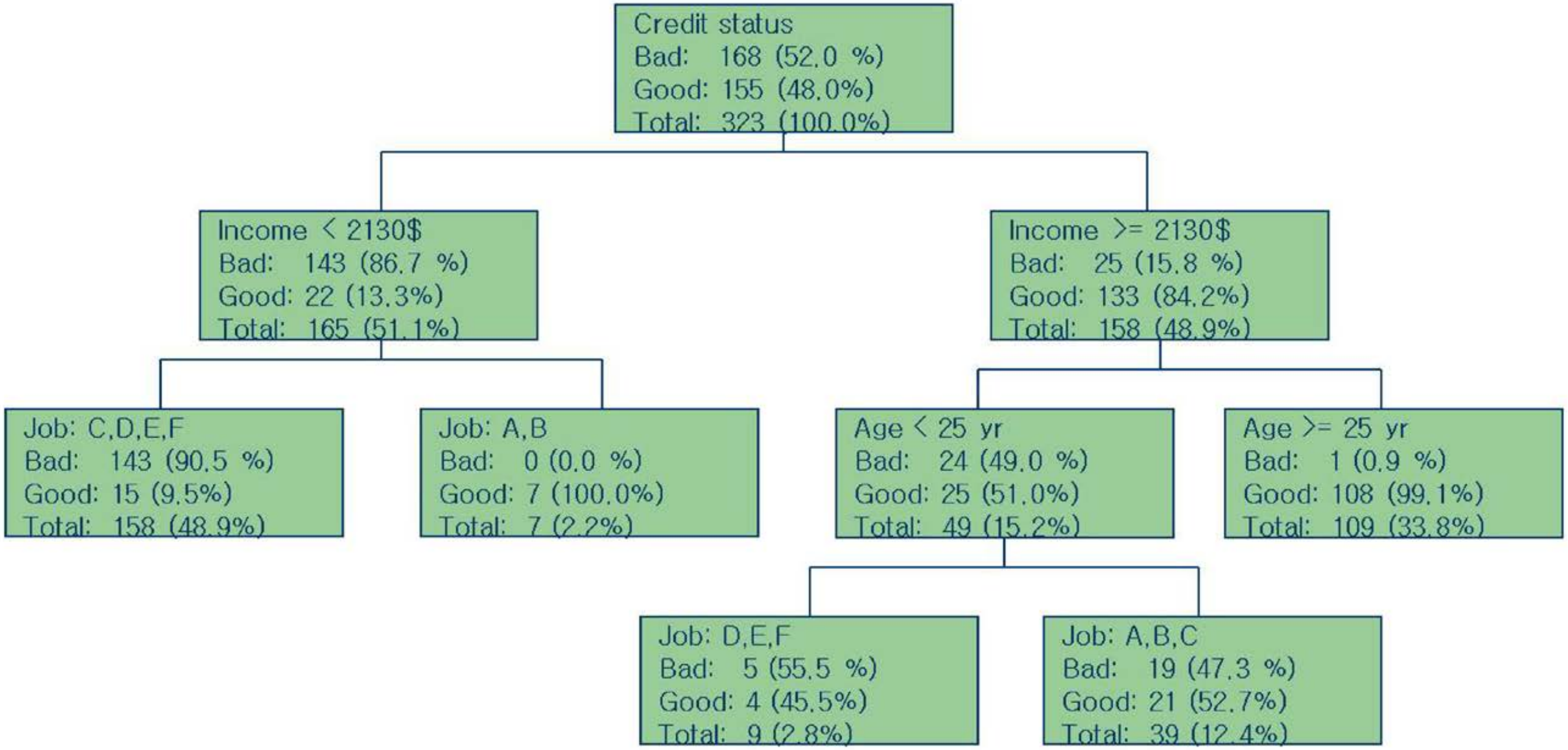
참조 : $Z \sim N(0,1) \rightarrow Z^2 \sim \chi^2(1)$

8. 의사결정나무 (Decision Tree)



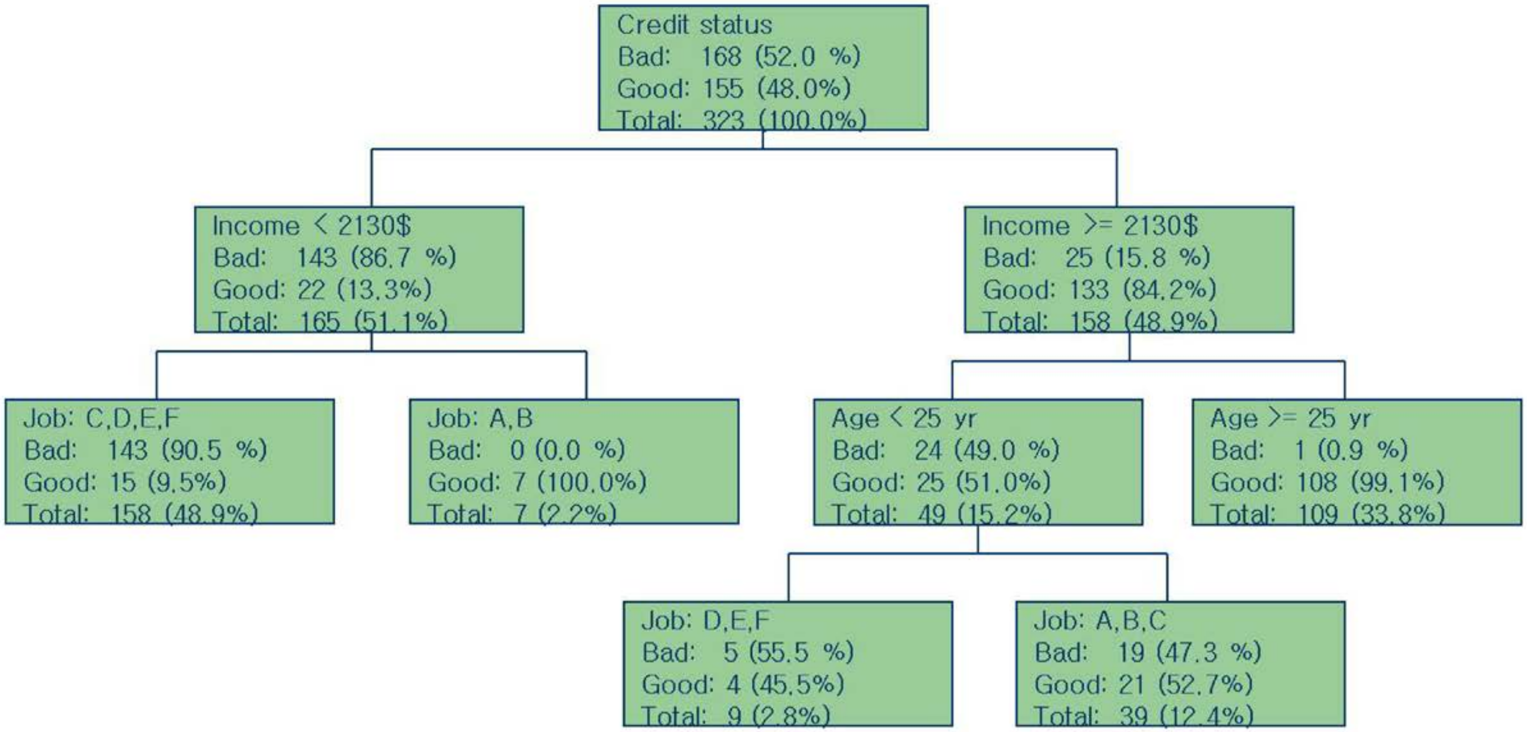
- **탐색**: 이상치의 검색, 분석에 필요한 변수 및 교호효과를 찾아냄
- **모형화**: 판별 및 예측 모형
- 적용결과에 의해 if-then으로 표현되는 규칙이 생성
- 해석이 쉬움 (많은 경우 결정을 내리게 된데 대한 이유를 설명하는 것(해석력)이 중요. 예: 은행의 대출심사 결과 부적격 판정이 나온 경우 고객에게 부적격 이유를 설명하여야 함)

의사결정나무(Decision Tree)의 예



뿌리마디(**root node**): 시작되는 마디로 전체 자료로 구성
자식마디(**child node**): 하나의 마디로부터 분리되어 나간 **2**개 이상의 마디들
부모마디(**parent node**): 주어진 마디의 상위마디
끝마디(**terminal node**): 자식마디가 없는 마디
중간마디(**internal node**): 부모마디와 자식마디가 모두 있는 마디
가지 (**branch**): 뿌리마디로부터 끝마디까지 연결된 마디들
깊이 (**depth**): 뿌리마디부터 끝마디까지의 중간마디의 수

의사결정나무 구축을 위한 질문



- 뿌리마디의 질문이 왜 소득인가? : 분할기준(splitting rule)의 선택
- 4번, 5번, 7번 마디들은 끝마디인 반면 6번 마디는 왜 중간마디인가?

: 분할을 계속할 것인가 그만둘 것인가 (stopping and pruning rule)

- 7번 마디에 속하는 자료는 신용상태를 어떻게 결정하여야 하는가?

: 각 끝마디에서의 예측값 할당

의사결정나무(Decision Tree)의 형성과정

- 나무의 성장(growing): 각 마디에서 적절한 최적의 분리규칙을 찾아서 나무를 성장 시킴. 정지규칙을 만족하면 중단.

- 가지치기(pruning): 오분류율을 크게 할 위험이 높거나 부적절한 추론규칙을 가지고 있는 가지를 제거. 또한, 불필요한 가지를 제거.

- 타당성 평가: 이익도표(gain chart)나 위험도표(risk chart) 또는 테스트 자료 (test sample)를 사용하여 의사결정나무를 평가

- 해석 및 예측: 구축된 나무모형을 해석하고 예측모형을 설정

의사결정나무(Decision Tree)의 분리규칙

- 각 마디에서 분리규칙은 분리에 사용될 입력변수 (분리변수, split variable)의 선택과 분리가 이루어 질 기준 (분리 기준, split criterion)을 의미
- 분리에 사용될 변수(X)가 연속 변수인 경우에는 X 가 분리기준 c 보다 작으면 왼쪽 자식마디로 X 가 c 보다 크면 오른쪽 자식마디로 자료를 분리
- 타당성 평가: 테스트 자료 (test sample)를 사용하여 의사결정나무를 평가
- 분리변수가 범주형인 경우에는 분리기준은 전체 범주를 두 개의 부분 집합으로 나누는 것이 됨. 예를 들면, 전체 범주가 {1,2,3,4}일때 분리 기준의 예로는 {1,2,4}과 {3}이 되고 분리변수가 범주 {1,2,4}에 속하면 왼쪽 자식마디로 범주 {3}에 속하면 오른쪽 자식마디로 자료를 분리

의사결정나무(Decision Tree)의 순수도와 불순도

- 각 마디에서 분리변수와 분리기준은 목표변수의 분포를 가장 잘 구별해주도록 정함
- 목표변수의 분포를 얼마나 잘 구별하는가에 대한 측정치로 순수도 (purity) 또는 불순도 (impurity)를 사용

- (예) 그룹0과 그룹 1의 비율이 45%와 55%인 마디는 각 그룹의 비율이 90%와 10%인 마디에 비하여 순수도가 낮다 또는 불순도가 높다라고 함

- 각 마디에서 분리변수와 분리 기준의 설정은 생성된 두 개의 자식마디의 순수도의 합이 가장 큰 (혹은 불순도의 합이 가장 작은) 분리변수와 분리기준을 선택

불순도의 측도

- 분류모형 (범주형 목표변수)

- 카이제곱 통계량 (chi-square statistics) (5장 참고)

- 지니지수 (Gini index)

- 엔트로피지수 (Entropy index)

- 회귀모형 (연속형 목표변수)

- 분산분석에 의한 F- 통계량 (F-Statistics)

- 불순도에 대한 참고사항

- 불순도는 각 마디마다 정의됨

- 불순도를 이용한 분리기준의 선택에서는 자식마디의 불순도의 합을 가장 작게 하는 분리를 선택

- 이 방법은 부모마디의 불순도에서 자식마디의 불순도의 합의 차이가 최대가 되는 분리를 찾는 것과 동일함

불순도의 측도 - 지니지수

주어진 분리변수와 분리기준에 의하여 다음의 표가 주어졌다고 하자.

	Good	Bad	Total
left	32	48	80
right	178	42	220
Total	210	90	300

$$
\begin{aligned}
\text{지니지수} \quad = \quad & 2\mathbb{P}(\text{left 에서 good})\mathbb{P}(\text{left 에서 bad})\mathbb{P}(\text{left}) \\
& +2\mathbb{P}(\text{right 에서 good})\mathbb{P}(\text{right 에서 bad})\mathbb{P}(\text{right})
\end{aligned}
$$

위의 표에서 지니지수를 구하면

$$
2\frac{32}{80}\frac{48}{80}\frac{80}{300} + 2\frac{178}{220}\frac{42}{220}\frac{220}{300} = 0.355
$$

- 모든 분리변수와 분리기준에서 지니지수를 가장 작게 하는 분리변수와 분리기준 선택

의사결정나무의 정지규칙과 가지치기

정지규칙: 현재의 마디가 더 이상 분리가 일어나지 못하게 하는 규칙

- 규칙의 종류

- 모든 자료가 한 그룹에 속할 때
- 마디에 속하는 자료가 일정 수 이하일 때
- 불순도의 감소량이 아주 작을 때
- 뿌리마디로부터의 깊이가 일정 수 이상일 때 등이 있음

가지치기

- 지나치게 많은 마디를 가지는 (복잡한 모형) 의사결정나무는 새로운 자료에 적용할 때 예측오차가 매우 클 가능성이 있음
- 성장이 끝난 나무의 가지를 제거하여 적당한 크기를 갖는 나무모형을 최종적인 예측모형으로 선택하는 것이 예측력의 향상에 도움이 됨
- 적당한 크기를 결정하는 방법은 평가용 자료(validation data)를 사용하여 예측에러를 구하고 이 예측에러가 가장 작은 나무모형을 선택

의사결정나무의 알고리즘

Classification And Regression Tree (CART)

- 1984년 Leo Breiman과 그의 동료들이 발명 (UC Berkeley)
- 가장 널리 사용되는 의사결정나무 알고리즘
- 이지분류(binary split)를 이용
- 불순도: 목표변수가 범주형인 경우 지니지수를 이용하고 목표변수가 연속형인 경우에는 분산을 이용
- 이해하기 쉬운 규칙을 생성
- 연속형 변수와 범주형 변수를 모두 다 취급할 수 있음
- 이상치에 덜 민감
- 모형의 가정 (선형성, 등분산성 등)이 필요 없는 비모수적 모형

Classification And Regression Tree (CART) in R

```
> library(MASS)
> library(tree)
> data(iris)
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa


> ir.tr = tree(Species ~.,data=iris) # Y변수는 ~ 왼쪽, X변수는 ~ 오른쪽에 표기. 여기서 “.”를 사용하면 모든 X 변수가 사용됨.data 이름을 명시해줘야 Y 변수명을 바로 표기해서 모델을 적합할 수 있음.
> summary(ir.tr)
```

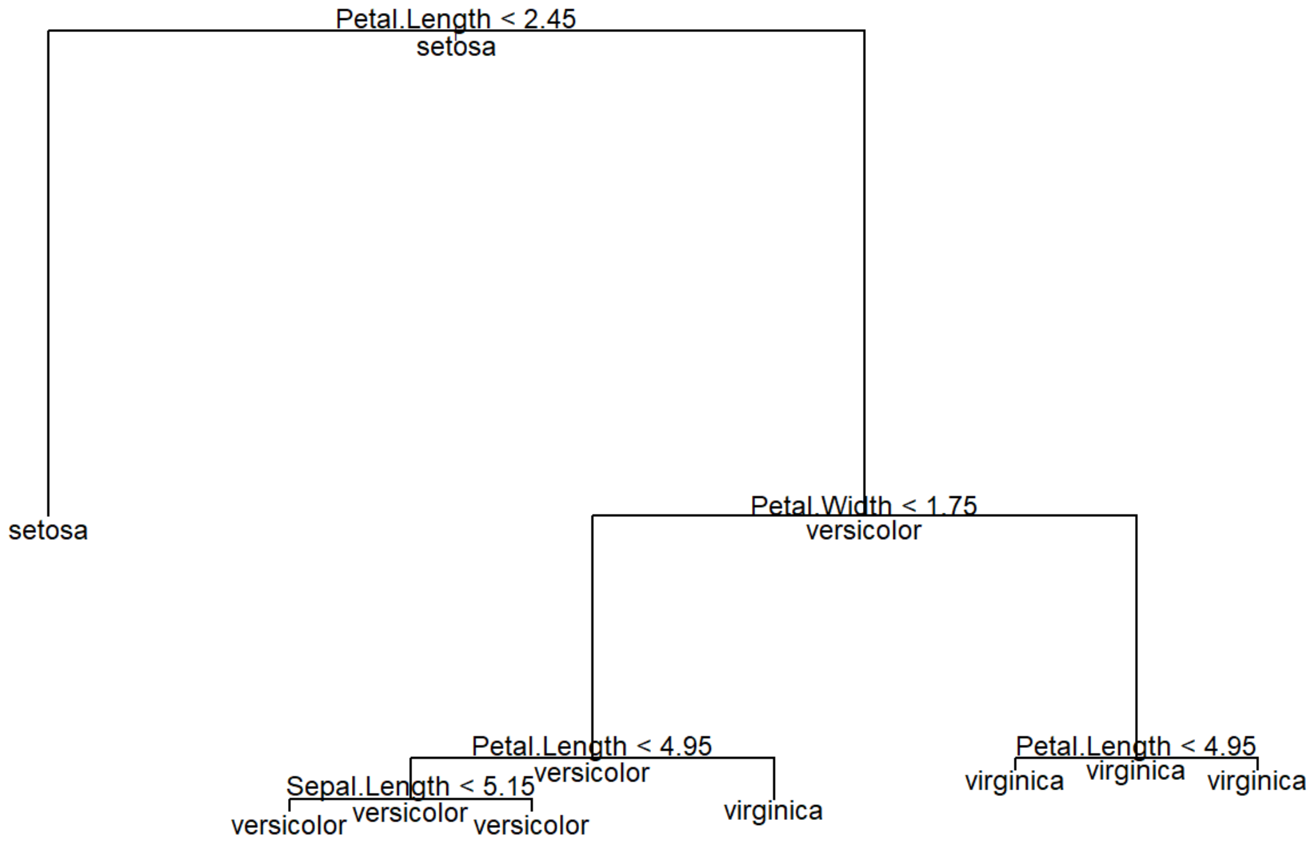
Classification And Regression Tree (CART) in R

```
> summary(ir.tr)

Classification tree:
tree(formula = Species ~ ., data = iris)
Variables actually used in tree construction:
[1] "Petal.Length" "Petal.Width"  "Sepal.Length"
Number of terminal nodes:  6
Residual mean deviance:  0.1253 = 18.05 / 144
Misclassification error rate: 0.02667 = 4 / 150
```

```
> ir.tr
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

 1) root 150 329.600 setosa ( 0.33333 0.33333 0.33333 )
   2) Petal.Length < 2.45 50   0.000 setosa ( 1.00000 0.00000 0.00000 ) *
   3) Petal.Length > 2.45 100 138.600 versicolor ( 0.00000 0.50000 0.50000 )
     6) Petal.Width < 1.75 54  33.320 versicolor ( 0.00000 0.90741 0.09259 )
      12) Petal.Length < 4.95 48   9.721 versicolor ( 0.00000 0.97917 0.02083 )
        24) Sepal.Length < 5.15 5   5.004 versicolor ( 0.00000 0.80000 0.20000 ) *
        25) Sepal.Length > 5.15 43   0.000 versicolor ( 0.00000 1.00000 0.00000 ) *
      13) Petal.Length > 4.95 6   7.638 virginica ( 0.00000 0.33333 0.66667 ) *
     7) Petal.Width > 1.75 46   9.635 virginica ( 0.00000 0.02174 0.97826 )
      14) Petal.Length < 4.95 6   5.407 virginica ( 0.00000 0.16667 0.83333 ) *
      15) Petal.Length > 4.95 40   0.000 virginica ( 0.00000 0.00000 1.00000 ) *
```

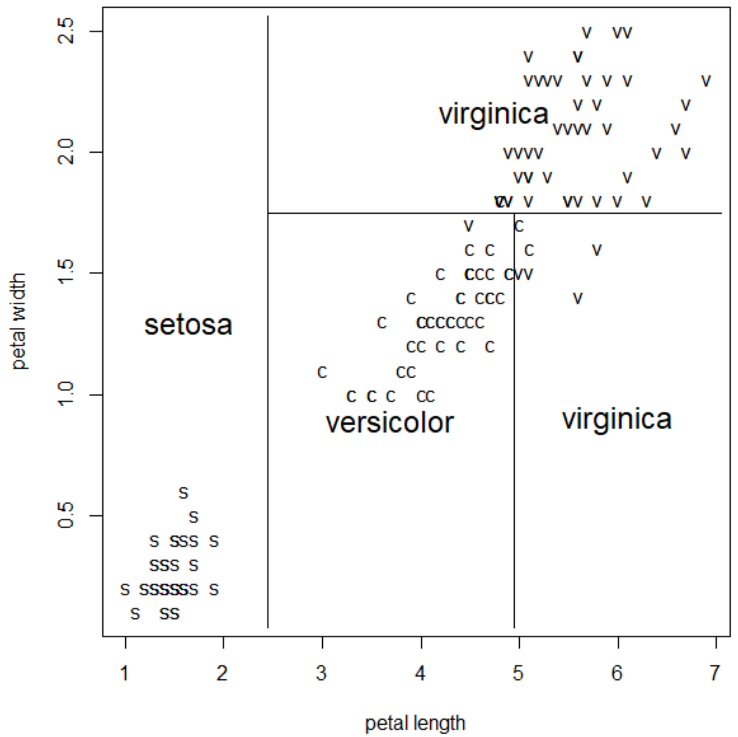
Classification And Regression Tree (CART) in R

```
> plot(ir.tr,lwd=2) # lwd 는 line width 조절
> text(ir.tr, all=T,cex=1.5) # cex 는 글씨크기 조절
```



Classification And Regression Tree (CART) in R

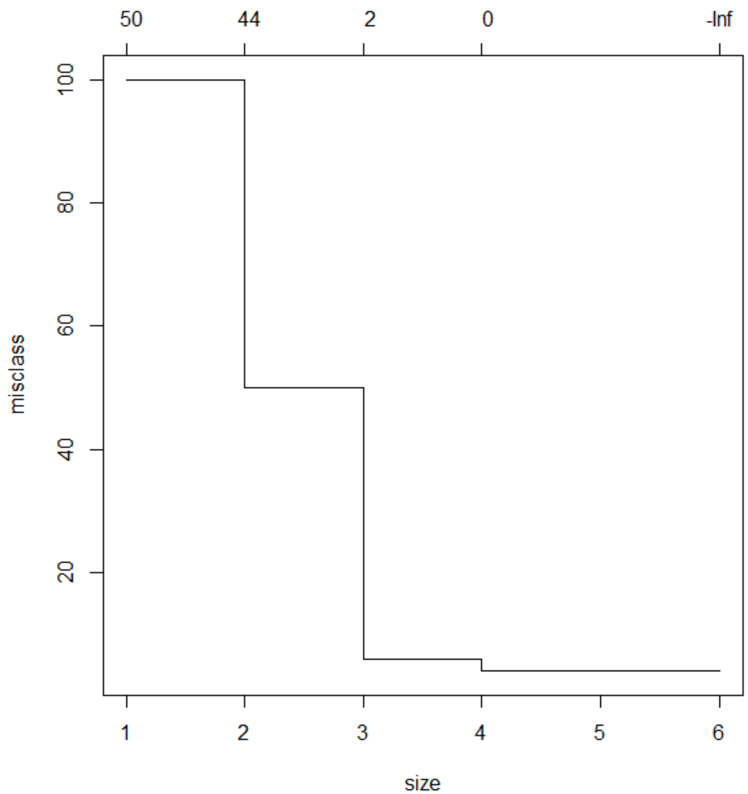
```
> ir.tr1 = snip.tree(ir.tr, nodes = c(12, 7))# 12와 7은 ir.tr의 결과에서 음영표시된 노드. 나무의 한 면을 보여준다. snip: 싹둑 자르다.
> par(pty = "s")  # s는 square plotting region 을 의미
> plot(iris[, 3],iris[, 4], type="n",xlab="petal length", ylab="petal width")# type = none 이기 때문에 그래프에 출력 안됨.
> text(iris[, 3], iris[, 4], c("s", "c", "v")[iris[, 5]])
> partition.tree(ir.tr1, add = TRUE, cex = 1.5)
```



Classification And Regression Tree (CART) in R

가지치기

```
>  plot(prune.misclass(ir.tr))
```
\# 어디서 가지치기를 할 것인지 판단한다. 오분류율도 줄이고 나무의 크기도 줄일 수 있는 적당한 싸이즈를 선택한다. **size=3** 또는 **4**가 좋아보임.

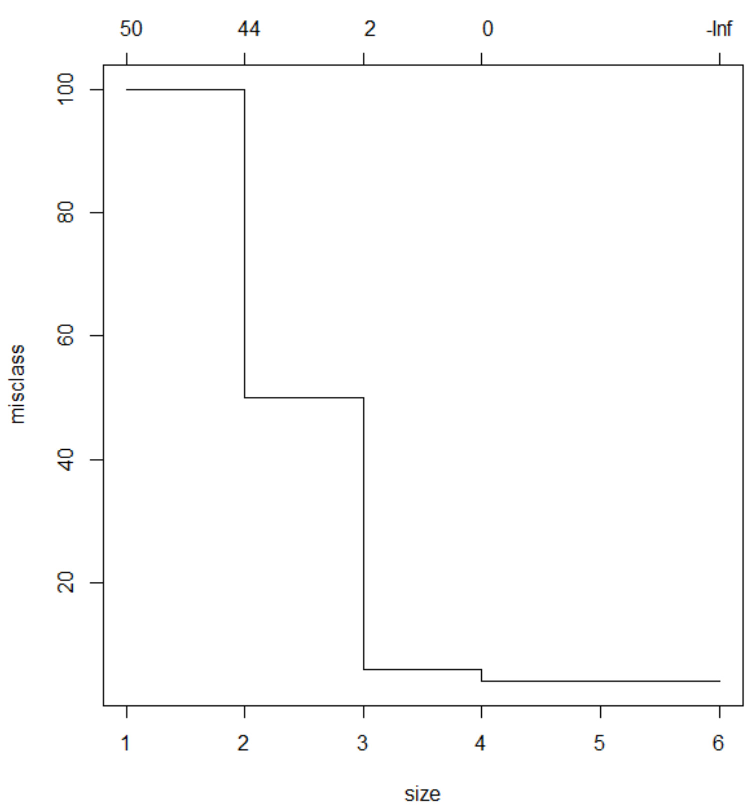


Classification And Regression Tree (CART) in R

가지치기

```
> final.tr1 = prune.misclass(ir.tr, best=3)
> final.tr2 = prune.misclass(ir.tr, best=4)
```

x축 위의 숫자(50 ~ -inf)는 pruning parameter 의 손실-복잡성을 나타내는 값. 가지치기 할 때에는 고려하지 않고 넘어가도 무방. (size에 내포되어 있음)



Classification And Regression Tree (CART) in R

가지치기한 후의 모형

```
> par(mfrow=c(1,2))
> plot(final.tr1,lwd=2)
> text(final.tr1,all=T,cex=1.5)
> plot(final.tr2,lwd=2)
> text(final.tr2,all=T,cex=1.5)
```