판별분석

- 판별분석(discriminant analysis)은 여러 집단에서 추출된 표본으로부터 집단을 잘 구별할 수 있는 분류함수 또는 판별규칙을 도출하고 이를 이용하여 새로운 관측대상이 어떤 집단에 속하는지 판별하고자 할 때 사용하는 통계분석

최적의 분류를 위한 분류규칙

- 판별분석에서 좋은 분류방법이란?
  - 관측대상들의 오분류(misclassification)를 최소로 하는 분류규칙
- 최적의 분류를 위해서 고려해야 할 사항
  - 오분류를 최소로 하기 위해서는 사전확률(prior probability)의 고려가 필요
    - 사전확률이란 임의의 한 관측값이 특정 집단에 속할 확률
  - 최적의 분류를 위해 고려해야할 사항은 오분류에 대한 손실비용

- 따라서 사전확률과 오분류 비용을 고려해 최적의 분류규칙을 설정
- 서로 다른 두 모집단 G_1과 G_2가 있고, 각 모집단에 대한 p개의 확률변수 $X_1, X_2, \cdots, X_p$로 이루어진 확률벡터 $X=(X_1, X_2, \cdots, X_p)^T$의 결합확률밀도함수를 각각 $f_1(x)$, $f_2(x)$
  - S는 표본공간으로 가능한 모든 관측값 x들의 집합
  - R_1과 R_2 각각 모집단 G_1과 G_2로 분류하게 되는 x값들의 집합

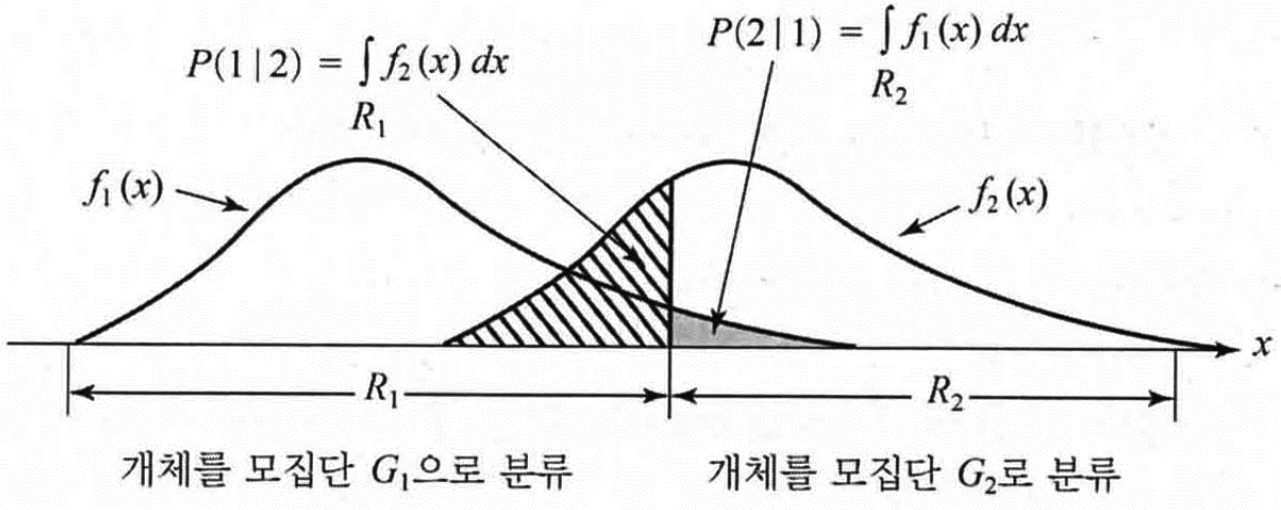
- 모든 관측값들은 모집단 G_1 또는 G_2 중 하나에 반드시 속하므로 R_1과 R_2는 상호배반이며, 표본공간 $S=R_1 \cup R_2$
- 임의의 관측값 x가 실제로 모집단 G_1에 속하지만 모집단 G_2에 속한다고 잘못 판단할 조건부 확률

$$P(2|1) = P(\mathbf{X} \in R_2|G_1) = \int_{R_2} f_1(\mathbf{x})d\mathbf{x}$$

- 임의의 관측값 x가 실제로 모집단 G_2에 속하지만 모집단 G_1에 속한다고 잘못 판단할 조건부 확률

$$P(1|2) = P(\mathbf{X} \in R_1|G_2) = \int_{R_1} f_2(\mathbf{x})d\mathbf{x}$$

- **잘못 판단할 조건부 확률이 일변량인 경우**



- **관측값 x가 모집단 G_1에 속할 사전확률$P(G_1)=p_1$, 모집단 G_2에 속할 사전확률$P(G_2)=p_2$**
- **이때 관측값을 올바르게 판별(정분류)할 확률과 잘못 판별(오분류)할 확률은 사전확률과 조건부 확률의 곱으로 표현**

$$\begin{aligned}
P(G_1\text{으로 정분류}) &= P(G_1 \cap \mathbf{X} \in R_1) = P(G_1)P(\mathbf{X} \in R_1|G_1) \\
&= p_1 P(1|1) \\
P(G_1\text{으로 오분류}) &= P(G_2 \cap \mathbf{X} \in R_1) = P(G_2)P(\mathbf{X} \in R_1|G_2) \\
&= p_2 P(1|2) \\
P(G_2\text{로 정분류}) &= P(G_2 \cap \mathbf{X} \in R_2) = P(G_2)P(\mathbf{X} \in R_2|G_2) \\
&= p_2 P(2|2) \\
P(G_2\text{로 오분류}) &= P(G_1 \cap \mathbf{X} \in R_2) = P(G_1)P(\mathbf{X} \in R_2|G_1) \\
&= p_1 P(2|1)
\end{aligned}$$

- **오분류 비용(cost of misclassification)**
  - **판별을 잘못함으로써 발생하게 되는 손실비용**

분류결과 / 실제 모집단	G_1	G_2
G_1	0	c(2|1)
G_2	c(1|2)	0

  - **c(2|1) : 실제로는 G_1에 속하는데 G_2로 분류하여 발생하는 오분류 비용**
  - **c(1|2) : 실제로는 G_2에 속하는데 G_1로 분류하여 발생하는 오분류 비용**

- **오분류 기대비용(expected cost of misclassification)**
  - **오분류 비용과 해당 오분류 확률의 곱**

$$\mathrm{ECM} = c(2|1)[p_1 P(2|1)] + c(1|2)[p_2 P(1|2)]$$

$$= c(2|1)p_1 \int_{R_2} f_1(\mathbf{x})d\mathbf{x} + c(1|2)p_2 \int_{R_1} f_2(\mathbf{x})d\mathbf{x}$$

  - **표본공간 $S=R_1 \cup R_2$이므로**

$$1 = \int_S f_1(\mathbf{x})d\mathbf{x} = \int_{R_1} f_1(\mathbf{x})d\mathbf{x} + \int_{R_2} f_1(\mathbf{x})d\mathbf{x}$$

- **오분류 기대비용**
  - **오분류 기대비용은**

$$\mathrm{ECM} = c(2|1)p_1\left[1 - \int_{R_1} f_1(\mathbf{x})d\mathbf{x}\right] + c(1|2)p_2\int_{R_1} f_2(\mathbf{x})d\mathbf{x}$$

$$= \int_{R_1}[c(1|2)p_2 f_2(\mathbf{x}) - c(2|1)p_1 f_1(\mathbf{x})]d\mathbf{x} + c(2|1)p_1$$

  - **여기에서 사전확률 p_1과 p_2, 오분류 비용 $c(1|2)$과 $c(2|1)$, 확률밀도함수 $f_1(x)$과 $f_2(x)$는 모두 음이 아닌 값**

- **오분류 기대비용**
  - **따라서 오분류 기대비용이 최소화가 되기 위해서는 영역 R_1이 부등식**

$$[c(1|2)p_2 f_2(\mathbf{x}) - c(2|1)p_1 f_1(\mathbf{x})] \le 0$$

  을 만족하는 x값들을 포함해야 함
  - **즉, 영역 R_1이 다음 부등식을 만족하는 x값들의 집합일 때 오분류 기대비용이 최소화 되는 것**

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \ge \left[\frac{c(1|2)}{c(2|1)}\right]\left[\frac{p_2}{p_1}\right]$$

(확률밀도함수 비) ≥ [비용 비][사전확률 비]

- **오분류 기대비용**
  - **또한 영역 R_2는 R_1의 여집합 이므로 다음 부등식을 만족하는 x값들의 집합일 때 오분류 기대비용이 최소화 되는 것**

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left[\frac{c(1|2)}{c(2|1)}\right]\left[\frac{p_2}{p_1}\right]$$

(확률밀도함수 비) < [비용 비][사잔확률 비]

- **따라서 오분류 기대비용은 확률밀도함수의 비, 오분류 비용의 비, 사전확률의 비에 의하여 최소화**

[오분류 기대비용(EMC)을 최소로 하는 분류규칙]

관측값 x_0에 대하여

- 사전확률이 동일한 경우 즉, $\frac{p_2}{p_1}=1$

  $\frac{f_1(\mathbf{x_0})}{f_2(\mathbf{x_0})} \geq \frac{c(1|2)}{c(2|1)}$이면 $\mathbf{x_0}$를 G_1으로 분류

  $\frac{f_1(\mathbf{x_0})}{f_2(\mathbf{x_0})} < \frac{c(1|2)}{c(2|1)}$이면 $\mathbf{x_0}$를 G_2로 분류

- 오분류 비용이 동일한 경우 즉, $\frac{c(1|2)}{c(2|1)}=1$

  $\frac{f_1(\mathbf{x_0})}{f_2(\mathbf{x_0})} \geq \frac{p_2}{p_1}$이면 $\mathbf{x_0}$를 G_1으로 분류

  $\frac{f_1(\mathbf{x_0})}{f_2(\mathbf{x_0})} < \frac{p_2}{p_1}$이면 $\mathbf{x_0}$를 G_2로 분류

[오분류 기대비용(EMC)을 최소로 하는 분류규칙]

- 사전확률과 오분류 비용이 동일한 경우

$\frac{f_1(\mathbf{x_0})}{f_2(\mathbf{x_0})} \geq 1$이면 $\mathbf{x}_0$를 G_1으로 분류

$\frac{f_1(\mathbf{x_0})}{f_2(\mathbf{x_0})} < 1$이면 $\mathbf{x}_0$를 G_2로 분류

- **오분류 비용이 동일할 경우의 판별기준은 사후확률 $P(G_i|\mathbf{x}_0)$이 큰 집단으로 관측값 $\mathbf{x}_0$를 분류하는 것과 동일**
- **사후확률은 베이즈 정리에 의해**

$$P(G_1|\mathbf{x}_0) = \frac{P(G_1)P(\mathbf{x}_0|G_1)}{P(G_1)P(\mathbf{x}_0|G_1) + P(G_2)P(\mathbf{x}_0|G_2)} = \frac{p_1 f_1(\mathbf{x}_0)}{p_1 f_1(\mathbf{x}_0) + p_2 f_2(\mathbf{x}_0)}$$

$$P(G_2|\mathbf{x}_0) = 1 - P(G_1|\mathbf{x}_0) = \frac{p_2 f_2(\mathbf{x}_0)}{p_1 f_1(\mathbf{x}_0) + p_2 f_2(\mathbf{x}_0)}$$

- 따라서 관측값 x_0가 주어졌을 때 집단 G_1에 속할 사후확률 $P(G_1|x_0)$가 집단 G_2에 속할 사후확률 $P(G_2|x_0)$보다 크거나 같을 때 다음 관계식이 성립

$$p_1 f_1(\mathbf{x}_0) \geq p_2 f_2(\mathbf{x}_0)$$

  - 이는 $\frac{f_1(\mathbf{x}_0)}{f_2(\mathbf{x}_0)} \geq \frac{p_2}{p_1}$과 동일
- 따라서 오분류 비용이 동일할 경우에는 사후확률을 이용하여 관측값을 분류해도 동일한 결과

두 모집단이 다변량 정규분포인 경우 분류규칙

- p개의 확률변수 $X_1, X_2, \cdots, X_p$로 이루어진 확률벡터 $X=(X_1, X_2, \cdots, X_p)^T$에서 j번째 확률변수 X_j의 평균을 μ_j로 표기
- 기대값 E(X)는 모든 확률변수들의 평균을 원소로 갖는 벡터(평균벡터)

$$\boldsymbol{\mu} = E(X) = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}$$

- **두 확률변수 X_j와 X_k의 공분산을 $Cov(X_j, X_k)=\sigma_{jk}$라 하면 p개의 확률변수 간에는 p(p−1)/2개의 공분산이 존재**
  - **특히 j=k이면 공분산은 확률변수 X_j의 분산 $Var(X_j)=\sigma_{jj}$**
- **공분산과 분산을 모아 행렬형태로 표현한 것을 공분산 행렬(Covariance matrix)**

$$\Sigma = \mathrm{E}[(\mathrm{X}-\mu)(\mathrm{X}-\mu)^{\mathrm{T}}] = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}$$

[정의] 다변량 정규분포(multivariate normal distribution)

p차원 확률벡터 $X=(X_1, X_2, \ldots, X_p)^T$의 결합 확률밀도 함수가

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

$$-\infty < x_j < \infty, j = 1,2,\cdots,p$$

와 같을 때, 확률벡터 X는 '모수가 $\boldsymbol{\mu}$, Σ인 다변량 정규분포를 따른다'라고 하며,

$\mathbf{X} = (X_1, X_2, \cdots, X_p)^{\mathrm{T}} \sim N_p(\boldsymbol{\mu}, \Sigma)$으로 표기한다.

- **두 모집단의 공분산 행렬이 동일한 경우 분류규칙**
  - **두 모집단 G_1과 G_2에서 확률벡터 $X=(X_1, X_2, \cdots, X_p)^T$가 각각 공분산 행렬이 동일한 다변량 정규분포 $N_p(\mu_i, \Sigma)$,i=1,2를 따를 때 결합확률밀도함수**

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right)$$
$$(i = 1,2)$$

  - **두 확률밀도함수 $f_1(x)$와 $f_2(x)$의 비**

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \frac{\frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right)}{\frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right)}$$
$$= \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right)$$

  - **따라서 오분류 기대비용이 최소화하는 분류규칙에 대입하면 로그를 취하면**

– 오분류 기대비용이 최소화하는 분류규칙은

$$\ln\left(\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}\right) \geq \ln\left(\left[\frac{c(1|2)}{c(2|1)}\right]\left[\frac{p_2}{p_1}\right]\right)$$

$$(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathrm{T}}\Sigma^{-1}\mathbf{x}-\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2) \geq \ln\left(\left[\frac{c(1|2)}{c(2|1)}\right]\left[\frac{p_2}{p_1}\right]\right)$$

$$\ln\left(\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}\right) < \ln\left(\left[\frac{c(1|2)}{c(2|1)}\right]\left[\frac{p_2}{p_1}\right]\right)$$

$$(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathrm{T}}\Sigma^{-1}\mathbf{x}-\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2) < \ln\left(\left[\frac{c(1|2)}{c(2|1)}\right]\left[\frac{p_2}{p_1}\right]\right)$$

– 일반적으로 모수 μ_1과 μ_2 그리고 Σ는 알려져 있지 않기 때문에 표본으로 부터 추정된 통계량을 사용

– 확률벡터 $X=(X_1, X_2, \cdots, X_p)^{\mathrm{T}}$에 대하여 두 모집단 G_1과 G_2로부터 각각 크기가 n_1과 n_2인 표본을 추출하여 얻은 자료행렬

$$X_1 = \begin{pmatrix} \mathbf{x}_{11}^{\mathrm{T}} \\ \mathbf{x}_{12}^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_{1n_1}^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} x_{111} & x_{112} & \cdots & x_{11p} \\ x_{121} & x_{122} & \cdots & x_{12p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1n_11} & x_{1n_12} & \cdots & x_{1n_1p} \end{pmatrix}$$

$$X_2 = \begin{pmatrix} \mathbf{x}_{21}^{\mathrm{T}} \\ \mathbf{x}_{22}^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_{2n_2}^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} x_{211} & x_{212} & \cdots & x_{21p} \\ x_{221} & x_{222} & \cdots & x_{22p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{2n_21} & x_{2n_22} & \cdots & x_{2n_2p} \end{pmatrix}$$

– 각 집단의 표본평균 벡터와 표본공분산 행렬

$$\bar{\mathbf{x}}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{x}_{1i} \quad , \quad S_1 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(\mathbf{x}_{1i}-\bar{\mathbf{x}}_1)(\mathbf{x}_{1i}-\bar{\mathbf{x}}_1)^T$$

$$\bar{\mathbf{x}}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2}\mathbf{x}_{2i} \quad , \quad S_2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(\mathbf{x}_{2i}-\bar{\mathbf{x}}_2)(\mathbf{x}_{2i}-\bar{\mathbf{x}}_2)^T$$

– 두 집단의 공분산 행렬이 동일하므로 두 표본공분산 행렬로 구한 합동 공분산 행렬 사용

$$S_{pooled} = \frac{(n_1-1)S_1 + (n_2-1)S_2}{n_1+n_2-2}$$

[다변량 정규분포를 따르면서 공분산 행렬이 동일한 경우의 분류규칙]

관측값 x_0에 대하여

$$(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T S_{pooled}^{-1}\mathbf{x}_0$$
$$\geq \frac{1}{2}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T S_{pooled}^{-1}(\bar{\mathbf{x}}_1+\bar{\mathbf{x}}_2) + \ln\left(\left[\frac{c(1|2)}{c(2|1)}\right]\left[\frac{p_2}{p_1}\right]\right)$$

이면 x_0를 G_1으로 분류, 그렇지 않으면 x_0를 G_2로 분류

– 이 분류규칙의 좌변은 선형결합식으로 표현

$$\hat{y} = (\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T S_{pooled}^{-1}\mathbf{x} = \hat{\mathbf{a}}^T\mathbf{x}$$

- 이를 선형판별함수(linear discriminant function)

– 분류규칙의 우변에서 오분류 비용과 사전확률이 각각 동일하면(c(1|2)=c(2|1), $p_1=p_2$)

$$\left[\frac{c(1|2)}{c(2|1)}\right]\left[\frac{p_2}{p_1}\right]=1$$

  • 따라서 ln(1)=0

– 이 때 다음과 같다면

$$\bar{y}_1=(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^{\mathrm{T}}S_{\mathrm{pooled}}^{-1}\bar{\mathbf{x}}_1=\hat{a}^{\mathrm{T}}\bar{\mathbf{x}}_1$$
$$\bar{y}_2=(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^{\mathrm{T}}S_{\mathrm{pooled}}^{-1}\bar{\mathbf{x}}_2=\hat{a}^{\mathrm{T}}\bar{\mathbf{x}}_2$$

– 분류규칙의 좌변은 다음이 성립

$$\hat{m}=\frac{1}{2}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^{\mathrm{T}}S_{\mathrm{pooled}}^{-1}(\bar{\mathbf{x}}_1+\bar{\mathbf{x}}_2)=\frac{1}{2}(\bar{y}_1+\bar{y}_2)$$

  • 이는 두 집단의 판별함수값의 평균 $\bar{y}_1$과 $\bar{y}_2$의 중간값을 의미

– **오분류 비용과 사전확률이 동일한 경우 분류규칙**

- **관측값 x_0로부터 계산되는 선형판별함수의 값 $\hat{y}_0 = \hat{a}^T x_0$를 $\hat{m} = \frac{1}{2}(\bar{y}_1 + \bar{y}_2)$와 비교하여 $\hat{y}_0 \geq \hat{m}$이면 x_0를 G_1으로, 그렇지 않으면 x_0를 G_2로 분류하는 것과 동일**

▪ **부그룹 자료 선택**

```
> subset(x, subset, select)
```

– **x : 부그룹 자료를 선택할 데이터 프레임**

– **subset : 부그룹 선택 조건**

– **select : 부그룹 자료로 선택할 변수 목록**

▪ **평균벡터와 공분산행렬**

```
> cov.wt(x, cor=FALSE, center=TRUE)
```

– **x : 데이터 프레임**

– **cor : 상관계수 행렬 구하기 선택**

– **center : 평균벡터 구하기 선택**

▪ 선형 판별함수(MASS 패키지)

```
> lda(x, grouping, prior=proportions)
```

- x : 자료 행렬 X
- grouping : 그룹 변수
- prior : 사전확률

▪ 판별그래프(klaR 패키지)

```
> partimat(x, grouping, method="lda",
+              plot.matrix=FALSE, imageplot=TRUE)
> partimat(formula, data, method="lda",
+              plot.matrix=FALSE, imageplot=TRUE)
```

- x : 모형분류를 위한 독립변수
- grouping : 집단변수(factor 유형)
- data : 독립변수와 집단변수의 데이터프레임
- method : 판별분석함수
- plot.matrix : 판별 그래프 행렬(TRUE)
- imageplot : 판별 그래프 이미지(색) 그래프 (TRUE)

예제 1.9

- 우리나라 20대 성인의 키(X_1)와 몸무게(X_2)는 남자의 경우 $X=(X_1, X_2)^T$는 다변량 정규분포 $N_p(\mu_1, \Sigma)$를 따르고, 여자의 경우 $X=(X_1, X_2)^T$는 다변량 정규분포 $N_p(\mu_2, \Sigma)$를 따른다고 한다.
  두 집단에서 각 크기가 각각 $n_1=n_2=20$인 표본을 측정한 결과가 다음과 같다고 할 때 다음 물음에 답하여라.(body.txt)

1) 남자와 여자의 모평균 μ_1과 μ_2, 공통 공분산 행렬 Σ의 추정값을 각각 구하여라.
2) 선형판별함수 $\hat{y}$을 구하여라.
3) 분류규칙을 설정하여라.

번호	남자		여자	
	키	체중	키	체중
1	166	72.5	118	64.5
2	143	73.3	147	65.0
3	172	68.8	146	69.0
4	134	66.3	138	64.5
5	172	68.8	175	66.0
6	151	70.0	118	64.5
7	155	69.1	155	70.5
8	178	73.5	146	66.0
9	180	70.0	135	68.0
10	166	71.4	127	68.5

번호	남자		여자	
	키	체중	키	체중
11	186	76.5	136	66.3
12	132	68.0	122	62.0
13	171	72.0	114	63.0
14	187	77.0	140	68.0
15	191	67.0	106	63.0
16	192	75.5	159	66.5
17	181	69.0	127	62.5
18	144	70.5	143	66.5
19	148	74.0	153	66.5
20	179	75.5	139	64.5

- **선형판별함수**

$$\hat{y} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathrm{T}} S_{\text{pooled}}^{-1} \mathbf{x} = \hat{\mathbf{a}}^{\mathrm{T}} \mathbf{x}$$

$$= (29.20 \quad 5.67) \begin{pmatrix} 0.0039 & -0.0116 \\ -0.0116 & 0.1643 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$= 0.0482x_1 + 0.5915x_2$$

- **분류점**

$$\hat{m} = \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathrm{T}} S_{\text{pooled}}^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

$$= \frac{1}{2}(29.20 \quad 5.67) \begin{pmatrix} 0.0039 & -0.0116 \\ -0.0116 & 0.1643 \end{pmatrix} \begin{pmatrix} 303.6 \\ 137.2 \end{pmatrix}$$

$$= 47.911$$

  - **따라서**

$$\mathbf{x}_0 = (x_{10} \quad x_{20})^{\mathrm{T}}$$

$$\hat{y} = 0.0482x_{10} + 0.5915x_{20} \geq 47.911 + \ln\left(\left[\frac{c(1|2)}{c(2|1)}\right]\left[\frac{p_2}{p_1}\right]\right)$$

이면 x_0를 G_1(남자)으로, 그렇지 않으면 x_0를 G_2(여자)로 분류

```
> h.data<-read.table('d:/mydata/body.txt', skip=4,
+                    header=T)
> h.data
   gender  height weight
1    남자     166      72.5
2    남자     143      73.3
3    남자     172      68.8
:      :       :       :
39   여자     153      66.5
40   여자     139      64.5
> Male<-subset(h.data, subset=(gender=='남자'),
+                  select=c(height, weight)) # 남자 자료
> Female<-subset(h.data, subset=(gender=='여자'),
+                    select=c(height, weight)) # 여자 자료
```

```
> library(MASS)
> lda.result<-lda(h.data[,c('height', 'weight')],
+                    grouping=h.data$gender,
+                    prior=c(0.5,0.5))
> # 또는
> lda.result<-lda(gender~height+weight,
+                    prior=c(0.5,0.5), data=h.data)
```

```
> lda.result
Call:
lda(h.data[, c("height", "weight")], grouping = h.data$gender,
    prior = c(0.5, 0.5))

Prior probabilities of groups:
남자 여자
 0.5  0.5

Group means:
     height weight
남자  166.4 71.435
여자  137.2 65.765

Coefficients of linear discriminants:
                LD1
height -0.02212371
weight -0.27102021
```

- 여기에서 판별함수 계수 (0.048289, 0.591555)와 MASS 패키지에서 lda()함수를 이용하여 구한 판별함수 계수 (-0.02212371, -0.27102021)가 차이
- 이러한 차이는 척도화(scale)에 따라서 나타나는 것
- 따라서 (0.048289, 0.591555)=k(-0.022124, -0.271020) 관계로 분류결과는 동일

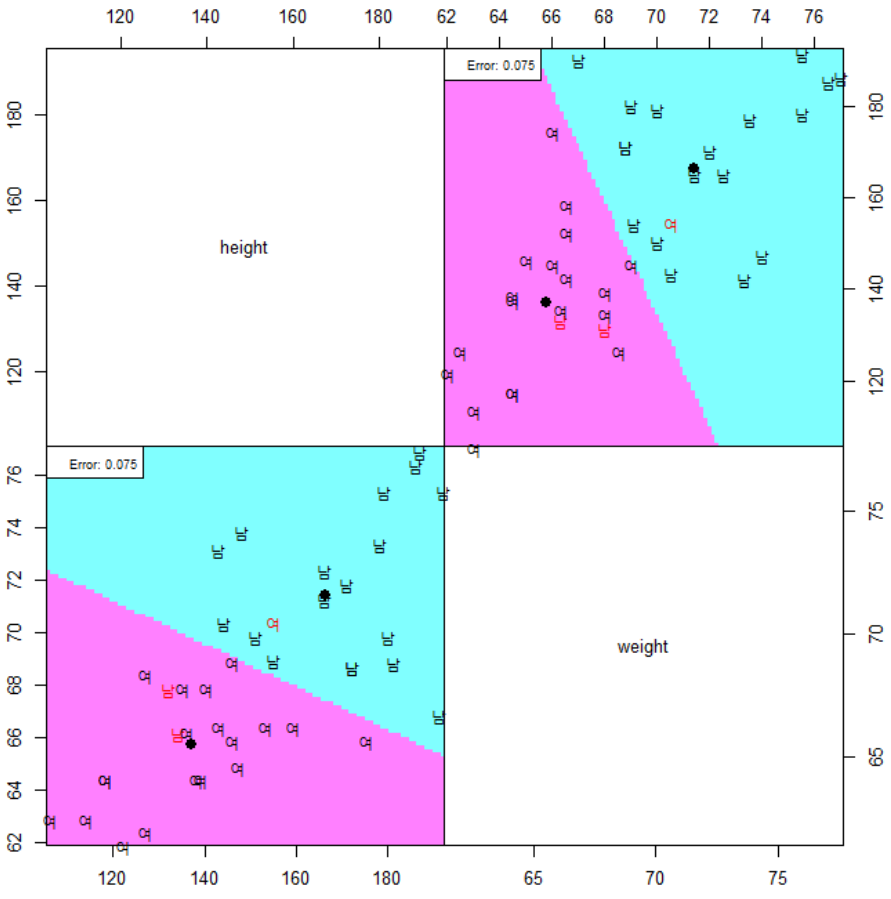
```
> predict(lda.result)
$class
 [1] 남자 남자 남자 여자 남자 남자 남자 남자 남자
[10] 남자 남자 여자 남자 남자 남자 남자 남자 남자
[19] 남자 남자 여자 여자 여자 여자 여자 여자 남자
[28] 여자 여자 여자 여자 여자 여자 여자 여자 여자
[37] 여자 여자 여자 여자
Levels: 남자 여자
```

```
$posterior
              남자         여자
 [1,] 0.952246583 0.047753417
 [2,] 0.913359493 0.086640507
 [3,] 0.749088356 0.250911644
 [4,] 0.097957573 0.902042427
 [5,] 0.749088356 0.250911644
 [6,] 0.687733733 0.312266267
 [7,] 0.610711380 0.389288620
   :          :           :
[37,] 0.008113643 0.991886357
[38,] 0.158796265 0.841203735
[39,] 0.234276651 0.765723349
[40,] 0.045499828 0.954500172
```

```
$x
               LD1
 [1,] -1.37113553
 [2,] -1.07910631
 [3,] -0.50110303
 [4,]  1.01714856
 [5,] -0.50110303
 [6,] -0.36172932
             :
[37,]  2.20189134
[38,]  0.76383111
[39,]  0.54259398
[40,]  1.39436638

> detach("package:MASS", unload=TRUE)
```

```
> library(klaR)
> partimat(h.data[,c('height', 'weight')],
+              grouping=h.data$gender, method="lda",
+              plot.matrix=TRUE, imageplot=TRUE)
> detach("package:klaR", unload=TRUE)
```



- **Fisher의 분류규칙**
  - Fisher는 모집단의 정규성 가정 없이 전혀 다른 접근방법으로 선형결합식을 도출

$$\hat{y} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^{\mathrm{T}} \mathrm{S}_{\mathrm{pooled}}^{-1} \mathbf{x} = \hat{\mathbf{a}}^{\mathrm{T}} \mathbf{x}$$

  - p개의 다변량 관측값 벡터 $\mathbf{x}=(x_1, x_2, \cdots, x_p)^{\mathrm{T}}$를 두 모집단 사이를 최대한 멀리 떨어지도록 만들어주는 일변량 관측값 변수 y로 변환한다는 기본 개념 하에서 $X_1, X_2, \cdots, X_p$의 선형 결합식인 Y의 분산에 비해 두 모집단의 중심인 μ_{1Y}와 μ_{2Y}간의 거리 제곱합이 최대가 되는 $X_1, X_2, \cdots, X_p$의 선형 함수식을 추정

  - 다변량 확률변수 $\mathbf{X}=(X_1, X_2, \cdots, X_p)^{\mathrm{T}}$의 분포와 무관하게 두 모집단 G_1과 G_2가 서로 동일한 공분산 행렬 Σ을 가진다고 가정
  - 확률벡터 X의 선형결합식을 $Y=\mathbf{a}^{\mathrm{T}}\mathbf{X}$이라 할 때, 모집단 G_1과 G_2의 기대값과 분산

$$\mu_{1Y} = \mathrm{E}(Y|G_1) = \mathrm{E}\left(\mathbf{a}^{\mathrm{T}}\mathbf{X}\middle|G_1\right) = \mathbf{a}^{\mathrm{T}}\boldsymbol{\mu}_1$$

$$\mu_{2Y} = \mathrm{E}(Y|G_2) = \mathrm{E}\left(\mathbf{a}^{\mathrm{T}}\mathbf{X}\middle|G_2\right) = \mathbf{a}^{\mathrm{T}}\boldsymbol{\mu}_2$$

$$\sigma_Y^2 = \mathrm{Var}(Y) = \mathrm{Var}\left(\mathbf{a}^{\mathrm{T}}\mathbf{X}\right) = \mathbf{a}^{\mathrm{T}}\mathrm{Cov}(\mathbf{X})\mathbf{a} = \mathbf{a}^{\mathrm{T}}\Sigma\mathbf{a}$$

– **Fisher의 선형결합식 Y는 Y값 전체 변화량에 대한 집단간 변화량의 비를 최대로 하는 a에 의하여 결정**

$$\frac{(\text{두 그룹 평균의 차이})^2}{\text{Y의 분산}} = \frac{(\mu_{1Y} - \mu_{2Y})^2}{\sigma_Y^2} = \frac{\left(\mathbf{a}^T\boldsymbol{\mu}_1 - \mathbf{a}^T\boldsymbol{\mu}_2\right)^2}{\mathbf{a}^T\Sigma\mathbf{a}}$$

$$= \frac{[\mathbf{a}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]^2}{\mathbf{a}^T\Sigma\mathbf{a}}$$

$$= \frac{\mathbf{a}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{a}}{\mathbf{a}^T\Sigma\mathbf{a}}$$

– **이를 최대로 하는 a는 최대화 정리를 이용**

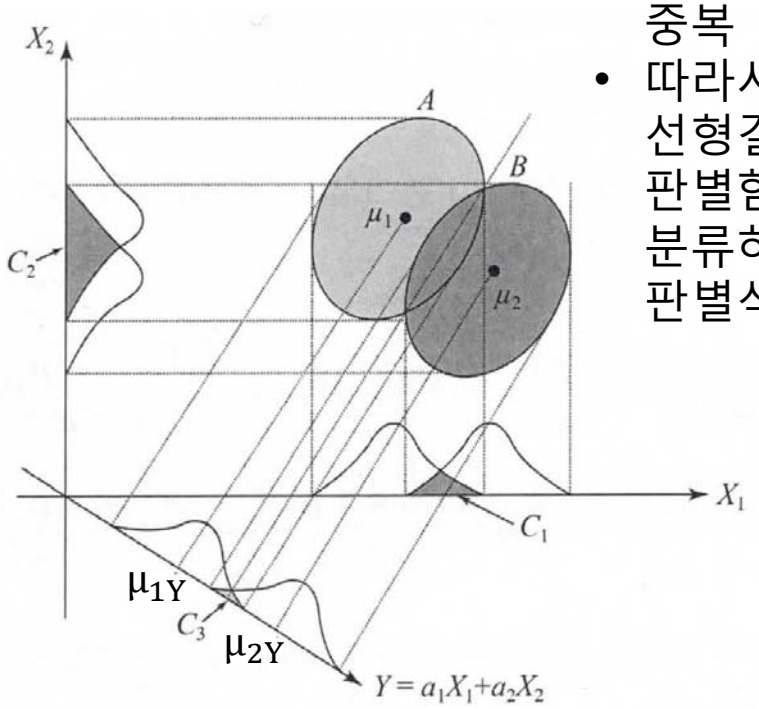
$$\mathbf{a} = c\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \quad , \quad \text{단}, c \neq 0$$

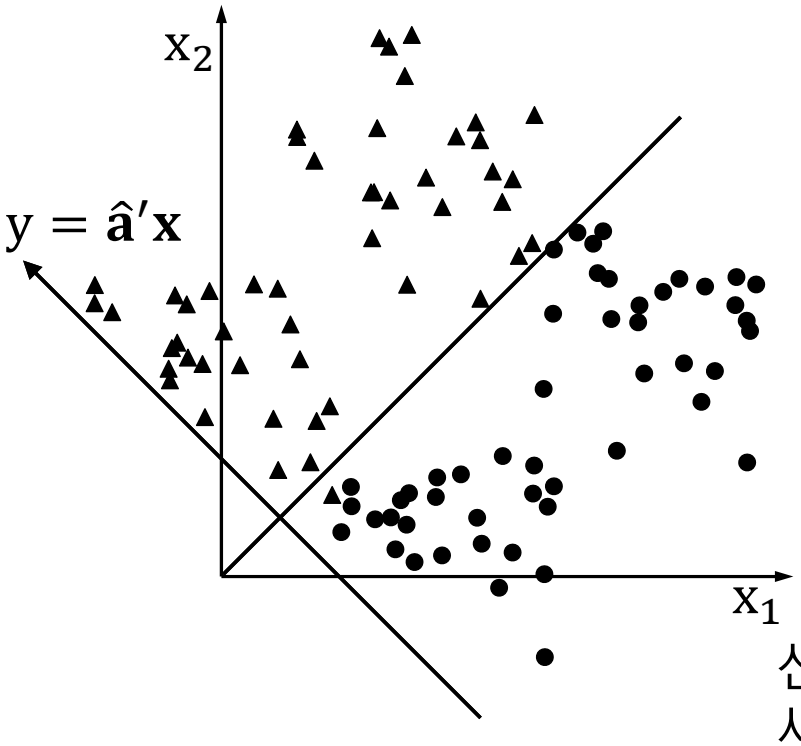
– **c=1로 놓을 때 Fisher의 선형 판별함수**

$$Y = \mathbf{a}^T\mathbf{X} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\Sigma^{-1}\mathbf{X}$$

- **선형판별함수 Y는 두 모집단 G_1과 G_2의 선형판별함수의 평균이 최대한 멀리 떨어지도록 만들어줌**
- **선형판별함수 Y는 두 집단 간의 거리를 최대로 해주는 새로운 축**

- X_1만으로 분류하면 C_1 만큼 중복
- X_2만으로 분류하면 C_2 만큼 중복
- Y에 의해서 분류하면 C_3 만큼 중복
- 따라서 설명변수 X_1과 X_2의 선형결합으로 만들어진 판별함수 Y가 두 집단을 분류하는데 있어 최적의 판별식이 되는 것





선형판별함수에 의해 새로운 y축이 형성되어 두 그룹의 중간점 $\hat{m}$을 기준으로 두 그룹이 나뉘어진 분포

- **선형 판별함수 Y를 이용하여 새로운 관측값 X=x_0에 대한 판별함수 Y의 값**

$$y_0 = \mathbf{a}^T\mathbf{x}_0 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\Sigma^{-1}\mathbf{x}_0$$

- **모집단 G_1과 G_2에서의 평균에 대한 중간값**

$$m = \frac{1}{2}(\mu_{1Y} + \mu_{2Y}) = \frac{1}{2}\left(\mathbf{a}^T\boldsymbol{\mu}_1 + \mathbf{a}^T\boldsymbol{\mu}_2\right)$$
$$= \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\Sigma^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$$

- **따라서 새로운 관측벡터 x_0에 대한 y_0가 중간값 m보다 크다면 G_1으로 분류, y_0가 중간값 m보다 작다면 G_2로 분류**
- **선형 판별함수와 중간값에서 모수 μ_1, μ_2, Σ는 알려져 있지 않으므로 표본의 추정량 $\bar{x}_1$, $\bar{x}_2$, S_{pooled}을 사용**

- **표본에 대한 Fisher의 선형 판별함수**

$$\hat{y} = \hat{\mathbf{a}}^T\mathbf{x} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T S_{pooled}^{-1}\mathbf{x}$$

- **두 모집단 분류점인 중간값**

$$\hat{m} = \frac{1}{2}(\bar{y}_1 + \bar{y}_2) = \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T S_{pooled}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

[Fisher의 분류규칙]

관측값 x_0에 대하여

$$\hat{y}_0 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T S_{pooled}^{-1}\mathbf{x}_0 \geq \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T S_{pooled}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

이면 x_0를 G_1으로 분류, 그렇지 않으면 x_0를 G_2로 분류

– Fisher의 분류규칙은 확률벡터 X가 다변량 정규분포를 따르는 분류규칙에서 오분류 비용과 사전확률이 같을 때와 동일

$$\left[\frac{c(1|2)}{c(2|1)}\right]\left[\frac{p_2}{p_1}\right] = 1$$

예제 1.10

- 새로 개발된 농기구 판매의 성공여부를 알아보기 위하여 A회사에서는 농촌 주민들을 대상으로 잠재적인 구매 여부를 소득(X_1)과 농지면적(X_2)에 따라 조사를 시행하였으며, 그 결과 아래와 같은 결과를 얻었다고 하자. 다음 물음에 답하여라.
- 소득(X_1)과 농지면적(X_2)에 대한 Fisher의 선형 결합식으로 구매와 비구매 집단으로 분류하여라.

▪ 데이터

관측대상	구매 집단(G_1)		비구매 집단(G_2)	
	X_1(소득)	X_2(면적)	X_1(소득)	X_2(면적)
1	20.0	9.2	25.0	9.8
2	28.5	8.4	17.6	10.4
3	21.6	10.8	21.6	8.6
4	20.5	10.4	14.4	10.2
5	29.0	11.8	28.0	8.8
6	36.7	9.6	16.4	8.8
7	36.0	8.8	19.8	8.0
8	27.6	11.2	22.0	9.2
9	23.0	10.0	15.8	8.2
10	31.0	10.4	11.0	9.4
11	17.0	11.0	17.0	7.0
12	27.0	10.0	21.0	7.4
평균	26.49	10.13	19.13	8.82

```
> p.data<-read.table('d:/mydata/farm.txt', skip=4, header=T)
> p.data
   Group    X1    X2
1    구매   20.0   9.2
2    구매   28.5   8.4
3    구매   21.6  10.8
4    구매   20.5  10.4
:       :      :     :
23 비구매  17.0   7.0
24 비구매  21.0   7.4
> # 구매 자료와 비구매 자료
> pur<-subset(p.data, subset=(Group=='구매'),
+                       select=c(X1, X2))
> n.pur<-subset(p.data, subset=(Group=='비구매'),
+                     select=c(X1, X2))
```

```
> Fisher.lda<-function(X1, X2, Group=c('Group1', 'Group2')) {
+   x1<-as.matrix(X1) # 그룹1의 데이터 행렬
+   x2<-as.matrix(X2) # 그룹2의 데이터 행렬
+   n1<-nrow(x1)  # 그룹1의 표본수
+   n2<-nrow(x2)  # 그룹2의 표본수
+
+   s1.result<-cov.wt(x1) # 그룹1의 평균벡터와 공분산행렬
+   s2.result<-cov.wt(x2) # 그룹2의 평균벡터와 공분산행렬
+   bar.x1<-s1.result$center  # 그룹1의 평균벡터
+   bar.x2<-s2.result$center  # 그룹2의 평균벡터
+   s.pooled<-((n1-1)*s1.result$cov+
+       (n2-1)*s2.result$cov)/(n1+n2-2) # 합동공분산행렬
+   inv.s.pooled<-solve(s.pooled) # 합동공분산행렬의 역행렬
+
+   x<-rbind(x1, x2)  # 자료 행렬 X
+   hat.coef<-(bar.x1-bar.x2)%*%inv.s.pooled  # 판별함수계수
+   hat.y<-hat.coef%*%t(x)  # 각 표본별 판별함수 추정값
+   hat.y<-as.vector(hat.y)
```

```
+   hat.m<-(1/2)*(bar.x1-bar.x2)%*%inv.s.pooled%*%
+              (bar.x1+bar.x2)  # 분류점
+   hat.m<-as.numeric(hat.m)
+
+   # 그룹 판정
+   hat.G<-ifelse(hat.y>=hat.m, Group[1], Group[2])
+
+   return(list(coefficients=hat.coef, hat.m=hat.m,
bar.x1=bar.x1, bar.x2=bar.x2, s.pooled=s.pooled,
hat.Group=hat.G))  # 결과값
+ }
```

```
> Fisher.lda(X1=pur, X2=n.pur, Group=c('구매', '비구매'))
$coefficients
             X1          X2
[1,] 0.3006909 1.570369

$hat.m
[1] 21.73876

$bar.x1
      X1         X2
26.49167 10.13333

$bar.x2
       X1          X2
19.133333   8.816667
```

```
$s.pooled
            X1          X2
X1 30.741629 -1.200606
X2 -1.200606  1.068333

$hat.Group
 [1] "비구매" "구매"   "구매"   "구매"   "구매"
 [6] "구매"   "구매"   "구매"   "구매"   "구매"
[11] "구매"   "구매"   "구매"   "비구매" "비구매"
[16] "비구매" "구매"   "비구매" "비구매" "비구매"
[21] "비구매" "비구매" "비구매" "비구매"
```

▪ Fisher의 판별함수

$$\hat{y} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T S_{pooled}^{-1} \mathbf{x} = \hat{\mathbf{a}}^T \mathbf{x}$$

$$= (7.358 \quad 1.317) \begin{pmatrix} 0.034 & 0.038 \\ 0.038 & 0.979 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

$$= 0.3007x_1 + 1.5704x_2$$

▪ 분류점

$$\hat{m} = \frac{1}{2}(\bar{y}_1 + \bar{y}_2) = \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T S_{pooled}^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

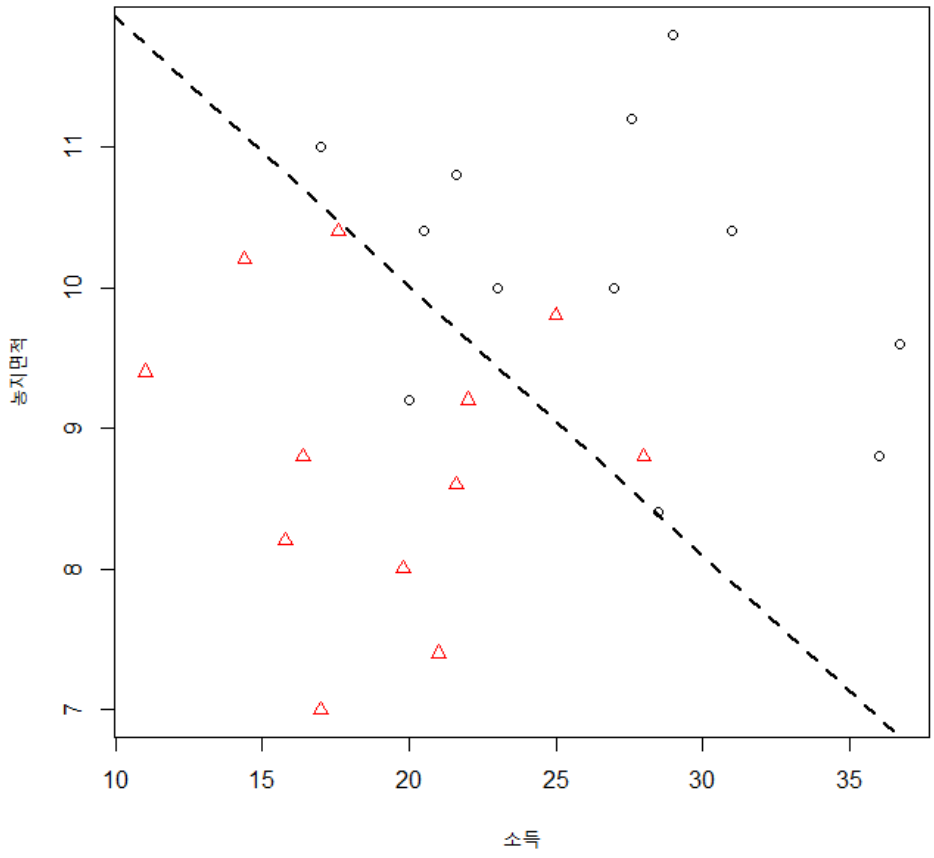
$$= \frac{1}{2}(7.358 \quad 1.317) \begin{pmatrix} 0.034 & 0.038 \\ 0.038 & 0.979 \end{pmatrix} \begin{pmatrix} 45.63 \\ 18.95 \end{pmatrix}$$

$$= 21.738$$

▪ 예측집단

소득 (X_1)	면적 (X_2)	y	실제 집단	예측 집단	소득 (X_1)	면적 (X_2)	y	실제 집단	예측 집단
20.0	9.2	20.461	구매	비구매	25.0	9.8	22.907	비구매	구매
28.5	8.4	21.761	구매	구매	17.6	10.4	21.624	비구매	비구매
21.6	10.8	23.455	구매	구매	21.6	8.6	20.000	비구매	비구매
20.5	10.4	22.496	구매	구매	14.4	10.2	20.348	비구매	비구매
29.0	11.8	27.250	구매	구매	28.0	8.8	22.239	비구매	구매
36.7	9.6	26.111	구매	구매	16.4	8.8	18.751	비구매	비구매
36.0	8.8	24.644	구매	구매	19.8	8.0	18.517	비구매	비구매
27.6	11.2	25.887	구매	구매	22.0	9.2	21.063	비구매	비구매
23.0	10.0	22.620	구매	구매	15.8	8.2	17.628	비구매	비구매
31.0	10.4	25.653	구매	구매	11.0	9.4	18.069	비구매	비구매
17.0	11.0	22.386	구매	구매	17.0	7.0	16.104	비구매	비구매
27.0	10.0	23.822	구매	구매	21.0	7.4	17.935	비구매	비구매

```
> result<-Fisher.lda(X1=pur, X2=n.pur,
+                            Group=c('구매', '비구매'))
> coef<-result$coefficients
> m<-result$hat.m
>
> attach(p.data)
> plot(X1, X2, xlab='소득', ylab='농지면적',
+      col=as.numeric(Group),
pch=as.numeric(Group))
> abline(a=m/coef[2], b=-coef[1]/coef[2])
> detach(p.data)
```



– **분할표**

예측집단 / 실제집단	구매	비구매	계
구매	11	1	12
비구매	2	10	12
계	13	11	24

- **구매 집단 12명 중에서 1명은 잘 못 분류**
- **비구매 집단 12명 중에서 2명은 잘 못 분류**

▪ **두 모집단의 공분산 행렬이 다른 경우 분류규칙**

– **두 모집단이 다변량 정규분포를 따르고 공분산 행렬이 서로 다른 경우의 분류규칙은 다소 복잡**

– **두 정규확률밀도함수의 비에서 공분산 비의 항이 소거되지 않으며 지수부분의 이차식도 단순하게 정리되지 않기 때문**

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \frac{|\Sigma_2|^{\frac{1}{2}}}{|\Sigma_1|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^{\mathrm{T}}\Sigma_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^{\mathrm{T}}\Sigma_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) \right)$$

– **양변에 로그를 취하면**

$$\ln\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = -\frac{1}{2}\mathbf{x}^{\mathrm{T}}(\Sigma_1^{-1} - \Sigma_2^{-1})\mathbf{x} + \left(\boldsymbol{\mu}_1^{\mathrm{T}}\Sigma_1^{-1} - \boldsymbol{\mu}_2^{\mathrm{T}}\Sigma_2^{-1}\right)\mathbf{x} - k$$

• **여기에서**

$$k = \frac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + \frac{1}{2}\left(\boldsymbol{\mu}_1^{\mathrm{T}}\Sigma_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^{\mathrm{T}}\Sigma_2^{-1}\boldsymbol{\mu}_2\right)$$

– **이 식의 우측 첫 번째 항이 x에 대한 이차식이 되므로 이차판별함수(quadratic discriminant function)**

– **일반적으로 모수 $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, Σ_1, Σ_2는 알려져 있지 않으므로 표본의 추정량 $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$, S_1, S_2를 사용**

[다변량 정규분포를 따르면서 공분산 행렬이 서로 다른 경우의 분류규칙]

관측값 x_0에 대하여

$$-\frac{1}{2}\mathbf{x}_0^{\mathrm{T}}(S_1^{-1} - S_2^{-1})\mathbf{x}_0 + \left(\bar{\mathbf{x}}_1^{\mathrm{T}}S_1^{-1} - \bar{\mathbf{x}}_2^{\mathrm{T}}S_2^{-1}\right)\mathbf{x}_0 \geq k + \ln\left(\left[\frac{c(1|2)}{c(2|1)}\right]\left[\frac{p_2}{p_1}\right]\right)$$

이면 x_0를 G_1으로 분류, 그렇지 않으면 x_0를 G_2로 분류

- 모집단이 정규분포를 따르지 않으면 정규분포에 근사하도록 변환
- 공분산 행렬의 동일성에 대한 검정을 통하여 (Box의 M 검정)

$$H_0: \Sigma_1 = \Sigma_2 \quad , \quad H_1: \Sigma_1 \neq \Sigma_2$$

  - 공분산 행렬이 동일하면 선형 판별함수
  - 공분산 행렬이 동일하지 않으면 이차 판별함수

▪ 이차판별함수(MASS 패키지)

```
〉 qda(x, grouping, prior=proportions)
```

- x : 자료 행렬 X
- grouping : 그룹 변수
- prior : 사전확률

▪ 공분산 행렬의 동일성 검정(biotools 패키지)

```
〉 boxM(data, grouping)
```

- data : 자료 행렬 X
- grouping : 그룹 변수

- **판별분석 변수선택(klaR 패키지)**

```
> stepclass(formula, data, method="lda",
+           direction=c("both", "forward", "backward"))
> stepclass(x, grouping, method,
+           direction=c("both", "forward", "backward"))
```

  - **formula : 판별함수 모형식(G~X1+X2+…)**
    - **또는 두 번째 형식으로 독립변수와 그룹변수**
      - **x : 모형분류를 위한 독립변수**
      - **grouping : 집단변수**
  - **method : 분류함수 종류("lda", "qda")**
  - **direction : 변수선택방법**

예제 1.11

- **연어는 강의 상류천에서 부화한 후 바다로 나아가 생활하게 된다. 그러다 산란기가 되면 알을 낳기 위하여 다시 자신이 태어난 곳으로 되돌아와 산란 후 최후의 죽음을 맞이하게 된다. 아래 표 (salmon.txt)는 알래스카와 캐나다 두 지역에서 부화한 연어의 크기를 측정한 결과로서 X_1은 강물에서, X_2는 바다물에서 성장한 길이를 각각 나타낸다.**
  두 집단에 대한 확률벡터 $X=(X_1, X_2)^T$가 각각 다변량 정규분포 $N_p(\mu_i, \Sigma_i)$, i=1,2를 따른다고 할 때, 다음 물음에 답하여라.

알래스카(group1)		캐나다(group2)		알래스카(group1)		캐나다(group2)	
X1	X2	X1	X2	X1	X2	X1	X2
108	368	129	420	102	429	145	376
131	355	148	371	101	469	115	354
105	469	179	409	85	444	134	383
86	506	152	381	109	397	117	355
99	402	166	377	106	442	126	345
87	423	124	389	82	431	118	379
94	440	156	419	118	381	120	369
117	489	131	345	105	388	153	403
79	432	140	362	121	403	150	354
99	403	144	345	85	451	154	390
114	428	149	393	83	453	155	349
123	372	108	330	53	427	109	325
123	372	135	355	95	411	117	344
109	420	170	386	76	442	128	400
112	394	152	301	95	426	144	403
104	407	153	397	87	402	163	370
111	422	152	301	70	397	145	355
126	423	136	438	84	511	133	375
105	434	122	306	91	469	128	383
119	474	148	383	74	451	123	349
114	396	90	385	101	474	144	373
100	470	145	337	80	398	140	388
84	399	123	364				

1) 오분류 비용이 동일하다고 가정할 때, 이차 판별계수 및 분류결과를 출력하기 위한 사용자 함수를 만들고, 이를 이용하여 관측개체의 집단을 추정하여라.

2) 추정된 이차 판별함수를 산점도 상에 나타내어라.

```
> s.data<-read.table('d:/mydata/salmon.txt', skip=4,
+                    header=T)
> s.data
     Area     X1  X2
1   Alaska  108 368
2   Alaska  131 355
3   Alaska  105 469
4   Alaska   86 506
:       :       :   :
89 Canada 144 373
90 Canada 140 388
> # 알래스카와 캐나다 자료
> Al<-subset(s.data,subset=(Area=='Alaska'),
+                 select=c(X1, X2))
> Ca<-subset(s.data,subset=(Area=='Canada'),
+                 select=c(X1, X2))
```

```
> # 공분산 행렬의 동일성 검정
> library(biotools)
> boxM(s.data[,2:3], grouping=s.data[,1])

Box's M-test for Homogeneity of Covariance
Matrices

data:  s.data[, 2:3]
Chi-Sq (approx.)=8.2141, df=3, p-value=0.04179

> detach("package:biotools", unload=TRUE)
```

유의확률=0.04179<0.05=유의수준 이므로
등분산 이라는 귀무가설 기각하므로 이분산

```
> library(MASS)
> result<-qda(s.data[,2:3], grouping=s.data[,1],
+                         prior=c(0.5,0.5))
> result
Call:
qda(s.data[, 2:3], grouping = s.data[, 1], prior = c(0.5, 0.5))

Prior probabilities of groups:
Alaska Canada
   0.5    0.5

Group means:
                X1       X2
Alaska    98.93333 426.5333
Canada   138.06667 369.2444
```

- **이차 판별함수**

$$\hat{y} = -\frac{1}{2}\mathbf{x}^{T}(S_1^{-1} - S_2^{-1})\mathbf{x} + \left(\bar{\mathbf{x}}_1^{T}S_1^{-1} - \bar{\mathbf{x}}_2^{T}S_2^{-1}\right)\mathbf{x}$$

$$= -\frac{1}{2}\mathbf{x}^{T}\begin{pmatrix} 0.00064 & 0.00102 \\ 0.00102 & -0.00031 \end{pmatrix}\mathbf{x} + (0.34424 \quad 0.05234)\mathbf{x}$$

- **분류점**

$$\hat{k} = \frac{1}{2}\ln\left(\frac{|S_1|}{|S_2|}\right) + \frac{1}{2}\left(\bar{\mathbf{x}}_1^{T}\Sigma_1^{-1}\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2^{T}\Sigma_2^{-1}\bar{\mathbf{x}}_2\right)$$

$$= \frac{1}{2}\ln\left(\frac{342297.2}{285786.4}\right) + \frac{1}{2}(237.2089 - 171.2343)$$

$$= 33.0775$$

$$\hat{m} = \hat{k} + \ln\left(\left[\frac{c(1|2)}{c(2|1)}\right]\left[\frac{p_2}{p_1}\right]\right) = 33.0775$$

  - **분류점 보다 크면 Group1(Alaska)로 분류, 그렇지 않으면 Group2(Canada)로 분류**

```
> predict(result)
$class
 [1] Alaska Canada Alaska Alaska Alaska Alaska Alaska
 [8] Alaska Alaska Alaska Alaska Canada Canada Alaska
[15] Alaska Alaska Alaska Alaska Alaska Alaska Alaska
[22] Alaska Alaska Alaska Alaska Alaska Alaska Alaska
[29] Alaska Canada Alaska Alaska Alaska Alaska Alaska
[36] Alaska Alaska Alaska Alaska Alaska Alaska Alaska
[43] Alaska Alaska Alaska Alaska Canada Canada Canada
[50] Canada Canada Canada Canada Canada Canada Canada
[57] Canada Canada Canada Canada Canada Canada Canada
[64] Canada Canada Alaska Canada Canada Canada Canada
[71] Canada Canada Canada Canada Canada Canada Canada
[78] Canada Canada Canada Canada Canada Canada Canada
[85] Canada Canada Canada Canada Canada Canada
Levels: Alaska Canada
```

```
$posterior
             Alaska       Canada
 [1,] 0.5178861536 4.821138e-01
 [2,] 0.0585504743 9.414495e-01
 [3,] 0.9996276639 3.723361e-04
 [4,] 0.9999998682 1.318078e-07
 [5,] 0.9654394131 3.456059e-02
 [6,] 0.9988092586 1.190741e-03
 [7,] 0.9992899706 7.100294e-04
  :           :            :
[87,] 0.1864292133 8.135708e-01
[88,] 0.0991848597 9.008151e-01
[89,] 0.0215347426 9.784653e-01
[90,] 0.0519785674 9.480214e-01
> detach("package:MASS", unload=TRUE)
```

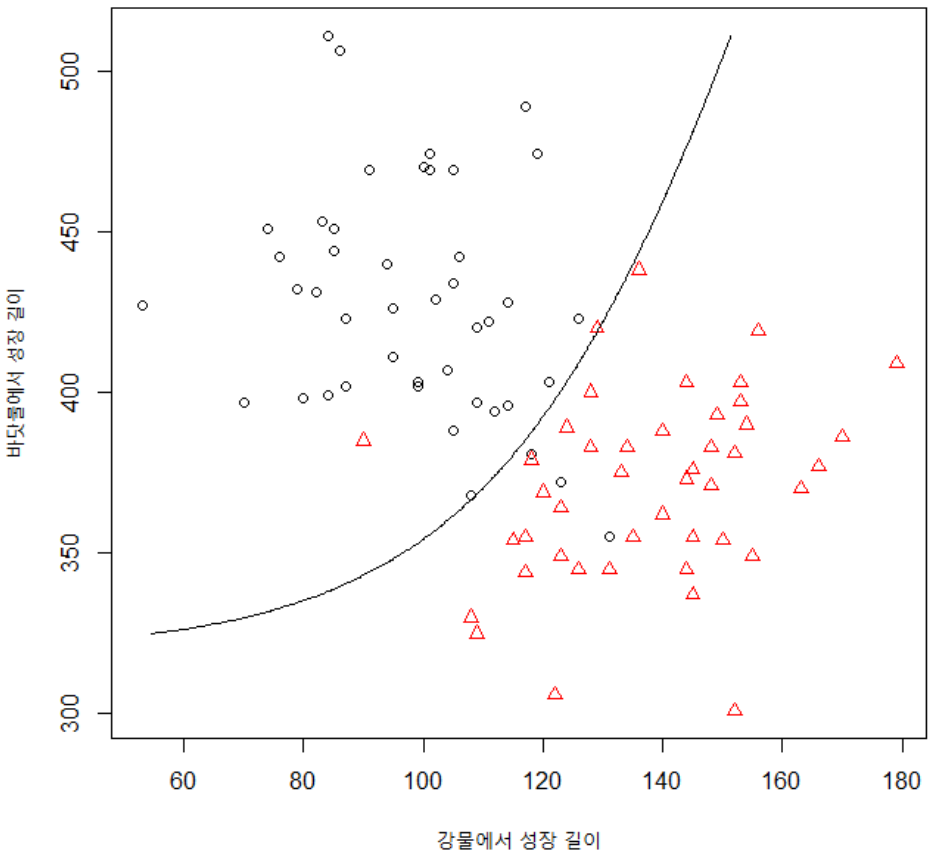
```
> # 이차 판별함수 그래프
> new.X1<-seq(min(s.data$X1), max(s.data$X1), by=0.5)
> new.X2<-seq(min(s.data$X2), max(s.data$X2), by=0.5)
> new.X<-cbind(rep(new.X1, length(new.X2)),
+                    rep(new.X2, each=length(new.X1)))
>
> m1<-colMeans(Al) # 그룹1 평균벡터
> m2<-colMeans(Ca) # 그룹2 평균벡터
> m1; m2
       X1         X2
 98.93333 426.53333
      X1        X2
138.0667 369.2444
>
> S1<-cov(Al) # 그룹1 공분산행렬
> S2<-cov(Ca) # 그룹2 공분산행렬
```

```
> S1; S2
          X1         X2
X1  285.3364 -192.7591
X2 -192.7591 1329.8455
         X1        X2
X1 326.3364 129.8015
X2 129.8015 927.3707
>
> det.S1<-det(S1) # 그룹 1 일반화 분산
> det.S2<-det(S2) # 그룹 2 일반화 분산
> det.S1; det.S2
[1] 342297.2
[1] 285786.4
>
> S1.Inv<-solve(S1) # 그룹1 공분산행렬의 역행렬
> S2.Inv<-solve(S2) # 그룹2 공분산행렬의 역행렬
```

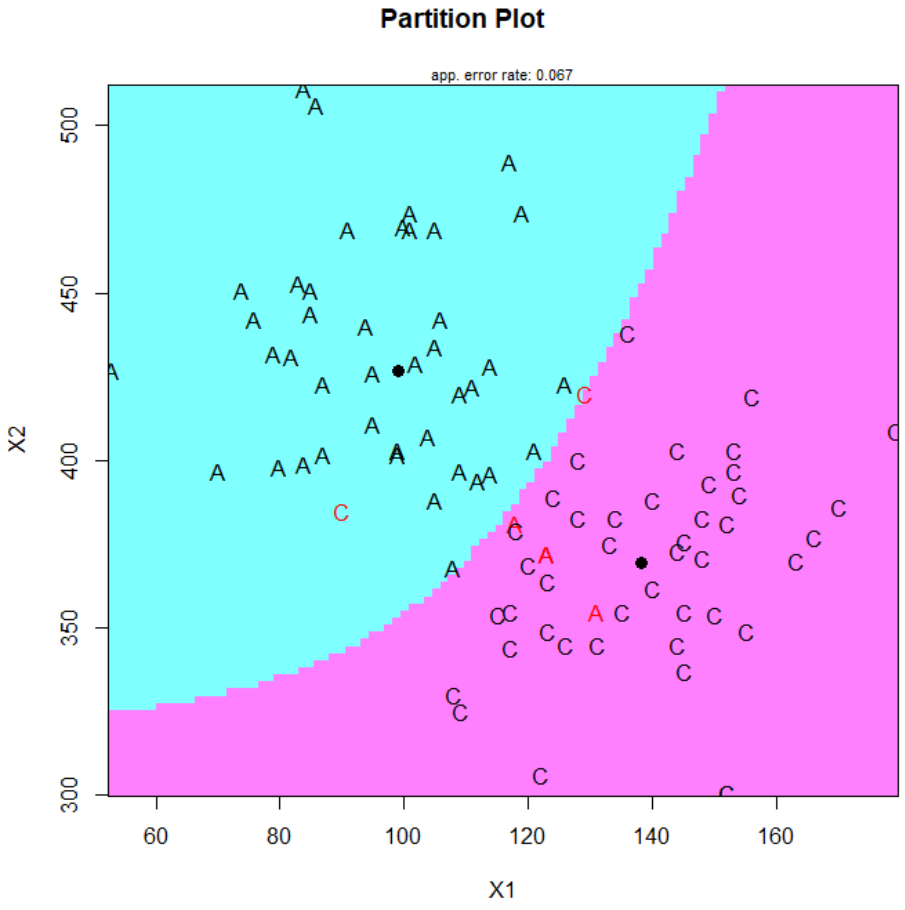
```
> S1.Inv; S2.Inv
             X1            X2
X1 0.0038850609 0.0005631337
X2 0.0005631337 0.0008335925
              X1             X2
X1  0.0032449790 -0.0004541907
X2 -0.0004541907  0.0011418893
>
> # 이차형식분류함수 Qs 항
> eq1<-(S1.Inv-S2.Inv)
> eq2<-(m1%*%S1.Inv-m2%*%S2.Inv)
> eq3<-(1/2)*log(det.S2/det.S1)
> eq4<-(1/2)*(m1%*%S1.Inv%*%m1-
+                     m2%*%S2.Inv%*%m2)
> k<-eq3+eq4
```

```
> eq1; eq2; eq3; eq4; k
              X1             X2
X1 0.0006400819  0.0010173245
X2 0.0010173245 -0.0003082968
             X1       X2
[1,] 0.3442413 0.05234
[1] -0.09021743
          [,1]
[1,] 32.98728
          [,1]
[1,] 32.89706
> score<-c()
> for(i in 1:nrow(new.X)) {
+   score[i]<-(-1/2)*new.X[i,]%*%eq1%*%new.X[i,]+
+                     eq2%*%new.X[i,]-k
+ }
```

```
> score.level<-matrix(score, nrow=length(new.X1),
+                                ncol=length(new.X2))
>
> attach(s.data)
> windows()
> plot(X1, X2, col=as.numeric(Area),
+      pch=as.numeric(Area), xlab='강물에서 성장 길이',
+      ylab='바닷물에서 성장 길이')
> contour(x=new.X1, y=new.X2, z=score.level, levels=0,
+              labels='', add=T)
> detach(s.data)
```



```
> library(klaR)
Warning message:
패키지 'klaR'는 R 버전 3.4.4에서 작성되었습니다
> windows()
> partimat(Area~X2+X1, data=s.data,
+              method="qda", plot.matrix=FALSE,
+              imageplot=TRUE)
> detach("package:klaR", unload=TRUE)
```



오류율의 계산

- 추정된 판별함수의 성능은 오분류 확률 (misclassification probability) 또는 오류율 (error rate)을 이용하여 평가
  - 오분류 확률은 모집단의 분포 형태가 완전히 알려져 있을 경우에 계산 가능
  - 일반적으로 모집단의 분포형태는 알려져 있지 않으므로 판별함수를 이용하여 구할 수 있는 오류율을 사용

- 오류율이란 판별함수에 의해서 잘못 분류된 관측값의 비율로 정의

실제 \ 예측	G_1	G_2	계
G_1	n_{1C}	n_{1M}	n_1
G_2	n_{2M}	n_{2C}	n_2

  - n_1, n_2 : 모집단 G_1과 G_2로 부터 추출된 표본의 수
  - n_{1M} : 실제 G_1인데 G_2로 분류한 표본의 수
  - n_{2M} : 실제 G_2인데 G_1로 분류한 표본의 수

$$\text{오류율} = \frac{n_{1M} + n_{2M}}{n_1 + n_2}$$

- 재대입법(resubstitution)
  - 오류율을 계산하기 위하여 판별함수의 추정에서 사용된 표본인 훈련 자료(training data)를 사용하여 오류율을 추정하는 방법
  - 이 방법으로 계산한 오류율은 명백한 오류율(apparent error rate : APER)라고 하며, 실제보다 낮게 추정
  - 따라서 좋지 않은 판별함수임에도 좋은 판별함수인 것 처럼 보임

- 재대입법의 문제점을 보완하기 위한 방법으로 전체 표본을 훈련 자료와 타당성 자료로 나누어 훈련 자료로 판별함수를 도출하고, 이 도출된 판별함수를 이용하여 타당성 자료로 오류율을 계산하는데 사용
  - 오류율에 대한 편이(bias) 문제를 해결
  - 두 자료로 구분해야 하므로 많은 표본이 필요
  - 전체가 아닌 일부 표본을 사용하므로 실제 구하고자 하는 판별함수와 차이가 있을 수 있음

- 판별함수를 보다 더 정확하게 평가하기 위해 교차타당성법(cross validation)을 사용
  - 대표적인 교차타당성법은 한 개씩 제거 방법(leave-one-out cross validation)과 k-폴더 방법(k-folder cross validation)
  - 한 개씩 제거 방법
    - 전체 표본의 수만큼 집단을 나누어 판별함수의 타당성 정도를 검사하는 방법
    - 한 개의 표본만 제거하고 나머지 표본으로 판별함수를 구하고, 이 판별함수로 제거되었던 표본을 예측하여 오류율을 계산

  - k-폴더 방법
    - 표본을 동일한 크기의 k개의 부분집합으로 나누고 하나의 부분집합을 제외하고 나머지 부분집합들로 판별함수를 구하고, 이 판별함수로 제외했던 부분집합에 대해 오류율을 계산하는 과정을 k번 반복
    - k가 전체 자료수가 되면 한 개씩 제거 방법과 동일

- 교차타당성법은 전체 표본수가 적어서 전체자료를 훈련 자료와 타당성 자료로 분할하기 어려울 때 유용하게 사용되는 방법
  - 모든 자료가 훈련자료가 되면서 또한 타당성 자료가 됨
  - k가 크면 여러 번 반복해야 한다는 단점
  - 그래도 다른 방법보다도 정확한 판별함수의 평가가 가능하다는 장점

- 표본추출함수

```
> sample(x, size, replace=FALSE, prob=NULL)
```

  - x : 표본으로 추출할 모집단 값
  - size : 추출할 표본의 크기
  - replace : 복원추출 여부
    - 복원추출(TRUE), 비복원추출(FALSE)
  - prob : 모집단 값의 확률

예제 1.12

- 연어는 강의 상류천에서 부화한 후 바다로 나아가 생활하게 된다. 그러다 산란기가 되면 알을 낳기 위하여 다시 자신이 태어난 곳으로 되돌아와 산란 후 최후의 죽음을 맞이하게 된다. 아래 표 (salmon.txt) 는 알래스카와 캐나다 두 지역에서 부화한 연어의 크기를 측정한 결과로서 X_1은 강물에서, X_2는 바다물에서 성장한 길이를 각각 나타낸다.
  두 집단에 대한 확률벡터 $X=(X_1, X_2)^T$가 각각 다변량 정규분포 $N_p(\mu_i, \Sigma_i)$, i=1,2를 따른다고 할 때, 다음 물음에 답하여라.

알래스카(group1)		캐나다(group2)		알래스카(group1)		캐나다(group2)	
X1	X2	X1	X2	X1	X2	X1	X2
108	368	129	420	102	429	145	376
131	355	148	371	101	469	115	354
105	469	179	409	85	444	134	383
86	506	152	381	109	397	117	355
99	402	166	377	106	442	126	345
87	423	124	389	82	431	118	379
94	440	156	419	118	381	120	369
117	489	131	345	105	388	153	403
79	432	140	362	121	403	150	354
99	403	144	345	85	451	154	390
114	428	149	393	83	453	155	349
123	372	108	330	53	427	109	325
123	372	135	355	95	411	117	344
109	420	170	386	76	442	128	400
112	394	152	301	95	426	144	403
104	407	153	397	87	402	163	370
111	422	152	301	70	397	145	355
126	423	136	438	84	511	133	375
105	434	122	306	91	469	128	383
119	474	148	383	74	451	123	349
114	396	90	385	101	474	144	373
100	470	145	337	80	398	140	388
84	399	123	364				

1) 훈련 자료(60%)와 타당성 자료(40%)로 나누어서 훈련자료로 판별함수를 구하고, 이 판별함수로 타당성 자료에 대한 분류표를 출력하여라.

2) 한 개씩 제거 교차타당성법을 이용하여 이차 판별함수의 분류표를 출력하여라.

```
> s.data<-read.table('d:/mydata/Exam07_05.txt', skip=4,
+                    header=T)
> s.data
     Area      X1   X2
1   Alaska  108 368
2   Alaska  131 355
3   Alaska  105 469
4   Alaska   86 506
:       :       :   :
89 Canada 144 373
90 Canada 140 388
> # 알래스카와 캐나다 자료
> Al<-subset(s.data,subset=(Area=='Alaska'),
+                    select=c(X1, X2))
> Ca<-subset(s.data,subset=(Area=='Canada'),
+                    select=c(X1, X2))
```

```
> # 알래스카 훈련 자료 index
> Al.index<-sort(sample(1:45, size=27))
> # 캐나다 훈련 자료 index
> Ca.index<-sort(sample(46:90, size=27))
>
> Al.index; Ca.index # 훈련 자료 index
 [1]  2  4  5  7  8 11 12 14 17 19 20 21 22 27 28 29 32
[18] 33 34 35 36 37 39 40 41 43 44
 [1] 46 49 50 52 54 55 56 58 61 63 64 67 68 70 71 73
[17] 74 76 78 79 81 82 84 85 88 89 90
>
> tra.data<-s.data[c(Al.index, Ca.index), ] # 훈련 자료
> val.data<-s.data[-c(Al.index, Ca.index), ] # 타당성 자료
```

```
> tra.data
       Area    X1  X2
2    Alaska  131 355
4    Alaska   86 506
:       :       :    :
89 Canada 144 373
90 Canada 140 388
> val.data
        Area   X1  X2
1    Alaska  108 368
3    Alaska  105 469
6    Alaska   87 423
:       :       :    :
86 Canada 133 375
87 Canada 128 383
```

```
> # MASS 패키지의 qda() 함수
> library(MASS)
> result<-qda(tra.data[,2:3],grouping=tra.data[,1],
+                         prior=c(0.5,0.5))
> result
Call:
qda(tra.data[, 2:3], grouping = tra.data[, 1], prior = c(0.5,
0.5))

Prior probabilities of groups:
Alaska Canada
   0.5    0.5

Group means:
                  X1          X2
Alaska     98.07407 433.3333
Canada 139.44444 373.1852
```

```
> val.Group<-predict(result, newdata=val.data[,2:3])
> detach("package:MASS", unload=TRUE)
>
> table(val.data$Area, val.Group$class)

                Alaska Canada
  Alaska         16          2
  Canada         1          17
```

```
> library(MASS)
> result<-qda(s.data[,2:3], grouping=s.data[,1], prior=c(0.5,0.5),
+                   CV=T)
> result
$class
 [1] Canada Canada Alaska Alaska Alaska Alaska Alaska Alaska
 [9] Alaska Alaska Alaska Canada Canada Alaska Alaska Alaska
[17] Alaska Alaska Alaska Alaska Alaska Alaska Alaska Alaska
[25] Alaska Alaska Alaska Alaska Alaska Canada Alaska Alaska
[33] Alaska Alaska Alaska Alaska Alaska Alaska Alaska Alaska
[41] Alaska Alaska Alaska Alaska Alaska Alaska Canada Canada
[49] Canada Canada Canada Canada Canada Canada Canada Canada
[57] Canada Canada Canada Canada Canada Canada Alaska Canada
[65] Canada Alaska Canada Canada Canada Canada Canada Canada
[73] Canada Canada Canada Canada Canada Canada Canada Canada
[81] Canada Canada Canada Canada Canada Canada Canada Canada
[89] Canada Canada
Levels: Alaska Canada
```

```
$posterior
             Alaska       Canada
 [1,] 0.4917262123 5.082738e-01
 [2,] 0.0379120685 9.620879e-01
 [3,] 0.9996014335 3.985665e-04
 [4,] 0.9999998186 1.814028e-07
 [5,] 0.9645716361 3.542836e-02
        :            :            :
[87,] 0.1907143631 8.092856e-01
[88,] 0.1021572439 8.978428e-01
[89,] 0.0220263563 9.779736e-01
[90,] 0.0531996690 9.468003e-01

$call
qda(x=s.data[, 2:3], grouping=s.data[, 1], prior=c(0.5, 0.5),
CV=T)
```

```
> table(s.data$Area, result$class)

                Alaska Canada
  Alaska        40      5
  Canada        3       42
> detach("package:MASS", unload=TRUE)
```

예제 1.13

- 다음 붓꽃(iris.csv)의 종(Species)을 판별분석을 이용하여 분류하시오.
- 등분산성 검정 및 변수선택방법을 사용하여 최종 판별함수 결정
  - 종 : setosa, versicolor, virginica

```
> iris.data<-read.csv("iris.csv")
> head(iris.data)
  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
```

```
> library(biotools)
---
biotools version 3.1

Warning message:
패키지 'biotools'는 R 버전 3.4.4에서 작성되었습니다
> boxM(iris.data[1:4], grouping=iris.data$Species)

Box's M-test for Homogeneity of Covariance Matrices

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

> detach("package:biotools", unload=TRUE)
```

```
> # MASS 패키지의 qda() 함수
> library(MASS)
> result<-qda(Species~., data=iris.data)
> result
Call:
qda(Species ~ ., data = iris.data)

Prior probabilities of groups:
    setosa versicolor  virginica
 0.3333333  0.3333333  0.3333333

Group means:
           Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa            5.006       3.428        1.462       0.246
versicolor        5.936       2.770        4.260       1.326
virginica         6.588       2.974        5.552       2.026
> detach("package:MASS", unload=TRUE)
```

```
> # 변수선택
> library(klaR) # klaR 패키지 실행
필요한 패키지를 로딩중입니다: MASS
Warning message:
패키지 'klaR'는 R 버전 3.4.4에서 작성되었습니다
> step.result<-stepclass(Species~., data=iris.data,
+                        method="qda", improvement=0.001)
 `stepwise classification', using 10-fold cross-validated correctness rate of
method qda'.
150 observations of 4 variables in 3 classes; direction: both
stop criterion: improvement less than 0.1%.
correctness rate: 0.96;  in: "Petal.Width";  variables (1): Petal.Width
correctness rate: 0.96667;  in: "Petal.Length";  variables (2): Petal.Width,
Petal.Length
correctness rate: 0.97333;  in: "Sepal.Length";  variables (3): Petal.Width,
Petal.Length, Sepal.Length ◀ 마지막 모형이 최종

 hr.elapsed min.elapsed sec.elapsed
       0.00        0.00        0.27
```

```
> step.result
method       : qda
final model : Species ~ Sepal.Length + Petal.Length +
Petal.Width
<environment: 0x0000000024acc748>

correctness rate = 0.9733
> detach("package:klaR", unload=TRUE)
>
> # MASS 패키지의 qda() 함수
> library(MASS)
> step.result$formula
Species ~ Sepal.Length + Petal.Length + Petal.Width
<environment: 0x0000000024acc748>
> result<-qda(step.result$formula, data=iris.data)
> detach("package:MASS", unload=TRUE)
```

```
> new.g<-predict(result)$class # 예측범주
>
> # 예측률
> xtabs(~new.g+iris.data$Species)  # 분류표
             iris.data$Species
new.g         setosa versicolor virginica
  setosa          50          0         0
  versicolor       0         48         1
  virginica        0          2        49
> sum(new.g==iris.data$Species)/NROW(iris.data)
[1] 0.98
```

```
> library(klaR)
필요한 패키지를 로딩중입니다: MASS
Warning message:
패키지 'klaR'는 R 버전 3.4.4에서 작성되었습니다
> windows()
> partimat(step.result$formula, data=iris.data,
+                 method="qda", plot.matrix=TRUE,
+                 imageplot=TRUE)
> detach("package:klaR", unload=TRUE)
```