연세대학교 통계 데이터 사이언스 학회 ESC 23-2 FALL WEEK4

Discriminant Analysis

[ESC 정규세션 임원진] 김시은 박정현

1. Introduction & 2. separation and classification for two populations

1. introduction

Discrimination :

두개 이상의 모집단에서 추출된 다변량 관측치들의 정보를 이용해, 다변량 관측치들(x) 이 어느 모집단에서 추출된 것인가를 결정해줄 수 있는 기준(discriminant function)을 찾는 분석.

Classification :

cost, prior probability 등 고려해 결정된 discrimination function을 통해,
새로 주어진 다변량 관측치가 어떤 모집단에서 추출된 것인가를 결정(분류) 하는 분석

-> 실제로는 상호보완적, 구분 모호

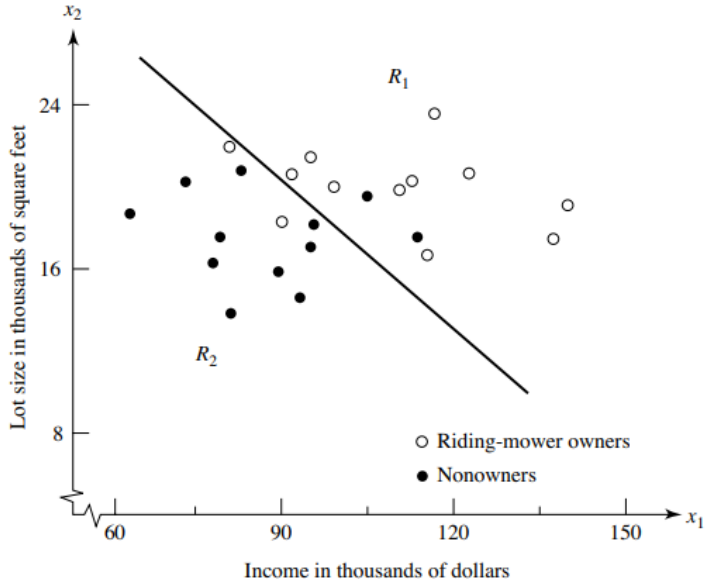
ECM , TPM ··· , separation 등의 기준

↓ (정규분포 , 공통공분산행렬 등 분포에 대한 가정)

discrimination rule 식 도출

2. Separation and classification for two populations

P개의 associated random variabes로 이루어진 $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ 를 basis로 하는 sample space Ω에서, discrimination rule 에 따라 새로운 개체 주어졌을 때 어느 모집단에서 추출된 것인지 분류



R_1에 **x** 존재 => π_1로 분류

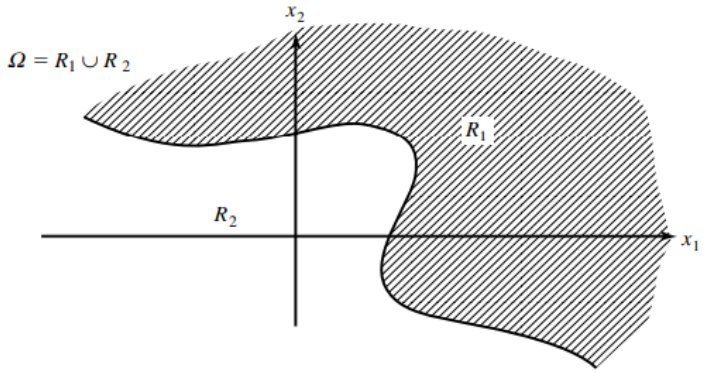
R_2에 **x** 존재 => π_2로 분류

$$R_1: \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$\begin{pmatrix}\text{density}\\ \text{ratio}\end{pmatrix} \geq \begin{pmatrix}\text{cost}\\ \text{ratio}\end{pmatrix}\begin{pmatrix}\text{prior}\\ \text{probability}\\ \text{ratio}\end{pmatrix}$$

$$R_2: \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$\begin{pmatrix}\text{density}\\ \text{ratio}\end{pmatrix} < \begin{pmatrix}\text{cost}\\ \text{ratio}\end{pmatrix}\begin{pmatrix}\text{prior}\\ \text{probability}\\ \text{ratio}\end{pmatrix}$$



2. Separation and classification for two populations

기본적인 분류 방법 : ECM을 최소화하는 discrimination function에 의한 분류

- expected cost of misclassification $\mathrm{ECM} = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2$

 - 정확하게 분류하기 위해, ECM을 구할 때 집단 각각의 사전확률과 cost 고려함.

- Minimum expected cost regions

$$R_1: \ \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$\begin{pmatrix}\text{density}\\ \text{ratio}\end{pmatrix} \geq \begin{pmatrix}\text{cost}\\ \text{ratio}\end{pmatrix}\begin{pmatrix}\text{prior}\\ \text{probability}\\ \text{ratio}\end{pmatrix}$$

$$R_2: \ \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$\begin{pmatrix}\text{density}\\ \text{ratio}\end{pmatrix} < \begin{pmatrix}\text{cost}\\ \text{ratio}\end{pmatrix}\begin{pmatrix}\text{prior}\\ \text{probability}\\ \text{ratio}\end{pmatrix}$$

EX 11.2)

$p_1 = 0.8, p_2 = 0.2$, $c(2|1) = 5$ units, $c(1|2) = 10$ units

$f_1(x_0) = 0.3, f_2(x_0) = 0.4$ (evaluated)

2. Separation and classification for two populations

1. 사전확률 " prior probabilities of occurrence"

   cf) 파산기업vs 건전기업

   - $p_1 + p_2 = 1$
   - $P(1|2) = P(\mathbf{X} \in R_1 | \pi_2) = \int_{R_1} f_2(\mathbf{x})\, d\mathbf{x}$

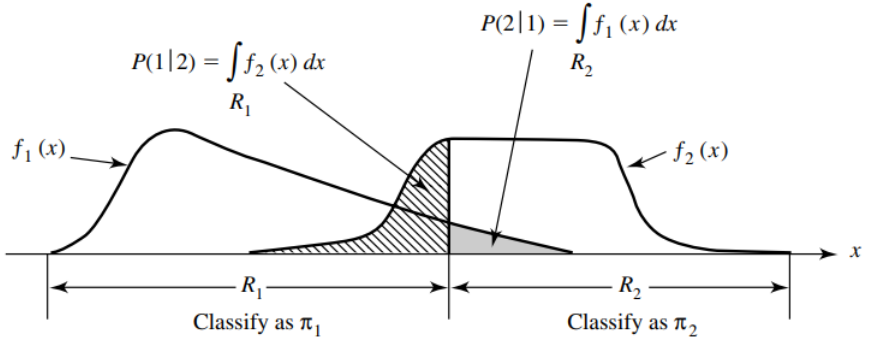


Figure 11.3 Misclassification probabilities for hypothetical classification regions when $p = 1$.

→ $P(\text{observation is misclassified as } \pi_1) = P(\text{observation comes from } \pi_2 \text{ and is misclassified as } \pi_1)$
$= P(\mathbf{X} \in R_1 | \pi_2) P(\pi_2) = P(1|2)p_2$

2. Cost

   cf) 건강한 사람을 심각한 질병이 있다고 잘못 분류

   vs 심각한 질병이 있는 사람을 건강하다고 잘못 분류

		Classify as:	
		π_1	π_2
True population:	π_1	0	$c(2\|1)$
	π_2	$c(1\|2)$	0

2. Separation and classification for two populations

- special cases of minimum expected cost regions

(a) $p_2/p_1 = 1$ (equal prior probabilities)

$$R_1: \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1|2)}{c(2|1)} \qquad R_2: \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{c(1|2)}{c(2|1)}$$

(b) $c(1|2)/c(2|1) = 1$ (equal misclassification costs)

$$R_1: \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{p_2}{p_1} \qquad R_2: \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{p_2}{p_1}$$

(c) $p_2/p_1 = c(1|2)/c(2|1) = 1$ or $p_2/p_1 = 1/(c(1|2)/c(2|1))$
(equal prior probabilities and equal misclassification costs)

$$R_1: \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq 1 \qquad R_2: \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < 1$$

Minimum TPM & largest posterior probability & Minimum ECM(equal misclassification costs) 통한 분류는 모두 같은 결과 나타냄.

cf) discrimination function을 구하는데 사용되는 또 다른 기준 :

TPM (Total probability of misclassification) & posterior probability

$$\begin{aligned}\text{TPM} &= P(\text{misclassifying a } \pi_1 \text{ observation } \textit{or} \text{ misclassifying a } \pi_2 \text{ observation}) \\ &= P(\text{observation comes from } \pi_1 \text{ and is misclassified}) \\ &\quad + P(\text{observation comes from } \pi_2 \text{ and is misclassified}) \\ &= p_1 \int_{R_2} f_1(\mathbf{x})\, d\mathbf{x} + p_2 \int_{R_1} f_2(\mathbf{x})\, d\mathbf{x} \qquad (11\end{aligned}$$

$$\begin{aligned}P(\pi_1|\mathbf{x}_0) &= \frac{P(\pi_1 \text{ occurs and we observe } \mathbf{x}_0)}{P(\text{we observe } \mathbf{x}_0)} \\ &= \frac{P(\text{we observe } \mathbf{x}_0|\pi_1)P(\pi_1)}{P(\text{we observe } \mathbf{x}_0|\pi_1)P(\pi_1) + P(\text{we observe } \mathbf{x}_0|\pi_2)P(\pi_2)} \\ &= \frac{p_1 f_1(\mathbf{x}_0)}{p_1 f_1(\mathbf{x}_0) + p_2 f_2(\mathbf{x}_0)}\end{aligned}$$

$$P(\pi_2|\mathbf{x}_0) = 1 - P(\pi_1|\mathbf{x}_0) = \frac{p_2 f_2(\mathbf{x}_0)}{p_1 f_1(\mathbf{x}_0) + p_2 f_2(\mathbf{x}_0)} \qquad (11\text{-}9)$$

3. Classification with two multivariate normal populations

3. Classification with two multivariate normal populations

- 다변량 정규분포

$\mathbf{X}' = [X_1, X_2, \cdots, X_p]$

$\mathbf{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$ $\quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix}$, $\quad \Sigma = \begin{bmatrix} \sigma_{11}^2 & \cdots & \sigma_{1p}^2 \\ \vdots & \ddots & \\ \sigma_{p1}^2 & \cdots & \sigma_{pp}^2 \end{bmatrix}$

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

Σ : Symmetric & positive definite

↳ Σ^{-1} 무조건 존재

- 모수 추정

$$\underset{(n_1 \times p)}{\mathbf{X}_1} = \begin{bmatrix} \mathbf{x}_{11}' \\ \mathbf{x}_{12}' \\ \vdots \\ \mathbf{x}_{1n_1}' \end{bmatrix}$$

$$\underset{(p\times 1)}{\bar{\mathbf{x}}_1} = \frac{1}{n_1}\sum_{j=1}^{n_1}\mathbf{x}_{1j}, \quad \underset{(p\times p)}{\mathbf{S}_1} = \frac{1}{n_1-1}\sum_{j=1}^{n_1}(\mathbf{x}_{1j}-\bar{\mathbf{x}}_1)(\mathbf{x}_{1j}-\bar{\mathbf{x}}_1)'$$

$$\underset{(p\times 1)}{\bar{\mathbf{x}}_2} = \frac{1}{n_2}\sum_{j=1}^{n_2}\mathbf{x}_{2j}, \quad \underset{(p\times p)}{\mathbf{S}_2} = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(\mathbf{x}_{2j}-\bar{\mathbf{x}}_2)(\mathbf{x}_{2j}-\bar{\mathbf{x}}_2)'$$

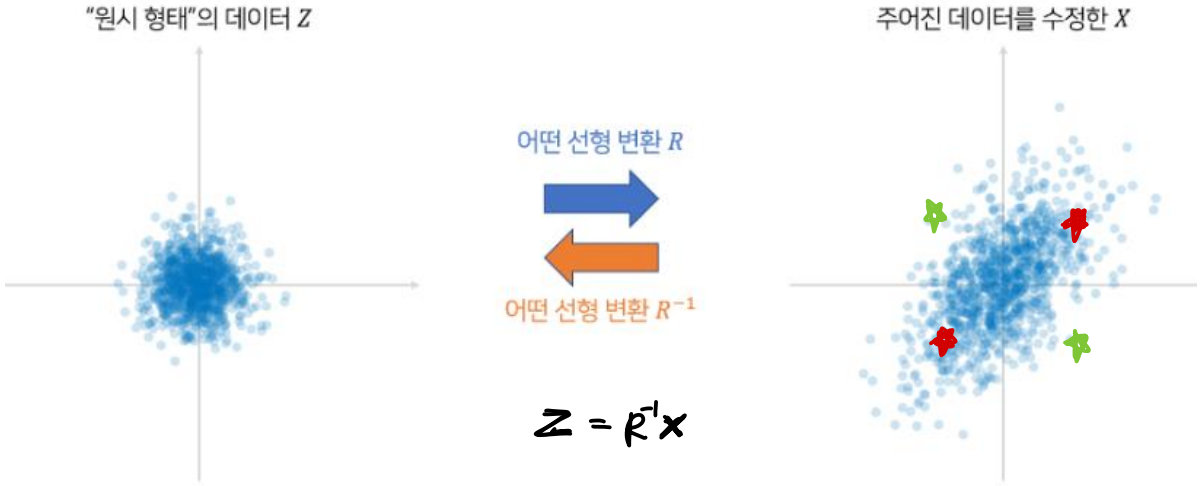
$$\underset{(n_2 \times p)}{\mathbf{X}_2} = \begin{bmatrix} \mathbf{x}_{21}' \\ \mathbf{x}_{22}' \\ \vdots \\ \mathbf{x}_{2n_2}' \end{bmatrix}$$

$$\mathbf{S}_{\text{pooled}} = \left[\frac{n_1-1}{(n_1-1)+(n_2-1)}\right]\mathbf{S}_1 + \left[\frac{n_2-1}{(n_1-1)+(n_2-1)}\right]\mathbf{S}_2$$

3. Classification with two multivariate normal populations

- Mahalanobis distance

$X = RZ$



$Z = R^{-1}X$

$$d_z = \sqrt{Z^T Z} = \sqrt{X^T (R^{-1})^T R^{-1} X} = \sqrt{X^T (RR^T)^{-1} X} = \sqrt{X^T \Sigma^{-1} X}$$

ⓥ 유클리드 거리

$$d_E = \sqrt{(x-y)'(x-y)}$$

$\bar{x}_1$, $\bar{x}_2$

$(p \times 1)$ $(p \times 1)$

0 0

ⓥ 마할라노비스 거리

$$d_z = \sqrt{(x-y)'\Sigma^{-1}(x-y)}$$

↑

표준편차

정규화해서 구한 거리 (분포 고려해 정규화한 후의 거리)

3. Classification with two multivariate normal populations

- Minimum ECM classification rule : (정규분포, 등분산성 , 모수 known 가정)

Allocate $\mathbf{x}_0$ to π_1 if

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x}_0 - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

Allocate $\mathbf{x}_0$ to π_2 otherwise.

Cf) Minimum ECM region

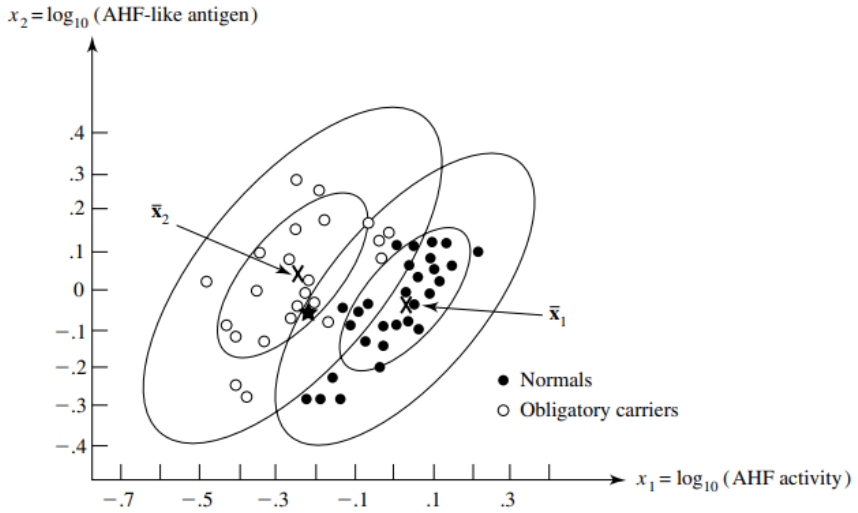
$$R_1: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

- 도출과정

$$\frac{f_1(x)}{f_2(x)} = \frac{(2\pi)^{p/2}|\Sigma|^{1/2}}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2)\right]$$

$$R_1: \quad -\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2) \geq \ln\left[\frac{c(1|2)}{c(2|1)} \cdot \frac{p_2}{p_1}\right]$$

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



3. Classification with two multivariate normal populations

- Estimated Minimum ECM rule (정규분포,등분산성 가정/ 모수 추정)

Allocate $\mathbf{x}_0$ to π_1 if

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \quad (11\text{-}18)$$

Allocate $\mathbf{x}_0$ to π_2 otherwise.

- special case : $\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) = 1$

$$\underbrace{(\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} x_0}_{= \hat{y}} \geq \underbrace{\frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} (\bar{x}_1 + \bar{x}_2)}_{\hat{m}}$$

An Allocation Rule Based on Fisher's Discriminant Function[5]

Allocate $\mathbf{x}_0$ to π_1 if

$$\hat{y}_0 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x}_0$$
$$\geq \hat{m} = \tfrac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

3. Classification with two multivariate normal populations

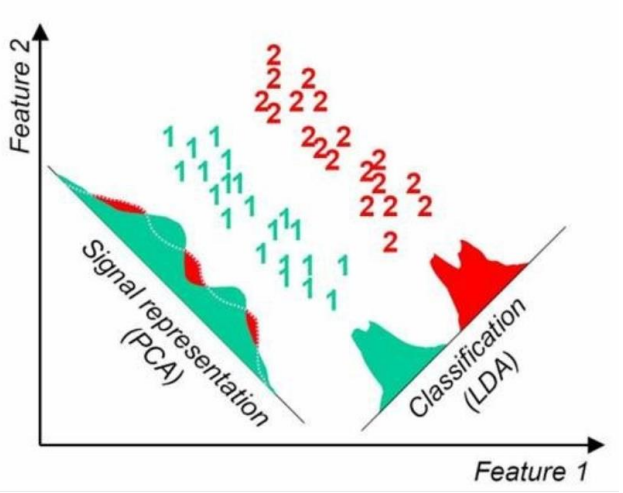
fisher's discriminant function 에 의한 분류

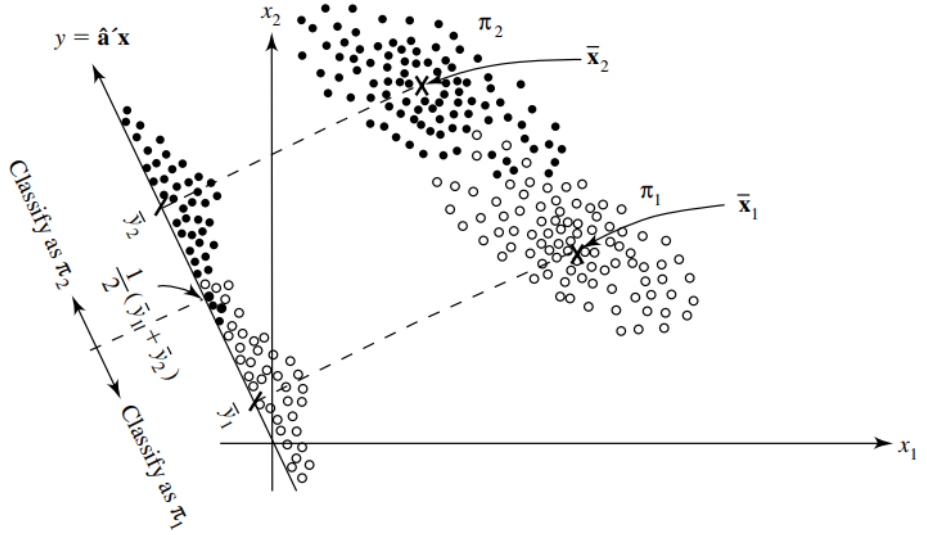
- multivariate observations x -> univaraiate observations y (linear combinations of x)

-projection 후에, $\bar{y}_1, \bar{y}_2$의 separation이 가장 크도록 하는 y를 찾으려는 것이 목적

$$\text{separation} = \frac{|\bar{y}_1 - \bar{y}_2|}{s_y}, \quad \text{where } s_y^2 = \frac{\sum_{j=1}^{n_1}(y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_2)^2}{n_1 + n_2 - 2}$$

- 정규분포 가정 x, 등분산성 가정 o





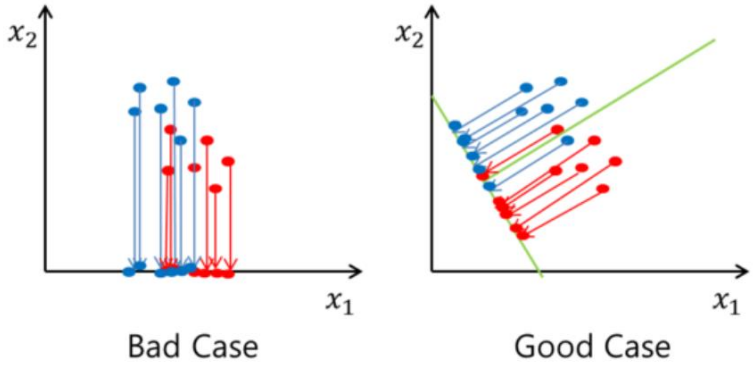
3. Classification with two multivariate normal populations

- fisher's discriminant function 도출과정

$$\frac{\left(\begin{array}{c}\text{squared distance}\\ \text{between sample means of } y\end{array}\right)}{(\text{sample variance of } y)} = \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2}$$

$$= \frac{(\hat{\mathbf{a}}'\bar{\mathbf{x}}_1 - \hat{\mathbf{a}}'\bar{\mathbf{x}}_2)^2}{\hat{\mathbf{a}}'\mathbf{S}_{\text{pooled}}\hat{\mathbf{a}}}$$

$$= \frac{(\hat{\mathbf{a}}'\mathbf{d})^2}{\hat{\mathbf{a}}'\mathbf{S}_{\text{pooled}}\hat{\mathbf{a}}}$$



$$\max_{\hat{\mathbf{a}}} \frac{(\hat{\mathbf{a}}'\mathbf{d})^2}{\hat{\mathbf{a}}'\mathbf{S}_{\text{pooled}}\hat{\mathbf{a}}} = \mathbf{d}'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{d} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = D^2$$

$$D^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

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

3. Classification with two multivariate normal populations

An Allocation Rule Based on Fisher's Discriminant Function[5]

Allocate $\mathbf{x}_0$ to π_1 if

$$\hat{y}_0 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x}_0 \geq \hat{m} = \tfrac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

or

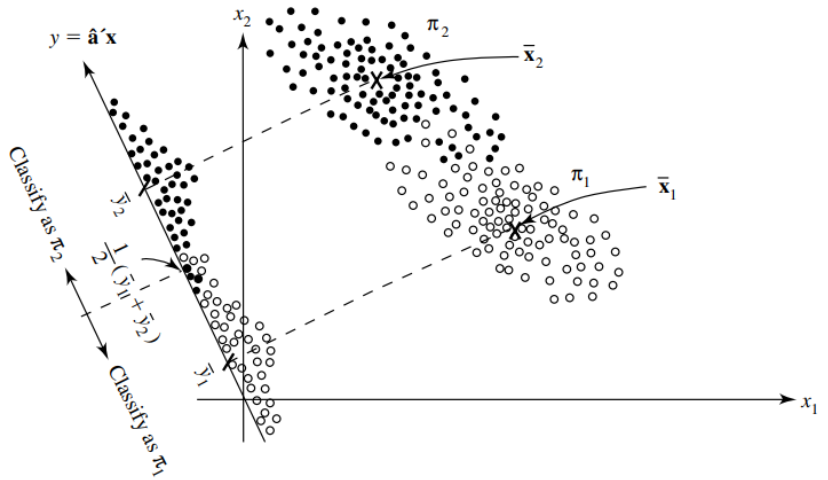
$$\hat{y}_0 - \hat{m} \geq 0 \qquad (11\text{-}25)$$

Allocate $\mathbf{x}_0$ to π_2 if

$$\hat{y}_0 < \hat{m}$$

or

$$\hat{y}_0 - \hat{m} < 0$$



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

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

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

$$\hat{m} = \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) = \frac{1}{2}(\bar{y}_1 + \bar{y}_2)$$

$$\hat{y} = \hat{\mathbf{a}}'\mathbf{x} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x} = 37.61x_1 - 28.92x_2$$

$$D^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = [.2418, \quad -.0652]\begin{bmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{bmatrix}\begin{bmatrix} .2418 \\ -.0652 \end{bmatrix} = 10.98$$

3. Classification with two multivariate normal populations

- Scailing

1. unit length

$$\hat{\mathbf{a}}^* = \frac{\hat{\mathbf{a}}}{\sqrt{\hat{\mathbf{a}}'\hat{\mathbf{a}}}}$$

2. 첫번째 원소가 1

$$\hat{\mathbf{a}}^* = \frac{\hat{\mathbf{a}}}{\hat{a}_1}$$

- 두 모평균 차이에 대한 검정

모평균이 유의미하게 달라야 거리로 분류하는 것이 의미가 있음.

a test of $H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ versus $H_1: \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$

$$\left(\frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p}\right)\left(\frac{n_1 n_2}{n_1 + n_2}\right)D^2$$

$$\sim F(p, n_1 + n_2 - p - 1)$$

다변량 정규분포, 등분산성 가정

3. Classification with two multivariate normal populations

두 집단의 covariance matrices가 다른 경우
=> x에 대한 quadratic functions를 이용해 classification regions 분류

Allocate $\mathbf{x}_0$ to π_1 if

$$-\frac{1}{2}\mathbf{x}_0'(\mathbf{S}_1^{-1}-\mathbf{S}_2^{-1})\mathbf{x}_0+(\bar{\mathbf{x}}_1'\mathbf{S}_1^{-1}-\bar{\mathbf{x}}_2'\mathbf{S}_2^{-1})\mathbf{x}_0-k\geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \quad (11\text{-}29)$$

w
constant

Allocate $\mathbf{x}_0$ to π_2 otherwise.

ln

$$\frac{f_1(x)}{f_2(x)}=\frac{(2\pi)^{p/2}|\Sigma_2|^{1/2}}{(2\pi)^{p/2}|\Sigma_1|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)+\frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2)\right]$$

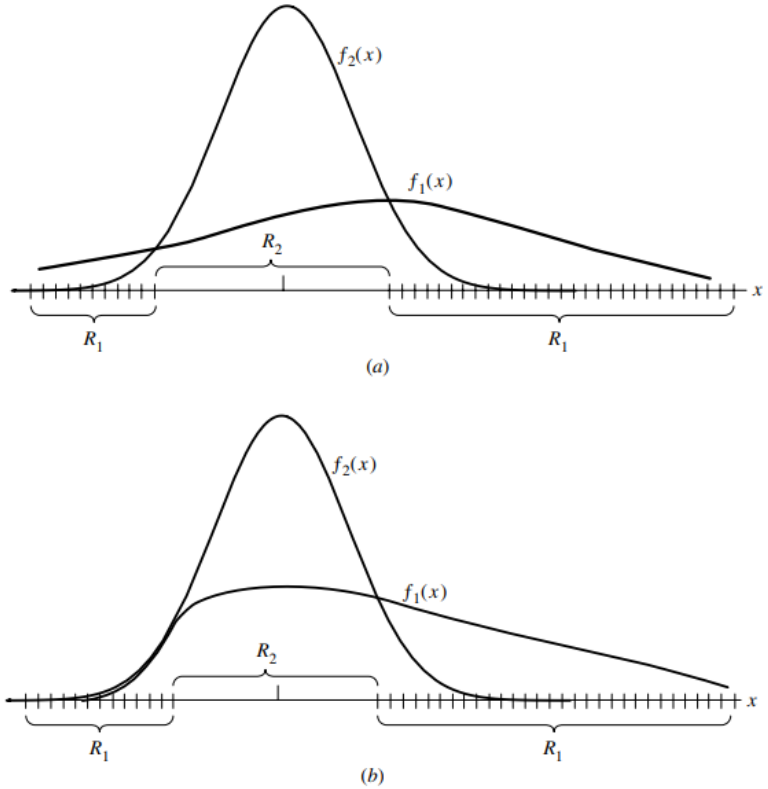


Figure 11.6 Quadratic rules for (a) two normal distribution with unequal variances and (b) two distributions, one of which is nonnormal—rule not appropriate.

- 등분산성 x -> 정규분포 가정 더 중요함.

4. evaluating classification functions

4. evaluating classification functions

- OER : error rate for the minimum TPM classification rule

$$\text{Optimum error rate (OER)} = p_1 \int_{R_2} f_1(\mathbf{x})\, d\mathbf{x} + p_2 \int_{R_1} f_2(\mathbf{x})\, d\mathbf{x} \qquad (11\text{-}30)$$

where R_1 and R_2 are determined by case (b) in (11-7).

(b) $c(1|2)/c(2|1) = 1$ (equal misclassification costs)

$$R_1: \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{p_2}{p_1} \qquad R_2: \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{p_2}{p_1} \qquad (11\text{-}7)$$

-ex.11.5

$p_1 = p_2 = 0.5$ 가정 시

minimum TPM rule 에서의 error rate 인 OER

$= \Phi\left(-\frac{\Delta}{2}\right)$;

Δ^2: fisher 에 따라 y로 변환 시의 각 집단의 분산 Δ_y^2

4. evaluating classification functions

- Actual error rate (AER) 추정 방법 두가지 $$\text{AER} = p_1 \int_{\hat{R}_2} f_1(\mathbf{x})\, d\mathbf{x} + p_2 \int_{\hat{R}_1} f_2(\mathbf{x})\, d\mathbf{x}$$

1. APER(apparent error rate) : 가정 필요 x, AER을 과소평가하는 경향

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

		Predicted membership: π_1: riding-mower owners	Predicted membership: π_2: nonowners	
Actual membership: π_1: riding-mower owners		$n_{1C} = 10$	$n_{1M} = 2$	$n_1 = 12$
Actual membership: π_2: nonowners		$n_{2M} = 2$	$n_{2C} = 10$	$n_2 = 12$

2. Lachenbruch's holdout procedure 로 구한 $\hat{E}(AER)$
   :nearly unbiased estimate

$$\hat{E}(\text{AER}) = \frac{n_{1M}^{(H)} + n_{2M}^{(H)}}{n_1 + n_2}$$

$$\hat{P}(2|1) = \frac{n_{1M}^{(H)}}{n_1}$$

$$\hat{P}(1|2) = \frac{n_{2M}^{(H)}}{n_2}$$

5. Classification with several populations

5. Classification with several populations

Minimum ECM classification rule

-conditional expected cost of misclassifying x from π_1 into $\pi_2, \ldots, \pi_g$

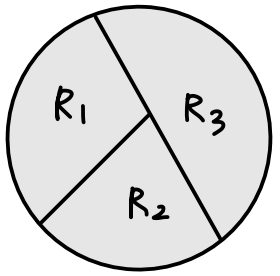
$$\begin{aligned} \text{ECM}(1) &= P(2|1)c(2|1) + P(3|1)c(3|1) + \cdots + P(g|1)c(g|1) \\ &= \sum_{k=2}^{g} P(k|1)c(k|1) \end{aligned}$$

-overall ECM

$$ECM = p_1 ECM(1) + \cdots + p_g ECM(g) = \sum_{i=1}^{g} p_i ECM(i)$$

↳ 최소화하도록 classification regions 결정

Ω: sample space



5. Classification with several populations

-Minimum ECM classification rule

Allocate $\mathbf{x_0}$ to π_k if $\sum_{\substack{i=1 \\ i \neq k}}^{g} p_i f_i(\mathbf{x}) c(k \mid i)$ is smallest.

-Minimum ECM classification rule <u>with Equal Misclassification costs</u>

Allocate $\mathbf{x}_0$ to π_k if

$$p_k f_k(\mathbf{x}) > p_i f_i(\mathbf{x}) \quad \text{for all } i \neq k$$

or, equivalently,

Allocate $\mathbf{x}_0$ to π_k if

$$\ln p_k f_k(\mathbf{x}) > \ln p_i f_i(\mathbf{x}) \quad \text{for all } i \neq k$$

5. Classification with several populations

-Estimated Minimum TPM rule for several Normal populations ,unequal $\sum_i$

Allocate $\mathbf{x}$ to π_k if
the quadratic score $\hat{d}_k^Q(\mathbf{x})$ = largest of $\hat{d}_1^Q(\mathbf{x}), \hat{d}_2^Q(\mathbf{x}), \ldots, \hat{d}_g^Q(\mathbf{x})$

$$\hat{d}_i^Q(\mathbf{x}) = -\tfrac{1}{2}\ln|\mathbf{S}_i| - \tfrac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_i)'\mathbf{S}_i^{-1}(\mathbf{x}-\bar{\mathbf{x}}_i) + \ln p_i, \quad i = 1, 2, \ldots, g$$

Cf) 거리기반

-Estimated Minimum TPM rule for equal-covariance normal populations

Allocate $\mathbf{x}$ to π_k if
the linear discriminant score $\hat{d}_k(\mathbf{x})$ = the largest of $\hat{d}_1(\mathbf{x}), \hat{d}_2(\mathbf{x}), \ldots, \hat{d}_g(\mathbf{x})$

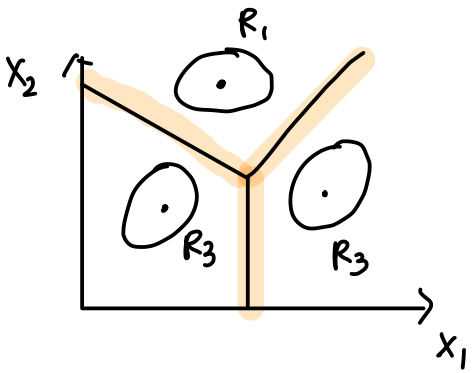
$$\hat{d}_i(\mathbf{x}) = \bar{\mathbf{x}}_i'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x} - \tfrac{1}{2}\bar{\mathbf{x}}_i'\mathbf{S}_{\text{pooled}}^{-1}\bar{\mathbf{x}}_i + \ln p_i \quad \text{for } i = 1, 2, \ldots, g$$

5. Classification with several populations

EX) linear discriminant score에 의한 영역 경계선 결정

- $d_k(x) \geq d_i(x)$ for all $i = 1, \cdots g$
- $(\mu_k - \mu_i)' \Sigma^{-1} x - \frac{1}{2}(\mu_k - \mu_i)' \Sigma^{-1}(\mu_k + \mu_i) + \ln\left(\frac{p_k}{p_i}\right) \geq 0$
- R_1: $(\mu_1 - \mu_2)' \Sigma^{-1} x - \frac{1}{2}(\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 + \mu_2) \geq \ln\left(\frac{p_2}{p_1}\right)$
  $(\mu_1 - \mu_2)' \Sigma^{-1} x - \frac{1}{2}(\mu_1 - \mu_3)' \Sigma(\mu_1 + \mu_3) \geq \ln\left(\frac{p_3}{p_1}\right)$

> intersecting hyperplanes 형성



EX) sample squared distance에 의한 분류

$n_1 = 31$ $n_2 = 28$ $n_3 = 26$

$\bar{x}_1 = \begin{bmatrix} 3.40 \\ 561.23 \end{bmatrix}$ $\bar{x}_2 = \begin{bmatrix} 2.48 \\ 447.07 \end{bmatrix}$ $\bar{x}_3 = \begin{bmatrix} 2.99 \\ 446.23 \end{bmatrix}$

$\bar{x} = \begin{bmatrix} 2.97 \\ 488.45 \end{bmatrix}$ $S_{pooled} = \begin{bmatrix} .0361 & -2.0188 \\ -2.0188 & 3655.9011 \end{bmatrix}$

(사전확률 equal, 공통공분산 가정)

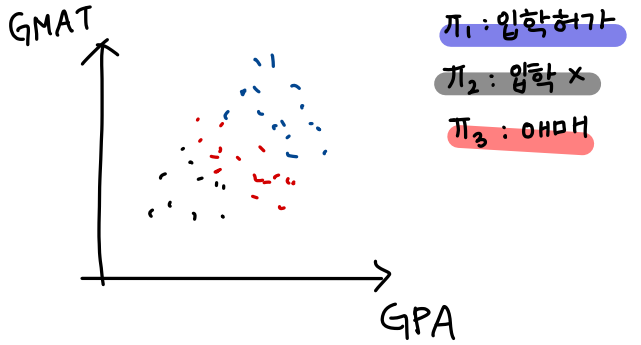
$x_0' = [\,3.21 \;,\; 4.97\,]$

$D_1^2(x_0) = (x_0 - \bar{x}_1)' S_{pooled}^{-1} (x_0 - \bar{x}_1)$

$= [3.21 - 3.40, \; \cdots \,] = 2.58$

$D_2^2(x_0) = (x_0 - \bar{x}_2)' \cdots = 17.10$

$D_3^2(x_0) = \cdots = 2.47$ ⇒ x_0를 π_3로 분류



$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} GPA \\ GMAT \end{pmatrix}$

6. Linear Discriminant Analysis

Linear Discriminant Analysis

#1 Bayes theorem and LDA

분류 문제에서 예측하고 싶은 것은 독립 변수 X가 특정 값으로 주어졌을 때 각 Class에 속할 확률, 즉 조건부 확률 $P(G = k|X = x)$

오늘 다룰 방법론 중 Logistic regression은 이를 보다 직접적으로, LDA는 베이즈 정리를 이용하여 간접적으로 예측 $P(B|A) = \frac{P(B)P(A|B)}{P(A)}$

$$P(G = k|X = x) = \frac{P(G = k)P(X = x|G = k)}{P(X = x)}$$

참고: $P(A) = P(A \cap B_1) + \cdots + P(A \cap B_k)$

$$= \frac{f_k(x)\pi_k}{\{P(G = 1)P(X = x|G = 1) + \cdots + P(G = K)P(X = x|G = K)\}}$$

$$= \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

π_k 는 주어진 전체 데이터에서 몇 개의 데이터가 k class label 가지고 있는지 계산하여 추정치로 사용

그러나 $f_k(x)$의 추정은 이렇게 간단하지 X : 이를 추정하는 것이 곧 목표로 하는 조건부 확률을 추정하는 것

LDA는 이를 추정하기 위해 이에 대한 (1) 정규 분포 가정, (2) 등분산 가정

Linear Discriminant Analysis

#2 LDA : Univariate Normal $\frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$

P=1일 때의 분포 가정 (정규 분포) $f_k(x) = \frac{1}{(\sqrt{2\pi}\sigma)}\exp(-\frac{1}{2\sigma^2}(x-\mu_k)^2)$ 평균은 다르고 분산은 동일한 정규분포가 class 개수인 k만큼 존재

조건부 확률 식에 대입하여 정리 *새로운 데이터를 classification 해야 한다면 이 확률을 가장 크게 만들어줄 수 있는 label k를 class로 선택

$P(G=k|X=x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{k} f_l(x)\pi_l} = \frac{\pi_k(\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(x-\mu_k)^2))}{\sum_{l=1}^{k}\pi_l\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(x-\mu_l)^2)}$ (1) 등분산 가정 = 분모 생각 X, (2) 로그 함수

$\log f\pi_k(\exp(-\frac{1}{2\sigma^2}(x-\mu_k)^2)) = \log\pi_k - \frac{x^2}{2\sigma^2} + \frac{\mu_k}{\sigma^2}x - \frac{\mu_k^2}{2\sigma^2}$ 등분산 가정 깨지면 소거 불가능

$\delta_k(x) = x * \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + log(\pi_k)$ 최종 도출된 Linear Discriminant Function, 각 class에 속할 확률을 비교

$\hat{\delta}_k(x) = x * \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + log(\hat{\pi}_k)$ 참값은 알 수 없고 추정이 필요

$\hat{\mu}_k = \frac{1}{n_k}\sum_{y_i=k} x_i$ sample mean

$\hat{\sigma}^2 = \frac{1}{n-k}\sum_{k=1}^{k}\sum_{i:y_i=k}(x_i-\hat{\mu}_k)^2$ pooled variance

$\hat{\pi}_k = \frac{n_k}{n}$ (kth labeled data) / (전체 data)

Linear Discriminant Analysis

#3 LDA : Multivariate Normal $\delta_k(x) = x * \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + log(\pi_k)$

P>1일 때의 분포 가정 (다변량 정규 분포) $f_k(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \mathbf{\Sigma}_k^{-1}(x-\mu_k)}$

동일과정 정리

$$\delta_k(x) = x^T \mathbf{\Sigma}^{-1} \mu_k - \frac{1}{2}\mu_k^T \mathbf{\Sigma}^{-1} \mu_k + \log \pi_k$$

$\hat{N}_k = \frac{1}{n_k} \Sigma_{y_i=k} x_i \rightarrow \Sigma_{g_i=k} x_i / N_k$

$\hat{\sigma}^2 = \frac{1}{n-k} \Sigma_{k=1}^{k} \Sigma_{i:y_i=k} (x_i - \hat{N}_k)^2$

$\hat{\Sigma} = \Sigma_{k=1}^{k} \Sigma_{g_i=k} (x - \hat{N}_k)(x - \hat{N}_k)^T / (N-k)$

$\hat{\pi}_k = \frac{n_k}{n} \rightarrow \frac{N_k}{N}$

결정 경계

$\log \frac{f_k(x)\pi_k / \sum_{i=1}^{K} f_i(x)\pi_i}{f_l(x)\pi_l / \sum_{i=1}^{K} f_i(x)\pi_i}$

$$\log \frac{\Pr(G=k|X=x)}{\Pr(G=\ell|X=x)} = \log \frac{f_k(x)}{f_\ell(x)} + \log \frac{\pi_k}{\pi_\ell}$$

$= \log \frac{\pi_k}{\pi_\ell} + \log \left\{ e^{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k) + \frac{1}{2}(x-\mu_\ell)^T \Sigma (x-\mu_\ell)} \right\}$

$= \log \frac{\pi_k}{\pi_\ell} + \log \left\{ e^{-\frac{1}{2}(\mu_k+\mu_\ell)^T \Sigma^{-1}(\mu_k-\mu_\ell) + x^T \Sigma^{-1}(\mu_k-\mu_\ell)} \right\}$

$$= \log \frac{\pi_k}{\pi_\ell} - \frac{1}{2}(\mu_k + \mu_\ell)^T \mathbf{\Sigma}^{-1}(\mu_k - \mu_\ell) + x^T \mathbf{\Sigma}^{-1}(\mu_k - \mu_\ell),$$

두 probability가 같아질 때,
즉 log ratio 0일 때가 선택의
기준이 되는 결정 경계

Linear Discriminant Analysis

#3 LDA : Multivariate Normal

$$\delta_k(x) = x^T \mathbf{\Sigma}^{-1} \mu_k - \frac{1}{2} \mu_k^T \mathbf{\Sigma}^{-1} \mu_k + \log \pi_k$$

어떤 class pair 선택하든 결정 경계는 x에 대한 linear function = Linear Discriminant Analysis

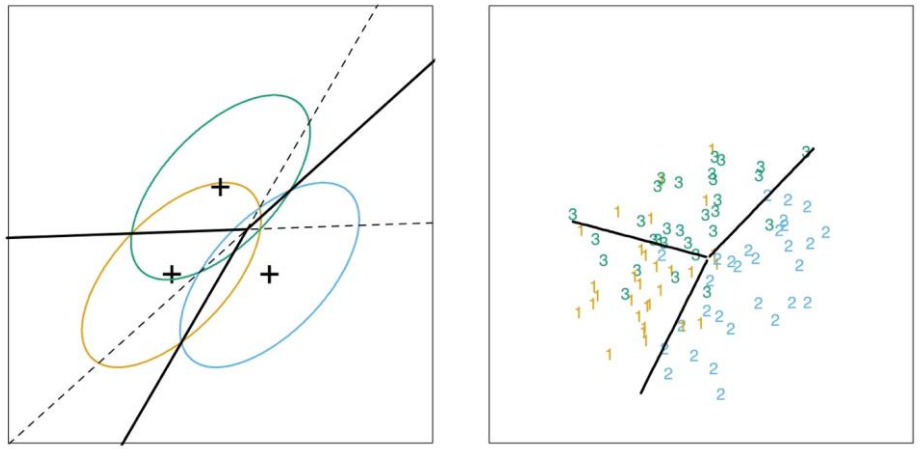
$$\log \frac{\pi_k}{\pi_\ell} - \frac{1}{2}(\mu_k + \mu_\ell)^T \mathbf{\Sigma}^{-1}(\mu_k - \mu_\ell) + x^T \mathbf{\Sigma}^{-1}(\mu_k - \mu_\ell),$$

(예) 새로운 점이 주어진 경우 class 1과 2 중 어디로 분류되어야 할까?

$$\delta_2(x) > \delta_1(x)$$

With two classes there is a simple correspondence between linear discriminant analysis and classification by linear regression, as in (4.5). The LDA rule classifies to class 2 if

$$x^T \hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{2}(\hat{\mu}_2 + \hat{\mu}_1)^T \hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) - \log(N_2/N_1), \qquad (4.11)$$



두 linear discriminant function 함숫값을 직접 비교

Linear Discriminant Analysis

#4 QDA

등분산 가정이 깨진다면? 각 Σ_k를 예측하여 사용해야 함

$$P(G=k|X=x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l} = \frac{\pi_k \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)}}{\sum_{l=1}^{K} \pi_l \frac{1}{(2\pi)^{p/2}|\Sigma_l|^{1/2}} e^{-\frac{1}{2}(x-\mu_l)^T\Sigma_l^{-1}(x-\mu_l)}}$$

$$\log\left[\pi_k \frac{1}{|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)}\right]$$

$$= \log\pi_k - \frac{1}{2}|\Sigma_k| - \frac{1}{2}(x-\mu_k)\Sigma_k^{-1}(x-\mu_k)$$

X에 대한 Quadratic form = Quadratic Discriminant Function

$$\delta_k(x) = -\frac{1}{2}\log|\mathbf{\Sigma}_k| - \frac{1}{2}(x-\mu_k)^T\mathbf{\Sigma}_k^{-1}(x-\mu_k) + \log\pi_k.$$

(Tip) 대각화 통한 계산 비용 절감

$$\hat{\Sigma}_k = U_k D_k U'.$$ *실수 대칭행렬은 고유값 분해가 가능하며 직교 행렬로 분해할 수 있음

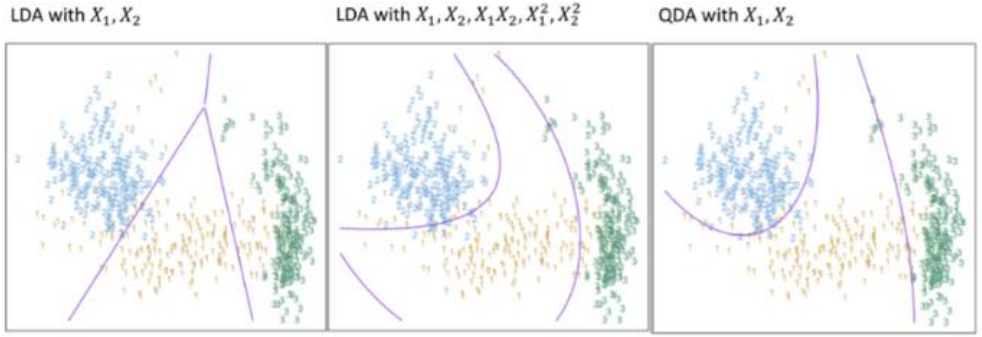
- $(x-\hat{\mu}_k)^T\hat{\mathbf{\Sigma}}_k^{-1}(x-\hat{\mu}_k) = [\mathbf{U}_k^T(x-\hat{\mu}_k)]^T\mathbf{D}_k^{-1}[\mathbf{U}_k^T(x-\hat{\mu}_k)];$
- $\log|\hat{\mathbf{\Sigma}}_k| = \sum_\ell \log d_{k\ell}.$

$$\det(\Sigma_k) = \det(U_k D_k U_k^T) = \det(U_k)\det(D_k)\det(U_k^T)$$

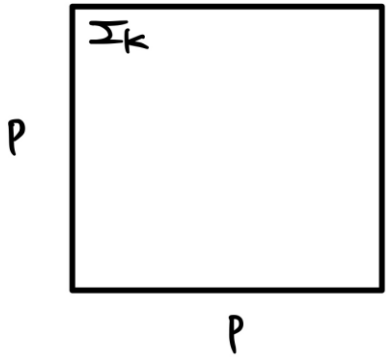
$$= \det(U_k)\det(D_k)\det(U_k)^{-1} = \det(D_k) = \pi_\ell d_{k\ell}$$

Linear Discriminant Analysis

#5 LDA vs QDA



QDA가 LDA에 비해 가정사항 적고 유연한 학습 가능하다면 왜 둘 다 사용?

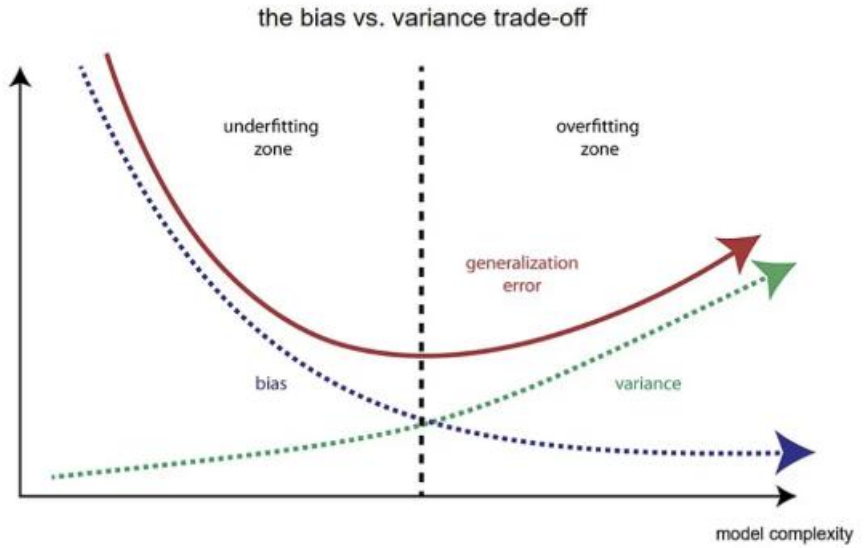


공분산 행렬에 대해 추정해야 하는 파라미터의 개수

LDA p(p+1)/2

QDA Kp(p+1)/2

-> LDA가 더 적음



Bias: 데이터 내 모든 정보 고려 못함, underfitting, 단순 모델

Variance: 실제 현상과 관련 없는 노이즈까지 학습, overfitting, 복잡 모델

데이터 개수 적어 Variance 통제 중요 = LDA

관측치 수 많아 Variance에 대한 우려 적고 등분산 가정 비현실적 = QDA

ESC

Linear Discriminant Analysis

#6 Fisher and LDA

1부에서 다룬 Fisher의 접근 방식이 2부의 LDA와 어떻게 수학적 연관?

$$\log \frac{\pi_k}{\pi_\ell} - \frac{1}{2}(\mu_k + \mu_\ell)^T \mathbf{\Sigma}^{-1}(\mu_k - \mu_\ell) + x^T \mathbf{\Sigma}^{-1}(\mu_k - \mu_\ell),$$

결정경계 수식에서 평균의 차이가 커질수록, 각 class 내의 분산 작아질수록 확률 차이 큼

= 분산 대비 평균의 차이가 최대가 되는 점을 찾는 것 (Fisher 관점과 동일)

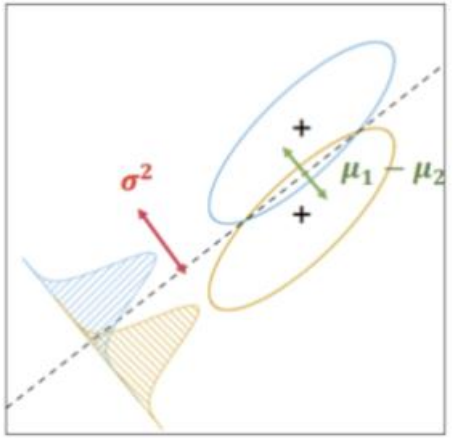
Fisher의 관점에서, FLDA는 분산을 최소화하면서 평균을 최대화하는 projection 찾는 것이 목표

$$a^T x_i,\ m_1 = a^T \mu_1,\ m_2 = a^T \mu_2$$

$$s_1^2 = \sum_{i \in \{i|y_i=1\}} (a^T x_i - m_1)^2 = \sum_{i \in \{i|y_i=1\}} a^T (x_i - \mu_1)(x_i - \mu_1)^T a$$

$$s_2^2 = \sum_{i \in \{i|y_i=2\}} (a^T x_i - m_2)^2 = \sum_{i \in \{i|y_i=2\}} a^T (x_i - \mu_2)(x_i - \mu_2)^T a$$

사영 시킬 대상 벡터를 a라고 가정할 때 사영이 되는 자료, 사영된 평균과 표본 분산



ESC

Linear Discriminant Analysis

#6 Fisher and LDA

분산을 최소화하면서 평균을 최대화하는 사영 찾는 문제는 아래와 같음

$$argmax_a \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2} = argmax_a \frac{a^T B a}{a^T W a}$$

$$B \equiv (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$$

$$W \equiv \sum_{i \in \{i | y_i = k\}} (x_i - \mu_1)(x_i - \mu_1)^T + \sum_{i \in \{i | y_i = k\}} (x_i - \mu_2)(x_i - \mu_2)^T$$

a에 대한 미분값을 0으로 두고 해 구하면 a가 eigenvector인 problem이 됨

$$\text{argmax}_a \frac{a^TBa}{a^TWa} \rightarrow \left(a^TBa \cdot (a^TWa)^{-1}\right)' = \frac{2Ba}{a^TWa} - \frac{a^TBa(2Wa)}{(a^TWa)^2} \overset{\text{set}}{=} 0$$

$$\frac{2Ba}{a^TWa} = \frac{a^TBa(2Wa)}{(a^TWa)^2} \Leftrightarrow Ba = W\left(\frac{a^TBa}{a^TWa}\right)a \Leftrightarrow W^{-1}Ba = \left(\frac{a^TBa}{a^TWa}\right)a$$

$$W^{-1}Ba = \lambda a \Leftrightarrow Ba = W\lambda a$$

$$W\lambda a = (M_1 - M_2)(M_1 - M_2)^T a. \text{ Let scalar } k = (M_1 - M_2)^T a$$

$$\text{Then, } W\lambda a = k(M_1 - M_2) \Leftrightarrow a = \frac{k}{\lambda} W^{-1}(M_1 - M_2) \rightarrow a = \Sigma^{-1}(M_1 - M_2)$$

Projection 통해 얻은 해는 eigenvalue와 eigenvector이며, eigenvector a는 위와 같이 정리

Linear Discriminant Analysis

#6 Fisher and LDA

LDA의 결정 경계 식을 해당 결과와 연관시키기

$$\text{argmax}_a \frac{a^T B a}{a^T W a} \rightarrow \left(a^T B a \cdot (a^T W a)^{-1}\right)' = \frac{2Ba}{a^T W a} - \frac{a^T B a (2Wa)}{(a^T W a)^2} \overset{\text{set}}{=} 0$$

$$\frac{2Ba}{a^T W a} = \frac{a^T B a (2Wa)}{(a^T W a)^2} \Leftrightarrow Ba = W\left(\frac{a^T B a}{a^T W a}\right)a \Leftrightarrow W^{-1}Ba = \left(\frac{a^T B a}{a^T W a}\right)a$$

$$W^{-1}Ba = \lambda a \Leftrightarrow Ba = W\lambda a$$

$$W\lambda a = (M_1 - M_2)(M_1 - M_2)^T a. \text{ Let scalar } k = (M_1 - M_2)^T a$$

$$\text{Then, } W\lambda a = k(M_1 - M_2) \Leftrightarrow a = \frac{k}{\lambda} W^{-1}(M_1 - M_2) \rightarrow a = \Sigma^{-1}(M_1 - M_2)$$

$$\log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k + \mu_l)^T \Sigma^{-1}(\mu_k - \mu_l) + x^T \Sigma^{-1}(\mu_k - \mu_l)$$

빨간색 (기울기, 법선 벡터: 앞서 구한 eigenvector) / 파란색 (직선이 지나는 한 점, 두 평균의 중점: Fisher의 관점 반영)

Fisher's discriminant를 통해 찾은 축을 기울기로 갖고, 동시에 class
집단의 평균을 지나는 초평면이 곧 LDA의 decision boundary

7. Logistic Regression and Classification

Logistic Regression and Classification

#1 The Logit Model

변수의 일부 또는 전부가 질적 변수인 경우에도 사용 가능

종속변수 Y가 여성과 남성 등의 두 개의 값만 갖는 이항형 – 두 값을 0, 1로 코딩 가능

전체 population 중 1로 코딩된 비율 = 1의 값을 가질 확률 = p

$$\text{mean} = 0 \times (1-p) + 1 \times p = p$$

$$\text{variance} = 0^2 \times (1-p) + 1^2 \times p - p^2 = p(1-p)$$

고정된 상수값이 아니며, p가 0 또는 1에 가까워지면 이 값은 0에 가까워짐

$$p = E(Y|z) = \beta_0 + \beta_1 z$$

(문제점 1) 종속변수 Y는 0, 1 값만 갖지만 단순선형회귀 적용 시 [0, 1] 벗어나는 결과값 나올 수밖에 없음

(문제점 2) 회귀분석 기본 가정 중 등분산성 만족 X

독립변수, 공변량을 모델에 집어넣는 다른 방식이 필요

Odds의 Logit 변환을 통해 독립 변수 범위 상관 없이 종속 변수 또는 결과값이 항상 [0, 1] 사이에 있도록 강제

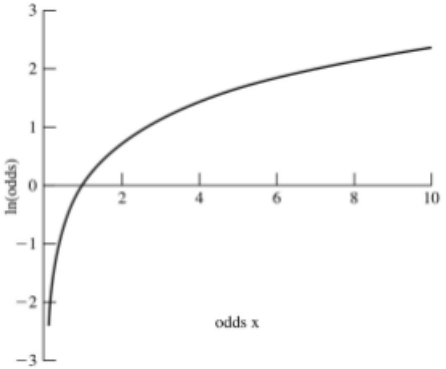
Logistic Regression and Classification

#1 The Logit Model

$$\text{odds} = \frac{p}{1-p}$$

사건 1의 확률과 사건 2의 확률 사이의 비율을 의미 (1보다 큰 값 O)

자연로그 취해주면 symmetry 부여 가능 (odds ratio 역수 = 절댓값 동일 & 부호 반대)



이러한 odds의 자연로그 값을 logit(p)로 정의

$$\text{logit}(p) = \ln(\text{odds}) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 z$$

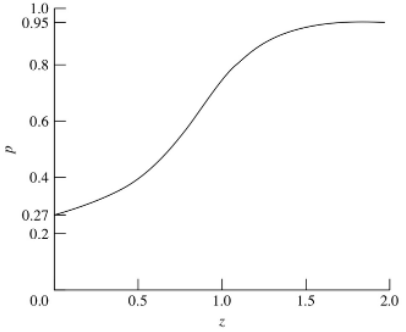
p가 아닌 이 logit 값 자체를 종속 변수로 설정

$$\theta(z) = \frac{p(z)}{1-p(z)} = \exp(\beta_0 + \beta_1 z)$$

식의 outpu이 확률값 p가 되도록 다시 정리

$$(1 + \exp(\beta_0 + \beta_1 z))p(z) = \exp(\beta_0 + \beta_1 z)$$

$$p(z) = \frac{\exp(\beta_0 + \beta_1 z)}{1 + \exp(\beta_0 + \beta_1 z)} = \frac{1}{1 + \exp(-\beta_0 - \beta_1 z)}$$



Logistic Regression and Classification

#2 Logistic Regression Analysis

독립 변수가 하나가 아닌 여러 개인 상황을 가정

종속 변수가 이진적 = 조건부 확률의 분포는 정규분포 X 이항분포 O, 해당 case는 베르누이 분포

$$P(Y_j = y_j) = p^{y_j}(\mathbf{z}_j)(1 - p(\mathbf{z}_j))^{1-y_j} \text{ for } y_j = 0, 1$$

$$E(Y_j) = p(\mathbf{z}_j) \text{ and } Var(Y_j) = p(\mathbf{z}_j)(1 - p(\mathbf{z}_j))$$

$$\ln\left(\frac{p(\mathbf{z})}{1 - p(\mathbf{z})}\right) = \beta_0 + \beta_1 z_1 + \cdots + \beta_r z_r = \beta' \mathbf{z}_j$$

Logit function 설명하는 회귀모형 적합

$$\frac{p(\mathbf{z}_j)}{1 - p(\mathbf{z}_j)} = \exp(\beta_0 + \beta_1 z_1 + \cdots + \beta_r z_r) = \exp(\beta' \mathbf{z}_j)$$

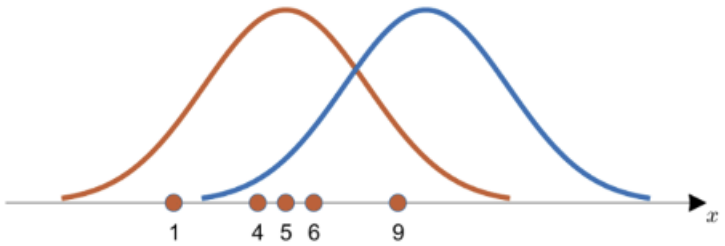
Input이 z, output p가 되도록 정리

$$(1 + \exp(\beta' \mathbf{z}_j))p(\mathbf{z}_j) = \exp(\beta' \mathbf{z}_j) \qquad p(\mathbf{z}_j) = \frac{\exp(\beta' \mathbf{z}_j)}{1 + \exp(\beta' \mathbf{z}_j)} = \frac{1}{1 + \exp(-\beta' \mathbf{z}_j)}$$

Logistic Regression and Classification

#3 Maximum Likelihood Estimation

지금 얻은 데이터가 특정 분포로부터 추출되었을 가능도



각 데이터 샘플에서 측정한 후보 분포에 대한 높이들을 전부 곱해줬을 때,
이 값이 가정 커지는 파라미터값을 추정값으로 사용: MLE

$$L(\beta) = \Pi_{j=1}^{n} \left(p(\mathbf{z}_j)^{y_j} (1 - p(\mathbf{z}_j))^{1-y_j}\right)$$

$$l(\beta) = \sum_{j=1}^{n} \{y_i \log p(\mathbf{z}_j) + (1 - y_i) \log (1 - p(\mathbf{z}_j))\}$$

$$p(z_j) = \frac{1}{1+\exp(-\beta' z_j)}, \quad 1-p(z_j) = \frac{\exp(-\beta' z_j)}{1+\exp(-\beta' z_j)}$$

$$\ell(\theta) = \sum_{j=1}^{n} \{ y_j \log p(z_j) + (1-y_j) \log (1-p(z_j)) \}$$

$$= \sum_{j=1}^{n} \{ y_j \log \left(\frac{1}{1+\exp(-\beta' z_j)}\right) + (1-y_j) \log \left(\frac{\exp(-\beta' z_j)}{1+\exp(-\beta' z_j)}\right) \}$$

$$= \sum_{j=1}^{n} \left[y_j \{ \log \left(\frac{1}{1+\exp(-\beta' z_j)}\right) - \log \left(\frac{\exp(-\beta' z_j)}{1+\exp(-\beta' z_j)}\right) \} + \log \left(\frac{\exp(-\beta' z_j)}{1+\exp(-\beta' z_j)}\right) \right]$$

$$= \sum_{j=1}^{n} \{ y_j \log \left(\frac{1}{\exp(-\beta' z_j)}\right) + \log \left(\frac{\exp(-\beta' z_j)}{1+\exp(-\beta' z_j)} * \frac{\exp(\beta' z_j)}{\exp(\beta' z_j)}\right) \}$$

$$= \sum_{j=1}^{n} \{ y_j \beta' z_j + \log \left(\frac{1}{1+\exp(\beta' z_j)}\right) \}$$

$$= \sum_{j=1}^{n} \{ y_j \beta' z_j - \log (1+e^{\beta' z_j}) \}$$

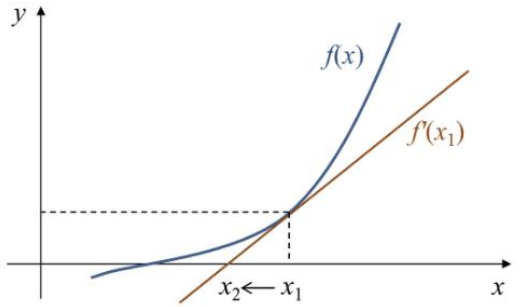
Logistic Regression and Classification

#4 Newton Raphson Method

$$x^{(k)} = x^{(k-1)} - \left(\nabla^2 f(x^{(k-1)})\right)^{-1} \nabla f(x^{(k-1)})$$

현재의 X값에서 접선을 그리고 접선이 X축과 만나는 지점으로

X를 이동시켜가며 점진적으로 해 찾는 방법



$$f_{approx}(y) = f(x) + \nabla f(x)^T (y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(x)(y-x).$$

$$\nabla f_{approx}(y) = \nabla f(x) + \frac{1}{2}\left((\nabla^2 f(x))^T (y-x) + (y-x)^T \nabla^2 f(x)\right)$$

$$= \nabla f(x) + \nabla^2 f(x)(y-x) \quad = 0, \quad \Leftrightarrow y = x - (\nabla^2 f(x))^{-1} \nabla f(x).$$

$$l(\beta) = \sum_{j=1}^{n} y_j \beta' \mathbf{z}_j - \log\left(1 + e^{\beta' \mathbf{z}_j}\right)$$ 최대화

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{j=1}^{n} y_j z_j - \frac{z_j e^{\beta' z_j}}{(1+e^{\beta' z_j})} = \sum_{j=1}^{n} z_j (y_j - P(z_j)) \overset{\text{set}}{=} 0$$

$$f(\beta) = \sum_{j=1}^{n} z_j (y_j - P(z_j)),\ P(z_j) = \frac{1}{1+e^{-\beta' z_j}},\ 1-P(z_j) = \frac{e^{-\beta' z_j}}{1+e^{-\beta' z_j}}$$

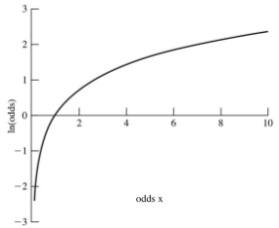
$$\frac{\partial P(z_j)}{\partial \beta} = \frac{-e^{-\beta' z_j}(-z_j)}{(1+e^{-\beta' z_j})^2} = \frac{z_j e^{-\beta' z_j}}{(1+e^{-\beta' z_j})^2} = z_j P(z_j)(1-P(z_j))$$

$$f'(\beta) = \frac{\partial f(\beta)}{\partial \beta} = \sum_{j=1}^{n} z_j z_j^T P(z_j)(1-P(z_j))$$

update $\beta^{(n+1)} = \beta^{(n)} - \frac{f(\beta^{(n)})}{f'(\beta^{(n)})}$

Logistic Regression and Classification

#5 Classification and Binomial Responses



Logistic function 결정 후 classification

$$\ln\left(\frac{p(\mathbf{z})}{1-p(\mathbf{z})}\right) = \beta_0 + \beta_1 z_1 + \cdots + \beta_r z_r = \beta' \mathbf{z}_j$$

Assign $\mathbf{z}$ to population 1 if the estimated odds ratio is greater than 1 or

$$\frac{\hat{p}(\mathbf{z})}{1-\hat{p}(\mathbf{z})} = \exp(\hat{\beta}_0 + \hat{\beta}_1 z_1 + \cdots + \hat{\beta}_r z_r) > 1$$

Assign $\mathbf{z}$ to population 1 if the linear discriminant is greater than 0 or

$$\ln \frac{\hat{p}(\mathbf{z})}{1-\hat{p}(\mathbf{z})} = \hat{\beta}_0 + \hat{\beta}_1 z_1 + \cdots + \hat{\beta}_r z_r > 0$$

각각의 관측값이 하나의 베르누이 시행이 아닌 n개의 독립적인, 동일하게 분포된 시험과 연관되어 있을 때 Binom(n, p)

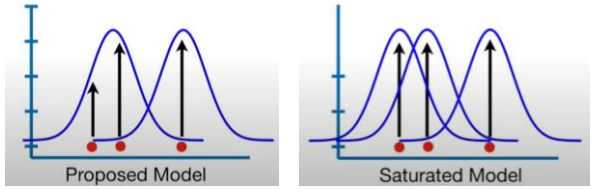
$$P(Y_j = y_j) = \binom{n_j}{y_j} p^{y_j}(\mathbf{z}_j)(1-p(\mathbf{z}_j))^{n_j - y_j} \text{ for } y_j = 0, 1$$

$$L(\beta_0, \beta_1, \ldots, \beta_r) = \Pi_{j=1}^{m} \binom{n_j}{y_j} p^{y_j}(\mathbf{z}_j)(1-p(\mathbf{z}_j))^{n_j - y_j}$$

$$\widehat{Cov}(\hat{\beta}) \approx [\sum_{j=1}^{m} n_j \hat{p}(\mathbf{z}_j)(1-\hat{p}(\mathbf{z}_j))\mathbf{z}_j \mathbf{z}_j']^{-1}$$

Logistic Regression and Classification

#6 Model Checking

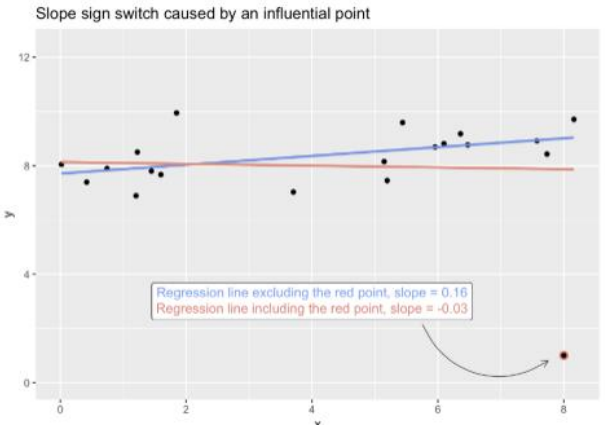


학습된 모델이 얼마나 적절한가?

1. 데이터의 측면에서 outlier의 존재 여부 및 강력한 영향력을 지닌 observation의 존재 여부 확인 (DFBETAS 등)

2. 우도비 검정 등을 통해 Full model과 Reduced model을 비교

3. 모델 성능 평가 지표로서 잔차를 활용



DEV = 2 * (포화모델 LL – 적합모델 LL)

특정 포인트 제외할 경우 모델이 얼마나 나빠지는지 평가

Deviance residuals (d_j):

$$d_j = \pm \sqrt{2\left[y_j \ln\left(\frac{y_j}{n_j \hat{p}(\mathbf{z}_j)}\right) + (n_j - y_j)\ln\left(\frac{n_j - y_j}{n_j(1-\hat{p}(\mathbf{z}_j))}\right)\right]}$$

where the sign of d_j is the same as that of $y_j - n_j\hat{p}(\mathbf{z}_j)$ and,

if $y_j = 0$, then $d_j = -\sqrt{2n_j|\ln(1-\hat{p}(\mathbf{z}_j))|}$

if $y_j = n_j$, then $d_j = -\sqrt{2n_j|\ln \hat{p}(\mathbf{z}_j))|}$

Pearson residuals(r_j): $r_j = \dfrac{y_j - n_j\hat{p}(\mathbf{z}_j)}{\sqrt{n_j\hat{p}(\mathbf{z}_j)(1-\hat{p}(\mathbf{z}_j))}}$

각 관측값에 대한 예측 확률을 계산한 후, 각 데이터 요소에 대해 관찰된 결과(0 or 1)와 예측 확률 간의 차이를 예측 확률의 분산으로 나눔

Logistic Regression and Classification

#7 Logistic Regression or LDA?

	Logistic Regression	Linear Discriminant Analysis
형태	Logit을 선형식으로 설명	Log ratio가 선형식으로 표현
방식	MLE 기반, 직접적 확률 예측	최소 제곱 기반, 조건부 정보 이용한 중간적 예측
가정	분포 형태에 구애 X, 공분산 행렬 요구사항 X	다변량 정규분포 및 등분산 가정 필요
특징	이상치, 클래스 내 샘플 수에 덜 민감	이상치에 민감하며 각 클래스들이 비슷한 n 가져야함
성능	Robust하고 준수한 예측성능	모든 요구사항 충족될 경우 BLR보다 좋은 성능

END