Classification 분류

Introduction to Statistical Learning

황성원

SW

데이터 타입: Quantitative(정량) VS. Qualitative(정성)

Quantitative: 숫자로 표시되는 값

→ 주식 값(1,120원, 1,223원, 1,402원, …), 성적(66점, 78점, 90점, …)

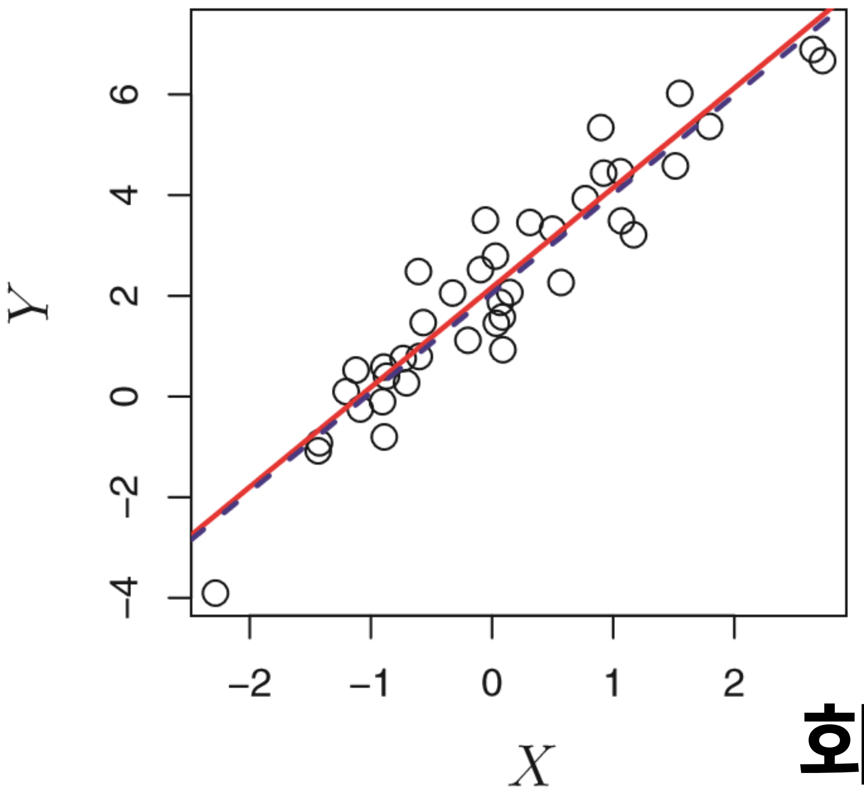
Qualitative: 카테고리(또는 Class)로 표시되는 값

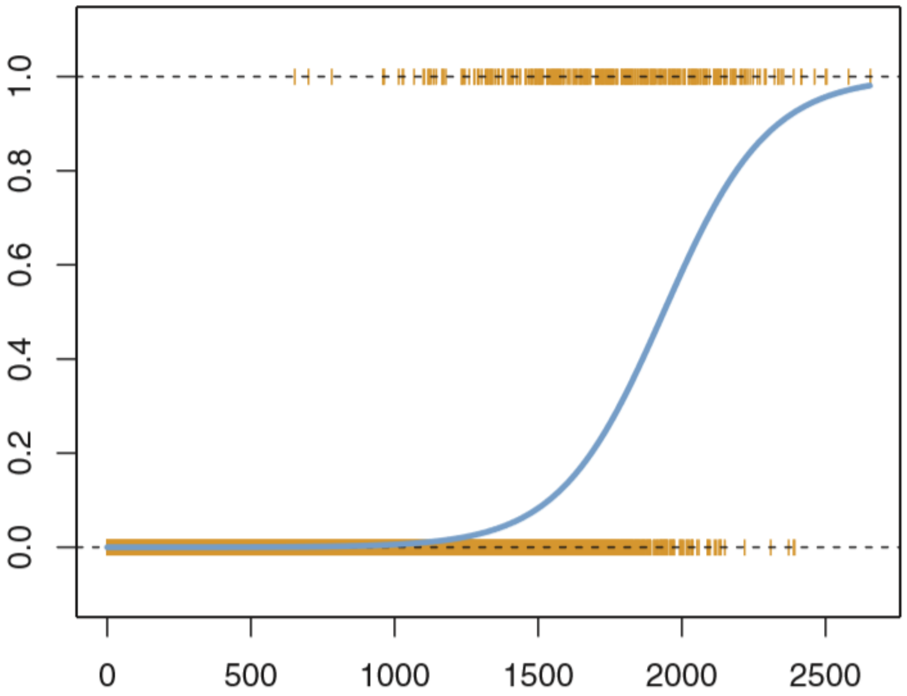
→눈의 색깔(파란색, 갈색, 초록색), 환자 상태(뇌졸증, 약물 과다, 간질)

SW

데이터 형태:
Quantitative(정량) VS. Qualitative(정성)

Quantitative → **회귀(Regression) 문제**: Fitting하여 값을 추정



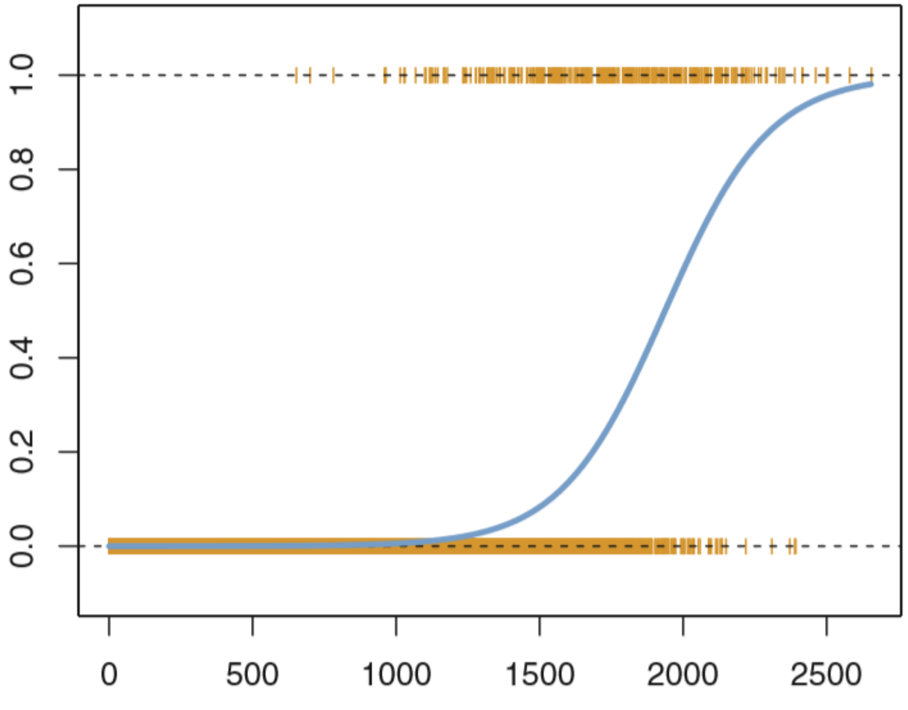


회귀 VS. 분류

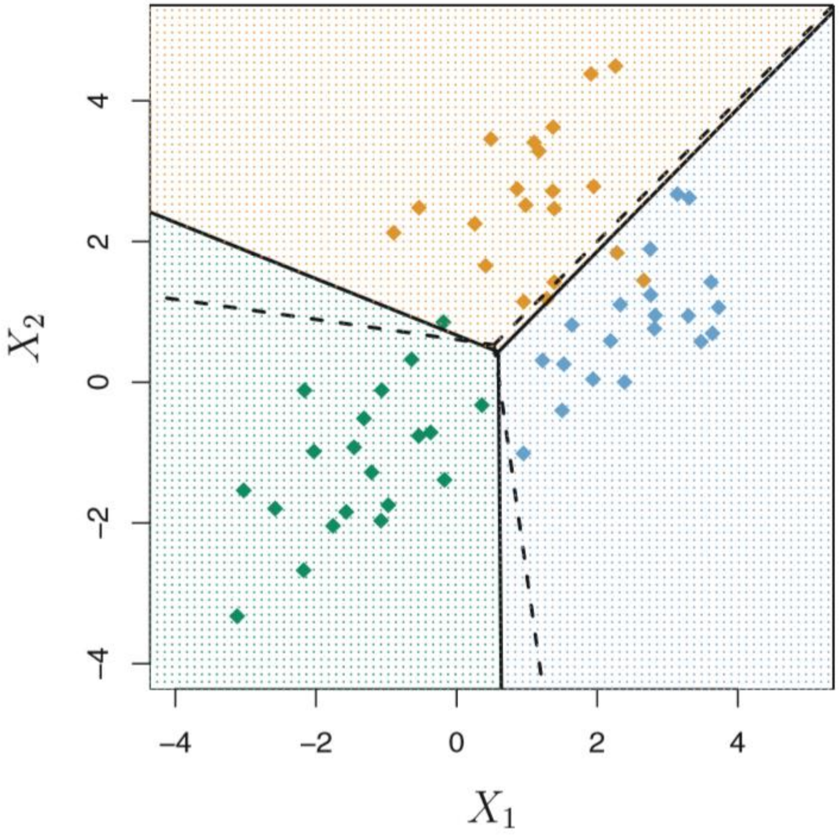
SW

데이터 형태: Quantitative(정량) VS. Qualitative(정성)

Qualitative → **분류(Classification) 문제**: Class로 할당(확률: 회귀)

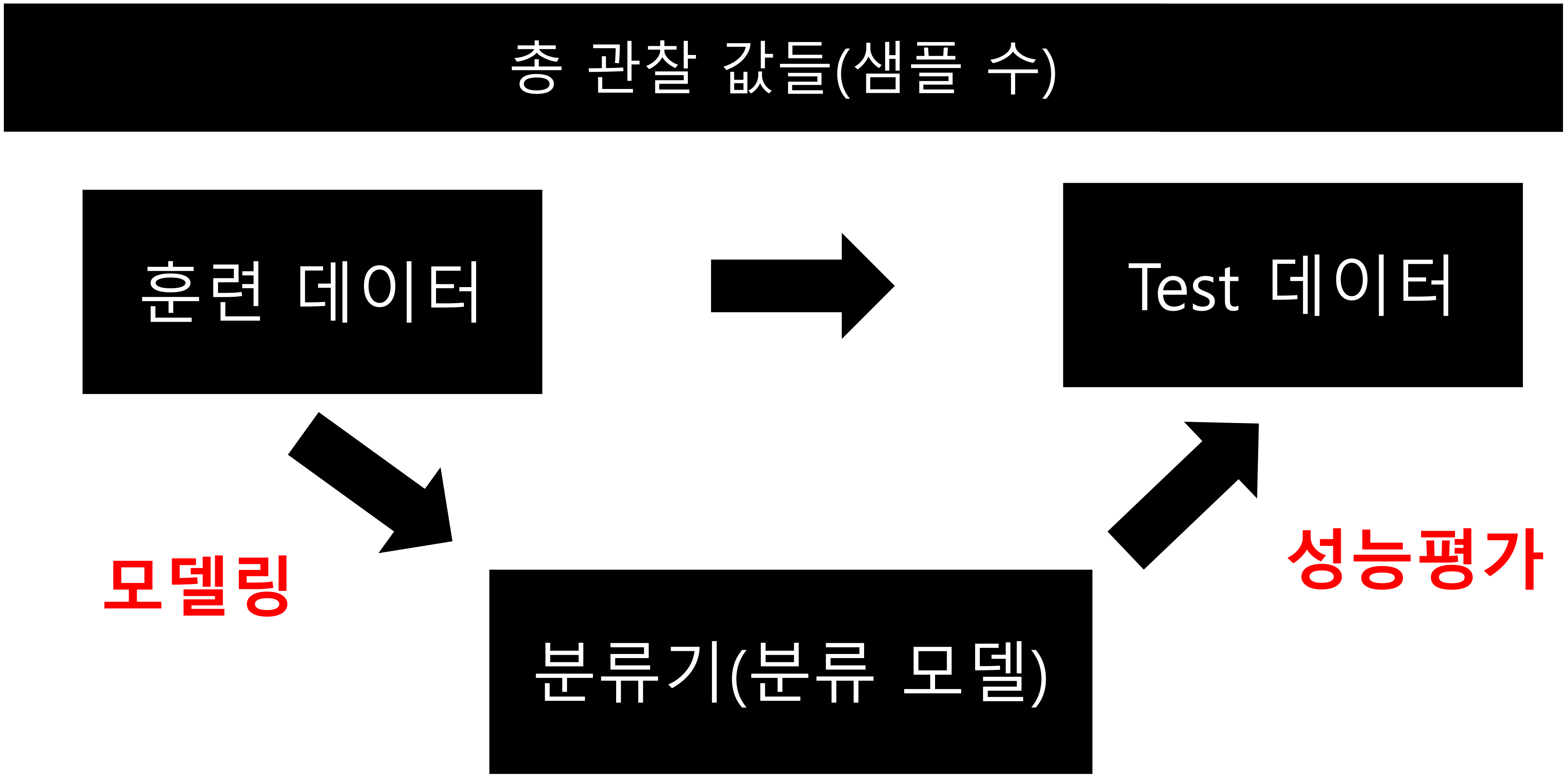


입력 변수 증가

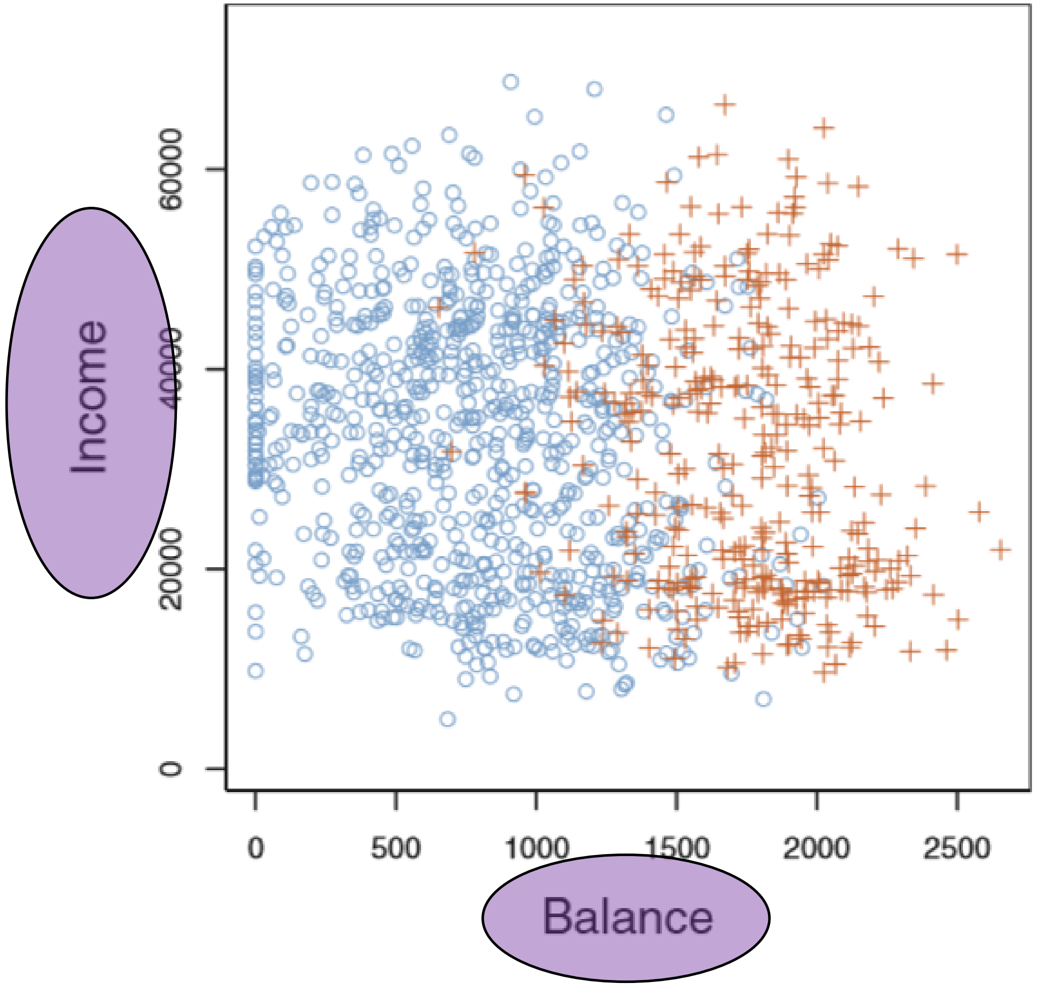


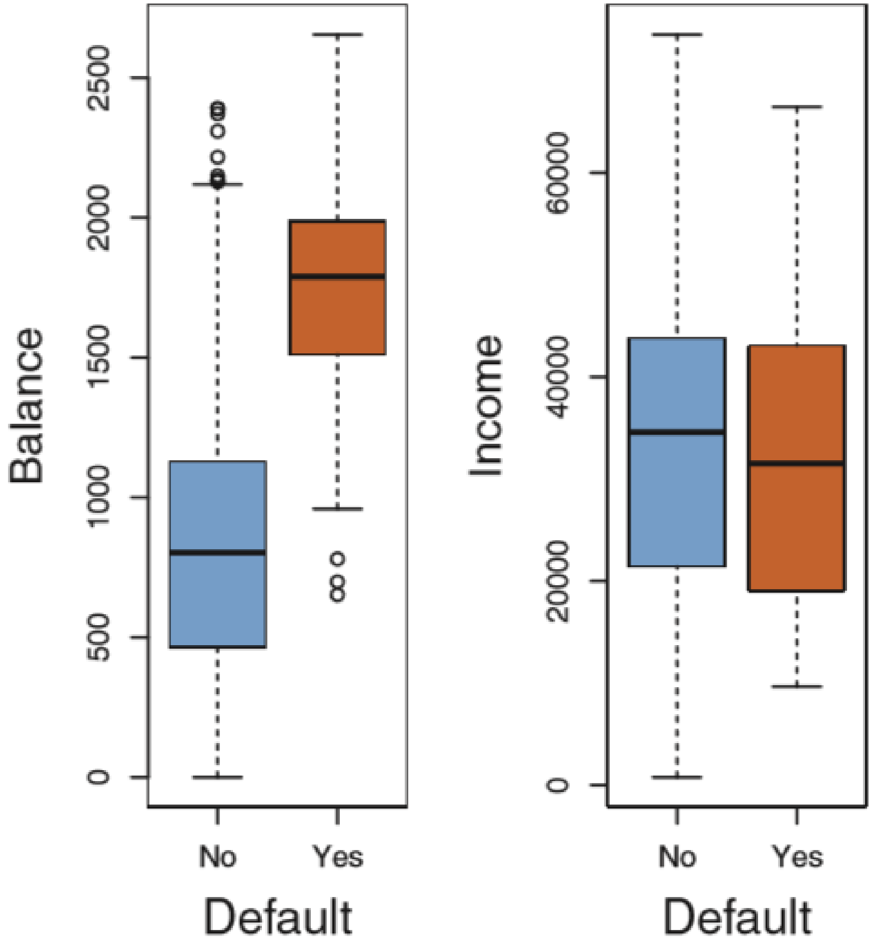
SW

데이터 세트(Dataset):
Observed = Training + Test



예시: 잔고, 수입 현황 → 채무 불이행 Yes or No





SW

분류문제에 선형 회귀는 적용이 안 될까?

예시: 응급실에 환자가 왔을 때, 환자의 증상을 보고 상태를 예측

입력　　　출력

정량적 입력: 열의 온도

20도, 25도, 30도…

정성적 입력: 열이 있다 or 없다

이진 변수가 된다.

SW

분류문제에 선형 회귀는 적용이 안 될까?

예시: 응급실에 환자가 왔을 때, 환자의 증상을 보고 상태를 예측

입력

출력

정성적 출력:

뇌졸증 or 약물 과다 복용 or 간질성 발작

분류문제에 선형 회귀는 적용이 안 될까?

예시: 응급실에 환자가 왔을 때, 환자의 증상을 보고 상태를 예측

입력

출력

정량적 출력으로 변환(암호화)!

$$Y = \begin{cases} 1 & \text{if 뇌졸증} \\ 2 & \text{if 약물 과다 복용} \\ 3 & \text{if 간질성발작} \end{cases}$$

정성적 출력:

뇌졸증 or 약물 과다 복용 or 간질성 발작

안 되는 이유: 순서 매기기를 통한 오해

$$Y = \begin{cases} 1 & \text{if 뇌졸증} \\ 2 & \text{if 약물 과다 복용} \\ 3 & \text{if 간질성발작} \end{cases} \quad \textbf{vs.} \quad Y = \begin{cases} 1 & \text{if 간질성발작} \\ 2 & \text{if 뇌졸증} \\ 3 & \text{if 약물 과다 복용} \end{cases}$$

여기서 1.5라는 값이 예측되면 무슨 의미일까?

대안: 자연적인 순서를 가지는 출력 Mild or Moderate or Severe

➡ **1, 2, 3**

일반적으로, 2개 이상의 Class를 가지는 정성 출력 변수를 정량 변수로 변환 하는 방법X

2개!의 정성 출력 변수의 경우!

확률 개념

$$Y = \begin{cases} 0 & \text{if 뇌졸증} \\ 1 & \text{if 약물 과다 복용} \end{cases} \qquad \hat{Y} > 0.5$$

Dummy Variable 접근법

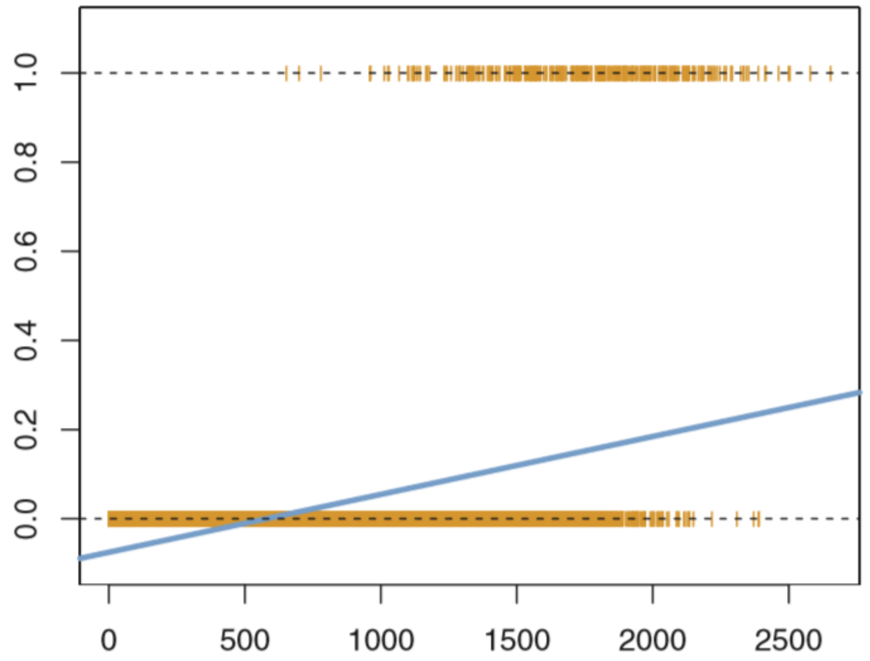
선형 회귀 방식으로 하기 위해
최소 자승법(Least Squares)를 사용해서 예측해보면,

$$Y \approx \beta_0 + \beta_1 X \quad = \quad P(\text{약물 과다 복용} \mid X)$$

위와 같이 계산하는 절차는 LDA(Linear Discriminant Analysis)와 동일!

SW

문제점



1. 마이너스 구역과 1 이상의 구역이 존재

2. Dummy Variable 접근법은 0과 1만을 취하므로 2개 이상의 Class를 가지는 정성적 출력에는 적합하지 않음

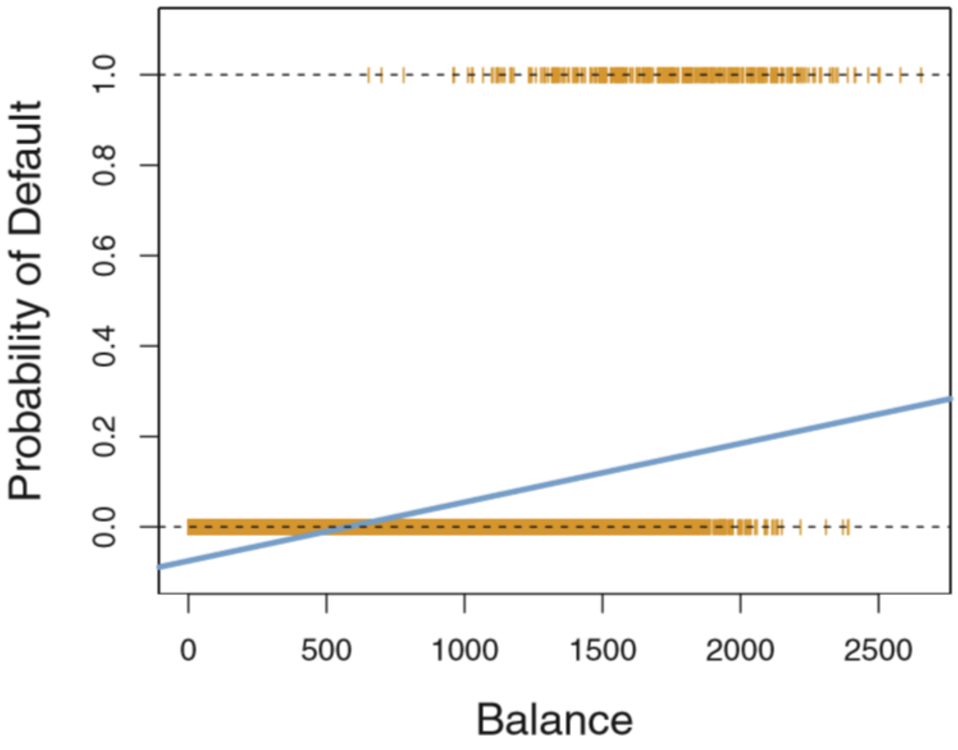
Logistic Regression(로지스틱 회귀)

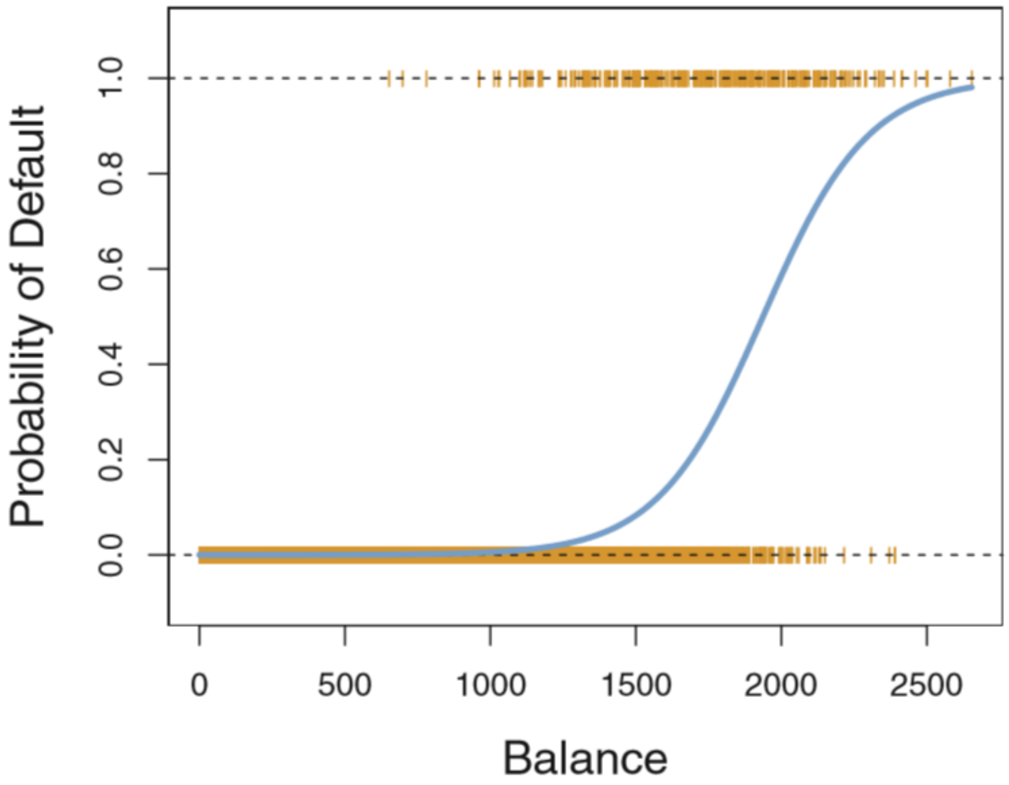
로지스틱 회귀

$$P(\text{잔고}) = P(X)$$

$$\|$$

$$\Pr(\text{채무불이행} = Yes \mid \text{잔고}) > 0.1 / 0.5 / 0.9$$





SW

로지스틱 모델

$$\boxed{P(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}}$$

$$odds = \frac{P(X)}{1 - P(X)} = e^{\beta_0 + \beta_1 X}$$

Scale 변환

0~1 → 0~무한대

로지스틱 모델

$$odds = \frac{P(X)}{1-P(X)} = e^{\beta_0 + \beta_1 X}$$

$$\log odds = \log(\frac{P(X)}{1-P(X)}) = \beta_0 + \beta_1 X$$

Log odds는 X에 대해서 선형이 된다.

회귀 계수 측정 방법

*선형 회귀 ← 최소 자승법 (최대우도법의 특수한 경우)

*로지스틱 회귀 ← 최대우도법

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} P(x_i) \prod_{i':y_{i'}=0} (1 - P(x_i'))$$

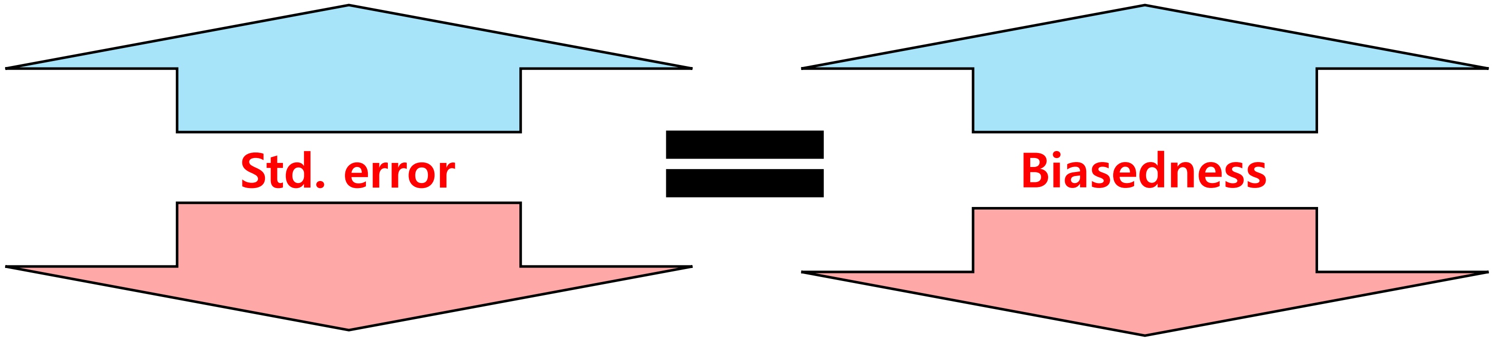
계수 분석

R을 이용해서 계산 수행 후 결과.

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.6513	0.3612	−29.5	<0.0001
balance	0.0055	0.0002	24.9	<0.0001

$$\log odds = \log(\frac{P(X)}{1-P(X)}) = \boxed{\beta_0} + \boxed{\beta_1} X$$

계수 분석



R을 이용해서 계산 수행 후 결과.

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.6513	0.3612	−29.5	<0.0001
balance	0.0055	0.0002	24.9	<0.0001

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\boxed{\hat{\beta}_1 \approx \beta_1}$$

계수 분석

Std. error

=

Biasedness

R을 이용해서 계산 수행 후 결과.

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.6513	0.3612	−29.5	<0.0001
balance	0.0055	0.0002	24.9	<0.0001

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\boxed{\hat{\beta}_1 \approx \beta_1}$$

SW

...

How far is far enough? This of course depends on the accuracy of $\hat{\beta}_1$—that is, it depends on $\mathrm{SE}(\hat{\beta}_1)$. If $\mathrm{SE}(\hat{\beta}_1)$ is small, then even relatively small values of $\hat{\beta}_1$ may provide strong evidence that $\beta_1 \neq 0$, and hence that there is a relationship between X and Y. In contrast, if $\mathrm{SE}(\hat{\beta}_1)$ is large, then $\hat{\beta}_1$ must be large in absolute value in order for us to reject the null hypothesis.

계수 분석

R을 이용해서 계산 수행 후 결과

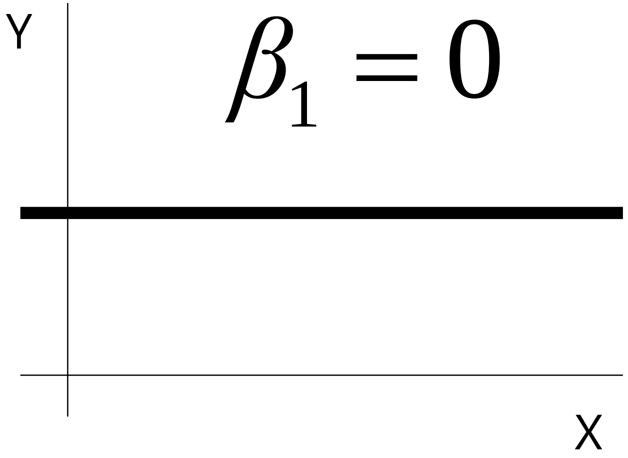
$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)} = $$

$\hat{\beta}_1$ 이 얼마나 0으로부터 멀리 떨어졌는지의 표준편차

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.6513	0.3612	−29.5	<0.0001
balance	0.0055	0.0002	24.9	<0.0001

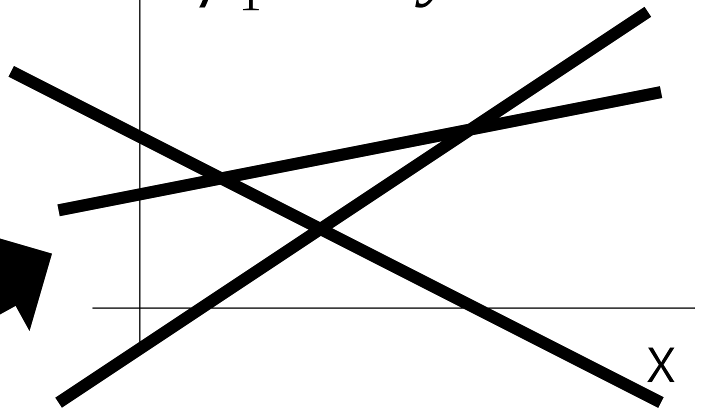
귀무가설(X,Y간에 관계가 없다)

$$\beta_1 = 0$$



대립가설(X,Y간에 관계가 있다)

$$\beta_1 = any\ values$$



$$Y = \beta_0 + \boxed{\beta_1} X$$

P-value < 0.05

예측 값 계산 해보기

잔고(X)가 1,000인 경우의 채무 불이행 확률은 다음과 같다.

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1,000}}{1+e^{-10.6513+0.0055\times 1,000}} = 0.00576$$

다른 입력(학생여부)일 경우.

학생 여부 → 0과 1로 학생일 때 1, 아닐 때 0으로

이번에는 입력에 대해서 Dummy Variable 접근법을 적용.

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−3.5041	0.0707	−49.55	<0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

$$\widehat{\Pr}(\texttt{default}=\texttt{Yes}|\texttt{student}=\texttt{Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431,$$

$$\widehat{\Pr}(\texttt{default}=\texttt{Yes}|\texttt{student}=\texttt{No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292.$$

W

다변량(여러 입력) 로지스틱 회귀

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	−0.6468	0.2362	−2.74	0.0062

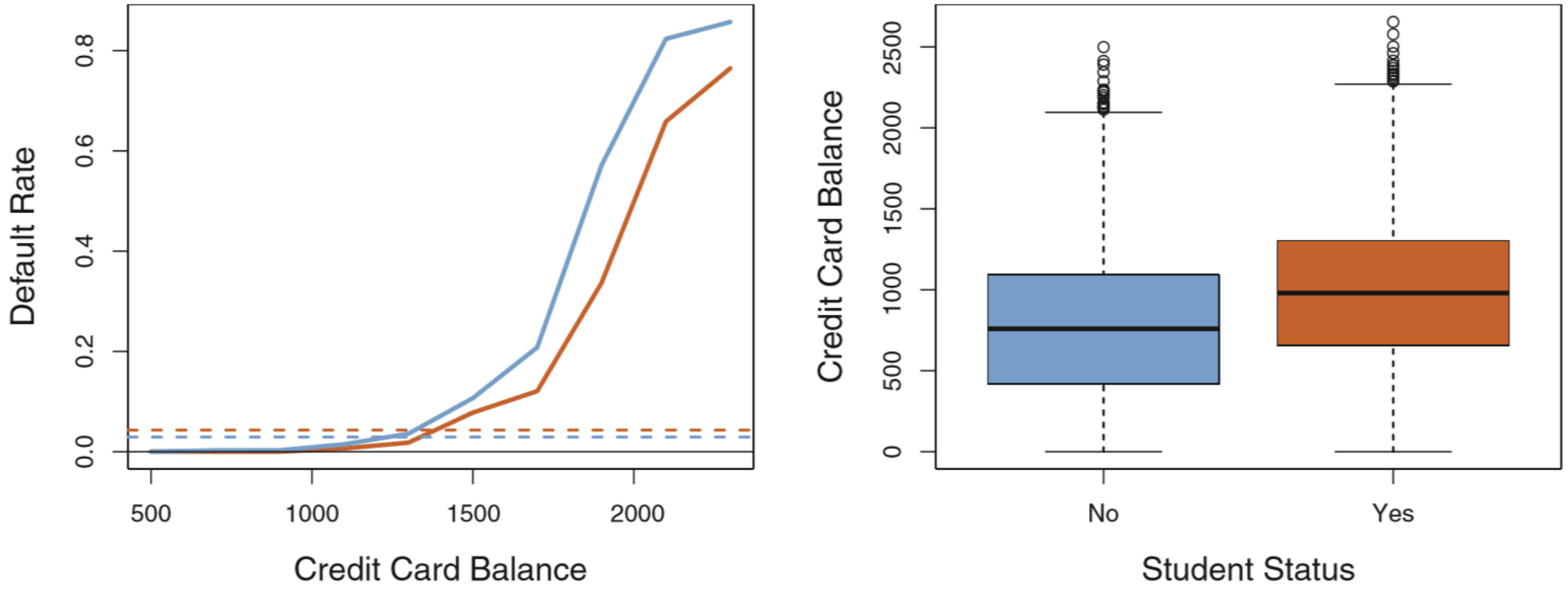
학생 여부에 관한 역설적 결과 분석

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−3.5041	0.0707	−49.55	<0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

VS.

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	−0.6468	0.2362	−2.74	0.0062

Confounding?



$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times \boxed{1}}}{1+e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times \boxed{1}}} = \boxed{0.058.}$$

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times \boxed{0}}}{1+e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times \boxed{0}}} = \boxed{0.105.}$$

SW

로지스틱 회귀(Class가 2개 이상인 경우)

- Discriminant Analysis 에서 다루게 되므로, 따로 다루지 않음.

- R에서 쉽게 구현 가능하다.

LDA (Linear Discriminant Analysis)

Bayes' Theorem

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)} = posterior = \frac{likelihood \times prior}{evidence}$$

$$f_k(x) = P(X = x \mid Y = k)$$

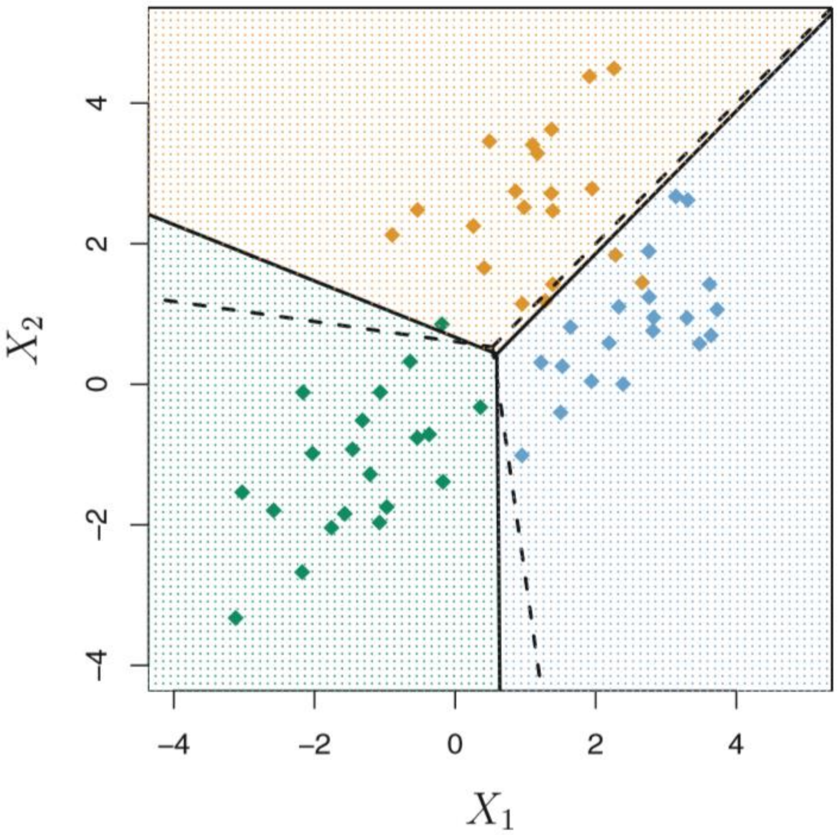
$$\pi_k = P(Y = k)$$

$$P(Y = k \mid X = x) = \frac{f_k(x) \times \pi_k}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

LDA (Linear Discriminant Analysis)

$$\pi_k = P(Y = k)$$

$$P(Y = k \mid X = x) = \frac{f_k(x) \times \pi_k}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

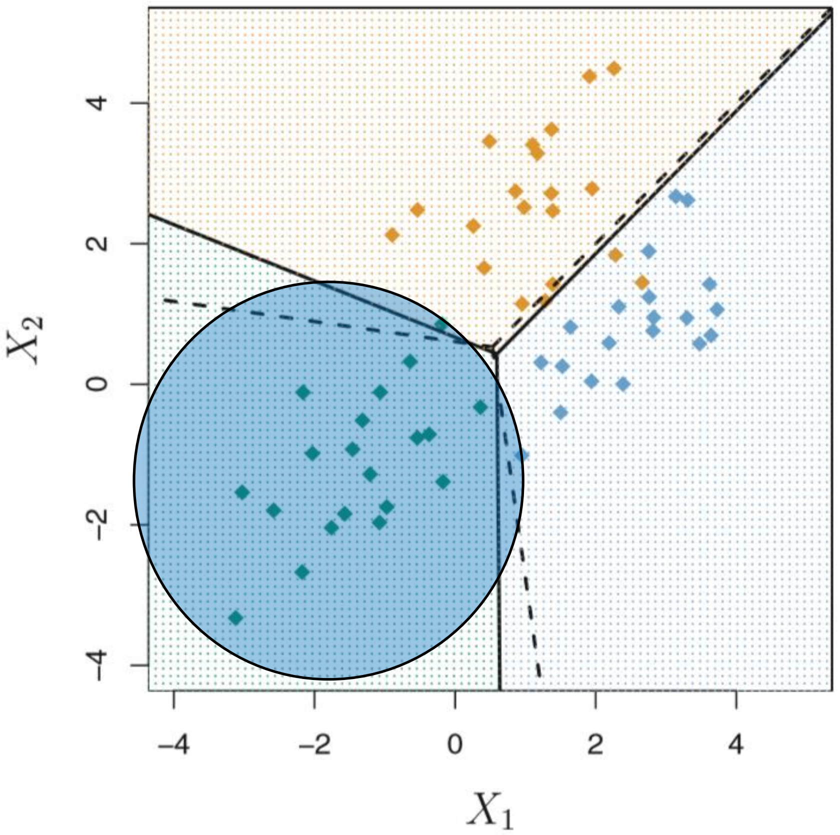


SW

LDA (Linear Discriminant Analysis)

$$f_k(x) = P(X = x \mid Y = k)$$

$$P(Y = k \mid X = x) = \frac{f_k(x) \times \pi_k}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

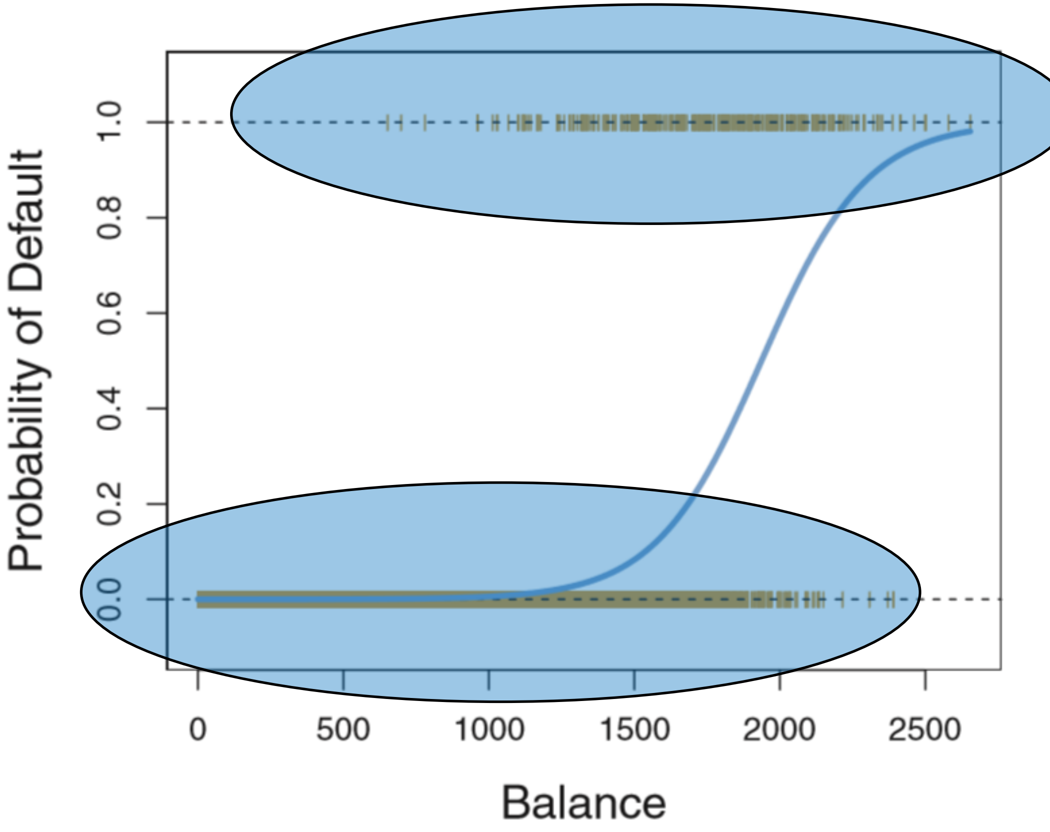


SW

LDA 입력 변수가 p=1개인 경우

$$f_k(x) = P(X = x \mid Y = k)$$

$$P(Y = k \mid X = x) = \frac{f_k(x) \times \pi_k}{\sum_{l=1}^{K} \pi_l f_l(x)}$$



가정 1

Normal or Gaussian Distribution

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right)$$

LDA 입력 변수가 p=1개인 경우

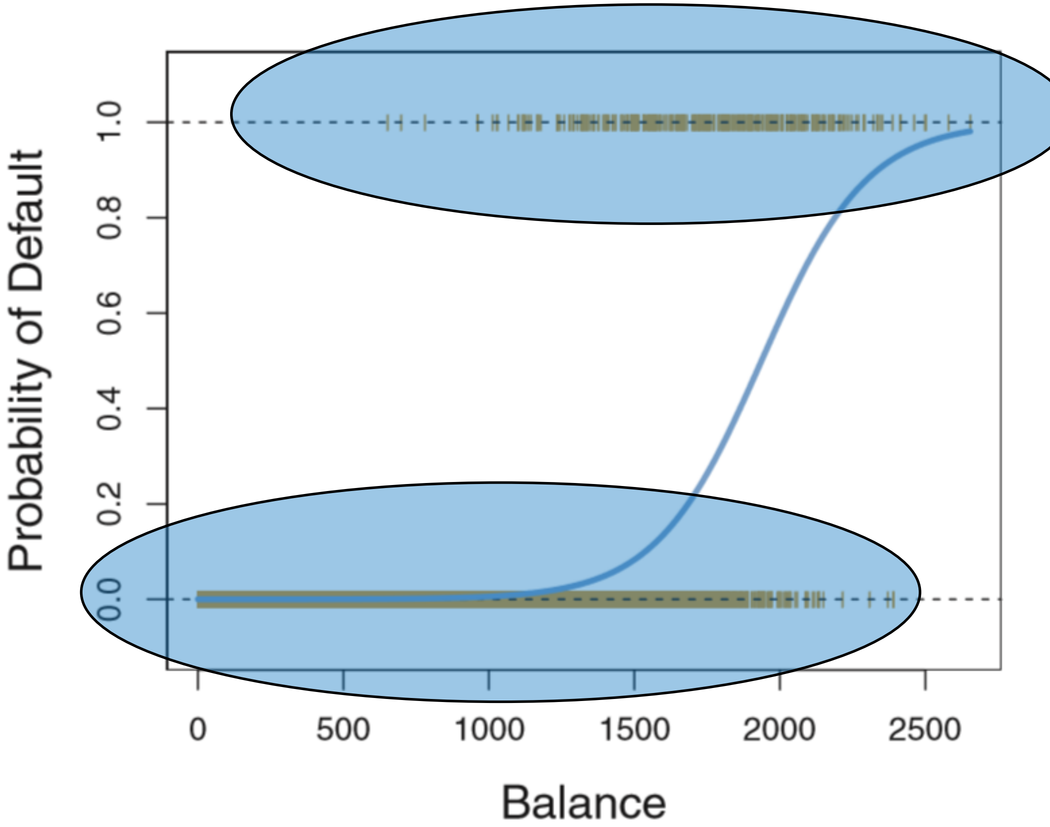
$$f_k(x) = P(X = x \mid Y = k)$$

$$P(Y = k \mid X = x) = \frac{f_k(x) \times \pi_k}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

가정 2

각 Class에서의 분산이 동일!

$$\sigma_1^2 = \ldots = \sigma_K^2$$



LDA 입력 변수가 p=1개인 경우

$$P(Y=k \mid X=x) = \frac{f_k(x) \times \pi_k}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

가정 1

Normal or Gaussian Distribution

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)$$

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}$$

가정 2

각 Class에서의 분산이 동일!

$$\sigma_1^2 = \ldots = \sigma_K^2$$

양변 Log

discriminant functions

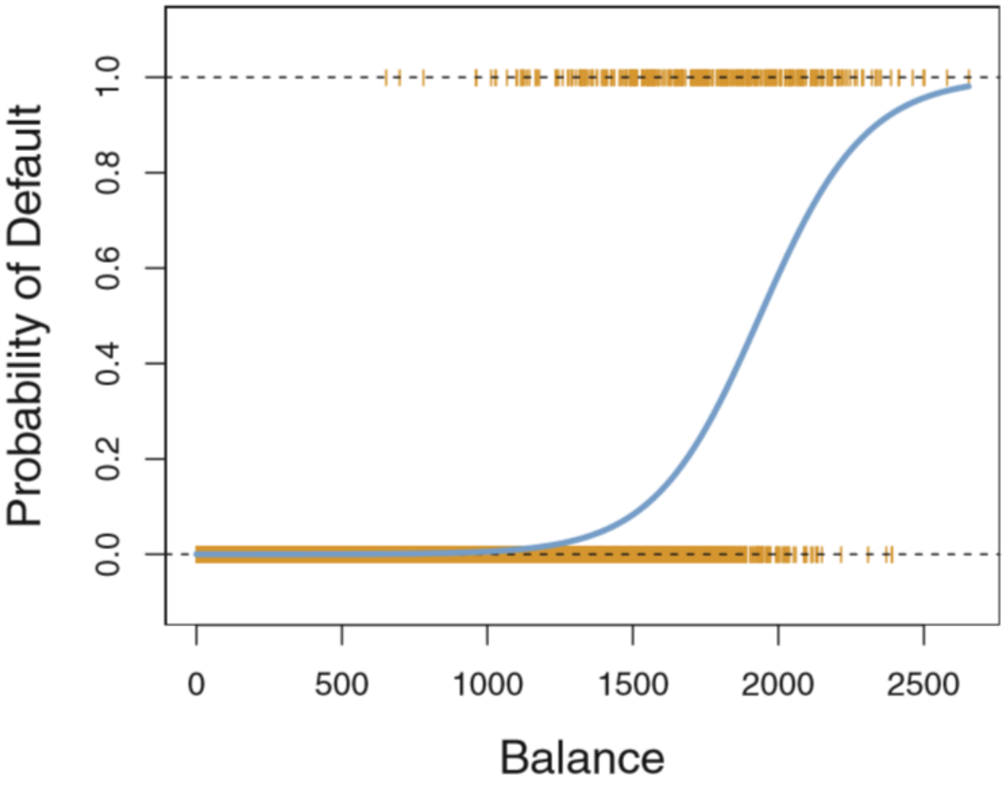
$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

회귀 계수 측정 방법

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

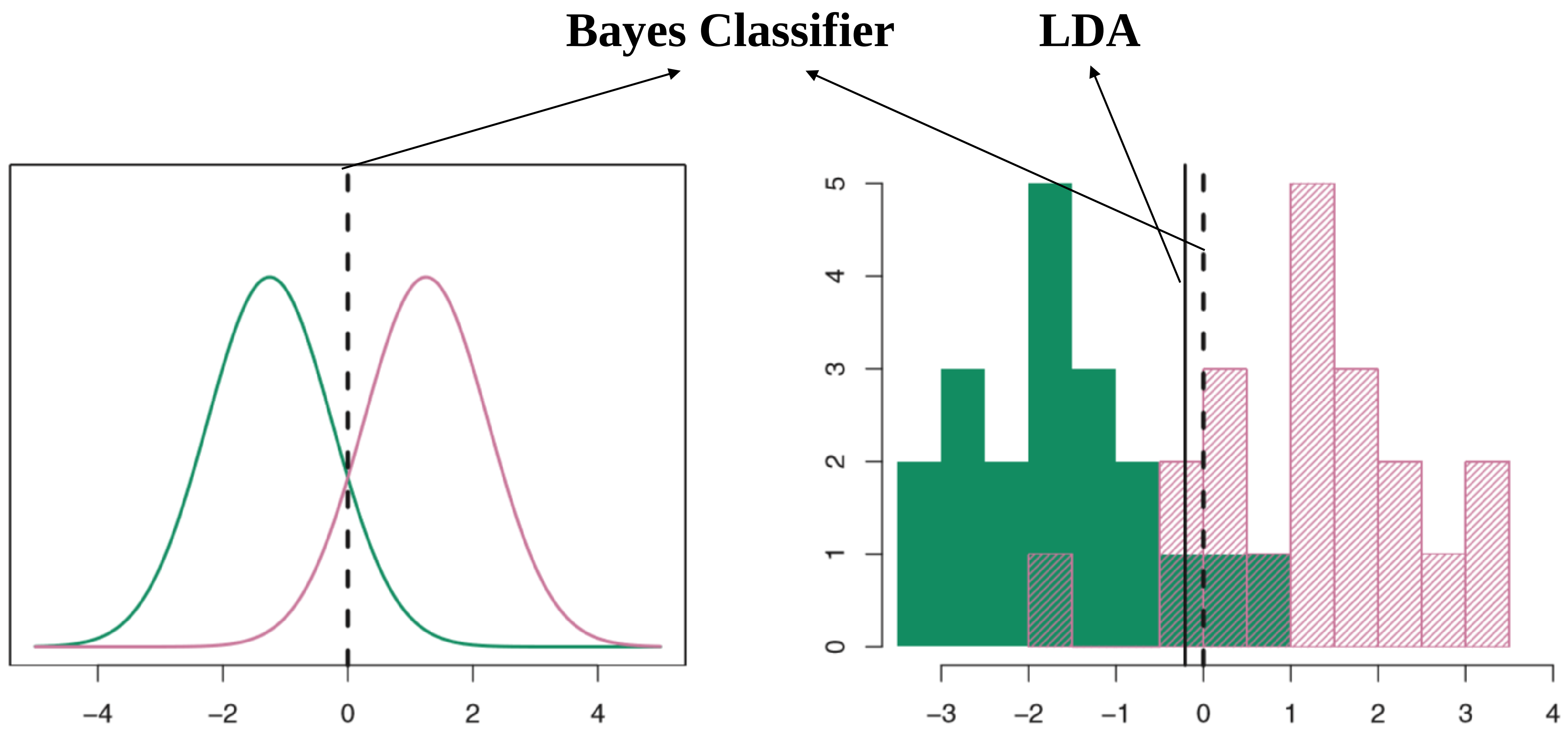
$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

$$\hat{\pi}_k = n_k/n$$



$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$$

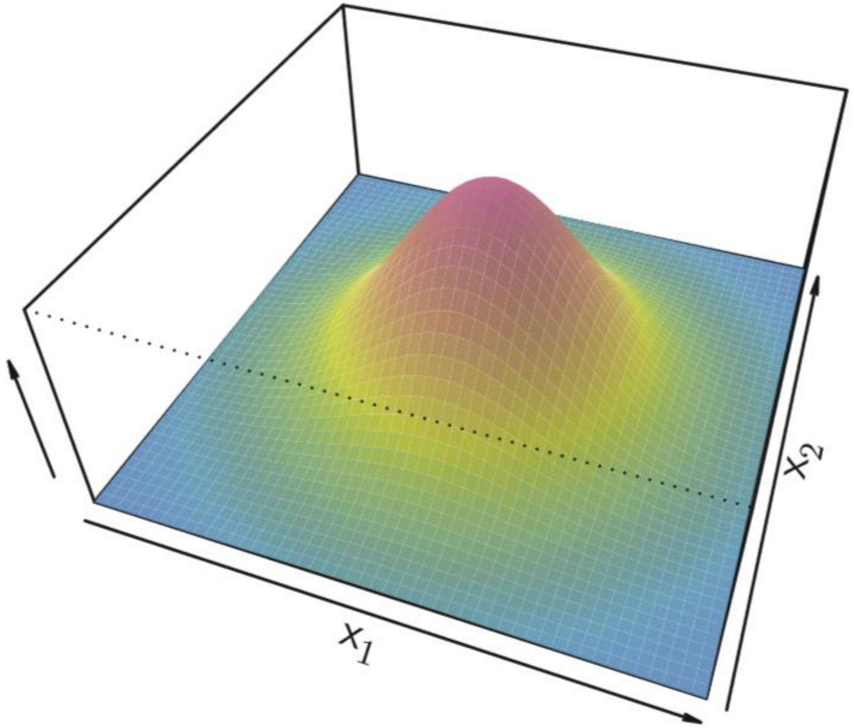
Bayes Classifier VS. LDA



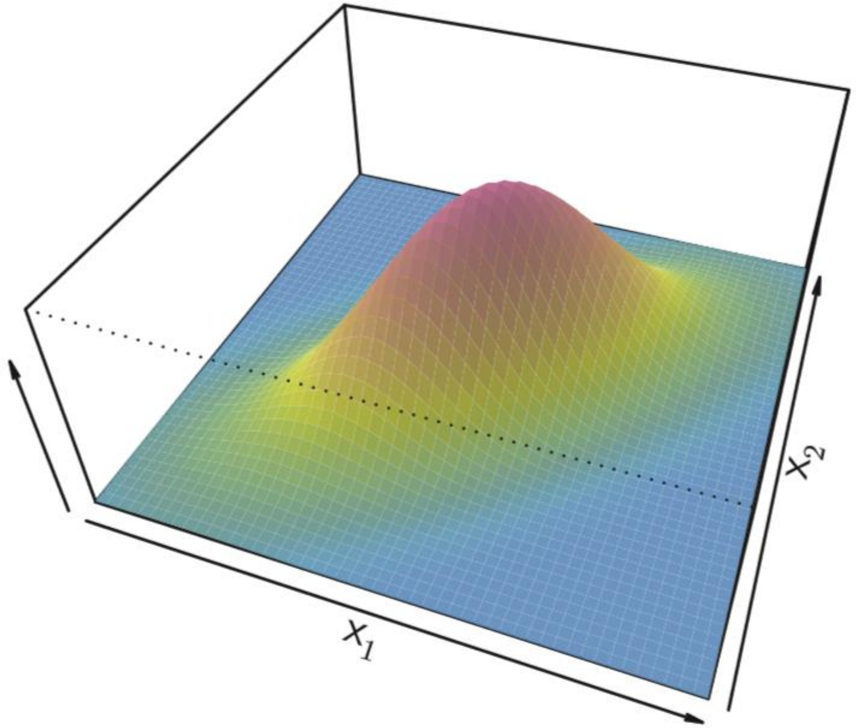
$$K = 2 \text{ and } \pi_1 = \pi_2$$

LDA 입력 변수가 p>1개인 경우

Multivariate Gaussian(normal) Distribution



입력끼리 상관 있다



$$\mathrm{Var}(X_1) = \mathrm{Var}(X_2) \ \ \mathrm{Cor}(X_1, X_2) = 0$$

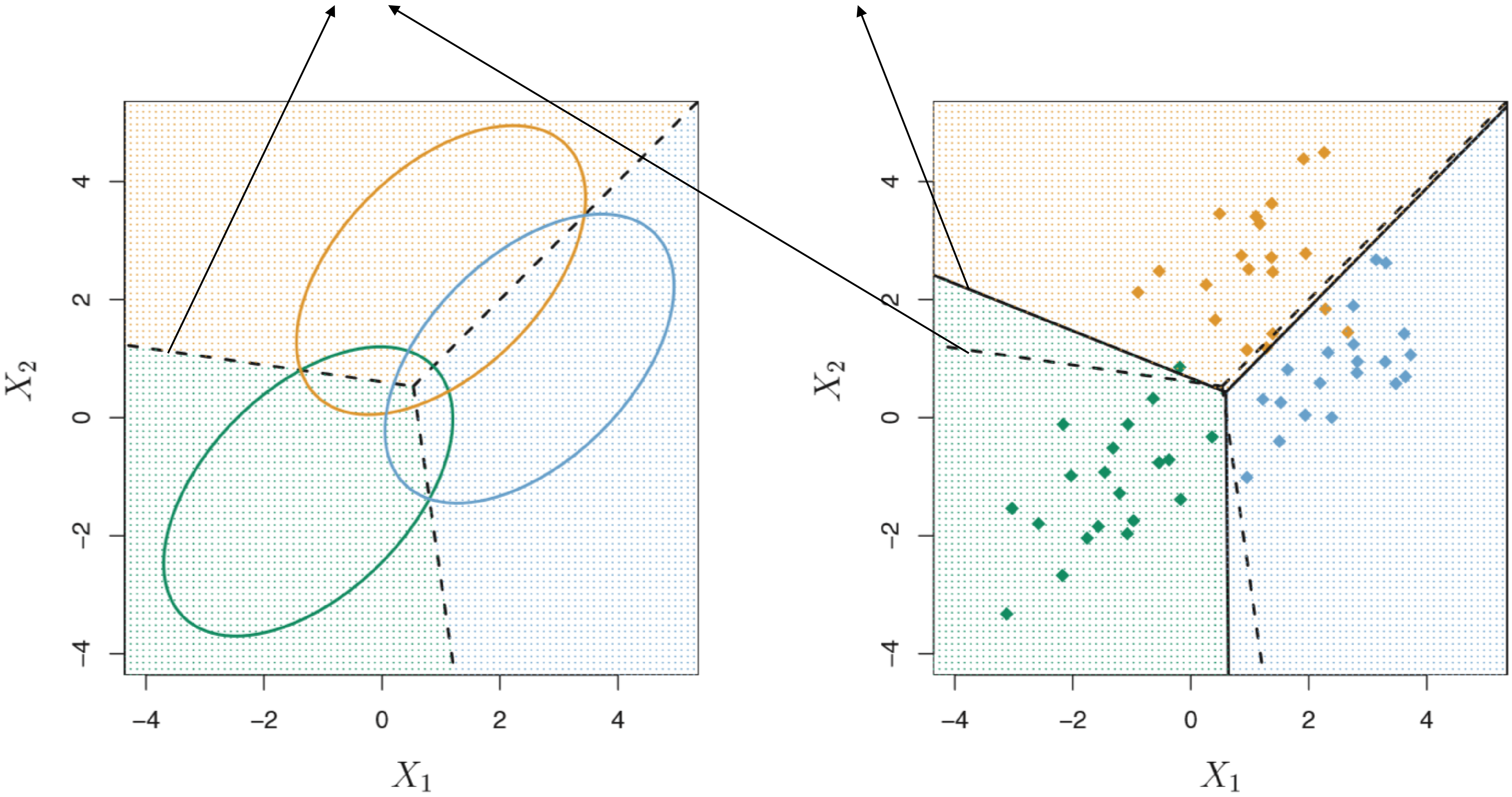
Discriminant Function 모델링

$$X \sim N(\mu, \mathbf{\Sigma}) \qquad \mathrm{Cov}(X) = \mathbf{\Sigma}$$

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

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

Bayes Classifier VS. LDA



SW

로지스틱이 있는데 왜 LDA?

1. Class가 많이 나눠진 경우,

2. 샘플 개수 n이 적으면서 입력 X 분포가 Gaussian과 흡사 할 때,

LDA가 로지스틱 보다 안정적!

QDA (Quadratic Discriminant Analysis)

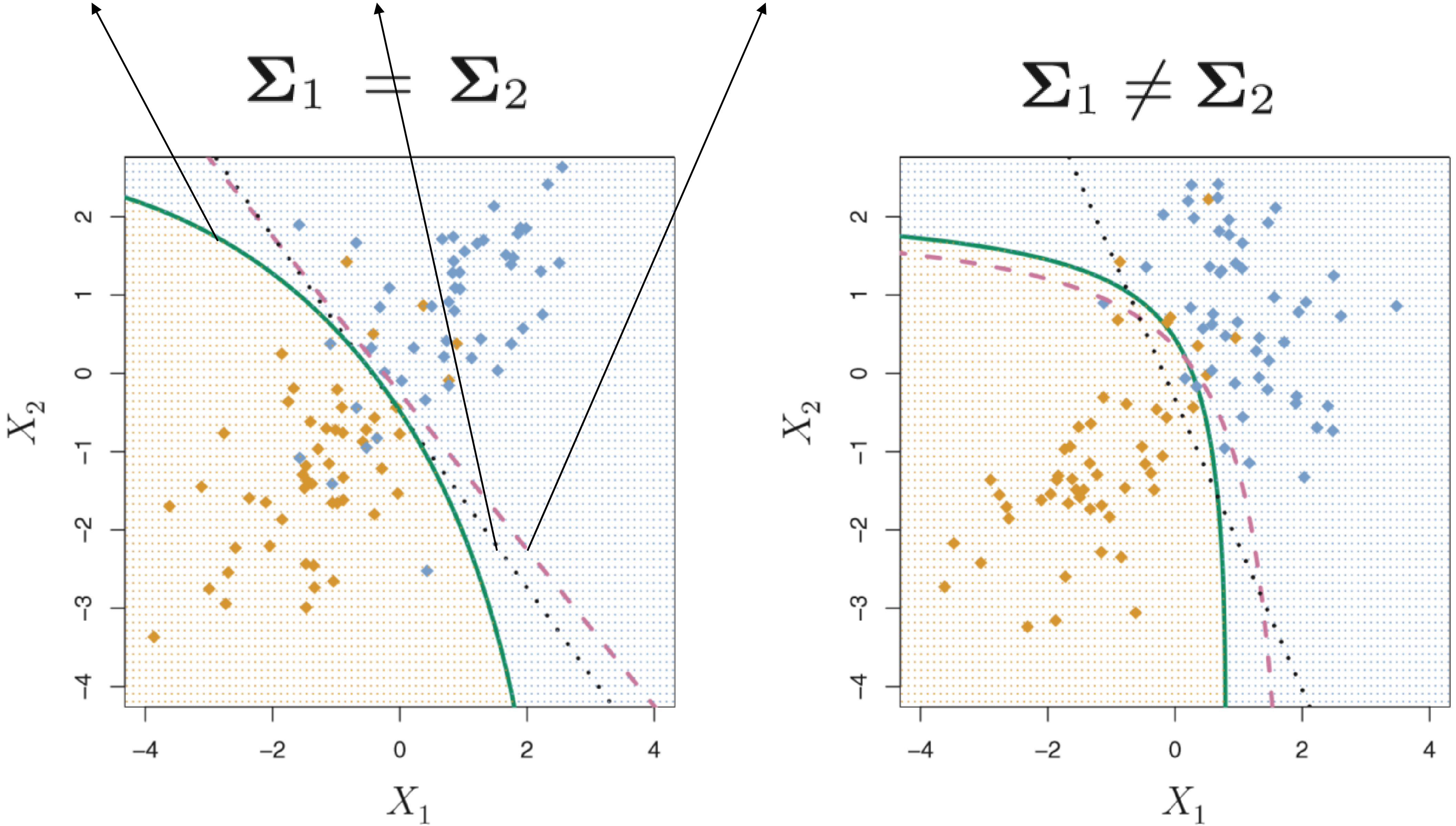
LDA = Gaussian 분포 가정 + 베이즈 정리 이용하여 추정 +

공분산 동일

QDA = LDA – 공분산 동일 $X \sim N(\mu_k, \boldsymbol{\Sigma}_k)$

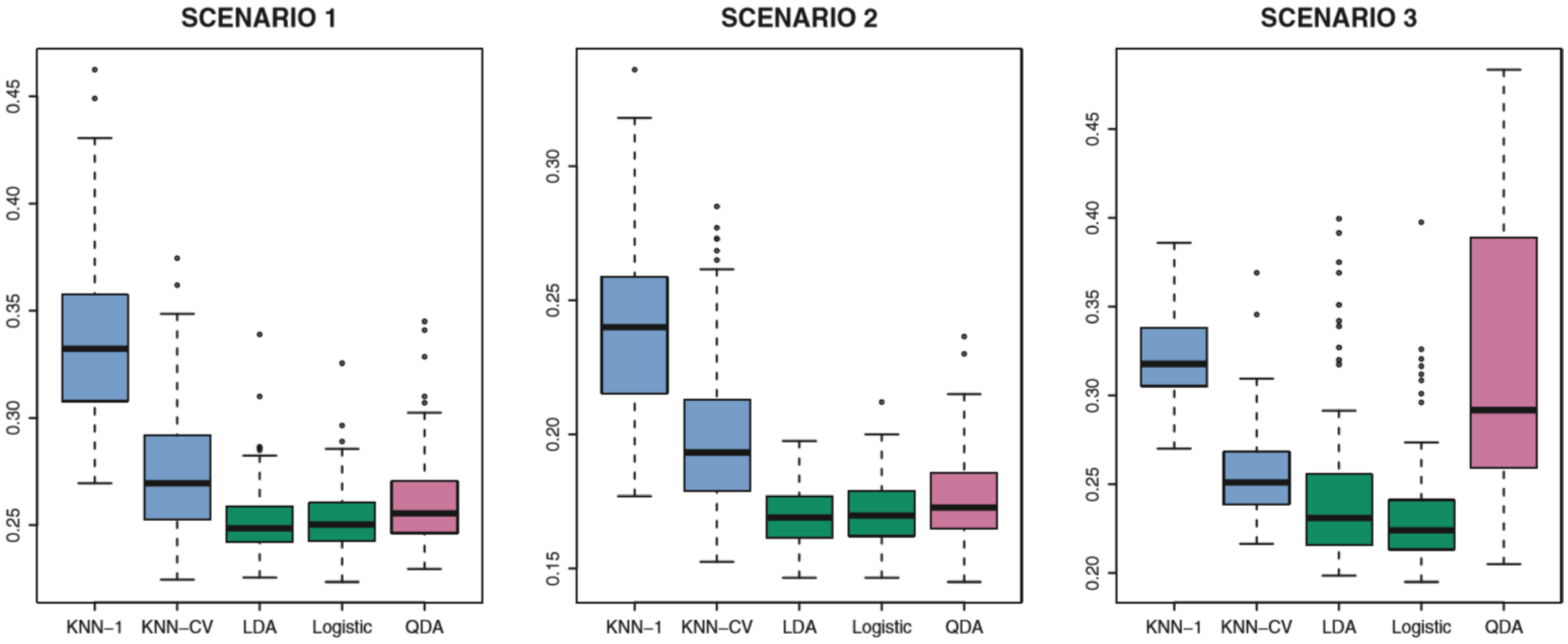
$$\begin{aligned} \delta_k(x) &= -\frac{1}{2}(x-\mu_k)^T \boldsymbol{\Sigma}_k^{-1}(x-\mu_k) + \log \pi_k \\ &= -\frac{1}{2} x^T \boldsymbol{\Sigma}_k^{-1} x + x^T \boldsymbol{\Sigma}_k^{-1} \mu_k - \frac{1}{2} \mu_k^T \boldsymbol{\Sigma}_k^{-1} \mu_k + \log \pi_k \end{aligned}$$

QDA VS. LDA VS. Bayes Classifier

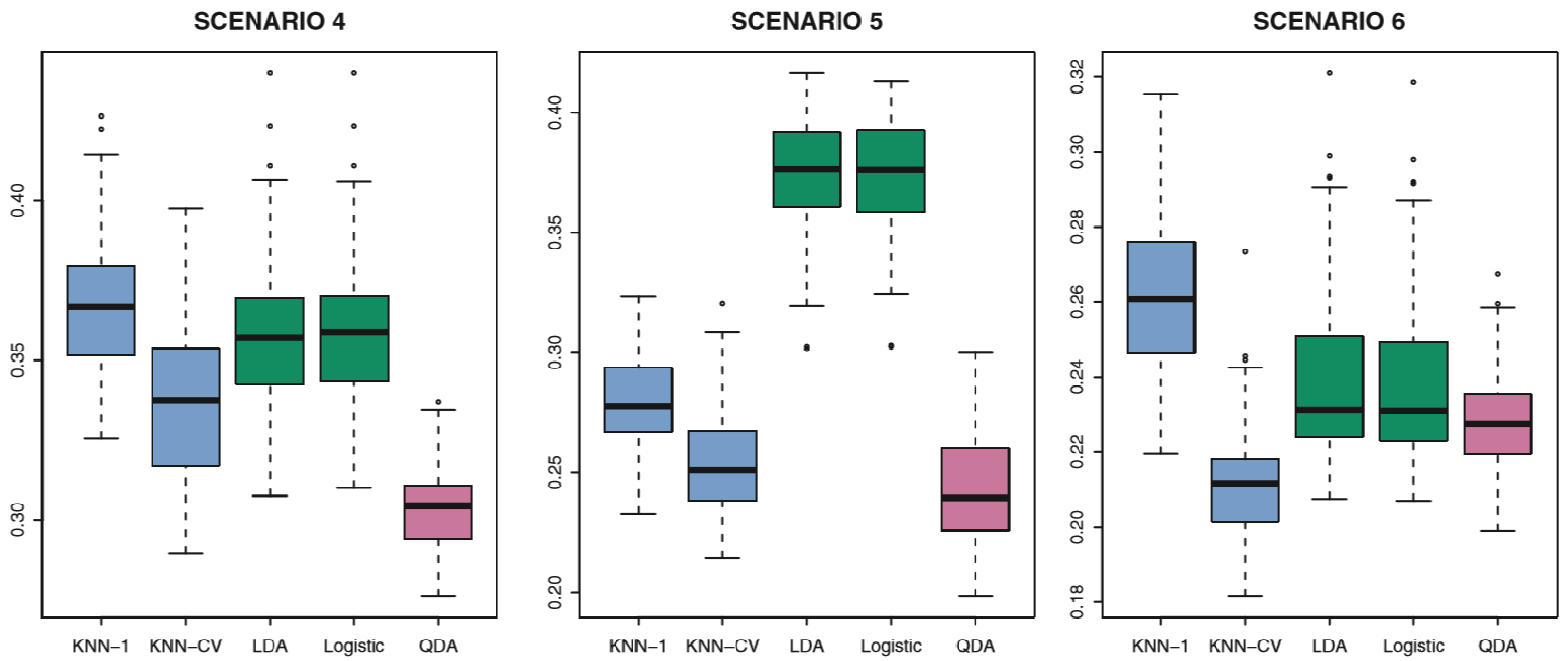


SW

선형 문제에 대한 예측 결과 에러 값 비교



비선형 문제에 대한 예측 결과 에러 값 비교



Thank you!