

데이터엔지니어를 위한 데이터 분석 기초

- 1) 확률분포
- 2) 회귀분석
- 3) Machine Learning 방법론

화물분포

◆ 확률변수(random variable)

- ✓ 표본공간에서 정의된 실수 함수
- ✓ 불확실성을 가지는 사회적·자연적 현상을 일종의 확률실험으로 이해
- ✓ 여기서 얻어진 표본공간을 숫자로 표시하여 불확실한 현상을 수학적
으로 모형화 함
- ✓ 이를 통해 구체적으로 계량화된 분석을 할 수 있음

- ✓ 확률변수는 변수가 취하는 값에 따라 이산확률변수와 연속확률변수로 나눔
- ✓ **이산확률변수(Discrete random variable):** 확률변수가 가질 수 있는 값들이 가산(countable) 또는 셀 수 있는 경우
 - '가산' 또는 '셀 수 있다'는 말은 확률변수의 값들이 자연수 1, 2, 3, ... 과 대응 관계를 가진다는 뜻
 - 예) 불량품의 개수, 사고건수, ...
- ✓ **연속확률변수(Continuous random variable):** 가질 수 있는 값이 셀 수 없을 정도로 많은 경우
 - 예) 수명, 신장, 체중
- ✓ 이산형과 연속형의 구분이 명확하지 않은 경우, 가정의 적절성이나 분석의 난이도 등을 고려하여 적절하게 선택

◆ 확률분포(Probability Distribution)

- ✓ 확률변수는 표본공간의 값을 숫자로 바꾼 함수이기 때문에 확률변수가 어떤 값을 가진다는 것은 표본공간 내에 대응하는 원소들이 존재
 - $X = x$ 이면 표본공간에 $\{\omega | X(\omega) = x, \omega \in \Omega\}$ 를 만족하는 사건이 존재
 - 임의의 상수 a, b 에 대해 $a \leq X \leq b$ 이면 이에 해당하는 사건 $\{\omega | a \leq X(\omega) \leq b, \omega \in \Omega\}$ 이 존재
- ⇒ 이는 확률변수에 대해 $X = x$ 또는 $a \leq X \leq b$ 에 대응하는 확률을 계산할 수 있음

- 동전을 세 번 던지기

$$P(X = 0) = P(\{TTT\}) = \frac{1}{8}$$

$$P(X = 1) = P(\{HTT, THT, TTH\}) = \frac{3}{8}$$

$$P(X = 2) = P(\{HHT, HTH, THH\}) = \frac{3}{8}$$

$$P(X = 3) = P(\{HHH\}) = \frac{1}{8}$$

- ✓ 확률변수는 숫자로 표시되어 특정 지점이나 영역에서의 확률을 표시할 수 있어 확률이 어떤 형태로 분포되었다는 말을 할 수 있음
- ✓ 확률변수가 가질 수 있는 값에 대해 확률을 표시한 것을 **확률분포 (probability distribution)**라고 함
- ✓ 확률분포표(probability distribution table): 확률변수의 확률을 표로 표시한 것
 - 예) 동전 세 번 던지기: 앞면의 수 X

X	0	1	2	3
$P(X = x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

✓ 확률은 모집단이 어떤 형태로 이루어져 있는지를 보여줌

⇒ 확률분포 또한 모집단을 숫자로 표시했을 때의 형태를 표시한 것

= 모집단의 확률 구조

✓ 모집단의 확률 구조를 표시하는 방법

- 이산확률변수: **확률질량함수(probability mass function)**, 누적분포함수
- 연속확률변수: **확률밀도함수(probability density function)**, 누적분포함수

◆ 확률질량함수(Probability Mass Function, pmf)

- ✓ 이산확률변수 X 가 임의의 값 x 일 확률 $P(X = x)$ 를 x 에 대한 함수로
생각

$$f(x) = P(X = x)$$

- ✓ 경우에 따라 확률변수 X 를 강조하기 위해 $f_X(x)$ 로 표시

■ 동전을 세 번 던지기

✓ X : 앞면의 수 $\rightarrow X$ 가 가질 수 있는 값은 $x = 0, 1, 2, 3$

$$f(0) = \frac{1}{8} \quad f(1) = \frac{3}{8} \quad f(2) = \frac{3}{8} \quad f(3) = \frac{1}{8}$$

✓ Y : 앞면과 뒷면의 수의 차이 $\rightarrow y = 1, 3$

$$f_Y(1) = \frac{6}{8} = \frac{3}{4} \quad f_Y(3) = \frac{2}{8} = \frac{1}{4}$$

■ 앞면이 나올 때 까지 동전을 던지기

✓ X : 던진 횟수

$$f(1) = P(X = 1) = P(\{H\}) = \frac{1}{2}$$

$$f(2) = P(X = 2) = P(\{TH\}) = \frac{1}{4} = \left(\frac{1}{2}\right)^2$$

$$f(3) = P(X = 3) = P(\{TTH\}) = \frac{1}{8} = \left(\frac{1}{2}\right)^3$$

⋮

$$\Rightarrow f(x) = \left(\frac{1}{2}\right)^x \quad x = 1, 2, 3, \dots$$

- 기하분포(Geometric Distribution)

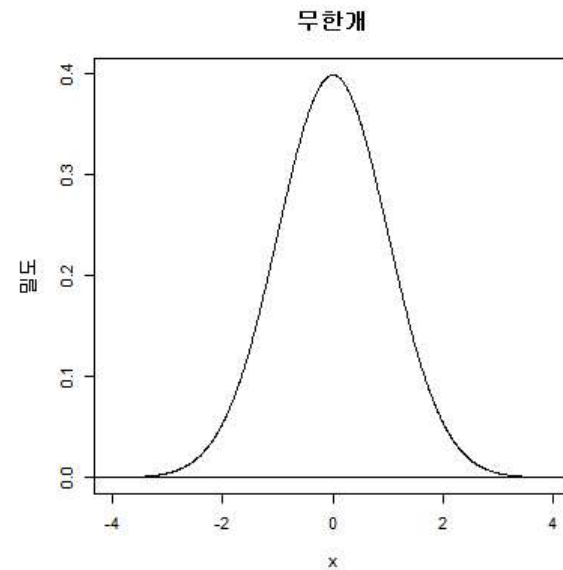
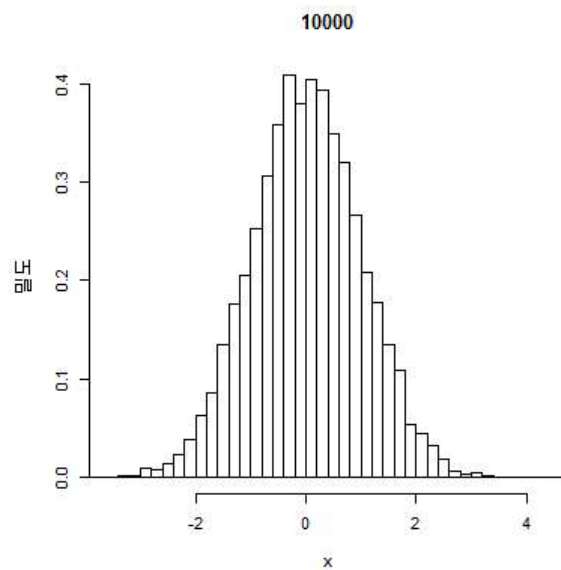
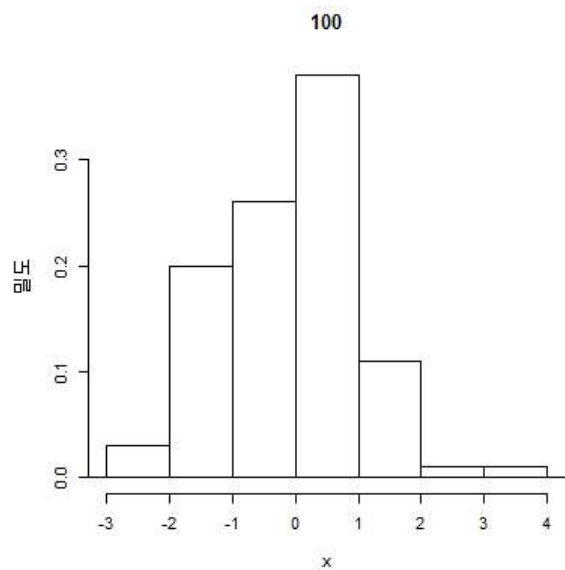
✓ Y : 뒷면의 수

- $Y = X - 1$ 의 관계를 가지며 해당 확률은 동일

$$f(0) = \frac{1}{2} \quad f(1) = \frac{1}{4} = \left(\frac{1}{2}\right)^2 \quad f(2) = \frac{1}{8} = \left(\frac{1}{2}\right)^3 \quad \dots$$

$$\Rightarrow f_Y(y) = \left(\frac{1}{2}\right)^{y+1} \quad y = 0, 1, 2, 3, \dots$$

◆ 확률밀도함수(Probability Density Function, pdf)



- ✓ 세 번째 그림은 연속확률변수 x 의 분포형태 모집단의 형태를 나타낸 것으로 임의의 지점 x 에서의 밀도를 $f(x)$ 라고 표시하면 $f(x)$ 를 확률 밀도함수라고 함

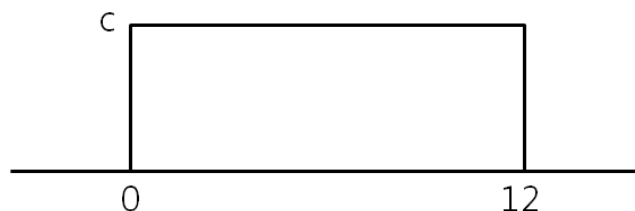
■ 0~12까지의 숫자가 표시된 돌림판

✓ 표본공간: $\Omega = \{x: 0 < x \leq 12\}$

✓ X : 바늘이 지적하는 위치

✓ 0에서 12사이에서 발생가능성이 동일

→ 밀도는 이 구간에서 동일: $f(x) = c$



✓ 전체 면적은 1이 되어야 하므로 $c = 1/12$

$$f(x) = \frac{1}{12} \quad 0 < x \leq 12$$

■ 확률밀도함수에서의 확률

- ✓ 히스토그램에서 면적이 해당 구간에서의 비율(상대도수)
- ✓ 확률밀도함수에서의 면적이 해당 구간에서의 확률
- ✓ X 가 구간 $[a, b]$ 에 속할 확률

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

- 예) X 가 3에서 6사이에 있을 확률

$$P(3 \leq X \leq 6) = \frac{3}{12} = \frac{1}{4}$$

- ✓ $X = 3$ 일 확률은?

- ✓ 어떤 점에서는 면적은 $f(x)$ 의 크기와 관계없이 항상 0
- ✓ X 가 연속확률변수일 때에는 모든 x 에 대해 $P(X = x) = 0$
- ✓ 확률밀도함수 $f(x)$ 는 x 에서의 확률이 아니라 상대적인 밀도를 나타내는 것
- ✓ X 가 연속확률변수이면

$$P(a < X < b) = P(a < X \leq b) = P(a \leq X < b) = P(a \leq X \leq b)$$

◆ 이산확률분포(Discrete probability distribution)

■ 베르누이(Bernoulli) 시행

✓ 각 실험에서 발생 가능한 결과는 단 2가지

- 예) (성공/실패), (앞면/뒷면)

✓ 각 실험이 독립적으로 수행

✓ 모든 실험에서 결과의 확률은 항상 동일


- $P(S) = p, P(F) = 1 - p = q$

■ 베르누이 확률변수

✓ 성공할 확률 = p 인 경우 $X \sim B(p)$ 로 표시함

- $X = \begin{cases} 1, & \text{성공} \\ 0, & \text{실패} \end{cases}$

- $P(X = 1) = P(\text{성공}) = p, \quad P(X = 0) = P(\text{실패}) = 1 - p$

 $f(x) = P(X = x) = p^x(1 - p)^{1-x}, \quad x = 0, 1$

✓ 기댓값

- $E(X) = p = P(\text{성공}) \qquad E(X^2) = p = P(\text{성공})$

$$Var(X) = p(1 - p) = P(\text{성공})P(\text{실패})$$

■ 이항분포(Binomial distribution)

✓ 성공할 확률이 p 인 베르누이 실험을 n 번 반복했을 때, 성공 횟수의 분포

✓ 성공 횟수 X 는 n 개의 베르누이 확률변수를 합한 것

$$\begin{array}{ccccccc}
 & X_1 & + & X_2 & + & \dots & + & X_n & = & X \\
 S & 1 & & 1 & & & & 1 & & \downarrow \\
 F & 0 & & 0 & & & & 0 & & \text{성공횟수}
 \end{array}$$

- $X_i \sim B(p)$

✓ 베르누이 시행은 독립을 의미함

- $E(X) = np$

- $Var(X) = np(1 - p)$

▪ 주사위 세 번 던지기

✓ X : 1이 나온 횟수(1이면 S , 아니면 F)

0

FFF

1

SFF
 FSF
 FFS

2

SSF
 FSS
 SFS

3

FFF

$$\binom{3}{0} \left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^3$$

$$\binom{3}{1} \left(\frac{1}{6}\right)^1 \left(\frac{5}{6}\right)^2$$

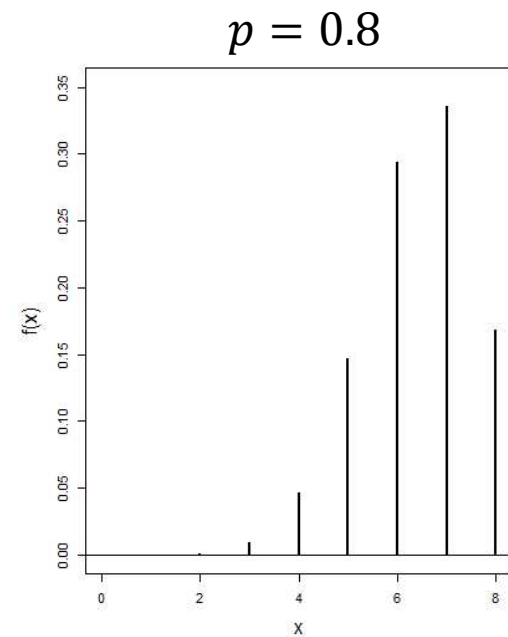
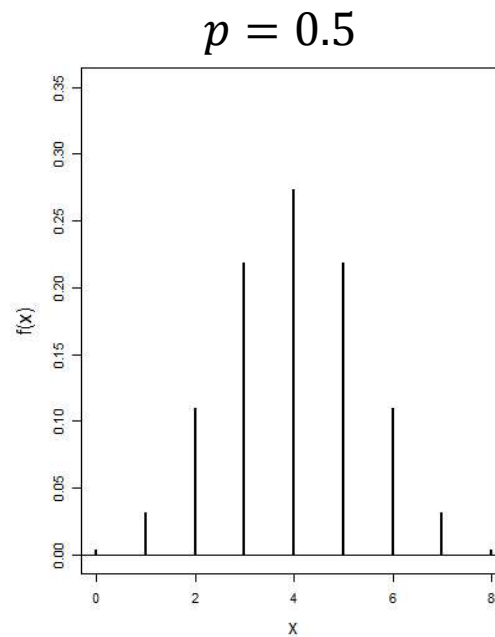
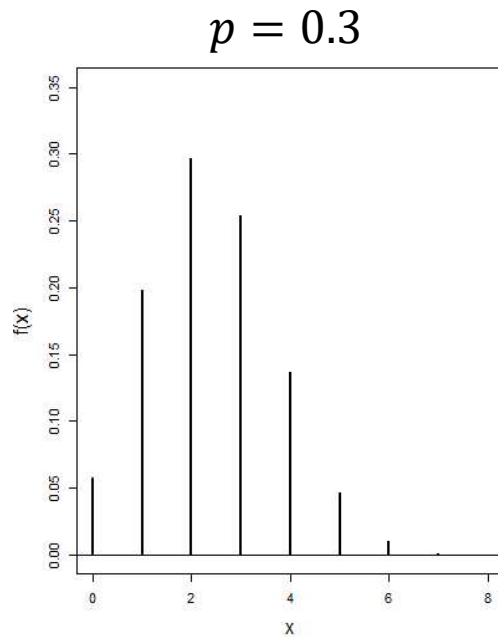
$$\binom{3}{2} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^1$$

$$\binom{3}{3} \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^0$$

✓ 일반식: $f(x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 1, 2, \dots, n$

✓ 표시: $X \sim B(n, p)$

- n 은 시행횟수, p 는 성공확률



✓ n 과 p 의 값은 분포의 모양을 결정함

⇒ 분포의 특성을 완전히 결정하는 값을 모수(parameter)라고 함

✓ 분포의 모수를 알면 해당 분포의 모든 것을 알 수 있음

■ 항암제 완치율

- ✓ 어떤 암에 대한 기존 항암제의 완치율은 50%
- ✓ 어느 제약회사에서 새로운 항암제를 개발하여 항암제의 효과를 확인하기 위해 15명의 환자를 대상으로 실험함
- ✓ 만약 새로운 항암제의 완치율이 기존과 같다면
 - 8명이 완치될 확률은?
 - 적어도 10명까지 치유될 확률은?
- ✓ 환자 중 12명의 환자가 완치되었다면, 기존보다 새로운 항암제의 효과가 있다고 할 수 있는가?

■ 포아송 분포(Poisson Distribution)

✓ 발생 가능성이 희박한 사건이 임의의 구간 안에서 평균적으로 λ 번 발생 할 때, 이 사건이 일어날 횟수의 분포

✓ 확률질량함수

$$f(x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, 3, \dots$$

✓ 표시: $X \sim P(\lambda)$

■ 컴퓨터 프로그램 버그

- ✓ 500개 모듈당 평균 한 개의 버그 발생
- ✓ 독립적으로 제작된 1500개 다른 모듈로 이루어진 프로그램 패키지에
서 버그가 2개 이하일 확률은?

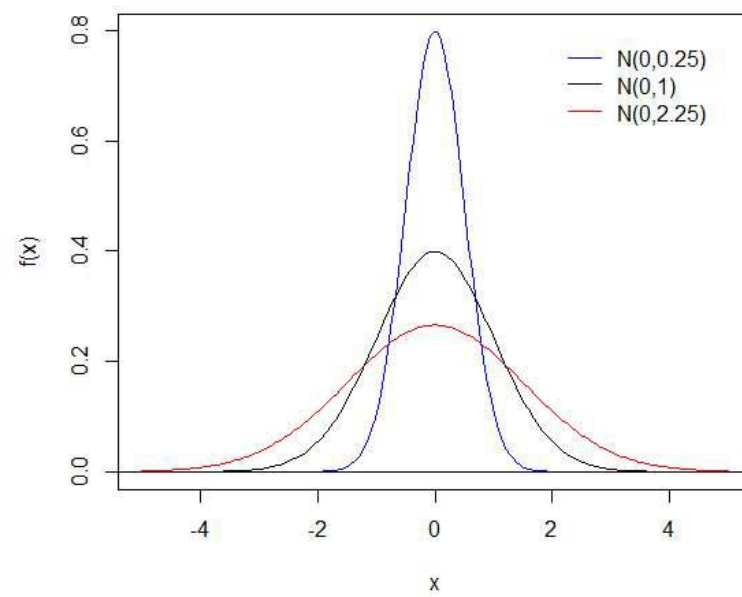
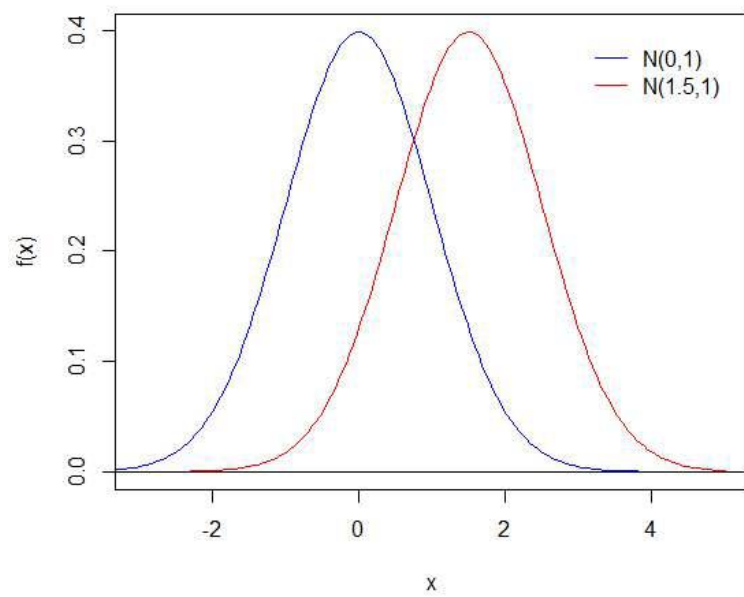
◆ 연속확률분포(Continuous probability distribution)

■ 정규분포(Normal Distribution)

- ✓ Gauss가 각종 물리실험을 수행할 때 발생하는 측정오차를 설명하기 위해 적용한 분포
- ✓ 모든 학문 분야에서 확률모형 또는 근사모형으로 사용
- ✓ 평균은 중심위치를 종모양(bell-shaped)의 대칭형태를 가짐
- ✓ 평균이 μ 이고 분산이 σ^2 인 정규분포의 확률밀도 함수

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \quad -\infty < x < \infty$$

- 표시: $X \sim N(\mu, \sigma^2)$



■ 표준정규분포(standard normal distribution)

✓ $\mu = 0$ 이고 $\sigma^2 = 1$ 인 경우의 정규분포

- 0을 중심으로 대칭

✓ 일반적으로 Z 로 표시: $Z \sim N(0,1)$

✓ 정규분포를 표준화하여 표준정규분포를 만듦

$$Z = \frac{X - \mu}{\sigma}$$

✓ 선형변환된 정규확률변수도 정규분포를 따름

$$Z \sim N(0,1) \rightarrow X = \sigma Z + \mu, X \sim N(\mu, \sigma)$$

■ 정규분포의 정리

✓ $X \sim N(\mu, \sigma)$ 이고 $a \neq 0$ 이면, $aX + b \sim N(a\mu + b, a^2\sigma^2)$

✓ 두 정규확률변수의 선형 결합도 정규분포를 따름

$$\text{- } X_1 \sim N(\mu_1, \sigma_1^2) \quad X_2 \sim N(\mu_2, \sigma_2^2) \quad \sigma_{12} = \text{COV}(X_1, X_2)$$

$$X_1 \pm X_2 \sim N(\mu_1 \pm \mu_2, \sigma_1^2 + \sigma_2^2 \pm 2\sigma_{12})$$

- $\sigma_{12} = 0$ 이면, X_1 과 X_2 는 독립

■ 아침식사 예제: 빵과 우유를 먹는다고 가정

- ✓ 빵의 열량은 평균 200kcal, 표준편차 15kcal인 정규분포
- ✓ 우유의 열량은 평균 80kcal, 표준편차 5kcal인 정규분포
- ✓ 아침식사에서 300 칼로리 이상 섭취할 확률은?

회귀분석

◆ 회귀분석(Regression Analysis)

✓ 상관계수와 상관분석

- 상관계수: 두 변수간 직선관계의 정도를 나타내는 척도

$$-1 \leq \rho \leq 1$$

음수: 음의 상관관계(반비례) -1: 반비례직선
양수: 양의 상관관계(정비례) +1: 정비례직선

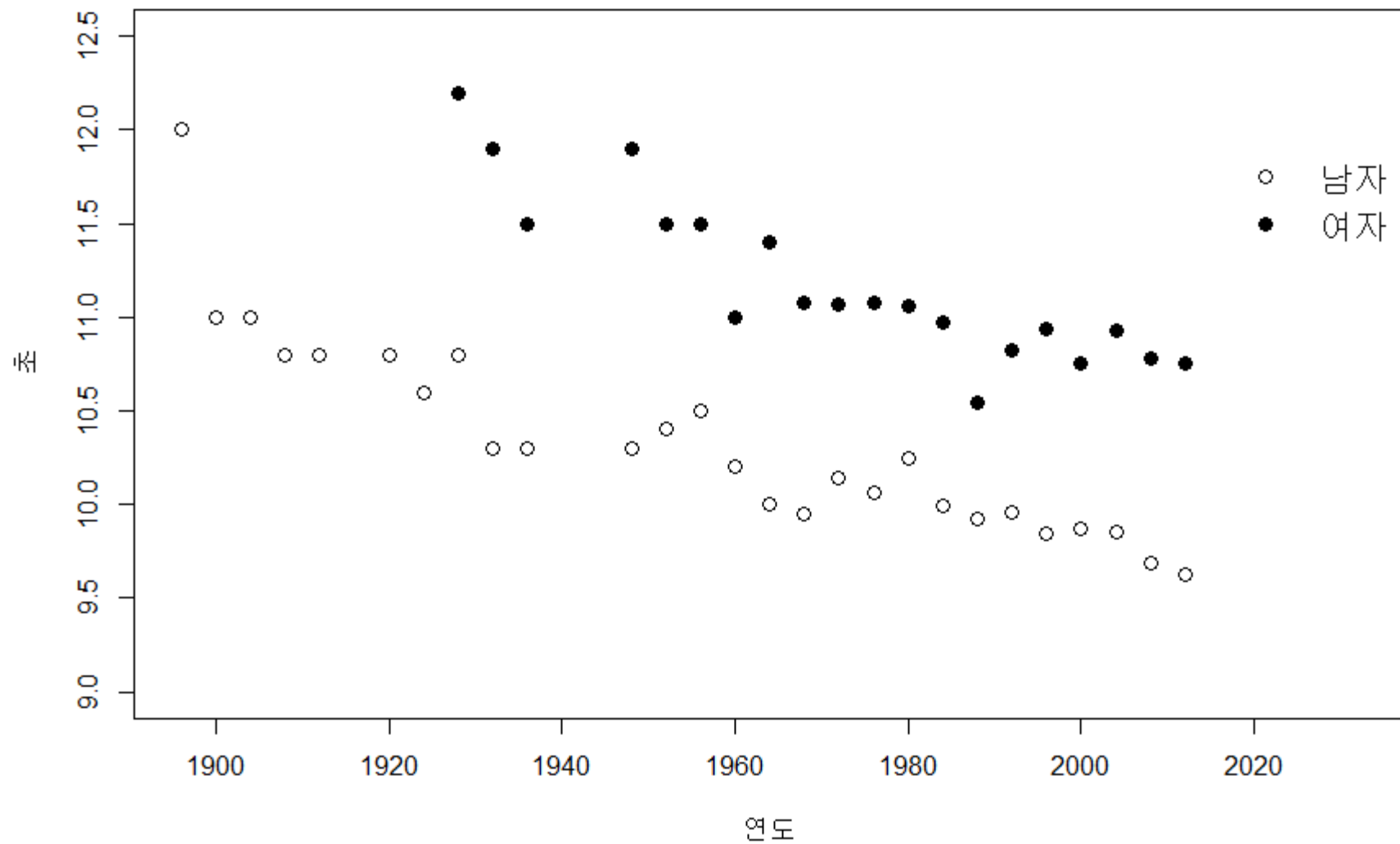
- Pearson 표본상관계수

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

✓ 1896년 아테네 올림픽부터 2012년 런던 올림픽까지 100미터 육상경기의 우승 기록

연도	우승기록		연도	우승기록		연도	우승기록		연도	우승기록	
	남자	여자		남자	여자		남자	여자		남자	여자
1896	12	-	1928	10.8	12.2	1964	10	11.4	1992	9.96	10.82
1900	11	-	1932	10.3	11.9	1968	9.95	11.08	1996	9.84	10.94
1904	11	-	1936	10.3	11.5	1972	10.14	11.07	2000	9.87	10.75
1908	10.8	-	1948	10.3	11.9	1976	10.06	11.08	2004	9.85	10.93
1912	10.8	-	1952	10.4	11.5	1980	10.25	11.06	2008	9.69	10.78
1920	10.8	-	1956	10.5	11.5	1984	9.99	10.97	2012	9.63	10.75
1924	10.6	-	1960	10.2	11.0	1988	9.92	10.54	2016	?	?

육상 100미터 올림픽 우승기록



통계량	남자		여자	
	기록	연도	기록	연도
표본수	24		18	
평균	10.318	1954.333	11.23	1968.667
제곱합	2558.401	91690624	2273.387	69771024
교차제곱합	483681.1		397789	

$$r_{\text{남자}} = \frac{-270.43}{\sqrt{24573.33}\sqrt{3.376}} = -0.939$$

$$r_{\text{여자}} = \frac{-157.32}{\sqrt{9352}\sqrt{3.355}} = -0.888$$

■ 회귀모형(Regression Model)

- ✓ 두 변수의 인과관계를 유도함

$$\begin{array}{ccccc} X & \rightarrow & f & \rightarrow & Y \\ \text{input} & & \text{system} & & \text{output} \end{array}$$

- 입력변수 X : 설명변수(explanatory variable),
독립변수(independent variable)
- 출력변수 Y : 반응변수(response variable),
종속변수(dependent variable)

- ✓ 단순선형 회귀모형(simple linear regression model)

$$Y = \beta_0 + \beta_1 X + e$$

- ✓ 다중회귀모형(multiple linear regression)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + e$$

- ✓ 다변량 회귀모형(multivariate regression)

$$(Y_1 \quad \cdots \quad Y_p) = XB + e$$

✓ 오차: 실제 관측값과 회귀직선간의 차

$$e = Y - \beta_0 - \beta_1 X$$

✓ 잔차: 실제 관측값과 추정된 회귀직선간의 차

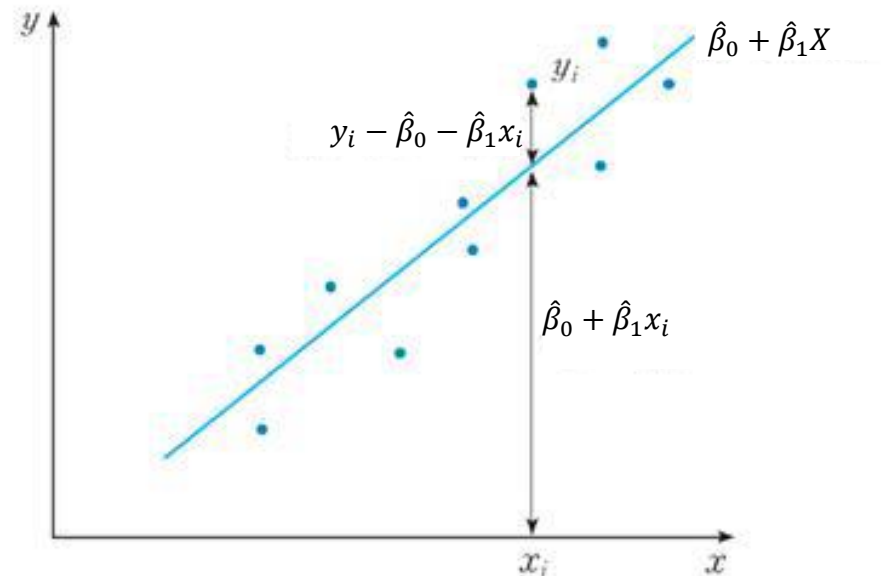
$$\hat{e} = Y - \hat{\beta}_0 - \hat{\beta}_1 X$$

✓ 최소제곱법(least square method)

$$\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \text{ 를 최소화}$$

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

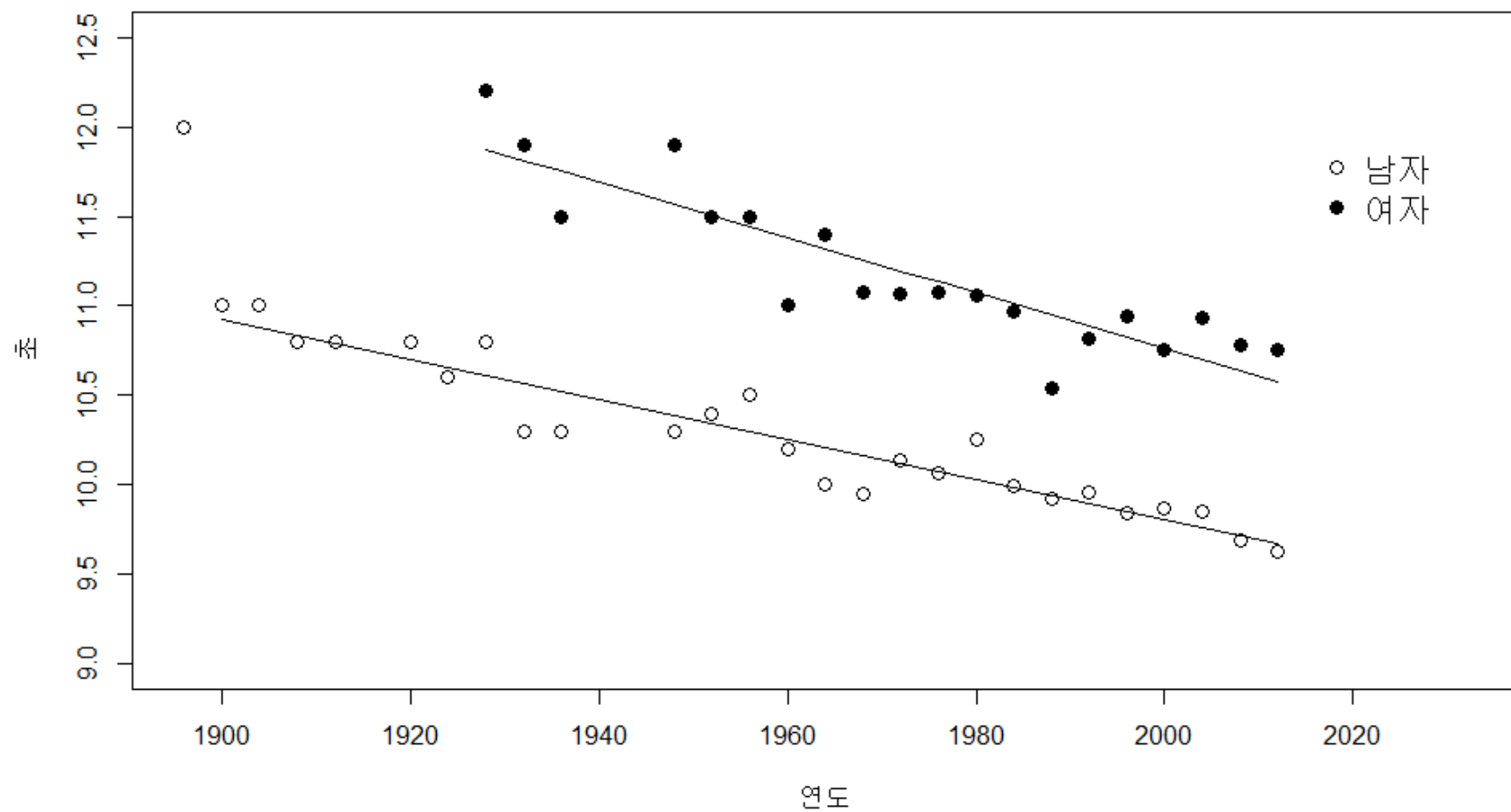
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$



■ 올림픽 육상 100m 우승기록(1900~2004)

	남자	여자
\bar{y}	10.318	11.23
\bar{x}	1954.333	1968.667
S_{xy}	-270.443	-157.32
S_{xx}	24573.33	9352
$\hat{\beta}_1$	$-270.443/24573.33 = -0.011$	$-157.32/9352 = -0.0168$
$\hat{\beta}_0$	$10.318+0.011 \times 1954.3 = 31.816$	$11.23+0.0168 \times 1968.7 = 44.3$
회귀식	$\hat{y}_i = 31.816 - 0.011x_i$	$\hat{y}_i = 44.3 - 0.0168x_i$

육상 100미터 올림픽 우승기록



■ 회귀모수의 추론

✓ 회귀계수 검정

$$- H_0: \beta_1 = 0 \text{ vs. } H_1: \beta_1 \neq 0$$

$$- \frac{\hat{\beta}_1 - \beta_1}{\sqrt{MSE/S_{xx}}} \sim t_{n-2}$$

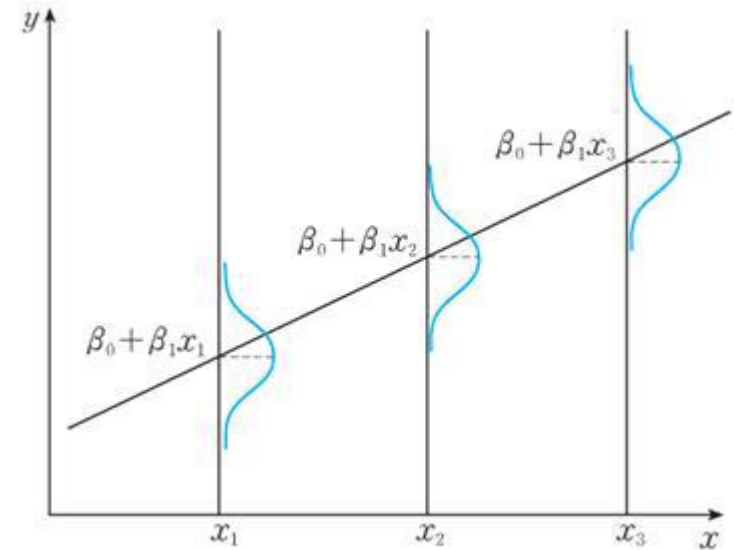
→ p-값이 0.05보다 작으면, 귀무가설을 기각

→ 기울기가 통계적으로 유의미함

■ 회귀모수의 추론

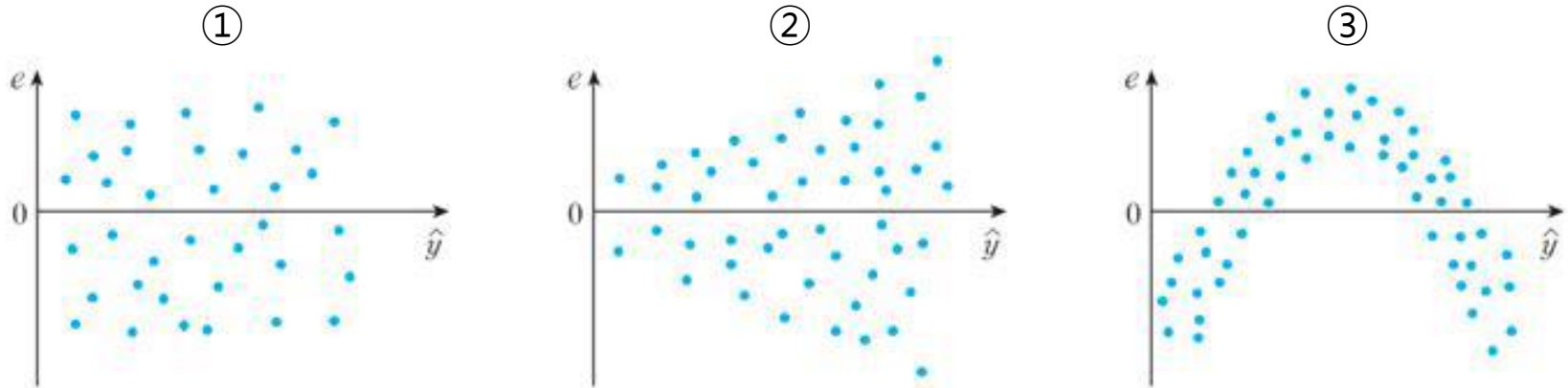
✓ 오차항 가정

- $E(\varepsilon_i) = 0$
- $Var(\varepsilon_i) = \sigma^2$ (등분산성)
- ε_i 는 서로 독립임(독립성)
- ε_i 는 정규분포를 따름(정규성)



➡ $Y_i \sim iid N(\beta_0 + \beta_1 x_i, \sigma^2)$

✓ 잔차 검정



① 특정 패턴이 없으며 등분산성을 만족함

② \hat{y} 가 커지면서 잔차의 표준편차가 커지는 경향이 있음

→ 등분산성 가정을 만족하지 않음 → 종속변수를 변환

③ 잔차가 \hat{y} 와 2차 곡선의 관계를 가짐

→ 모형의 관계식이 잘못됨

※ 정규성검정: 히스토그램, QQ-plot, shapiro-wilk's 검정, K-S 검정

※ 독립성검정: durbin-watson 검정

■ 회귀모수의 추론

✓ 결정계수(coefficient of determination)

- 모형이 어느 정도 적합한지를 나타내는 척도
- 변동 분해

$$\sum_{\text{SST}} (y_i - \bar{y})^2 = \sum_{\text{SSR}} (\hat{y}_i - \bar{y})^2 + \sum_{\text{SSE}} (y_i - \hat{y}_i)^2$$

- SST: y의 총변동
- SSR: 모형으로 설명되는 변동
- SSE: 모형으로 설명되지 않는 변동

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad \leftarrow \text{상관계수의 제곱}$$

■ 회귀모수의 추론

✓ 결정계수(coefficient of determination)

- $0 \leq R^2 \leq 1$

- 1에 가까운 경우 회귀모형이 관측 결과를 잘 설명한다고 할 수 있음

- 0에 가까운 경우 두 변수간의 관계가 선형관계가 아니거나 독립변수가 종속변수에 영향을 미치지 못한다고 할 수 있음

- 올림픽 육상 100m 우승기록(1900~2004)

	남자	여자
회귀식	$\hat{y}_i = 31.816 - 0.011x_i$	$\hat{y}_i = 44.3 - 0.0168x_i$
기울기 검정통계량	-12.85	-7.72
P-value	5.28e-12	4.38e-07
R^2	0.882	0.789

■ 다중공선성

다중공선성이란 입력변수들 간의 상관관계가 존재하여 회귀 계수의 분산을 크게 하기 때문에, 회귀 분석 시 추정 회귀 계수를 믿을 수 없게 되는 문제가 발생하는 것을 말한다. 다중 회귀 모형에서 회귀 계수란 독립 변수의 변화에 따른 종속 변수의 변화량을 나타내기 때문에, 설명 변수들 사이에 유의한 상관관계가 존재하는 경우 한 설명변수를 다른 설명변수와의 함수 관계로 표시할 수 있다. 이러한 경우 회귀 계수의 분산이 증가하며, 회귀 계수 추정치가 불안하고 해석하기 어려워진다.

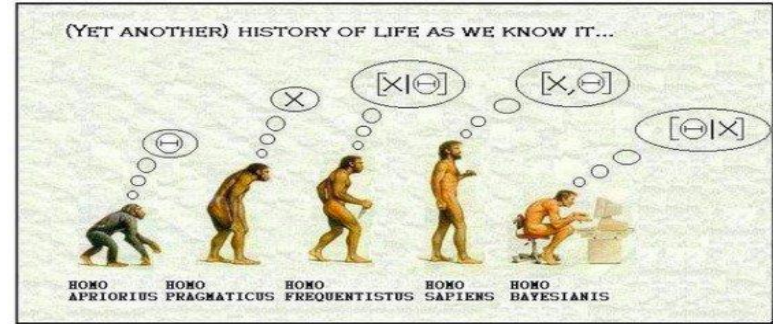
이러한 다중공선성을 측정하기 위해서는 변수의 상관구조를 조사할 수 있는데, 분산 팽창 인수(Variance Inflation Factor), 공차 한계(Tolerance), 상태지수를 조사하는 방법 등이 있다. 또한 다중공선성 문제를 해결하기 위해서는 문제를 일으키는 설명변수를 제거하거나 주성분 분석(PCA) 혹은 능형회귀분석(Ridge Regression)과 같은 다른 추정 방법을 이용할 수 있다.

Machine Learning 방법론

- 지도학습(Supervised Learning)
- 비지도학습(Unsupervised Learning)
- 기타방법론

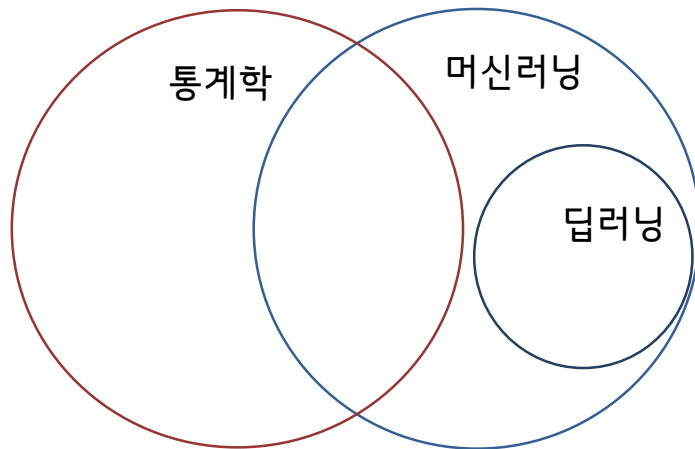
• 통계학(통계분석)

- 전수조사를 할 수 없었던 과거에 일부 표본을 이용하여 모집단의 상태를 진단/표현하기 위해 만들어진 분석 방법론
- 데이터의 요약, 변수간의 관계 등을 규명
- 현재 수집되지 않은 데이터 값을 추정/예측
- Frequentist Statistics, Bayesian Statistics



• 머신러닝

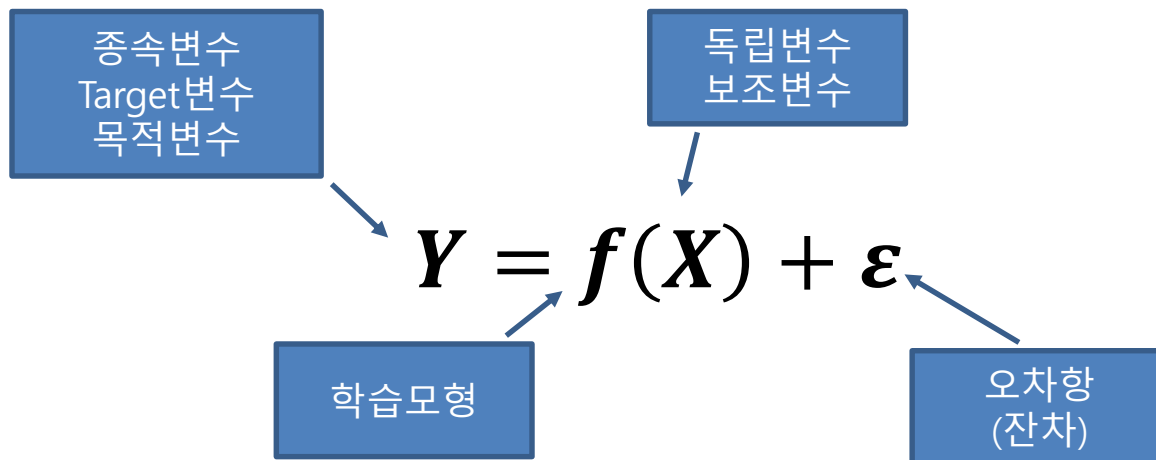
- 변수간 관계, 수집되지 않은 값의 추정/예측을 함수화(일반화) 하여 자동화에 적합한 형태로 만드는 분석방법론
- 지도학습, 비지도학습, 강화학습



통계학	머신러닝
Estimation	Learning
Data Point	Example/ Instance
Regression	Supervised Learning
Classification	Supervised Learning
Covariate	Feature
Response	Label

The the two fields are blending together more and more. 'Larry Wasserman.'

지도학습(Supervised Learning)



- 함수 $f()$ 의 형태에 따라 학습모형의 이름이 달라짐
- 종속변수 Y 와 독립변수 X 가 연속형 변수인 경우와 범주형 변수인 경우로 나누어 사용 가능한 $f()$ 를 고려해야 함

Y변수 형태	X변수 형태	사용 가능한 모형
연속형	명목형	분산분석모형(ANOVA)
	연속+명목형	선형회귀모형(Linear Regression)
범주형	연속+명목형	로지스틱 회귀모형, 포아송 회귀모형 (GLM), SVM
연속 / 범주형	연속+명목형	의사결정나무(Decision Tree), RandomForest 인공신경망 모형(Neural Network, Deep Learning)

비지도학습 (Unsupervised Learning)

군집화(Clustering) : 유사한 값을 갖는 관측치끼리 동일한 그룹으로 묶는 방법

ID	X1	X2	X3	X4	X5	군 집
1	1.03	0.15	8.12	4.14	0.55	A
2	1.05	0.54	9.14	4.11	0.41	A
3	1.33	1.43	7.56	4.09	0.56	B
4	1.54	1.12	9.41	4.08	0.88	B
5	2.65	0.54	7.13	3.11	3.12	C
6	2.88	0.98	8.41	3.00	3.01	C
7	2.56	0.34	8.00	3.09	2.98	C
8	3.41	1.44	9.14	3.03	2.79	D
9	3.50	0.99	7.32	1.00	3.21	D
10	3.31	1.21	8.03	1.01	2.97	D

※ 일반적으로 clustering은 군집의 개수를 지정해야 하며, 세분화 할수록 점점 군집 내 동질성이 강해짐

변수 유형에 따라 선택할 수 있는 (machine learning) 모델의 종류가 달라지므로 변수가 어떤 분포를 따르고 있는지 확인하는 과정이 중요함

[변수 유형에 따른 분석 모형의 종류]

X \ Y	연속형	이산형(범주형)
연속형	상관분석, 회귀분석, Tree, RandomForest, XGBoost, Light GBM, Deep Learning	GLM(로지스틱 회귀), TreeRandomForest, XGBoost, Light GBM, Deep Learning
이산형(범주형)	회귀분석(t-test, ANOVA), Tree, RandomForest, XGBoost, Light GBM, Deep Learning	GLM(로지스틱 회귀, 카이제곱 검정), Tree, RandomForest, XGBoost, Light GBM, Deep Learning
연속형 + 이산형	다중 회귀분석, Tree, RandomForest, XGBoost, Light GBM, Deep Learning	Tree, RandomForest, XGBoost, Light GBM, Deep Learning

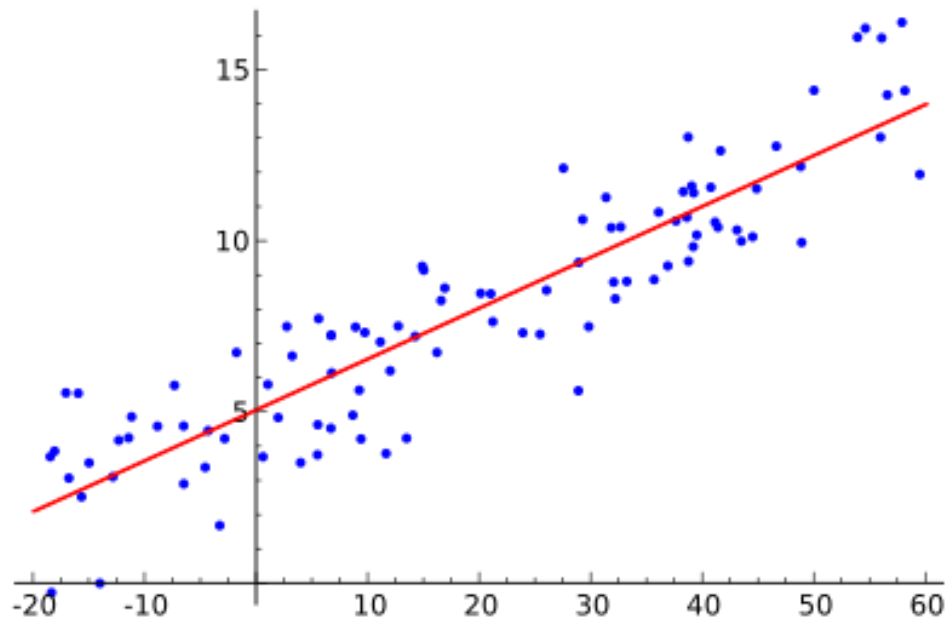
지도학습(Supervised Learning)

- 선형회귀모형(Linear Regression Model)

- $Y = \alpha + \beta X + \varepsilon$

X: 독립변수, 영향인자, Y: 종속변수, 목표변수

- 독립변수의 개수에 따라 단순선형회귀(simple linear regression, SLR), 다중선형회귀(Multiple linear regression)으로 나뉨
- $\beta=0$ 에 대해 통계적 유의성 검정을 실시하여 X가 Y에 유의한 영향을 미치는지 아닌지 판단함
- 목표변수는 추정된 관계식으로 얻어진 예측값(Yhat)과 오차(e)의 합으로 정의하고, 오차를 최소화 하는 함수식을 찾음.

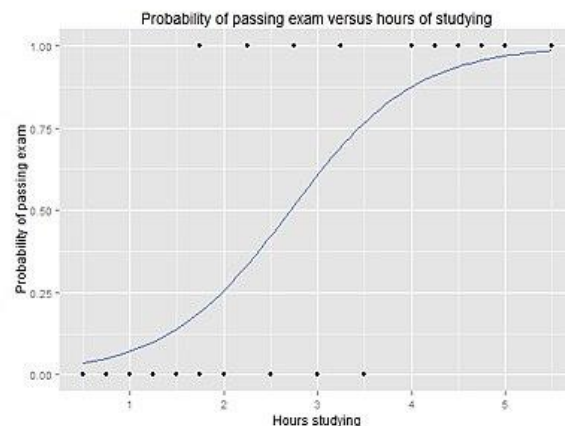


- 일반화 선형 모형(GLM; Generalized Linear Model)

- 선형회귀모형, 로지스틱회귀모형, 포아송회귀모형의 일반화된 형태(통칭)
- 종속변수의 평균을 정해진 함수로 변환한 값이 독립변수들의 선형결합 형태임을 가정한 모형
$$g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \dots$$
- $g()$ 를 link function이라고 부르며 종속변수의 분포에 따라 최적 link function이 정해져 있음
(이항분포 = logit, 포아송분포 = log)

- 로지스틱 회귀모형(Logistic Regression)

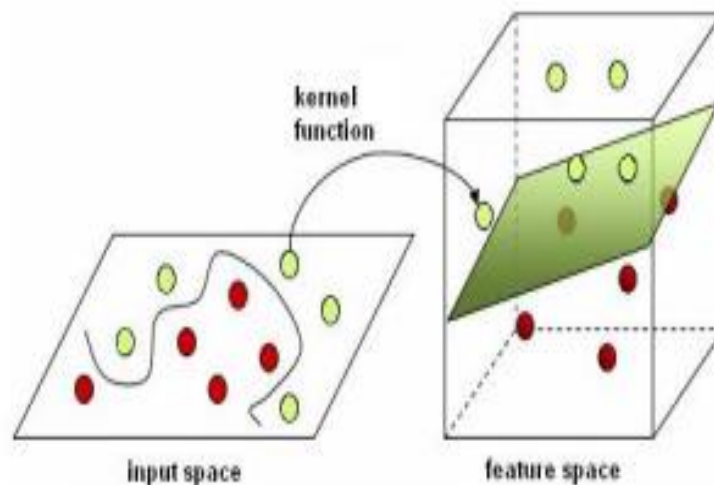
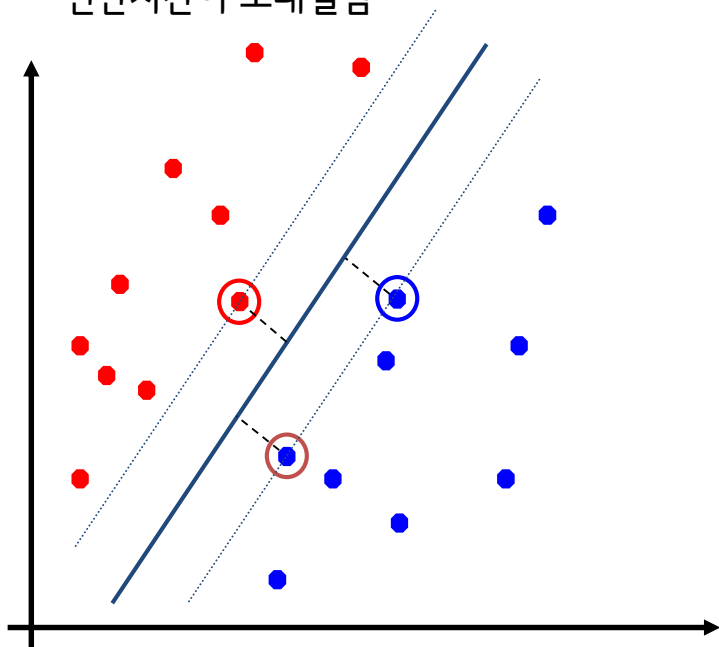
- $$P(y = 1) = \frac{\exp(\alpha + \beta X)}{1 + \exp(\alpha + \beta X)}, \quad y = \begin{cases} 1 \\ 0 \end{cases}$$
- 종속변수가 남/여, 성공/실패, 불량/정상등과 같이 두 가지 값을 갖는 경우 이항로지스틱모형(Binomial Logistic Regression Model)이라고 함
- 종속변수가 불교/천주교/기독교, 삼성/LG/롯데 등과 같이 세 종류 이상의 값을 갖는 경우 다항로지스틱모형(Multinomial Logistic Regression Model)이라고 함
- 독립변수가 종속변수의 특정 항목이 나타날 확률에 영향을 미치는 모델



지도학습(Supervised Learning)

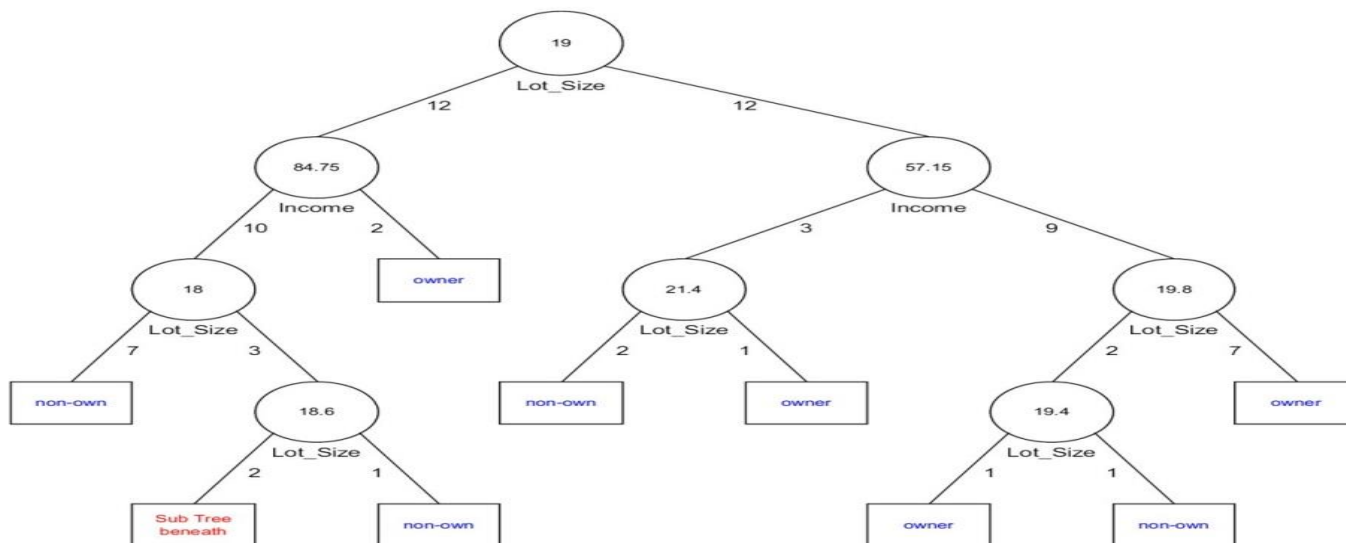
- SVM(Support Vector Machine)

- 군집을 분류하는 Support Vector들 중에서 Margin을 최대화 하는 분류기준을 찾음
- Kernel function으로 차원을 왜곡하여 최적 분류를 찾음
- LDA(Linear Discriminant Analysis)나 QDA(Quadratic Discriminant Analysis)에 비해 해석이 어려움
- 분류/수치예측 문제에 모두 적용 가능
- 과대적합(overfitting)경향이 적음
- Text 분류 문제에 가장 성능이 좋은 것으로 알려짐
- 연산시간이 오래걸림



• 의사결정나무(Decision Tree)

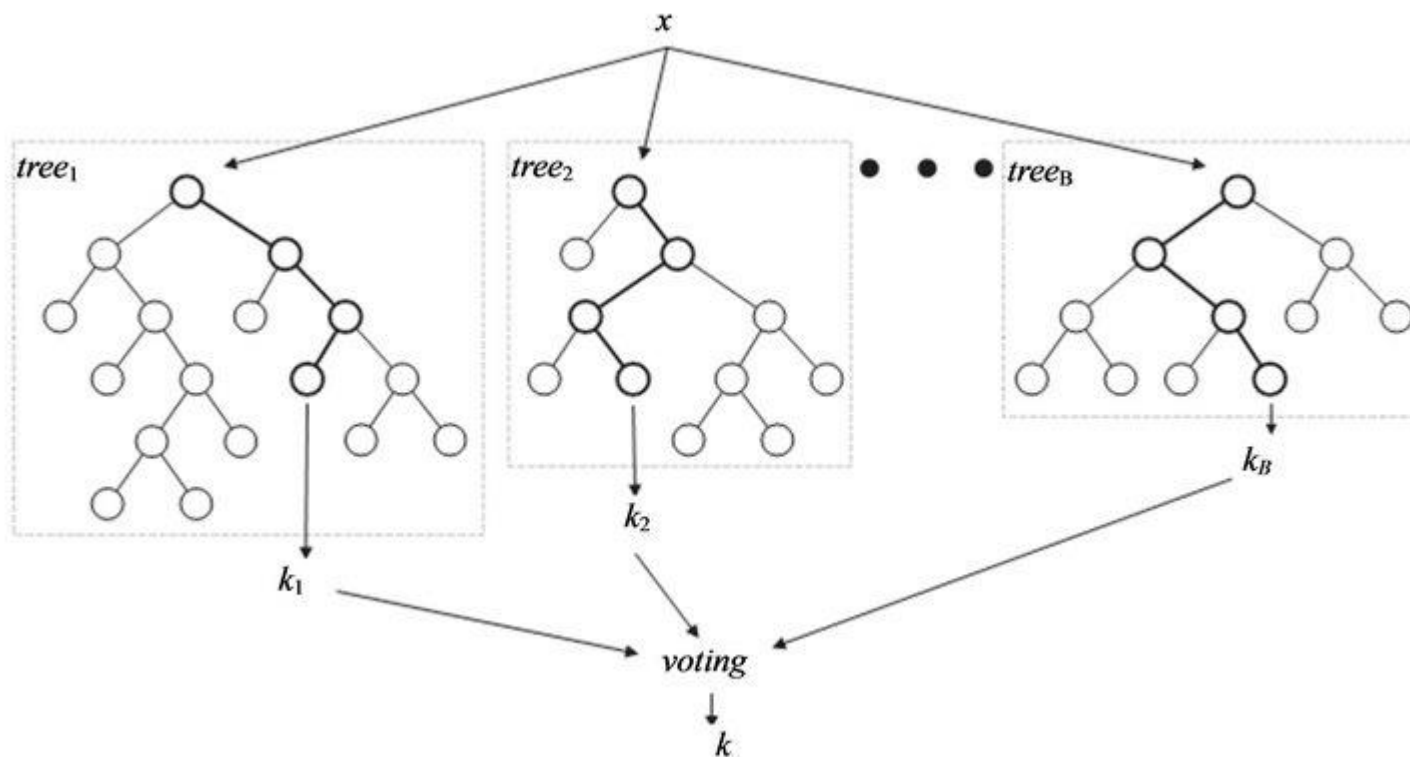
- 종속변수의 형태는 연속형, 이산형 모두 사용가능
- 그룹별 종속변수 평균의 차이가 가장 커지는 분리지점을 찾고 이를 반복적으로 시행하면서 데이터를 여러 조각으로 분류함
- 분리되기 전 최초의 데이터상태를 root, 분리된 각 덩어리를 leaf 혹은 node, 분리가 종료된 후 제일 마지막 덩어리를 terminal node라고 부름
- 각 덩어리간 종속변수 평균의 차이가 커지도록 분리하므로 terminal node내의 데이터들은 서로 동질적인 값을 갖게 됨
- 분리기준은 독립변수로 정의되고, 가장 먼저 분리시킨 독립변수가 가장 영향력이 큰 변수라고 할 수 있음
- 여러 독립변수들간의 상호연관성이 있는 경우 활용도가 높음(각 분리기준이 AND 조건으로 연결됨)



지도학습(Supervised Learning)

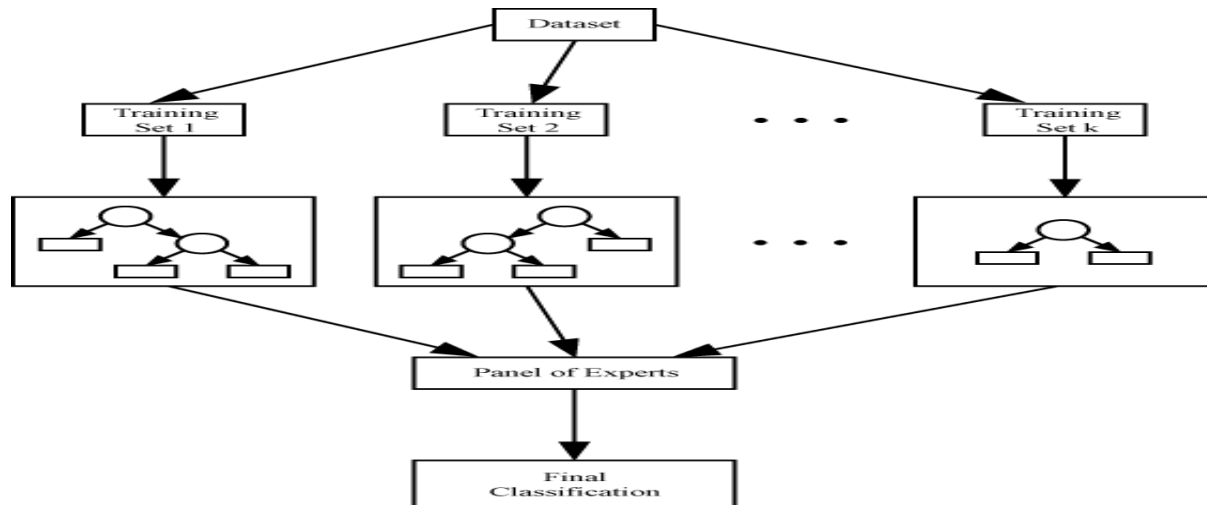
- 랜덤 포레스트(Random Forest)

- 여러 개의 트리를 생성하여 얻어진 결과들을 종합하여 최종결과를 도출하는 앙상블 기법
- 랜덤하게 독립변수를 선택하고 랜덤 표본으로 트리를 생성하는 과정을 반복하여 일반화 과정을 거치므로 다른 방법론에 비해 과적합(overfitting)가능성이 적고, 예측정확도가 높음
- 신경망 모형과 비슷한 정확도를 보이거나 구현이 더 쉬움(독립변수를 표준화할 필요 없음)



• Bagging(Bootstrap Aggregating)

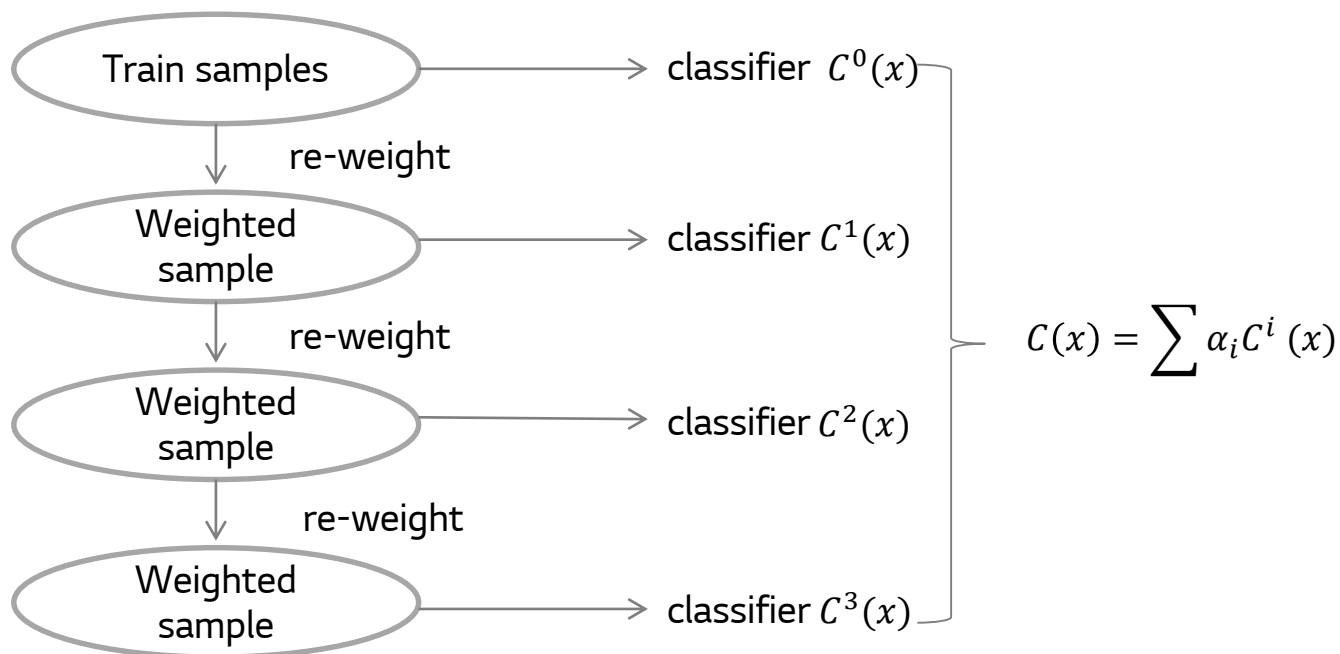
- 주어진 데이터를 모집단으로 가정하고 부표본(subsample)을 반복추출하여(복원추출) 표본이 얻어진 상황을 재현함
- 부 표본으로부터 분석을 진행하고 얻어진 분석결과를 부표본의 반복추출시 반복 재현, 기대값으로 최종결과 결정
- 한번에 처리할 수 없는 대용량 데이터에 대해 처리 가능한 크기의 부표본(subsample)을 이용하여 분석 수행 가능
- 반복수행이 많아 질수록, 원 데이터의 크기가 클수록, 표본의 크기가 커질수록 추정된 Bootstrap 추정치의 결과는 참값에 가까워짐
- 반복수를 많이 늘리거나 표본의 크기를 키워도 결국 원데이터의 모수로 수렴하게 되고 따라서 원 데이터의 크기가 작으면 Bootstrap추정치의 신뢰도가 떨어짐



지도학습(Supervised Learning)

• Boosting

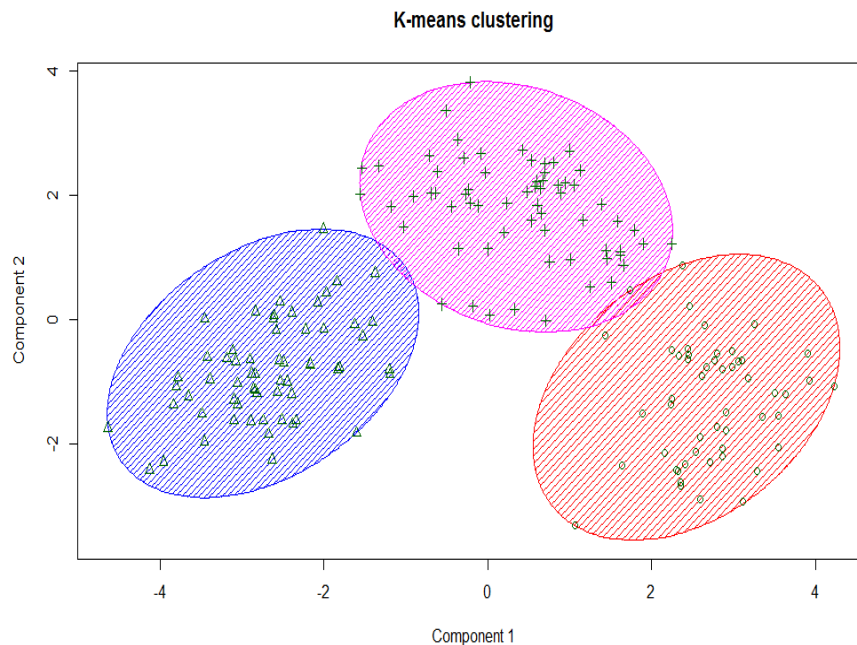
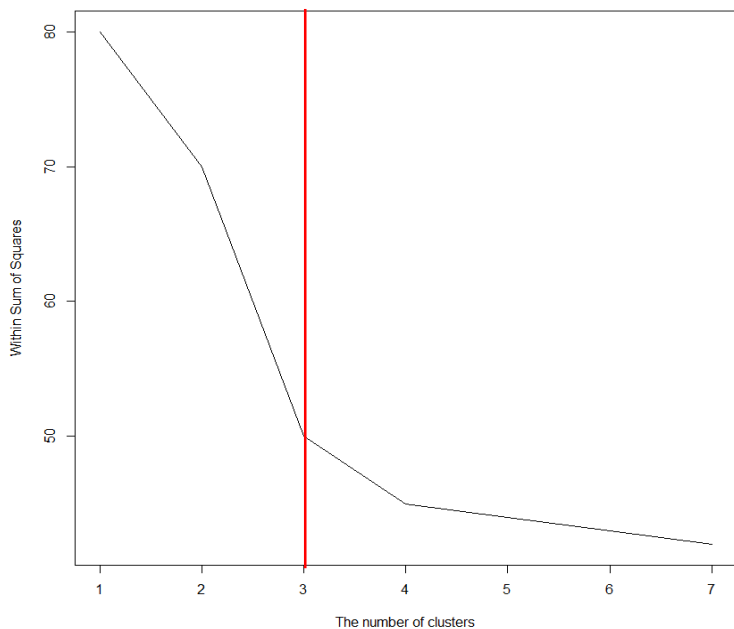
- Weak learner를 결합하여 strong learner를 만드는 머신러닝 알고리즘.
- 분류 문제에서 Weak learner는 데이터의 작은 변화가 분류 모델의 큰 차이를 야기하는 분류 알고리즘을 지칭함.
- Weak learner를 얻은 후, 오분류된 관측치에 더 높은 가중치를 주는 방식으로 분류 알고리즘이 학습됨.
- Ex) Ada-Boost, XGBoost, Gradient boosting



비지도학습(Unsupervised Learning)

- K-means clustering

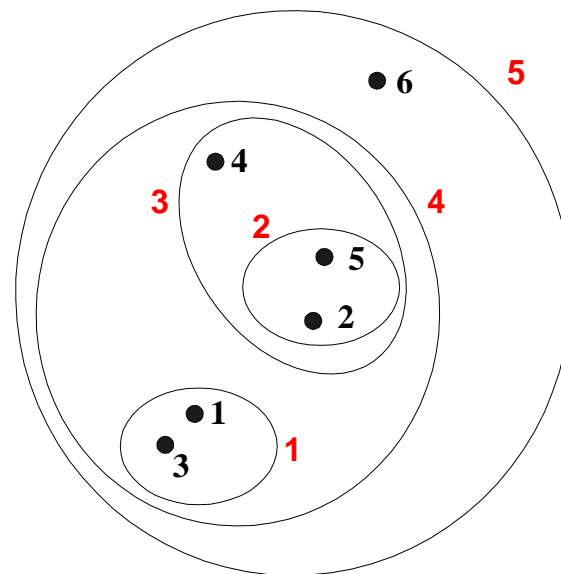
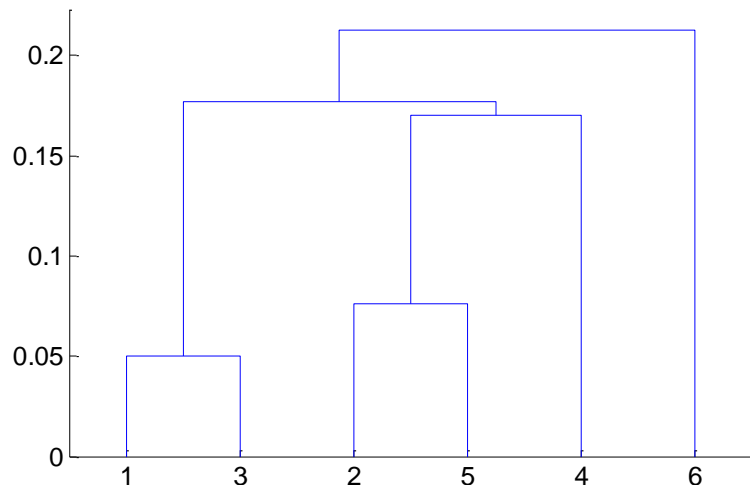
- 데이터를 K개의 그룹으로 나누는 방법론
- 그룹의 평균으로부터 각 관측치까지의 거리를 산출하여 거리가 가까운 관측치끼리 묶음
- 현재 데이터의 분할 이므로 함수를 추정하는 것이 아니며 따라서 새로운 데이터에 대해 군집 배정이 불가능함
- 연산 속도가 빠름
- 군집 내 관측 값들은 서로 동질적이며 군집간 이질적인 상태임
- 군집의 개수를 알아야 분할할 수 있음
- 적절한 군집의 개수는 elbow 방법을 통해 결정 가능



비지도학습(Unsupervised Learning)

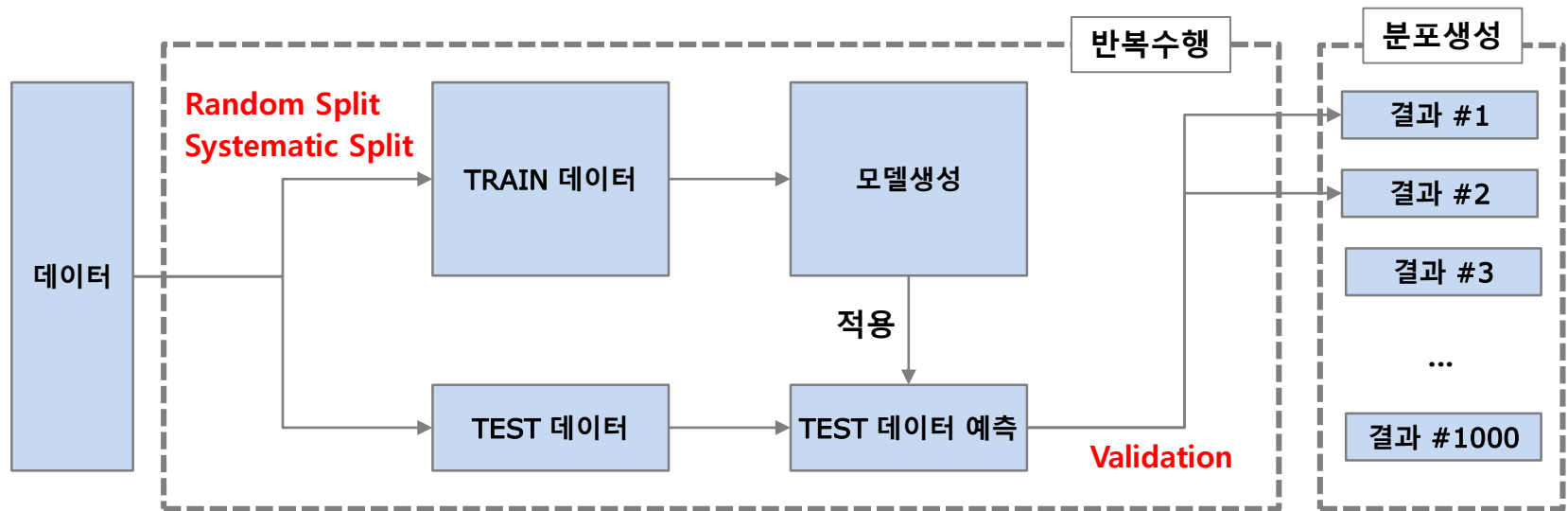
- Hierarchical clustering

- 데이터를 2개 이상 그룹으로 나누는 방법론
- Tree와 반대로 데이터의 성격이 유사한 값들을 묶고 거리가 가까운 군집들을 서로 묶으면서 전체 데이터가 하나가 될 때까지 반복함
- 군집의 개수를 미리 정의할 필요 없고 연산이 빠름



• Cross-Validation

- 데이터를 훈련(train)용 데이터와 검증(test)용 데이터로 반복하여 분리, 훈련용 데이터로 모델을 생성하고, 검증용 데이터로 모델의 성능을 평가하는 과정을 반복함
- 훈련/검증 데이터의 분리를 반복하여 얻어진 여러 개의 평가결과를 종합(일반적으로 평균)하여 평가를 일반화함
- Ex) Leave-p-out CV, K-fold CV, Repeated random subsample CV
- CV의 목적은 검증결과의 일반화에 있음. 즉, 현재 적용하려는 방법론의 과대적합가능성을 점검함
- CV과정 중에 생성된 여러 개의 모형 중 검증결과가 좋은 모형은 모형의 성능이 아닌 “데이터의 성질”에 의한 결과일 가능성이 높음

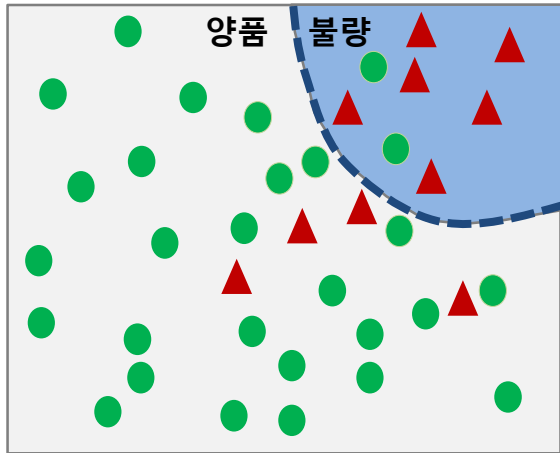


• Confusion Matrix

- 종속변수가 명목형(Categorical) 변수인 경우 예측정확도를 판별하기 위해 예측값과 실측값을 표형태로 표현한 것
- 재현율(recall): 양성 중 실제 양성으로 맞춘 비율
- 특이도(specificity): 음성 중 실제 음성으로 맞춘 비율
- 정밀도(precision): 양성으로 예측된 관측치 중 실제 양성인 비율.
- 정확도(accuracy): 전체 데이터 수 대비 실측값의 개별 항목들을 맞게 예측한 비율
- 특정 항목에 몰아서 예측되는 경우 잘못된 예측모형을 설계하였음에도 불구하고 실측값의 개별 항목 비율에 따라 정확도가 높게 산출될 수 있으므로 정밀도와 재현율을 모두 확인하여 모형을 평가해야 함
- 로지스틱회귀와 같이 확률로 예측한 경우 항목을 분리하기 위한 cut-off결정이 필요하고, 이 경우 재현율과 정밀도, 정확도를 고려하여 cut-off를 찾는 것이 바람직함

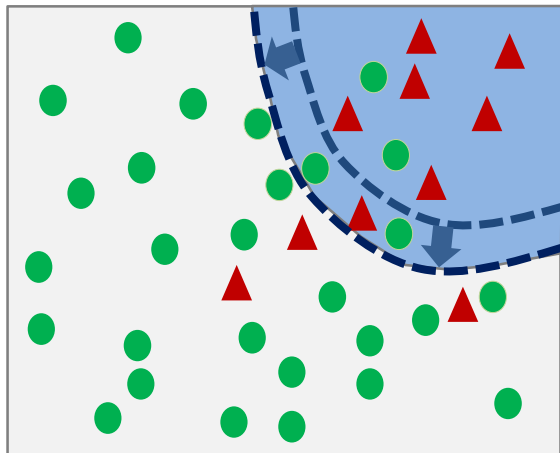
예측값 \ 실측값	0	1	합계	
0	105	5	120	
1	15	108	123	정밀도 0.88
합계	120	113	243	
	특이도 0.88	재현율 0.96	0.88	정확도 (Accuracy)

Measurement



정확도의 종류

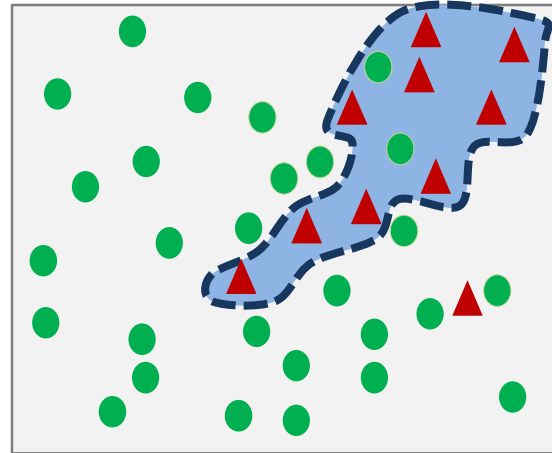
- Recall
: 6 / 10 = 60%
- Precision
: 6 / 8 = 75%



Cut-off 를 낮춘다면?

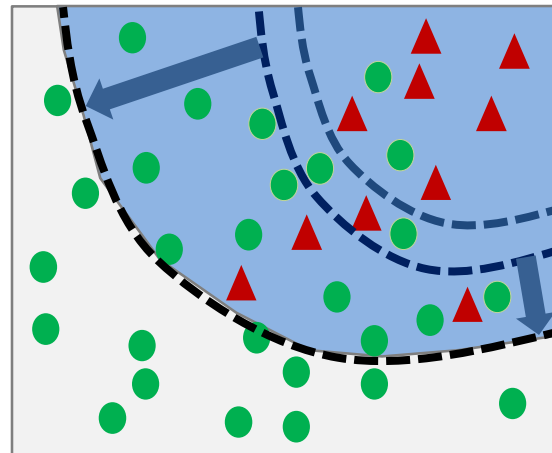
- Recall
: 7 / 10 = 70%
- Precision
: 7 / 11 = 63%

Pitfall



과적합 (Overfitting)

- Recall
: 9 / 10 = 90%
- Precision
: 9 / 11 = 81%



Cut-off 를 너무 낮추면?

- Recall
: 10 / 10 = 100%
- Precision
: 10 / 25 = 40%

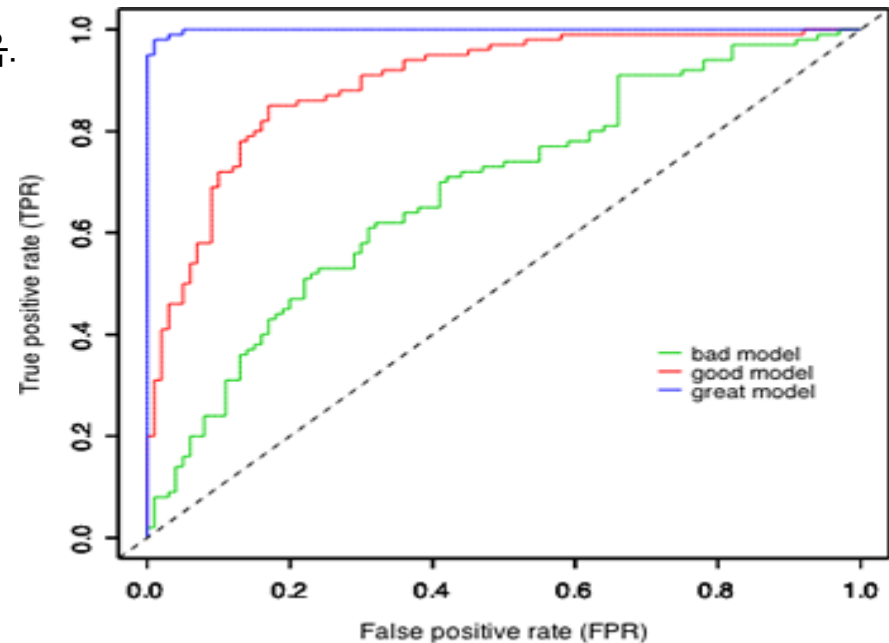
● 양품
▲ 불량
--- Cut-off

• ROC(Receiver Operating Characteristic) Curve

- 0부터 1사이의 모든 값을 cutoff로 재현율과 특이도를 계산 후, x축과 y축에 1-특이도와 재현율을 각각 표시하여 그린 커브.
- AUC(Area Under Curve): ROC 커브의 면적
- AUC는 0과 1사이의 값이며, 1과 가까울수록 모델의 예측력이 좋다고 판단할 수 있음.

• AIC(Akaike Information Criterion)

- 하나의 데이터 셋에 대하여 다양한 설명변수를 이용하여 모델을 만드는 경우, 모델을 비교하는 방법.
- 예를들어, 동일한 데이터셋에 대해 서로 다른 설명변수를 이용하여 두 로지스틱 회귀모형을 만드는 경우 두 모델을 비교하기 위하여 AIC를 사용할 수 있음.
- AIC의 값이 작을 수록 좋은 통계모형이라고 할 수 있음.
- AIC는 서로 다른 모형을 비교할 때는 사용할 수 없음.

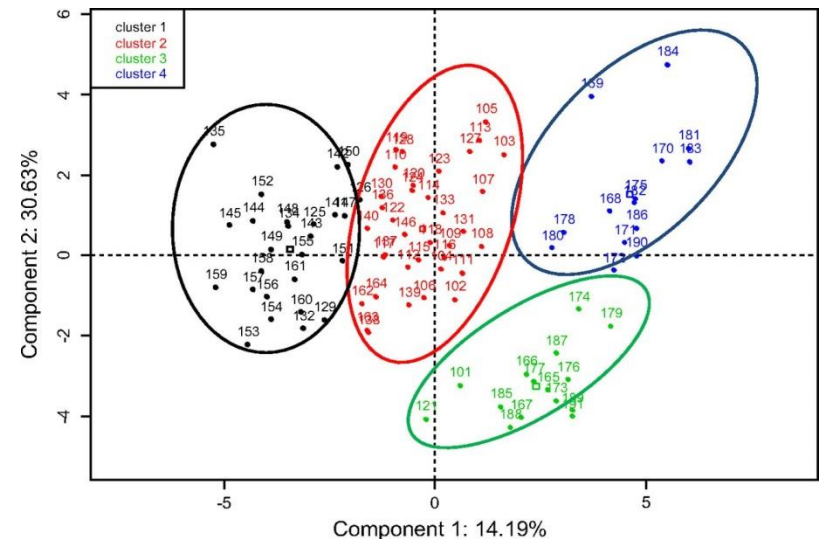


• 주성분 분석(Principal Component Analysis)

- 데이터의 정보량을 최대한 유지하면서 컬럼의 수를 줄여주는 방법.
- 첫 번째 주성분은 가장 많은 정보량을 포함하는 주성분(변수의 선형 결합)이며, 두 번째 주성분은 두 번째로 많은 정보량을 포함하는 주성분임. 누적 정보량을 통하여 몇 개의 주성분이 필요한지를 결정함.
- 데이터를 축약한 후에는 다중공선성, 자유도 등의 문제로 인하여 불가능한 회귀분석 등을 시행할 수 있다는 장점이 있음.
- PCA는 차원을 축소하는 목적으로 개발된 방법이지만 2개 차원으로 축소한 후 xy-좌표에 표시하여 군집화에 활용할 수 있음

	변수 1	변수2	...	변수 1000
관측치 1				
관측치 2				
...				
관측치 500				

누적 정보량	40%	70%	80%		100%
	주성분 1	주성분 2	주성분 3	주성분1000
관측치 1					
관측치 2					
...					
관측치 500					

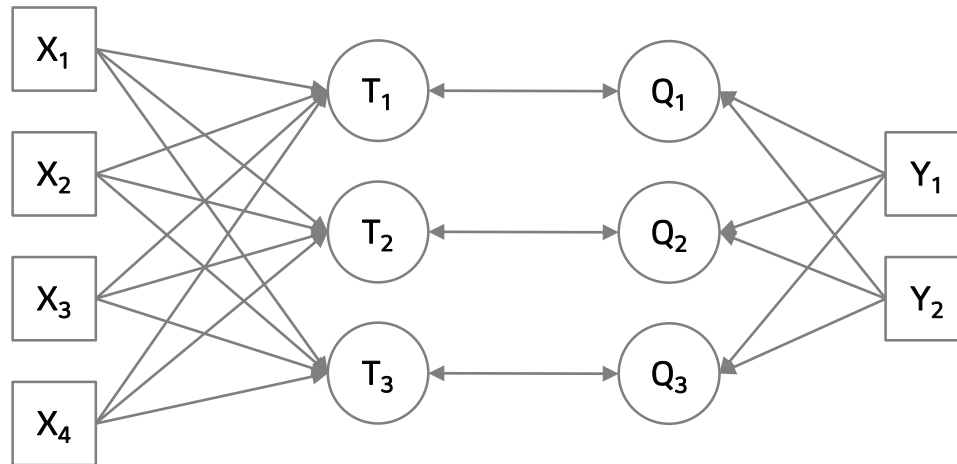


• PLS regression (Partial Least Square)

- PCA regression과 유사하나 PCAR은 독립변수만으로 차원을 축소한 후 회귀모형을 적합하는 반면, PLS regression은 종속변수와의 상관관계를 고려하여 차원을 축소함
- Matrix decomposition 방법을 이용하여 독립변수와 종속변수를 각각 분할 하는데 분할된 projection term들이 서로 공분산을 최대화 하도록 분할을 결정하는 방법

$$X = TP^T + E \quad Y = UQ^T + F$$

- T, U는 각각 X, Y에 대한 projection, P, Q는 회귀계수, E, F는 white noise이고, T와 U의 공분산이 최대가 되도록 분할을 결정함
- Chemometrics(계량분석화학) 분야에서 가장 많이 활용됨



- **연관분석(Association Rule)**

- 장바구니 분석(Market Basket Analysis)로도 널리 알려져 있음.
- 상품 혹은 설비간의 관계를 살펴보고 이로부터 유용한 연관관계를 찾아내고자 할 때 이용될 수 있는 기법
- 따라서 연관규칙 평가 척도를 통하여 높은 불량이 발생(사건B)하는 특정 설비(사건A)를 찾고자 함

IF A(특정설비) Then B(불량발생)

1. **LIFT (향상도)**

: 사건A가 일어난 조건 하에서 사건B가 일어날 조건부확률을 사건B가 일어날 확률

즉, 평균적인 불량률 대비 특정 설비에서 불량률이 얼마나 더 많이 발생하는지를 보이는 지표. 값이 높을수록 혐의 설비

$$\frac{P(A \cap B)}{P(A)P(B)} = \frac{P(B | A)}{P(B)}$$

2. **SUPPORT (지지도) → 보조지표로 활용**

: 전체 사건에서 특정 사건 A, B가 동시에 발생하는 확률

즉, 전체 LOT수 대비 특정 설비에서 불량률이 발생하는 비중을 보이는 지표

$$P(A \cap B)$$

3. **CONFIDENCE (신뢰도) → 보조지표로 활용**

: A라는 사건이 발생했을 때 B가 발생할 확률이 얼마나 높은지를 보이는 지표

$$\frac{P(A \cap B)}{P(A)} = P(B | A)$$

• 연관분석 예시 1

ID	Items
1	빵, 우유
2	빵, 기저귀, 맥주, 계란
3	우유, 기저귀, 맥주, 콜라
4	빵, 우유, 기저귀, 맥주
5	빵, 우유, 기저귀, 콜라

• {빵, 기저귀} -> {맥주}

- Support : 2/5
- Confidence : 2/3
- Lift: $(2/5)/(3/5 * 3/5) = 10/9$

• 연관분석 예시 2

Association Rule 분석 결과 시각화

