간단한 기계 학습

임 경 덕

다양한 데이터 분석

- ❖ 데이터 요약(aggregation)
- ❖ 검정(test)
- ❖ 모의실험(simulation)
  - 확률 분포 등을 활용하여 관심 대상에 대한 기댓값을 확률적으로 예측

- ❖ 데이터 기반 의사결정(data-driven decision)
  - ▪ 규칙기반 시스템(rule-based system)
    - 확인된 사실, 가설, 분석 결과를 바탕으로 조건이나 규칙을 설정
    - 금융 상품 가입 심사 등 정보가 제한적일 때 주로 활용
  - ▪ 기계학습 알고리즘(machine learning algorithm)
    - 다양한 변수의 관계 속에서 의미 있는 정보와 패턴을 파악, 활용
    - 신용등급 등 활용가능한 정보가 많을 때 활용

(복습) 보험료 청구 데이터의 요약

❖ 어떤 실손 보험의 고객별 청구금액 데이터

고객의 신상, 건강정보에 따른 청구금액(charges)의 차이 확인

	A	B	C	D	E	F	G
1	age	sex	bmi	children	smoker	region	charges
2	19	female	27.9	0	yes	southwest	16884.92
3	18	male	33.77	1	no	southeast	1725.552
4	28	male	33	3	no	southeast	4449.462
5	33	male	22.705	0	no	northwest	21984.47
6	32	male	28.88	0	no	northwest	3866.855
7	31	female	25.74	0	no	southeast	3756.622
8	46	female	33.44	1	no	southeast	8240.59
9	37	female	27.74	3	no	northwest	7281.506
10	37	male	29.83	2	no	northeast	6406.411

다양한 변수 형식의 이해

❖ 각 변수에 적합한 형식을 지정

표현	형식	비고
chr	문자형(character)	주소 등에 활용
factor	범주형	지역, 상품 등 범주형 변수에 활용
num	수치형(numerical)	금액 등 소수점을 포함하거나 큰 숫자에 활용
int	정수형(integer)	건수 등 정수형 변수에 활용
Date	날짜형	
POSIXlt	날짜시간형	날짜 요소의 조합
POSIXct	날짜시간형	1970년 1월 1일 0시 0분 0초부터 누적 초

변수 형식 변환

❖ 함수를 활용한 변수 변환

as.factor(*변수이름*) : factor 형식으로 변환

as.character(…) : character 형식으로 변환

as.numeric(…) : numeric 형식으로 변환

as.interger(…) : integer 형식으로 변환

❖ 날짜 형식의 활용

as.Date(*chr변수이름*) : Date 형식으로 변환

- 'YYYY-MM-DD' 혹은 'YYYY/MM/DD' 형식을 날짜 형식으로 변환

strptime(*chr변수이름* , *format*) : format에 따라 Date로 변환

format(*Date변수이름*, *format*) : format에 따라 chr으로 변환

데이터 요약과 검정의 한계

❖ 일반적인 데이터 요약

그룹별 관심 변수의 평균 계산 등 단순한 집계 중심

- 예) 월별 매출액, 상품별 판매건수, 성/연령대별 반응률

저차원의 분석일 뿐만 아니라 설명 변수 간의 관계를 무시

❖ 통계 검정(test)의 한계

표본의 크기(관측치 수)가 증가함에 따라 작은 차이도 유의하다고 판단

→ 대부분의 변수의 관계가 유의하다는 결론을 내릴 수 있음

데이터 요약과 검정의 한계(2)

❖ 새로운 관측치에 대한 예측 불가

요약과 검정은 현재 상황에 대한 파악이 목적

차이 및 인사이트는 확인할 수 있지만 미래 활용은 제한적

예제) 데이터와 예측

	A	B	C	D	E	F	G
1	age	sex	bmi	children	smoker	region	charges
2	19	female	27.9	0	yes	southwest	16884.92
3	18	male	33.77	1	no	southeast	1725.552
4	28	male	33	3	no	southeast	4449.462
5	33	male	22.705	0	no	northwest	21984.47
6	32	male	28.88	0	no	northwest	3866.855

	A	B	C	D	E	F	G
1	age	sex	bmi	children	smoker	region	charges
2	21	female	27.9	0	yes	southwest	
3	18	male	33.77	1	no	southeast	?
4	30	male	33	3	no	southeast	
5	33	female	22.705	0	yes	northwest	
6	43	male	32.12	2	yes	northeast	

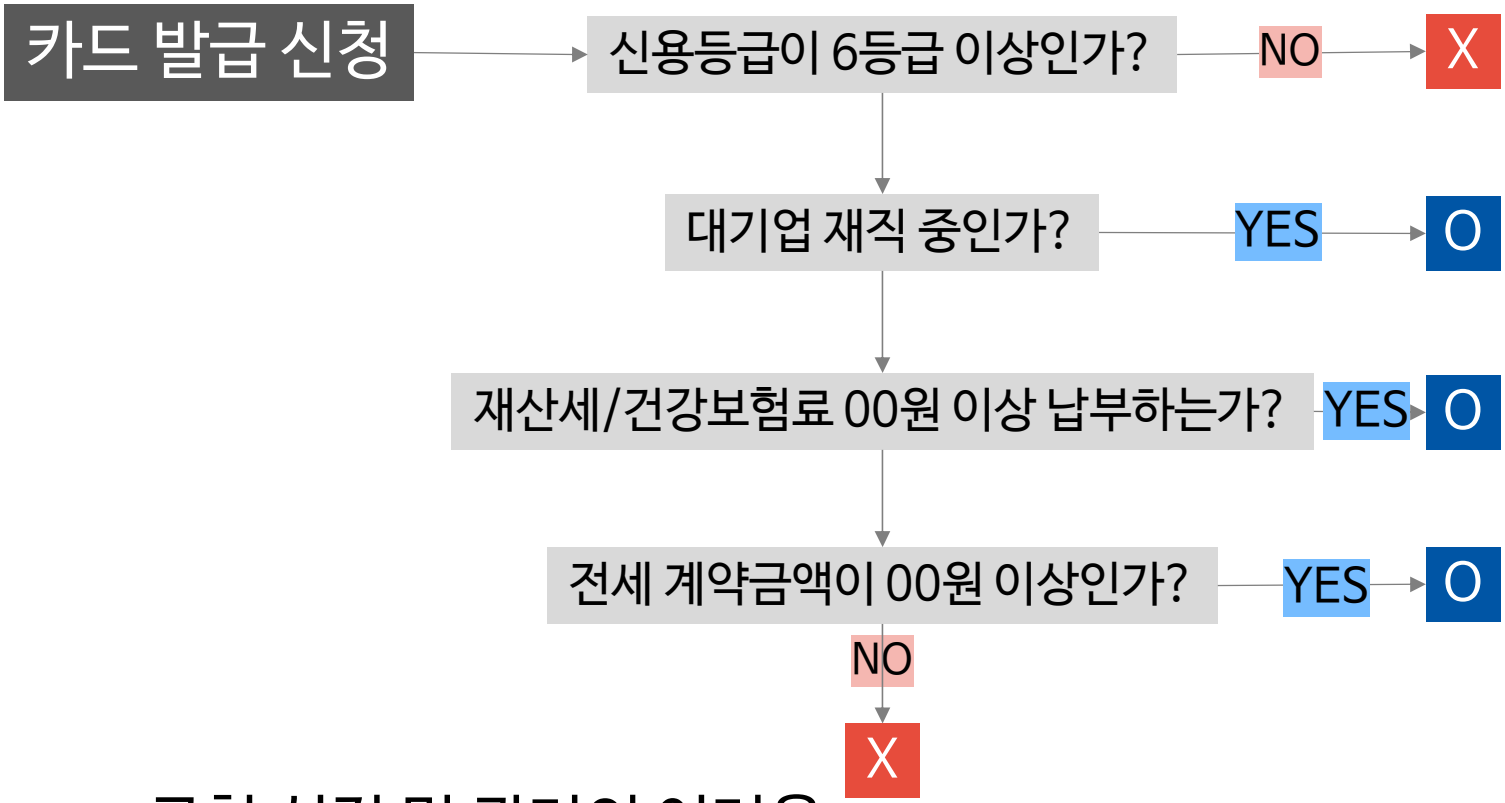
	A	B	C	D
1	admit	gre	gpa	rank
2	0	380	3.61	3
3	1	660	3.67	3
4	1	800	4	1
5	1	640	3.19	4

	A	B	C
1	gre	gpa	rank
2	750	4.01	4
3	800	3.52	1
4	500	4.22	2
5	650	3.76	3

규칙기반 알고리즘의 활용

❖ 금융권의 다양한 심사 과정에서 주로 활용

- (예제) 카드 발급 심사



- 규칙 설정 및 관리의 어려움

기계 학습의 이해

❖ 기계 학습(machine learning)

주어진 문제를 해결하기 위해 알고리즘과 기계의 저장, 연산 능력을 활용
데이터 속에 존재하는 다양하고 복잡한 패턴을 컴퓨터로 파악하는 과정

- 데이터 속 차이를 확인하고 설명

❖ 지도학습(supervised learning) : 관심변수 y의 차이를 설명

- 회귀(regression) : 연속형(continuous) 관심변수 y (1, 100의 차이)
- 분류/판별(classification) : 범주형(categorical) 관심변수 y (0, 1의 차이)

❖ 비지도학습(unsupervised learning) : 관측치 간 차이를 일반화

- 군집화(clustering) : 가까운 관측치들끼리 묶어 군집화

지도학습과 변수의 구분

❖ 설명변수와 반응변수

변수의 관계를 설명할 때 영향을 주는 변수와 영향을 받는 변수를 구분

- **설명변수(explanatory variable)**
  - 독립변수(independent variable)
  - 결과에 영향을 줄 수 있는 변수

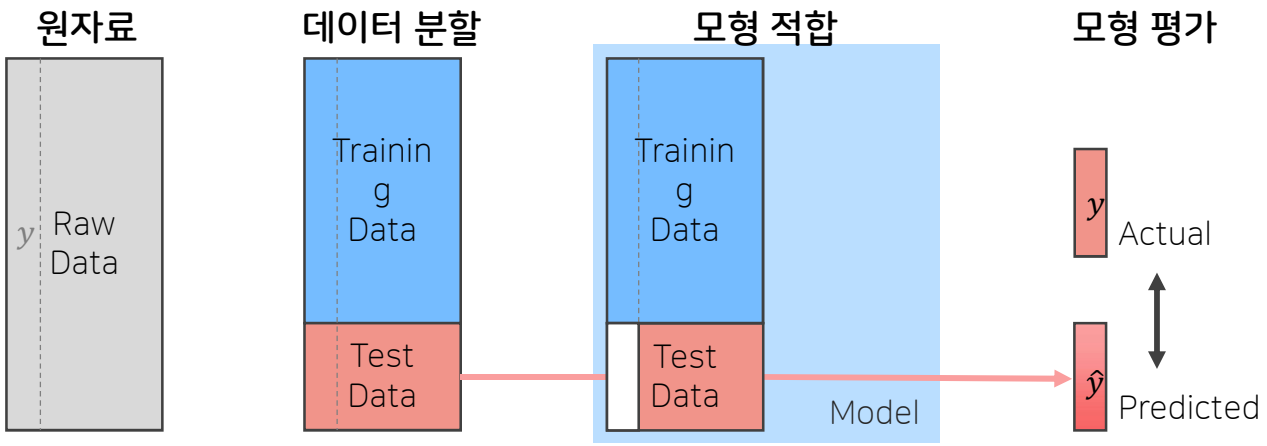
- **관심변수**
  - 종속변수(dependent variable), 반응변수(response variable)
  - 설명 변수의 변화에 따라 값이 결정되는 변수

❖ 모형 적합 : 설명변수로 관심변수의 차이를 설명하는 과정

교차 검증을 통한 모형 평가

❖ 데이터 분할을 활용한 모형 평가 과정

1. Training(훈련) 데이터와 Test(검증) 데이터 분할
2. Training 데이터를 활용한 모형 적합(fitting)
3. 적합된 분류모형을 Test 데이터를 활용해서 평가



선형 회귀와 로지스틱 회귀

간단한 선형 회귀 예제

❖ 아빠키-아들키 예제

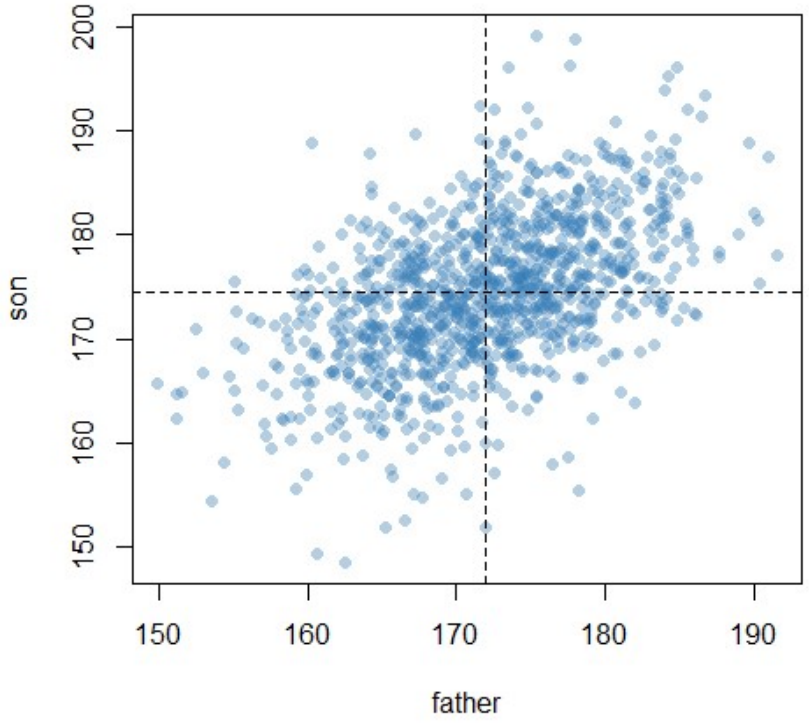
아들키는 아빠키와 관련이 있을 가능성이 큼

- 두 변수의 상관계수 : 0.5

❖ 일차함수를 활용한 관계식의 표현

아들키 = a + 아빠키 × b ± 개인차

❖ 선형 회귀 : 회귀 계수를 계산/추정



단순 선형 회귀 모형

❖ 단순 선형 회귀(simple linear regression)

관심 변수 Y와 설명 변수 X에 대해 직선 관계를 가정

- 모형식 : $y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon \sim Normal(0, \sigma^2)$
- 모형 적합 : $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \longrightarrow \hat{\sigma}^2$

❖ 회귀 모형의 적합

n개의 관측치 $(Y_i, X_i), \quad i = 1,2,\cdots,n$을 활용하여

회귀 계수 β_0, β_1에 대한 추정 값 $\hat{\beta}_0, \hat{\beta}_1$을 계산하는 과정

다중 선형 회귀 모형

❖ 다중 선형 회귀(multiple linear regression)

관심 변수 Y와 p개 설명 변수$X_1, X_2, \cdots X_p$에 대해 다음의 관계식을 가정

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_p X_p + \varepsilon,\ \varepsilon \sim N(0, \sigma^2)$$

관심 변수 Y는 설명 변수X_j에 β_j만큼 비례하며, 설명할 수 없는 불확실성 ε가 존재

관심 변수 Y는 p개 설명 변수$X_1, X_2, \cdots X_p$의 효과의 합으로 표현

단, ε는 평균이 0이고 분산이 σ^2인 정규분포를 따른다고 가정

로지스틱 회귀분석

❖ 로지스틱 회귀분석(logistic regression)

선형모형에 근거한 확률 계산

0/1의 이진 관심변수에 대해 '1'이 될 확률을 기준으로 관측치를 분류

관심확률 $\pi = P(Y = 1|X)$: 설명변수 X에 따라 $Y = 1$일 확률

▪ 관심확률의 로짓에 대한 선형회귀모형 적합

$$logit(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

$$\Rightarrow \quad \frac{\pi}{1-\pi} = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon}$$

x_i이 한 단위 증가하면 사건 $Y = 1$의 $odds$가 e^{β_i}배 증가

즉, β_i이 (+)이면 $Y = 1$일 확률이 증가

범주형 설명변수의 활용

❖ 가변수(dummy variable)의 활용

회귀모형에서 연속형 설명변수만 회귀계수를 추정가능

k개 수준을 갖는 범주형 변수에 대해 $k-1$개 가변수를 생성/활용

❖ 가변수의 생성

가변수는 0 또는 1 (혹은 1 또는 -1)의 두개의 값만 가짐

생성된 $k-1$개 가변수를 조합하면 원래 범주형 변수의 수준 유추 가능

가변수를 활용한 회귀모형의 결과는 분산분석과 동일

예제)

Rank
A
B
C
D

RankB	RankC	RankD
0	0	0
1	0	0
0	1	0
0	0	1

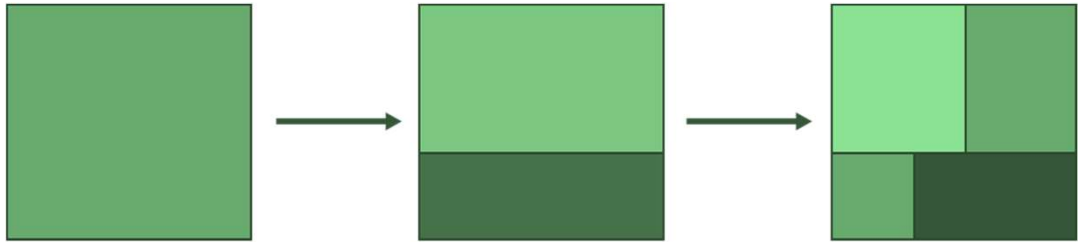
의사결정 나무 모형의 활용

의사결정 나무의 개념

❖ 의사결정 나무(decision tree)

관측치를 가장 잘 분류하는 조건으로 가지를 뻗어 모형확장

- 순도(purity)가 0/1에 가깝게 공간을 분할
- 1회 분할(partitioning) 마다 하위 그룹의 수가 하나 씩 증가
- 모든 가능한 분할 중 알고리즘 지표 개선이 가장 큰 분할 순으로 적용



조건부 평균과 조건부 확률

❖ 조건을 활용한 관측치 분할

특정 조건을 활용하여 조건과 일치하는 관측치와 나머지 관측치로 분할
조건에 따라 2분할 된 관측치 그룹에 대해 관심변수에 대한 요약 값 계산

- 수치형 관심 변수 : 조건부 평균
- 범주형 관심 변수 : 조건부 확률(비율)

❖ 최적 분할 기준 조건의 탐색

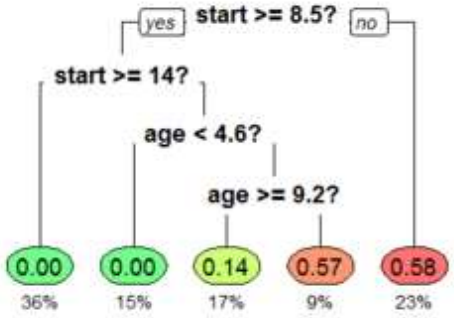
전체 평균/확률과 조건부 평균/확률의 차이 발생
알고리즘을 활용하여 차이를 극대화하는 기준 변수 및 임계 값 탐색

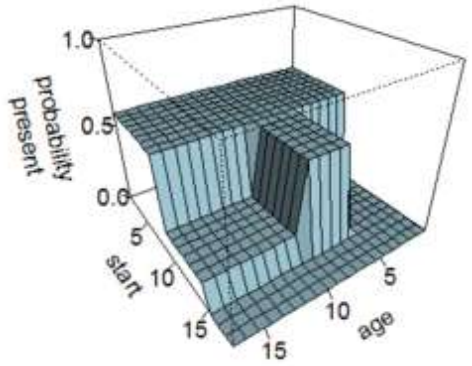
의사결정 나무의 장점

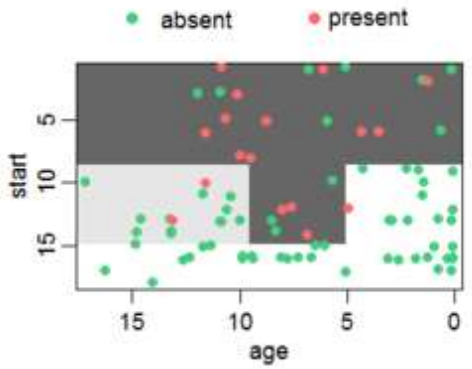
❖ 의사결정 나무의 장점

- 자료 가공이 거의 필요 없고, 설명변수의 역할 등 모형 해석이 용이함
- 비선형적인 관계를 설명(교호작용 등이 반영)
- 연속형 관심변수와 범주형 관심변수를 모두 설명 가능

예시) 의사결정 나무 모형 적합의 시각화







주요 알고리즘

❖ 지니 불순도(Gini impurity)

- CART 알고리즘에서 활용
- 특정 그룹에 이질적인 것이 얼마나 섞였는지를 계산하는 지표

  0에 가까울 수록 순도가 높음

$$I_G(p) = \sum_{i=1}^{J} p_i \sum_{k \neq i} p_k = \sum_{i=1}^{J} p_i(1-p_i) = \sum_{i=1}^{J} (p_i - {p_i}^2) = \sum_{i=1}^{J} p_i - \sum_{i=1}^{J} {p_i}^2 = 1 - \sum_{i=1}^{J} {p_i}^2$$

❖ 정보 획득량(information gain)

- 정보 이론의 엔트로피(entropy)에 기반한 ID3, C4.5 알고리즘에서 활용
- 엔트로피 : $H(T) = I_E(p_1, p_2, \ldots, p_J) = -\sum_{i=1}^{J} p_i \log_2 p_i$
- 특정 그룹의 엔트로피와 분할된 하위 그룹의 엔트로피 차이를 계산, 활용

$$\overbrace{IG(T,a)}^{\text{Information Gain}} = \overbrace{H(T)}^{\text{Entropy(parent)}} - \overbrace{H(T|a)}^{\text{Weighted Sum of Entropy(Children)}}$$

$$= -\sum_{i=1}^{J} p_i \log_2 p_i - \sum_{a} p(a) \sum_{i=1}^{J} -Pr(i|a) \log_2 Pr(i|a)$$

의사결정 나무 모형의 이해(CART)

❖ 재귀분할(recursive partitioning)과 가지치기(pruning)

- 재귀분할 : 분할된 하위 그룹을 반복하여 분할
- 가지치기 : 지표를 활용하여 유의하지않은 하위그룹 분할을 제거

❖ 비용 복잡도(cost complexity)를 활용한 가지치기

$$R_{cp}(T) \equiv R(T) + cp * |T| * R(T_1)$$

- $R(T)$: 트리 T 의 오분류율
- $|T|$: 가지의 개수
- T_1 : 분할되지 않은 기본 트리, 모두 0으로 예측
- cp : 복잡도 모수(complex parameter)
- cp에 따라 모형의 복잡도가 결정

  $cp = 0$: 최대 가지 모형

  $cp = 1$: 평균값 모형

군집화의 활용

군집화를 활용한 그룹 생성

❖ 고객/상품의 수치적 특성에 군집화(clustering) 적용

예제)카드사 데이터를 가정한 일반적인 고객 군집화 데이터의 구성

ID	백화점	마트	편의점	지하철	주유소	한식	일식	…	합계
1	100	200	100	100	0	0	50		700
2	800	400	0	0	400	0	0		2000
3	0	0	0	0	200	200	0		1000
4	0	50	0	0	0	0	0		50

- 각 관측치의 특성을 연속형 변수로 표현
- 관측치 간 거리를 활용한 군집화

유사도의 개념

❖ 유사도의 계산

각 관측치 간의 거리를 활용하여 관측치 간 유사도 측정 가능

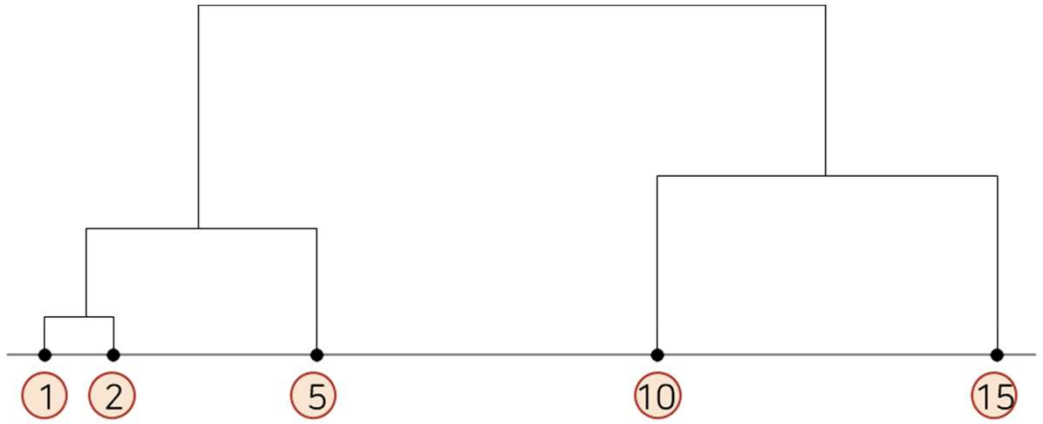
- 예제) 유클리드 거리

  두 점 (x_1, y_1), (x_2, y_2) 간의 유클리드 거리 : $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

유사도가 높은 관측치들을 묶어 그룹을 생성

그룹 간의 거리는 멀고, 그룹 내 관측치 간 거리는 가까운 방향으로 생성

- 예제) 수직선 상의 5개의 점을 유사도 순서로 하나씩 묶기



k-평균 군집화

❖ k-평균 군집화(k-means clustering)

초기 설정된 k개의 중심 중 가장 가까운 그룹으로 각 관측치를 할당
그룹 간 거리를 최대화/그룹 내 거리를 최소화하는 방향으로 군집 생성

❖ k-평균 군집화 알고리즘

- 관측치 중 임의로(Randomly) k개의 관측치를 선택 (중심 초기값)
- n개의 관측치를 k개 중심 중 가장 가까운 그룹으로 할당
- 할당된 관측치의 평균으로 k개 중심을 업데이트
- n개의 관측치를 업데이트 된 k개 중심 중 가장 가까운 그룹으로 재할당
- k개 중심 업데이트와 n개 관측치 재할당을 k개 중심이 움직이지 않을 때까지 반복

감사합니다.