Chapter. 11

비슷한 애들 모여라: 군집화

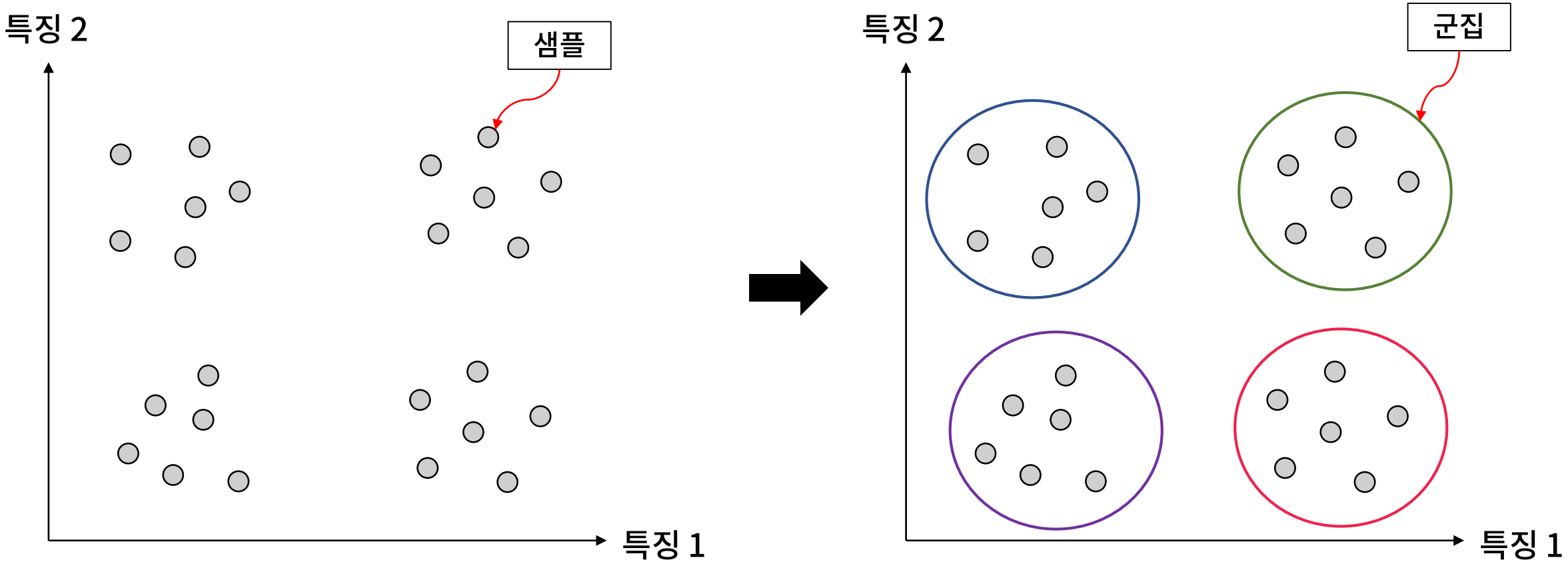
군집화 기본 개념

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

I 군집화란?

- 군집화: 하나 이상의 특징을 바탕으로 **유사한 샘플을 하나의 그룹**으로 묶는 작업



군집화의 목적

- 많은 샘플을 **소수의 군집**으로 묶어 **각 군집의 특성을 파악**하여 데이터의 특성을 이해하기 위함
- 군집의 특성을 바탕으로 각 군집에 속하는 샘플들에 대한 **세분화된 의사결정**을 수행하기 위함

I 군집화 활용 사례: Code 9

- 신한카드에서는 **카드 사용 패턴에 대한 군집화 결과**를 바탕으로 고객을 총 18개의 군집으로 구분함

source: https://www.shinhancardblog.com/

I 거리와 유사도

- 유사한 샘플을 하나의 군집으로 묶기 위해서는 **거리** 혹은 **유사도**의 개념이 필요

> 두 샘플이 **유사하다** = 두 샘플 간 **유사도가 높다** = 두 샘플 간 **거리가 짧다**

- Tip. 대부분의 거리 척도와 유사도 척도는 **수치형 변수**에 대해서 정의되어 있으므로, 문자를 숫자로 바꿔주는 작업이 반드시 선행되어야 함

범주형 변수의 숫자화: 더미화

- 가장 일반적인 범주형 변수를 변환하는 방법으로, 범주형 변수가 특정 값을 취하는지 여부를 나타내는 더미 변수를 생성하는 방법

#1의 종교 변수가 기독교 값을 취하므로, 기독교 변수가 1을 가짐

불교 변수는 나머지 변수로 완벽히 추론 가능하므로 변수간 상관성 제거 및 계산량 감소를 위해 제거

샘플	종교
#1	기독교
#2	천주교
#3	불교
#4	기독교
#5	기독교
#6	천주교

더미화 →

레코드	기독교	천주교	불교
#1	1	0	0
#2	0	1	0
#3	0	0	1
#4	1	0	0
#5	1	0	0
#6	0	1	0

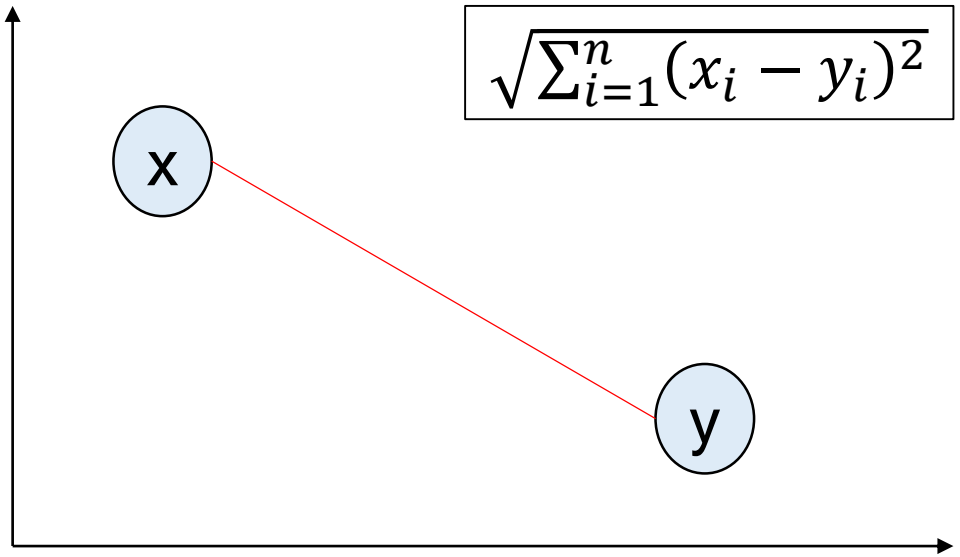
#1의 종교 변수가 기독교 값을 취하지 않으므로, 기독교 변수가 0을 가짐

I 관련 함수: Pandas.get_dummies

- DataFrame이나 Series에 포함된 범주 변수를 더미화하는 함수
- 주요 입력
  - ➢ data: 더미화를 수행할 Data Frame 혹은 Series
  - ➢ drop_first: 첫 번째 더미 변수를 제거할지 여부 (특별한 경우를 제외하면 True라고 설정)
- 사용시 주의사항: 숫자로 표현된 범주 변수 (예: 시간대, 월, 숫자로 코드화된 각종 문자)를 더미화하려면, 반드시 astype(str)을 이용하여 컬럼의 타입을 str 타입으로 변경해야 함

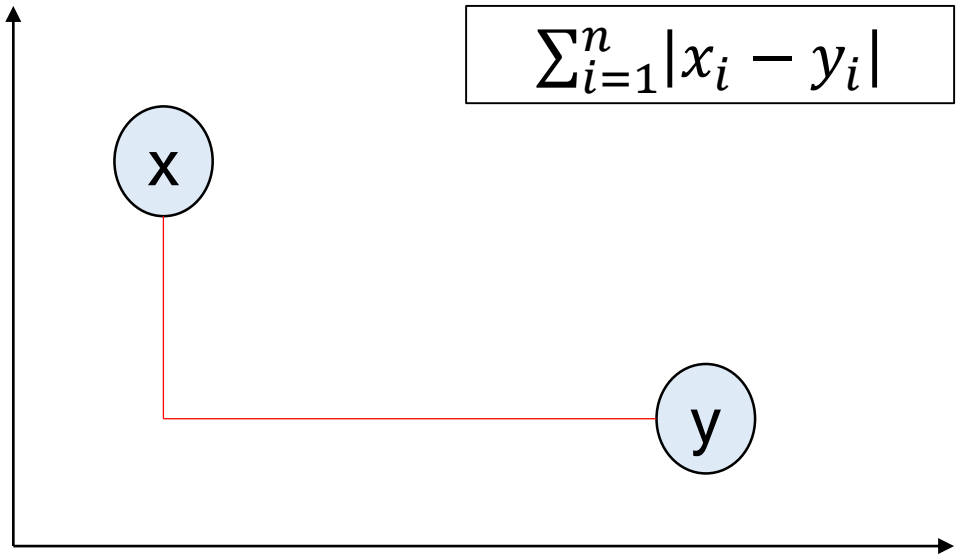
I 다양한 거리 / 유사도 척도: (1) 유클리디안 거리

- 가장 **흔하게 사용**되는 거리 척도로 빛이 가는 거리로 정의됨



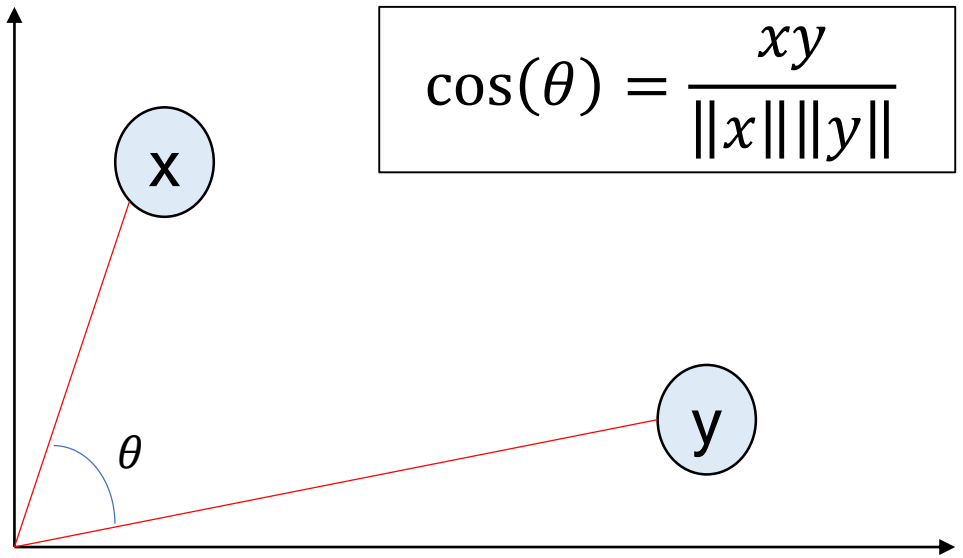
I 다양한 거리 / 유사도 척도: (2) 맨하탄 거리

- 정수형 데이터(예: 리커트 척도)에 적합한 거리 척도로, 수직/수평으로만 이동한 거리의 합으로 정의됨



I 다양한 거리 / 유사도 척도: (3) 코사인 유사도

- 스케일을 고려하지 않고 방향 유사도를 측정하는 상황(예: 상품 추천 시스템)에 주로 사용



$$\cos(\theta) = \frac{xy}{\|x\|\|y\|}$$

I 다양한 거리 / 유사도 척도: (4) 매칭 유사도

- 이진형 데이터에 적합한 유사도 척도로 전체 특징 중 일치하는 비율을 고려함

x_1	x_2	x_3	x_4	x_5
1	0	1	0	0
0	0	1	0	1

$$\frac{N(x_i = y_i)}{n} = \frac{3}{5}$$

I 다양한 거리 / 유사도 척도: (5) 자카드 유사도

- 이진형 데이터에 적합한 유사도 척도로 둘 중 하나라도 1을 가지는 특징 중 일치하는 비율을 고려함

x_1	x_2	x_3	x_4	x_5
1	0	1	0	0
0	0	1	0	1

$$\frac{N(x_i = y_i = 1)}{N(x_i + y_i \geq 1)} = \frac{1}{3}$$

- 희소한 이진형 데이터에 적합한 유사도 척도임

Chapter. 11

비슷한 애들 모여라: 군집화

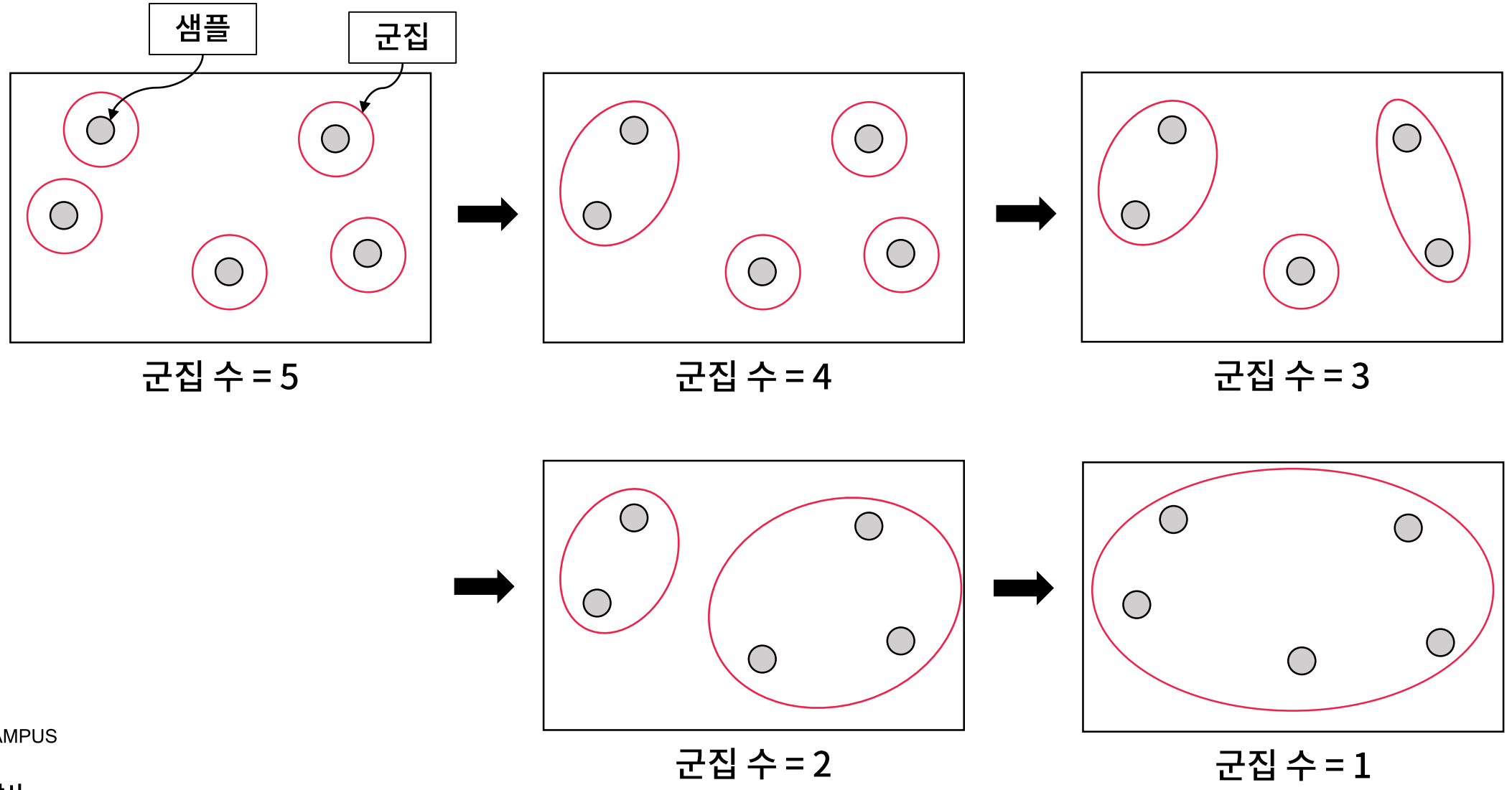
계층적 군집화

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

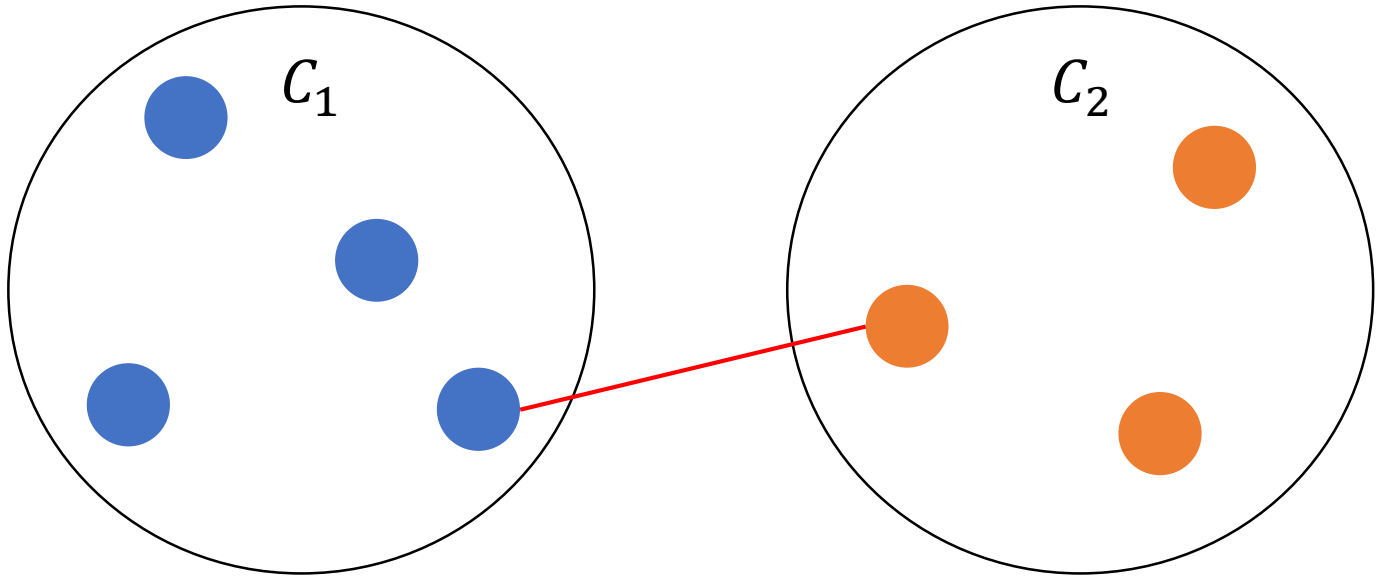
I 기본 개념

- 개별 샘플을 군집으로 간주하여, 거리가 가장 가까운 두 군집을 순차적으로 묶는 방식으로 큰 군집을 생성



군집 간 거리: 최단 연결법

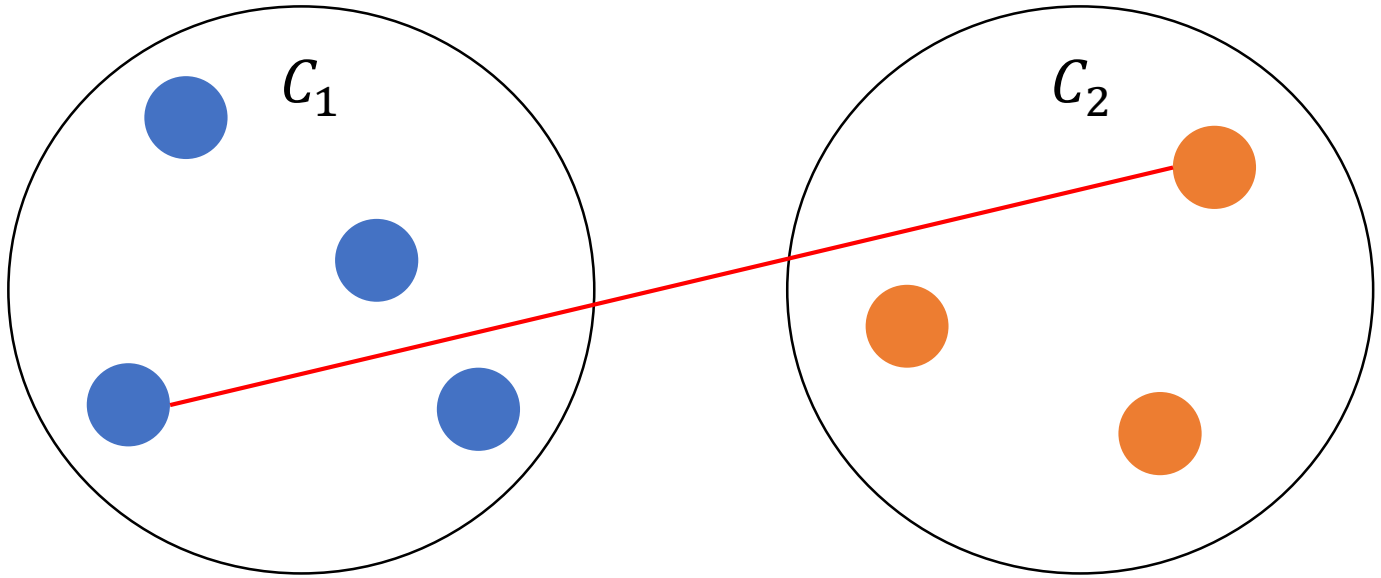
- $\min_{(x_1 \in C_1, x_2 \in C_2)} dist(x_1, x_2)$



- 이상치에 민감
- 계산량 많은 편

군집 간 거리: 최장 연결법

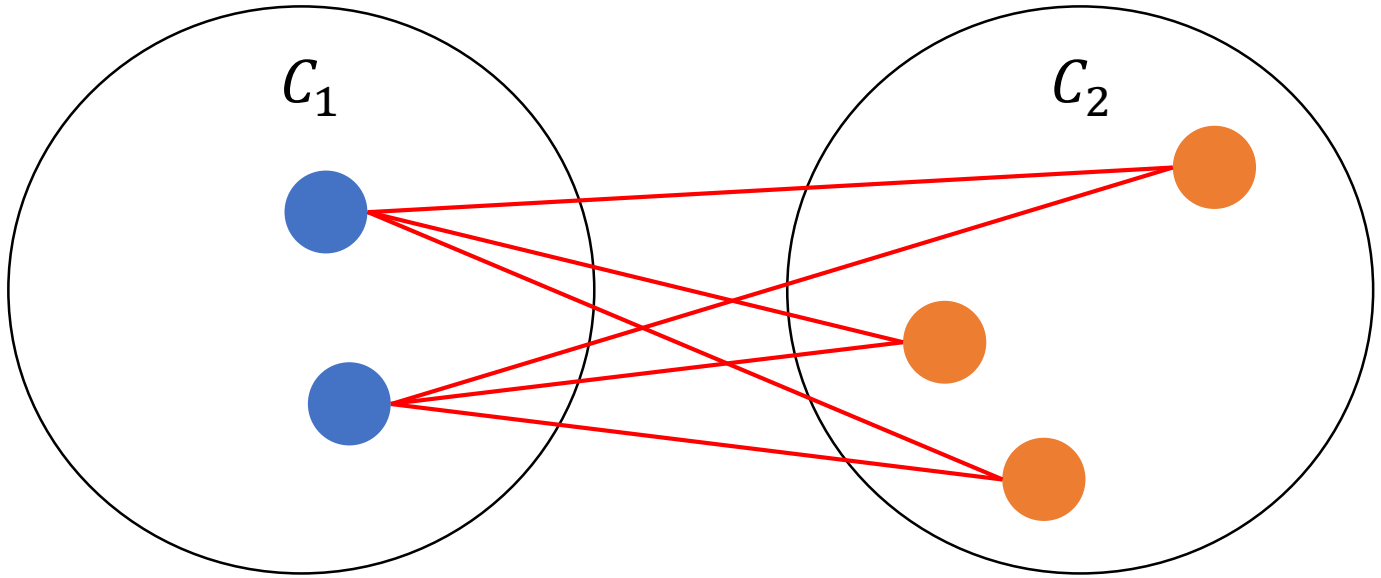
- $\max_{(x_1 \in C_1, x_2 \in C_2)} dist(x_1, x_2)$



- 이상치에 민감
- 계산량 많은 편

군집 간 거리: 평균 연결법

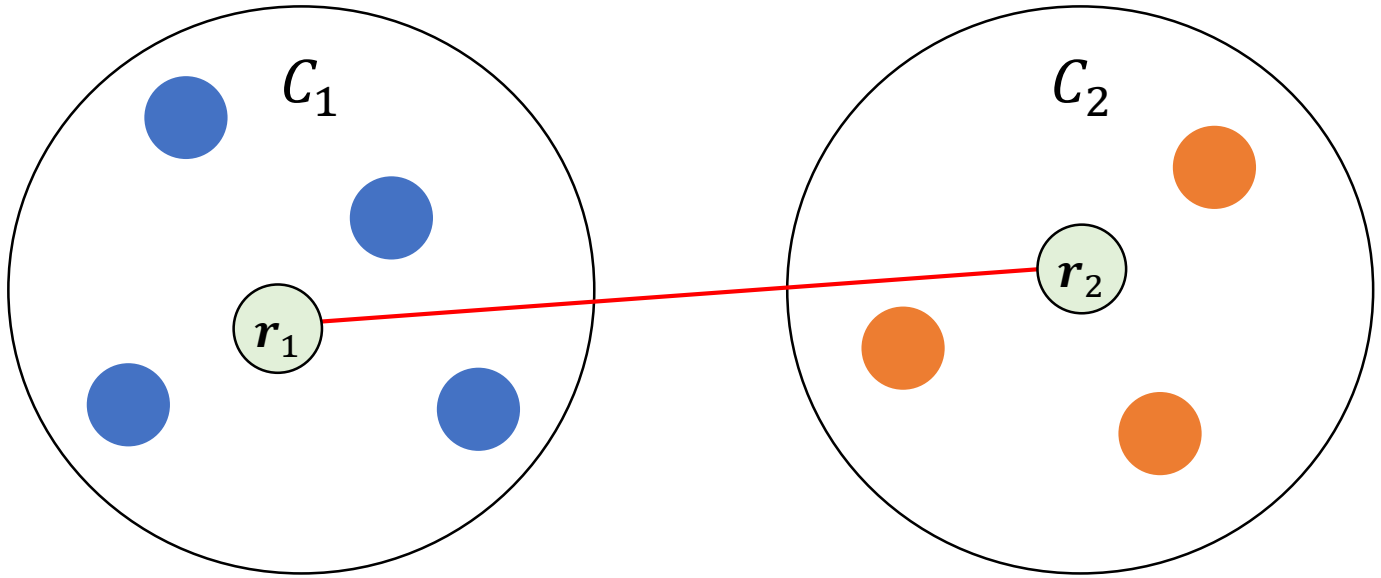
- $\frac{\sum_{x_1 \in C_1} \sum_{x_2 \in C_2} dist(x_1, x_2)}{|C_1| \times |C_2|}$



- 이상치에 둔감
- 계산량 많은 편

군집 간 거리: 중심 연결법

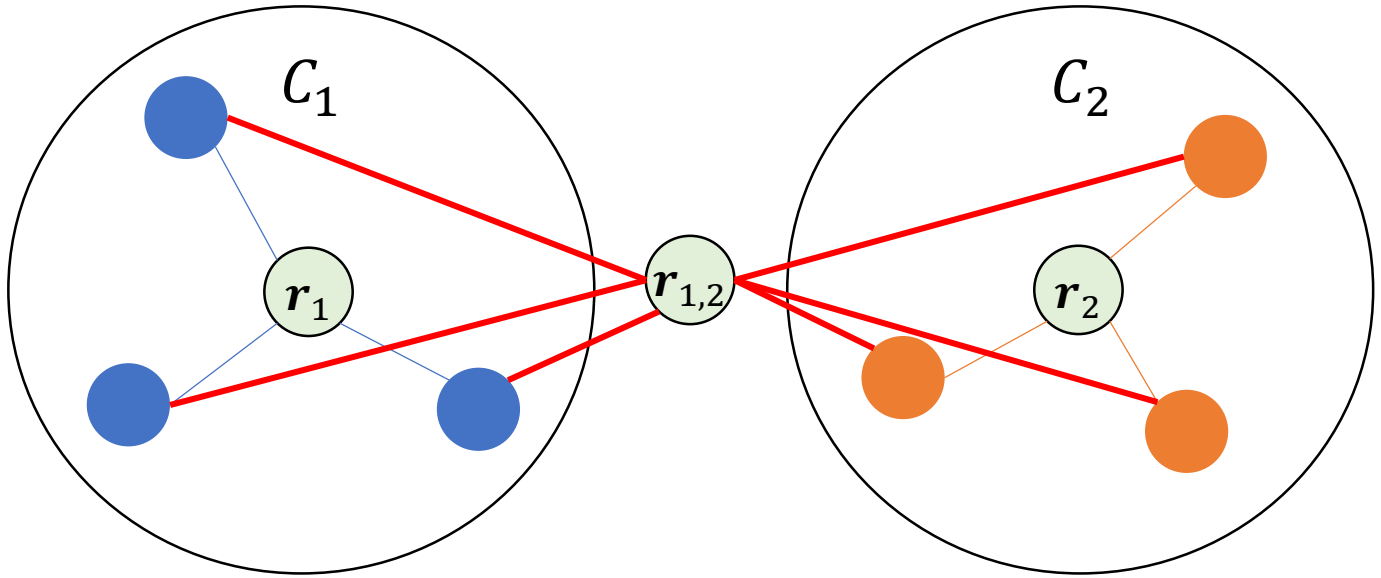
- $dist(\boldsymbol{r}_1, \boldsymbol{r}_2)$, 여기서 $\boldsymbol{r}_1$과 $\boldsymbol{r}_2$는 C_1과 C_2의 중심 $(\boldsymbol{r}_1 = mean_{\boldsymbol{x}_1 \in C_1}(\boldsymbol{x}_1), \boldsymbol{r}_2 = mean_{\boldsymbol{x}_2 \in C_2}(\boldsymbol{x}_2))$



- 이상치에 둔감
- 계산량 적은 편

군집 간 거리: 와드 연결법

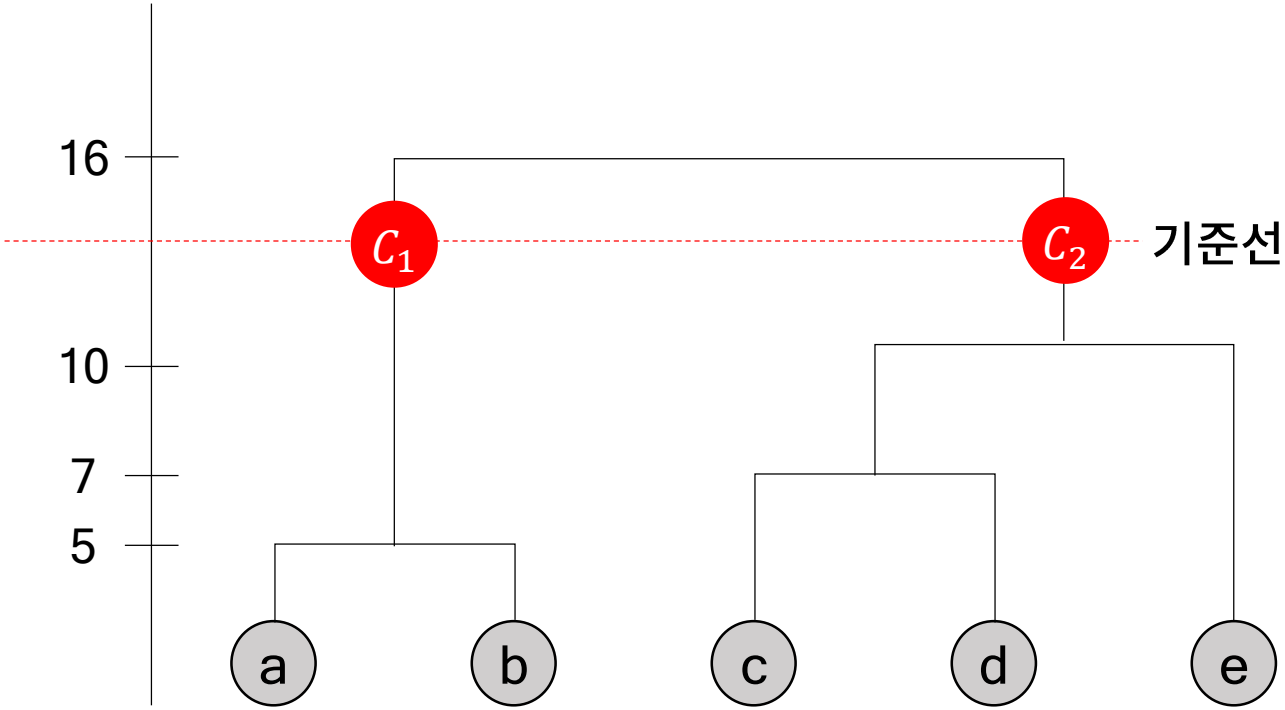
- $\sum_{x_1 \in C_1} dist(x_1, r_1) + \sum_{x_2 \in C_2} dist(x_2, r_2) - \sum_{x \in C_1 \cup C_2} dist(x_1, r_{1,2})$

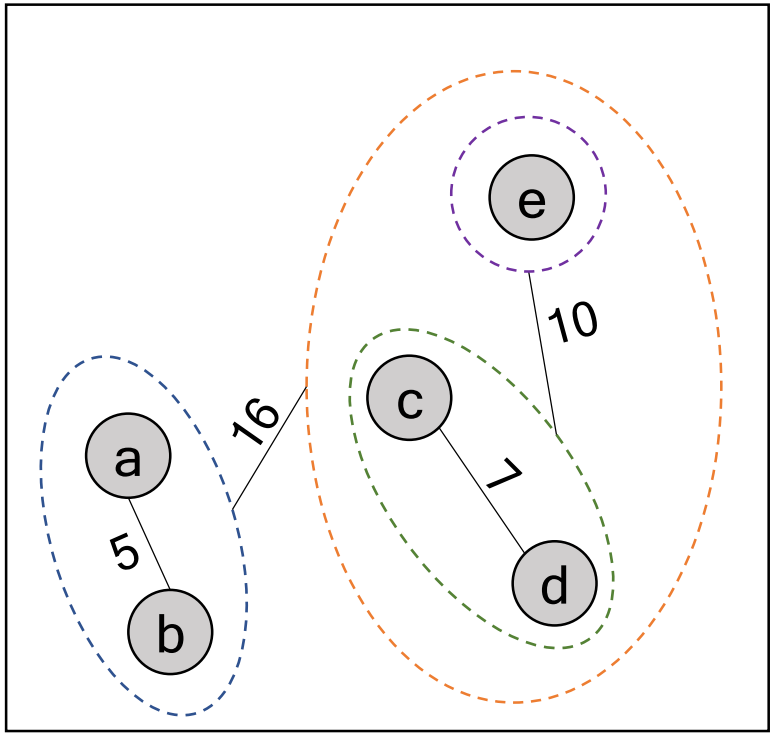


- 이상치에 매우 둔감
- 계산량 매우 많은 편
- 군집 크기 비슷하게 만듦

덴드로그램

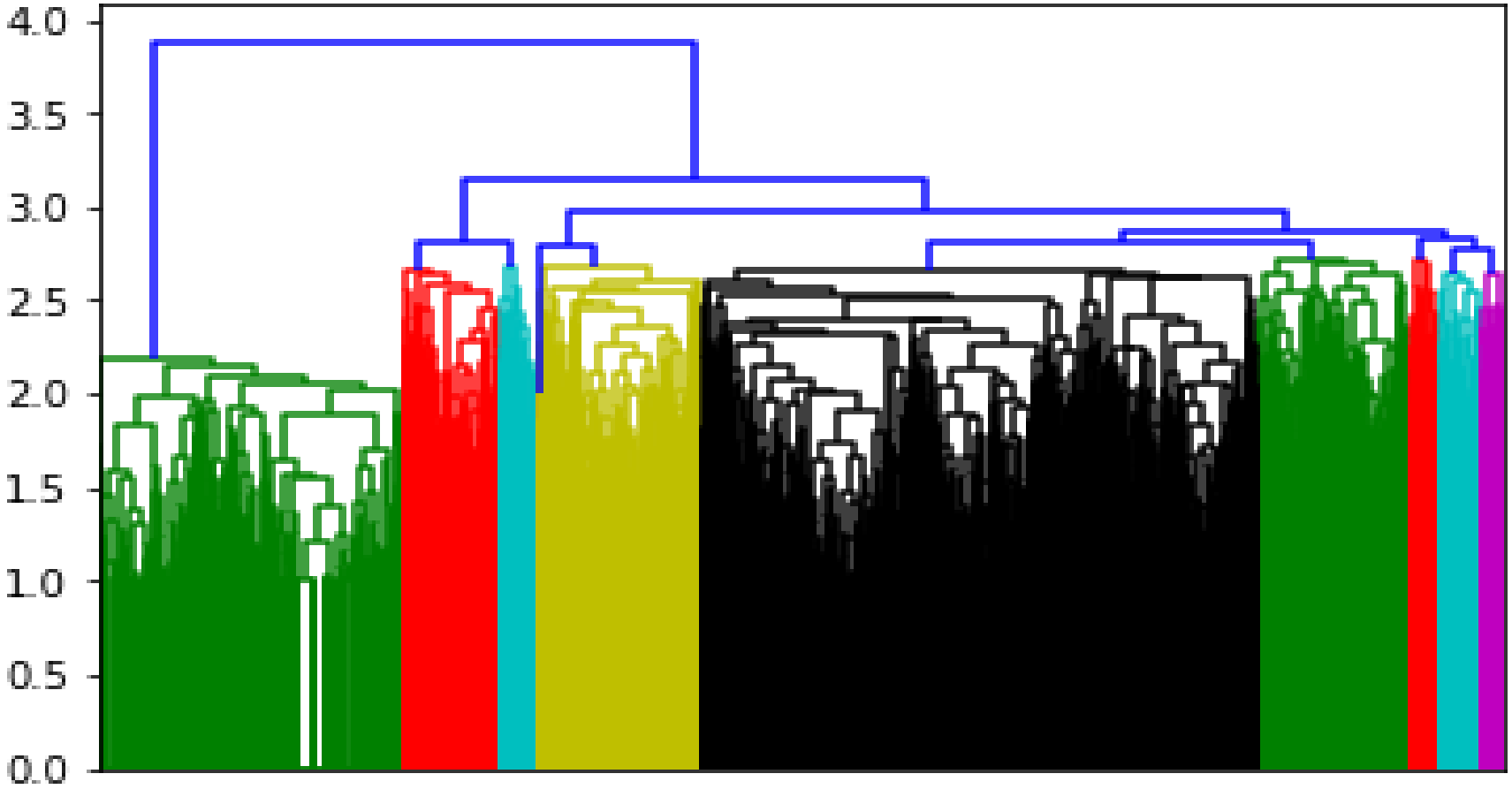
- 계층 군집화 과정을 트리 형태로 보여주는 그래프





덴드로그램의 현실

- 덴드로그램은 샘플 수가 많은 경우에는 **해석이 불가능**할 정도로 복잡해진다는 문제가 있음



계층 군집화의 장단점

- 장점 (1) **덴드로그램**을 이용한 군집화 과정 확인 가능
- 장점 (2) **거리/유사도** 행렬만 있으면 군집화 가능
- 장점 (3) **다양한** 거리 척도 활용 가능
- 장점 (4) 수행할 때마다 **같은 결과**를 냄 (임의성 존재 X)

- 단점 (1) 상대적으로 **많은 계산량** $O(n^3)$
- 단점 (2) **군집 개수** 설정에 대한 **제약** 존재

sklearn.cluster.AgglomerativeClustering

- 주요 입력
  - ➢ n_clusters: 군집 개수
  - ➢ affinity: 거리 척도 {“Euclidean”, “manhattan”, “cosine”, “precomputed”}
    - ✓ linkage가 ward로 입력되면 “Euclidean”만 사용 가능함
    - ✓ “precomputed”는 거리 혹은 유사도 행렬을 입력으로 하는 경우에 설정하는 값
  - ➢ linkage: 군집 간 거리 {“ward”, “complete”, “average”, “single”}
    - ✓ complete: 최장 연결법
    - ✓ average: 평균 연결법
    - ✓ single: 최단 연결법

sklearn.cluster.AgglomerativeClustering (계속)

- 주요 메서드
  - fit(X): 데이터 X에 대한 군집화 모델 학습
  - fit_predict(X): 데이터 X에 대한 군집화 모델 학습 및 라벨 반환

- 주요 속성
  - labels_: fitting한 데이터에 있는 샘플들이 속한 군집 정보 (ndarray)

(Tip) Pandas.crosstab

- 일반적인 거래 데이터를 교차 테이블 형태로 변환하는데 사용하는 함수

회원 ID	구매상품
001	A
002	C
003	A
001	B
002	C
003	B
003	A

df

pd.crosstab(
index = df['회원 ID'],
columns = df['구매상품'])

회원 ID	A	B	C
001	1	1	0
002	0	0	2
003	2	1	0

- 참고: 카이제곱 검정

Chapter. 11

비슷한 애들 모여라: 군집화

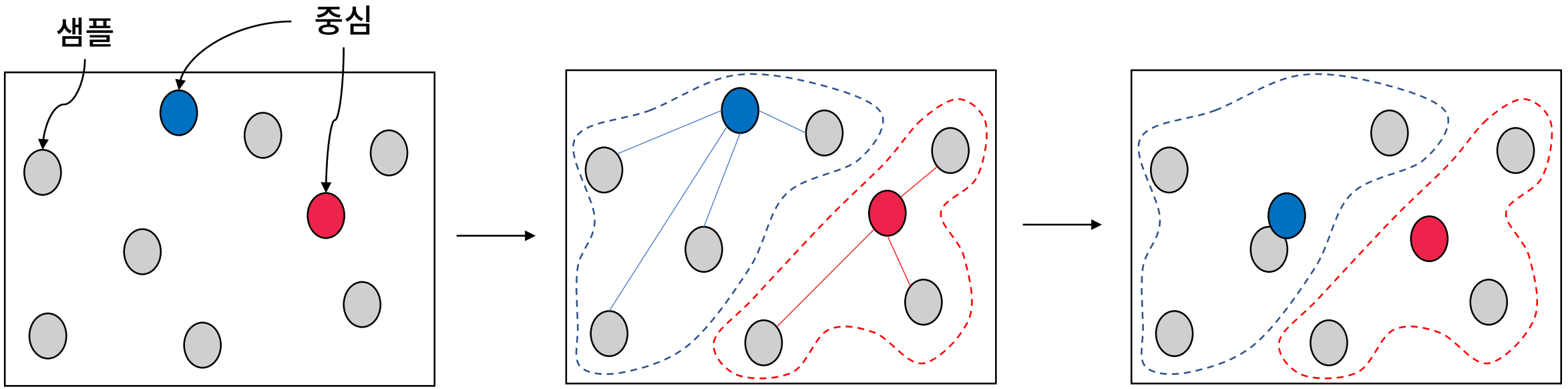
k-평균 군집화

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

I 기본 개념

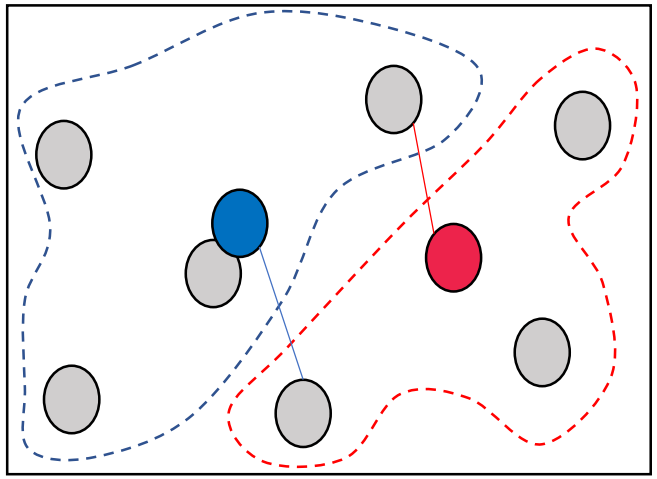
- (1) k개의 중심점 설정, (2) 샘플 할당, (3) 중심점 업데이트를 반복하는 방식으로 k개의 군집을 생성하는 알고리즘



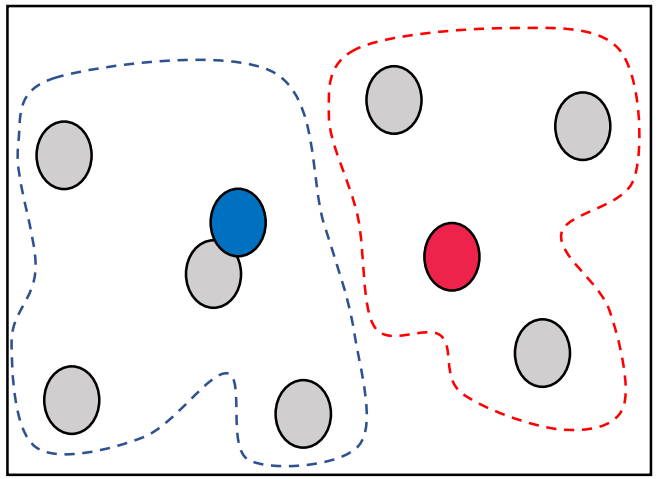
(1) 중심점 임의 설정 (2) 샘플 할당 (3) 중심 업데이트

I 기본 개념

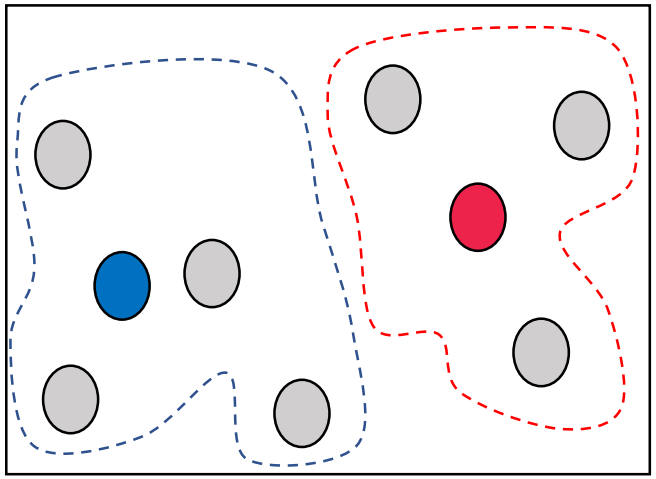
- (1) k개의 중심점 설정, (2) 샘플 할당, (3) 중심점 업데이트를 반복하는 방식으로 k개의 군집을 생성하는 알고리즘 (계속)



(4) 샘플 재할당



(5) 샘플 재할당에 따른 군집 업데이트



(6) 중심점 업데이트 및 종료

k-평균 군집화의 장단점

- 장점 (1) 상대적으로 **적은 계산량** $O(n)$
- 장점 (2) **군집 개수** 설정에 **제약이 없고** 쉬움

- 단점 (1) 초기 중심 설정에 따른 수행할 때마다 **다른 결과**를 낼 가능성 존재 (임의성 존재 O)
- 단점 (2) 데이터 **분포가 특이**하거나 **군집별 밀도 차이**가 존재하면 좋은 성능을 내기 어려움
- 단점 (3) 유클리디안 거리만 사용해야 함
- 단점 (4) 수렴하지 않을 가능성 존재

sklearn.cluster.KMeans

- 주요 입력
  - n_clusters: 군집 개수
  - max_iter: 최대 이터레이션 횟수

- 주요 메서드
  - fit(X): 데이터 X에 대한 군집화 모델 학습
  - fit_predict(X): 데이터 X에 대한 군집화 모델 학습 및 라벨 반환

- 주요 속성
  - labels_: fitting한 데이터에 있는 샘플들이 속한 군집 정보 (ndarray)
  - cluster_centers_: fitting한 데이터에 있는 샘플들이 속한 군집 중심점 (ndarray)

Chapter.

비슷한 애들 모여라: 군집화

감사합니다

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승