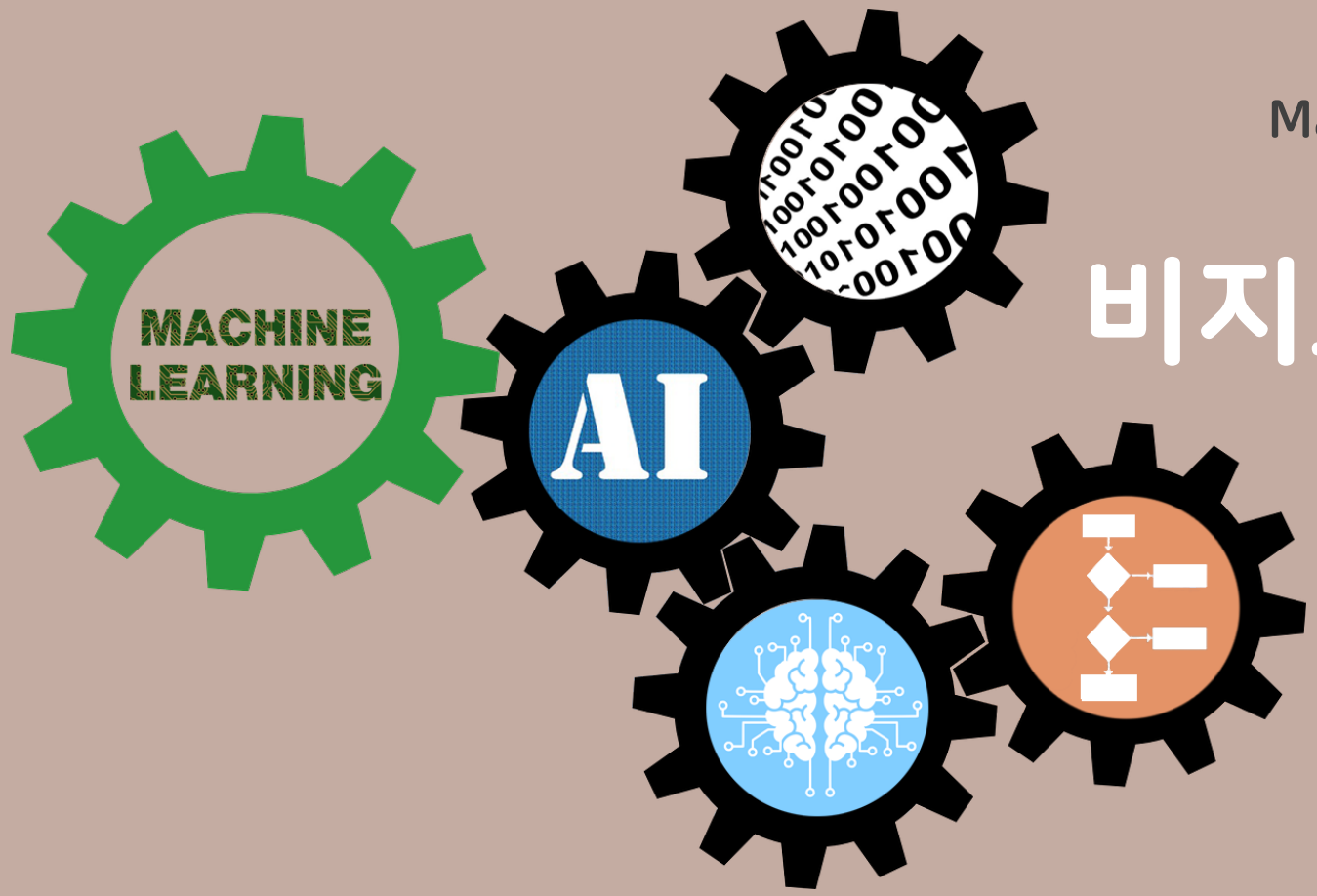


Machine Learning 4강

비지도학습: 군집화

컴퓨터과학과 이관용 교수

국립한국방송통신대학교
Korea National Open University

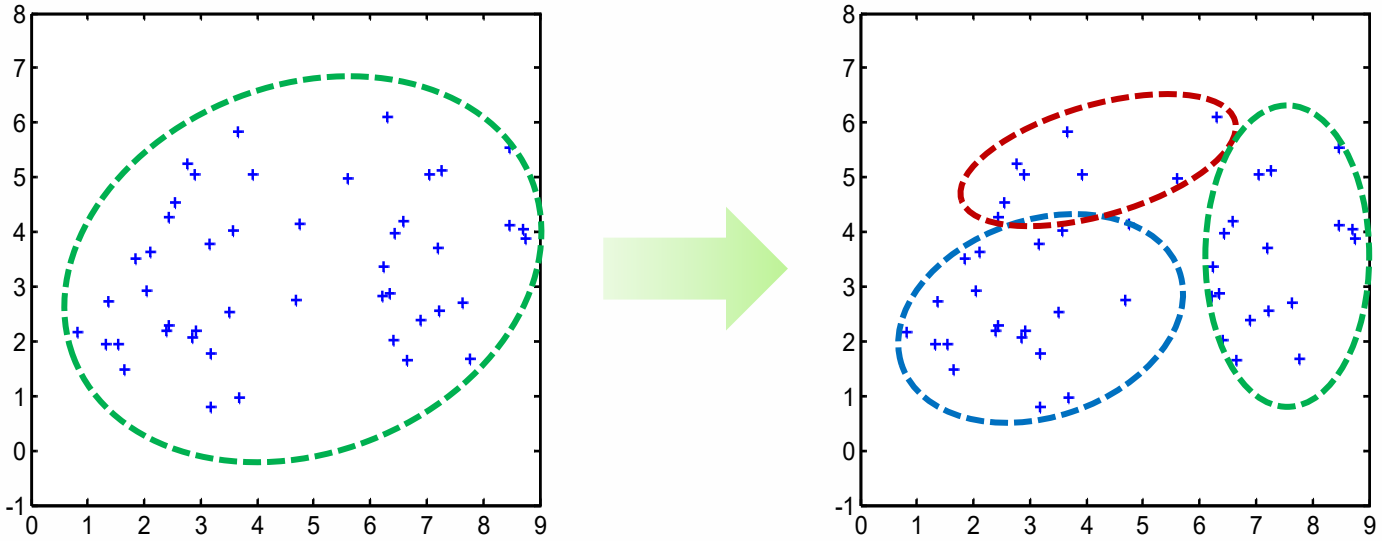
학습목차

1

군집화의 개념

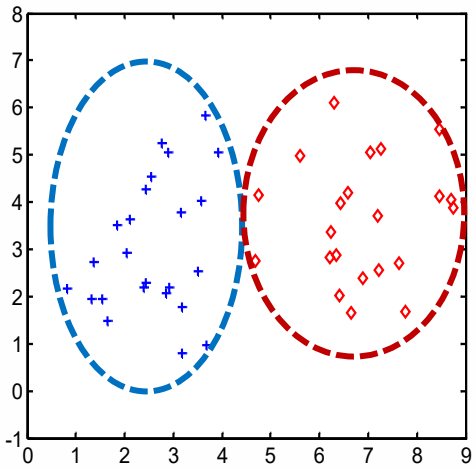
군집화?

- 데이터 집합의 분포 특성을 분석하여 서로 교차하지 않는 복수 개의 부분집합("군집", cluster)으로 나누는 문제
  - 입력 데이터로부터 추출된 특징 공간에서 특징값의 유사성에 따라 비슷한 데이터들끼리 묶음



분류 vs. 군집화

분류

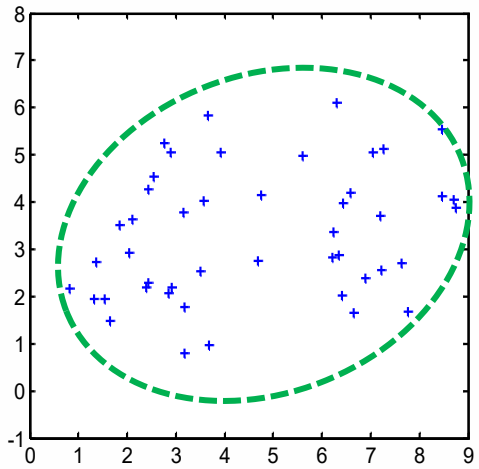


입력 → $\{(\boldsymbol{x}_i, y_i)\}_{i=1,\cdots,N}$

↓

지도학습

군집화

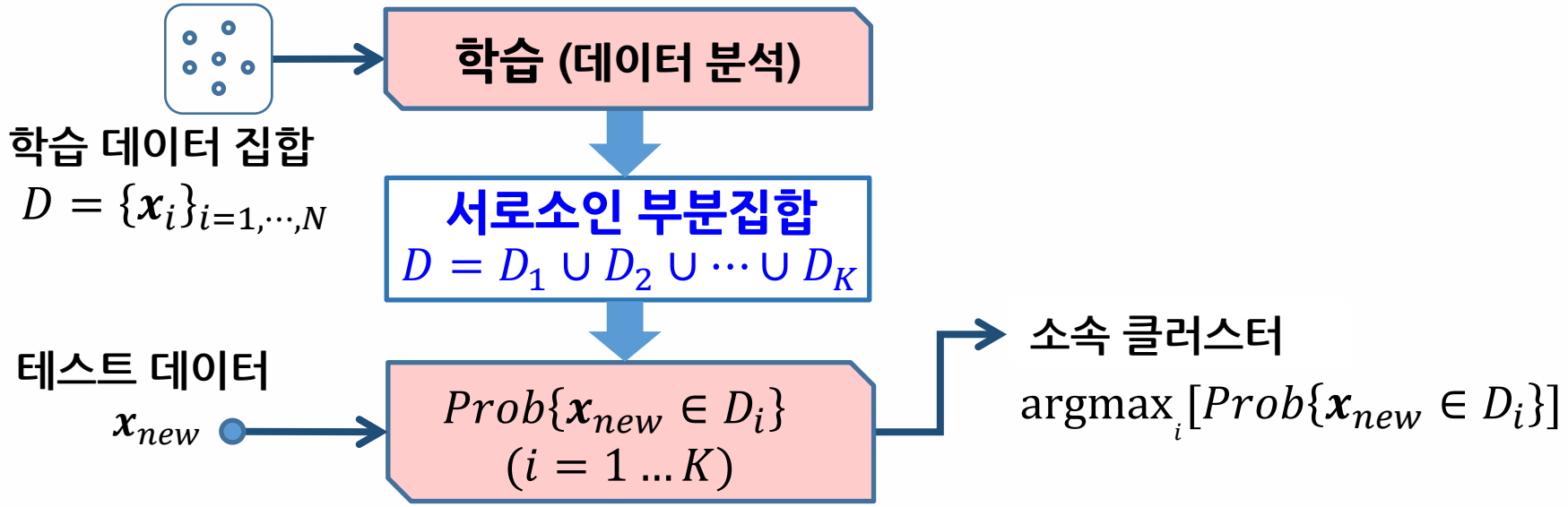


입력 → $\{\boldsymbol{x}_i\}_{i=1,\cdots,N}$

↓

비지도학습

군집화



적용 방법론

- K-평균 군집화 K-means clustering, 계층적 군집화 hierarchical clustering
- 가우시안 혼합 Gaussian mixture 모델, SOM Self Organizing Feature Map

군집화의 적용 예

장면 영상 데이터의 군집화

https://www.flickr.com/

영상 화소의 군집화에 의한 영상분할

http://cs.brown.edu/people/pfelzens/segment/

- 군집화가 적용 가능한 데이터?
  - 데이터에 대한 클래스 레이블이 주어지지 않는 경우
  - 데이터에 대한 클래스 레이블링에 비용이 많이 드는 경우

2 K-평균 군집화

K-평균 군집화 알고리즘

- 주어진 데이터 집합을 K개의 그룹으로 묶는 알고리즘

- 수행 단계

① 시작(초기화) → ② 데이터 그룹핑 → ③ 대표 벡터 수정 → ④ 반복 여부 결정

  - 데이터 집합 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$으로부터
    임의로 K개의 벡터를 선택하여
    K개의 초기 대표 벡터 집합 $\{\boldsymbol{m}_1, \boldsymbol{m}_2, \cdots, \boldsymbol{m}_K\}$를 생성함

알고리즘 수행 단계

① 시작(초기화) → ② 데이터 그룹핑 → ③ 대표 벡터 수정 → ④ 반복 여부 결정

$$C_k = \{\boldsymbol{x}_j | d(\boldsymbol{x}_j, \boldsymbol{m}_k) \leq d(\boldsymbol{x}_j, \boldsymbol{m}_i), i = 1, \cdots, K\}$$

- 각 데이터 $\boldsymbol{x}_j$ $(j = 1, \cdots, N)$에 대해 K개의 대표 벡터들과의 거리 $d(\boldsymbol{x}_j, \boldsymbol{m}_k)(k = 1, \cdots, K)$를 계산함
- 만약 데이터 $\boldsymbol{x}_j$가 대표 벡터 $\boldsymbol{m}_k$에 가장 가깝다면 이 데이터를 클러스터 C_k에 속하도록 레이블링함
- 이 과정을 통해 데이터 집합을 K개의 클러스터 $\{C_1, C_2, \cdots, C_K\}$로 나눔

알고리즘 수행 단계

① 시작(초기화) → ② 데이터 그룹핑 → **③ 대표 벡터 수정** → ④ 반복 여부 결정

- 단계 ②에서 구한 새로운 클러스터들에서 각각의 대표 벡터를 갱신함

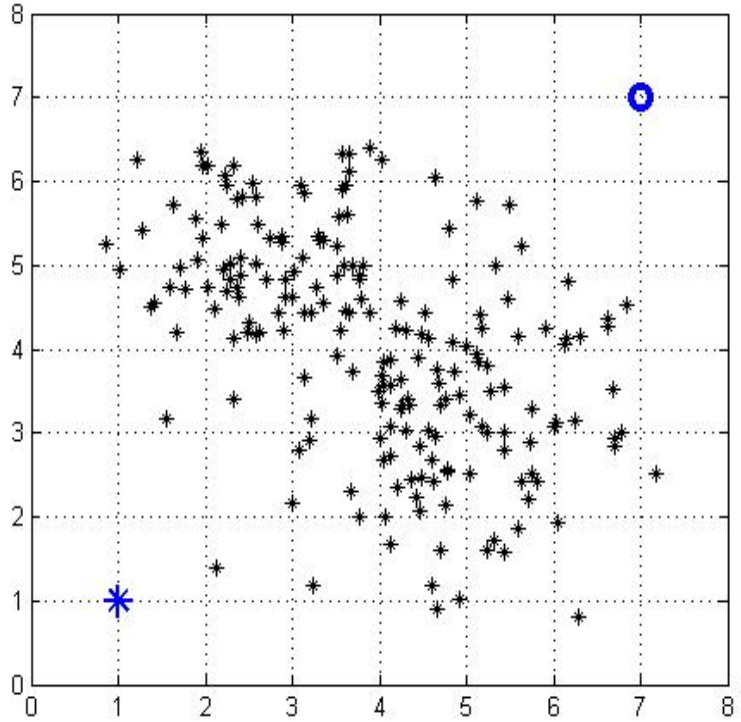
$$\boldsymbol{m}_k^{new} = \frac{1}{|C_k|} \sum_{x_j \in C_k} \boldsymbol{x}_j$$

① 시작(초기화) → ② 데이터 그룹핑 → ③ 대표 벡터 수정 → **④ 반복 여부 결정**

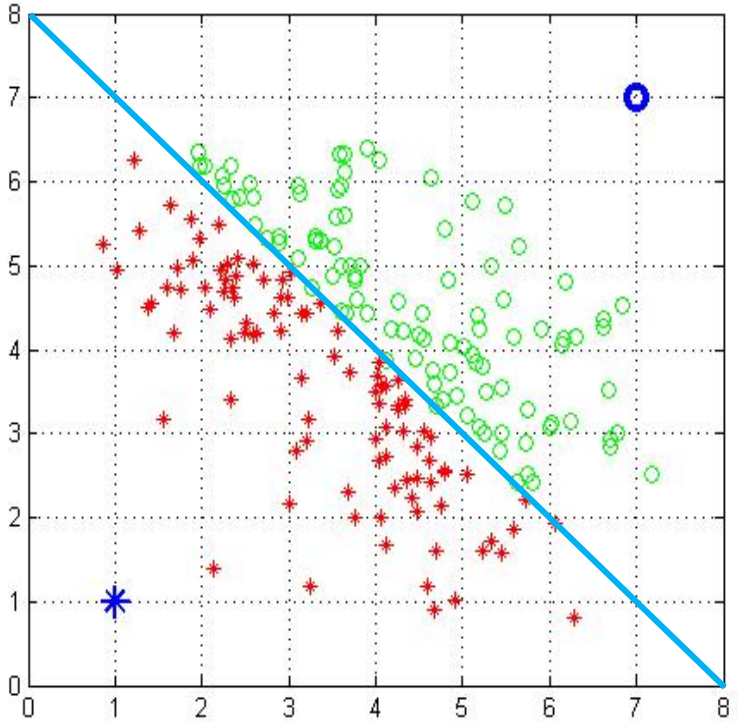
- 수정 전의 대표 벡터 $\boldsymbol{m}_k$와 수정 후의 대표 $\boldsymbol{m}_k^{new}$ 벡터의 차이를 계산하여 그 값에 변화가 없거나 설정된 반복 횟수에 도달할 때까지 단계 ②~④를 반복함

적용 과정의 예 (K=2인 경우)

초기 상태

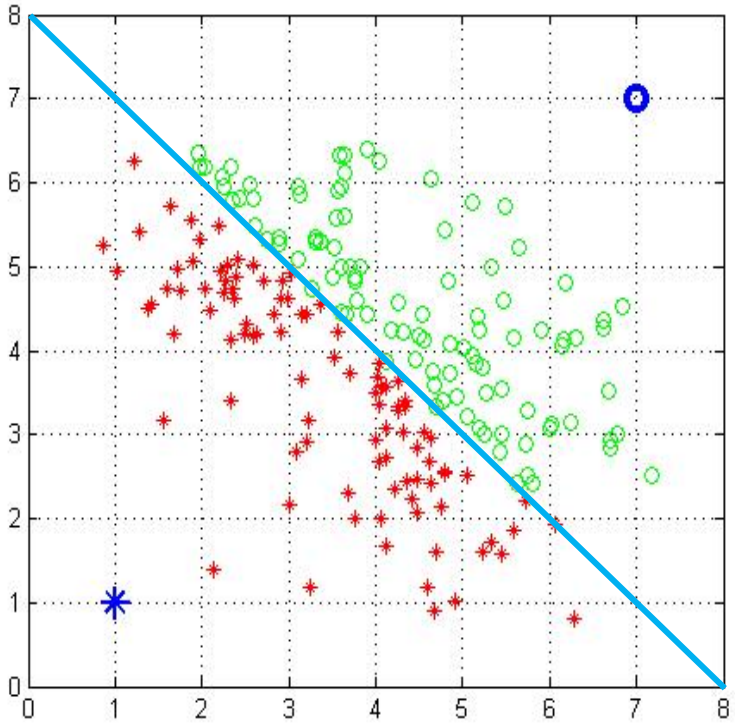


첫번째 데이터 그룹핑

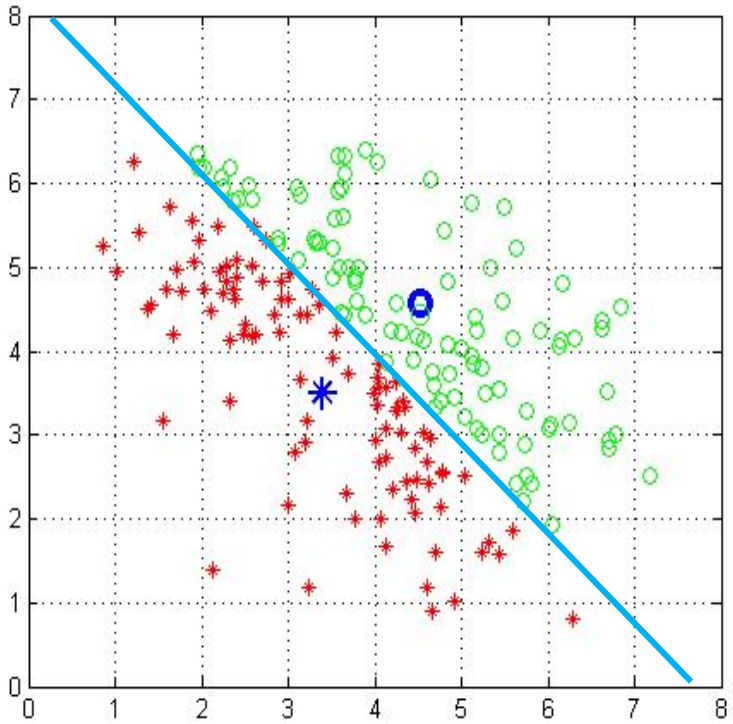


적용 과정의 예 (K=2인 경우)

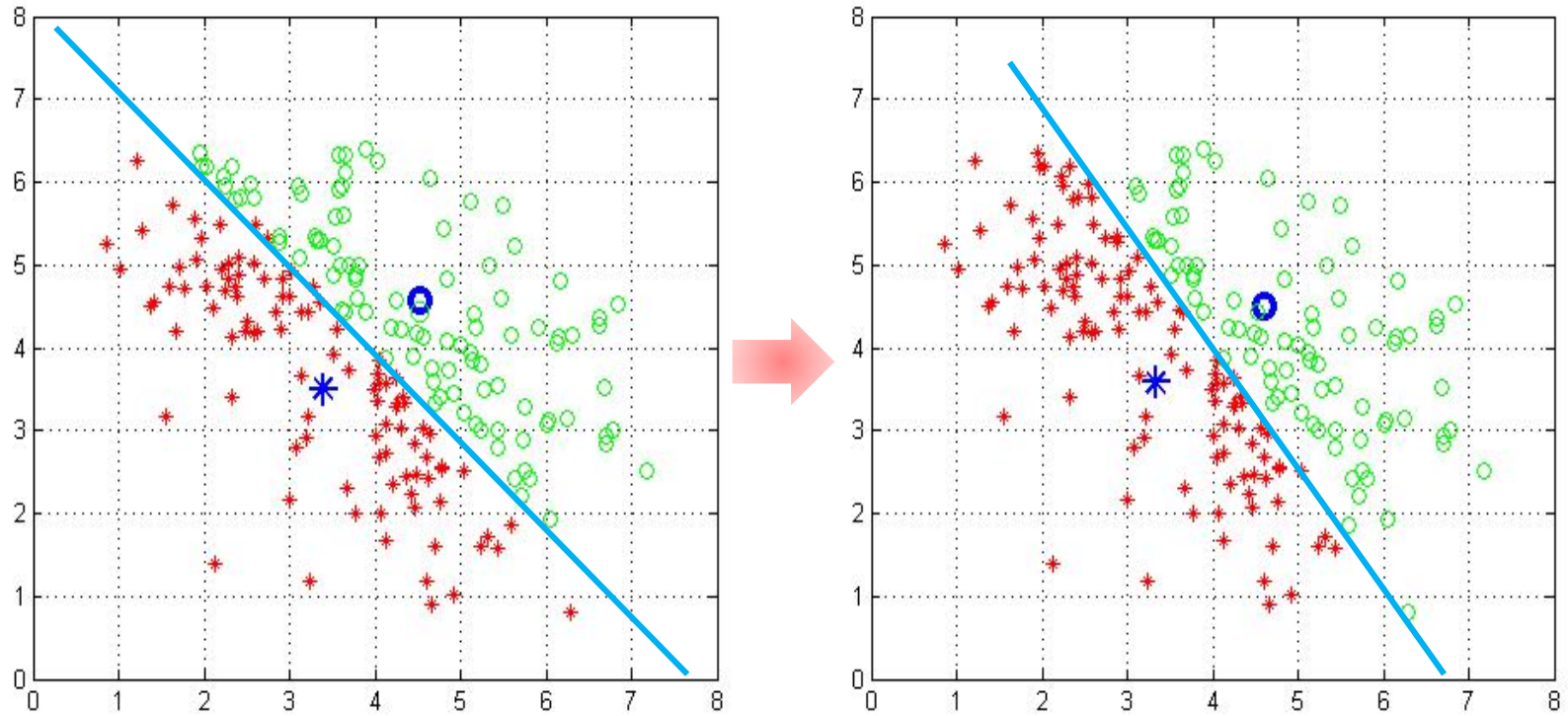
첫번째 데이터 그룹핑



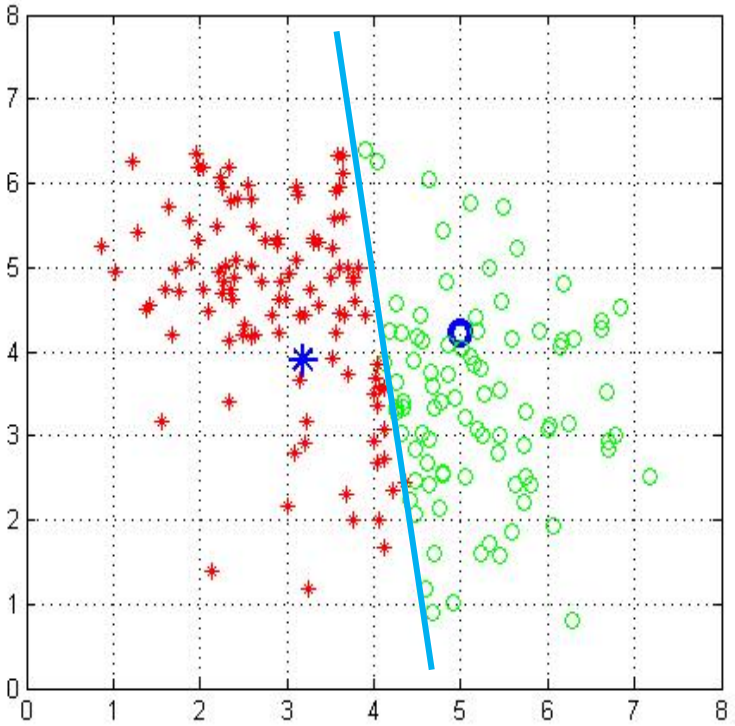
대표벡터 수정후 그룹핑

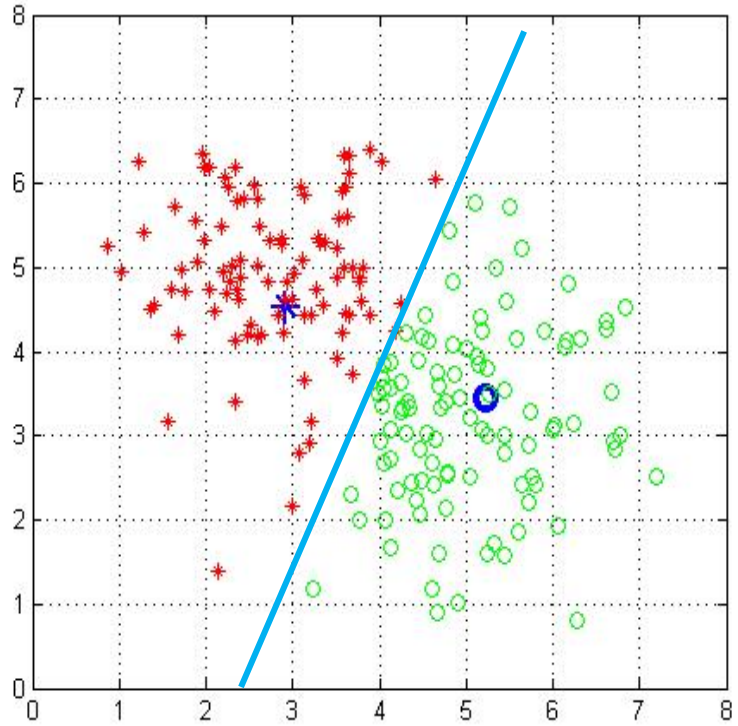


적용 과정의 예 (K=2인 경우)

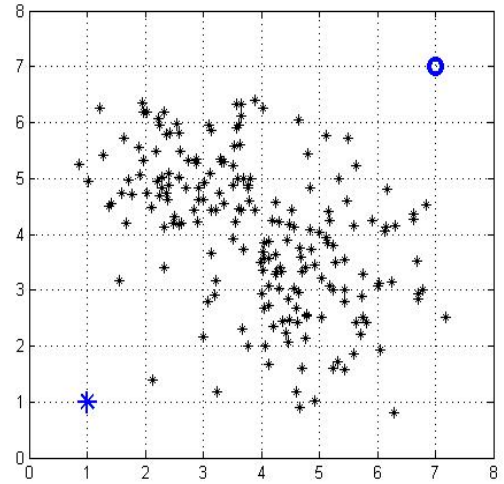


적용 과정의 예 (K=2인 경우)





초기 상태



알고리즘의 특성

- 실제 문제에 적용할 때 고려해야 할 사항

① 대표 벡터 계산과 데이터 그룹핑 과정의 반복적인 수행을 통해 좋은 군집을 찾는 것이 확실히 보장되는가?

② 초기 대표 벡터의 설정이 군집화의 성능에 미치는 영향은?

③ 데이터에 의존하는 적절한 K값을 어떻게 선택할 것인가?

알고리즘의 특성

① 반복수행 과정의 의미

K-평균 군집화 알고리즘의 목적함수

$$J = \sum_{n=1}^{N} \sum_{i=1}^{K} r_{ni} \left\| \boldsymbol{x}_n - \boldsymbol{m}_i \right\|^2$$

$$r_{ni} = \begin{cases} 1 & \text{if } i = \operatorname{argmin}_j \left\| \boldsymbol{x}_n - \boldsymbol{m}_j \right\|^2 \\ 0 & \text{otherwise} \end{cases}$$

J → 각 클러스터의 분산을 모두 더한 값

$J \uparrow$ → 각 클러스터 내의 데이터들이 서로 뭉쳐있지 않음

$J \downarrow$ → 각 클러스터 내에서는 데이터들이 잘 결집되어 있음

알고리즘의 특성

한 번 반복할 때마다 J의 값이 줄어드는 방향으로 학습이 진행

$$J = \sum_{n=1}^{N} \sum_{i=1}^{K} r_{ni} \left\| \boldsymbol{x}_n - \boldsymbol{m}_i \right\|^2$$

J의 값을 결정하는 파라미터 — 대표 벡터 $\boldsymbol{m}_i \ (i = 1, \cdots, K)$ — 각 데이터에 대한 클러스터 레이블 r_{ni}

1. $\boldsymbol{m}_i$가 결정되어 있을 때 r_{ni}를 결정하는 경우

J의 값이 최소화되기 위해서는 각 데이터로부터 가장 가까운 대표 벡터까지 거리의 합이 더해질 수 있도록 r_{ni}값 결정 ⇒ 『K-평균 알고리즘의 그룹핑 과정』

2. r_{ni}가 고정되어 있을 때 $\boldsymbol{m}_i$를 수정하는 경우

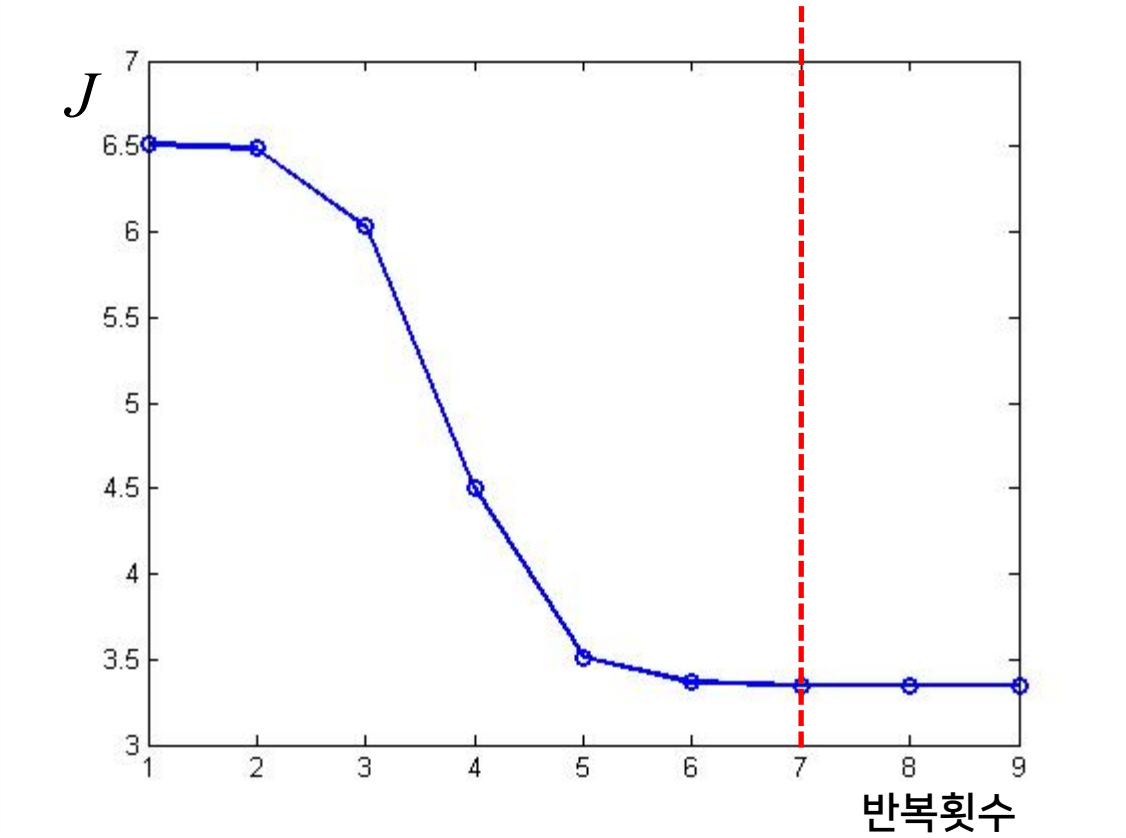
$$\frac{\partial J}{\partial \boldsymbol{m}_i} = 0 \longrightarrow \boldsymbol{m}_i = \frac{\sum_n r_{ni} \boldsymbol{x}_n}{\sum_n r_{ni}}$$

⇒ 『K-평균 알고리즘의 대표 벡터 수정식』

⇒ K-평균 군집화 알고리즘은 목적함수 J를 극소화하는 지역 극소점을 찾는 것을 보장

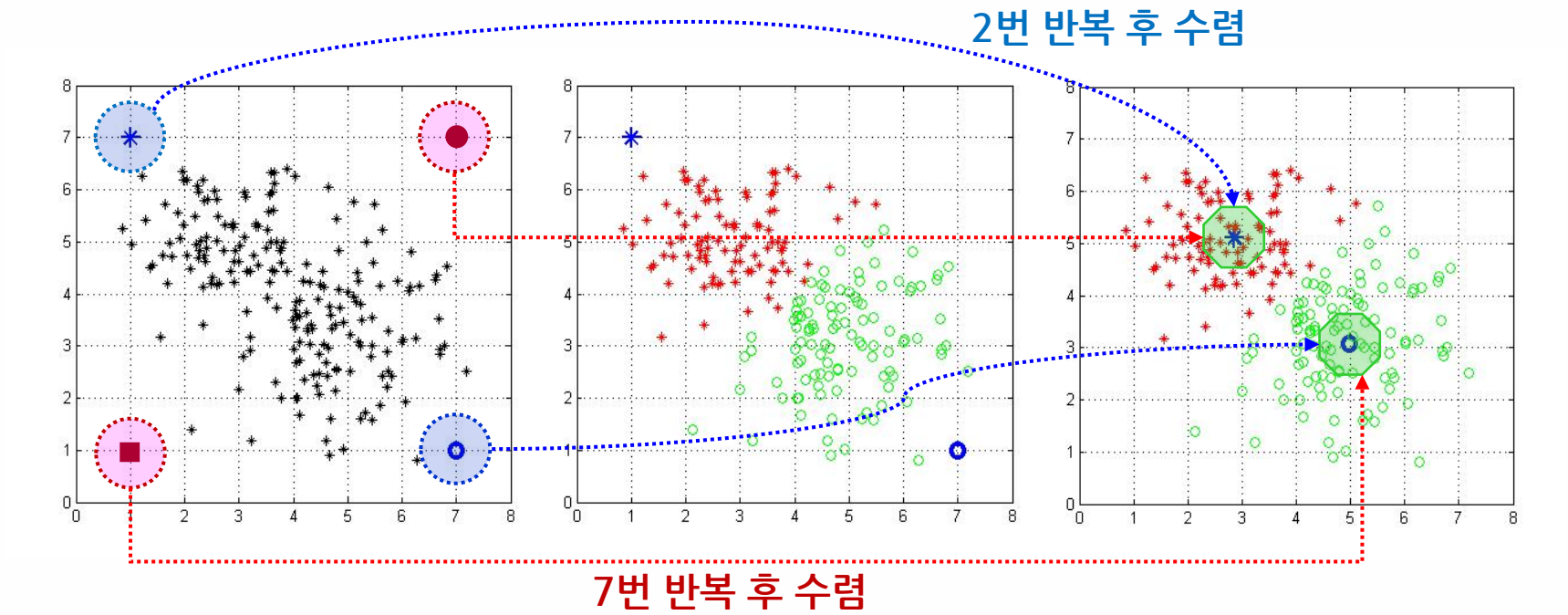
알고리즘의 특성

- 반복 횟수에 따른 J값의 변화



알고리즘의 특성

② 초기값에 대한 의존성 문제



- 초기에 임의로 결정하는 대표 벡터에 따라 최종적으로 찾아지는 해가 달라짐

알고리즘의 특성

③ K값에 따른 변화

- 적절한 K값의 선정은 주어진 문제에 의존적
  - ✓ 다양한 K값에 대해 군집화 결과들을 비교하여 선택?
  - ✓ 계층적 군집화 알고리즘?

3 계층적 군집화

계층적 군집화 알고리즘

전체 데이터를 몇 개의 배타적인 그룹으로 나누는 대신, 큰 군집이 작은 군집을 포함하는 형태로 계층을 이루도록 군집화를 수행하여 그 구조를 살펴보는 방법

병합적 방법 agglomerative or bottom up

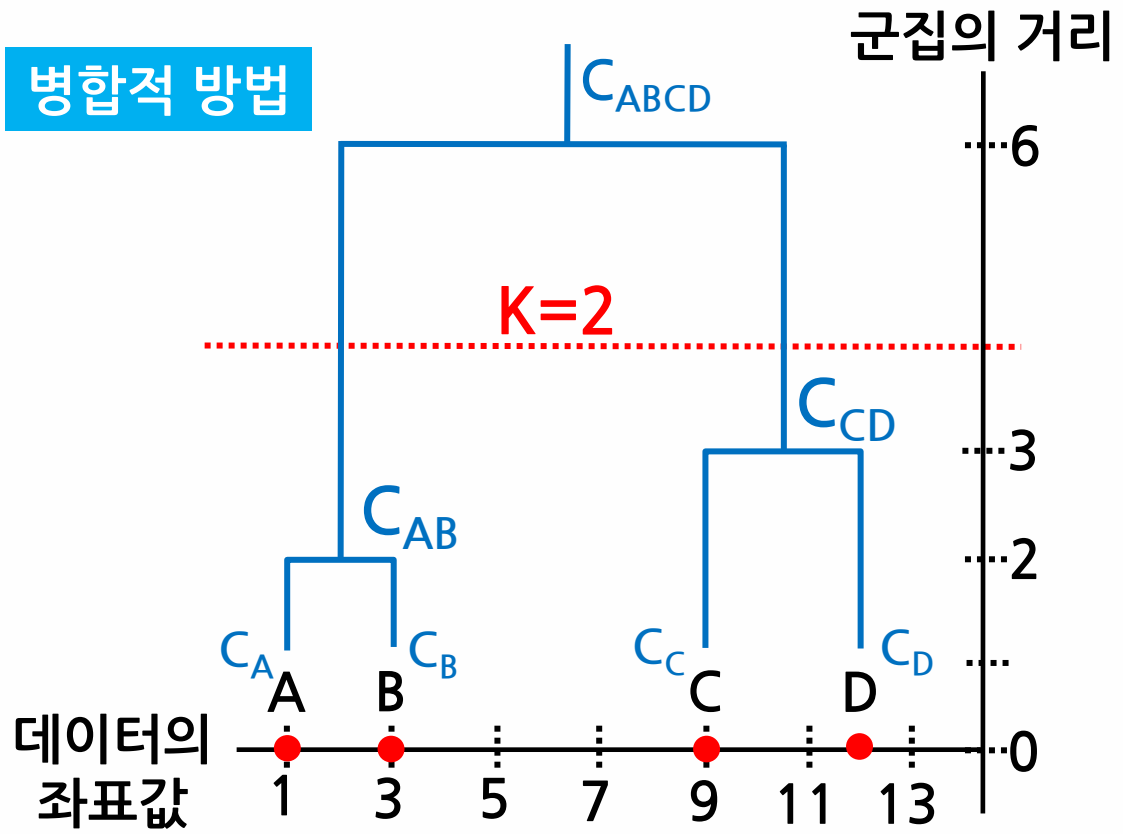
- 각 데이터가 하나의 군집을 이루는 최소 군집에서 시작하여 가까운 군집끼리 단계적으로 병합하여 더 큰 군집을 만들어 가는 방법
- N-1번의 병합 과정이 필요

분할적 방법 divisive or top down

- 모든 데이터가 하나의 군집에 속하는 최대 군집에서 시작하여 특정 기준에 따라 군집들을 분할해 가는 방법
- 가능한 분할 방법의 가지수 $2^{N-1}-1$개 → 비실용적

알고리즘 수행 과정의 예

dendrogram → 계층적인 군집화 결과를 보여주는 그림



알고리즘의 수행 단계 (병합적 방법)

① 데이터 집합 $\{x_1, x_2, \cdots, x_N\}$으로부터 각 데이터가 각각의 군집이 되도록 N개의 군집 $\{C_1, C_2, \cdots, C_N\}$을 설정함

② 가능한 모든 군집 쌍에 대해 군집 간의 거리를 계산함

$$d(C_i, C_j) = \min_{x_i \in C_i, x_j \in C_j} \{d(x_i, x_j)\}$$

③ 거리가 가장 가까운 두 군집 C_i, C_j를 선택하고 병합하여 새로운 클러스터 C_{ij}를 생성함

$$C_{ij} = C_i \cup C_j$$

④ 새로운 클러스터 C_{ij}를 클러스터 풀에 넣고, 원래 클러스터 C_i, C_j를 제거함

⑤ 오직 하나의 클러스터가 남을 때까지 ②~⑤의 과정을 반복함

계층적 군집화 알고리즘의 특성

① 군집 간의 거리를 계산하는 방식

최단 연결법 minimum 또는 single linkage	정의	$d(C_i, C_j) = \min_{x_i \in C_i, x_j \in C_j} \{d(x_i, x_j)\}$
	의미	가장 가까운 데이터 쌍 간의 거리
	특징	고립된 군집을 찾는 데 유용

최장 연결법 maximum 또는 complete linkage	정의	$d(C_i, C_j) = \max_{x_i \in C_i, x_j \in C_j} \{d(x_i, x_j)\}$
	의미	가장 멀리 떨어진 데이터 쌍 간의 거리
	특징	응집된 군집을 찾는 데 중점을 둠

계층적 군집화 알고리즘의 특성

① 군집 간의 거리를 계산하는 방식

중심 연결법 centroid linkage	정의	$d(C_i, C_j) = d(\boldsymbol{m}_i, \boldsymbol{m}_j) \quad \boldsymbol{m}_i = \frac{1}{\lvert C_i \rvert}\sum_{x \in C_i} \boldsymbol{x}, \ \boldsymbol{m}_j = \frac{1}{\lvert C_j \rvert}\sum_{x \in C_j} \boldsymbol{x}$
	의미	두 군집의 평균 간의 거리
	특징	특이값에 강건함
평균 연결법 mean 또는 average linkage	정의	$d(C_i, C_j) = \frac{1}{\lvert C_i \rvert \lvert C_j \rvert}\sum_{x_i \in C_i}\sum_{x_j \in C_j} d(\boldsymbol{x}_i, \boldsymbol{x}_j)$
	의미	모든 데이터 쌍 간 거리의 평균
	특징	작은 분산을 가지는 군집을 형성함

계층적 군집화 알고리즘의 특성

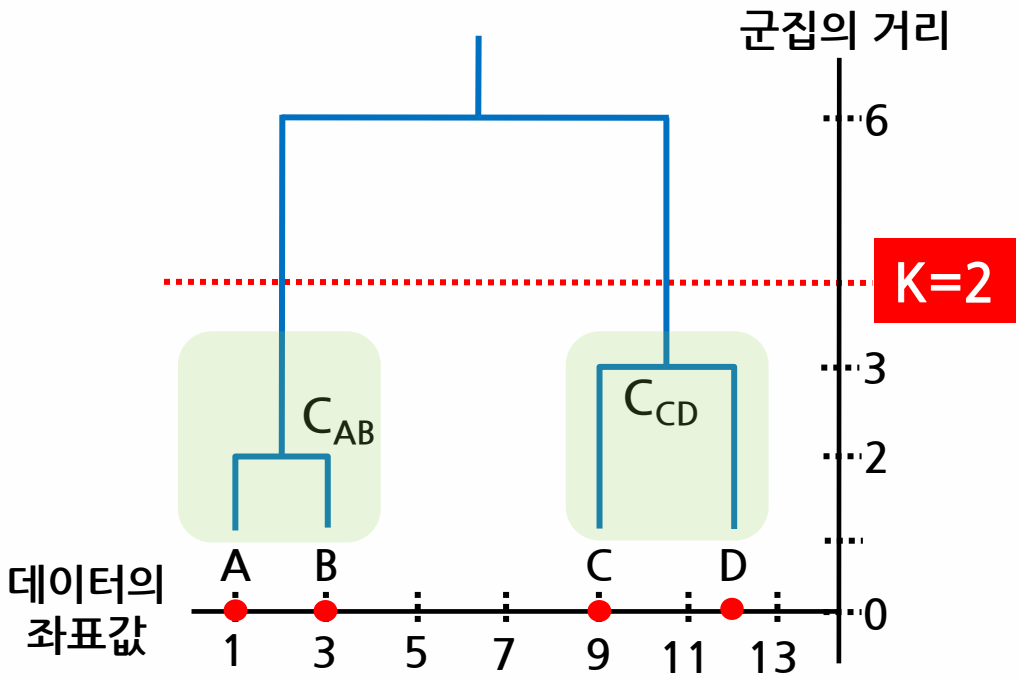
① 군집 간의 거리를 계산하는 방식

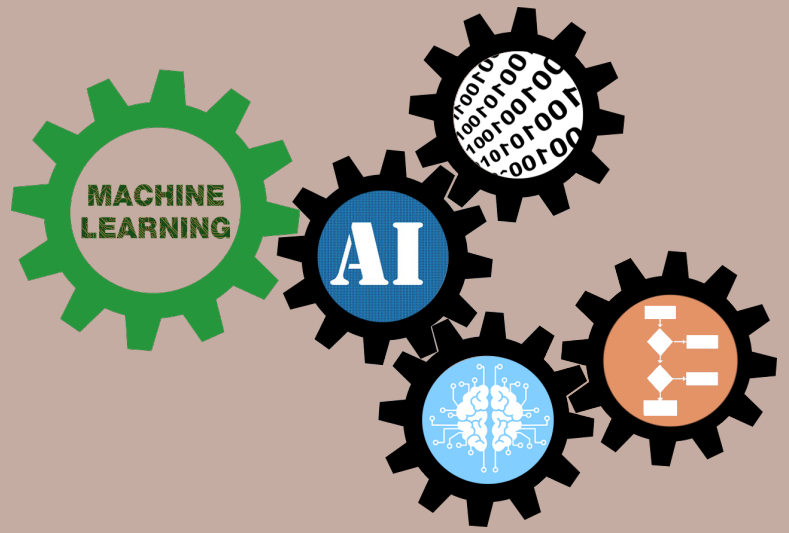
Ward's 방법	정의	$d(C_i, C_j) = \|C_i\|\|\boldsymbol{m}_i - \boldsymbol{m}\|^2 + \|C_j\|\|\boldsymbol{m}_j - \boldsymbol{m}\|^2$ $\boldsymbol{m}_i = \frac{1}{\|C_i\|}\sum_{\boldsymbol{x}\in C_i}\boldsymbol{x}, \boldsymbol{m}_j = \frac{1}{\|C_j\|}\sum_{\boldsymbol{x}\in C_j}\boldsymbol{x}, \ \boldsymbol{m} = \frac{1}{\|C_i\|+\|C_j\|}\sum_{\boldsymbol{x}\in C_i\cup C_j}\boldsymbol{x}$
	의미	병합 후의 클러스터 내부의 분산값
	특징	비슷한 크기의 군집을 병합함

계층적 군집화 알고리즘의 특성

② 덴드로그램으로부터 적합한 군집의 수를 결정하는 방법

- 덴드로그램에서 클러스터 간의 거리가 증가하는 동안
  클러스터의 수가 늘어나지 않고 일정 기간 유지되는 지점을 선택





다음시간안내

제5강

데이터 표현: 특징추출