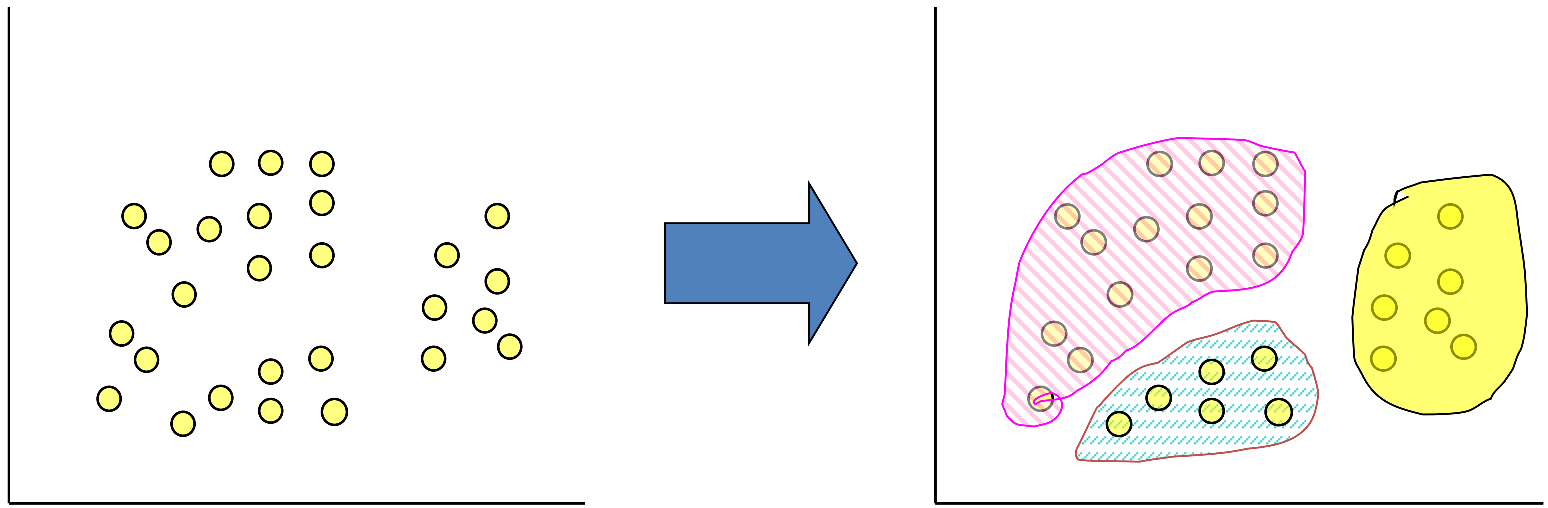
군집분석 (Clustering Analysis)

군집화 (Clustering) 개념

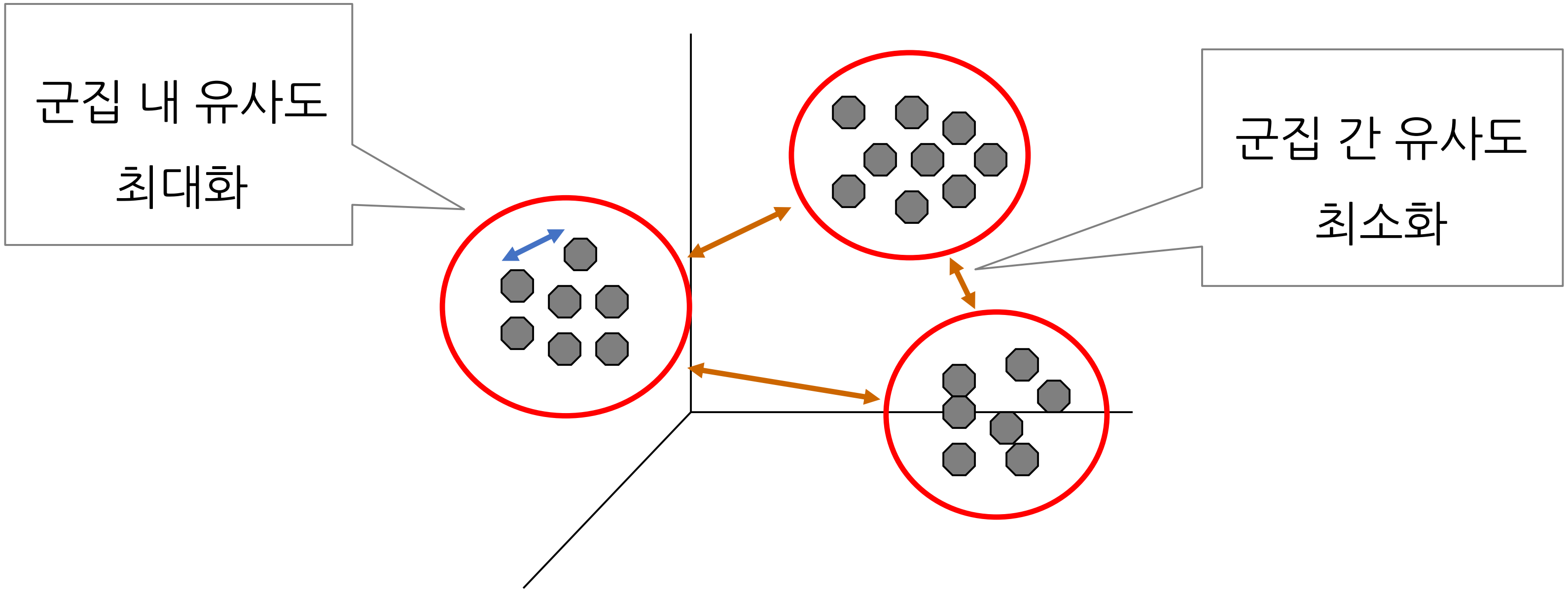
❖ 유사한 속성들을 갖는 관측치들을 묶어 전체 데이터를 몇 개의 군집(그룹)으로 나누는 것



군집화 개념

❖ 군집화 기준

- 동일한 군집에 소속된 관측치들은 서로 유사할수록 좋음
- 상이한 군집에 소속된 관측치들은 서로 다를수록 좋음

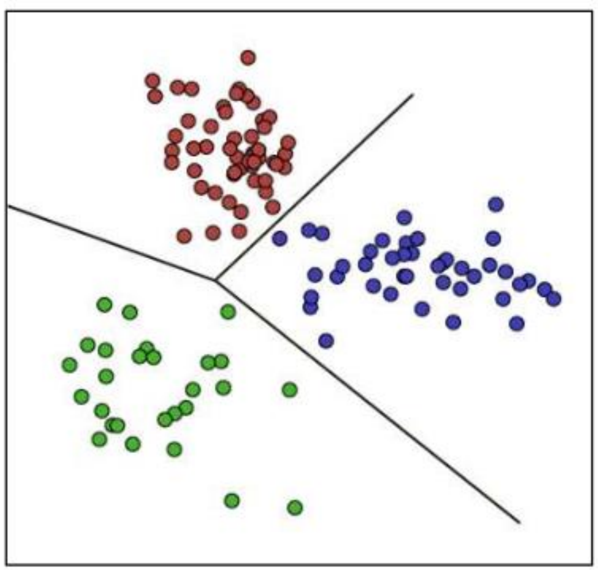


군집화 개념

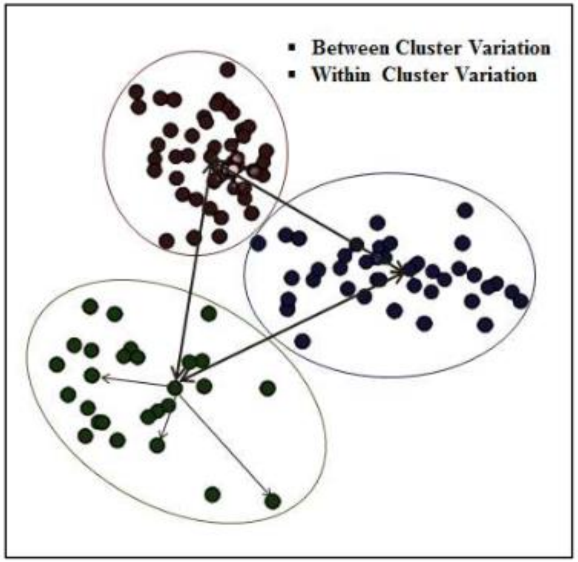
❖ 분류 (Classification) vs. 군집화 (Clustering)

- 분류: 사전 정의된 범주가 있는 (labeled) 데이터로부터 예측 모델을 학습하는 문제 (지도학습)
- 군집화: 사전 정의된 범주가 없는 (unlabeled) 데이터에서 최적의 그룹을 찾아나가는 문제 (비지도학습)

분류



군집

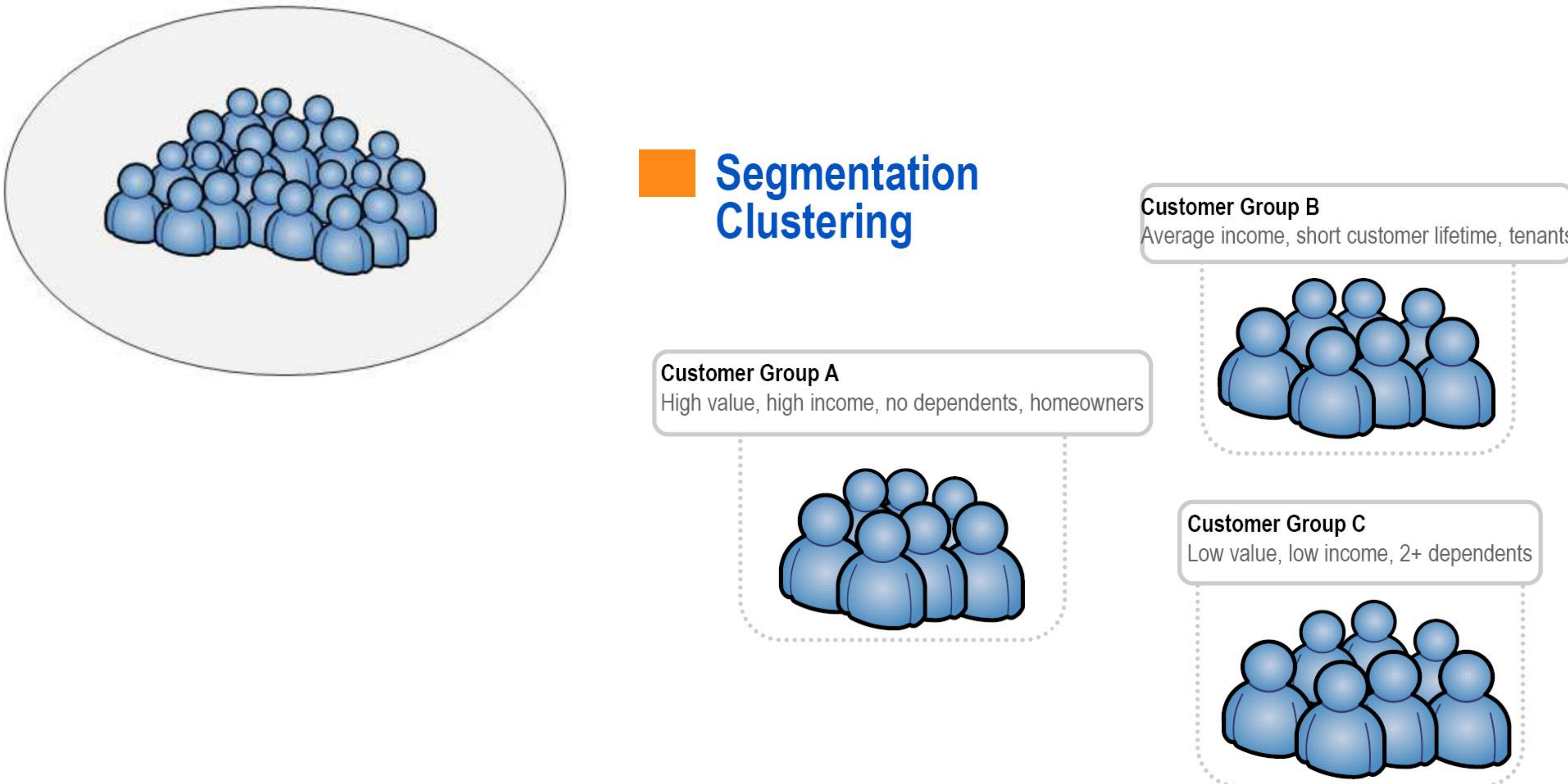


고려대학교 KOREA UNIVERSITY

군집화 적용사례

❖ 군집화 적용 사례

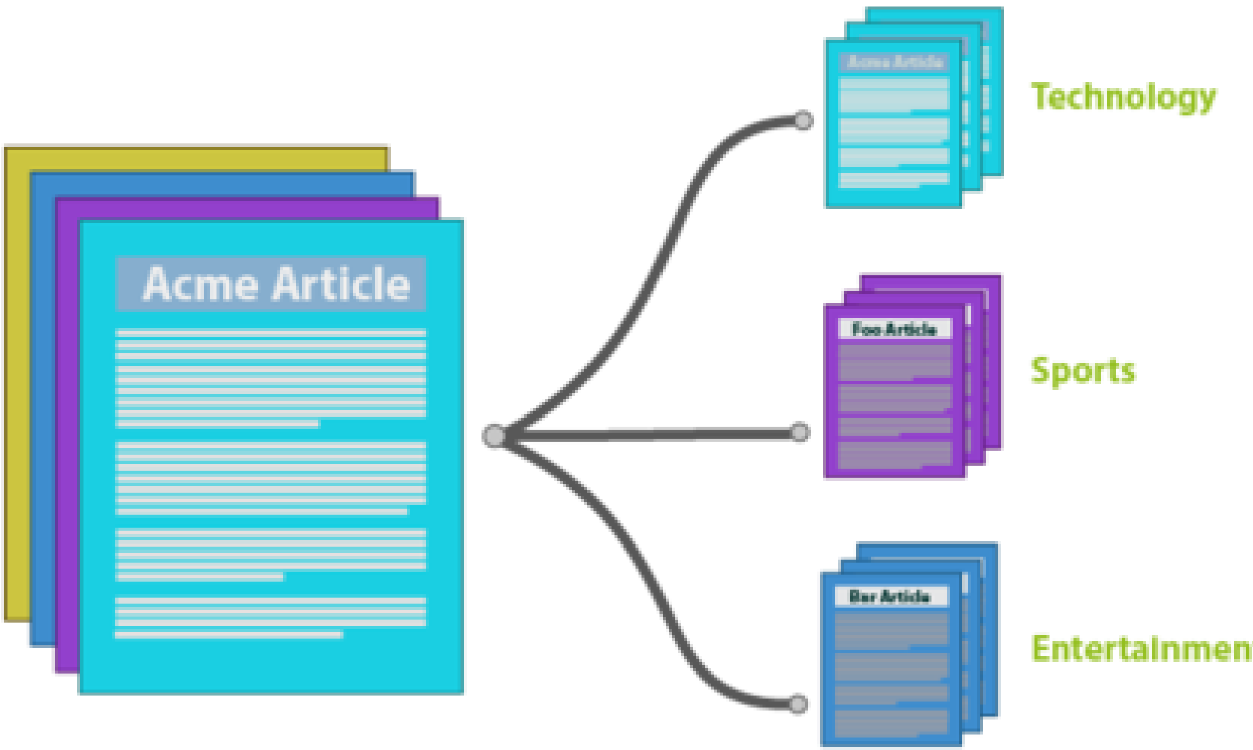
- 특성 별 고객 군집 (Customer Segmentation)



군집화 적용사례

❖ 군집화 적용 사례

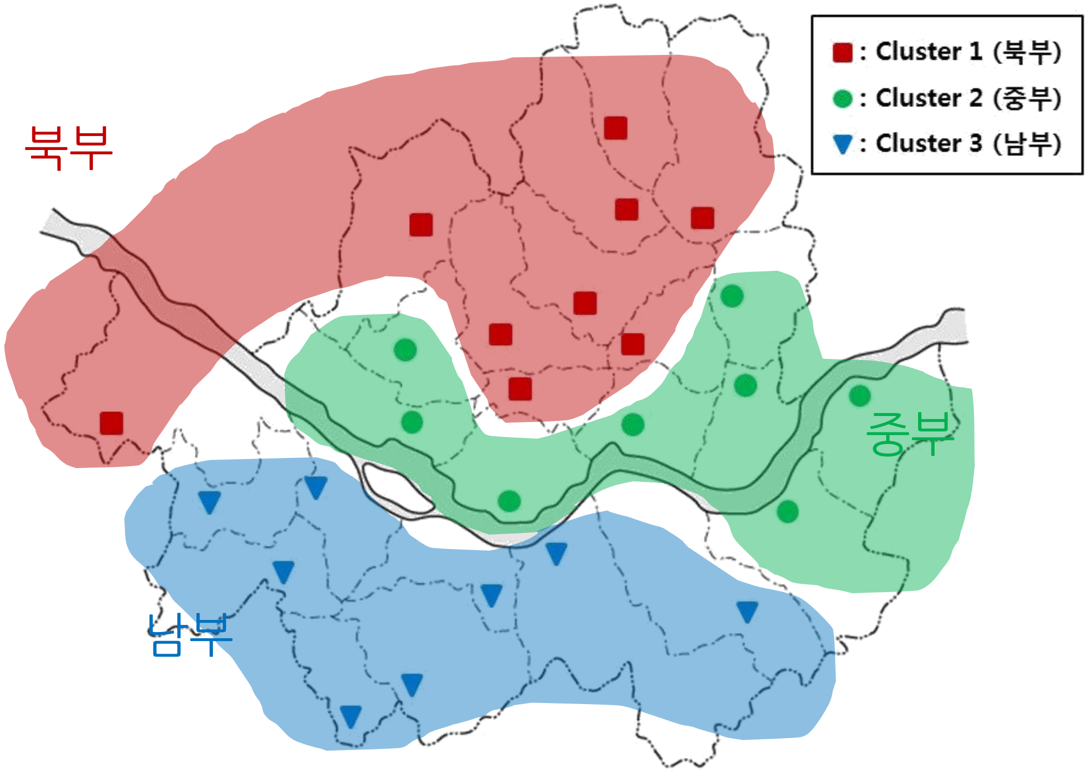
- 유사문서 군집화



군집화 적용사례

❖ 군집화 적용 사례

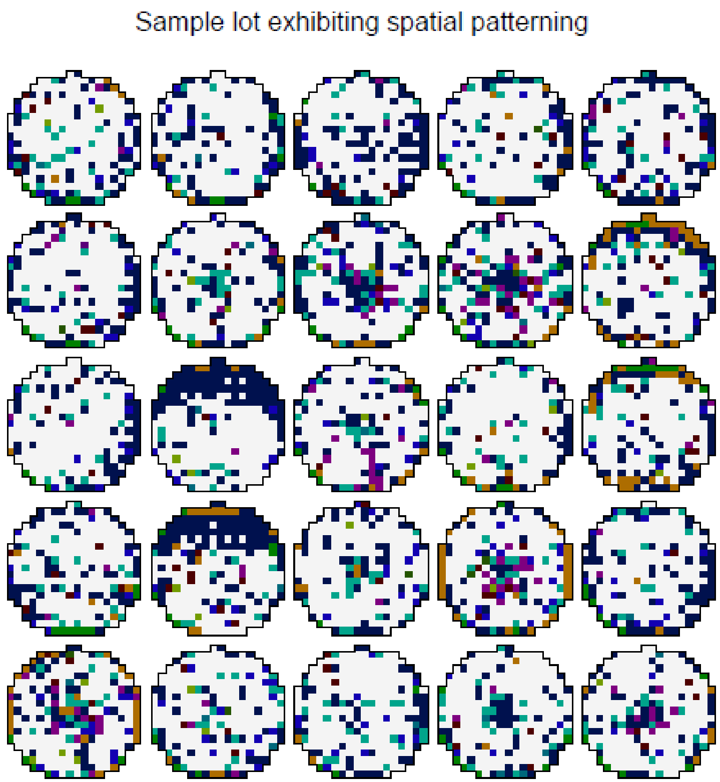
- 서울시 오존농도 패턴 군집화 (25개 구)

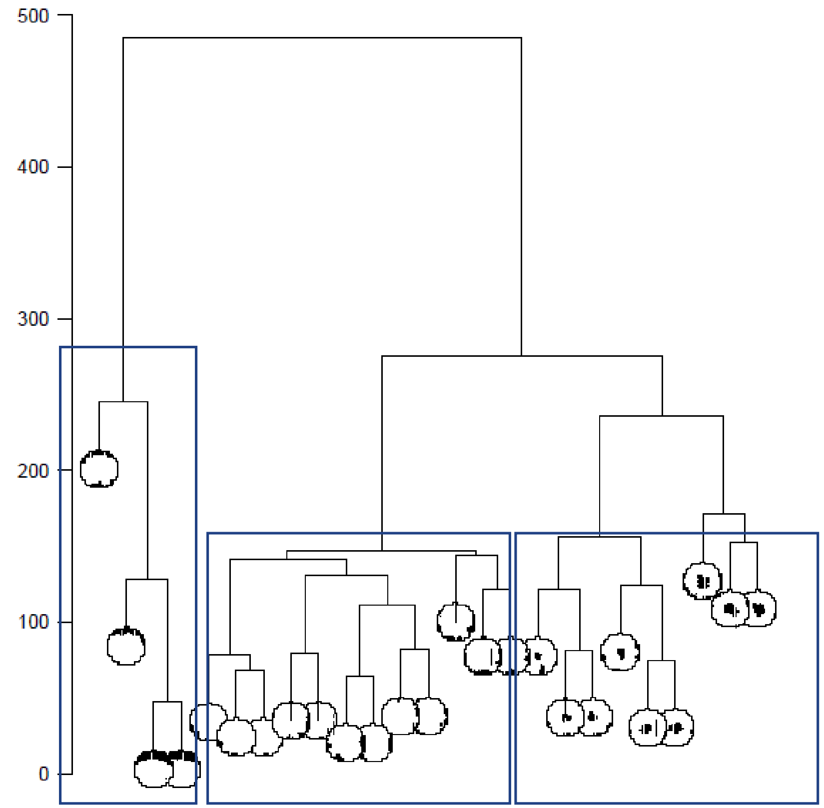


군집화 적용사례

군집화 적용 사례

- 웨이퍼 Fail bit map 군집화





군집화 수행 시 주요 고려사항

- ❖ 어떤 거리 척도를 사용하여 유사도를 측정할 것인가?
- ❖ 어떤 군집화 알고리즘을 사용할 것인가?
- ❖ 어떻게 최적의 군집 수를 결정할 것인가?
- ❖ 어떻게 군집화 결과를 측정/평가할 것인가?

군집화: 유사도 척도

- 어떤 거리 척도를 사용하여 유사도를 측정할 것인가?
  - 유클리디안 거리 (Euclidean Distance)
  - 맨하탄 거리 (Manhattan Distance)
  - 마할라노비스 거리 (Mahalanobis Distance)
  - 상관계수 거리 (Correlation Distance)

군집화: 유사도 척도

❖ 유클리디안 거리

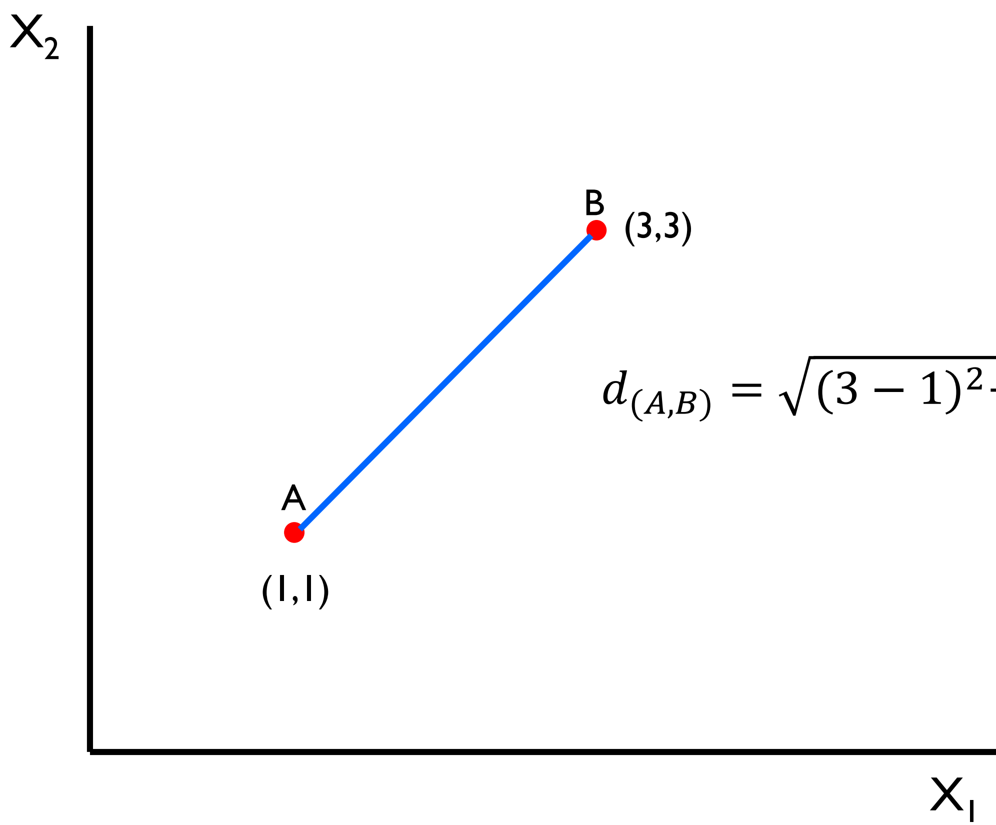
- 일반적으로 사용하는 거리 척도
- 대응되는 관측치 X, Y 값 간 차이 제곱합의 제곱근으로써, 두 관측치 사이의 직선 거리를 의미함

$$d_{(X,Y)} = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$$

군집화: 유사도 척도

❖ 유클리디안 거리

- 2 차원



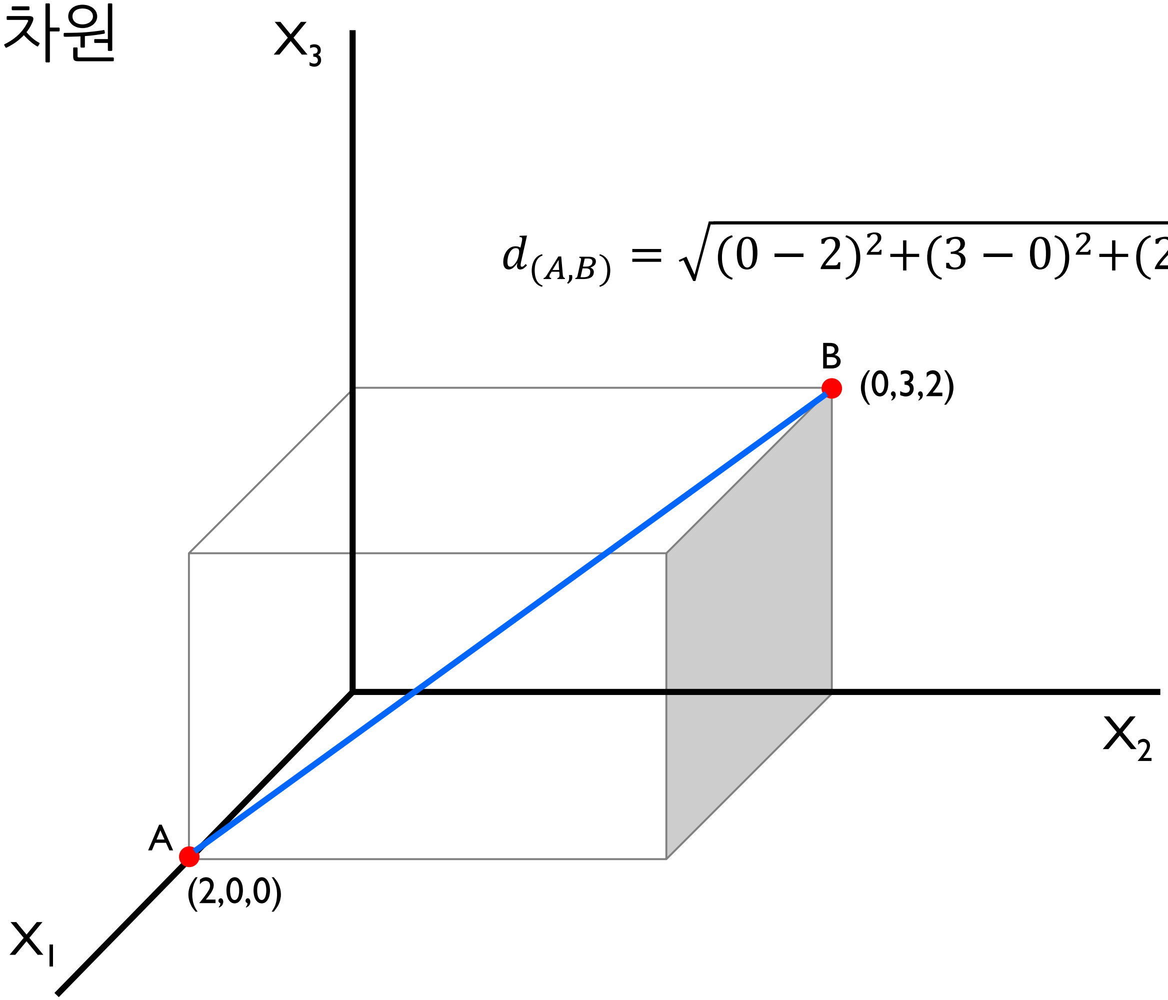
$$d_{(A,B)} = \sqrt{(3-1)^2 + (3-1)^2} = \sqrt{8}$$

군집화: 유사도 척도

❖ 유클리디안 거리

- 3 차원

$$d_{(A,B)} = \sqrt{(0-2)^2+(3-0)^2+(2-0)^2} = \sqrt{17}$$



군집화: 유사도 척도

❖ 유클리디안 거리

- p 차원

$$A = (a_1, a_2, \cdots, a_p)$$
$$B = (b_1, b_2, \cdots, b_p)$$

$$d_{(A,B)} = \sqrt{(a_1 - b_1)^2 + \cdots + (a_p - b_p)^2} = \sqrt{\sum_{i=1}^{p} (a_i - b_i)^2}$$

군집화: 유사도 척도

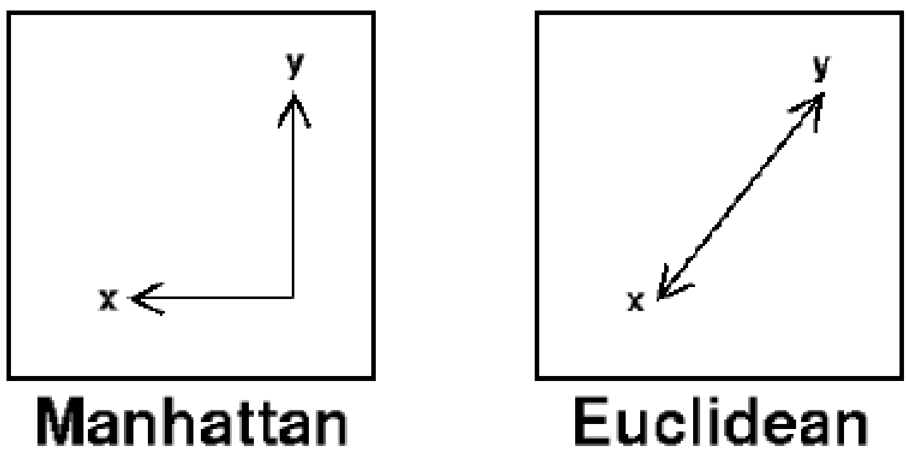
❖ 맨하탄 거리 (Manhattan Distance)

고려대학교 KOREA UNIVERSITY

군집화: 유사도 척도

- 맨하탄 거리 (Manhattan Distance)
  - X에서 Y로 이동 시 각 좌표축 방향으로만 이동할 경우에 계산되는 거리

$$d_{Manhattan(X,Y)} = \sum_{i=1}^{p} |x_i - y_i|$$



군집화: 유사도 척도

❖ 마할라노비스 거리

$$d_{Mahalanobis(X,Y)} = \sqrt{(X-Y)^T \Sigma^{-1} (X-Y)}$$

where Σ^{-1}= Inverse of covariance matrix

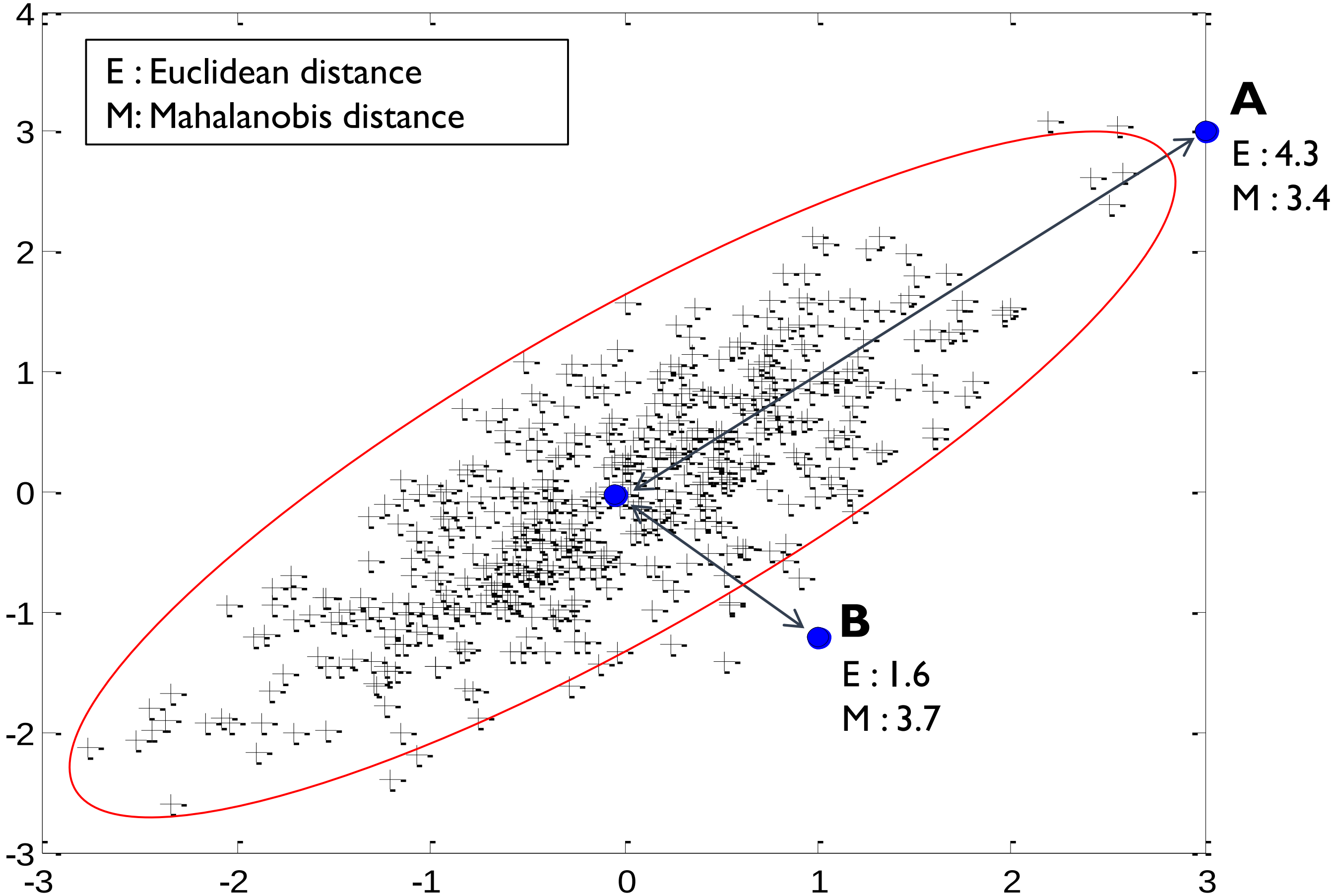
- 변수 내 분산, 변수 간 공분산을 모두 반영하여 *X*, *Y* 간 거리를 계산하는 방식

- 데이터의 covariance matrix가 identity matrix인 경우는 Euclidean distance와 동일함

고려대학교 KOREA UNIVERSITY

군집화: 유사도 척도

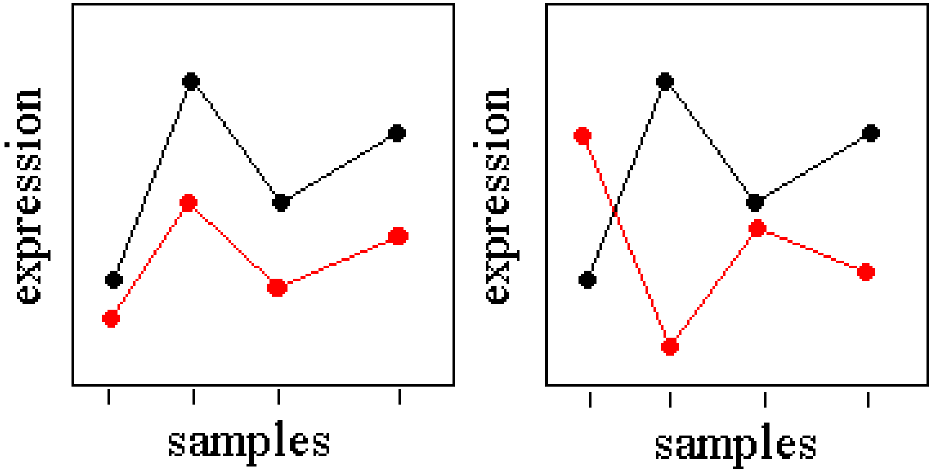
❖ 마할라노비스 거리



군집화: 유사도 척도

❖ 상관계수 거리

$$d_{Corr(X,Y)} = 1 - r$$
$$\text{where } r = \sigma_{XY}$$



• 데이터 간 Pearson correlation을 거리 척도로 직접 사용하는 방식으로, 데이터 패턴의 유사도 / 비유사도를 반영할 수 있음

군집화: 유사도 척도

❖ 스피어만 상관계수 거리

$$d_{Spearman(X,Y)} = 1 - \rho,$$

$$\text{where } \rho = 1 - \frac{6\sum_{i=1}^{n}(rank(x_i) - rank(y_i))^2}{n(n^2-1)}$$

- ρ를 Spearman correlation이라 하며, 이는 데이터의 rank를 이용하여 correlation distance를 계산하는 방식임
- ρ의 범위는 -1 부터 1로, Pearson correlation과 동일

군집화: 유사도 척도

지역	봄	여름	가을	겨울
서울	17.06	28.43	19.07	3.50
뉴욕	16.32	28.22	18.37	5.43
시드니	22.23	17.03	21.90	25.63

계절 평균 낮 최고 기온

지역	봄	여름	가을	겨울
서울	3	1	2	4
뉴욕	3	1	2	4
시드니	2	4	3	1

지역 별 계절 기온 순위

서울 - 뉴욕 간 Spearman correlation distance:

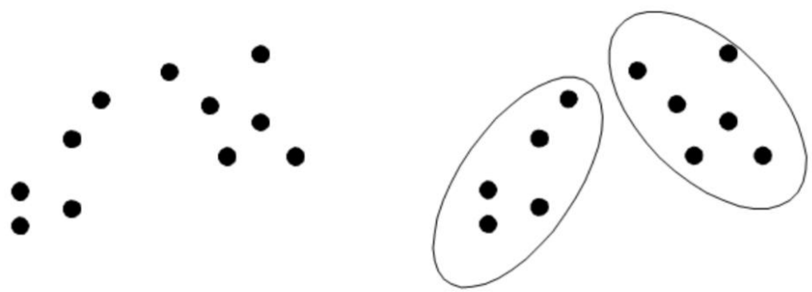
$$\rho = 1 - \frac{6\{(3-3)^2 + (1-1)^2 + (2-2)^2 + (4-4)^2\}}{4(4^2-1)} = 1 \;\longrightarrow\; d_{(서울,뉴욕)} = 1 - 1 = 0$$

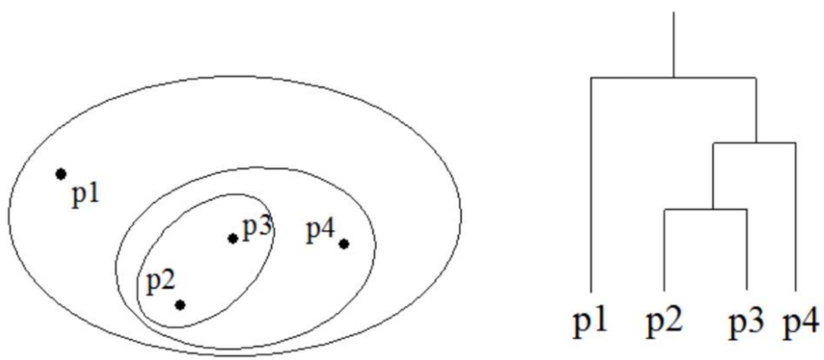
서울 - 시드니 간 Spearman correlation distance:

$$\rho = 1 - \frac{6\{(3-2)^2 + (1-4)^2 + (2-3)^2 + (4-1)^2\}}{4(4^2-1)} = -1 \;\longrightarrow\; d_{(서울,시드니)} = 1 - (-1) = 2$$

군집화: 알고리즘

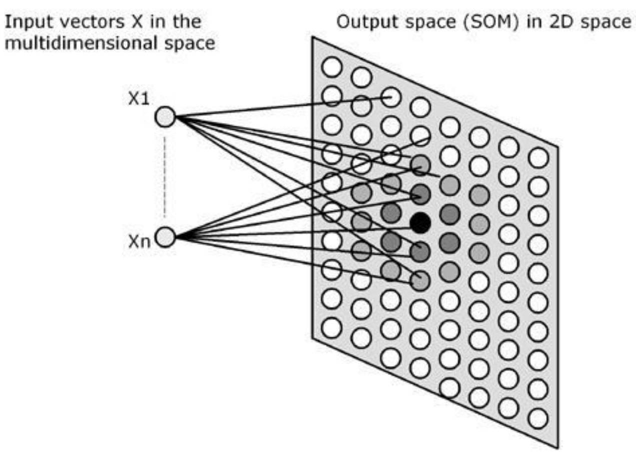
- 어떤 군집화 알고리즘을 사용할 것인가?
- 군집화 알고리즘의 종류
  - 분리형 군집화
    - 전체 데이터의 영역을 특정 기준에 의해 동시에 구분
    - 각 개체들은 사전에 정의된 개수의 군집 중 하나에 속하게 됨
  - 계층적 군집화
    - 개체들을 가까운 집단부터 차근차근 묶어나가는 방식
    - 군집화 결과 뿐만 아니라 유사한 개체들이 결합되는 dendrogram도 생성

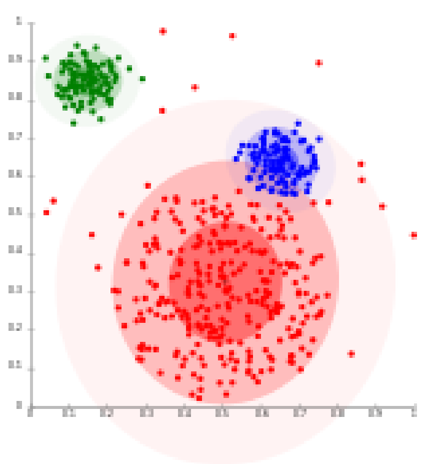




군집화: 알고리즘

- ❖ 어떤 군집화 알고리즘을 사용할 것인가?
- ❖ 군집화 알고리즘의 종류
  - ▪ 자기조직화 지도
    - • 2차원의 격자에 각 개체들이 대응하도록 인공신경망과 유사한 학습을 통해 군집 도출
  - ▪ 분포 기반 군집화
    - • 데이터의 분포를 기반으로 높은 밀도를 갖는 세부 영역들로 전체 영역을 구분

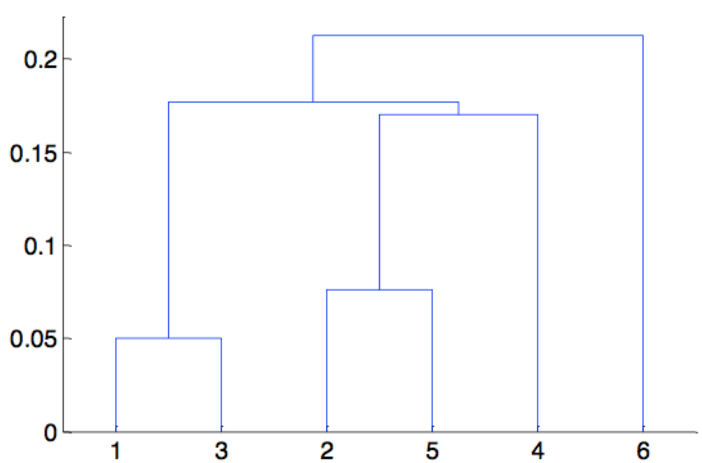


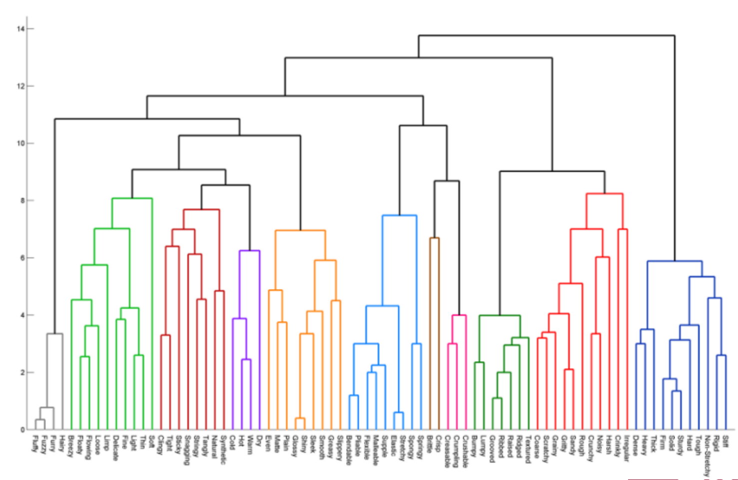


계층적 군집화 (Hierarchical Clustering)

❖ 계층적 군집화

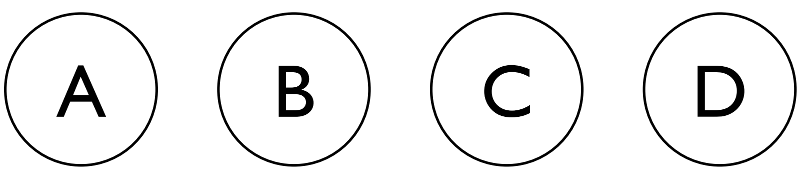
- 계층적 트리모형을 이용하여 개별 개체들을 순차적/계층적으로 유사한 개체/군집과 통합
- 덴드로그램(Dendrogram)을 통해 시각화 가능
  - ✓ 덴드로그램: 개체들이 결합되는 순서를 나타내는 트리형태의 구조
- 사전에 군집의 수를 정하지 않아도 수행 가능
  - ✓ 덴드로그램 생성 후 적절한 수준에서 자르면 그에 해당하는 군집화 결과 생성





계층적 군집화 (Hierarchical Clustering)

- 계층적 군집화 수행 예시
  - 모든 개체들 사이의 거리에 대한 유사도 행렬 계산

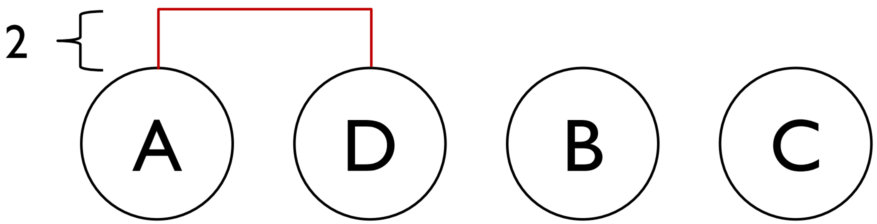


	A	B	C	D
A		20	7	2
B			10	25
C				3
D				

계층적 군집화

❖ 계층적 군집화 수행 예시

- 모든 개체들 사이의 거리에 대한 유사도 행렬 계산
- 거리가 인접한 관측치끼리 군집 형성

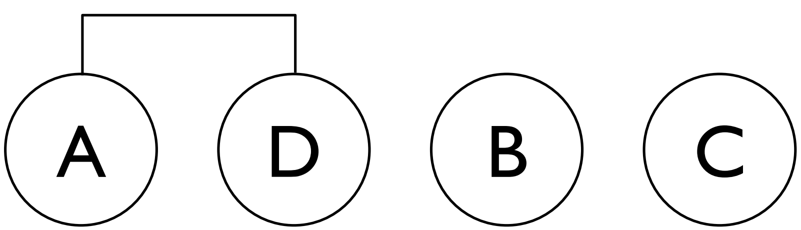


	A	B	C	D
A		20	7	2
B			10	25
C				3
D				

계층적 군집화

- 계층적 군집화 수행 예시
  - 모든 개체들 사이의 거리에 대한 유사도 행렬 계산
  - 거리가 인접한 관측치끼리 군집 형성
  - 유사도 행렬 업데이트

	AD	B	C	
AD		20	3	
B			10	
C				

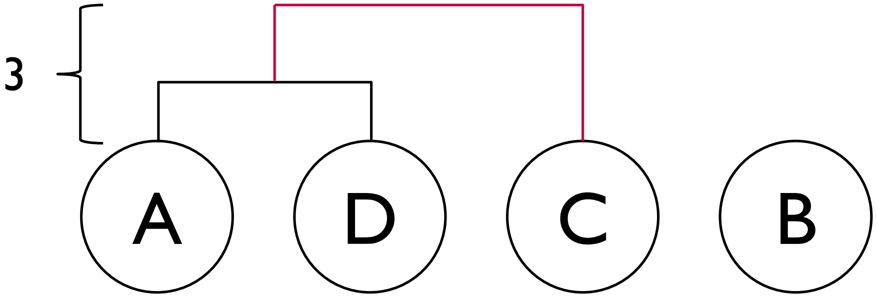


계층적 군집화

❖ 계층적 군집화 수행 예시

- 모든 개체들 사이의 거리에 대한 유사도 행렬 계산
- 거리가 인접한 관측치끼리 군집 형성
- 유사도 행렬 업데이트
- 위의 과정 반복

	AD	B	C	
AD		20	3	
B			10	
C				

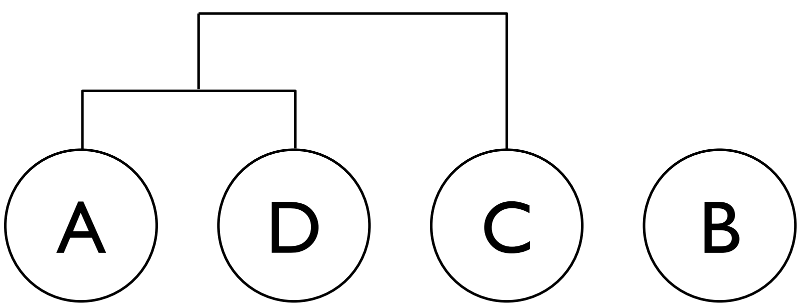


계층적 군집화

❖ 계층적 군집화 수행 예시

- 모든 개체들 사이의 거리에 대한 유사도 행렬 계산
- 거리가 인접한 관측치끼리 군집 형성
- 유사도 행렬 업데이트
- 위의 과정 반복

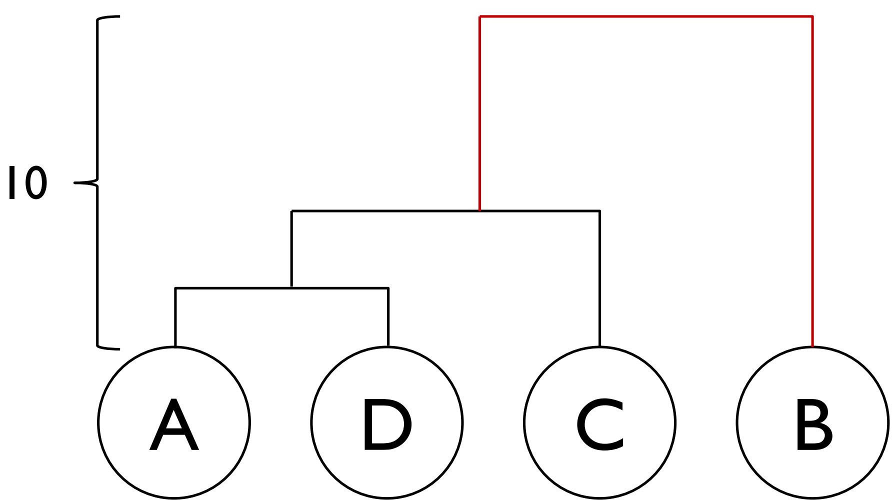
	AD C	$\boxed{B}$		
$\boxed{AD\ C}$		10		
B				



계층적 군집화

❖ 계층적 군집화 수행 예시

- 모든 개체들 사이의 거리에 대한 유사도 행렬 계산
- 거리가 인접한 관측치끼리 군집 형성
- 유사도 행렬 업데이트
- 위의 과정 반복

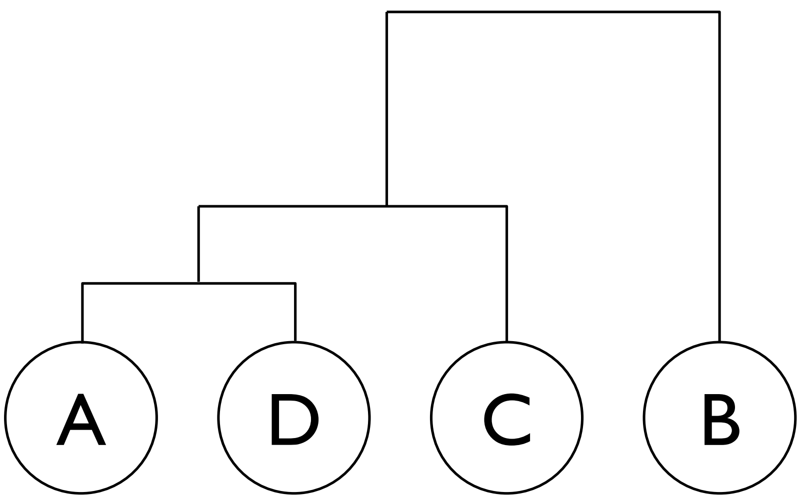


	ADC	B		
ADC		10		
B				

계층적 군집화

❖ 계층적 군집화 수행 예시

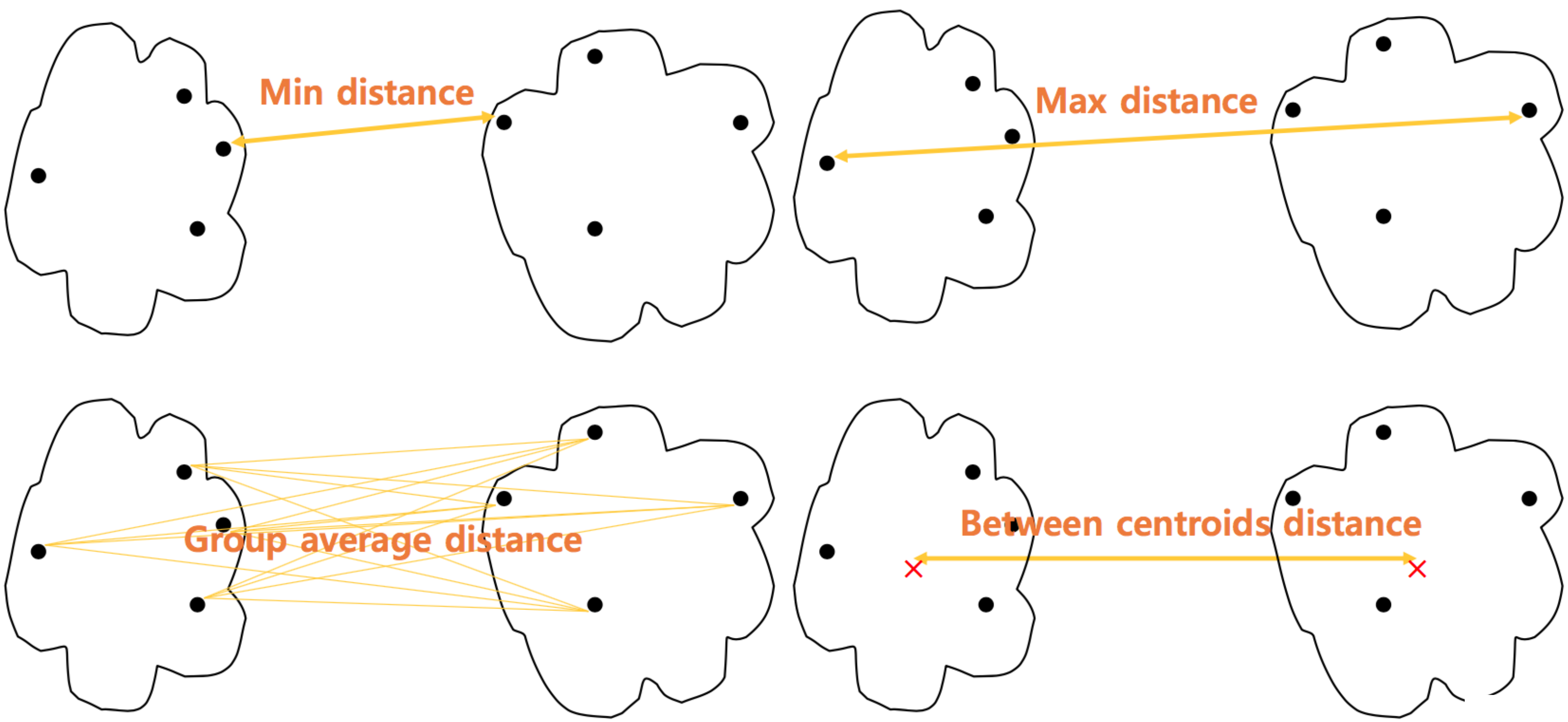
- 최종결과



	AD CB			
AD CB				

계층적 군집화

- 핵심 수행 절차: 두 군집 사이의 유사성/거리 측정
  - ✓ Min (단일연결), max (완전연결), group average (평균연결), between centroid, Ward's, …



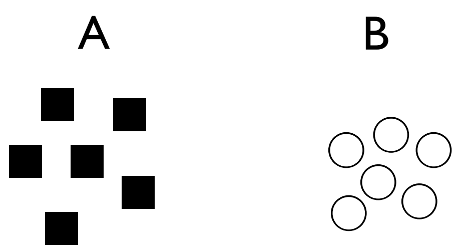
계층적 군집화

- Ward's method: Distance between two clusters, A and B, is how much the sum of squares will increase when they are merged.

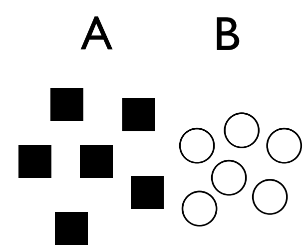
$$Ward\ Distance = \sum_{i \in A \cup B} \|x_i - m_{A \cup B}\|^2 - \left\{ \sum_{i \in A} \|x_i - m_A\|^2 + \sum_{i \in B} \|x_i - m_B\|^2 \right\}$$

m_A is the center of cluster A.

Ward's distance can be considered as the merging cost of combining the clusters A and B



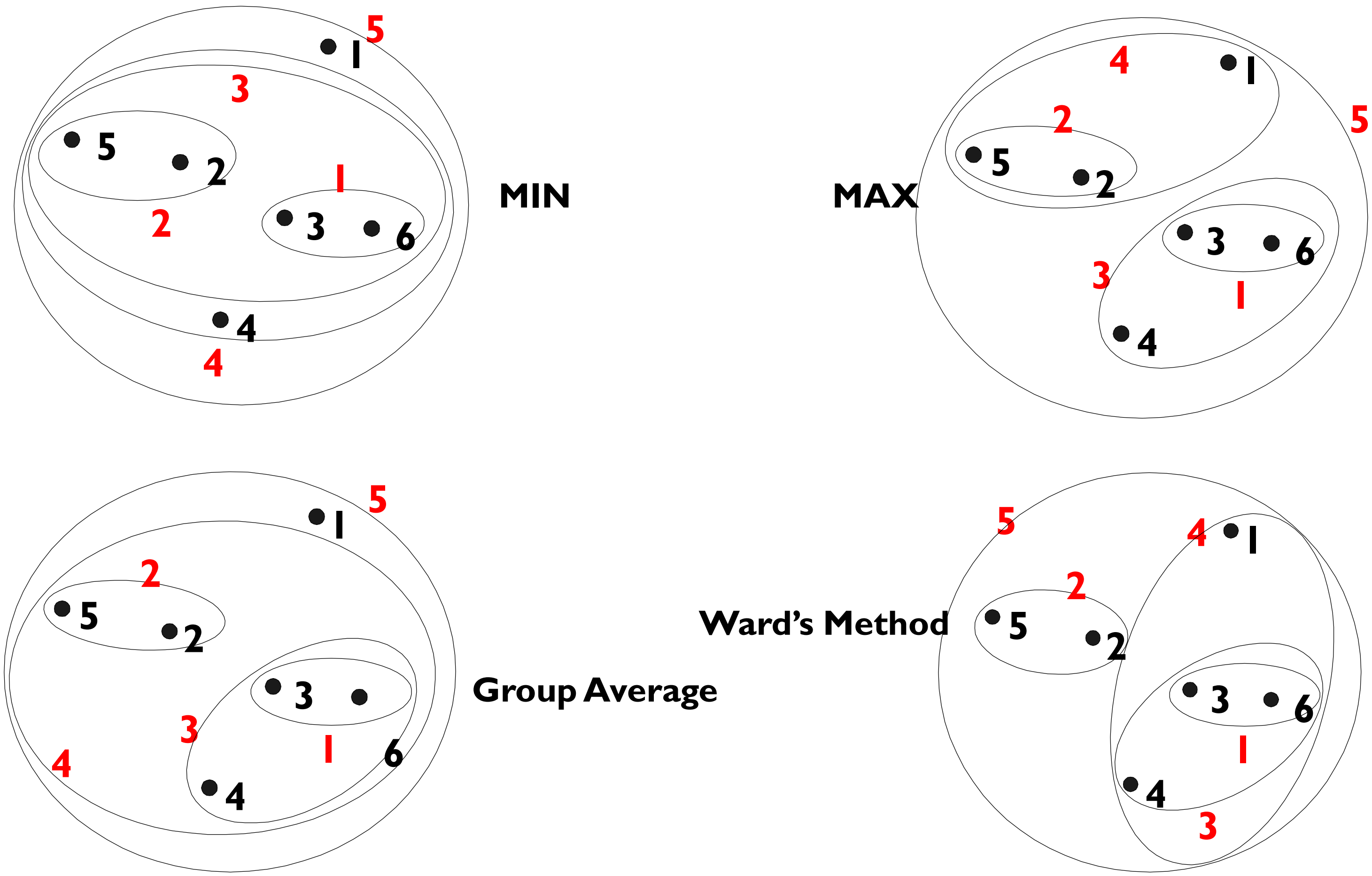
Ward's distance = 10 – (3+2) = 5



Ward's distance = 7 – (3+2) = 2

계층적 군집화

❖ 유사성/거리 행렬 계산 방식에 따른 결과 차이



K-평균 군집화 (K-Means Clustering)

❖ K-평균 군집화

- 대표적인 분리형 군집화 알고리즘
  - ✓ 각 군집은 하나의 중심(centroid)을 가짐
  - ✓ 각 개체는 가장 가까운 중심에 할당되며, 같은 중심에 할당된 개체들이 모여 하나의 군집을 형성
  - ✓ 사전에 군집의 수 K가 정해져야 알고리즘을 실행할 수 있음

$$X = C_1 \cup C_2 \cdots \cup C_k,\ C_i \cap C_j = \emptyset,\ i \neq j$$

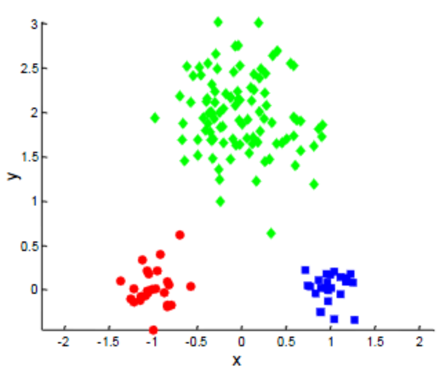
$$\underset{c}{\operatorname{argmin}} \sum_{i=1}^{K} \sum_{x_j \in C_i} \left\| x_j - c_i \right\|^2$$

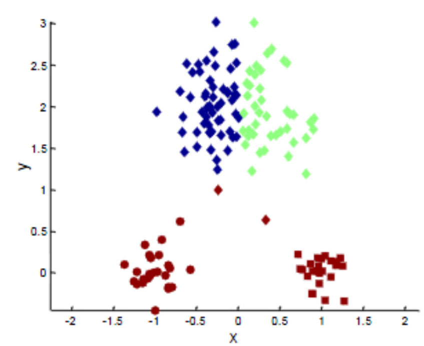
K-평균 군집화

❖ K-평균 군집화 수행 절차

1. 초기 중심을 K개 임의로 생성
2. 개별 관측치로부터 각 중심까지의 거리를 계산 후, 가장 가까운 중심이 이루는 군집에 관측치 할당
3. 각 군집의 중심을 다시 계산
4. 중심이 변하지 않을 때까지 **2, 3**의 과정을 반복

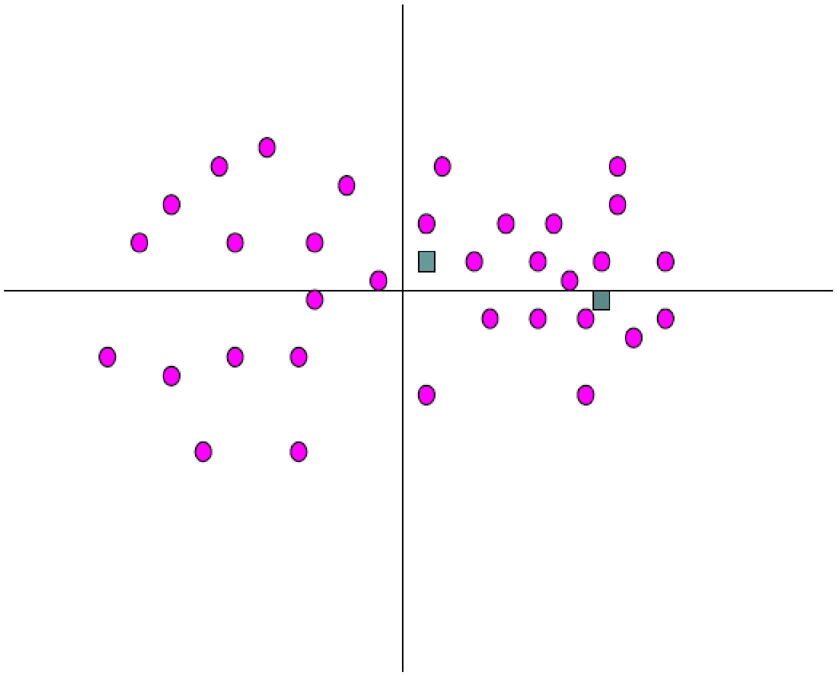
✓ 초기 중심은 종종 무작위로 설정됨: 군집화 결과가 초기 중심 설정에 따라 다르게 나타나는 경우가 발생할 수도 있음





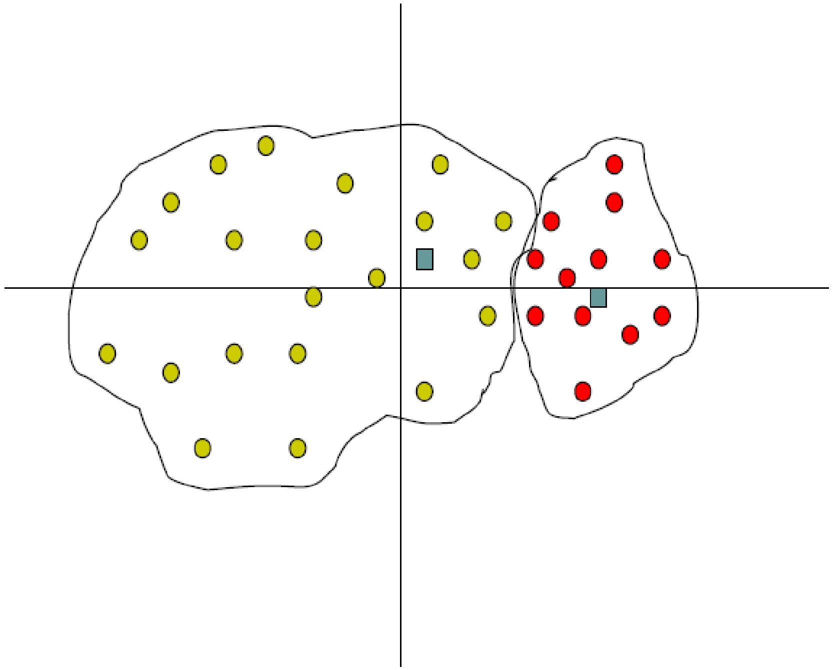
K-평균 군집화

- K-평균 군집화 수행 예시 (K=2)
  1. 2개의 중심을 임의로 생성



K-평균 군집화

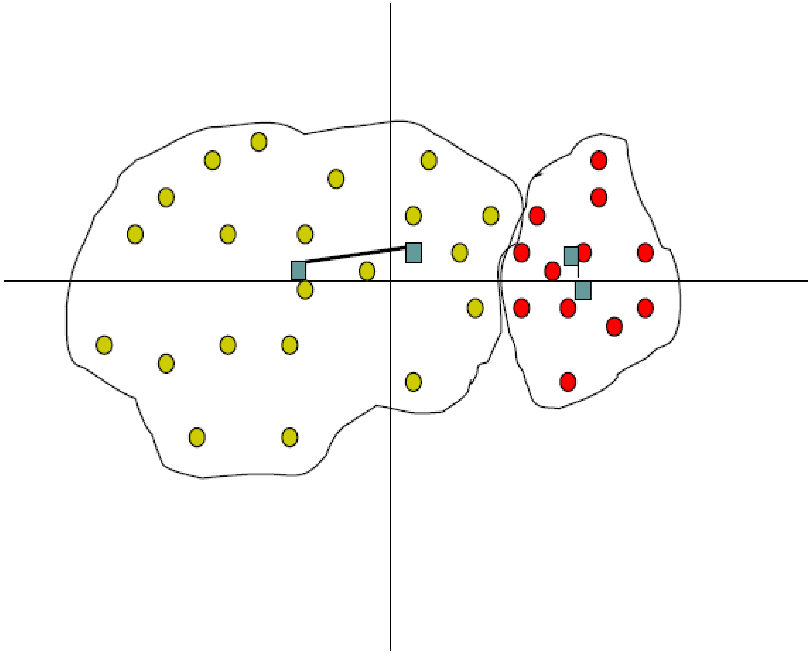
- K-평균 군집화 수행 예시 (K=2)
  1. 2개의 중심을 임의로 생성
  2. 생성된 중심을 기준으로 모든 관측치에 군집 할당



K-평균 군집화

❖ K-평균 군집화 수행 예시 (K=2)

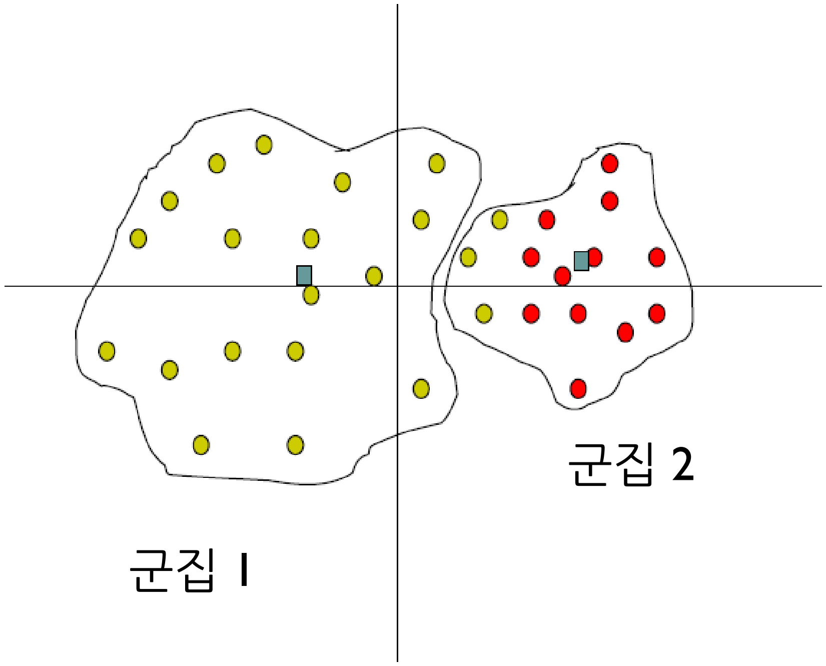
1. 2개의 중심을 임의로 생성
2. 생성된 중심을 기준으로 모든 관측치에 군집 할당
3. 각 군집의 중심을 다시 계산



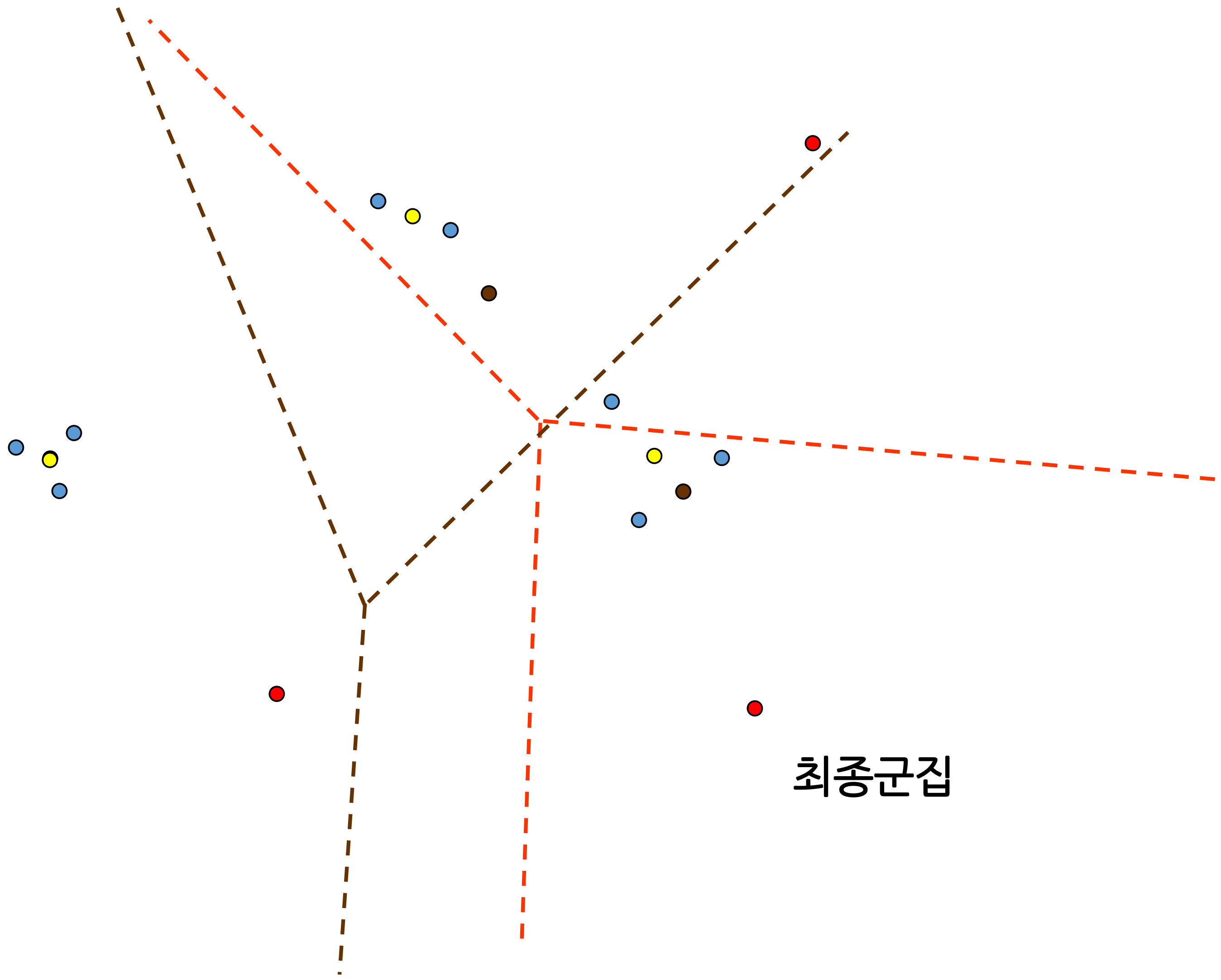
K-평균 군집화

- K-평균 군집화 수행 예시 (K=2)
  1. 2개의 중심을 임의로 생성
  2. 생성된 중심을 기준으로 모든 관측치에 군집 할당
  3. 각 군집의 중심을 다시 계산
  4. 중심이 변하지 않을 때까지

  위의 과정을 반복



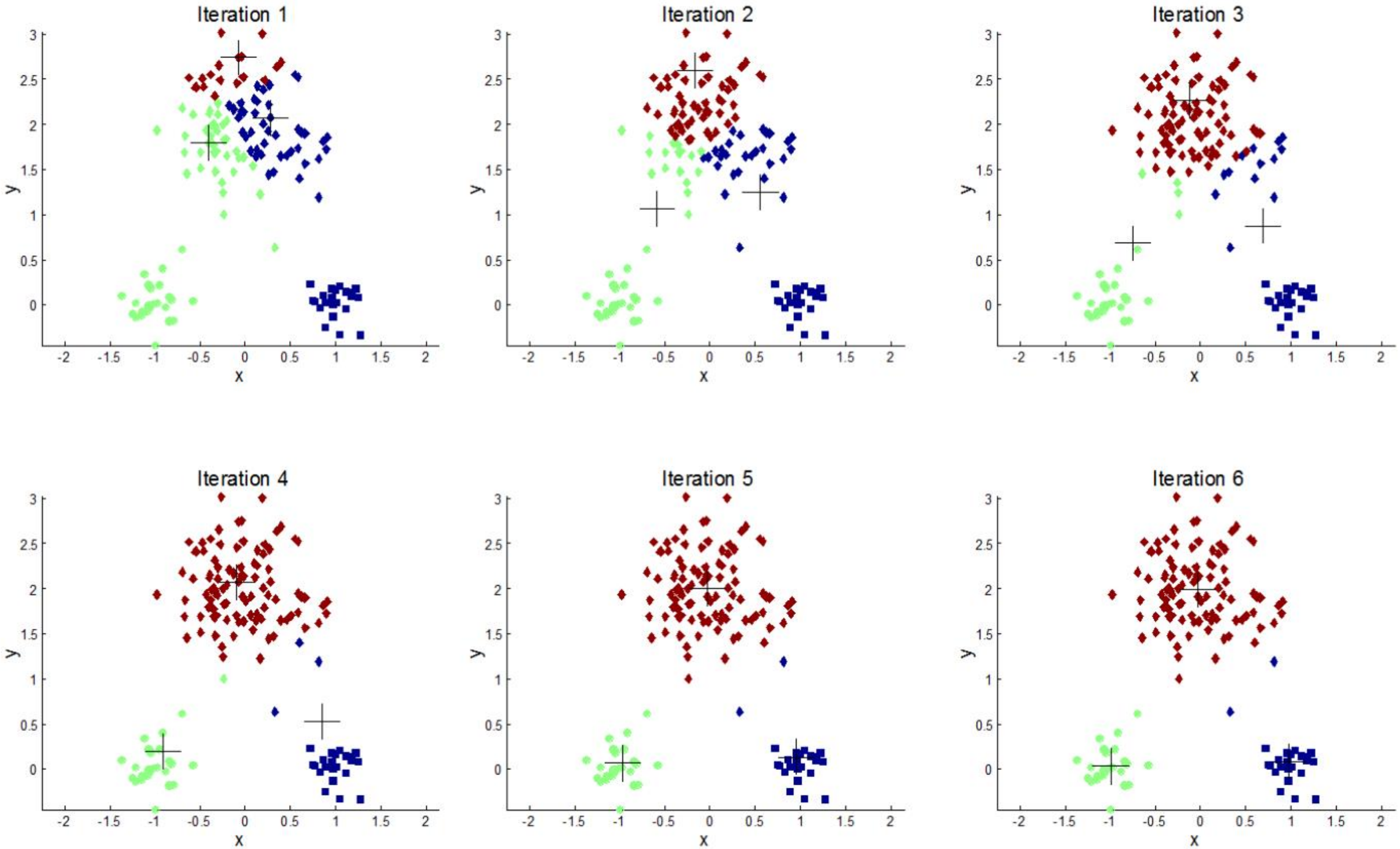
K-평균 군집화



K-평균 군집화

❖ 초기 중심 설정이 최종 결과에 어떤 영향을 미치는가?

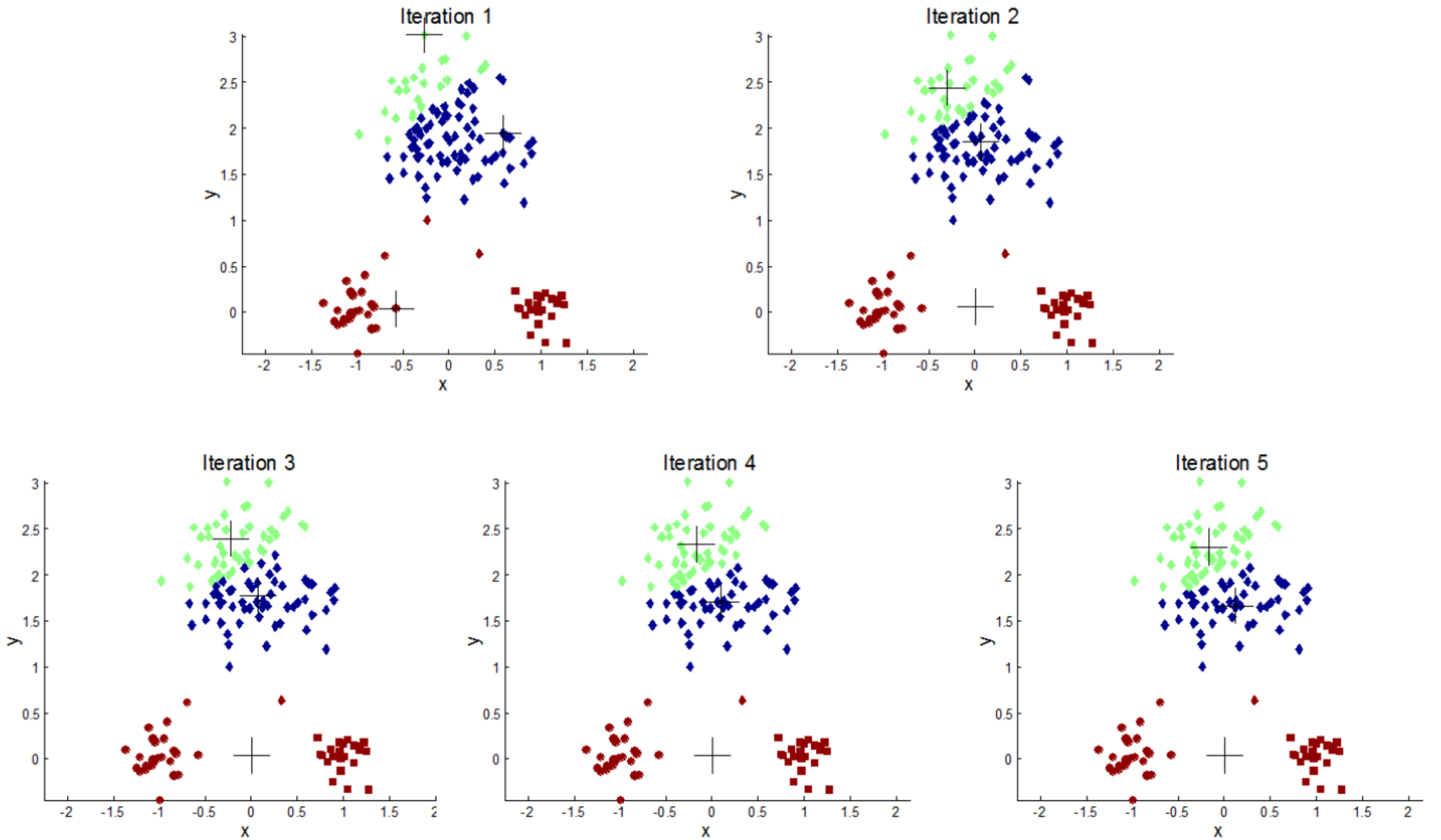
- 바람직한 결과



K-평균 군집화

❖ 초기 중심 설정이 최종 결과에 어떤 영향을 미치는가?

- 바람직하지 않은 결과



K-평균 군집화

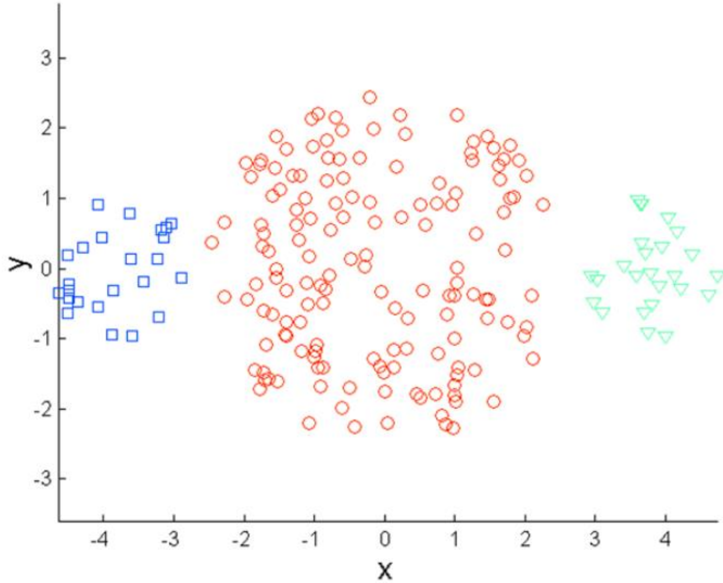
- 무작위 초기 중심 설정의 위험을 피하고자 다양한 연구 존재
  - 반복적으로 수행하여 가장 여러 번 나타나는 군집을 사용
  - 전체 데이터 중 일부만 샘플링하여 계층적 군집화를 수행한 뒤 초기 군집 중심 설정
  - 데이터 분포의 정보를 사용하여 초기 중심 설정
  - 하지만 많은 경우 초기 중심 설정이 최종 결과에 큰 영향을 미치지는 않음

K-평균 군집화

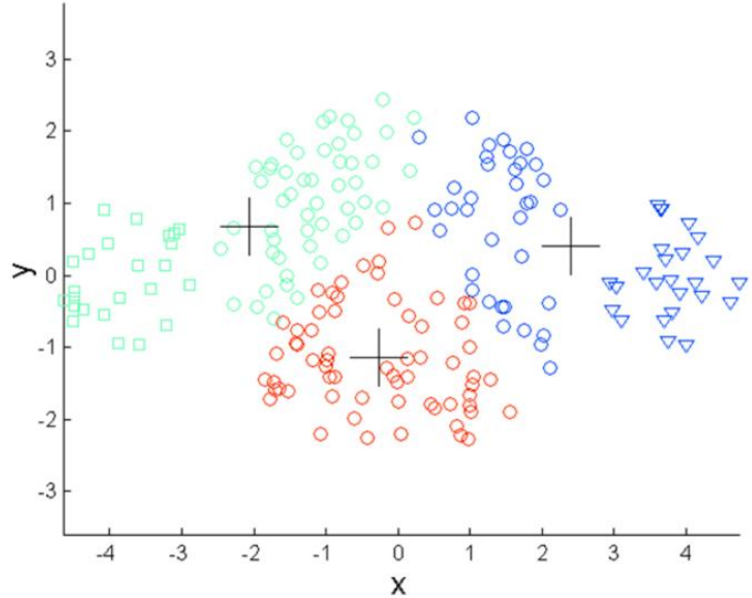
❖ K-평균 군집화의 문제점

- 문제점 I : **서로 다른 크기**의 군집을 잘 찾아내지 못함

정답



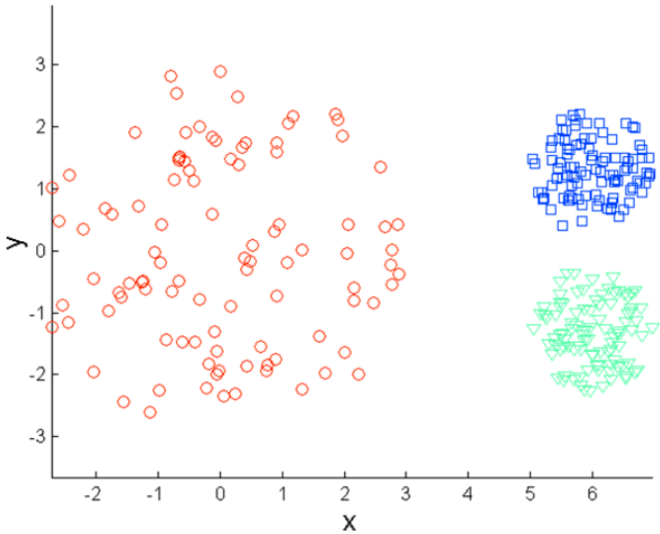
K-평균 군집화 결과



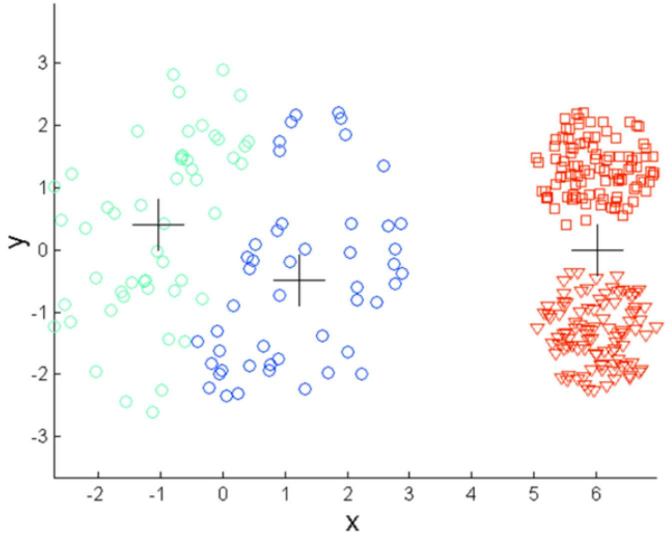
K-평균 군집화

- K-평균 군집화의 문제점
  - 문제점2: **서로 다른 밀도**의 군집을 잘 찾아내지 못함

정답



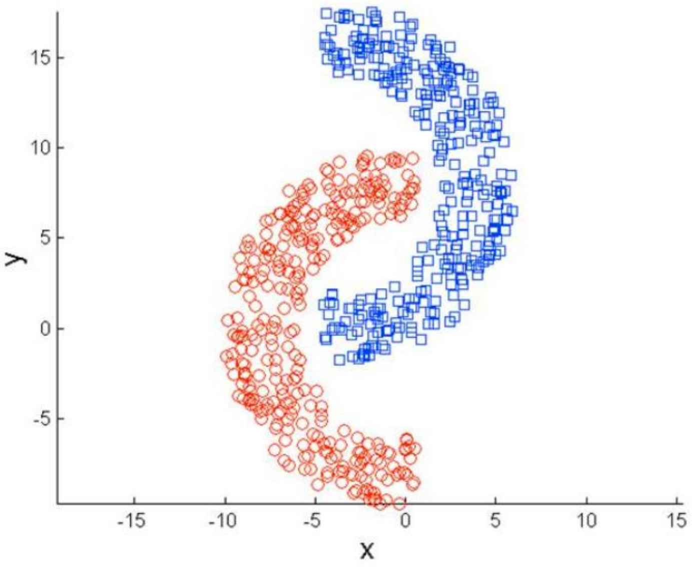
K-평균 군집화 결과



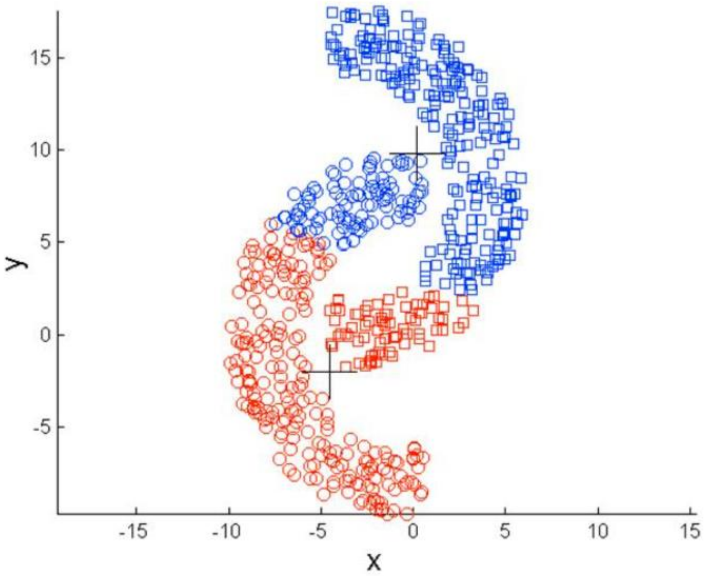
K-평균 군집화

- K-평균 군집화의 문제점
  - 문제점3: **지역적 패턴이 존재**하는 군집을 판별하기 어려움

정답



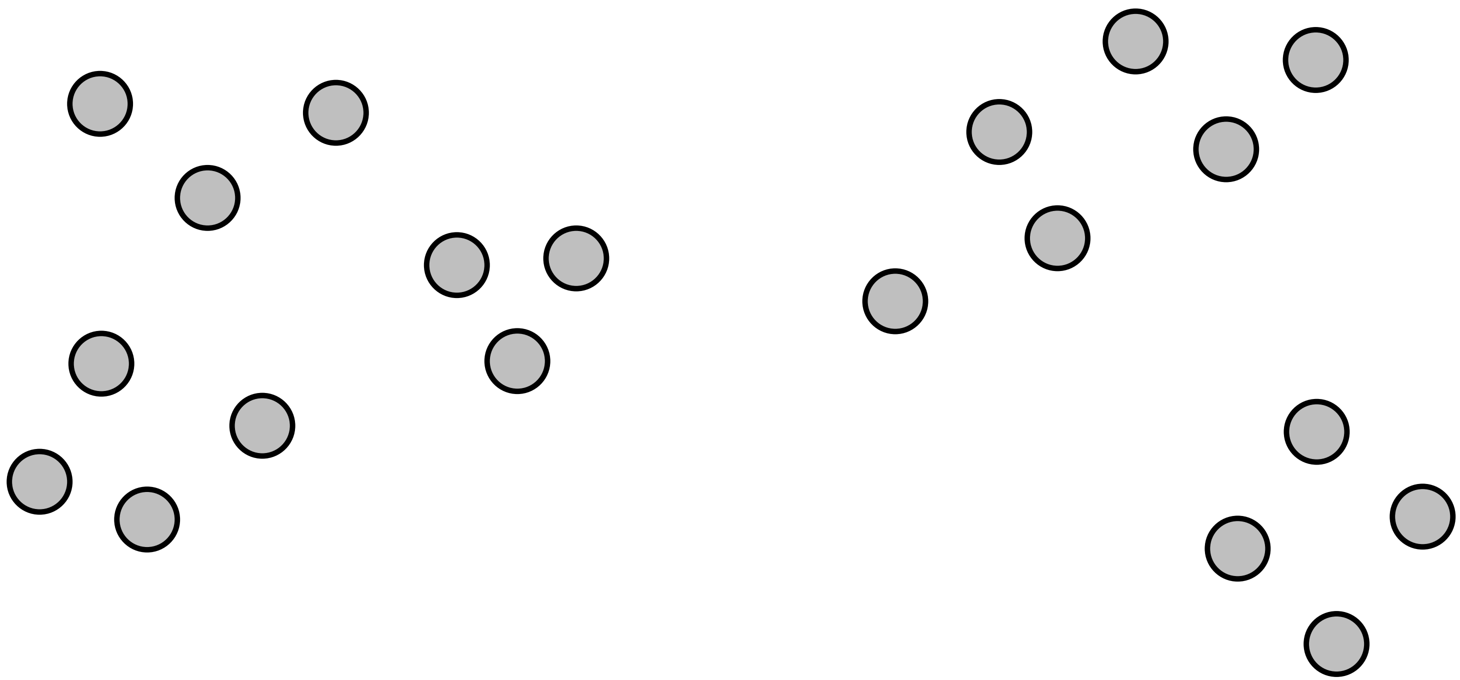
K-평균 군집화 결과



군집화: 최적의 군집 수 결정

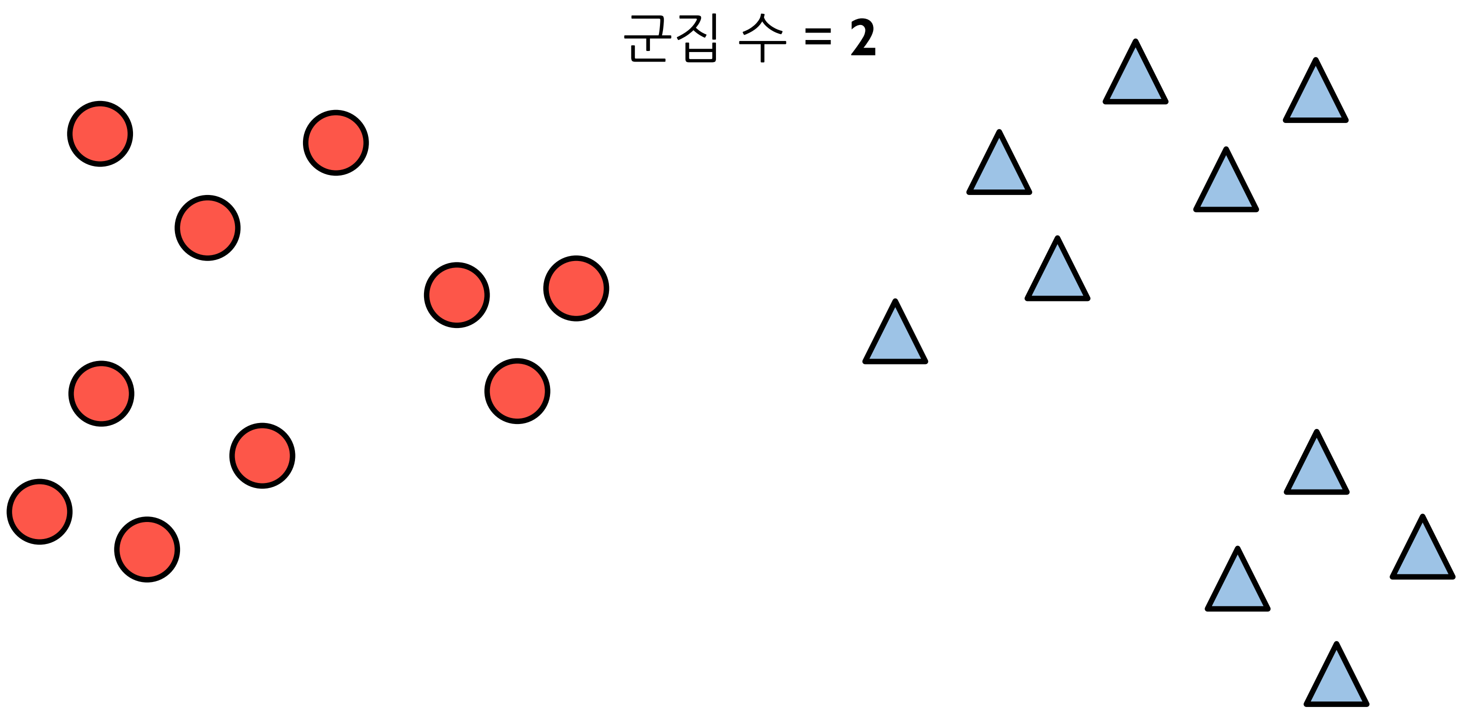
❖ 어떻게 최적의 군집 수를 결정할 것인가?

- 예시) 20개의 관측치가 존재할 때, 최적의 군집 수는?



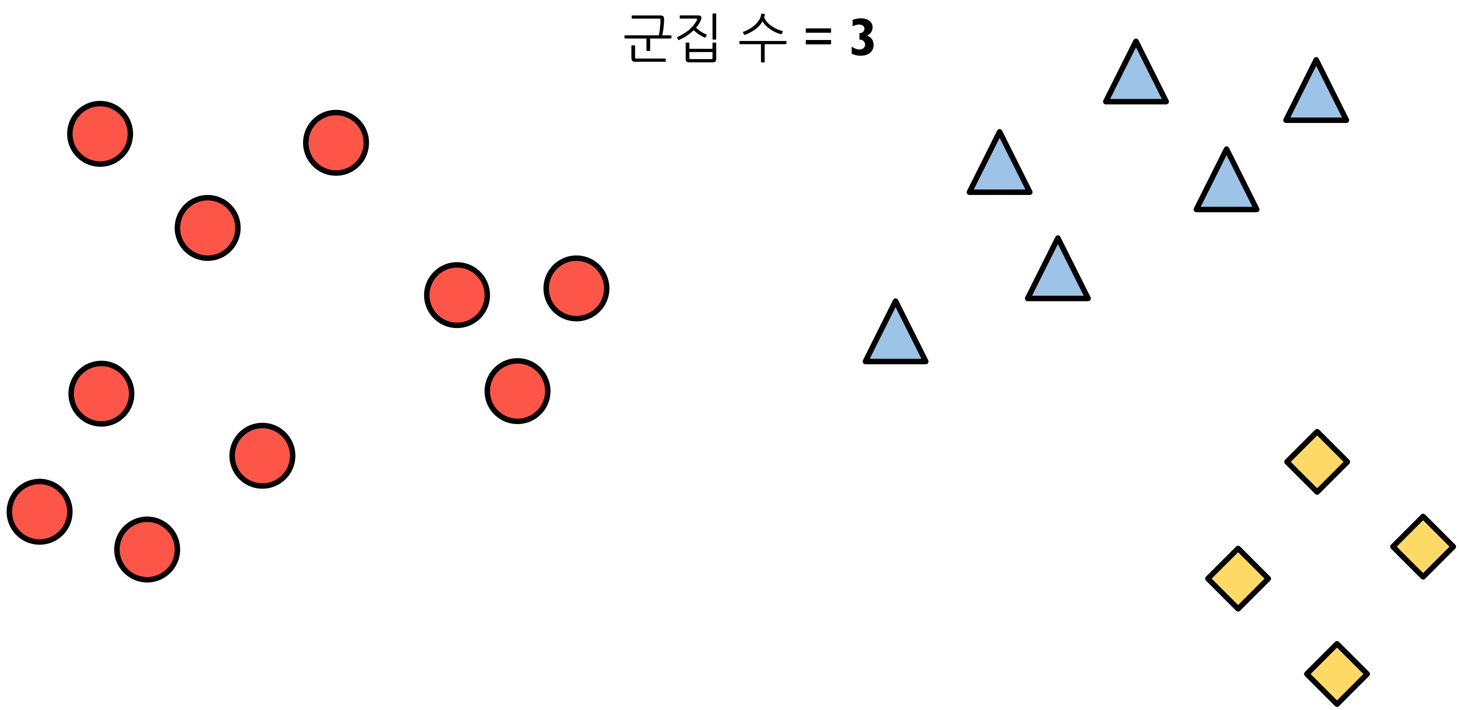
군집화: 최적의 군집 수 결정

- 어떻게 최적의 군집 수를 결정할 것인가?
  - 예시) 20개의 관측치가 존재할 때, 최적의 군집 수는?



군집화: 최적의 군집 수 결정

- 어떻게 최적의 군집 수를 결정할 것인가?
  - 예시) 20개의 관측치가 존재할 때, 최적의 군집 수는?

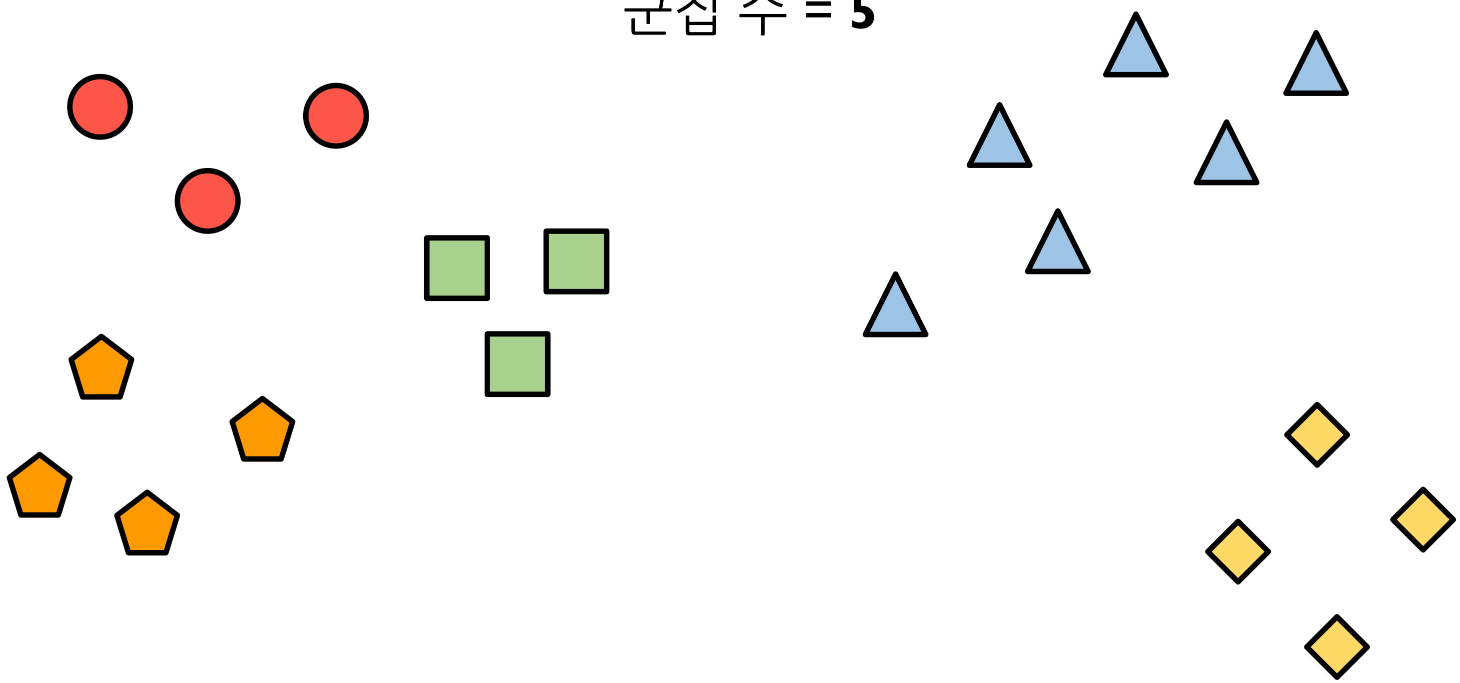


군집화: 최적의 군집 수 결정

❖ 어떻게 최적의 군집 수를 결정할 것인가?

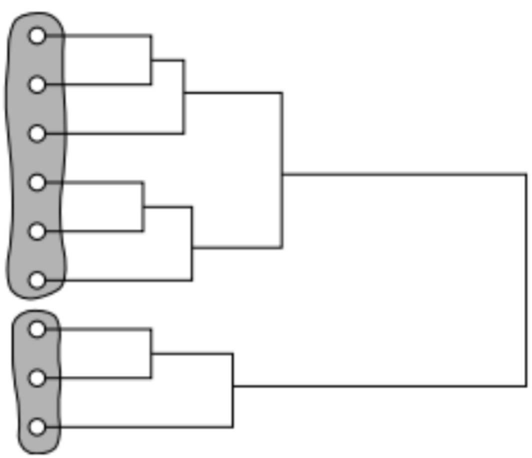
- 예시) 20개의 관측치가 존재할 때, 최적의 군집 수는?

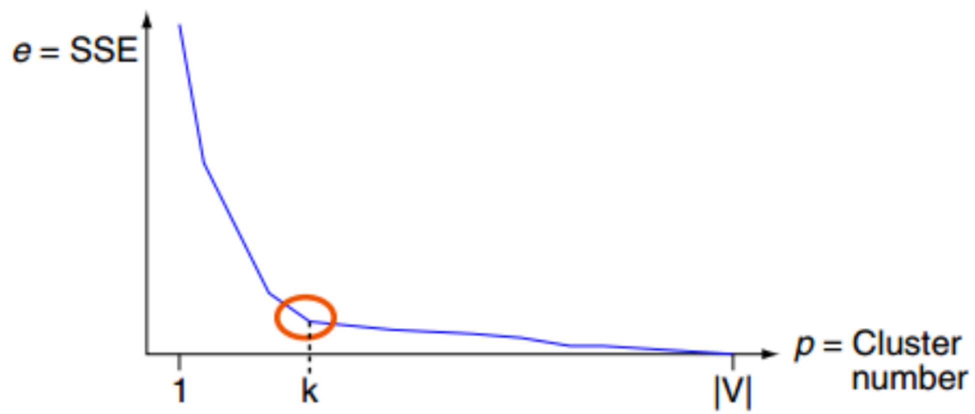
군집 수 = **5**



군집화: 최적의 군집 수 결정

- 어떻게 최적의 군집 수를 결정할 것인가?
  - 다양한 군집 수에 대해 성능 평가 지표를 도시하여 최적의 군집 수 선택
  - Elbow point에서 최적 군집 수가 결정되는 경우가 일반적





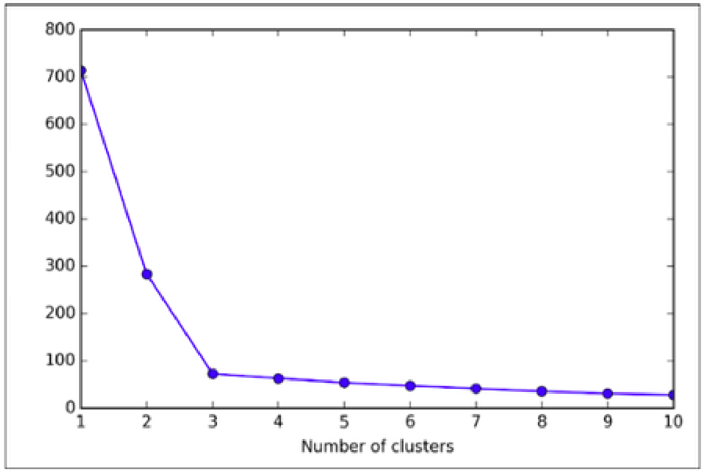
군집화: 결과 측정 및 평가

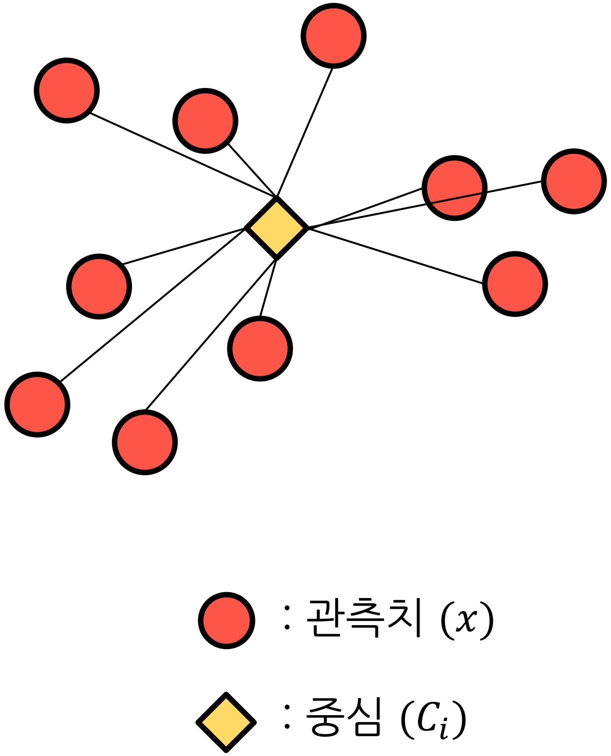
- 어떻게 군집화 결과를 측정/평가할 것인가?
- 분류 알고리즘처럼 모든 상황에 적용가능한 평가 지표 부재
  - 내부 평가 지표
    - ✓ Dunn Index, Silhouette, Sum of Squared Error, …
  - 외부 평가 지표
    - ✓ Rand Index, Jaccard Coefficient, Folks and Mallows Index, …

군집화: 결과 측정 및 평가

❖ 군집화 평가 지표 1: Sum of Squared Error (SSE)

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist(x, c_i)^2$$





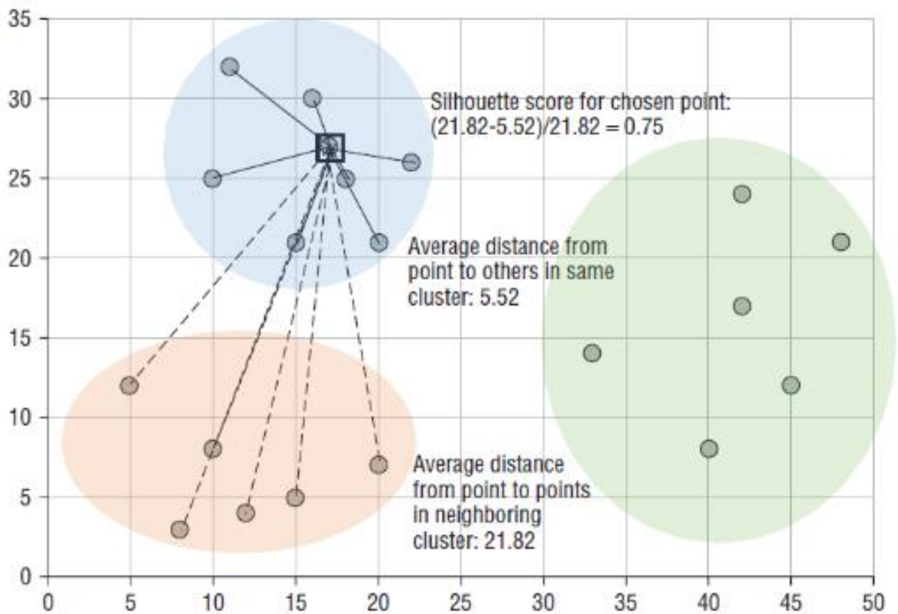
군집화: 결과 측정 및 평가

- 군집화 평가 지표 2: Silhouette 통계량
  - *a*(*i*): 관측치*i*로부터 같은 군집 내에 있는 모든 다른 개체들 사이의 평균 거리
  - *b*(*i*): 관측치*i*로부터 다른 군집 내에 있는 개체들 사이의 평균 거리 중 최솟값
  - 일반적으로$\bar{S}$의 값 0.5보다 크면 군집 결과가 타당하다고 볼 수 있음
  - -1에 가까우면 군집이 전혀 되지 않음

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

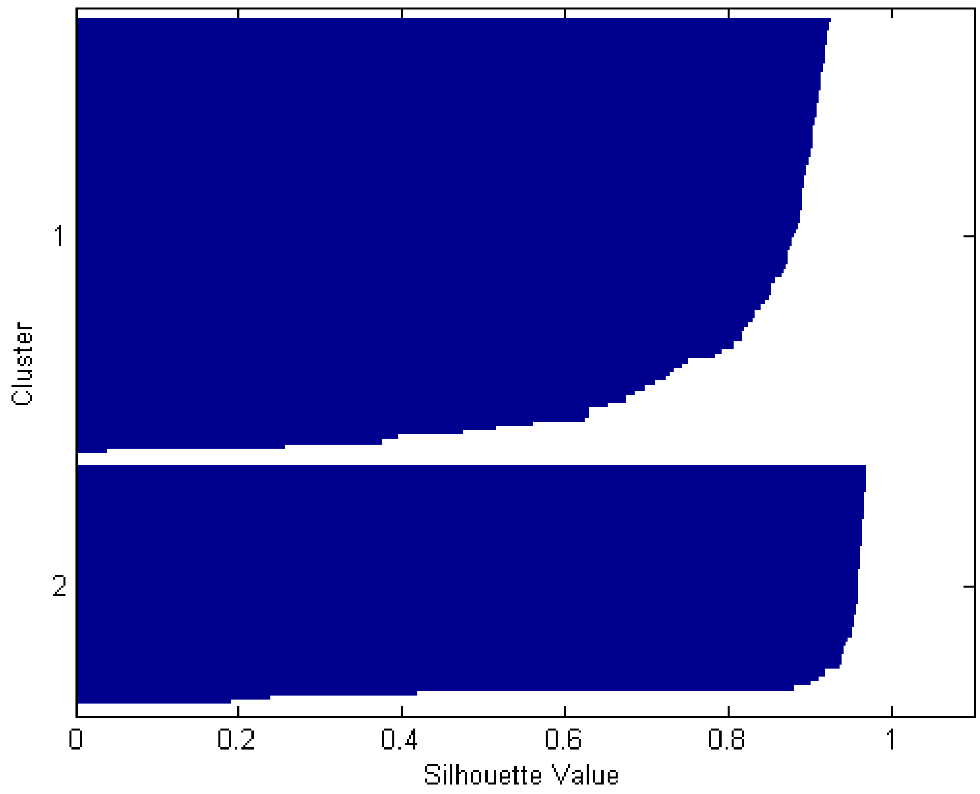
$$-1 \le s(i) \le 1$$

$$\bar{S} = \frac{1}{n}\sum_{i=1}^{n} S(i)$$



군집화: 결과 측정 및 평가

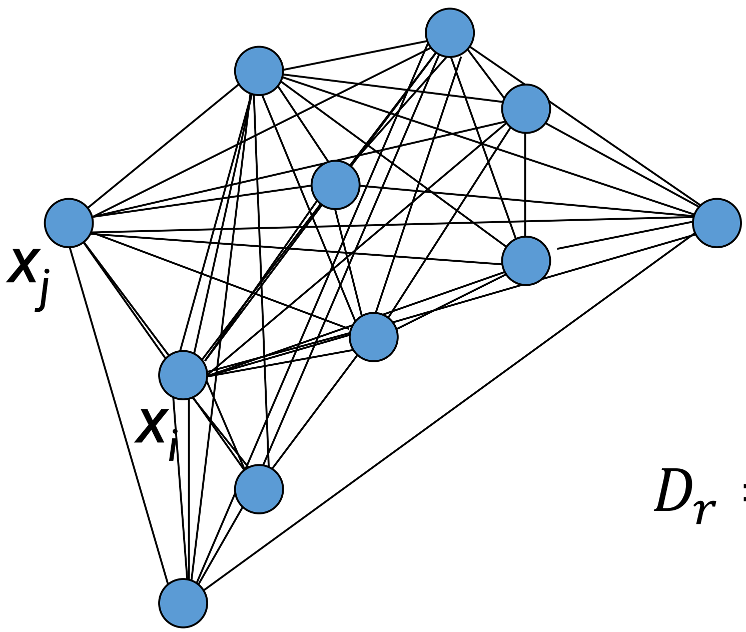
$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$



군집화: 결과 측정 및 평가

❖ 군집화 평가 지표 3: Gap 통계량

Within-Cluster Sum of Squares



$$D_r = \sum_{i \in C_r} \sum_{i \in C_r} \left\| x_i - x_j \right\|^2$$

군집화: 결과 측정 및 평가

- 군집화 평가 지표 3: Gap 통계량

Sum of the pairwise distances for all points in cluster *r*.

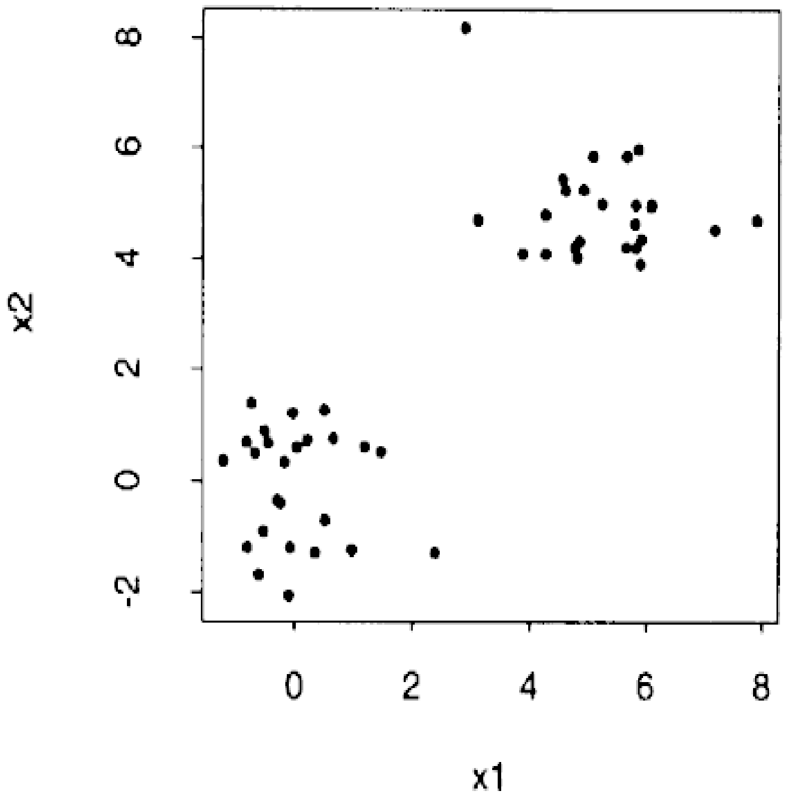
$$D_r = \sum_{i \in C_r} \sum_{i \in C_r} \left\| x_i - x_j \right\|^2$$

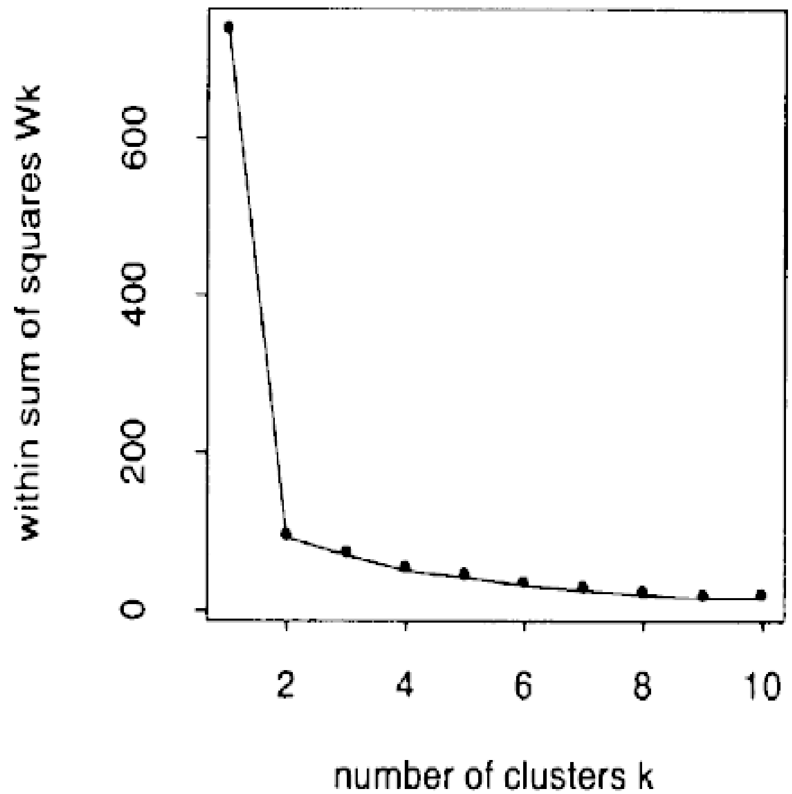
Define W_k

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r$$

군집화: 결과 측정 및 평가

❖ 군집화 평가 지표 3: Gap 통계량





군집화: 결과 측정 및 평가

❖ 군집화 평가 지표 3: Gap 통계량

Problem w/ using the L-Curve method:

– no reference clustering to compare

Gap Statistic:

– Cluster the data to obtain partitions (k=1,2,…,K) using any desired clustering method.
– For each partition with k clusters, calculate $\log W_k$
– Generate the reference distribution (no clusters)
– Calculate W^*_k from the reference distribution
– $\text{Gap}(k) = E(\log W^*_k) - \log W_k$
– Find the k that maximizes Gap(k) (within some tolerance)

군집화: 결과 측정 및 평가

❖ 군집화 평가 지표 3: Gap 통계량

- Gap-Uniform: For each of the *i* variable, generate *n* observations that are uniformly distributed over the range x_i^{min} to x_i^{max}, where x_i represents *i*th variable of X.

- Gap-PC: Singular value decomposition technique.

$$X = UDV^T$$
$$X' = XV$$

Generate a matrix of random variates Z' as in the gap-uniform case using the range of the columns of X' instead.

$$Z = Z'V^T$$

군집화: 결과 측정 및 평가

- 군집화 평가 지표 3: Gap 통계량

$$Gap(k) = \frac{1}{B}\sum_{b=1}^{B} \log W^*{}_{kb} - \log W_k$$

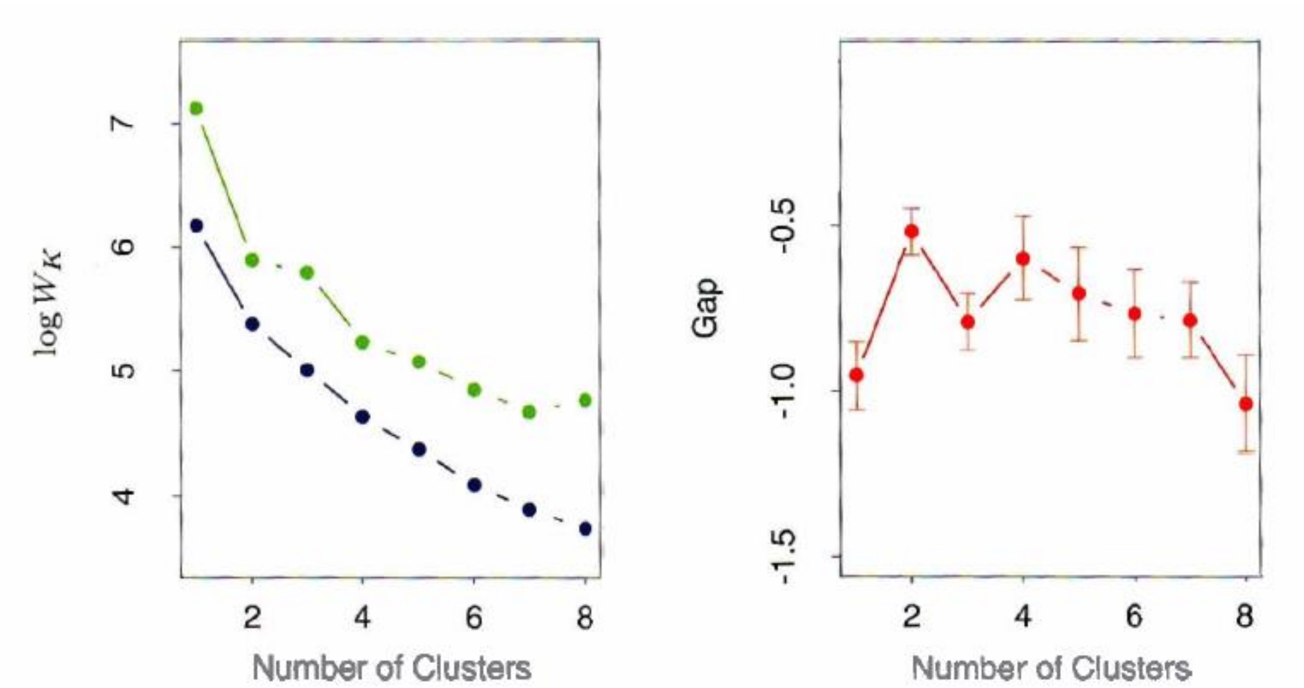
No specific guideline for what value of *B* to use, but *B* > 10.

$$\bar{W}_k = \frac{1}{B}\sum_{b} \log(W^*{}_{kb})$$

$$sd_k = \sqrt{\frac{1}{B}\sum_{b} [\log(W^*{}_{kb}) - \bar{W}_k]^2}$$

군집화: 결과 측정 및 평가

❖ 군집화 평가 지표 3: Gap 통계량



Green: Observed values of log W_k.

Blue: Expected values of log W_k from the reference distribution.

군집화: 결과 측정 및 평가

❖ 군집화 평가 지표 3: Gap 통계량 - 2 Cluster 예제

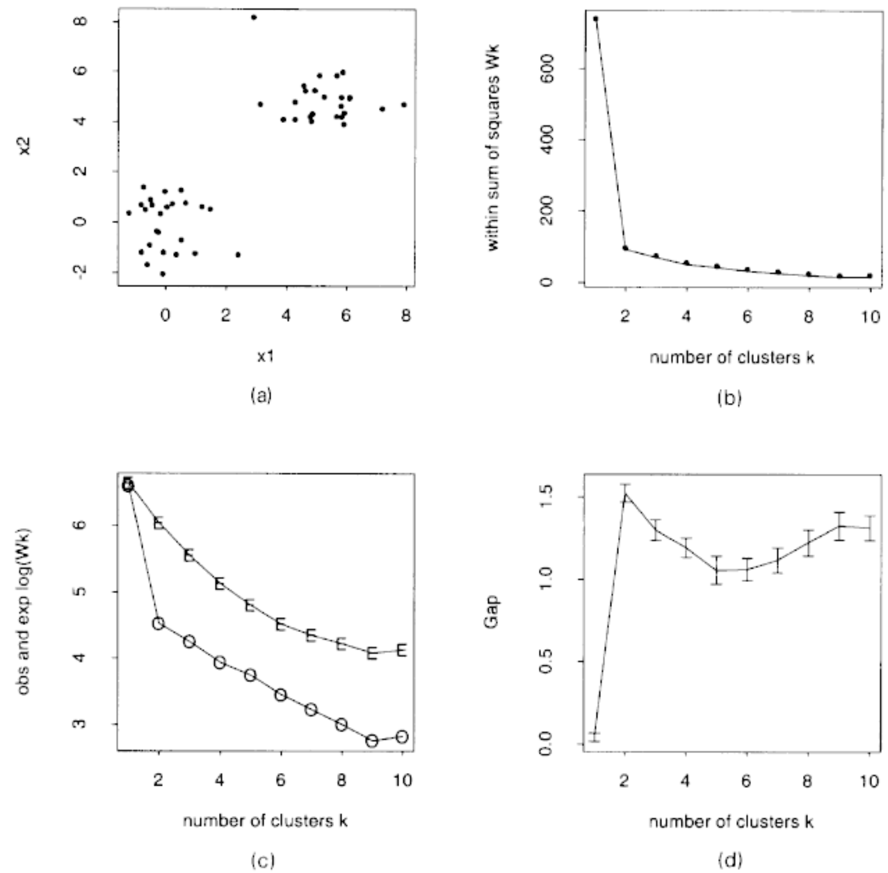


Fig. 1. Results for the two-cluster example: (a) data; (b) within sum of squares function W_k; (c) functions $\log(W_k)$ (O) and $\hat{E}^*_n\{\log(W_k)\}$ (E); (d) gap curve

EOD