정규 교육 세미나

ToBig's 11기 정혜인

Clustering

클러스터링

Contents

Unit 01 | Clustering

Unit 02 | Hierarchical Clustering

Unit 03 | K-Means Clustering

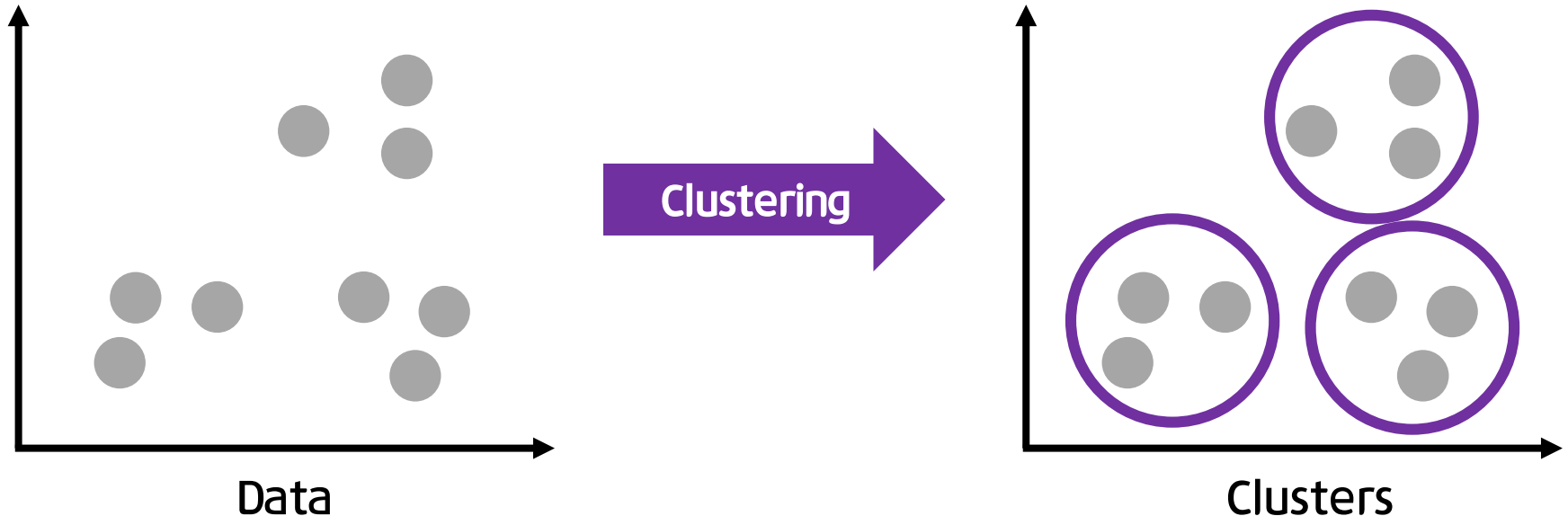
Unit 04 | DBSCAN

Unit 05 | 모델 평가

Unit 01 | Clustering

Clustering (군집화)

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



Unit 01 | Clustering

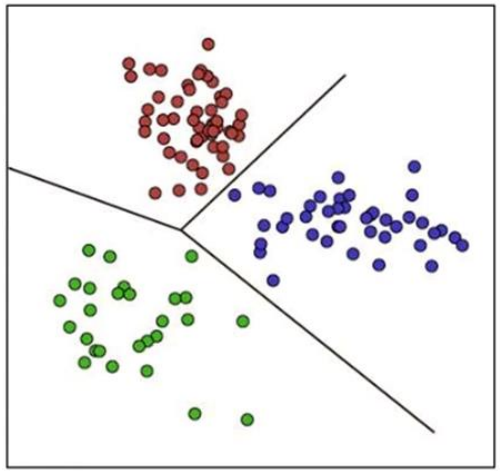
Classification vs. Clustering

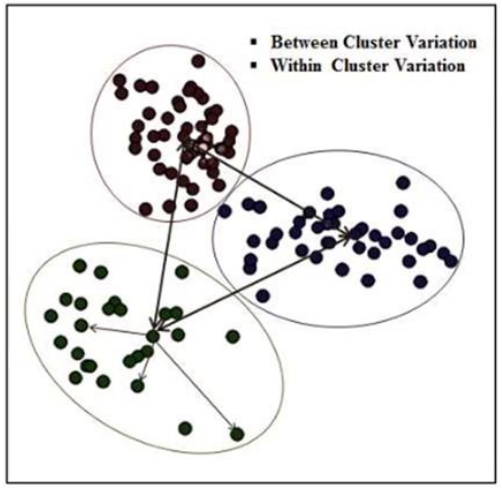
Classification : **Supervised** Learning

- 소속 집단의 정보를 이미 알고 있는 상태에서 비슷한 집단으로 묶는 방법
- Label이 있는 data를 나누는 방법

Clustering : **Unsupervised** Learning

- 소속집단의 정보가 없고, 모르는 상태에서 비슷한 집단으로 묶는 방법
- Label이 없는 Data를 군집단위로 나누는 것

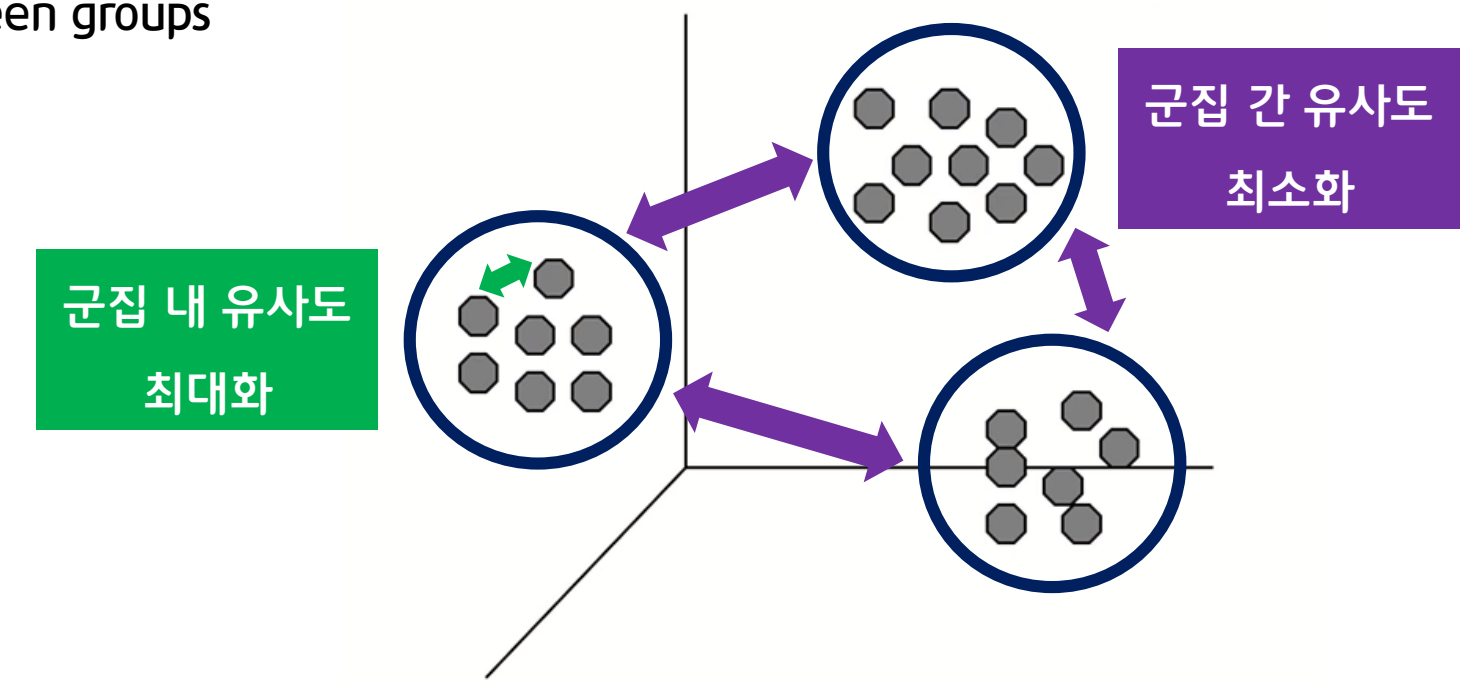




Unit 01 | Clustering

좋은 Clustering?

1. Maximizes the similarity within a group
2. Maximizes the difference between groups

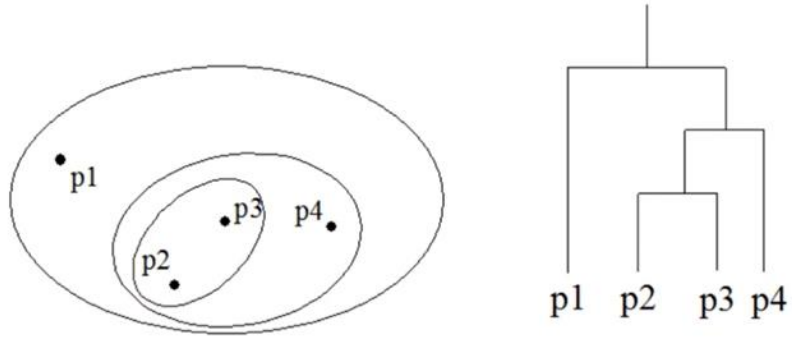


Unit 01 | Clustering

Clustering 방법

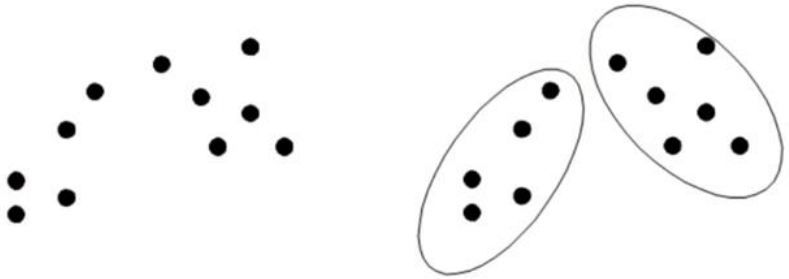
1. Hierarchical agglomerative clustering (계층적 군집화)

- 개체들을 가까운 집단부터 차근차근 묶어나가는 방식
- 유사한 개체들이 결합되는 dendrogram 생성
- Clusters have subclusters, as if in a tree



2. Partitioning clustering (분리형 군집화)

- 전체 데이터의 영역을 특정 기준에 의해 동시에 구분
- 각 개체들은 사전에 정의된 개수의 군집 중 하나에 속하게 됨
- Clusters are non-overlapping and each object is in exactly one cluster

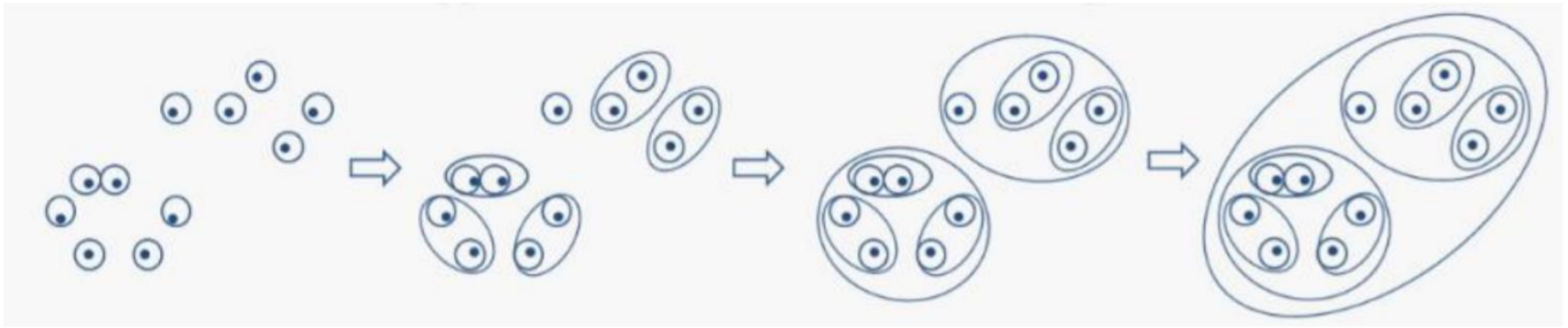


Unit 02 | Agglomerative Hierarchical Clustering

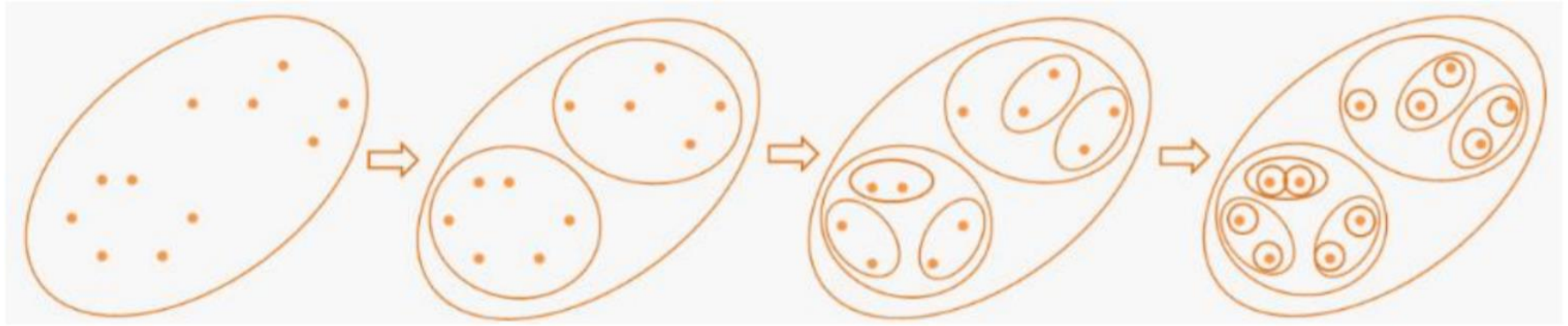
Hierarchical clustering

Agglomerative

(most common)



Divisive



Unit 02 | Agglomerative Hierarchical Clustering

Agglomerative Hierarchical Clustering

① 계층적 트리모형을 이용하여 개별 개체들을 순차적/계층적으로 유사한 개체/군집과 통합

② 덴드로그램(Dendrogram)을 통해 시각화 가능

- 덴드로그램 : 개체들이 결합되는 순서를 나타내는 트리형태의 구조

③ 사전에 군집의 개수를 정하지 않아도 수행 가능

- 덴드로그램 생성 후 적절한 수준에서 자르면 그에 해당하는 군집화 결과 생성

Agglomerative Hierarchical clustering

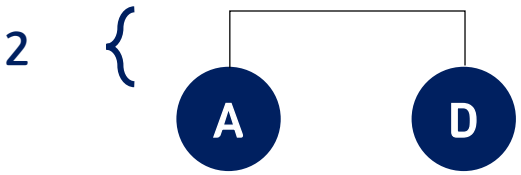
Algorithm 8.3 Basic agglomerative hierarchical clustering algorithm.

1: Compute the proximity matrix, if necessary.
2: **repeat**
3: Merge the closest two clusters.
4: Update the proximity matrix to reflect the proximity between the new cluster and the original clusters.
5: **until** Only one cluster remains.

Unit 02 | Agglomerative Hierarchical Clustering

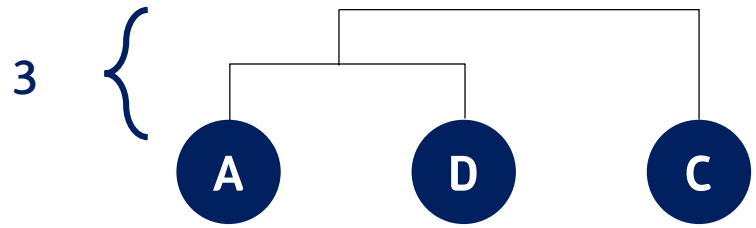
- 모든 개체들 사이의 거리에 대한 유사도 행렬 계산
- 거리가 인접한 관측치끼리 cluster 형성
- 유사도 행렬 update

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



Unit 02 | Agglomerative Hierarchical Clustering

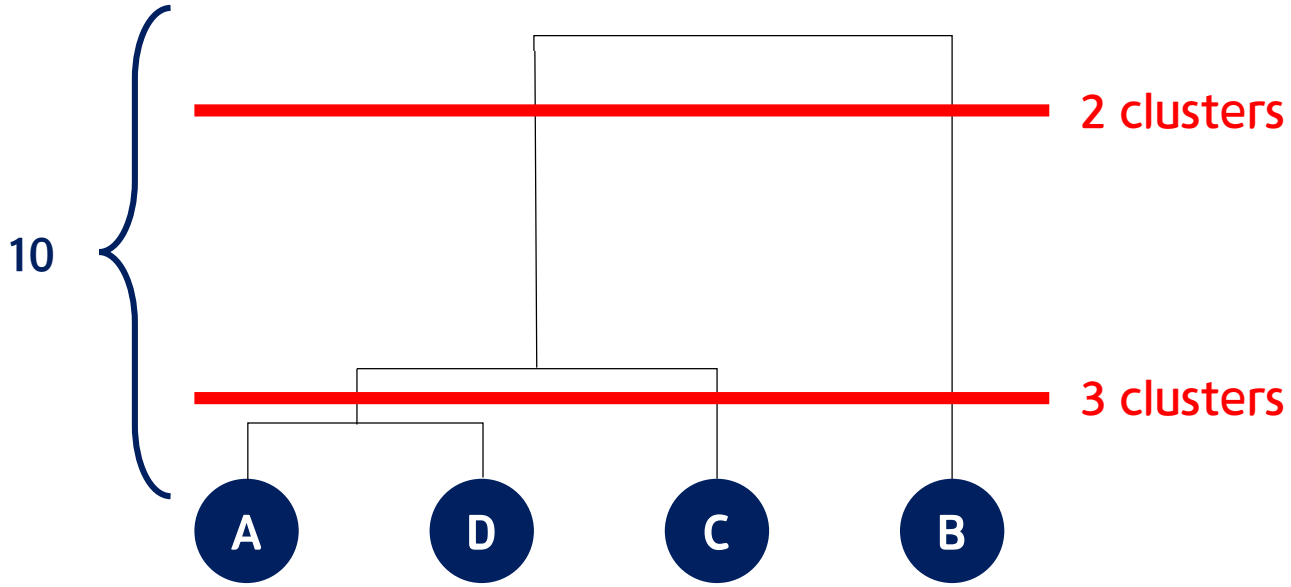
- 모든 개체들 사이의 거리에 대한 유사도 행렬 계산
- 거리가 인접한 관측치끼리 cluster 형성
- 유사도 행렬 update



	AD	B	C	
AD		20	3	
B			10	
C				

Unit 02 | Agglomerative Hierarchical Clustering

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



	ADC	B		
ADC		10		
B				

단일 data 간의 거리

For points in **Euclidean space** : (e.g., x = (3.2, 1.8, 5), y = (2.1, 3.5, 4))

- has some number of real-valued dimensions and "dense" points
- There is a notion of "average" of two points
- based on the **locations** of points in such a space

• Euclidean (L2) distance, Manhattan (L1) distance, ···

For points in **Non-Euclidean space** : **documents** (e.g., x = (3, 2, 0, 0, ···, 5, 0), y = (1, 1, 0, 4, ···, 0, 2))

- based on **properties** of points, but not their "location" in a space

• Cosine distance, Jaccard distance, ···

Cluster 간의 거리

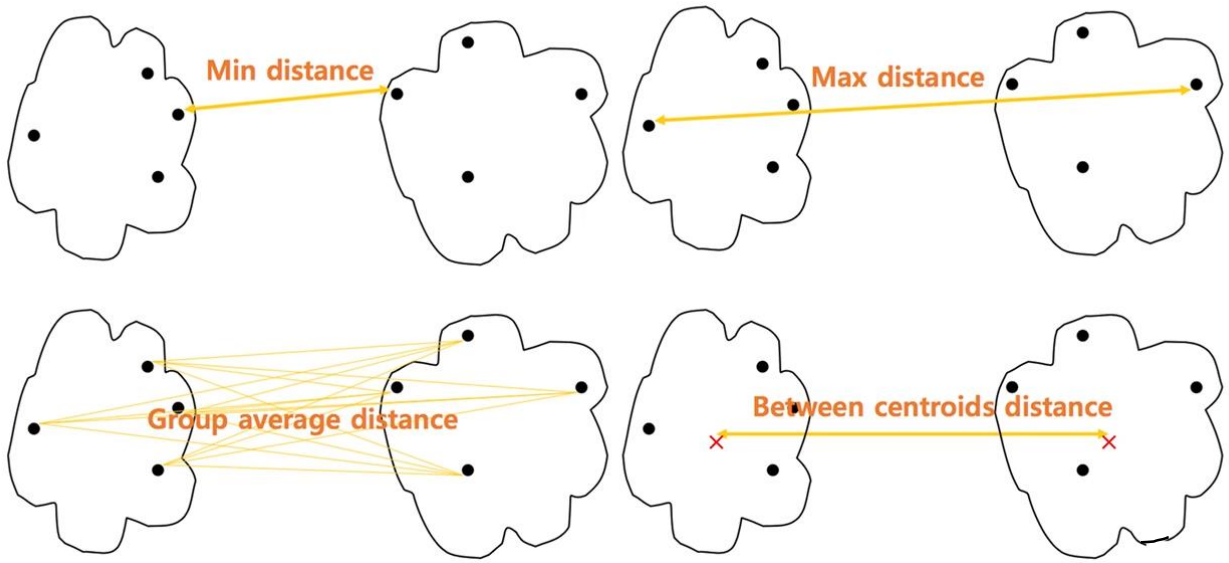
MIN (Single Link) : 두 군집에서 가장 가까운 멤버들의 거리를 잰다. 긴 chain을 만드는 경향

MAX (Complete Link) : 두 군집에서 가장 먼 멤버들의 거리를 잰다. 구형 (spherical)으로 뭉치는 경향

Average Link

Centroids

Ward's method

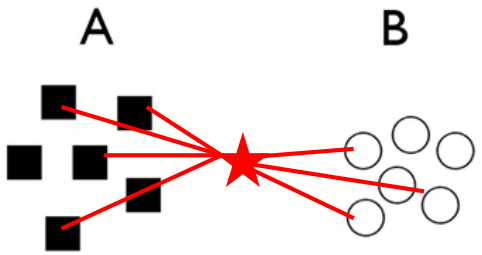


Ward's method

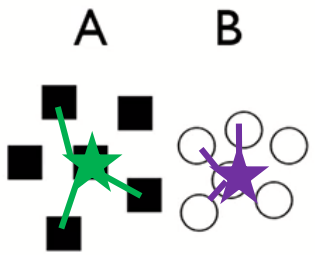
Distance between two clusters, A and B, is how much the sum of squares will increase when they are merged

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

m_A is the center of cluster A.



Ward's distance = 10 – (3+2) = 5



Ward's distance = 7 – (3+2) = 2

Unit 02 | Agglomerative Hierarchical Clustering

Hierarchical Clustering

① **단일 데이터 간 거리**를 정의하고

- Euclidean (L2) distance, Manhattan (L1) distance …

② **군집-군집 or 군집-개체 간 거리**를 정의하고

- MIN (Single Link), MAX (Complete Link), Ward's method …

③ 돌리자!

Initial points → deterministic

Merging decisions are local and final

The computational cost is very high

K-means Clustering

Algorithm 8.1 Basic K-means algorithm.

1: Select K points as initial centroids.
2: **repeat**
3: Form K clusters by assigning each point to its closest centroid.
4: Recompute the centroid of each cluster.
5: **until** Centroids do not change.

Unit 03 | K-means Clustering

K-means Clustering

① 각 군집은 하나의 중심(Centroid)을 가짐

② 각 개체는 가장 가까운 중심에 할당, 같은 중심에 할당된 개체들이 모여 하나의 군집을 형성

③ 사전에 군집의 수, K가 정해져야 함.

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

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

Unit 03 | K-means Clustering

① 데이터 내 객체 중 임의로 K개의 군집 중심점(Centroid) 설정

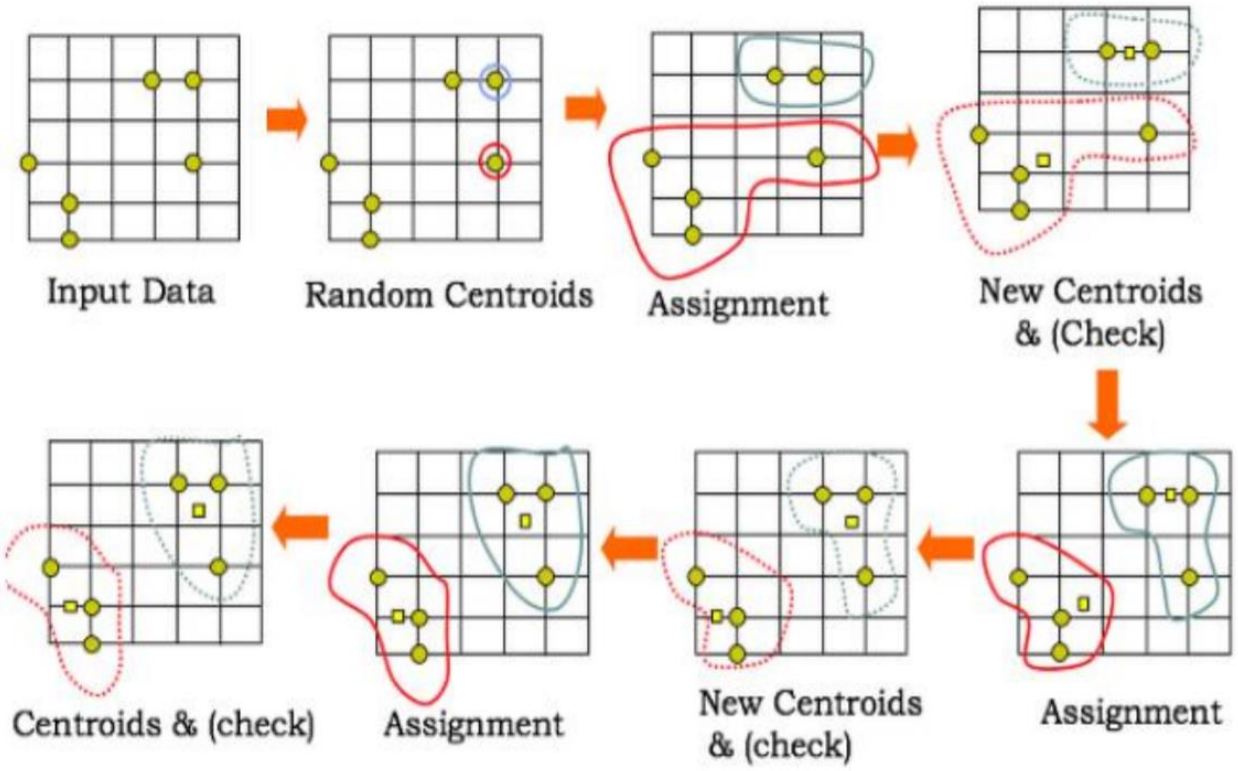
② 모든 객체에 대해 각 군집 중심점까지의 거리 계산

③ 모든 객체를 가장 가까운 군집 중심점이 속한 군집으로 할당

④ 각 군집의 중심점 재설정

⑤ 군집의 중심점이 변경되지 않을 때까지 위 과정 반복

(또는 적당한 범위 내로 수렴하거나 적당한 반복회수에 도달할 때까지 반복)



K-means Clustering

Algorithm 8.1 Basic K-means algorithm.

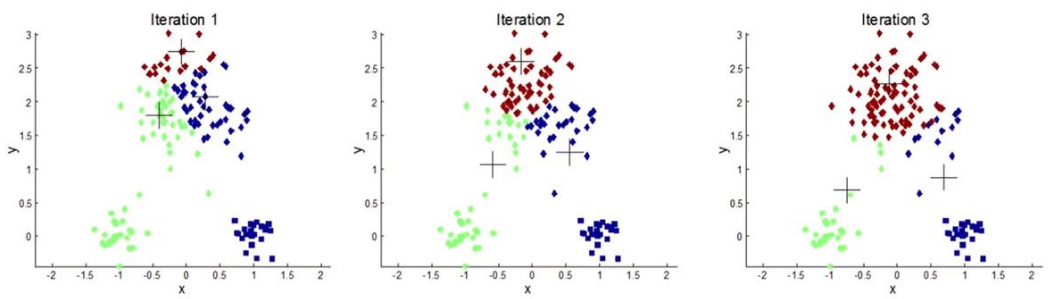
1: Select K points as initial centroids.
2: **repeat**
3: Form K clusters by assigning each point to its closest centroid.
4: Recompute the centroid of each cluster.
5: **until** Centroids do not change.

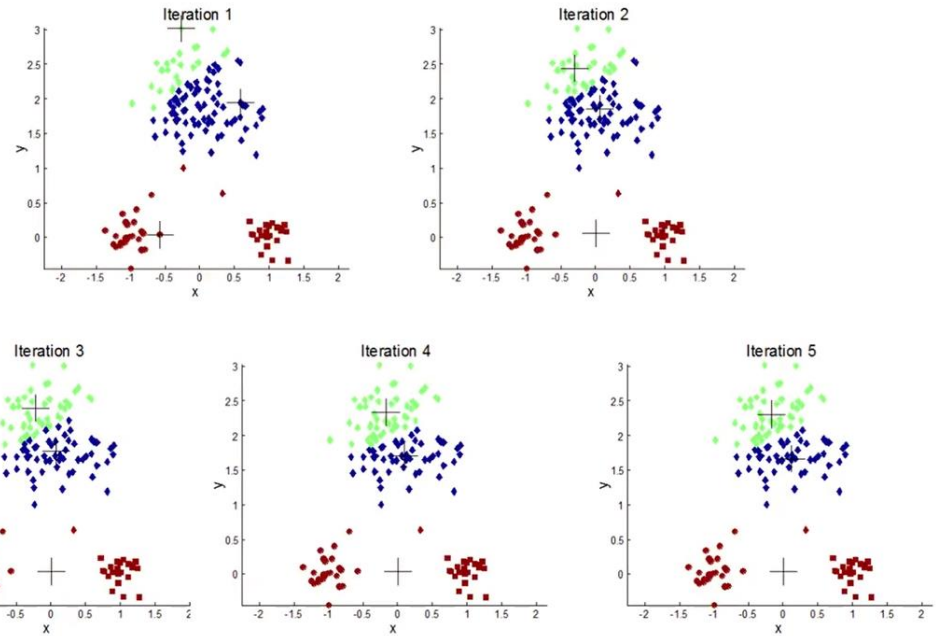
$$\underset{c}{\operatorname{argmin}} \sum_{i=1}^{K} \sum_{x_j \in C_i} \left\| x_j - c_i \right\|^2$$

Unit 03 | K-means Clustering

K-means의 주요 변수

① 초기 군집 중심점 (Centroid) 설정





K-means의 주요 변수

① 초기 군집 중심점 (Centroid) 설정

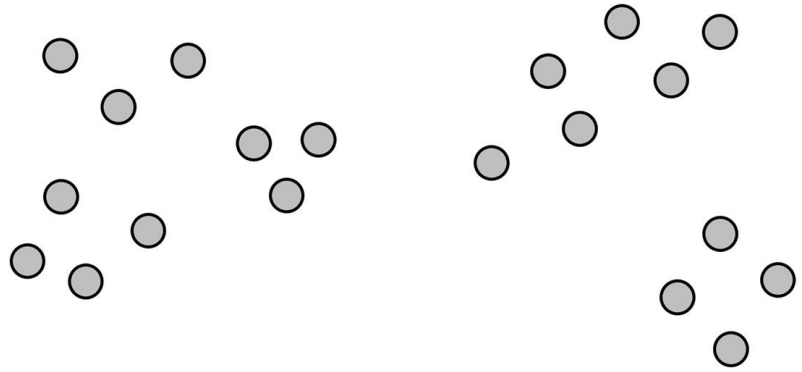
무작위 초기 중심 설정의 위험을 피하고자 다양한 연구 존재

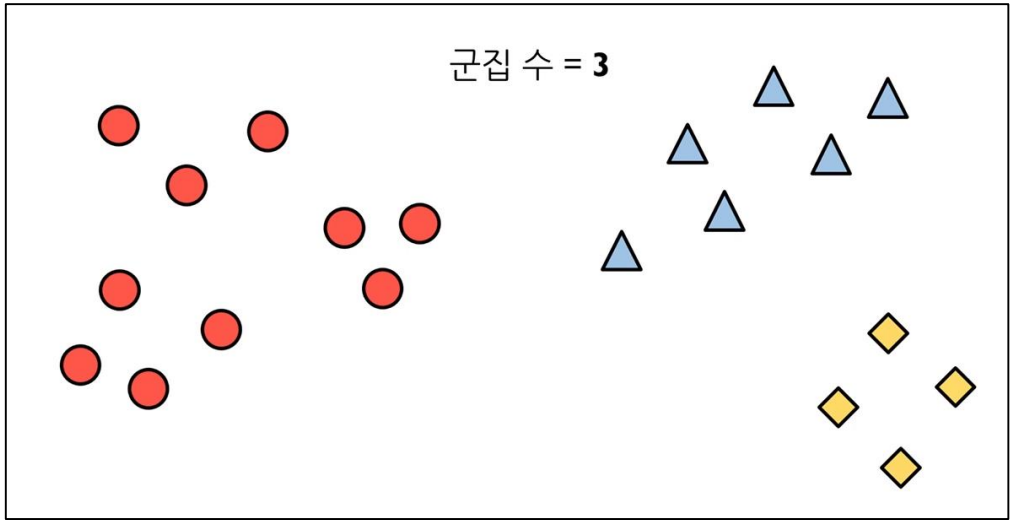
- 반복적으로 수행하여 가장 여러 번 나타나는 군집 사용
- 전체 데이터 중 일부만 샘플링하여 계층적 군집화를 수행한 뒤 초기 군집 중심 설정
- 데이터 분포의 정보를 사용하여 초기 중심 설정

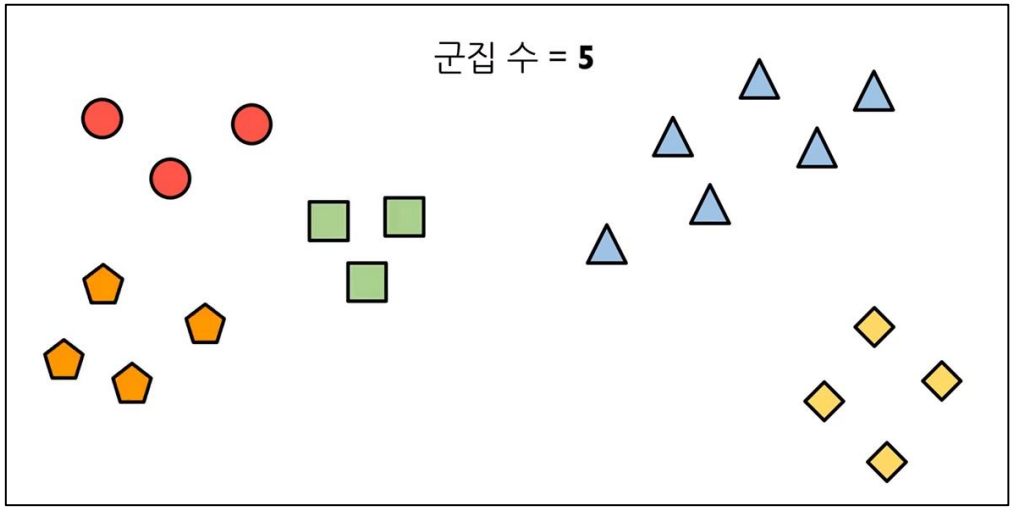
Unit 03 | K-means Clustering

K-means의 주요 변수

② 군집의 개수 (K)



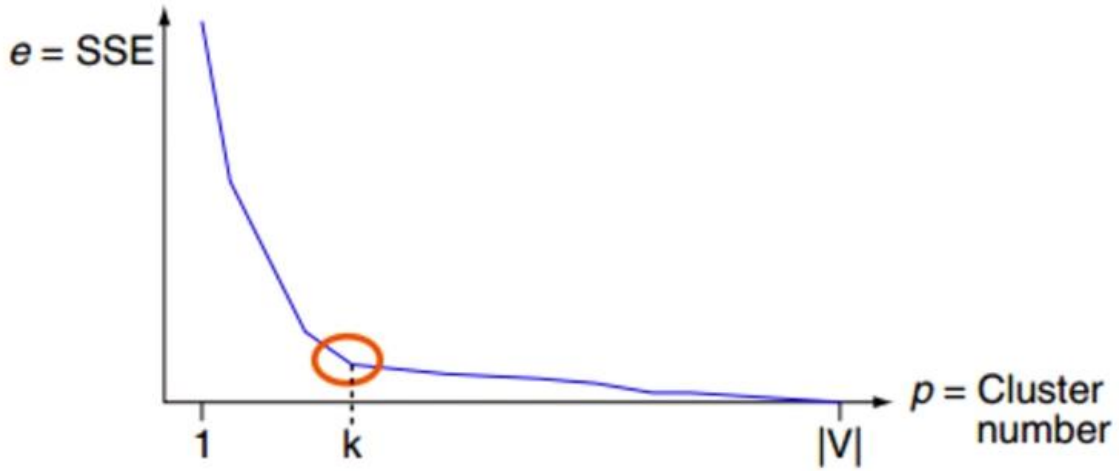




K-means의 주요 변수

② 군집의 개수 (K)

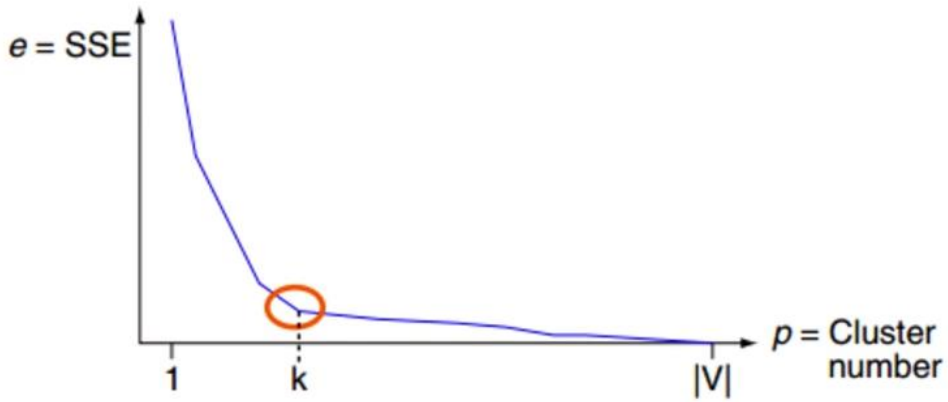
Elbow point에서 최적 군집수가 결정되는 경우가 일반적

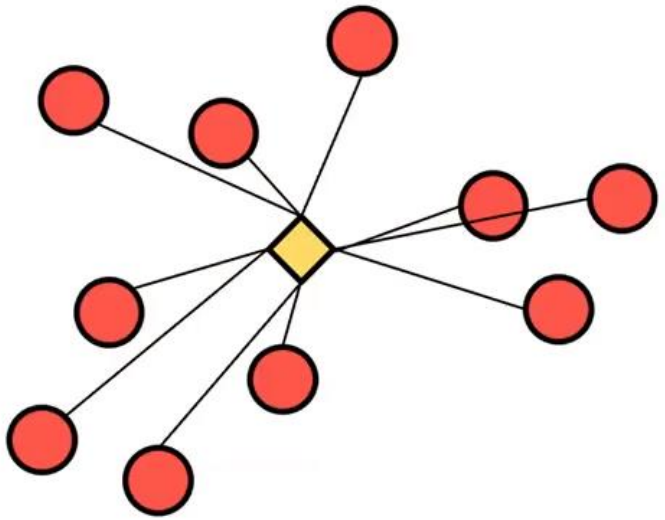


Unit 03 | K-means Clustering

SSE (Sum of Squared Error)

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



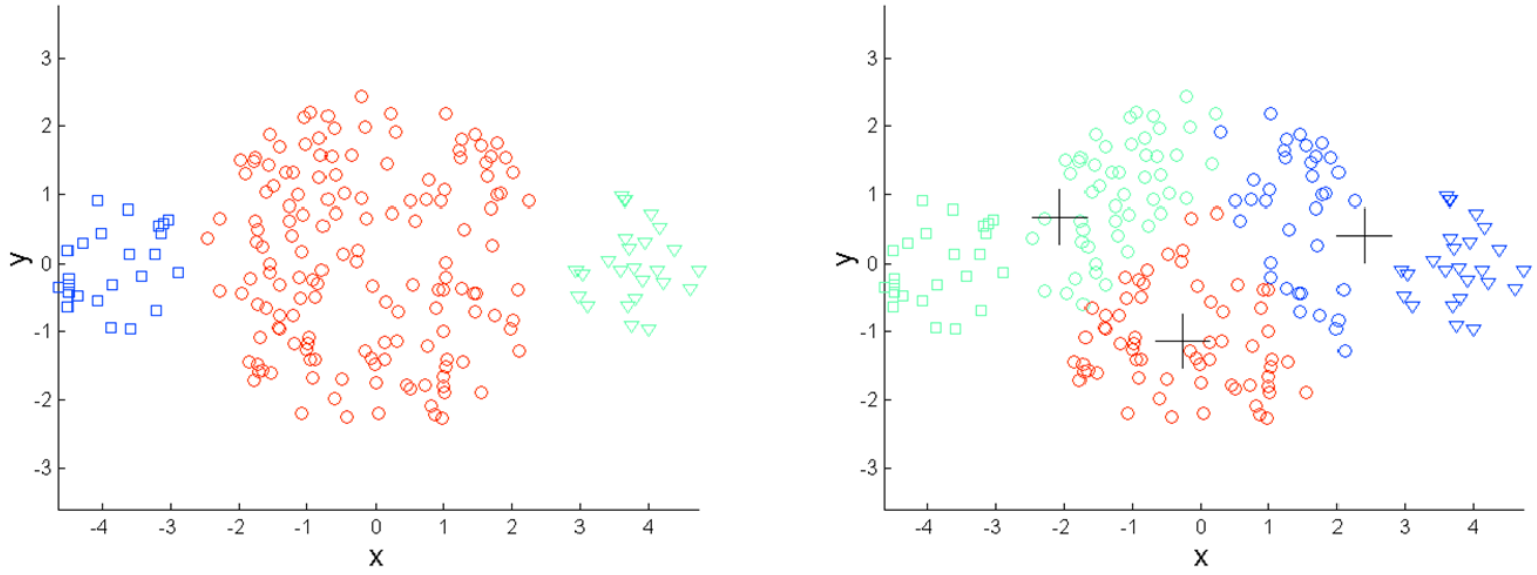


: 관측치 (x)

: 중심 (C_i)

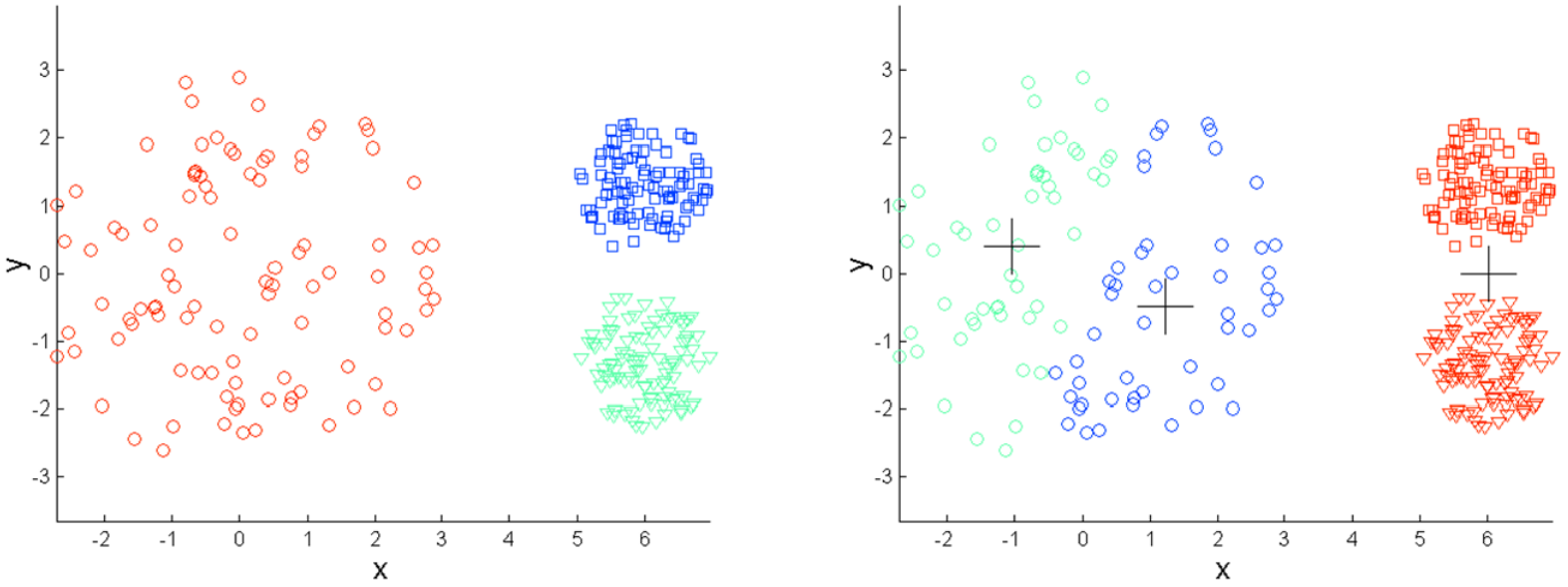
K-means Clustering의 문제점

① 서로 다른 크기의 군집을 잘 찾아내지 못함



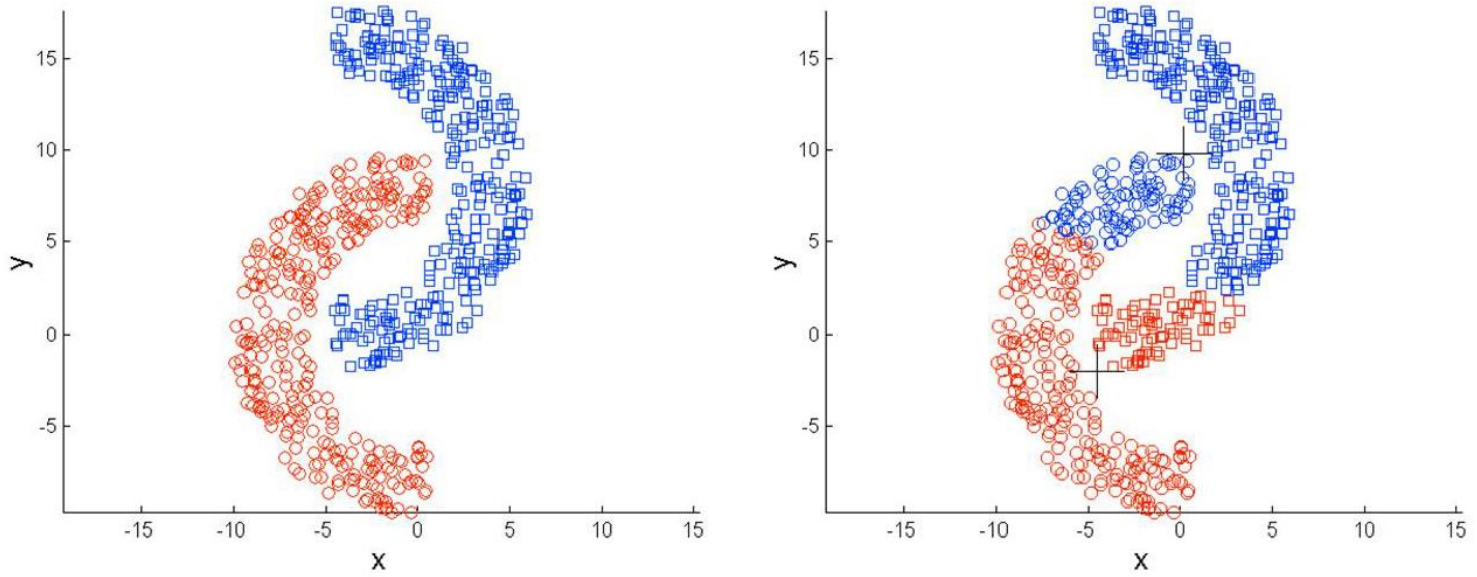
K-means Clustering의 문제점

② 서로 다른 밀도의 군집을 잘 찾아내지 못함



K-means Clustering의 문제점

③ 지역적 패턴이 존재하는 군집을 판별하기 어려움

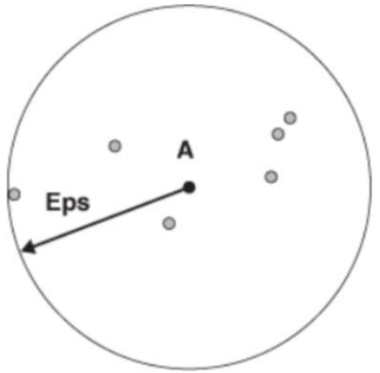


Unit 04 | DBSCAN

DBSCAN : Density-based spatial clustering of applications with noise

점 P에서부터 거리 Eps 내에 점이 MinPts개 이상 있으면 하나의 군집으로 인식

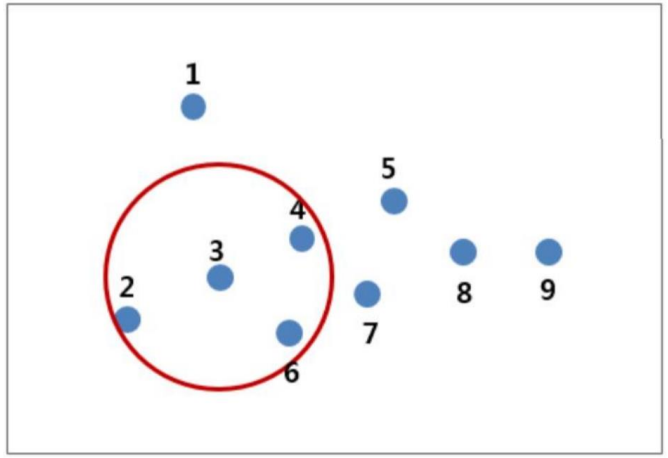
- Example
  - The density of A = 7
    - Because the number of points within a radius of *Eps* of A is 7, including A itself



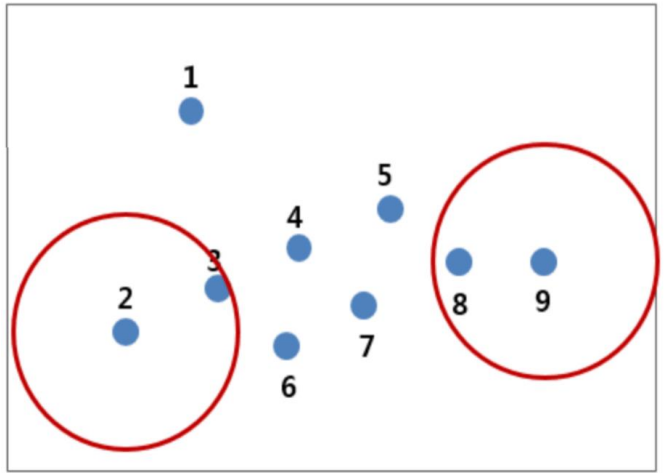
Unit 04 | DBSCAN

MinPts = 4

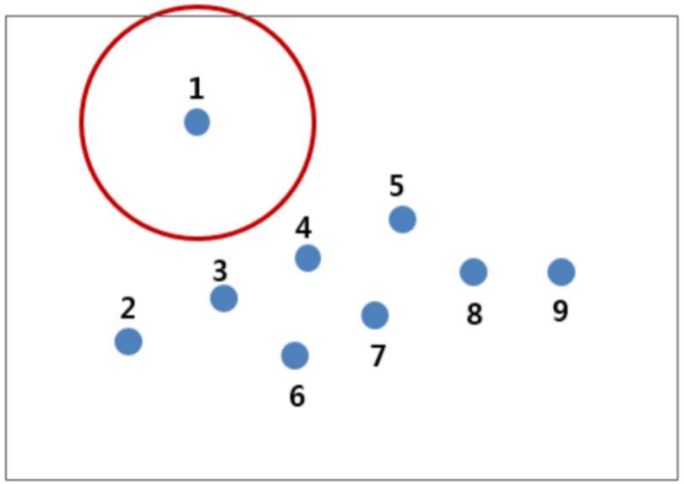
① Core points



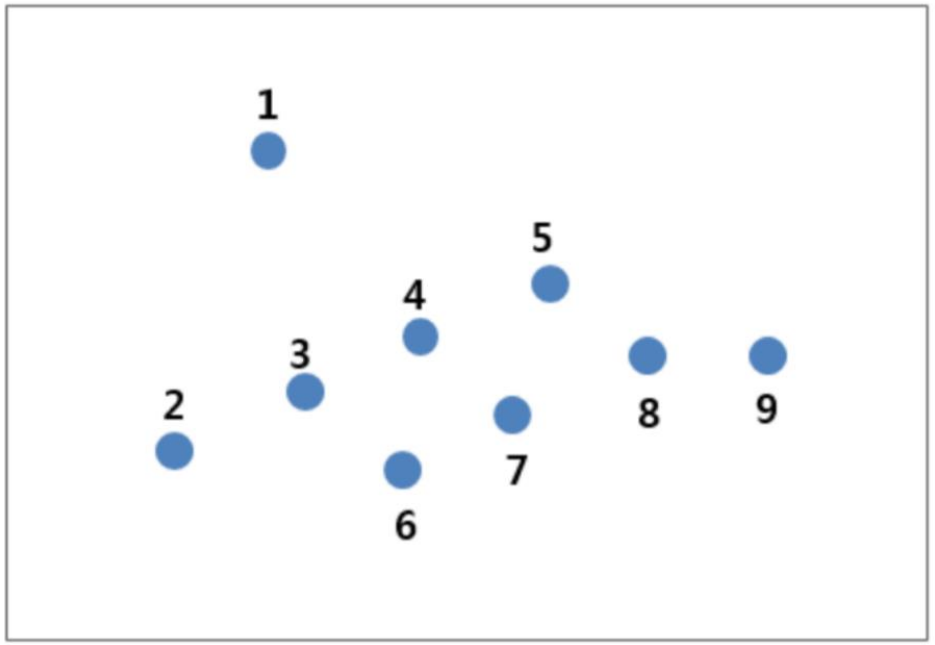
② Border points

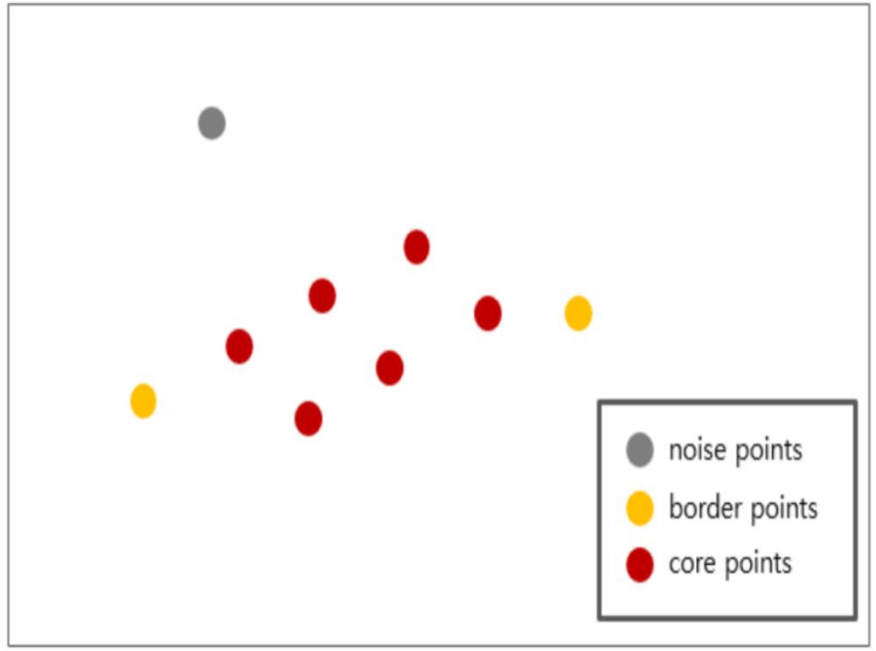


③ Noise points



Unit 04 | DBSCAN





DBSCAN

Algorithm 8.4 DBSCAN algorithm.

1: Label all points as core, border, or noise points.
2: Eliminate noise points.
3: Put an edge between all core points that are within *Eps* of each other.
4: Make each group of connected core points into a separate cluster.
5: Assign each border point to one of the clusters of its associated core points.

Selection of DBSCAN Parameters

① Eps

② MinPts

- k-dist : the distance from a point to its kth neighbor

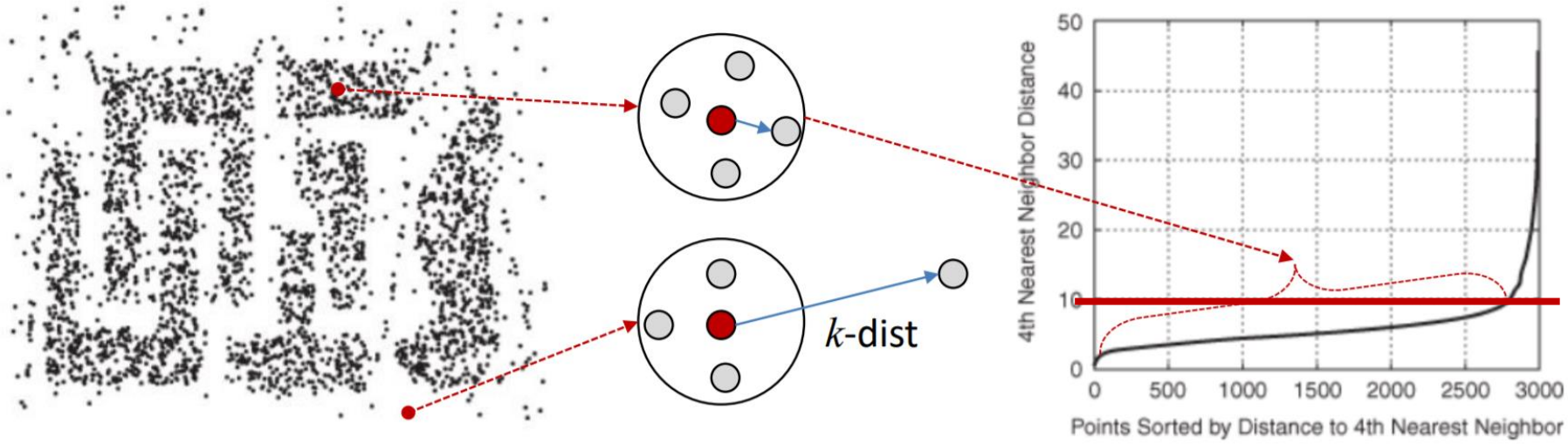
Unit 04 | DBSCAN

① Compute k-dist for all pointe for some k (e.g., k = 4)

② Sort them in increasing order, and then plot the sorted values

③ If we see sharp change at the value of k-dist, then

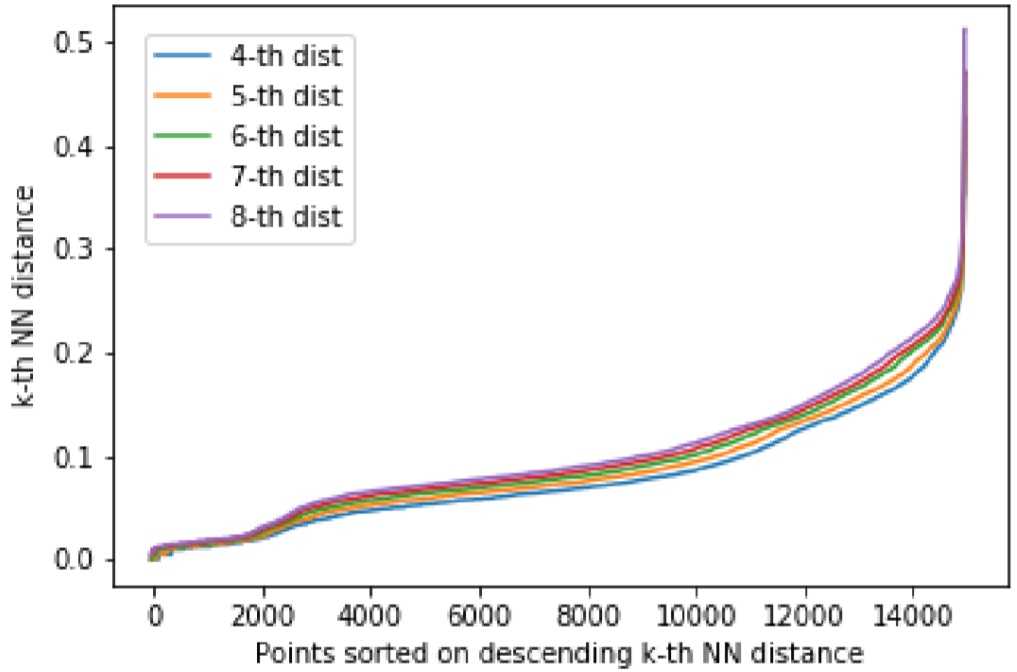
we take this distance as Eps (e.g., Eps = 10) and the value of k as MinPts (e.g., MinPts = 4)



Most points within a cluster have at least **4** points within a radius of **10**

Selection of DBSCAN Parameters

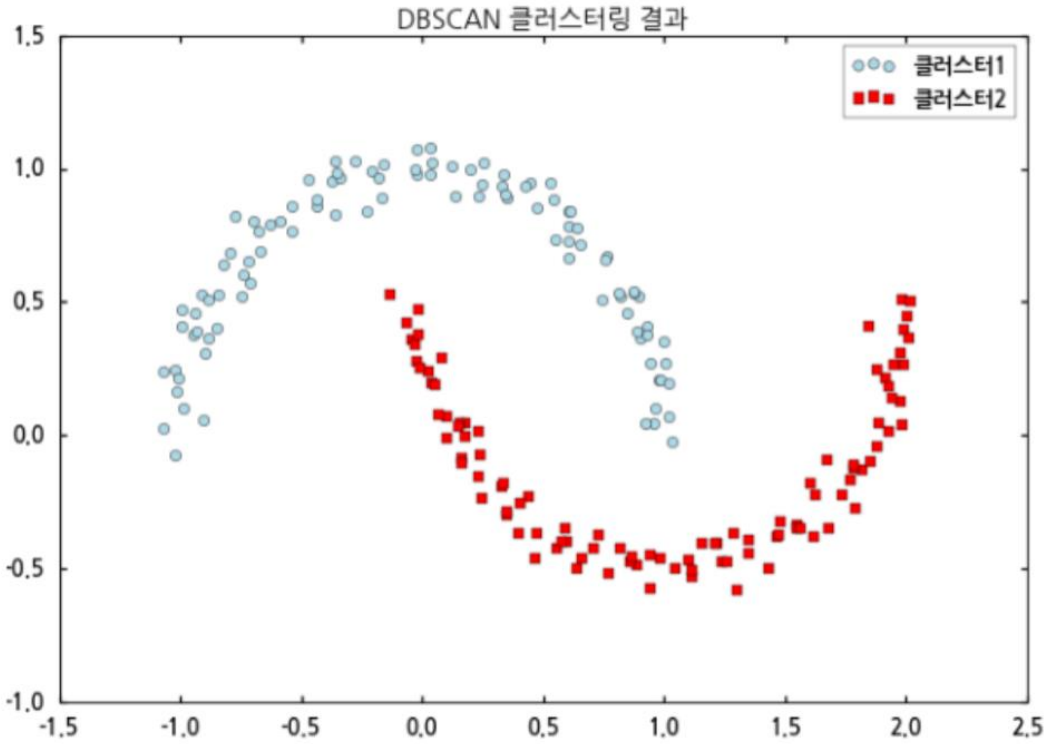
The original DBSCAN algorithm used a value of k = 4 (which appears to be a reasonable value for most 2-dimensional data sets)



https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

DBSCAN needs two parameters, Eps and MinPts. However, our experiments indicate that the k-dist graphs for $k > 4$ do not significantly differ from the 4-dist graph and, furthermore, they need considerably more computation. Therefore, we eliminate the parameter MinPts by setting it to 4 for all databases (for 2-dimensional data). We propose the following interactive approach for determining the parameter Eps of DBSCAN:

DBSCAN의 특징



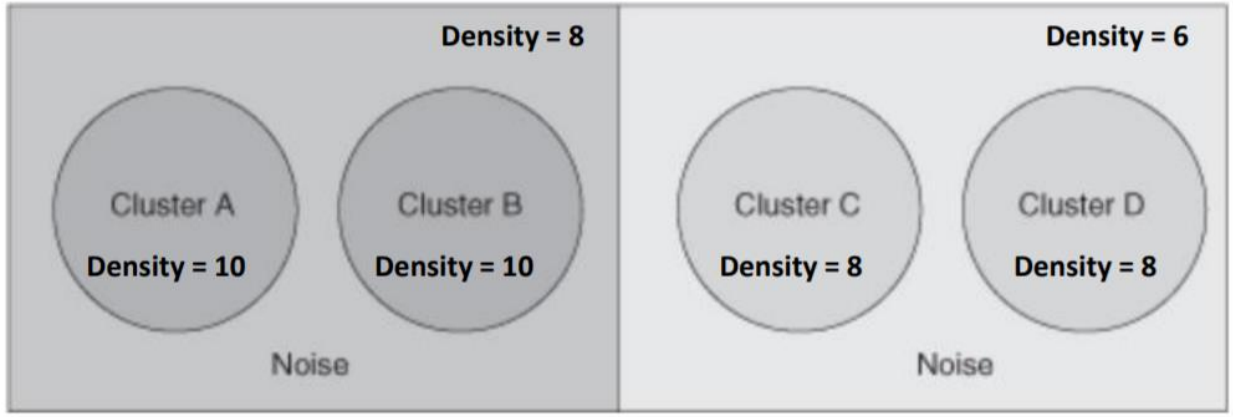
- 클러스터의 수를 정하지 않아도 됨

- 비선형 경계의 군집을 구하는 것도 가능
  (밀도에 따라 클러스터를 서로 연결하기 때문)

- 노이즈 데이터를 따로 분류하여 노이즈 데이터들이 군집에 영향을 주지 않음

DBSCAN의 문제점 : density가 구역에 따라 바뀌는 경우

- If *MinPts* is chosen to find C and D as dense clusters (e.g., $MinPts = 7$)
  - Then A, B, and the points surrounding them will become a ***single*** cluster
- If *MinPts* is chosen to find A and B as dense clusters (e.g., $MinPts = 9$)
  - Then C, D, and the points surrounding them will be marked as ***noise***

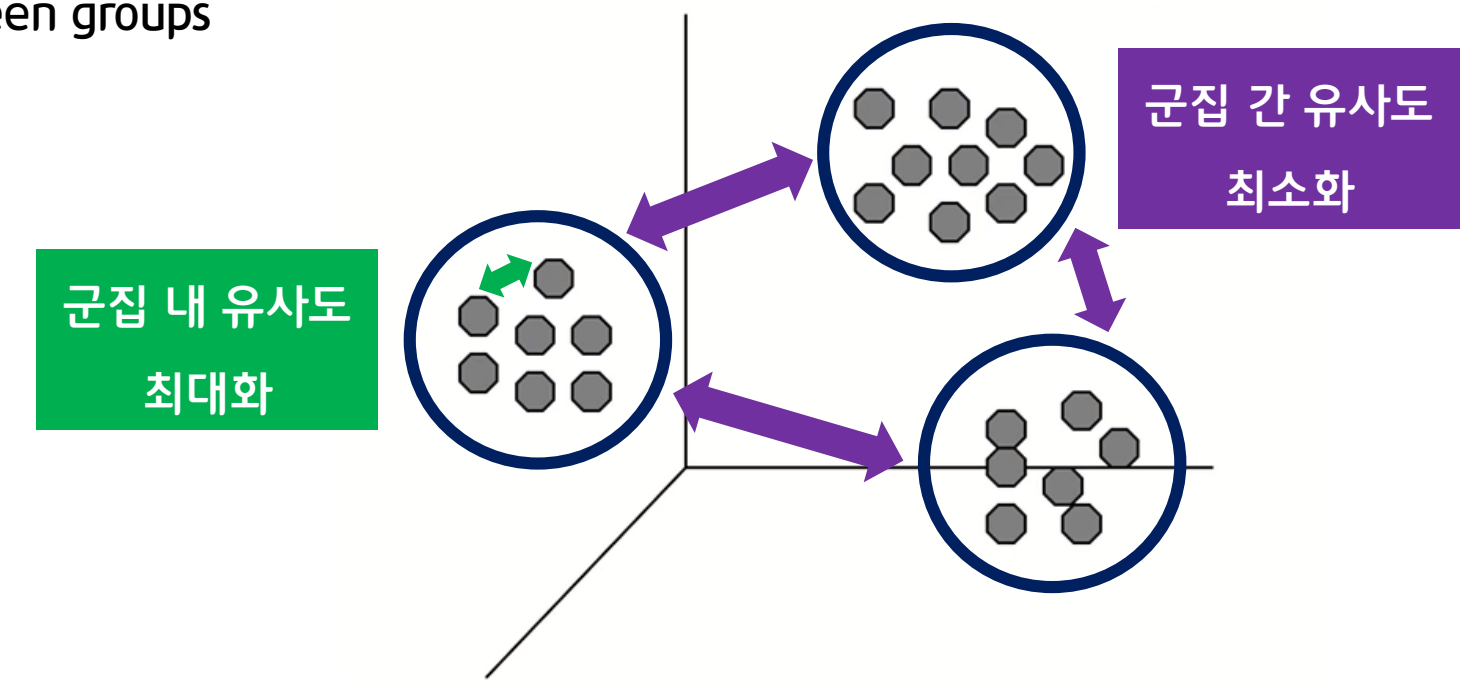


Assume that Eps is fixed

Unit 05 | 모델 평가

좋은 Clustering?

1. Maximizes the similarity within a group
2. Maximizes the difference between groups



Unit 05 | 모델 평가

내부 평가

클러스터링한 그 자체를 놓고 평가하는 방식

1. Dunn Index
2. 실루엣 (Silhouette)

외부 평가

클러스터링에 사용되지 않은 데이터로 평가하는 방식

1. Rand Measure
2. F-Measure
3. Jaccard Index

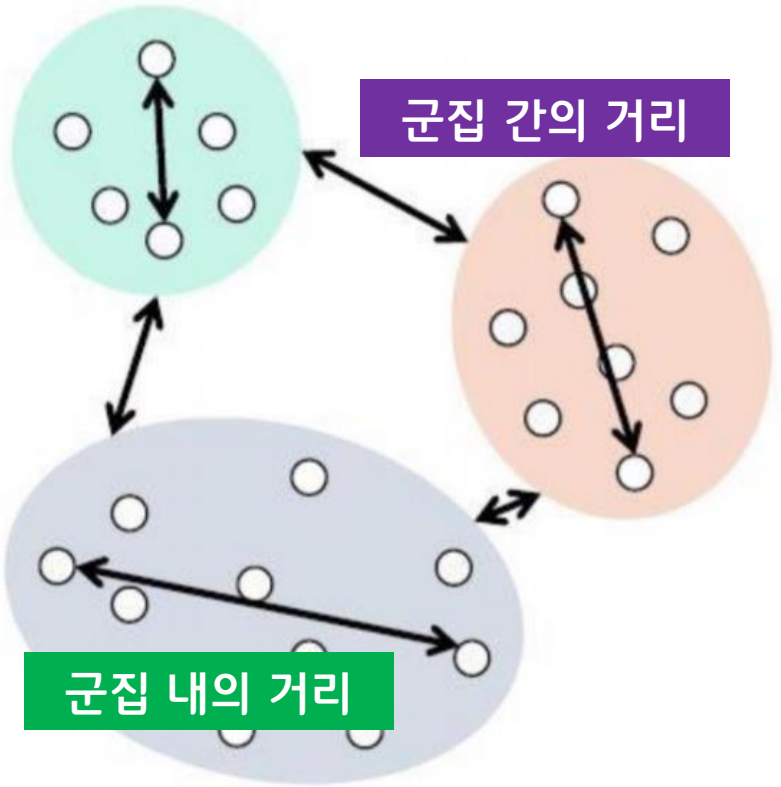
Unit 05 | 모델 평가

1. Dunn Index

$$DI = \frac{\text{군집과 군집 사이의 거리 중 최솟값}}{\text{군집 내 객체 간 거리 중 최댓값}}$$

군집과 군집 사이의 거리가 클수록,

군집 내 객체간 거리가 작을수록 좋은 모델 → DI가 큰 모델



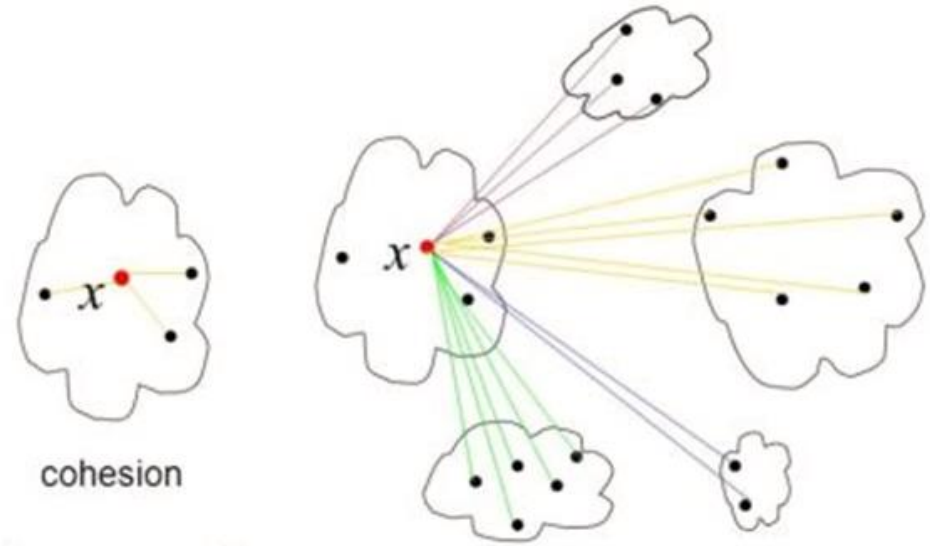
Unit 05 | 모델 평가

2. 실루엣 (Silhouette)

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

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

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



$a(x)$: average distance in the cluster

군집 내의 거리

$b(x)$: average distances to others clusters, find minimal

군집 간의 거리

일반적으로 $\bar{S}$의 값이 0.5보다 크면 군집 결과가 타당하다고 볼 수 있음

-1에 가까우면 군집이 전혀 되지 않음

과 제

Clustering 해보기 : Mall_Customers.csv

1) Preprocessing / EDA

2) Clustering

- **Hierarchical agglomerative Clustering, K-Means Clustering, DBSCAN, …**

3) Evaluation

CustomerID : Unique ID assigned to the customer

Gender : Gender of the customer

Age : Age of the customer

Annual Income(k$) : Annual Income of the customer

Spending Score(1-100) : Score assigned by the mall based on customer behavior and spending nature

참고자료

- 투빅스 10기 임진혁 강의자료
- 투빅스 9기 박송은 강의자료
- https://www.youtube.com/watch?v=8zB-_LrAraw
- Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar, "Introduction to Data Mining," 2nd Edition, Pearson, 2018.
- https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68
- https://gentlej90.tistory.com/64?category=682471
- https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

들어주셔서 감사합니다.