분할적 군집분석

10강. 분할적 군집분석

- K-Means와 DBSCAN
- 퍼지군집화 / EM 알고리즘 / SOM
- 엘보우 기법

비지도학습의 종류

분할적 군집분석

계층적 관계가 없는 다수의 군집들을 만드는 방법

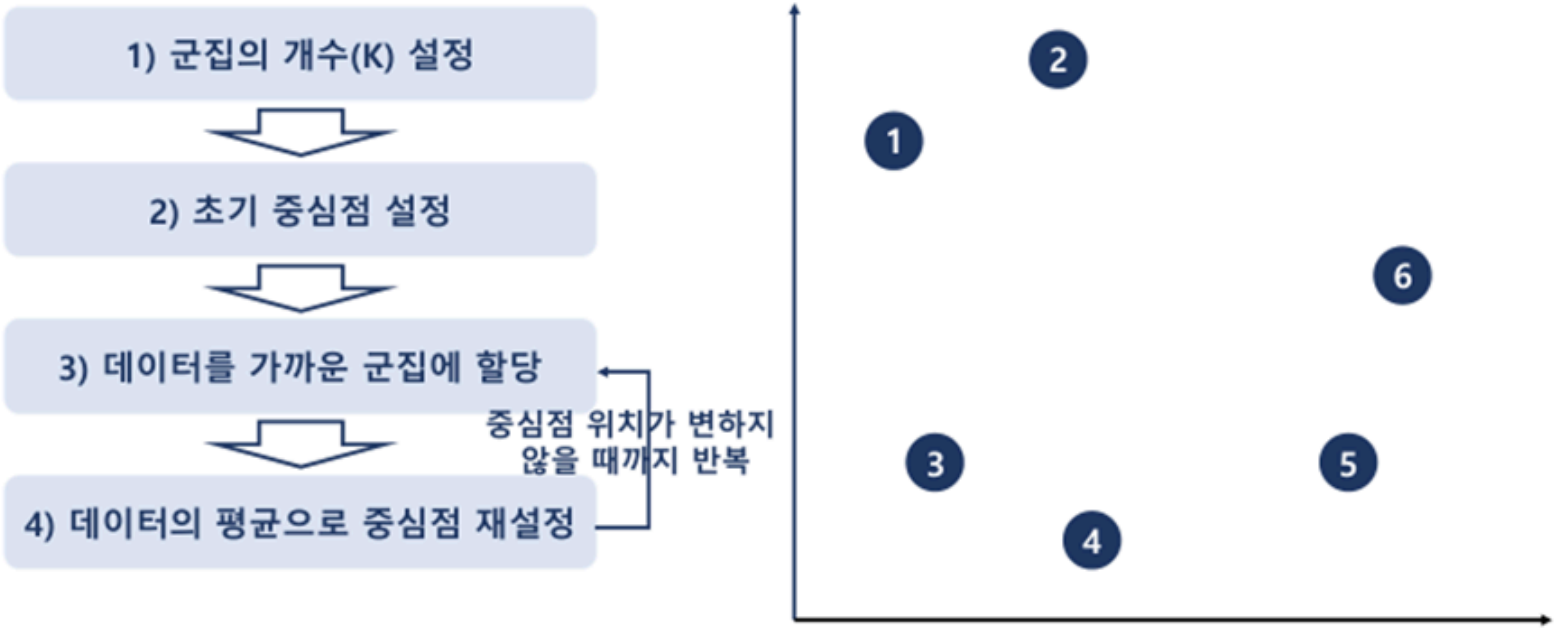
- 중심점 기반
  - K-Means 군집화
- 밀도 기반
  - DBSCAN 군집화
- 확률 기반
  - 퍼지군집화
- 분포 기반
  - EM알고리즘
- 그래프 기반
  - 자기조직화지도(SOM)

K-Means 군집화

각 데이터와 중심점의 거리를 측정 후 가장 가까운 그룹에 할당하여 K개의 군집으로 묶는 방법

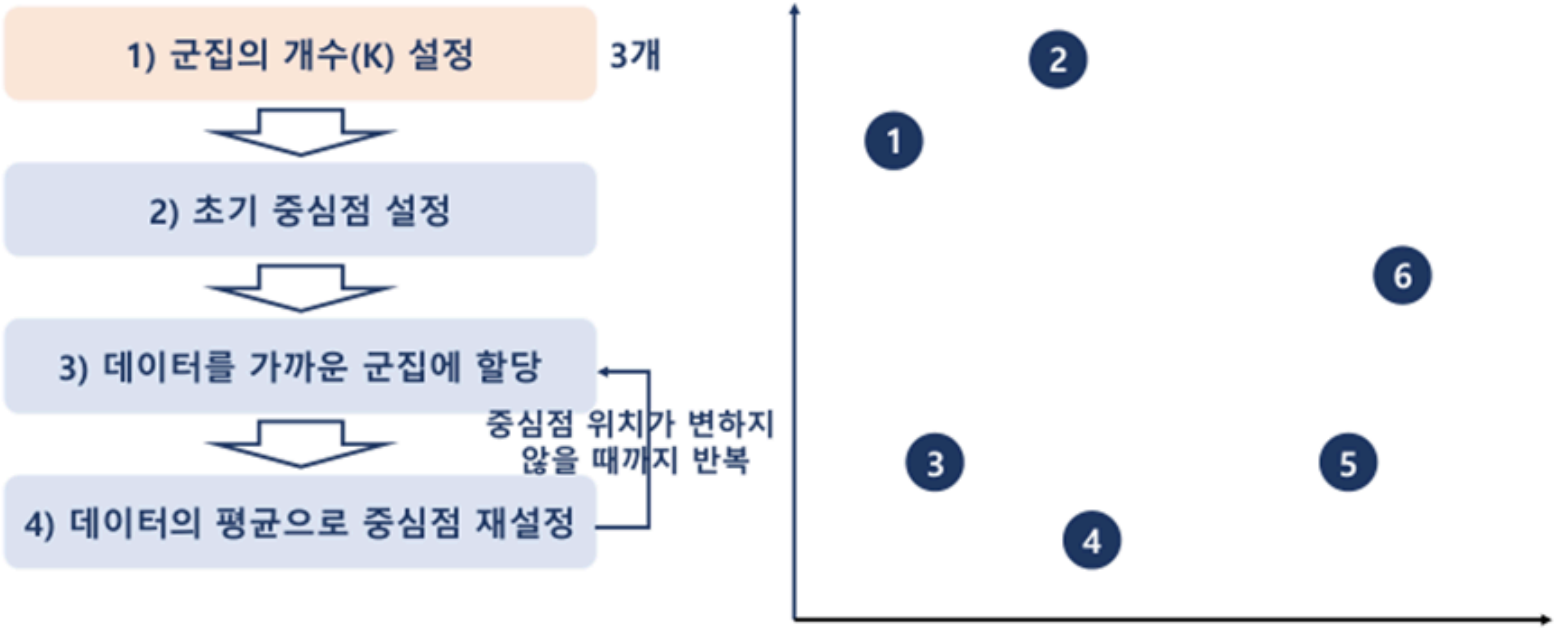
K-Means 군집화

각 데이터와 중심점의 거리를 측정 후 가장 가까운 그룹에 할당하여 K개의 군집으로 묶는 방법



K-Means 군집화

각 데이터와 중심점의 거리를 측정 후 가장 가까운 그룹에 할당하여 K개의 군집으로 묶는 방법

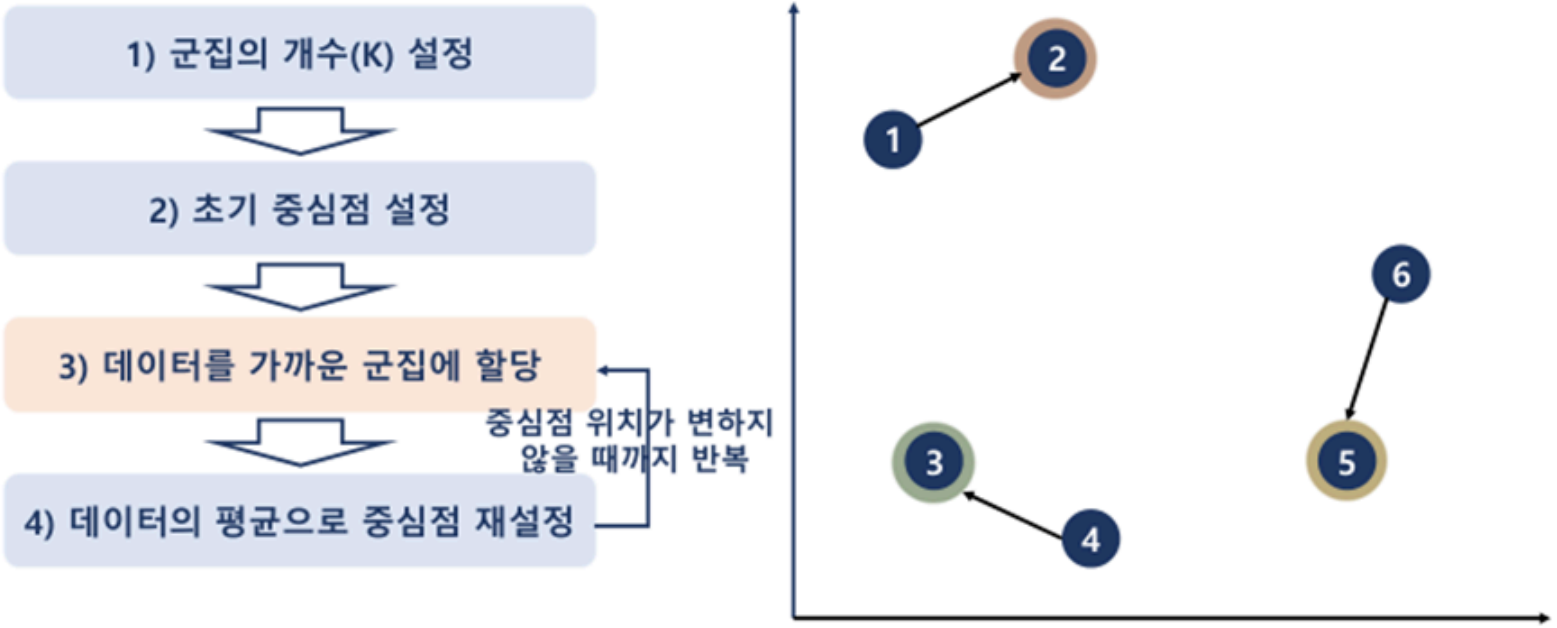


K-Means 군집화

각 데이터와 중심점의 거리를 측정 후 가장 가까운 그룹에 할당하여 K개의 군집으로 묶는 방법

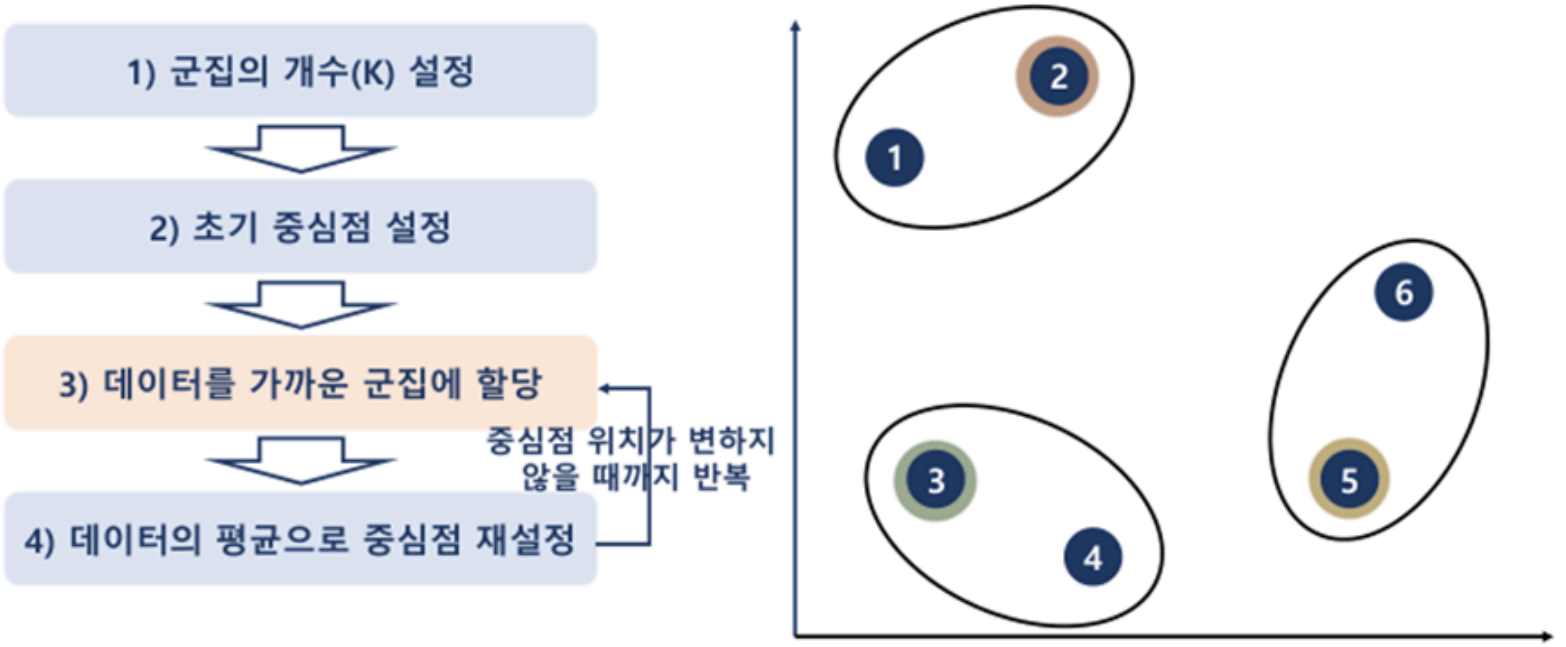
K-Means 군집화

각 데이터와 중심점의 거리를 측정 후 가장 가까운 그룹에 할당하여 K개의 군집으로 묶는 방법



K-Means 군집화

각 데이터와 중심점의 거리를 측정 후 가장 가까운 그룹에 할당하여 K개의 군집으로 묶는 방법



K-Means 군집화

각 데이터와 중심점의 거리를 측정 후 가장 가까운 그룹에 할당하여 K개의 군집으로 묶는 방법

1) 군집의 개수(K) 설정

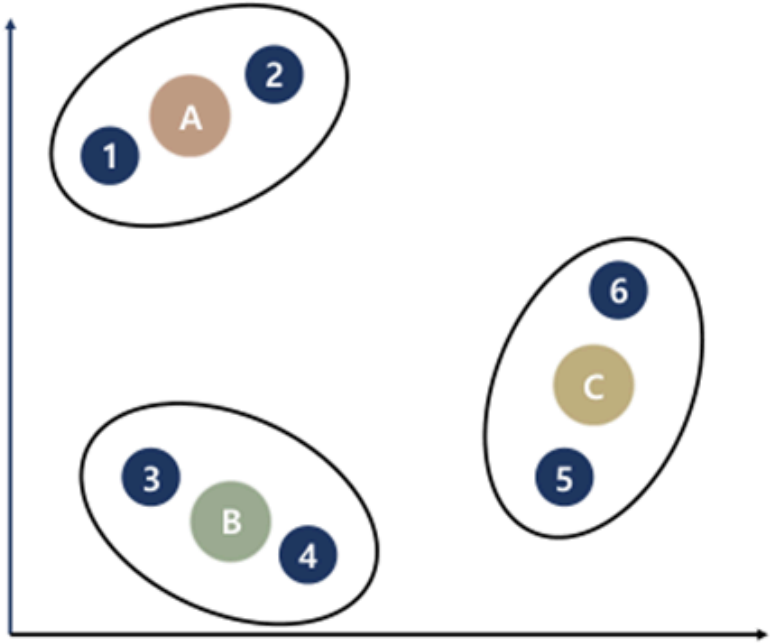
2) 초기 중심점 설정

3) 데이터를 가까운 군집에 할당

4) 데이터의 평균으로 중심점 재설정

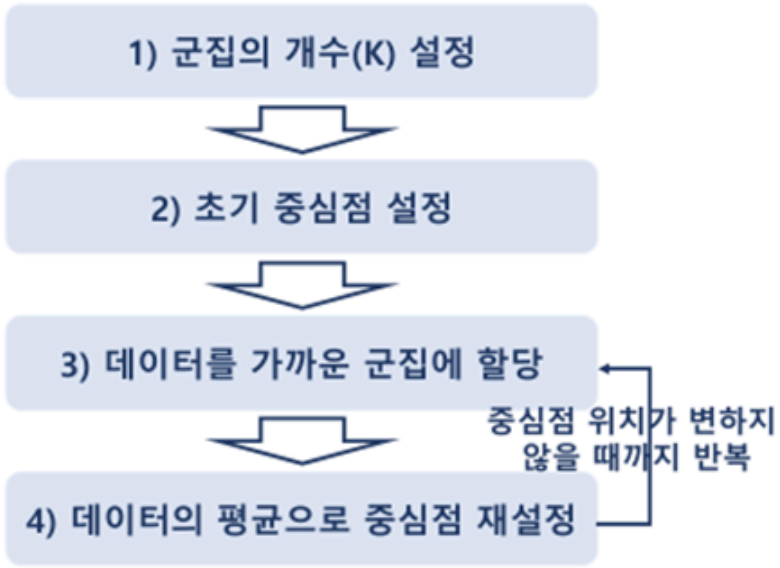
중심점 위치가 변하지 않을 때까지 반복

중심점 위치가 변하지 않으므로 종료



K-Means 군집화

각 데이터와 중심점의 거리를 측정 후 가장 가까운 그룹에 할당하여 K개의 군집으로 묶는 방법



- 알고리즘의 이해와 적용이 쉬움
- 데이터에 대한 사전 정보가 특별히 필요 없음
- 초기 군집의 수를 정의하는데 어려움
- 초기 중심점 설정에 따라 결과가 다르게 나옴
- 노이즈나 이상치에 매우 민감
- 범주형 데이터에서 사용 불가
  → K-Modes 군집화 활용

K-Means 군집화

최적의 군집 개수 K를 결정하는 방법?

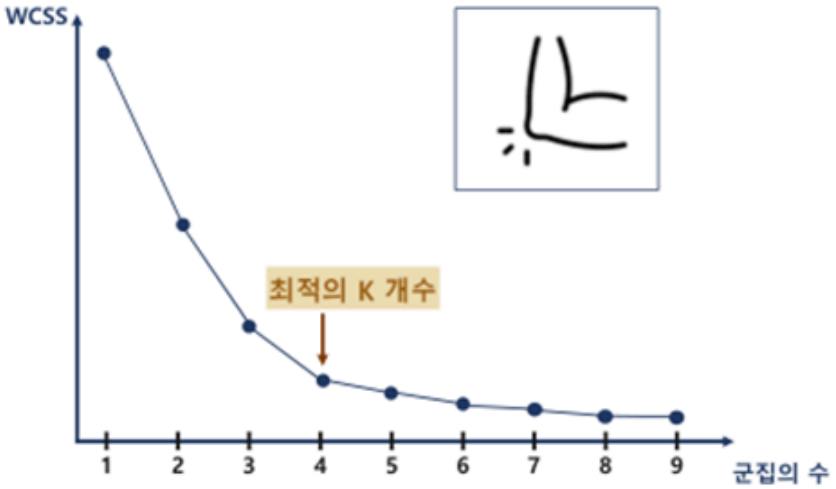
엘보우 기법(Elbow Method)

WCSS 값과 군집의 개수를 두고 비교 한 그래프를 통해 최적의 K 값을 선택하는 기법

$$WCSS = \sum_{C_k}^{C_n} \left(\sum_{d_i in C_i}^{d_m} distance(d_i, C_k)^2 \right)$$

(Within Clusters Sum of Squares)

- C : 군집(Cluster)의 중심 값
- d : 클러스터 내에 있는 데이터



경사가 완만해지는 지점이 최적의 K 개수

DBSCAN 군집화

데이터의 밀도를 기반으로 서로 가까운 데이터들을 군집으로 묶는 방법

DBSCAN 군집화

데이터의 밀도를 기반으로 서로 가까운 데이터들을 군집으로 묶는 방법

1) 포인트 임의로 선택

2) Epsilon 거리 내 모든 데이터 탐색

3) Min Points 이상이면 군집에 할당

4) Min Points 이하이나 군집에 속한 포인트는 Border Point

5) 어느 군집에도 속하지 않는 포인트는 이상치

DBSCAN 군집화

데이터의 밀도를 기반으로 서로 가까운 데이터들을 군집으로 묶는 방법

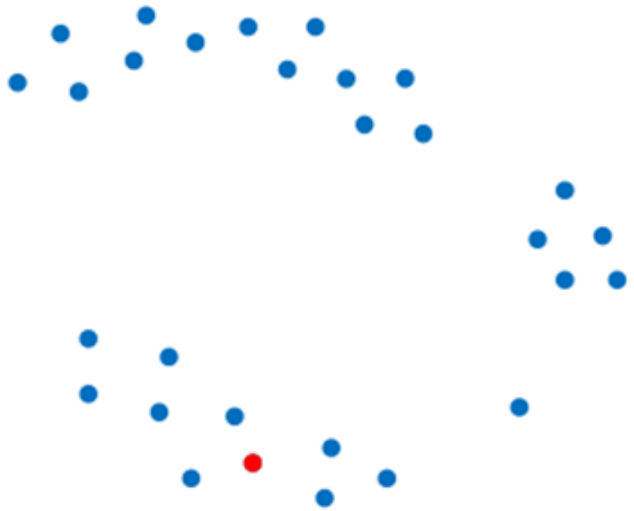
1) 포인트 임의로 선택

2) Epsilon 거리 내 모든 데이터 탐색

3) Min Points 이상이면 군집에 할당

4) Min Points 이하이나 군집에 속한 포인트는 Border Point

5) 어느 군집에도 속하지 않는 포인트는 이상치



DBSCAN 군집화

데이터의 밀도를 기반으로 서로 가까운 데이터들을 군집으로 묶는 방법

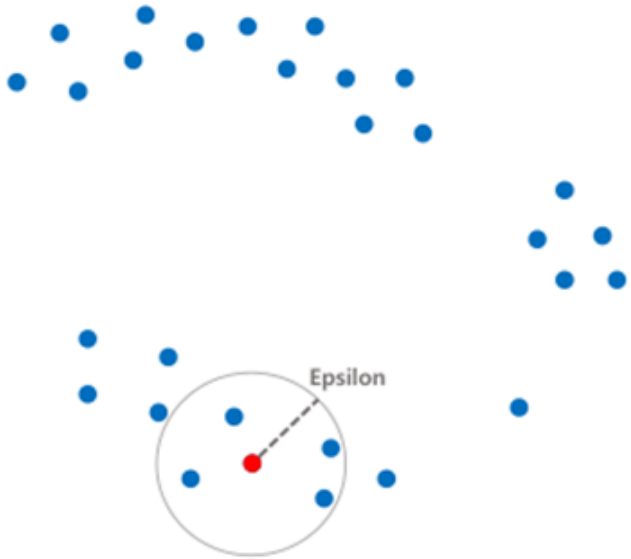
1) 포인트 임의로 선택

2) Epsilon 거리 내 모든 데이터 탐색

3) Min Points 이상이면 군집에 할당

4) Min Points 이하이나 군집에 속한 포인트는 Border Point

5) 어느 군집에도 속하지 않는 포인트는 이상치



DBSCAN 군집화

데이터의 밀도를 기반으로 서로 가까운 데이터들을 군집으로 묶는 방법

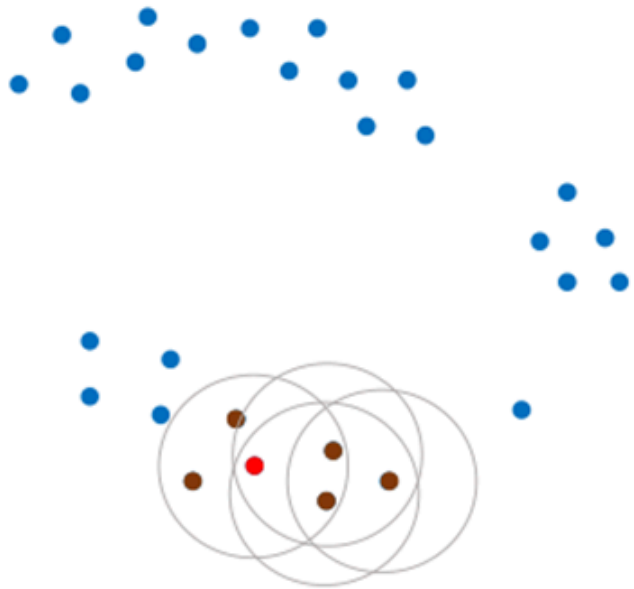
1) 포인트 임의로 선택

2) Epsilon 거리 내 모든 데이터 탐색

3) Min Points 이상이면 군집에 할당 4개

4) Min Points 이하이나 군집에 속한 포인트는 Border Point

5) 어느 군집에도 속하지 않는 포인트는 이상치



DBSCAN 군집화

데이터의 밀도를 기반으로 서로 가까운 데이터들을 군집으로 묶는 방법

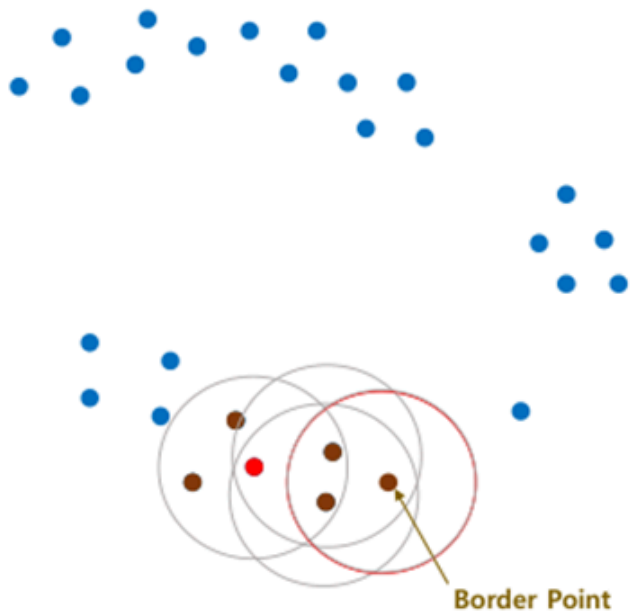
1) 포인트 임의로 선택

2) Epsilon 거리 내 모든 데이터 탐색

3) Min Points 이상이면 군집에 할당

4) Min Points 이하이나 군집에 속한 포인트는 Border Point

5) 어느 군집에도 속하지 않는 포인트는 이상치



DBSCAN 군집화

데이터의 밀도를 기반으로 서로 가까운 데이터들을 군집으로 묶는 방법