군집분석이란?

- 데이터 간의 유사도를 정의하고, 그 유사도에 가까운 것부터 순서대로 합쳐 가는 방법으로 그룹(군집)을 형성한 후 각 그룹의 성격을 파악하거나 그룹 간의 비교분석을 통해서 데이터 전체의 구조에 대한 이해를 돕고자 하는 탐색적인 분석 방법

군집 분석의 목적

- 데이터 셋 전체를 대상으로 서로 유사항 개체 들을 몇 개의 군집으로 세분화하여 대상 집단을 정확하게 이해하고,효율적으로 활용하기 위함.

  - 군집 분석으로 그룹화된 군집은 변수의 특성이 그룹 내적으로는 동일하고, 외적으로는 이질적인 특성을 가짐
  - 군집 분석의 용도는 고객의 충성도에 따라서 몇 개의 그룹으로 분류하고, 그룹별로 맞춤형 마케팅 및 프로모션 전략을 수립하는 데 활용됨

계층적 군집분석

- 데이터를 유사도에 따라 계층적으로 묶어주는 군집분석 방법
- 예를 들어, 고객 데이터를 구매 패턴과 구매 금액에 따라 묶어주는 것이나, 상품 데이터를 재료, 맛, 가격 등에 따라 묶어주는 것이 대표적인 예임
- 이렇게 묶인 군집들은 덴드로그램 형태로 시각화되어 나타남

비계층적 군집분석(Non-Hierarchical Clustering)

- 계층 구조를 형성하지 않고 데이터 포인트들을 클러스터로 그룹화하는 방법으로, 주로 K-means와 같은 알고리즘이 사용됨

  - K-means의 경우, 미리 정해진 클러스터 개수(K)를 기준으로 중심점을 설정하고, 각 데이터 포인트가 가장 가까운 중심점에 속하게 클러스터링하는 방식으로 진행됨

- 비계층적 군집분석 사례:

- 마케팅 - 고객 데이터를 분석하여 고객 세분화(customer segmentation)를 수행하고, 각 세분화된 고객 그룹에 맞춘 마케팅 전략을 수립할 수 있음
- 이미지 분석 - 이미지 픽셀을 클러스터링하여 각 클러스터의 대표 색상을 추출하여 이미지 색상 분석이나 색상 분류를 할 수 있음
- 유전자 데이터 분석 - 유전자 데이터에서 유전자 클러스터를 식별하고, 유전자 간의 관련성을 분석할 수 있음
- 문서 데이터 분석 - 문서 데이터를 분석하여 유사한 내용을 가진 문서들을 클러스터링하여 문서 분류를 할 수 있음
- 고객 행동 분석 - 웹사이트 방문 고객의 행동 패턴을 분석하여 고객 세분화 및 맞춤형 서비스 제공에 활용할 수 있음

군집분석 사례 - 신한카드(고객들을 군집 분석해서 고객마다 원하는 카드혜택이 다르게 하고, 고객 특성을 새롭게 정의함으로써 고객 특성별로 특화된 카드 상품을 만드는 데 활용)

- 신한카드는 2014년부터 빅데이터를 활용하여 고객의 소비 특성을 파악하여 고객 분류를 하고, 각 고객 그룹에 맞는 카드 혜택과 유형을 정했음
- 그래서 나온 것이 바로 Code 9(코드나인)이라고 하는 것임
- 고객정보와 카드 결제내역을 군집분석하여 남녀 고객을 위와 같이 총 18가지로 분류하고 각 분류별로 Rookie, It-Girl 같은 명칭을 지었음
- 각 고객군집의 특성을 파악해서 카드 이름도 지은 것 같음


M
W
Rookie
새롭고 다양한 분야에 관심이 많은 프레쉬맨
LOEL
패션 및 명품 브랜드에 관심이 많으며 외모가꾸길 즐기는 센스남
Friend Daddy
자녀와의 친밀감을 중시하는 친구처럼 가까운 아빠
It-Girl
최신 유행 및 트렌드를 선도하는 패셔니스타
Prima Donna
문화와 여가를 즐기는 전문직 싱글 여성
Trend Setter
럭셔리한 삶을 추구하는 세련된 감각의 여성
Smart Saver
합리적 가격을 중요시하며 계획적인 소비를 즐기는 플랜맨
Mr. Routine
하루하루 최선을 다하며 소소한 행복을 추구하는 가장 standard한 직장인
BOBOS
일과 여가를 즐기며, 독특한 소비감각을 지닌 여유로운 남성
Ruby
자기계발에 적극적인 젊은 감각의 여성
Alpha Mom
자녀교육에 매진하는 똑똑한 엄마
Queen of House
경제관이 뚜렷하며 가족을 적극 보살피는 내조의 여왕
Bravo Life
사회적 기여에 관심이 많고, 젊고 댄디한 감성을 잃지 않은 남성
Realist
건강과 웰빙, 일과 여가의 균형을 추구하는 이성적 남성
Gray Gentleman
필수적 소비에 집중하며 삶의 질을 중시하는 시니어 남성
줌마챌라
외모와 건강에 관심이 많은 사교적인 여성
Grace Woman
레저와 여가를 즐기며, 기부활동에 적극적인 여성
Silver Lady
건강을 유지하며 삶 자체를 즐기는 시니어 여성


ShinhanCard
23.5°


ShinhanCard
S-Line
Check S-Line
VISA

10강. 분할적 군집분석

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

비지도학습의 종류

분할적 군집분석

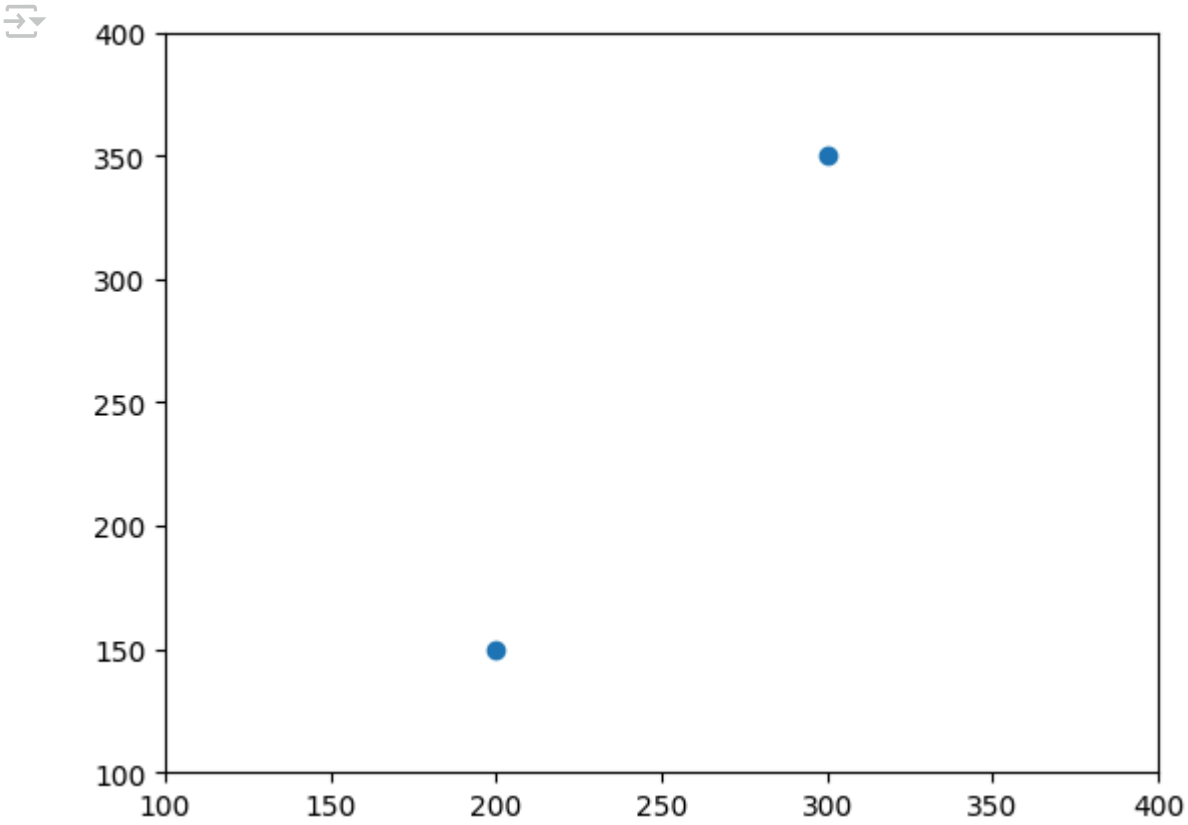
계층적 관계가 없는 다수의 군집들을 만드는 방법

- 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

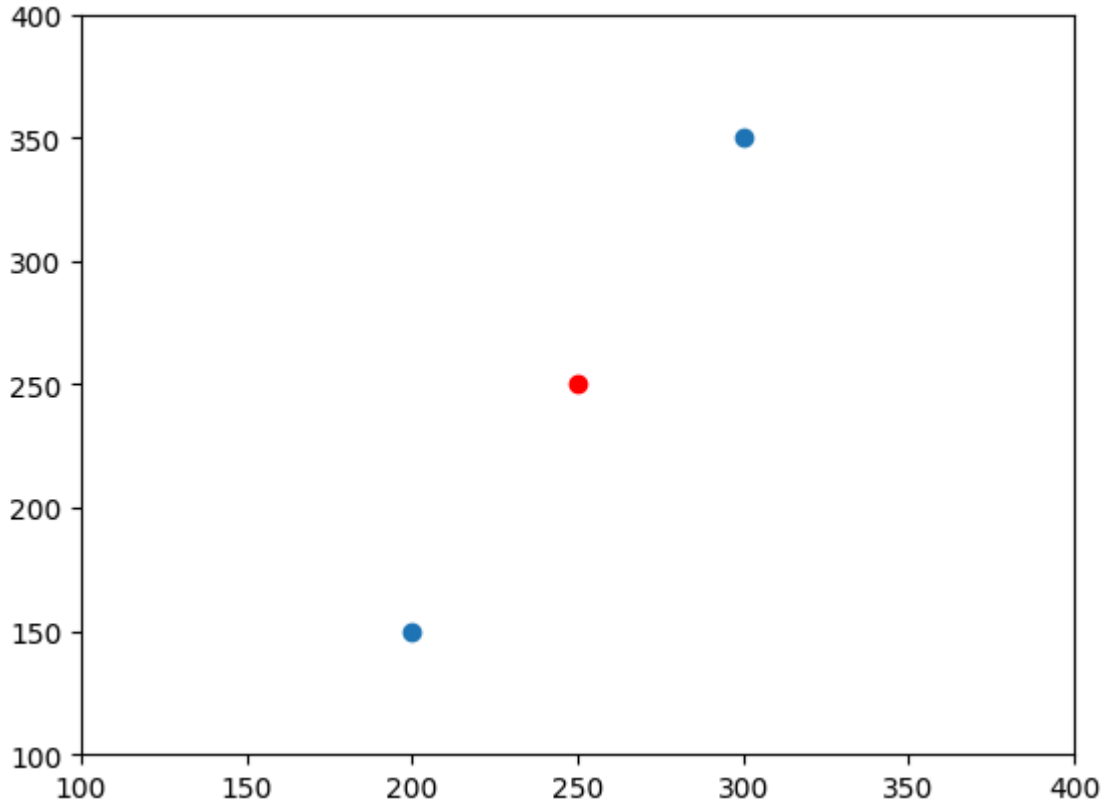
```
import matplotlib.pyplot as plt

x1, y1 = 300, 350
x2, y2 = 200, 150

plt.xlim(100, 400)
plt.ylim(100, 400)
plt.scatter([x1, x2], [y1, y2])
plt.show()
```




```
center_coordinate_x = (x1 + x2)/2
center_coordinate_y = (y1 + y2)/2
print(center_coordinate_x, center_coordinate_y)
```

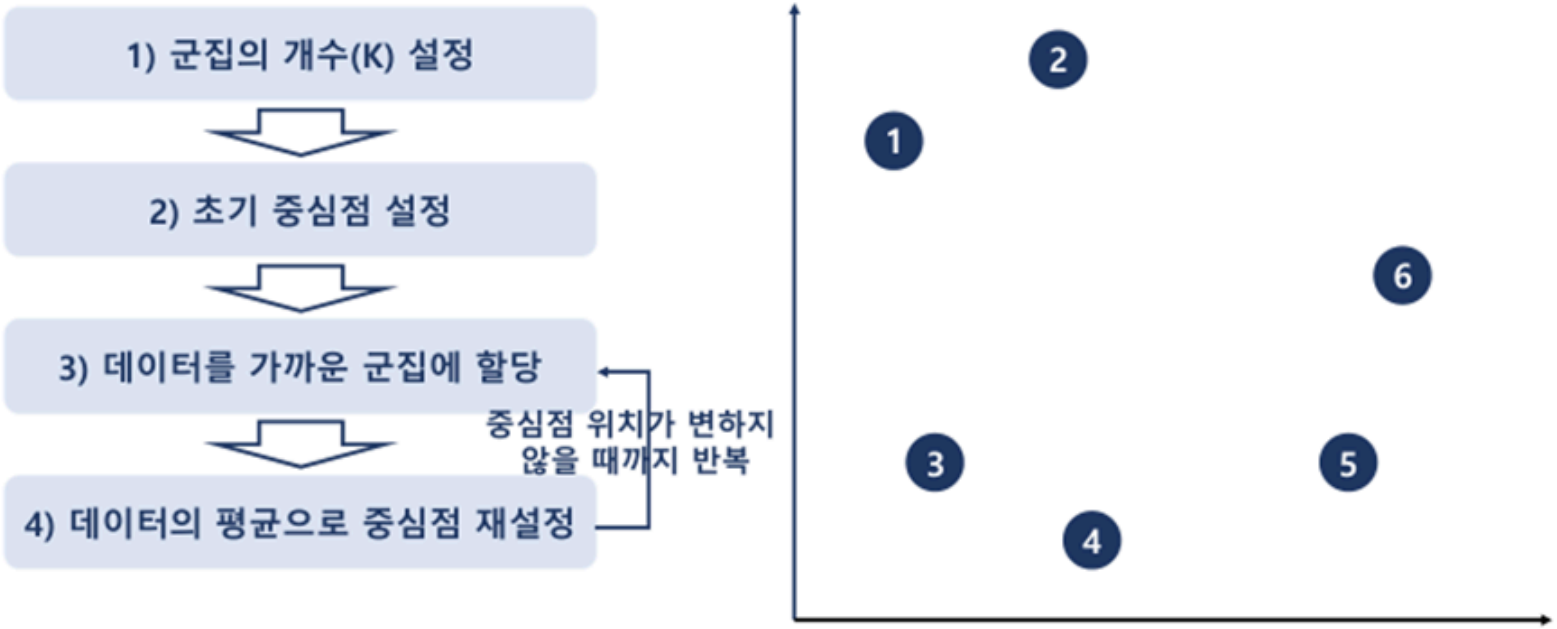
```
plt.xlim(100, 400)
plt.ylim(100, 400)
plt.scatter(center_coordinate_x, center_coordinate_y, color = 'red')
plt.scatter([x1, x2], [y1, y2])
plt.show()
```

```
250.0 250.0
```



K-Means 군집화

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



K-Means 군집화

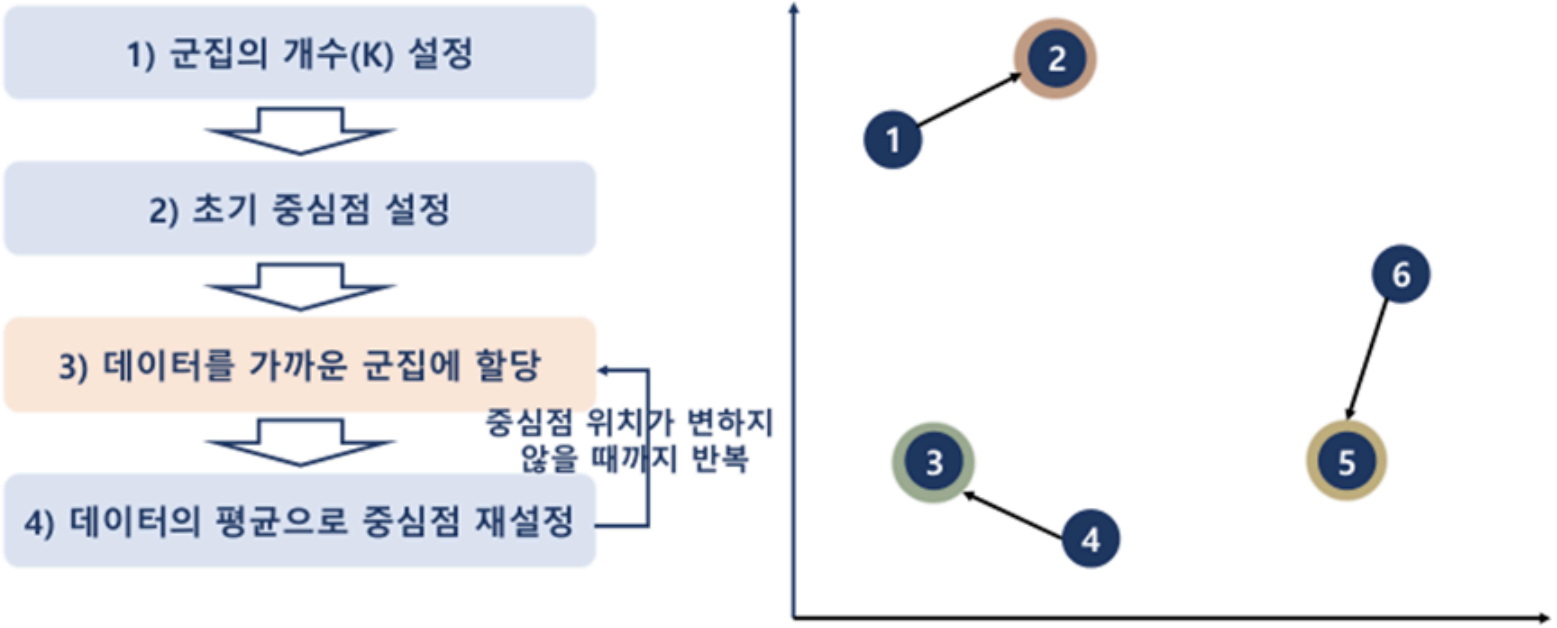
각 데이터와 중심점의 거리를 측정 후 가장 가까운 그룹에 할당하여 K개의 군집으로 묶는 방법

K-Means 군집화

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

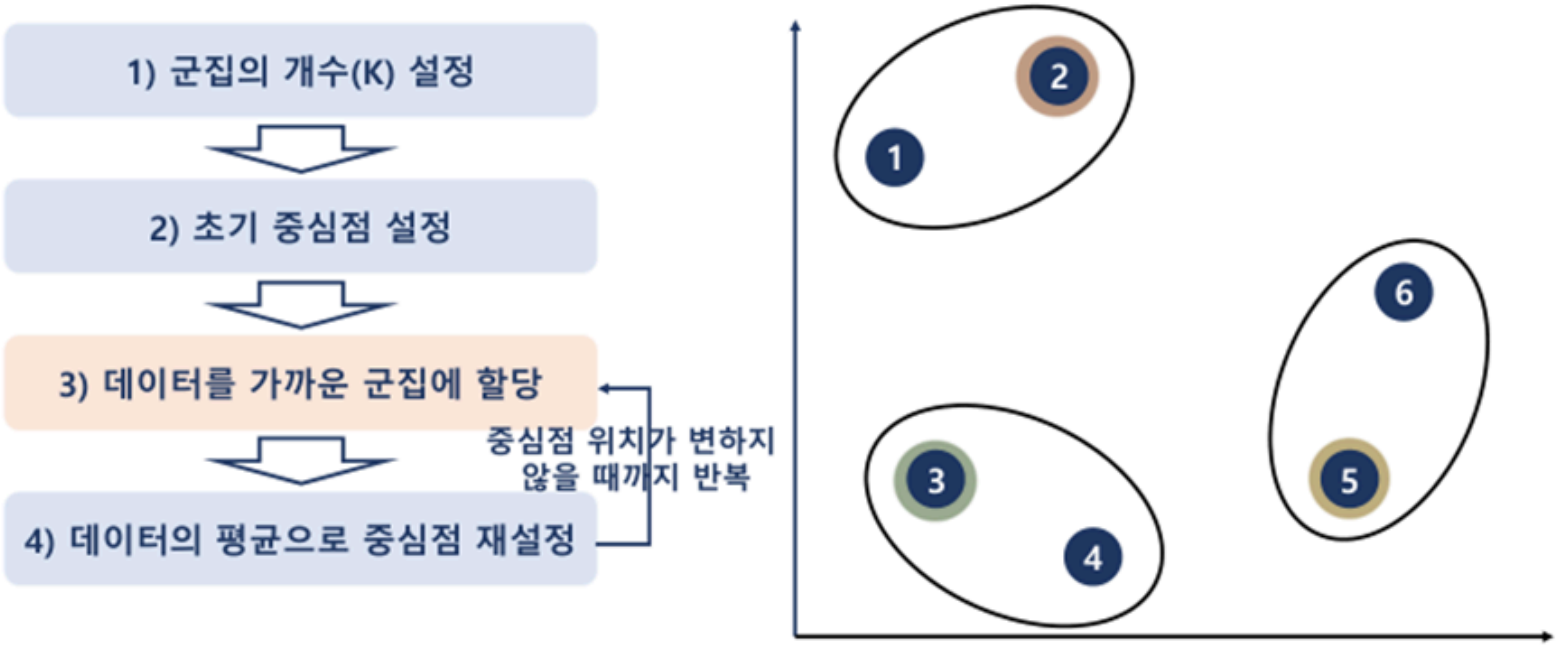
K-Means 군집화

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

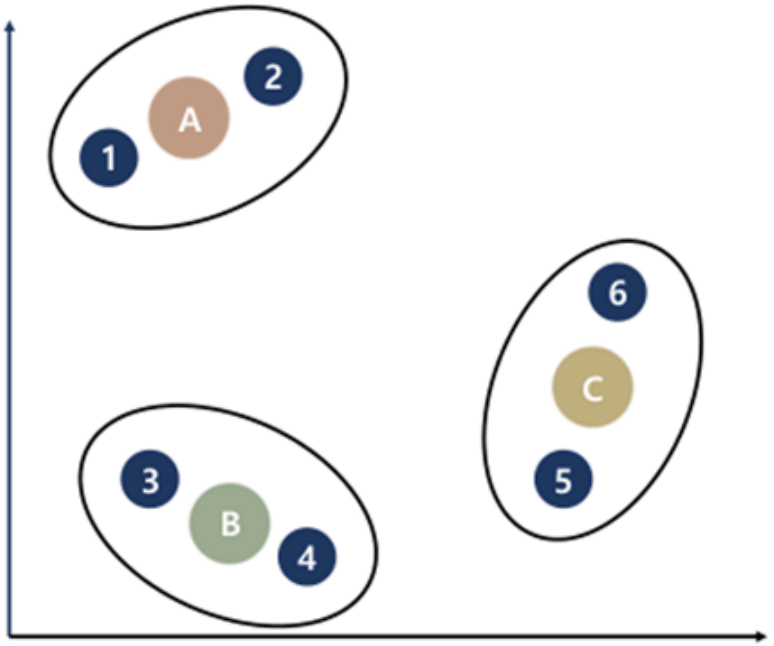
2) 초기 중심점 설정

3) 데이터를 가까운 군집에 할당

4) 데이터의 평균으로 중심점 재설정

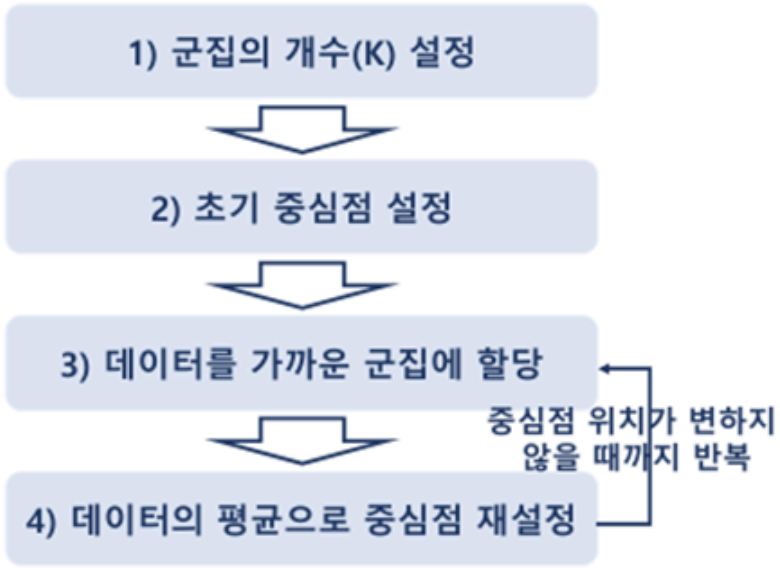
중심점 위치가 변하지 않을 때까지 반복

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?

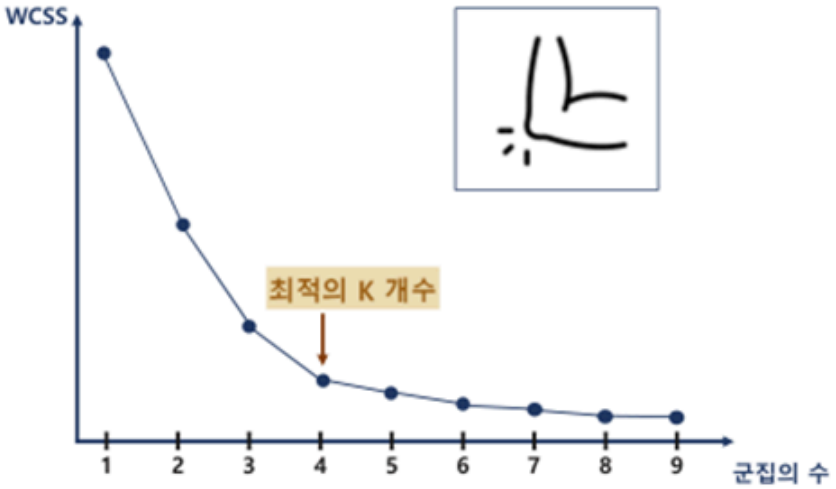
엘보우 기법(Elbow Method)

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

- 데이터 밀도: 정보 작성의 기초가 되는 데이터에서 단위 길이, 단위 면적, 단위 부피마다 기억된 문자 수. (어휘 혼종어 정보·통신)

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

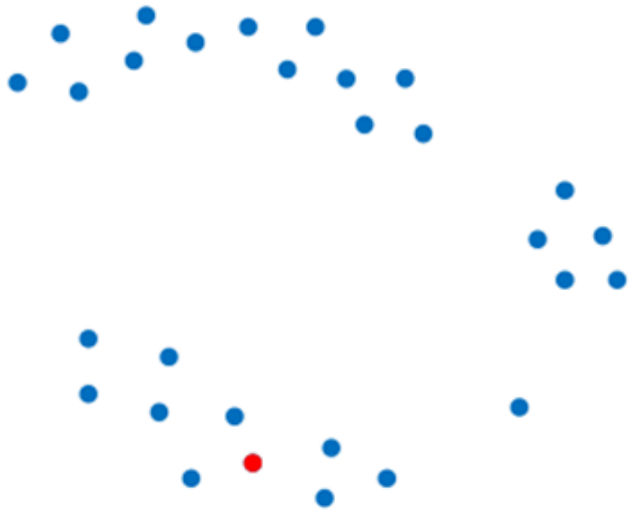
1) 포인트 임의로 선택

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

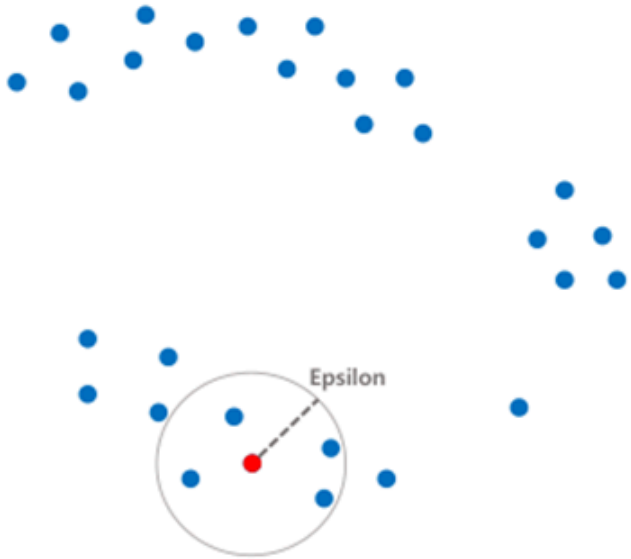
1) 포인트 임의로 선택

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

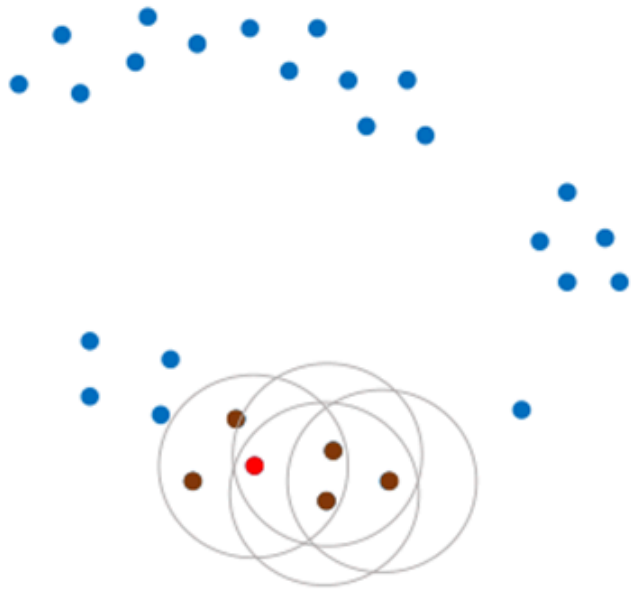
1) 포인트 임의로 선택

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

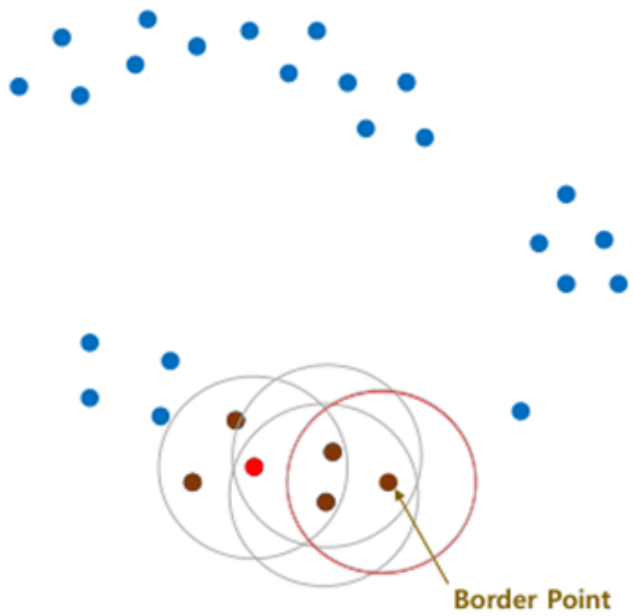
1) 포인트 임의로 선택

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

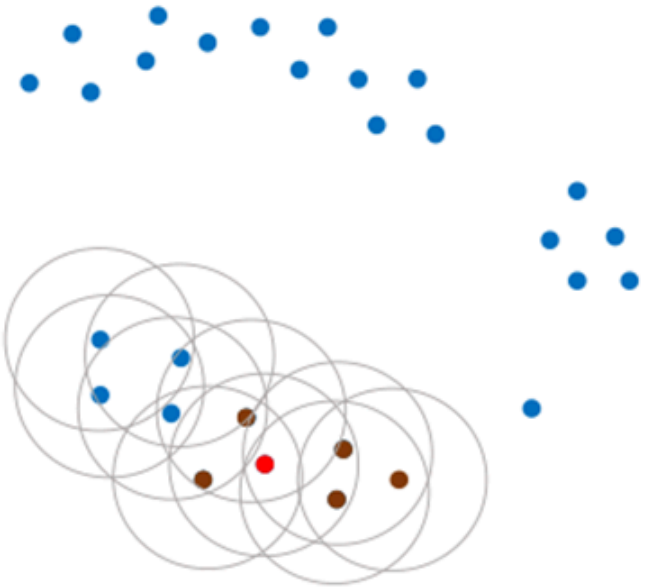
1) 포인트 임의로 선택

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

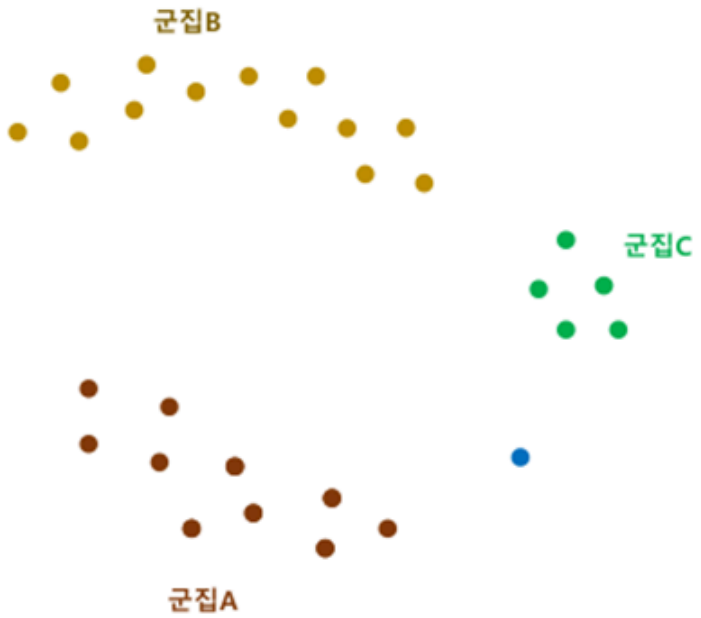
1) 포인트 임의로 선택

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

K-Means와 DBSCAN 비교