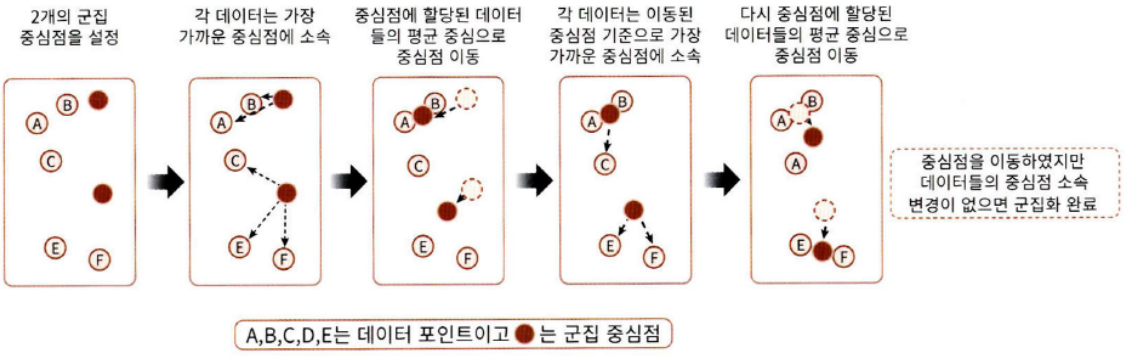
7. 군집화

작성일시	@2024년 12월 14일 오전 8:43
강의 번호	Euron
유형	스터디 그룹
복습	☐

1. K-평균 알고리즘 이해

- K-평균
  - **군집 중심점**을 선택, 중심에 가장 가까운 포인트를 선택하는 군집화 기법
    1. 군집화 개수만큼 중심을 임의의 위치에 놓음
    2. 각 데이터를 가장 가까운 중심점에 소속
    3. 중심점을 선택된 포인트의 평균 지점으로 이동
    4. 이동한 중심점에 대해 다시 데이터의 소속 변경함
    5. 중심점 이동해도 데이터의 소속 변경 없으면 멈춤



장점	단점
1. 가장 많이 활용됨	1. 속성(피처) 개수 많으면 군집화 정확도 떨어짐 - 차원 감소 필요
2. 쉽고 간결함	2. 반복 횟수 많을수록 수행 시간 길어짐
	3. 군집 수 결정하기 어려움

군집화 알고리즘 테스트 위한 데이터 생성

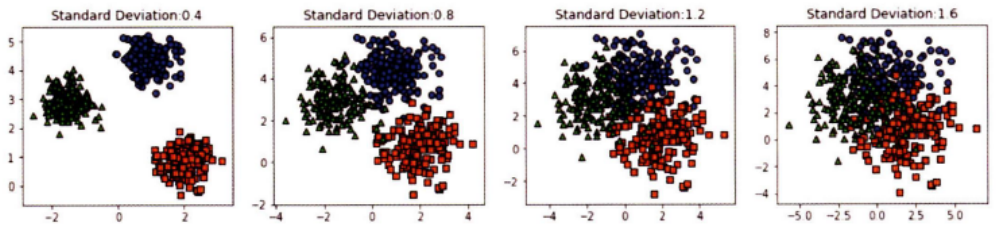
- 군집화용 데이터 생성기

API	특징
make_blobs( )	- 개별 군집의 중심점, 표준편차 제어 기능 - 피처, 타깃 데이터 세트 반환 형태: tuple
make_classification( )	노이즈 포함한 데이터 생성에 사용됨

- make_blobs( ) 파라미터

n_samples	생성할 데이터 개수
n_features	데이터 피처(속성) 개수 - 보통 시각화를 위해 2개 (x좌표, y좌표) 설정
centers	정수값 → 군집 개수 ndarray→ 개별 군집 중심점 좌표
cluster_std	군집 내 데이터 표준편차 - ex. [0.8, 1.2, 0.6] → 각 군집의 표준편차 0.8, 1.2, 0.6

  - cluster_std 로 분포도 조절

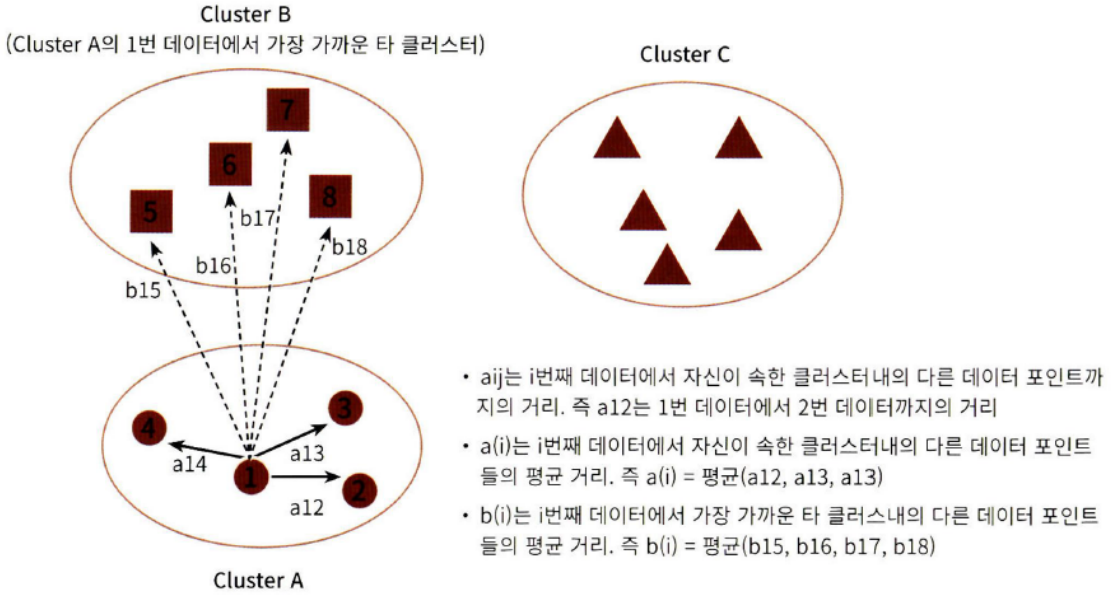


2. 군집 평가

- 대부분의 군집화 데이터는 비지도학습
  - (타킷 레이블 없음)
  - 성능 평가하기 어려움

실루엣 분석

- 각 군집 간의 거리가 얼마나 효율적으로 분리 되었는지
  - 같은 그룹은 가까이, 다른 군집은 멀리
  - 군집화 잘 되었음 = 각 그룹이 비슷한 여유 공간을 갖고 있을 것
- 실루엣 계수
  - 개별 데이터의 군집화 지표
  - 같은 군집 내 데이터와 얼마나 가깝고, 다른 군집 데이터와 얼마나 멀리 있는지 나타냄



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

- 0 < s(i) < 1: 1에 가까울수록 근처 군집과 떨어져 있고, 0에 가까울수록 근처 군집과 가까움
- -1 < s(i) < 0: 다른 군집에 데이터 할당됨

메서드	설명
sklearn.metrics.silhouette_samples(X, labels, metric = 'euclidean', **kwds)	X: 피처 데이터, label: 피처 데이터 속한 군집 레이블 값 → 반환값: 각 데이터 실루엣 계수
sklearn.metrics.silhouette_score(X, labels, metric = 'euclidean', sample_size = None, **kwds)	X: 피처 데이터, label: 피처 데이터 속한 군집 레이블 값 → 반환값: 전체 데이터 실루엣 계수 평균 **np.mean(silhouette_samples())** - 절대적인 평가 척도는 아니지만, 반환값 높을수록 군집화 잘된 것으로 판단 (특정 군집만 평균이 높을 수 있음)

- 좋은 군집화 기준
  1. silhouette_score( ) 1에 가까울수록 좋음
     - 0~1 사이 값
  2. 개별 군집 실루엣 계수 평균이 전체 실루엣 계수 평균과 크게 다르지 않아야 함
     - 특정 군집의 실루엣 계수 평균만 높고 나머지는 낮으면 좋은 군집화 조건 아님

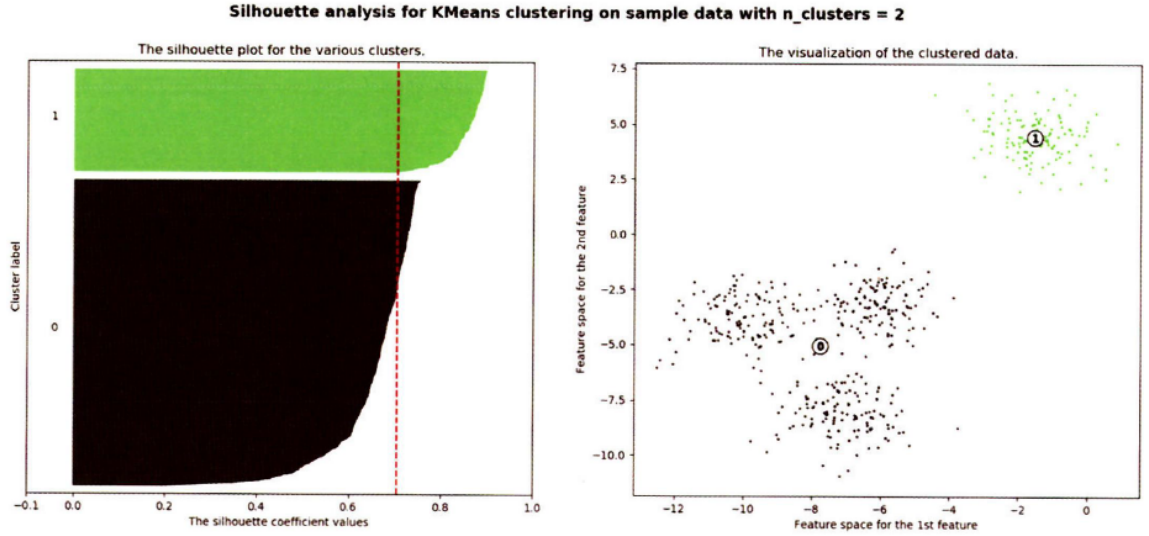
장점	단점
직관적으로 이해하기 쉬움	각 데이터별 다른 데이터와의 거리 반복 계산 → 데이터 수 늘면 수행시간 늘어남

군집별 평균 실루엣 계수 시각화로 군집 개수 최적화

- 군집별 적당히 분리된 거리 & 군집 내 데이터 뭉쳐 있을 때 → 적절한 군집 개수 설정됨!

1. 군집 2개 (silhouette_score( ) = 0.704)
    - 1번 군집 - 대부분 평균 실루엣 계수 이상

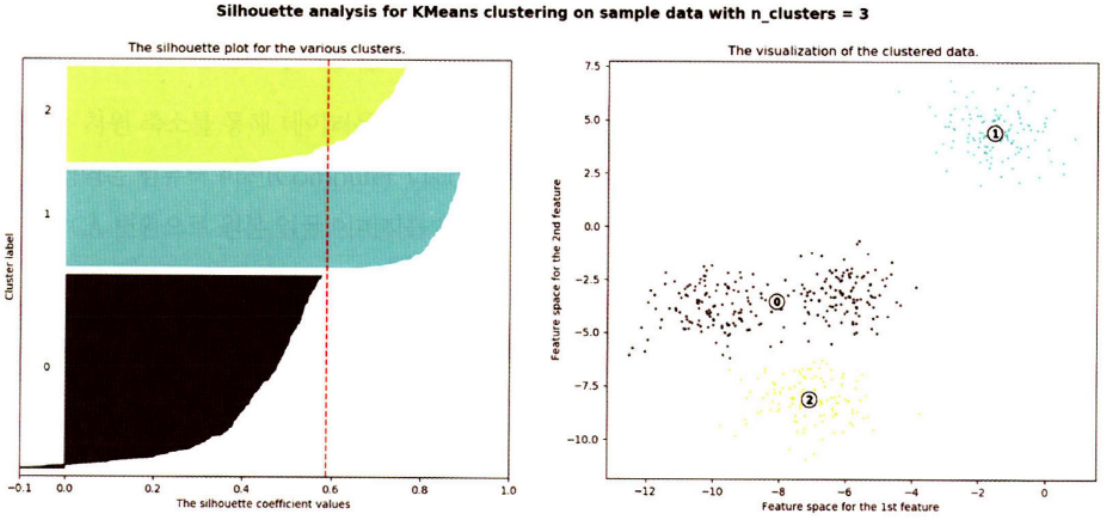
      2번 군집 - 평균 실루엣 계수보다 적은 데이터 많음
    - 1번 군집은 서로 잘 뭉쳐 있지만, 2번 군집은 내부 데이터끼리 떨어져 있음



군집이 2개일 경우 평균 실루엣 계수 값: 0.704

2. 군집 3개 (silhouette_score( ) = 0.588)
    - 1, 2번 군집 - 평균 실루엣 계수보다 높은 값

      0번 군집 - 모든 데이터 평균 실루엣 계수보다 낮음
    - 0번 군집은 내부 데이터끼리 떨어져 있고, 2번 군집과 가까움

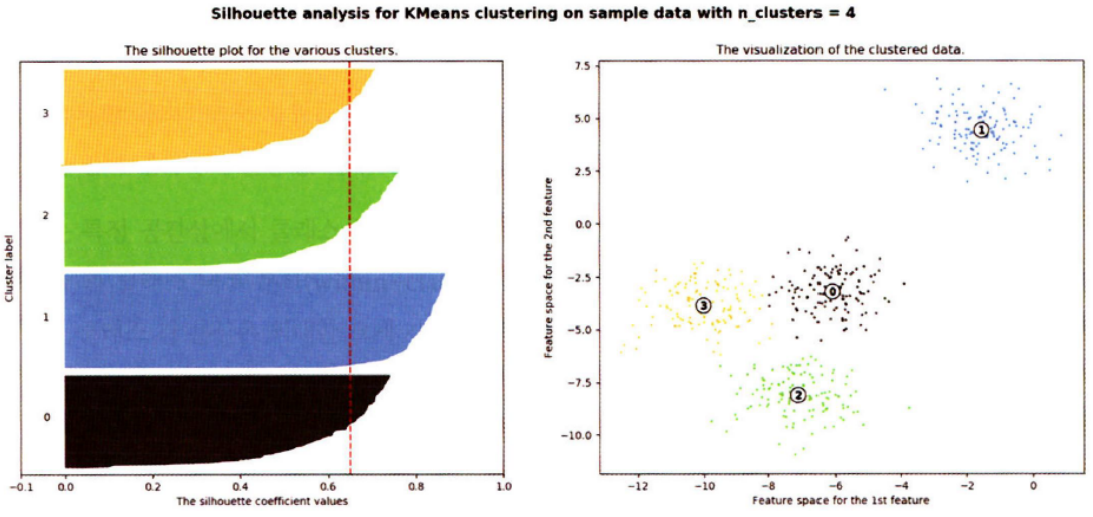


군집이 3개일 경우 평균 실루엣 계수 값: 0.588

3. 군집 4개 (silhouette_score( ) = 0.65)
    - 1번 군집 - 모든 데이터가 평균 실루엣 계수보다 높은 값

      0, 2번 군집 - 절반 이상 평균보다 높음

      3번 군집 - 1/3 데이터 정도 평균보다 높음
    - 군집 2개일 때보다 평균 실루엣 계수가 낮지만, 이상적인 군집화 개수



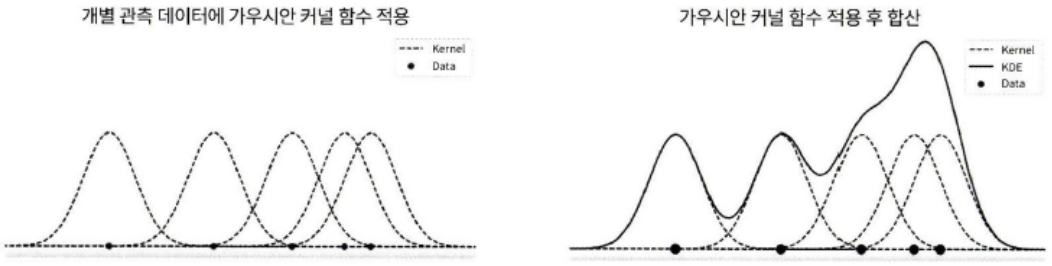
군집이 4개일 경우 평균 실루엣 계수 값: 0.65

3. 평균 이동

개요

- 평균 이동 군집화

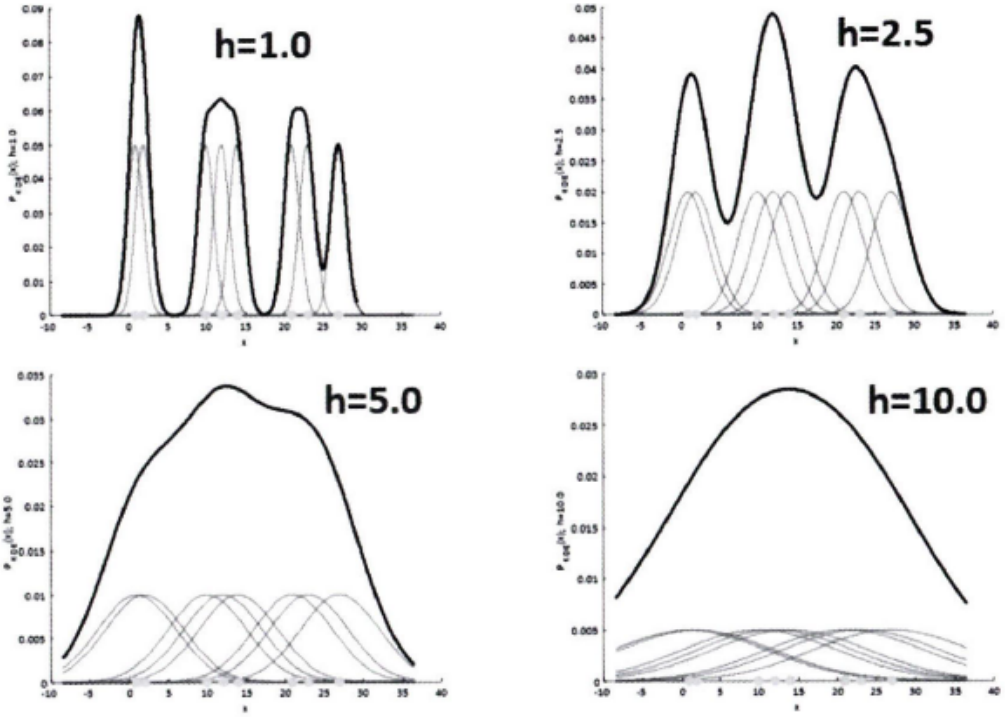
- **데이터 밀도**가 가장 높은 곳으로 중심 이동
  - cf) K-평균: 데이터 평균 거리로 중심 이동
- KDE 이용해서 **확률 밀도 함수** 추정
  1. 관측 데이터에 커널 함수 적용한 값 모두 더함
  2. 데이터 건수로 나눔 → 확률 밀도 함수 추정



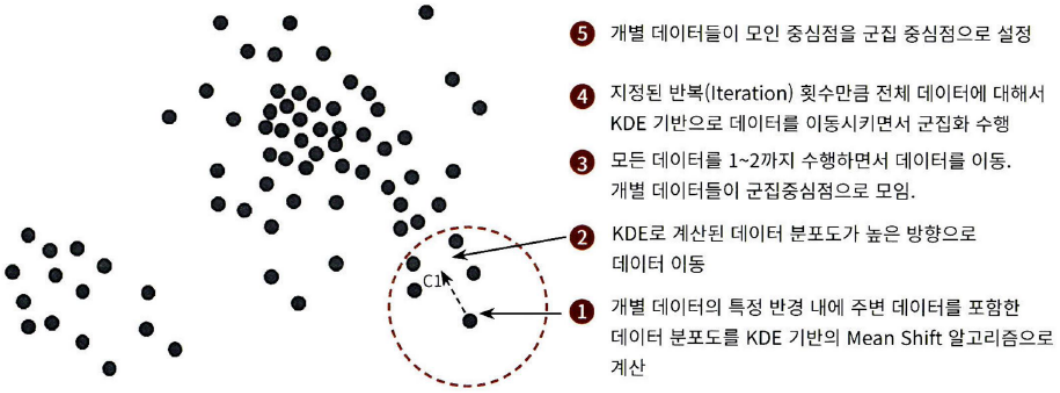
$$\text{KDE} = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

  - K: 커널함수(ex. 가우시안 분포 함수), x: 확률변수값, xi: 관측값, h: 대역폭(bandwidth)
  - 대역폭: KDE 형태 평활화(부드럽게)하는 데 적용됨
    - h = 1.0 일때, 변동성이 커서 과적합하기 쉬움

      h = 10.0 일때, 지나치게 단순화됨 (과소적합하기 쉬움)
    - 일반적으로 대역폭 클수록 군집 중심점 적음, 대역폭 적을수록 군집 중심점 많음
    - 군집의 개수 지정 x, 대역폭 크기에 따라 수행



- 특정 데이터와 주변 데이터의 거리 입력 → 반환값 업데이트하면서 개별 데이터 이동



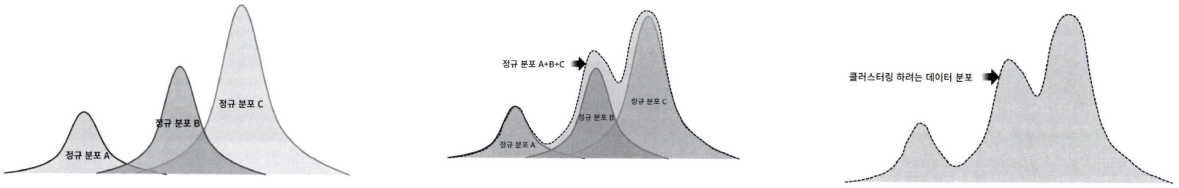
장점	단점
1. 유연한 군집화 가능 - 특정 형태/분포도 데이터 세트 가정 x	1. 수행 시간 길다
2. 이상치 영향력 크지 않음	2. bandwidth 크기에 따른 영향도 매우 큼
3. 군집 개수 정할 필요 없음	

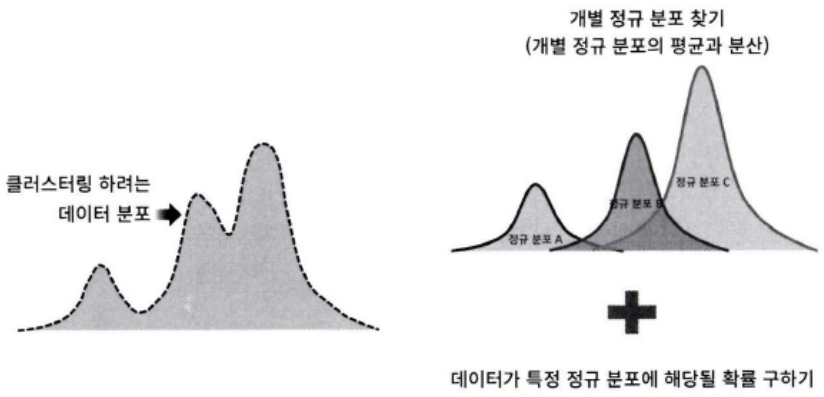
4. GMM (Gaussian Mixture Model)

GMM

- 가정: 군집화 적용 데이터가 여러 개의 가우시안 분포를 가진 데이터들이 섞여 생성됨

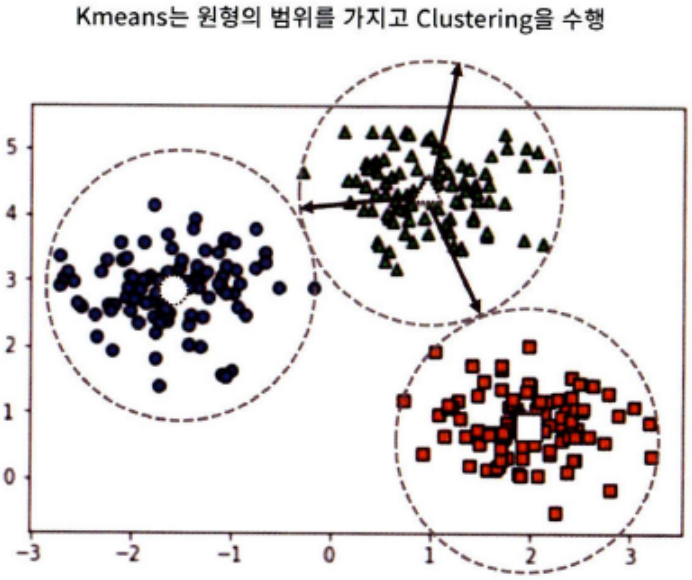
    - 가우시안 분포 = 정규 분포
- 섞인 데이터 분포에서 개별 가우시안 분포 추출
    - 데이터 세트를 구성하는 여러 정규분포 곡선 추출 → 개별 데이터 어디에 속하는지 결정
    - GMM 모수 추정
        - 정규분포의 평균, 분산
        - 각 데이터가 어떤 정규 분포에 해당하는지 확률





GMM과 K-평균 비교

- K-평균
    - 원형 범위 갖는 데이터일 수록 군집화 효율 높아짐
        - 원형 범위로 퍼져 있지 않으면 군집화 수행 잘 못함 (ex. 타원형 데이터)
    - 군집 내 데이터 뭉치게 유도한 후 평균 적용함
        - cluster_std 낮게 설정 → 데이터 원형 형태로 분산



5. DBSCAN

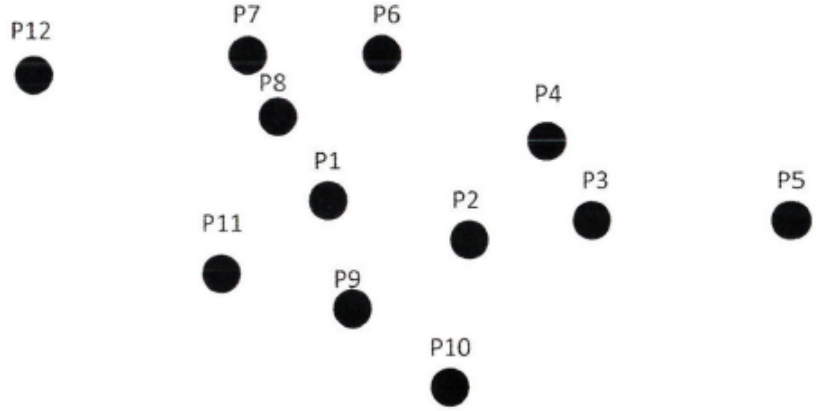
DBSCAN

- **밀도** 기반 군집화
  - 밀도 기준 충족 시키는 핵심 포인트 (core) 연결하며 군집화
- 기하학적으로 복잡한 데이터 효과적으로 군집화
- 주요 파라미터

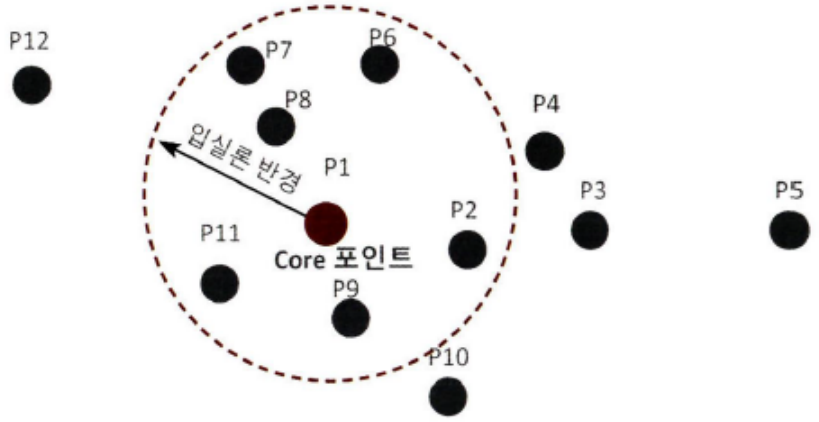
파라미터명	설명
입실론 주변 영역 (epsilon)	개별 데이터 중심으로 입실론 반경 갖는 원형 영역
최소 데이터 수(min points)	입실론 주변 영역에 포함된 데이터 개수 (개별 데이터 제외)

핵심 포인트 (core)	입실론 주변 영역 내 최소 데이터 수 이상의 (개별 데이터 제외) 데이터 갖고 있을 때
이웃 포인트 (neighbor)	입실론 주변 영역 내 개별 데이터 외의 데이터 (타 데이터)
경계 포인트 (border)	입실론 주변 영역 내 최소 데이터 수를 충족하지 않고, 핵심 포인트를 이웃 포인트로 갖는 데이터
잡음 포인트 (noise)	입실론 주변 영역 내 최소 데이터 수를 충족하지 않고, 핵심 포인트를 이웃 포인트로 갖지 않는 데이터

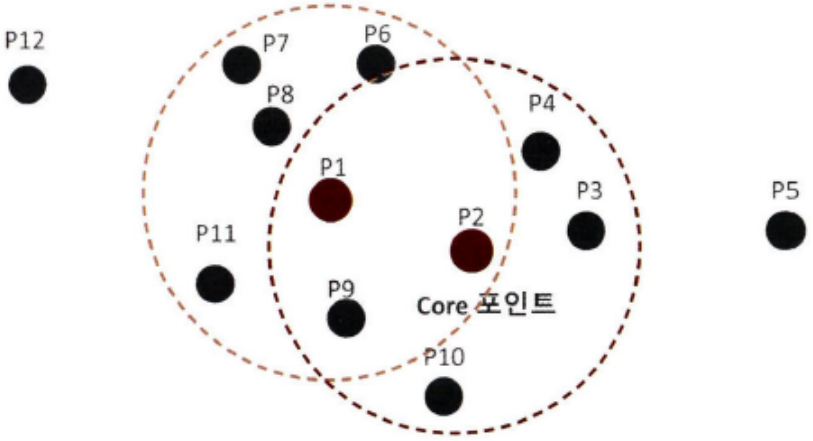
- DBSCAN 적용



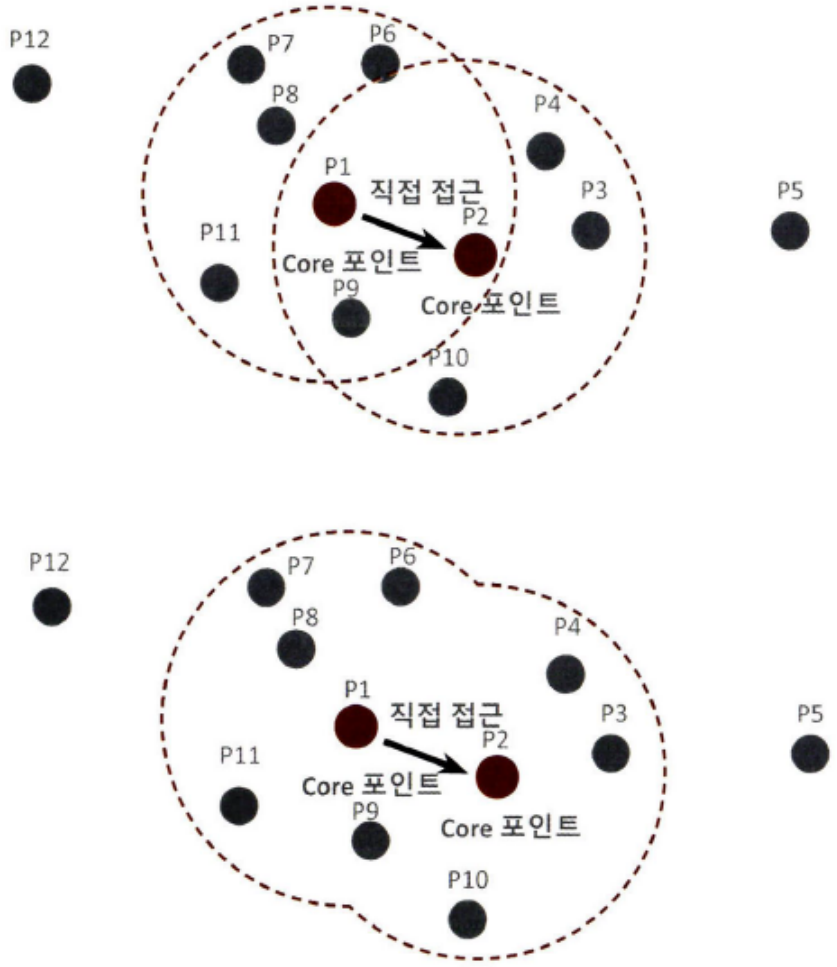
1. P1 데이터: 핵심 포인트



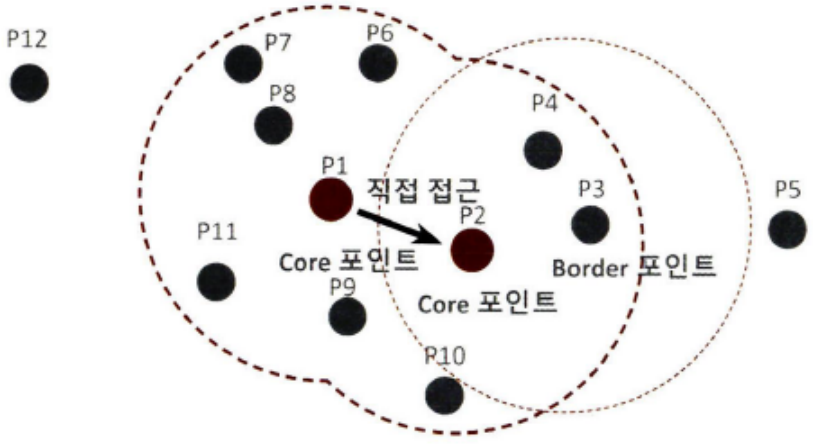
2. P2 데이터: 핵심 포인트



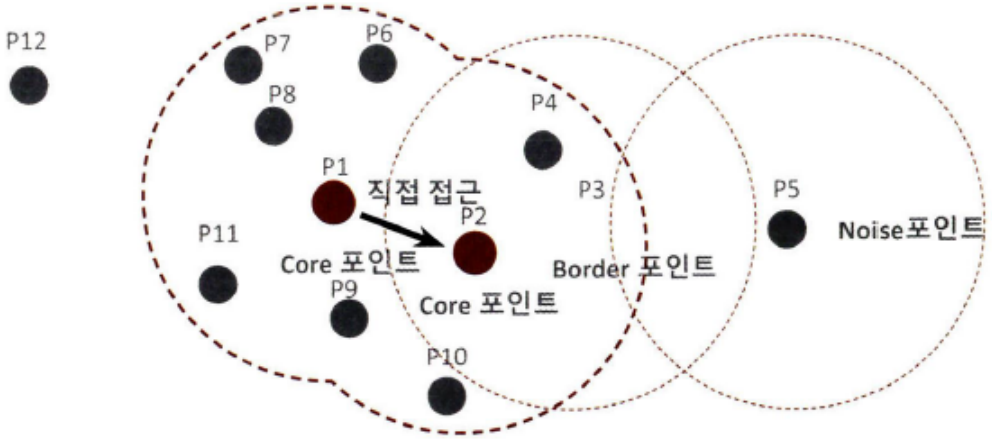
3. P1 (core point)의 이웃 포인트 P2도 핵심 포인트 → 직접 접근 가능 ⇒ 군집화 (군집 확장)



4. P3 데이터: 경계 포인트
    - 군집의 외곽 형성



5. P5 데이터: 잡음 포인트



- DBSCAN 클래스

파라미터명	설명
eps	입실론 주변 영역 반경
min_samples	핵심 포인트 요건을 충족하기 위해 입실론 주변 영역에 포함되어야 할 최소 데이터 수 (개별 데이터 포함)

6. 군집화 실습 - 고객 세그먼테이션

고객 세그먼테이션 정의, 기법

- 고객 세그먼테이션
  - 다양한 기준으로 고객 분류하는 기법
    - 어떤 상품?
    - 얼마나 많은 비용 지불?
    - 얼마나 자주 사용?
  - 목표
    - 타켓 마케팅 → 맞춤형 서비스 제공
- RFM 기법
  1. R (recency): 가장 최근 상품 구입일부터 오늘까지 기간
  2. F (frequency): 구매 횟수
  3. M (monetary value): 구매 금액