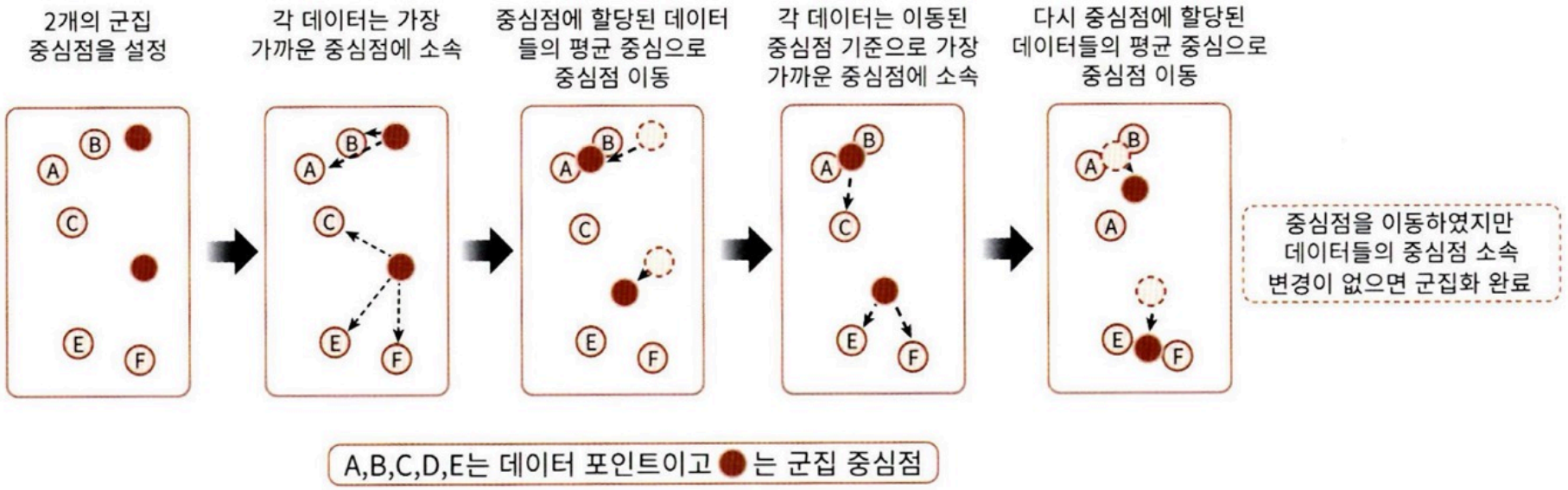
Chapter 07. 군집화

01. K-평균 알고리즘 이해

K-평균: 군집 중심점이라는 특정한 임의의 지점을 선택해 해당 중심에 가장 가까운 포인트들을 선택하는 군집화 기법

- 군집 중심점: 선택된 포인트의 평균 지점으로 이동하고 이동된 중심점에서 다시 가까운 포인트를 선택, 다시 중심점을 평균 지점으로 이동하는 프로세스를 반복적으로 수행



- 장점: 일반적인 군집화에서 가장 많이 활용, 알고리즘이 쉽고 간결
- 단점: 속성의 개수가 많으면 정확도가 떨어짐, 반복 횟수가 많으면 수행 시간이 느려짐, 몇 개의 군집을 선택해야 할지 가이드 하기가 어려움

KMeans 클래스: K-평균을 구현하기 위해 사이킷런에서 제공

- 주요 파라미터
    - n_clusters: 가장 중요한 파라미터, 군집 중심점의 개수
    - init: 군집 중심점의 좌표를 설정할 방식
    - max_iter: 최대 반복 횟수
- 주요 속성
    - labels_: 각 데이터 포인트가 속한 군집 중심점 레이블
    - cluster_centers_: 각 군집 중심점 좌표

군집화용 데이터 생성기: make_blobs(), make_classification()

- make_blobs(): 개별 군집의 중심점과 표준 편차 제어 기능이 추가
    - 호출 파라미터: n_samples, n_features, centers, cluster_std
- make_classification(): 노이즈를 포함한 데이터를 만드는 데 유용

02. 군집 평가

실루엣 분석: 각 군집 간의 거리가 얼마나 효율적으로 분리되어 있는지를 나타냄

- 효율적으로 잘 분리 -> 다른 군집과의 거리는 떨어져 있고 동일 군집끼리의 데이터는 가깝게 잘 뭉쳐 있음
- 실루엣 계수를 기반으로 함: 개별 데이터가 가지는 군집화 지표
    - a(i): 군집 내 다른 데이터 포인트와의 거리를 평균한 값
    - b(i): 가장 가까운 군집과의 평균 거리
    - s(i): (b(i)-a(i)) / (max(a(i),b(i)))

- 좋은 군집화의 조건
  - 전체 실루엣의 평균값이 0에서 1사이의 값을 가짐
  - 개별 군집의 평균값의 편차가 크지 않아야 함

03. 평균 이동

평균 이동: 중심을 군집의 중심으로 움직이면서 군집화를 수행, 밀도가 가장 높은 곳으로 이동

- 확률 밀도 함수가 피크인 점을 군집 중심점으로 선정, 확률 밀도 함수를 찾기 위해 KDE 이용
  - KDE: 관측된 데이터 각각에 커널 함수를 적용한 값을 모두 더한 뒤 데이터 건수로 나눠 확률 밀도 함수를 추정
    - 대표적인 커널 함수: 가우시안 분포 함수

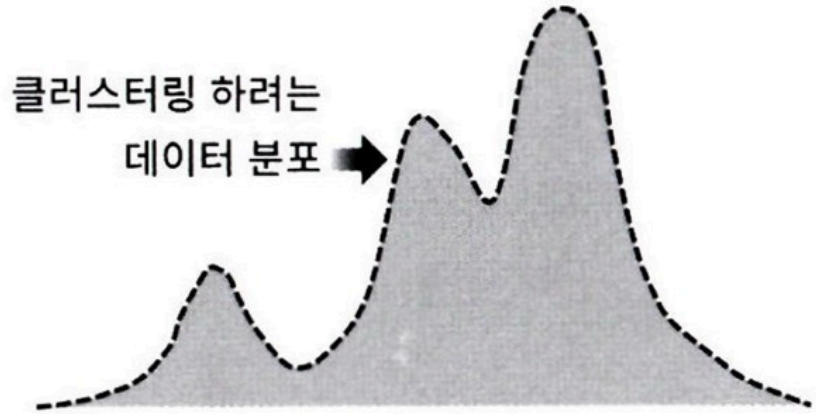
$$\mathrm{KDE} = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

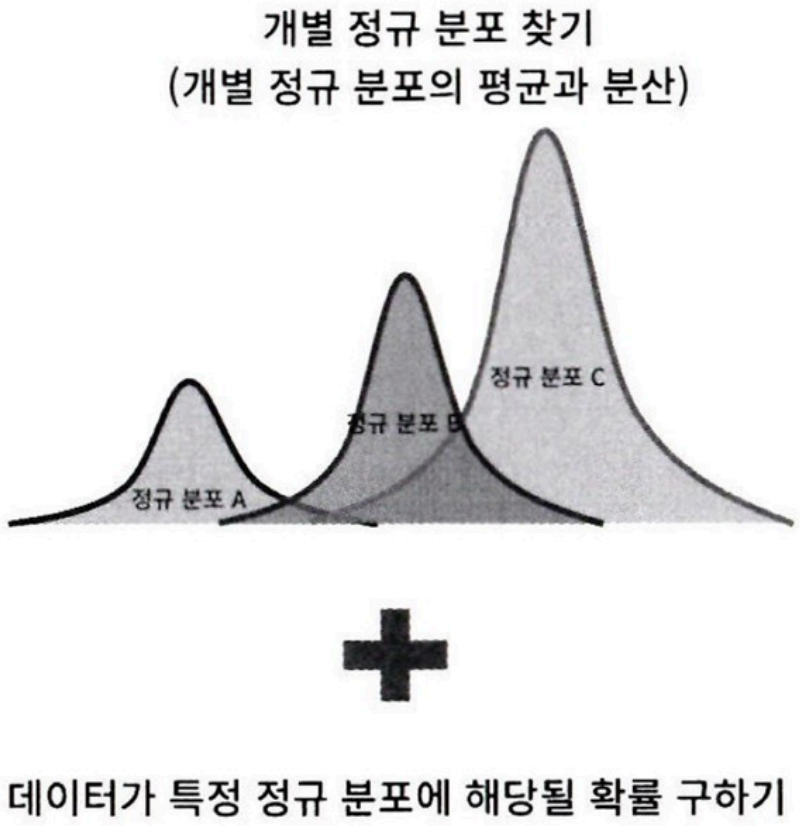
- K는 커널 함수, x는 확률 변숫값, xi는 관측값, h는 대역폭
  - 대역폭은 형태를 부드러운 형태로 평활화하는 데 적용
    - 작은 h 값: 과적합 가능성 있음
    - 큰 h 값: 과소적합 가능성 있음
- 사이킷런에서는 평균 이동 군집화를 위해 MeanShift 클래스 제공
- 평균 이동의 장단점
  - 장점: 유연한 군집화 가능, 이상치의 영향력이 크지 않음, 미리 개수를 정할 필요 X

- 단점: 수행 시간이 오래 걸림, bandwidth의 크기에 따른 군집화 영향도가 매우 큼

04. GMM(Gaussian Mixture Model)

GMM: 데이터가 여러 개의 가우시안 분포를 가진 데이터 집합들이 섞여서 생성된 것이라는 가정하에 군집화 수행





모수 추정: 대표적으로 개별 정규 분포의 평균과 분산/ 각 데이터가 어떤 정규 분포에 해당되는지의 확률을 추정

- GMM vs K-평균
    - K-평균: 원형의 범위에서 군집화 수행
    - GMM: 군집의 중심 좌표를 구할 수 없음

05. DBSCAN

DBSCAN: 밀도 차이 기반 알고리즘, 복잡한 기하학적 분포도를 가진 데이터 세트에 대해서도 군집화를 잘 수행

- 주요 파라미터: 입실론 주변 영역(epsilon), 최소 데이터 개수(min points)
- 데이터 포인트: 핵심 포인트, 이웃 포인트, 경계 포인트, 잡음 포인트
- 주요 초기 파라미터: eps, min_samples

06. 군집화 실습 - 고객 세그먼테이션

고객 세그먼테이션: 다양한 기준으로 고객을 분류하는 기법

- 타깃 마케팅이 주요 목표, 고객을 여러 특성에 맞게 세분화해서 그 유형에 따라 맞춤형 마케팅이나 서비스를 제공
- RFM 기법: 기본적인 고객 분석 요소
    - R(RECENCY): 가장 최근 상품 구입 일에서 오늘까지의 기간
    - F(FREQUENCY): 상품 구매 횟수
    - M(MONETARY VALUE): 총 구매 금액