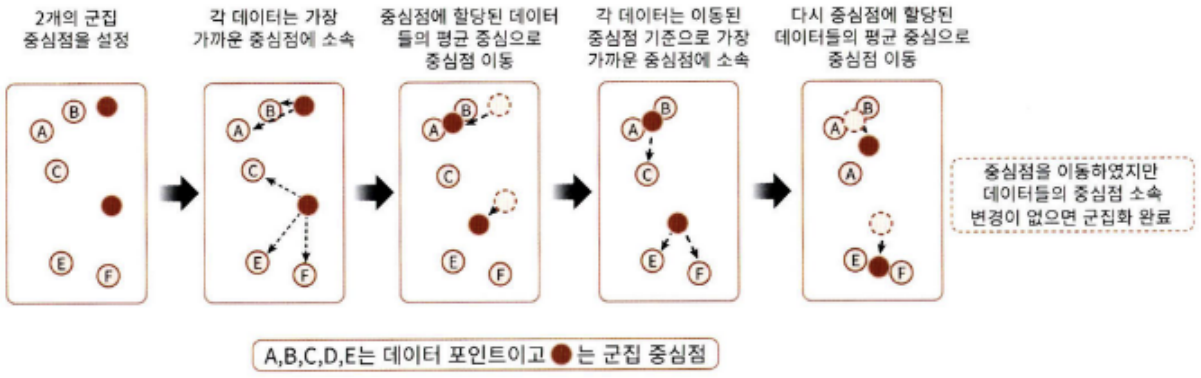
K-평균 알고리즘 이해

- **K-평균 알고리즘** : *군집 중심*이라는 특정한 임의의 지점을 선택해 해당 중심에 가장 가까운 포인트들을 선택하는 군집화 기법



- **K-평균의 장점**
  - 일반적인 군집화에서 가장 많이 활용되는 알고리즘
  - 알고리즘이 쉽고 간결
- **K-평균의 단점**
  - 거리 기반 알고리즘으로 속성의 개수가 매우 많은 경우 군집화 정확도가 떨어짐
  - 반복 횟수가 많을 경우 수행 시간이 매우 느려짐
  - 몇 개의 군집을 선택해야 할지 가이드하기 어려움

사이킷런 KMeans 클래스 소개

- **주요 파라미터**
  - n_clusters : 군집화할 개수
  - int : 군집 중심점의 좌표를 설정할 방식, 일반적으로는 k-means++ 방식으로 설정
  - max_iter : 최대 반복 횟수
- **K-Means 객체의 주요 속성**
  - labels_ : 각 데이터 포인트가 속한 군집 중심 레이블
  - clusters*centers* : 각 군집 중심점 좌표

군집화 알고리즘 테스트를 위한 데이터 생성

- make_blobs() : 개별 군집의 중심점과 표쥰 편차 제어 기능 추가
- make_classification() : 노이즈를 포함한 데이터를 만드는 데 유용

군집 평가

실루엣 분석의 개요

- **실루엣 분석** : 각 군집 간의 거리가 얼마나 효율적으로 분리돼 있는지 나타냄
- 효율적으로 잘 분리됐다는 것
  - 다른 군집과의 거리는 떨어져 있음
  - 동일 군집끼리의 데이터는 서로 가깝게 잘 뭉쳐 있음
- **실루엣 계수** : 개별 데이터가 가지는 군집화 지표
  - -1에서 1 사이의 값을 가짐
  - 1로 가까워짐 -> 근처의 군집과 더 멀리 떨어져 있음
  - 0에 가까움 -> 근처의 군접과 가까워짐
  - • : 아예 다른 군집에 데이터 포인트가 할당됨

군집별 평균 실루엣 계수의 시각화를 통한 군집 개수 최적화 방법

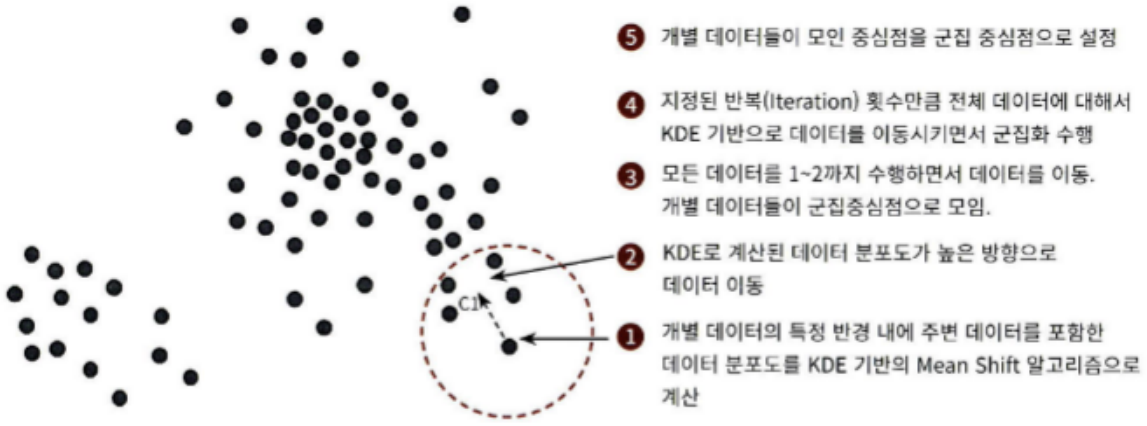
- 개별 군집별로 *적당히 분리된 거리*를 유지하면서도 군집 내의 데이터가 *서로 뭉쳐 있는 경우*에 적절한 군집 개수가 설정됐다고 판단할 수 있음

평균 이동

평균 이동의 개요

- **평균 이동**
  - 중심을 군집의 중심으로 지속적으로 움직이면서 군집화를 수행
  - 중심을 데이터가 모여 있는 밀도가 가장 높은 곳으로 이동시킴
- **군집 중심점**
  - 데이터 포인트가 모여있는 곳이라는 생각에서 착안한 것

- 확률 밀도 함수를 사용 -> KDE 이용



- **KDE** : 커널 함수를 통해 어떤 변수의 확률 밀도 함수를 추정하는 대표적인 방법

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

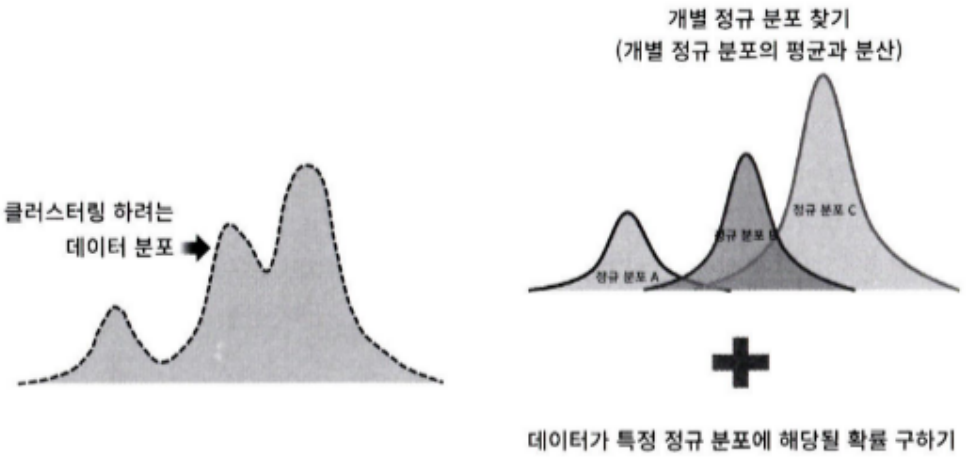
- **PDF** : 확률 변수의 분포를 나타내는 함수(정규분포 함수, 감마 분포, t-분포)
- 대역폭 h에 따라 확률 밀도 추정 성능을 크게 좌우함.
  - 작은 h값 -> 변동성이 큰 방식으로 확률 밀도 함수 추정 -> *과적합* 쉬움
  - 큰 h값 -> 단순화된 방식으로 확률 밀도 함수 추정 -> *과소적합* 쉬움
- -> h 값을 결정하는 것 : KDE 기반의 평균 이동 군집화에서 매우 중요

GMM(Gaussian Mixture Model)

GMM 소개

- **GMM 군집화** : 군집화를 적용하고자 하는 데이터가 여러 개의 가우시안 분포를 가진 데이터 집합들이 섞여서 생성된 것이라는 가정하에 군집화를 수

행하는 방식



- **모수 추정**
  - 개별 정규 분포의 평균과 분산 추정
  - 각 데이터가 어떤 정규 분포에 해당되는지의 확률
- -> EM 알고리즘 적용

GMM과 K-평균의 비교

- KMeans : 원형의 범위에서 군집화 수행 -> 데이터 세트가 원형의 범위를 가질 수록 효율 증가
- GMM : KMeans보다 유연하게 다양한 데이터 세트에 잘 적용될 수 있음, but 수행시간 오래 걸림

DBSCAN

DBSCAN 개요

- 데이터 밀도 차이를 기반한 알고리즘 사용 -> 복잡한 기하학적 분포도를 가진 데이터 세트에 대해서도 군집화를 잘 수행
- **입실론 주변 영역** : 개별 데이터를 중심으로 입실론 반경을 가지는 원형의 영역
- **최소 데이터 개수** : 개별 데이터의 입실론 주변 영역에 해당되는 타 데이터의 개수
- **핵심 포인트** : 주변 영역에 최소 데이터 개수 이상의 타 데이터를 가지고 있을 경우 해당 데이터를 핵심 포인트로 정의
- **이웃 포인트** : 주변 영역 내에 위치한 타 데이터

- **경계 포인트** : 주변 영역 내에 최소 데이터 개수 이상의 이웃 포인트를 가지고 있지 않지만 핵심 포인트를 이웃 포인트로 가지고 있는 데이터
- **잡음 포인트** : 최소 데이터 이상의 이웃 포인트를 가지고 있지 않으며, 핵심 포인트도 이웃 포인트로 가지고 있지 않은 데이터
- -> 핵심 포인트를 연결하면서 군집화를 구성

군집화 실습 - 고객 세그먼테이션

고객 세그먼테이션의 정의와 기법

- **고객 세그먼테이션** : 다양한 기준으로 고객을 분류하는 기법
- 주요 목표 : 타깃 마켓팅 -> 고객을 여러 특성에 맞게 세분화해서 그 유형에 따라 맞춤형 마케팅이나 서비스를 제공하는 것