* **군집화(Clustering)**

데이터 포인트들을 별개의 군집으로 그룹화 하는 것을 의미함.

유사성이 높은 데이터들을 동일한 그룹으로 분류, 서로 다른 군집들이 상이성을 가지도록 그룹화

* **군집화(Clustering) 활용 분야**

고객, 마켓, 브랜드, 사회 경제 활동 세분화(Segmentation)

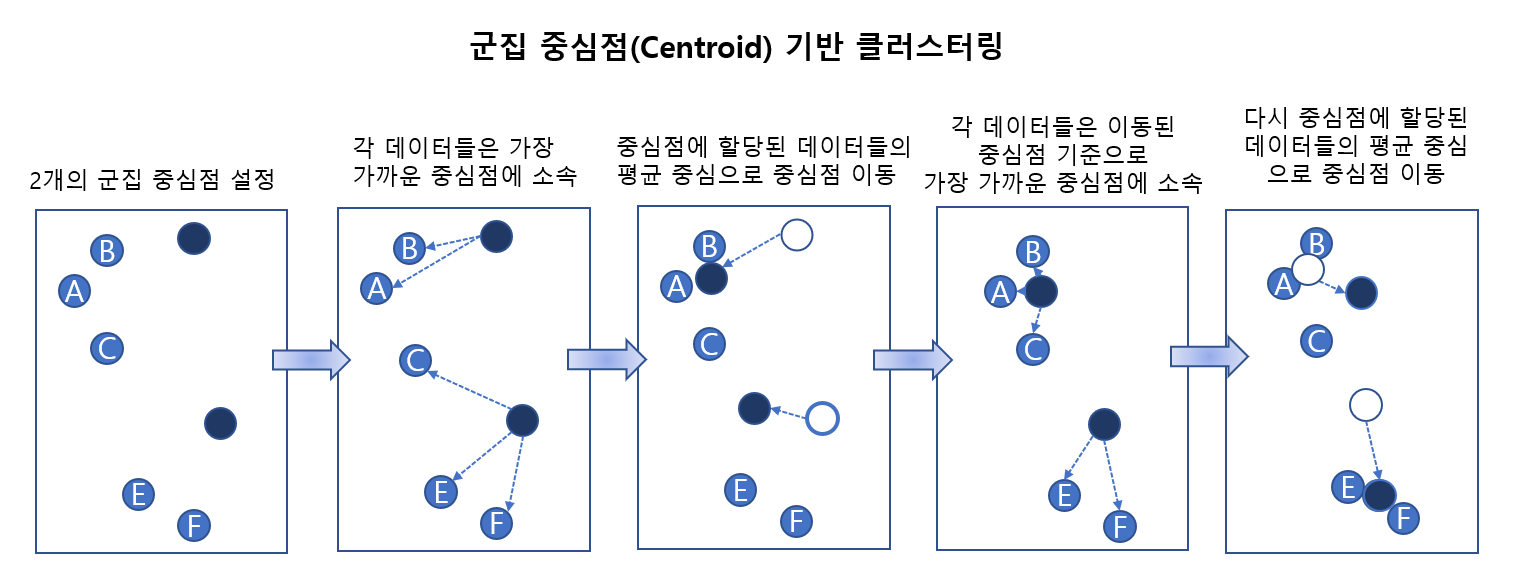
Image 검출, 세분화, Tracking

이상 검출(Abnomaly detection)

* **군집화(Clustering) 알고리즘**

K-Means / Mean Shift / Gaussian Mixture Model / DBSCAN

* **K-Means Clustering**

****

중심점을 이동하여 데이터들의 중심점 소속 변경이 없으면 군집화 완료.

* **K-Means Clustering 장점과 단점**

**K-Means 장점 :** 일반적인 군집화에서 가장 많이 활용되는 알고리즘

**알고리즘이 쉽고 간결함**

**대용량 데이터**에도 활용 가능

**K-Means 단점 :** 거리기반 알고리즘으로 **속성의 개수가 많을 경우** **군집화 정확도가 떨어짐**

🡪 PCA 차원축소 적용

반복 횟수가 많을 경우 **수행 시간이 느려짐**

**이상치(outlier)에 취약함**

* **사이킷런 K-Means 클래스**

class sklearn.cluster.KMeans(n\_clusters= 8, init = ‘k-means++’, n\_init = 10, max\_iter = 300, tol = 0.0001, precompute\_distances = ‘auto’, verbose = 0, random\_state = None, copy\_x = True, n\_jobs = 1, algorithm = ‘auto’ )

**n\_clusters :** 군집화할 개수, 군집 중심점의 개수

**init :** 초기 군집 중심점의 좌표 설정 방식

**max\_iter :** 최대 반복 횟수 (but 최대 반복횟수 이전에 모든 데이터의 중심점 이동이 없으면 종료)

**labels\_** : 각 데이터 포인트가 속한 군집 중심점 레이블

**cluster\_centers\_** : 각 군집 중심점 좌표 ( shape : [군집개수, 피처개수] )

* 각 군집 중심점 좌표 시각화
* **군집 평가 – 실루엣 분석**

실루엣 분석은 **각 군집 간의 거리가 얼마나 효율적으로 분리**되어 있는지를 나타냄.

**실루엣 계수(silhouette coefficient) :** 같은 군집내의 데이터와 얼마나 가깝게 군집화 되어있고,

다른 군집내의 데이터와 얼마나 멀리 분리돼 있는지 나타내는 지표

텍스트이(가) 표시된 사진

자동 생성된 설명

실루엣 계수는 -1에서 1사이의 값을 가짐

**실루엣 계수 -1 :** 아예 다른 군집에 데이터 포인트가 할당됨

**실루엣 계수 0 :** 0에 가까울수록 군집과 가까운 것임

**실루엣 계수 1 :** 1에 가까울수록 군집과 멀리 떨어져 있음

**sklearn.metrics.sillhouette\_samples**( X, labels, metric=’euclidean’, \*\*kwds ) : 실루엣 계수 반환

**sklearn.metrics.silhouette\_score**( X, labels,metric=’euclidean’,

sample\_size = None, \*\*kwds ) : 실루엣 계수 평균