CODINGON x posco

K-Digital Training 스마트 팩토리 3기

클러스터링(Clustering)

클러스터링(Clustering)

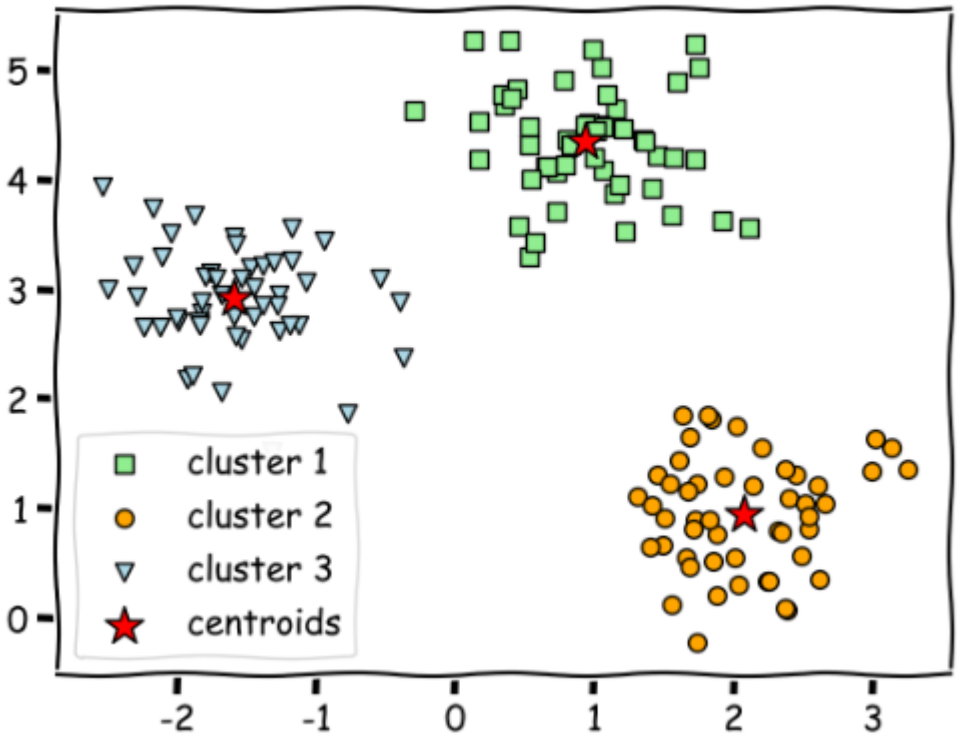
- 군집화(Clustering)는 비지도학습의 대표적 기술로 레이블이 지정 되어있지 않은 데이터를 그룹핑하는 분석 알고리즘
- 클러스터란 비슷한 특성을 가진 데이터들의 집단

클러스터링(Clustering)

- 알고리즘 종류
  - K 평균 (K-Means)
  - 평균이동(Mean Shift)
  - GMM(Gaussian Mixture Model)
  - DBSCAN(Density-Based Spatial Clustering of Applications with Noise)

K-Means

- 군집 중심점(centroid)라는 특정한 임의이 지점을 선택해 해당 중심에 가장 가까운 포인트들을 선택하는 군집화 기법
- 선택된 포인트의 평균지점으로 이동하고 이동된 중심점에서 다시 가까운 포인트를 선택, 이를 반복적으로 수행

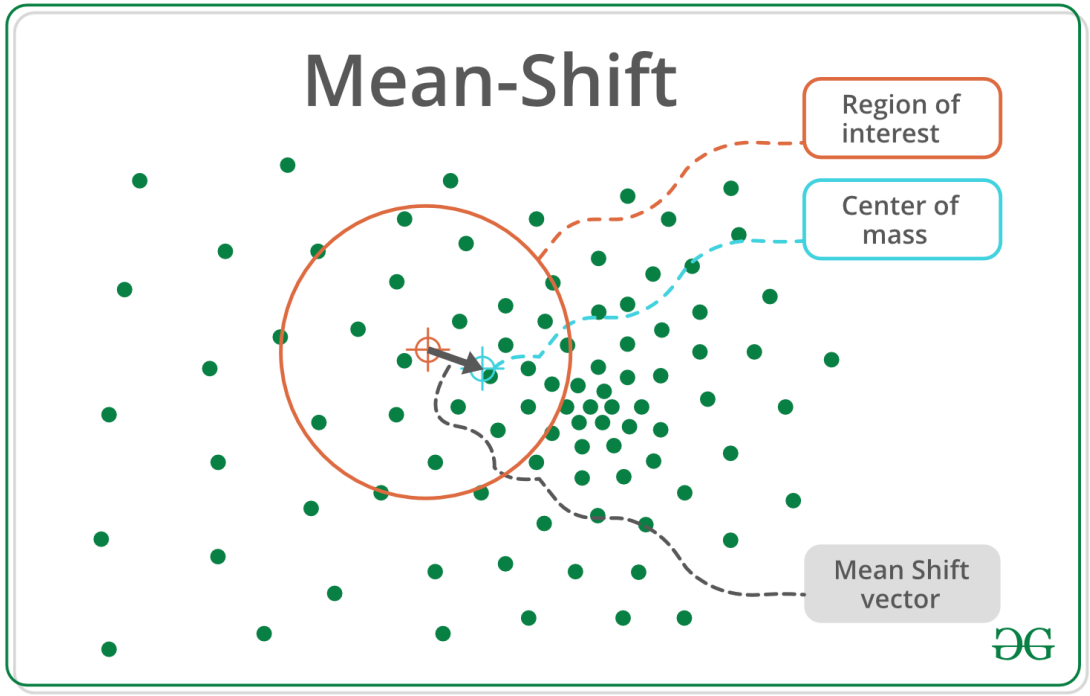


K-Means

- 장점
  - 일반적으로 군집화에서 가장 많이 사용되는 알고리즘
  - 쉽고 간결
- 단점
  - 거리기반 알고리즘으로 속성의 개수가 많을수록 정확도가 떨어짐
  - 반복횟수가 많을 경우 느려짐
  - 몇개의 군집을 선택해야 할 지 가이드하기 어려움

평균이동(Mean Shift)

- K 평균과 유사하지만 거리 중심이 아니라 데이터가 모여있는 밀도가 가장 높은쪽으로 군집 중심점 이동하면서 군집화 수행
- 일반 업무 기반의 정형 데이터 세트보다 컴퓨터 비전 영역에서 이미지나 영상 데이터에서 특정 개체를 구분하거나 움직임을 추적하는데 뛰어난 역할을 수행하는 알고리즘
- 컴퓨터 비전 영역에서 잘 사용됨



평균이동(Mean Shift)

- 장점
  - 데이터 세트의 형태를 특정형태나 특정 분포도 기반의 모델로 가정하지 않기 때문에 좀 더 유연한 군집화 가능
  - 이상치의 영향력 크지 않음(데이터 특성에 영향을 덜 받는다)
  - 미리 군집의 개수 정할 필요 없음
- 단점
  - 알고리즘의 수행시간이 오래 걸림 (O(N^2))
  - Bandwidth의 크기에 따른 군집화 영향도가 매우 크다

DBSCAN

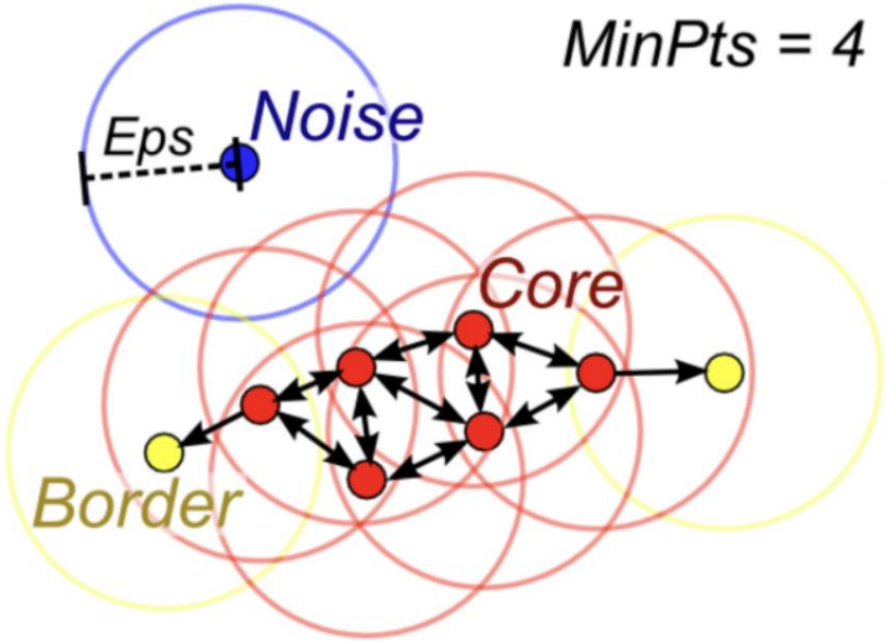
- Density-Based Spatial Clustering of applications with noise
- 밀도 기반 클러스터링
- 데이터를 밀도 기반으로 서로 가까운 데이터 포인트를 함께 그룹화
- 매개 변수 :
  - 엡실론(ε) : 인접한 점을 찾기 위해 점 주위의 반경을 결정하는 거리 매개변수
  - 최소 포인트(MinPts) : 코어 포인트로 간주되기 위해 포인트의 ε 근처 내의 필요한 최소 포인트 수

DBSCAN

- 포인트 카테고리
  - 핵심 포인트 : ε 이웃 내에 MinPts개 이상의 포인트가 있는 클러스터 내의 포인트
  - 경계 포인트 : 핵심 포인트의 ε 이웃 내에 있지만 코어 포인트 자체가 되기에는 이웃이 충분하지 않은 포인트
  - 노이즈 포인트(아웃라이어) : 핵심 포인트도 경계 포인트도 아닌 포인트

DBSCAN

- 알고리즘
  1. 방문하지 않은 지점 무작위 선택, 충분한 포인트가 포함되어있으면 핵심 포인트로 지정, 그렇지 않으면 노이즈
  2. 핵심 포인트의 ε 이웃 내의 포인트를 동일한 클러스터에 추가하여 클러스터 확장, 포인트가 핵심 포인트인 경우 계속해서 탐색
  3. 모든 포인트가 클러스터에 할당되거나 노이즈로 표시될 때까지 반복



Red: Core Points

Yellow: Border points. Still part of the cluster because it's within epsilon of a core point, but not does not meet the min_points criteria

Blue: Noise point. Not assigned to a cluster

DBSCAN

- 장점
  - 노이즈에 강함 : 이상값이나 잡음을 별도의 엔터티로 지정하여 처리
  - 클러스터 수 지정이 필요하지 않음 : K 평균과 달리 클러스터 수 지정 불필요
  - 클러스터 모양이 불규칙하고 밀도가 다양한 시나리오의 경우 유용
- 단점
  - 매개변수에 민감함 : 적절한 ε 및 MinPts 값을 선택하는 것이 어렵고 결과에 영향을 미침
  - 다양한 밀도에 대한 성능 : 단일 ε이 모든 클러스터에 적합하지 않을 수 있음

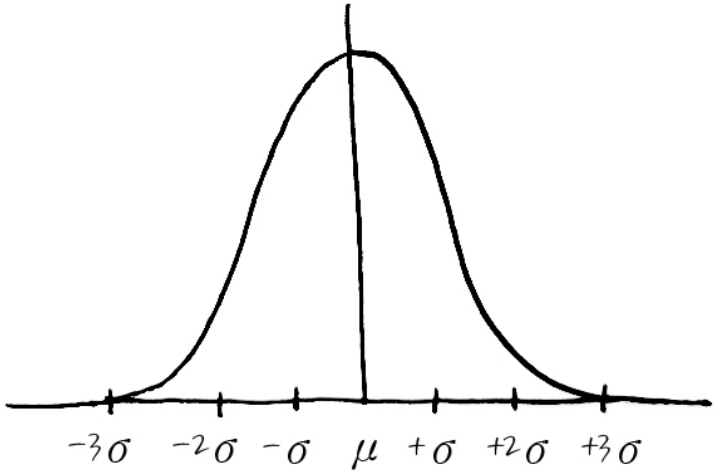
DBSCAN

- Mean shift와의 차이

특징	DBSCAN	Mean Shift
클러스터링 유형	밀도 기반 클러스터링	중심 기반 클러스터링
클러스터 수 결정	ε와 MinPts 설정에 의존	밀도 피크에 의한 자동 클러스터 결정
핵심 포인트 정의	MinPts 개수 이상의 이웃을 가진 포인트	밀도 피크가 수렴하는 중심으로 이동
이상치 처리	노이즈 포인트를 별도로 처리	피크에서 멀리 떨어진 이상치 포인트 처리
클러스터 모양과 크기	상대적으로 유연한 형태와 크기에 적응	가우시안 커널 함수를 사용하여 클러스터 모양 결정
하이퍼파라미터 영향	ε와 MinPts 설정에 영향을 받음	이동 크기(hyperparameter)에 영향을 받음

GMM(Gaussian Mixture Model)

- 군집화를 적용하는 데이터가 여러 개의 가우시안 분포를 섞어서 생성된 모델로 가정해 수행하는 방식
- 가우시안 분포
  - 좌우 대칭형의 bell 형태를 가진 통계학에서 가장 잘 알려진 연속 확률 함수



GMM(Gaussian Mixture Model)

- GMM은 데이터를 여러 개의 가우시안 분포가 섞인것으로 간주
  - 섞인 데이터 분포에서 개별 유형의 가우시안 분포를 추출한다
- 전체 데이터 세트는 서로 다른 정규분포를 가진 여러가지 확률 분포곡선으로 구성되어 있으며, 정규분포에 기반에 군집화를 수행하는 것이 GMM 군집화 방식

GMM(Gaussian Mixture Model)

- 일정한 데이터 세트가 있으면 이를 구성하는 여러 개의 정규 분포 곡선 추출 뒤, 개별 데이터가 이 중 어떤 정규분포에 속하는지 결정
- 이러한 방식을 모수추정 이라고 하며 “개별 정규분포의 평균과 분산“, “각 데이터가 어떤 정규 분포에 해당되는지의 확률“ 을 구하기 위해서 추정
- GMM은 모수 추정을 하기 위해 EM(Expectation and Maximization) 방법을 적용