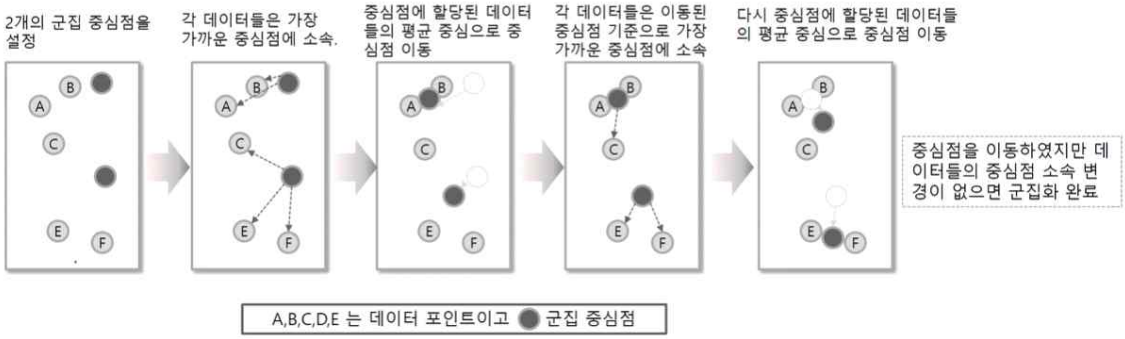
DA_Week10_preview_박보영



k-평균은 군집중심점을 선택해 해당 중심에 가장 가까운 포인트들을 선택하는 군집화 기법.

장점 : 군집화에서 가장 많이 활용됨. 쉽고 간결
단점 : 속성 개수가 많을 경우 정확도 떨어짐(PCA 필요할 수도) 반복횟수 많을수록 수행시간 길어짐. 몇 개의 군집을 선택해야 할지 가이드하기 어렵.

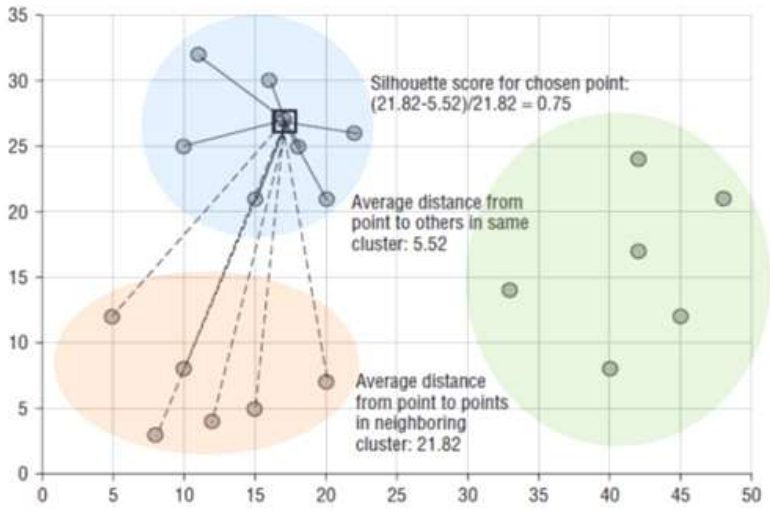
KMeans: n_cluster 군집화할 개수, init 초기 군집 중심점 좌표, max_iter최대반복횟수

fit & transform

labels_ 각 데이터 포인트가 속한 군집 중심점 레이블,
cluster_centers_ 각 군집 중심점 좌표

군집화 알고리즘 테스트를 위한 데이터 생성기 make_blobs() & make_classification()
전자는 개별 군집의 중심점과 표준 편차 제어 기능이 추가됨
후자는 노이즈를 포함한 데이터 만드는데 유용

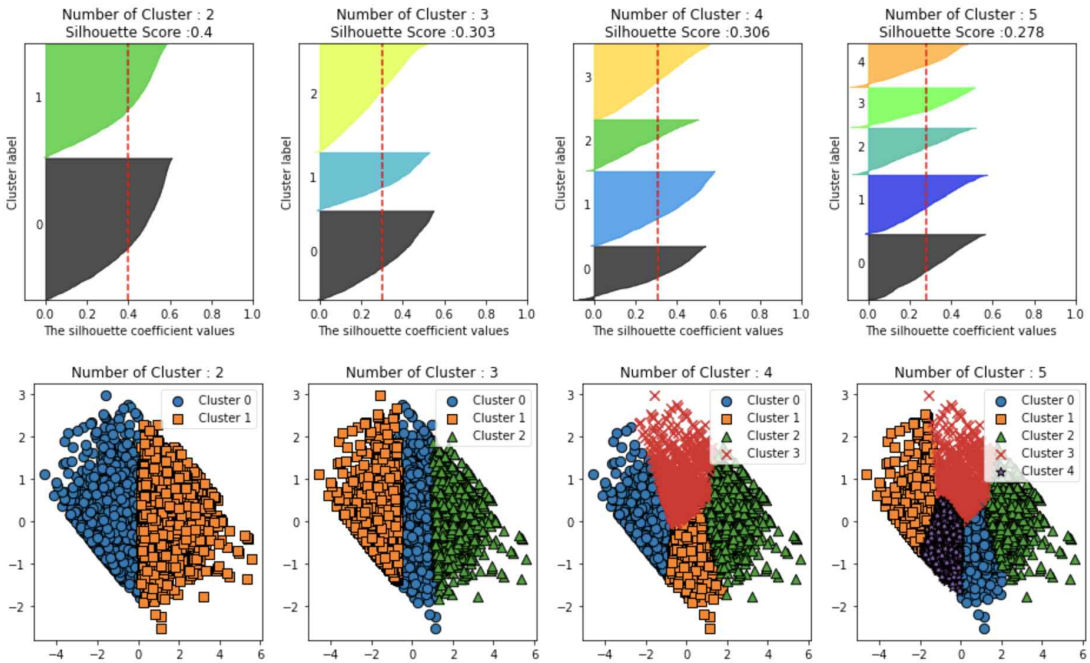
군집화 평가방법으로 실루엣 분석이 있음. 각 군집 간의 거리가 얼마나 효율적으로 분리돼 있는지를 나타낸다. 실루엣 계수. 해당데이터가 같은 군집 내의 데이터와 얼마나 가깝게 군집화돼 있고, 다른 군집에 있는 데이터와는 얼마나 멀리 분리돼 있는지를 나타내는 지표이다.



좋은 군집화 조건

전체 실루엣 계수의 평균은 1에 가깝다/ 개별 군집의 평균값의 편차가 크지 않아야 한다.

군집별 평균 실루엣 계수의 시각화를 통한 군집 개수 최적화 방법

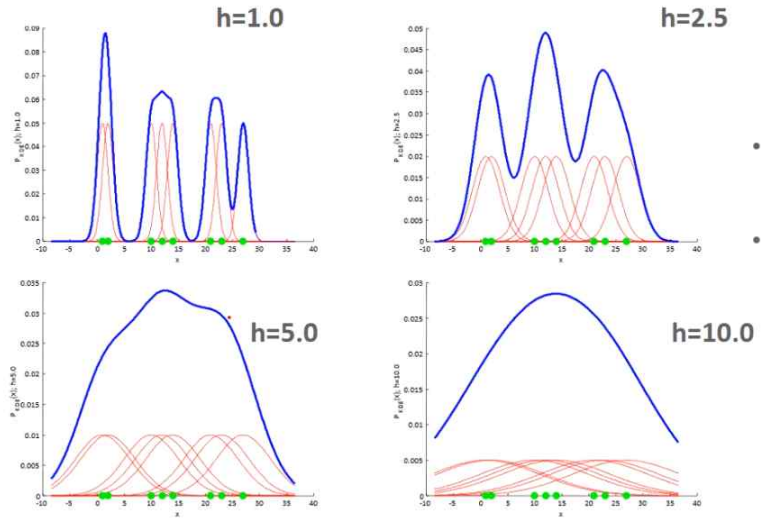


시각화

visualize_silhouette(cluster_lists, X_features)

평균이동(mean shift)
군집의 중심으로 지속적으로 움직이면서 군집화-밀도가 가장 높은 곳으로

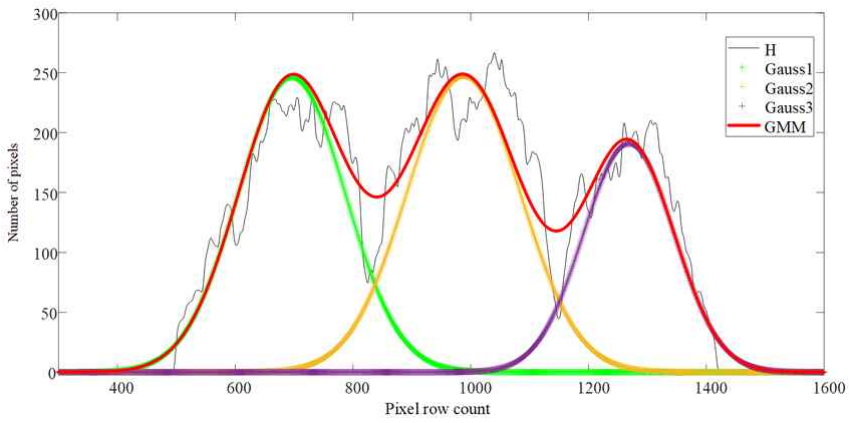
확률 밀도 함수를 이용. 함수를 찾기 위해서 KDE이용.
주변 데이터와의 거리값을 KDE 함수값으로 입력한 뒤 그 반환값을 현재 위치에서 업데이트하면서 이동하는 방식을 취함.



대역폭 h를 어떻게 설정하느냐에 따라 확률 밀도 추정 성능을 크게 좌우할 수 있다.
작은 h는 과적합. 큰 h는 평활화

MeanShift - bandwith=h

GMM(Gaussian Mixture Model)
데이터가 여러개의 가우시안 분포를 가진 데이터 집합들이 섞여서 생성된 것이라는 가정하에 군집화를 수행하는 방식.



개별 데이터가 어떤 전규분포에 속하는지 결정하는 방식.(모수 추정)
개별 정규 분포의 평균과 분산/ 각 데이터가 어떤 정규 분포에 해당하는지의 확률 -> EM

*K-means는 원형의 범위에서 수행

DBSCAN 밀도 기반 군집화는 기하학적으로 복잡한 데이터세트에도 효과적.
epsilon, min points