07장. 군집화

01. K-평균 알고리즘 이해

- 군집화 (clustering) 에서 가장 일반적으로,
- 특정한 임의의 점을 선택, 해당 중심에 가장 가까운 포인트를 선택. } 이 과정을 반복.
- 선택된 포인트의 평균 지점으로 중심점이 이동
→ 중심점의 이동이 없을 때까지 위의 과정을 반복.

K평균의 장점	K평균의 단점
· 일반적 군집화에 가장 많이 활용.	· 개수가 많으면 정확도↓ · 몇개의 군집을 선택할지 어려움.
· 쉽고 간결.	· 반복횟수가 많으면 느려진다.

· 사이킷런 kmeans 파라미터.

· n_clusters : 군집화할 개수 (군집 중심점의 개수)

· init : 초기에 군집 중심점의 좌표를 설정할 방식.

· max_iter: 최대 반복 횟수

· 군집화 알고리즘 테스트를 위한 데이터 생성.

· make_blobs()
→ 개별 군집의 중심점과 표준 편차 제어 기능이 추가.

· make_classification()
→ 노이즈를 포함한 데이터 생성에 유용.

군집화용 데이터 생성기

make_blobs() 파라미터

n_samples: 생성할 총 데이터 개수

n_features: 데이터 피처 개수

centers: int 값.

cluster_std: 생성될 군집 데이터의 표준 편차.

02. 군집평가

- 실루엣 분석. (각 군집 간의 거리가 얼마나 효율적으로 분리되 있는지..)

↳ 좋은 군집끼리는 잘 뭉쳐있고, 다른 군집끼리는 떨어져 있음.

실루엣 계수를 통해 표현.

실루엣 계수 $s(i) = \frac{(b(i) - a(i))}{\max(a(i), b(i))}$

· -1과 1 사이의 값.
· 1로 가까워질 수록 근처 군집과 멀리, 0에 가까울 수록 가까워진다.

· 전체 실루엣 계수의 평균값 : silhouette_score() 는 0~1의 값, 1에 가까울 수록 좋다.

03. 평균이동 (Mean Shift)

· K평균과 유사하게 군집의 중심으로 점적으로 움직이면서 군집화를 수행.

· 데이터의 분포도를 이용해 군집 중심점을 찾음.

데이터 포인트가 몰려있는 곳. (확률 밀도 함수를 이용) KDE (Kernel Density Estimation) 이용.

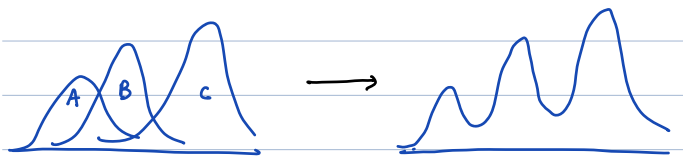
· 관측된 데이터 각각에 커널함수를 적용하여 합한다. → 비모수적인 분포 함수 사용.

· 데이터 건수로 나눠 확률밀도 함수를 추정.

대역폭 h: smoothing 에 영향. (h가 크면: 적은 수의 군집 중심점
h가 작으면: 많은 수의 〃)

04. GMM (Gaussian Mixture Model)

· 군집화를 적용하고자 하는 데이터가 여러 개의 가우시안 분포를 가진 데이터 집합들의 섞여서 생성된 것이라는 가정 하..



정규분포 A, B, C 클러스터링 하려는 데이터 분포.

여러 데이터 set → 이를 구성하는 여러 개의 정규분포 곡선 추출. → 개별 데이터가 이 중 어떤 정규 분포에 속하는지.

모수추정 ① 개별 정규 분포의 평균 / 분산

② 각 데이터가 어떤 정규분포에 해당하는지..

* GMM은 KMeans 보다 유연하게 다양한 데이터 셋에 잘 적용될 수 있다는 장점 - but 시간↑

다원형은 잘 X

05. DBSCAN

· 밀도 기반 군집화의 대표적인 알고리즘.

파라미터

· 입실론 주변 영역 (epsilon) : 개별 데이터를 중심으로 입실론

· 최소 데이터 개수 (min points) : 개별 데이터의 입실론 주변 영역에 포함되는 다 데이터의 개수.