EURON 11주차 과제

01. K-평균 알고리즘 이해

K-평균

- 군집화에서 가장 일반적으로 사용되는 알고리즘
- 군집 중심점이라는 특정한 임의의 지점을 선택해 해당 중심에 가까운 포인트들을 선택하는 군집화 기법
    - 군집 중심점 : 선택된 포인트의 평균 지점으로 이동하고 이동된 중심점에서 다시 가까운 포인트를 선택, 다시 중심점을 평균 지점으로 이동하는 프로세스 반복 수행
    - 모든 데이터 포인트에서 더 이상 중심점의 이동이 없을 경우, 반복을 멈추고 해당 중심점에 속하는 데이터 포인트들을 군집화하는 기법
- 장점
    - 일반적인 군집화에서 가장 많이 활용되는 알고리즘
    - 알고리즘이 쉽고 간결
- 단점
    - 거리 기반 알고리즘으로 속성의 개수가 매우 많을 경우 군집화의 정확도가 떨어짐 (이를 위해 PCA로 차원 감소를 적용해야 할 수도 있음)
    - 반복을 수행하는데, 반복 횟수가 많을 경우 수행 시간이 매우 느려짐
    - 몇 개의 군집을 선택해야 할 지 가이드하기가 어려움

사이킷런 KMeans 클래스 소개

다음과 같은 초기화 파라미터를 가짐

```
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'
```

이 중 중요한 파라미터는 다음과 같음

- KMeans 초기화 파라미터 중 가장 중요한 파라미터는 n_clusters이며, 군집화할 개수, 즉 군집 중심점의 개수를 의미
- init는 초기에 군집 중심점의 좌표를 설정할 방식을 말하며 보통은 임의로 중심을 설정하지 않고 일반적으로 k-means++ 방식으로 최초 설정
- max_iter는 최대 반복 횟수이며, 이 횟수 이전에 모든 데이터의 중심점 이동이 없으면 종료

KMeans는 사이킷런의 비지도학습 클래스와 마찬가지로 fit 또는 fit_transform 메서드를 이용해 수행

군집화와 관련된 주요 속성은 다음과 같음

- labels_ : 각 데이터 포인트가 속한 군집 중심점 레이블
- cluster_centers_ : 각 군집 중심점 좌표. 이를 이용하면 군집 중심점 좌표가 어디인지 시각화할 수 있음

군집화 알고리즘 테스트를 위한 데이터 생성

대표적인 군집화용 데이터 생성기

- make_blobs() : 개별 군집의 중심점과 표준 편차 제어 기능이 추가
- make_classification() : 노이즈를 포함한 데이터를 만드는 데 유용하게 사용 가능

두 API 모두 여러 개의 클래스에 해당하는 데이터 세트를 만드는데, 하나의 클래스에 여러 개의 군집이 분포될 수 있게 데이터를 생성할 수 있음

02. 군집 평가(Cluster Evaluation)

- 대부분의 군집화 데이터는 비교할 만한 타깃 레이블을 가지고 있지 않음
- 군집화는 분류와 유사해보일 수 있으나 성격이 많이 다름
- 군집화의 성능을 평가하는 대표적인 방법 → 실루엣 분석

실루엣 분석의 개요

- 각 군집 간의 거리가 얼마나 효율적으로 분리돼 있는지 나타냄
- 효율적으로 잘 분리됐다는 것은 다른 군집과의 거리는 떨어져 있고 동일 군집끼리의 데이터는 서로 가깝게 잘 뭉쳐있다는 의미
- 실루엣 분석은 실루엣 계수를 기반으로 함
    - 실루엣 계수는 개별 데이터가 가지는 군집화 지표
    - 개별 데이터가 가지는 실루엣 계수는 해당 데이터가 같은 군집 내의 데이터와 얼마나 가깝게 군집화돼 있고, 다른 군집에 있는 데이터와는 얼마나 멀리 분리돼 있는지를 나타내는 지표
    - -1에서 1 사이의 값을 가지며, 1로 가까워질수록 근처의 군집과 더 멀리 떨어져 있는 것이고 0에 가까울수록 근처의 군집과 가까워진다는 것
    - -값은 아예 다른 군집에서 데이터 포인트가 할당됐음을 의미

군집별 평균 실루엣 계수의 시각화를 통한 군집 개수 최적화 방법

- 전체 데이터의 평균 실루엣 계수 값이 높다고 해서 반드시 최적의 군집 개수로 군집화가 잘 됐다고 볼 수는 없음
- 특정 군집 내의 실루엣 계수 값만 너무 높고, 다른 군집은 내부 데이터끼리의 거리가 너무 떨어져 있어 실루엣 계수 값이 낮아져도 평균적으로 높은 값을 가질 수 있음
- 개별 군집별로 적당히 분리된 거리를 유지하면서도 군집 내의 데이터가 서로 뭉쳐 있는 경우에 K-평균 의 적절한 군집 개수가 설정됐다고 판단할 수 있음

03. 평균 이동

평균 이동(Mean Shift)의 개요

- 평균 이동은 K-평균과 유사하게 중심을 군집의 중심으로 지속적으로 움직이면서 군집화를 수행
- 데이터가 모여 있는 밀도가 가장 높은 곳으로 이동시킴

평균 이동 군집화

- 데이터의 분포도를 이용해 군집 중심점을 찾음
- 군집 중심점 : 데이터 포인트가 모여 있는 곳, 이를 위해 확률 밀도 함수 이용
- 확률 밀도 함수가 피크인 점 : 군집 중심점 → KDE 이용

KDE(Kernel Density Estimation)

- 커널 함수를 통해 어떤 변수의 확률 밀도 함수를 추정하는 대표적인 방법
- 관측된 데이터 각각에 커널 함수를 적용한 값을 모두 더한 뒤 데이터 건수로 나눠 확률 밀도 함수를 추정
- 대표적인 커널 함수로 가우시안 분포 함수가 사용
- 적절한 KDE의 대역폭 h를 계산하는 것은 KDE 기반의 평균 이동 군집화에서 매우 중요

04. GMM(Gaussian Mixture Model)

- 군집화를 적용하고자 하는 데이터가 여러 개의 가우시안 분포를 가진 데이터 집합들이 섞여서 생성된 것이라는 가정하에 군집화 수행
- 데이터를 여러 개의 가우시안 분포가 섞인 것으로 간주
- 모수 추정을 위해 GMM은 EM 방법을 적용

05. DBSCAN

DBSCAN 개요

- 밀도 기반 군집화의 대표적인 알고리즘
- 간단하고 직관적인 알고리즘으로 되어있음에도 데이터의 분포가 기하학적으로 복잡한 데이터 세트에도 효과적인 군집화가 가능
- 특정 공간 내에 데이터 밀도 차이를 기반 알고리즘으로 하고 있어서 복잡한 기하학적 분포도를 가진 데이터 세트에 대해서도 군집화를 잘 수행
- DBSCAN을 구성하는 가장 중요한 두 가지 파라미터는 입실론으로 표기하는 주변 영역과 이 입실론 주변 영역에 포함되는 최소 데이터의 개수 min points
    - 입실론 주변 영역 : 개별 데이터를 중심으로 입실론 반경을 가지는 원형의 영역
    - 최소 데이터 개수 : 개별 데이터의 입실론 주변 영역에 포함되는 타 데이터의 개수
- DBSCAN은 입실론 주변 영역의 최소 데이터 개수를 포함하는 밀도 기준을 충족시키는 데이터인 핵심 포인트를 연결하면서 군집화를 구성