Week 11_예습과제_김정은

07. 군집화

K- 평균

- K - 평균
  - 군집화에서 가장 일반적으로 사용되는 알고리즘
  - 군집 중심점이라는 특정한 임의의 지점을 선택해 해당 중심에 가장 가까운 포인트들을 선택하는 군집화 기법
  - 군집 중심점은 선택된 포인트의 평균 지점으로 이동하고 이동된 중심점에서 다시 가까운 포인트를 선택, 다시 중심점을 평균 지점으로 이동하는 프로세스를 반복적으로 수행
  - 모든 데이터 포인트에서 더 이상 중심점의 이동이 없을 경우에 반복을 멈추고 해당 중심점에 속하는 데이터 포인트들을 군집화

- K- 평균 장점
  - 일반적인 군집화에서 가장 많이 활용
  - 알고리즘이 쉽고 간결
- K- 평균 단점
  - 거리 기반 알고리즘으로 속성의 개수가 매우 많을 경우 군집화 정확도가 떨어짐 (PCA로 차원 감소를 적용해야 할 수도)
  - 반복을 수행하는데 반복 횟수가 많을 경우 수행시간이 느려짐
  - 몇 개의 군집을 선택해야할지 가이드하기 어려움

- 사이킷런은 다양한 유형의 군집화 알고리즘을 테스트해 보기 위한 간단한 데이터 생성기를 제공
- 대표적인 데이터 생성기로는 make_blobs( ) 와 make_classification( ) API 존재
- 이들은 비슷하게 여러 개의 클래스에 해당하는 데이터 세트를 만드는데 하나의 클래스에 여러 개의 군집이 분포될 수 있게 데이터 생성 가능
- make_blobs( )는 개별 군집의 중심점과 표준 편차 제어 기능이 추가되어 있음
- make_classification( )은 노이즈를 포함한 데이터를 만드는 데 유용하게 사용 가능

군집 평가(Cluster Evaluation)

- 군집화의 성능을 평가하는 대표적인 방법으로 실루엣 분석 사용

- 실루엣 분석
  - 각 군집 간의 거리가 얼마나 효율적으로 분리되어 있는지를 나타냄.
  - 효율적으로 잘 분리되었다는 것은 다른 군집과의 거리는 떨어져 있고 동일 군집끼리의 데이터는 서로 가깝게 잘 뭉쳐 있다는 의미
  - 군집화가 잘 될수록 개별 군집은 비슷한 정도의 여유 공간을 가지고 떨어져 있을 것
  - 실루엣 계수(silhouette coefficient)를 기반으로 실행 : 개별 데이터가 가지는 군집화 지표
    - 해당 데이터가 같은 군집 내의 데이터와 얼마나 가깝게 군집화돼 있고, 다른 군집에 있는 데이터와는 얼마나 멀리 분리돼 있는지를 나타냄

- 좋은 군집화가 되기 위한 조건
  - 전체 실루엣 계수의 평균값은 0~1 사이의 값을 가지며 1에 가까울수록 좋다.
  - 전체 실루엣 계수의 평균값과 더불어 개별 군집의 평균값의 편차가 크지 않아야 한다.

평균 이동

- 평균 이동
  - K-평균과 유사하게 중심을 군집의 중심으로 지속적으로 움직이면서 군집화를 수행
  - 중심을 데이터가 모여 있는 밀도가 가장 높은 곳으로 설정
  - 데이터의 분포도를 이용하여 군집 중심점을 찾는다.
    - 군집 중심점은 데이터 포인트가 모여있는 곳이라는 생각에서 착안한 것이며, 이를 위해 확률 밀도 함수 이용
    - 가장 집중적으로 데이터가 모여있어 확률 밀도 함수가 피크인 점을 군집 중심점으로 선정
    - 일반적으로 주어진 모델의 확률 밀도 함수를 찾기 위해 KDE를 이용
      - KDE는 커널 함수를 통해 어떤 변수의 확률 밀도함수를 추정하는 대표적인 방법
      - 관측된 데이터 각각에 커널 함수를 적용한 값을 모두 더한 뒤 데이터 건수로 나눠 확률 밀도 함수를 추정
      - 대표적인 커널 함수로서는 가우시안 분포 존재.
    - PDF는 확률 변수의 분포를 나타내는 함수, 감마 분포, t- 분포 등이 있음

GMM(Gaussian Mixture Model)

- GMM 군집화
  - 군집화를 적용하고자 하는 데이터가 여러 개의 가우시안 분포를 가진 데이터 집합들이 섞여서 생성된 것이라는 가정하에 군집화를 수행
  - 데이터를 여러 개의 가우시안 분포가 섞인 것으로 간주
  - 섞인 데이터 분포에서 개별 유형의 가우시안 분포 추출
  - 서로 다른 정규 분포에 기반하여 군집화를 수행
  - 1000개의 데이터 세트가 있다면 이를 구성하는 여러 개의 정규 분포 곡선을 추출하고 개별 데이터가 이 중 어떤 정규 분포에 속하는지 결정 → 모수 추정
  - 모수 추정
    - 개별 정규 분포의 평균과 분산
    - 각 데이터가 어떤 정규 분포에 해당되는지의 확률

        - 모수 추정을 위해 EM 방법 적용
    - GMM의 경우 KMeans보다 유연하게 다양한 데이터 세트에 잘 적용될 수 있다는 장점
    - 시간이 오래 걸린다는 단점

DBSCAN

- DBSCAN
    - 밀도 기반 군집화의 대표적인 알고리즘
    - 데이터의 분포가 기하학적으로 복잡한 데이터 세트에도 효과적인 군집화 가능
    - 특정 공간 내에 데이터 밀도 차이를 기반 알고리즘으로 하고 있어서 복잡한 기하학적 분포를 가진 데이터 세트에 대해서도 군집화 잘 수행
    - 구성하는 두 가지 파라미터는 입실론(epsilon) : 주변 영역, 그리고 입실론 주변 영역에 포함되는 최소 데이터 개수 min points
    - 입실론 주변 영역 내에 포함되는 최소 데이터 개수를 충족시키는가 아닌가에 따라 데이터 포인트를 정의
        - 핵심 포인트
        - 이웃 포인트
        - 경계 포인트
        - 잡음 포인트