CHAPTER 07 군집화

K-평균 알고리즘 이해

K-평균

- 군집화에서 가장 일반적으로 사용되는 알고리즘
- 군집 중심점이라는 특정 임의의 지점을 선택, 해당 중심에 가장 가까운 포인트를 선택하는 군집화 기법
- 군집 중심점은 선택된 포인트의 평균 지점으로 이동, 이동 중심점에서 다시 가까운 포인트를 선택, 다시 중심점을 평균 지점으로 이동하는 프로세스를 반복적으로 수행
- 더 이상 중심점의 이동이 없을 경우, 반복을 멈추고 군집화

사이킷런 KMeans 클래스

- K-평균을 구현하기 위한 클래스
- 사이킷런의 비지도학습 클래스와 마찬가지로 데이터 세트 또는 데이터 세트 메서드를 이용해 수행

K-평균을 이용한 붓꽃 데이터 세트 군집화

- 실습

군집화 알고리즘 테스트를 위한 데이터 생성

- Make_blobs(): 개별 군집의 중심점과 표준 편차 제어 기능이 추가
  make_classification(): 노이즈를 포함한 데이터를 만드는 데 유용
- 하나의 클래스에 여러 개의 군집이 분포될 수 있게 데이터 생성이 가능

군집 평가

- 군집화의 성능을 평가하는 대표적인 방법으로, 실루엣 분석을 이용

실루엣 분석의 개요

- 각 군집 간의 거리가 얼마나 효율적으로 분리돼 있는지 나타냄
- 효율적으로 잘 분리됐다는 것은 다른 군집과의 거리는 떨어져 있고, 동이일 군집 끼리의 데이터는 서로 가깝게 잘 뭉쳐 있다는 의미
- 실루엣 계수(개별 데이터가 가지는 군집화 지표)를 기반으로 함
- 개별 데이터가 가지는 실루엣 계수는 해당 데이터가 같은 군집 내의 데이터와 얼마나 가깝게 군집화되어있고, 다른 군집에 있는 데이터와는 얼마나 멀리 분리돼있는지를 나타내는 지표

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

- 실루엣 계수는 -1에서 1사이의 값을 가지며, 1로 가까워질수록 근처의 군집과 더 멀리 떨어져있고, 0에 가까울수록 근처의 군집과 가까워짐

붓꽃 데이터 세트를 이용한 군집 평가

- 실습

군집별 평균 실루엣 계수의 시각화를 통한 군집 개수 최적화 방법

- 개별 군집별로 적당히 분리된 거리를 유지하면서도 군집 내의 데이터가 서로 뭉쳐 있는 경우에 K-평균의 적절한 군집 개수가 설정됐다고 판단

평균 이동

- K-평균과 유사하게 중심을 군집의 중심으로 지속적으로 움직이면서 군집화를 수행
- 중심을 데이터가 모여 있는 밀도가 가장 높은 곳으로 이동시킴
- 평균 이동 군집화는 데이터의 분포도를 이용해 군집 중심점을 찾음(확률 밀도 함수를 이용)
- 평균 이동 군집화는 특정 데이터를 반경 내의 데이터 분포 확률 밀도가 가장 높

은 곳으로 이동하기 위해 주변 데이터와의 거리 값을 KDE 함수 값으로 입력한 뒤 그 반환 값을 현재 위치에서 업데이트하면서 이동하는 방식을 취함
- KDE는 커널 함수를 통해 어떤 변수의 확률 밀도 함수를 추정하는 대표적인 방법
- 확률 밀도 함수 PDF는 확률 변수의 분포를 나타내는 함수로 널리 알려진 정규 분포 함수를 포함해 감마 분포, t-분포 등이 있음

KDE

- 개별 관측 데이터에 커널 함수를 적용한 뒤, 이 적용 값을 모두 더한 후 개별 관측 데이터의 건수로 나눠 확률 밀도 함수를 추정, 가우시안 분포 함수가 사용됨

GMM

- 군집화를 적용하고자 하는 데이터가 여러 개의 가우시안 분포를 가진 데이터 집합들이 섞여서 생성된 것 이라는 가정하에 군집화를 수행하는 방식
- 데이터를 여러 개의 가우시안 분포가 섞인 것으로 간주
- 확률 기반 군집화

GMM을 이용한 붓꽃 데이터 세트 군집화

- 실습

GMM과 K-평균의 비교

- 실습

DBBSCAN

- 밀도 기반 군집화의 대표적인 알고리즘
- 간단하고 직관적인 알고리즘
- 데이터 분포가 기하학적으로 복잡한 데이터 세트에도 효과적인 군집화가 가능
- DBSCAN을 구성하는 가장 중요한 두 가지 파라미터는 입실론으로 표기하는 주변 영역과 입실론 주변 영역에 포함되는 최소 데이터의 개수
- 데이터 포인트: 핵심, 이웃, 경계, 잡음 포인트

- 입실론 주변 영역의 최소 데이터 개수를 포함하는 밀도 기준으로 충족시키는 데이터인 핵심 포인트를 연결하면서 군집화를 구성

DBSCAN 적용하기 – 붓꽃 데이터 세트

- 실습

군집화 실습 – 고객 세그먼테이션

- 실습