파이썬 머신러닝 완벽가이드 7장 필사&개념정리 (7.1~7.6)

[Chapter 07. 군집화]

1. K- 평균 알고리즘 이해

1) K-평균: 군집 중심점이라는 특정한 임의의 지점을 선택해 해당 중심에 가장 가까운 포인트들을 선택하는 군집화 기법

2) 군집 중심점은 선택된 포인트의 평균 지점으로 이동
➔ 이동된 중심점에서 다시 가까운 포인트를 선택
➔ 다시 중심점을 평균 지점으로 이동하는 프로세스 반복
➔ 모든 데이터 포인트에서 더 이상 중심점의 이동이 없을 경우에 반복을 멈추고 해당 중심점에 속하는 데이터 포인트들을 군집화

3) 사이킷런 KMeans 클래스 소개

(1) 중요 파라미터
n_clusters: 군집화할 개수
init: 초기에 군집 중심점의 좌표를 설정할 방식
max_iter: 최대 반복 횟수

(2) 주요 속성 정보
labels_: 각 데이터 포인트가 속한 군집 중심점 레이블
cluster_centers: 각 군집 중심점 좌표

4) K-평균을 이용한 붓꽃 데이터 세트 군집화
꽃받침과 꽃잎과 너비에 따른 품종을 분류하는 데이터 세트

필요 모듈, 데이터 세트 로드

➔ DataFrame으로 변경

➔ n_cluster는 3, 초기 중심 설정 방식은 디폴트, 최대 반복 횟수도 디폴트로 설정한 KMeans 객체를 만들고 fit() 수행

➔ kmeans 객체 변수로 반환됨

➔ labels_ 속성값 출력

➔ 군집화가 효과적으로 이루어졌는지 확인하기 위해 target과 cluster 값 개수 비교

➔ 군집화 시각화 (2개로 차원 축소)

➔ 맷플롯립의 산점도는 서로 다른 마커를 한 번에 표현할 수 없으므로 마커별로 별도의 산점도를 수행함

5) 군집화 알고리즘 테스트를 위한 데이터 생성

make_blobs(), make_classification() API

make_blobs()는 개별 군집의 중심점과 표준 편차 제어 기능이 추가되어 있고, make_classification()은 노이즈를 포함한 데이터를 만드는 데 유용하게 사용 가능

+) make_circle(), make_moon() API

make_blobs()

n_samples	생성할 총 데이터 개수, 디폴트=100개
n_features	데이터의 피처 개수
centers	int 값
cluster_std	생성될 군집 데이터의 표준 편차

2. 군집평가

1) 실루엣 분석: 각 군집 간의 거리가 얼마나 효율적으로 분리되어 있는지를 나타냄

(1) 실루엣 계수: 개별 데이터가 가지는 군집화 지표

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

1로 가까워질수록 근처 군집과의 거리가 멀다

-는 아예 다른 군집에 데이터 포인트가 할당되었음을 뜻함

(2) 좋은 군집화의 조건

- 평균값이 1에 가까울수록 좋음
- 개별 군집의 평균값의 편차가 크지 않아야 함

2) 군집별 평균 실루엣 계수의 시각화를 통한 군집 개수 최적화 방법

- 붓꽃 데이터를 K-평균으로 군집화할 경우에는 군집 개수를 2개로 하는 것이 가장 좋아 보임
- 데이터별로 다른 데이터와의 거리를 반복적으로 계산해야 하므로 데이터 양이 늘어나면 수행 시간이 크게 늘어남
-

3. 평균 이동

평균 이동: 평균 이동 군집화는 데이터의 분포도를 이용하여 군집 중심점을 찾음

군집 중심점은 데이터 포인트가 모여있는 곳임

확률 밀도 함수 이용

KDE 이용

- 과정

  개별 데이터들이 모인 중심점을 군짐 중심점으로 설정

➔ 지정된 반복 횟수만큼 전체 데이터에 대해서 KDE 기반으로 데이터를 이동시키면서 군집화 수행

➔ 모든 데이터를 1~2까지 수행하면서 데이터 이동, 개별 데이터들이 군집중심점으로 모임

➔ KDE로 계산된 데이터 분포도가 높은 방향으로 데이터 이동

➔ 개별 데이터의 특정 반경 내에 주변 데이터를 포함한 데이터 분포도를 KDE 기반의 Mean Shift 알고리즘으로 계산

K: 커널 함수
x: 확률 변숫값
xi: 관측값
h: 대역폭

- 평균 이동 군집화는 대역폭이 클수록 평활화된 KDE로 인해 적은 수의 군집 중심점을 가지며 대역폭이 적을수록 많은 수의 군집 중심점을 가짐
- 유연한 군집화 가능
- 이상치 영향력도 적고 미리 군집 개수 정할 필요도 없음

4. GMM
   1) GMM: 군집화를 적용하고자 하는 데이터가 여러 개의 가우시안 분포를 가진 데이터 집합들이 섞여서 생성된 것이라는 가정 하에 군집화 수행
      - 가우시안 분포 == 정규분포
      - 섞인 데이터에서 개별 유형의 가우시안 분포를 추출
      - 클러스터링 하려는 데이터 분포에서 개별 정규 분포 찾고 데이터가 특정 정규 분포에 해당될 확률 구함
      - 모수 추정
        (1) 개별 정규 분포의 평균과 분산
        (2) 각 데이터가 어떤 정규 분포에 해당되는지의 확률
   2) GMM과 K-평균
      KMeans는 원형의 범위에서 군집화를 수행함
      but 길쭉한 타원형일 때는 수행을 잘 못함

5. DBSCAN
   내부의 원 모양과 외부의 원 모양 형태의 분포를 가진 데이터 세트를 군집화한다

입실론 주변 영역	개별 데이터를 중심으로 입실론 반경을 가지는 원형의 영역
최소 데이터 개수	개별 데이터의 입실론 주변 영역에 포함되는 타 데이터의 개수

핵심 포인트	주변 영역 내 최소 데이터 개수 이상의 타 데이터를 가지고 있을 경우 해당 데이터를 핵심 포인트라고 함
이웃 포인트	주변 영역 내에 위치한 타 데이터를 이웃 포인트라고 함
경계 포인트	주변 영역 내에 최소 데이터 갯 이상의 이웃 포인트를 가지고 있지 않지만 핵심 포인트를 이웃 포인트로 가지고 있는 데이터를 경계 포인트라고 함
잡음 포인트	최소 데이터 개수 이상의 이웃 포인트를 가지고 있지 않으며, 핵심 포인트도 이웃 포인트로 가지고 있지 않는 데이터

eps	입실론 주변 영역의 반경을 의미
Min_samples	핵심 포인트가 되기 위해 입실론 주변 영역 내에 포함되어야 할 데이터의 최소 개수를 의미

6. 실습 – 고객 세그멘테이션
   1) 고객 세그먼테이션: 다양한 기준으로 고객을 분류하는 기법
   2) 타깃 마케팅이 중요
   3) RFM 이용
      가장 최근 상품 구입 일에서 오늘까지의 기간, 상품 구매 횟수, 총 구매 금액