11주차 예습

날짜	@2024년 5월 21일
태그	예습

7장. 군집화

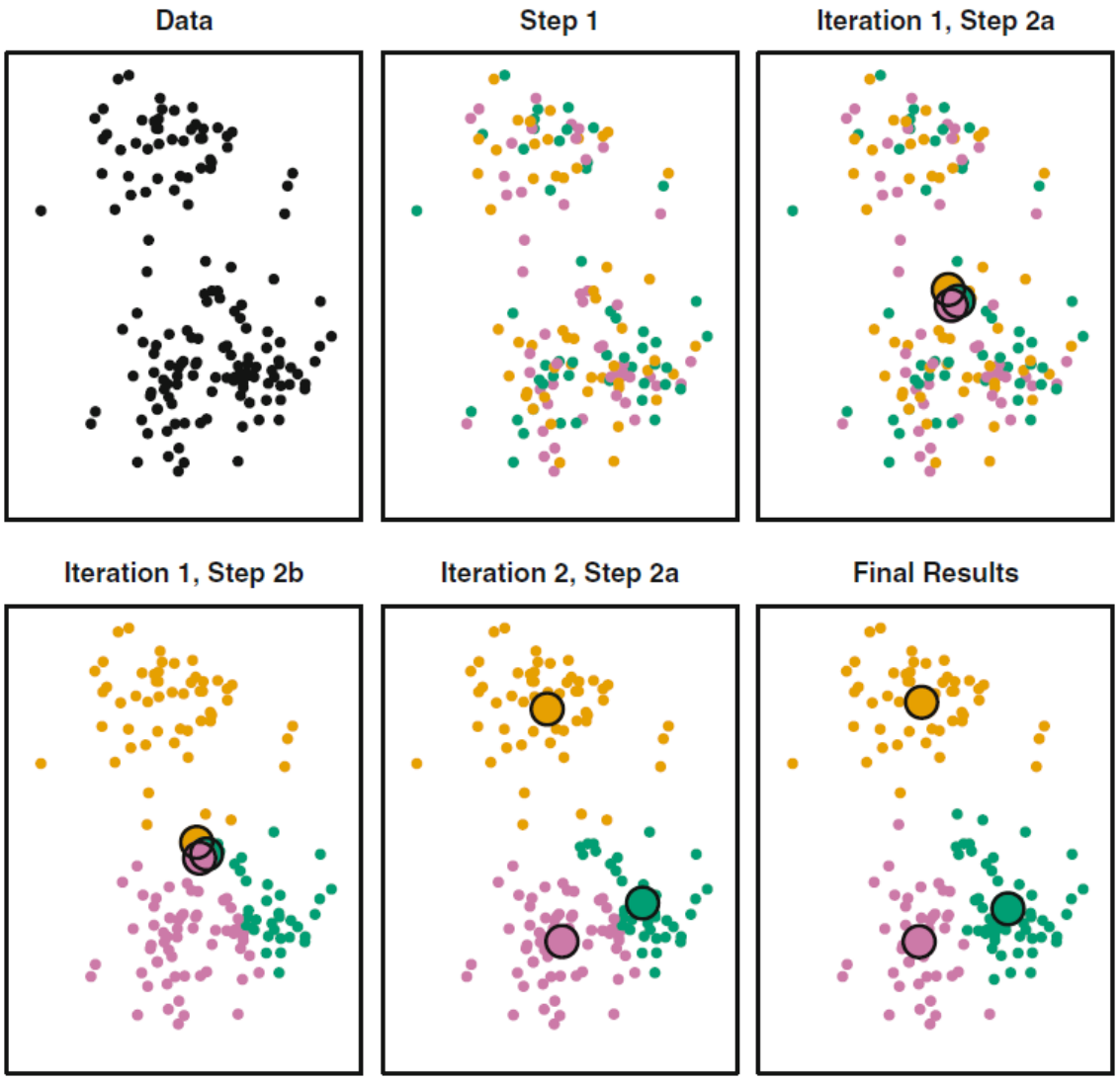
▼ 7.1 K-평균 알고리즘 이해

K-평균

- 군집화(Clustering)에서 가장 일반적으로 사용되는 알고리즘
- **군집 중심점**이라는 특정한 임의의 지점을 선택해 해당 중심에 가장 가까운 포인트들을 선택하는 군집화 기법

K-평균 작동 프로세스

1. 몇 개의 덩어리로 clustering할 지 정한다.
2. 1번에서 정한 개수만큼 중심점을 정한다. 이때, 중심점이 군집 중심점(centroid)이다.
3. 각 점(sample/point)마다 가까운 군집 중심점을 정한다.
4. 매핑된 점들을 바탕으로 하여 군집 중심점을 이동한다. 이때, 새로 만들어지는 군집 중심점은 같은 색으로 매핑된 점들의 평균이 된다.
5. 3번~4번의 과정을 더 이상 새로 매핑되지 않을 때까지 반복한다.



장점

- 일반적인 군집화에서 가장 많이 활용되는 알고리즘
- 알고리즘이 쉽고 간결

단점

- 거리 기반 알고리즘으로 속성의 개수가 매우 많을 경우 군집화 정확도가 떨어짐
  > PCA로 차원 감소를 적용해야할 수도 있음
- 반복을 수행하는데, 반복 횟수가 많을 경우 수행 시간이 매우 느려짐
- 몇 개의 군집을 선택해야 할지 가이드하기가 어려움

사이킷런 KMeans 클래스 소개

KMeans 클래스의 초기화 파라미터

```
class sklearn.cluster.KMeans(n_cluster=8, init='k-means++', n_init=10, max_iter=300, tol=0.001, precompute_distances='auto', verbose=0, random_state=None, copy_x=True, n_jobs=1, algorithm='auto')
```

n_clusters

군집화할 개수, 즉 군집 중심점의 개수를 의미

init

초기에 군집 중심점의 좌표를 설정할 방식
일반적으로 k-means++ 방식으로 최초 설정

max_iter

최대 반복 횟수
이 횟수 이전에 모든 데이터의 중심점 이동이 없으면 종료함

✅ **fit**, **fit_transform** 메서드를 이용해 수행

군집화와 관련된 주요 속성

- labels_: 각 데이터 포인트가 속한 군집 중심점 레이블
- cluster_centers_: 각 군집 중심점 좌표 [군집 개수, 피처 개수]
  이를 이용해 군집 중심점 좌표 시각화 가능

군집화 알고리즘 테스트를 위한 데이터 생성

군집화용 데이터 생성기

- make_blobs( )
  개별 군집의 중심점과 표준 편차 제어 기능

- make_classification( )
  노이즈를 포함한 데이터를 만드는 데 유용

여러 개의 클래스에 해당하는 데이터 세트 생성
하나의 클래스에 여러 개의 군집이 분포될 수 있게 데이터 생성 가능

- make_circle( )
- make_moon( )

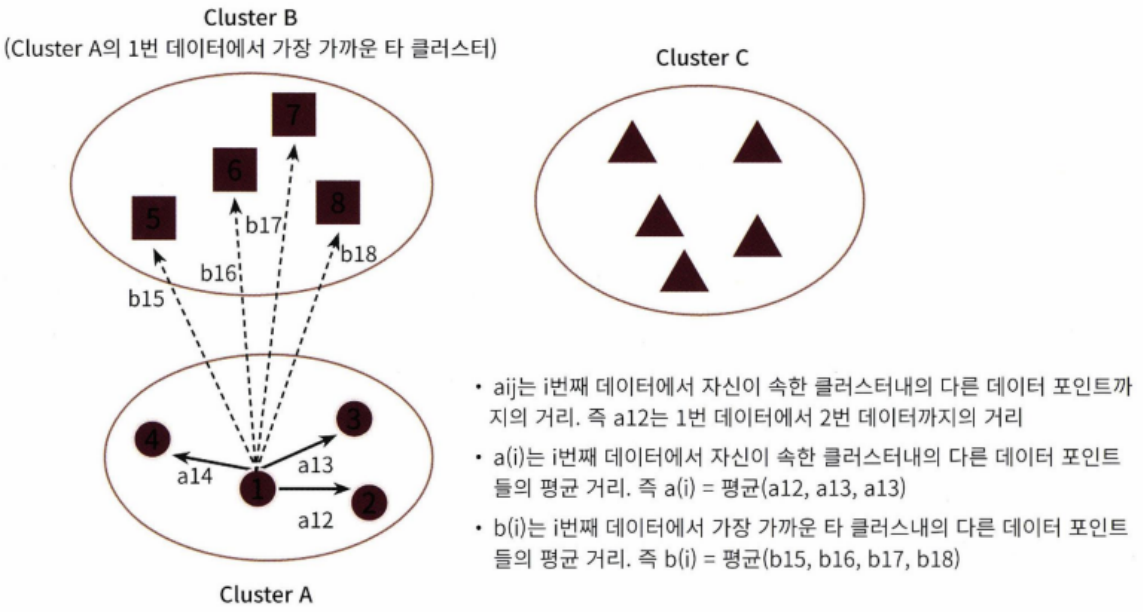
중심 기반의 군집화로 해결하기 어려운 데이터 세트를 만드는 데 사용

▼ 7.2 군집 평가

실루엣 분석

- 각 군집 간의 거리가 얼마나 효율적으로 분리돼 있는지 나타냄
- 즉, 다른 군집과의 거리는 떨어져 있고 동일 군집끼리의 데이터는 서로 가깝게 잘 뭉쳐져 있는지를 확인하여 나타냄
- 군집화가 잘 될 수록 개별 군집은 비슷한 정도의 여유공간을 가지고 떨어져 있을 것임
- **실루엣 계수**(개별 데이터가 가지는 군집화 지표)를 기반으로 진행

실루엣 계수 $s(i)$



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

- -1과 1 사이의 값을 가지며, 1로 가까워질수록 근처의 군집과 더 멀리 떨어져 있다는 뜻이고 0에 가까울수록 근처의 군집과 가까워진다는 것이다.
- -값은 아예 다른 군집에 데이터 포인트가 할당됐음을 뜻한다.

사이킷런에서의 실루엣 분석

실루엣 분석 메서드

```
sklearn.metric.silhouette_samples(X, lables, metric='euc
lidean', **kwds)
```

인자로 X feature 데이터 세트와 각 피처 데이터 세트가 속한 군집 레이블 값인 labels 데이터를 입력해주면 각 데이터 포인트의 실루엣 계수를 계산해 반환

```
sklearn.metrics.silhouette_score(X, labels, metric='ducl
idean', sample_size=None, **kwds)
```

인자로 X feature 데이터 세트와 각 피처 데이터 세트가 속한 군집 레이블 값인 labels 데이터를 입력해주면 전체 데이터 실루엣 계수 값을 평균해 반환
즉, np.mean(silhouette_samples()). 높을수록 군집화가 어느정도 잘 됐다고 판단 가능하나 무조건 높다고 군집화가 잘 된 것은 아님.

좋은 군집화의 조건

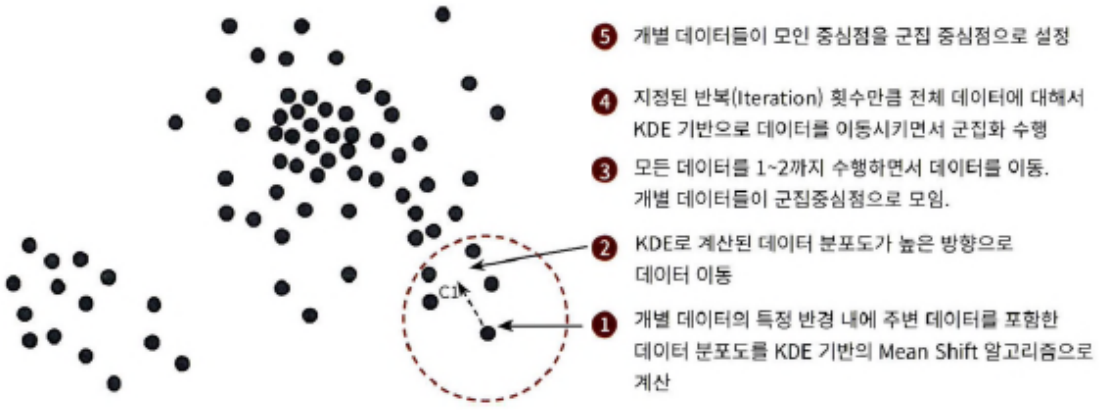
1. 전체 실루엣 계수의 평균값, 즉 silhouette_score( ) 값은 0~1 사이의 값을 가지며, 1에 가까울수록 좋음
2. 전체 실루엣 계수의 평균과 더불어 개별 군비의 평균값의 편차가 크지 않아야 함
   즉, 개별 군집의 실루엣 계수의 평균값이 전체 실루엣 계수의 평균값에서 크게 벗어나지 않아야 함

▼ 7.3 평균 이동

- K-평균과 유사하게 중심을 군집의 중심으로 지속적으로 움직이면서 군집화를 수행
- 이동 시 데이터가 모여 있는 밀도가 가장 높은 곳으로 이동

평균 이동 군집화

- 데이터의 분포도를 이용해 군집 중심점을 찾는다
- 확률 밀도 함수가 가장 피크인 점을 군집 중심점으로 선정
- 특정 데이터를 반경 내 데이터 분포 확률 밀도가 가장 높은 곳으로 이동하기 위해 주변 데이터와의 거릭 밧을 KDE 함수값으로 입력한 뒤 그 반환 값을 현재 위치에서 업데이트 하면서 이동하는 방식을 가짐

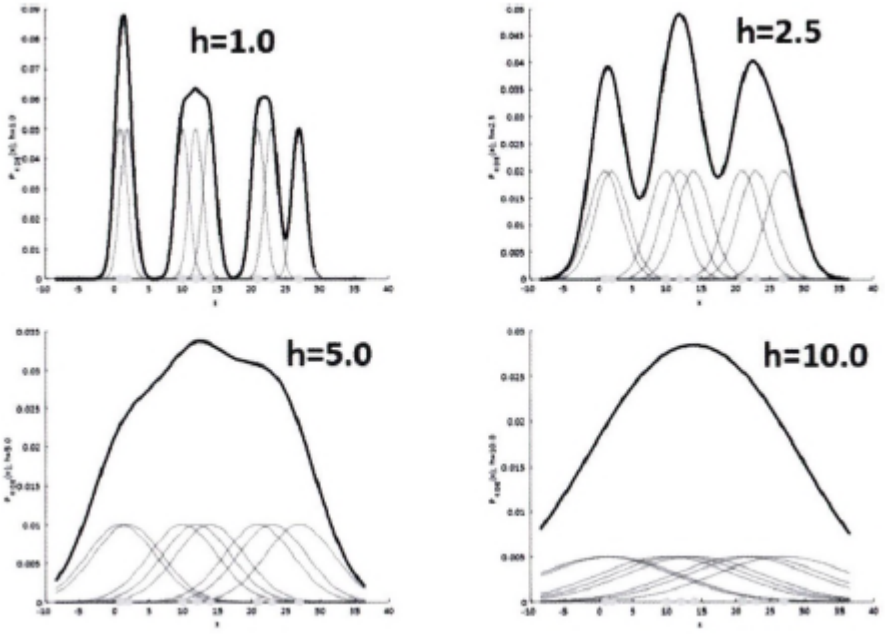


KDE(Kernel Density Estimation)

: 커널 함수를 통해 어떤 변수의 확률 밀도 함수를 추정하는 대표적 방법

$$KDE = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h})$$

- 관측된 데이터 각각에 커널 함수 적용 값을 모두 더한 후 데이터 건수로 나눠 확률 밀도 함수 추정
- 확률 밀도 함수를 통해 변수의 특성, 확률 분포 등 변수의 다양한 요소 알 수 있음
- 대표적인 커널 함수: 가우시안 분포 함수
- h: KDE 형태를 부드러운 형태로 평활화하는 데 적용되며 h 설정 값에 따라 확률 밀도 추정 성능을 크게 좌우할 수 있음



▼ 7.4 GMM(Gaussian Mixture Model)

GMM 군집화

: 군집화를 적용하고자 하는 데이터가 여러 개의 가우시안 분포를 가진 데이터 집합들이 섞여서 생성된 것이라는 가정하에 군집화를 수행하는 방식

- 모수 추정: 섞인 데이터 분포에서 개별 유형의 가우시안 분포 추출 후 개별 데이터가 그 중 어느 정규 분포에 속하는지 결정하는 방식
- 대표적으로 2가지를 추정 함
  1. 개별 정규 분포의 평균과 분산
  2. 각 데이터가 어떤 정규 분포에 해당되는지의 확률

GMM과 K-평균의 비교

KMeans

- 평균 거리 기반으로 군집화
- 원형의 범위에서 군집화 수행
  데이터 세트가 원형의 범위를 가질 수록 군집화 효율 증가
- 원형의 범위가 아닌 경우 군집화 정확성이 떨어짐

GMM

- 유연하게 다양한 데이서 셑트에 잘 적용 가능
- 수행 시간이 오래 걸림

▼ 7.5 DBSCAN

DBSCAN(Density Based Spatial Clustering of Applications with Noise)

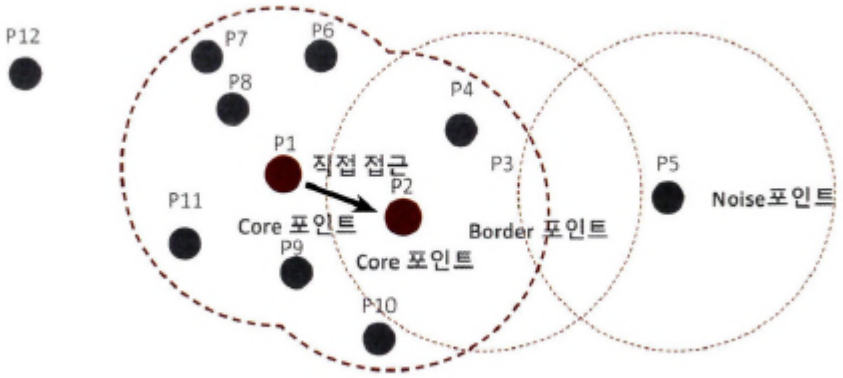
- 밀도 기반 군집화의 대표적 알고리즘
- 간단하고 직관적인 알고리즘
- 데이터의 분포가 기하학적으로 복잡한 데이터 세트에도 효과적인 군집화 가능

중요 파라미터

- 입실론 주변 영역(epsilon): 개별 데이터를 중심으로 입실론 반경을 가지는 원형의 영역
- 최소 데이터 개수(min points): 개별 데이터의 입실론 주변 영역에 포함되는 타 데이터의 개수

데이터 포인트

- 핵심 포인트(Core Point): 주변 영역 내에 최소 데이터 개수 이상의 타 데이터를 가지고 있는 경우
- 이웃 포인트(Neighbor Point): 주변 영역 내 위치한 타 데이터
- 경계 포인트(Border Point): 주변 영역 내 치소 데이터 개수 이상의 이웃 포인트를 가지고 있지 않지만 핵심 포인트를 이웃 포인트로 가지고 있는 데이터
- 잡음 포인트(Noise Point): 최소 데이터 개수 이상의 이웃 포인트를 가지고 있지 않으며 핵심 포인트도 이웃 포인트로 가지고 있지 않은 데이터



1. 최소 데이터 세트를 6개로 가정
2. P1: 입실론 반경 내 포함된 데이터 7개 > 최소 데이터 6개 이므로 핵심 포인트
3. P2: 반경 내 6개의 데이터, 핵심 포인트
4. P1의 이웃 데이터 포인트 P2가 핵심 포인트이므로 직접 접근 가능
5. 직접 접근이 가능한 다른 핵심 포인트를 서로 연결하면서 군집화 구성
6. P3: 반경 내 아웃 데이터는 2개(P2, P4)지만 이웃 데이터 중 핵심 포인트를 가지므로 경계 포인트
7. P5: 반경 내 최소 데이터를 가지지도 않으며 핵심 포인트를 이웃 포인트로 가지고 있지도 않는 데이터이므로 잡음 포인트

사이킷런: DBSCAN 클래스

주요 초기화 파라미터

- eps: 입실론 주변 영역의 반경
- min_samples: 핵심 포인트가 되기 위해 입실론 주변 영역에 포함돼야 할 최소 데이터 개수(자신의 데이터 포함)

▼ 7.6 군집화 실습 - 고객 세그먼테이션

: 다양한 기준으로 고객을 분류하는 기법

- 중요 분류 요소: 어떤 상품을 얼마나 많은 비용을 써서 얼마나 자주 사용하는가에 기반한 정보로 분류
  즉, 얼마나 많은 매출을 발생하는가
- 주요 목표; 타깃 마케팅

    - 타킷 마케팅: 고객을 여러 특성에 맞게 세분화해서 그 유형에 따라 맞춤형 마케팅이나 서비스를 제공하는 것
- RFM: 가장 기본적인 고객 분석 요소
    - RECENCY(R): 가장 최근 상품 구입 일에서 오늘까지의 기간
    - FREQUENCY(F): 상품 구매 횟수
    - MONETARY VALUE(M): 총 구매 금액