파머완 7장: 군집화

상태	진행 중
담당자	시현 이
마감일	@11/17/2025
작업 유형	개념정리 및 필사
설명	파이썬 머신러닝 완벽 가이드_개정2판_제7장 개념 정리+ 필사 링크
업데이트 시간	@November 18, 2025 1:06 PM

01. K-평균 알고리즘 이해

K-평균

- 군집화(Clustering)에서 가장 일반적으로 사용되는 알고리즘
- 군집 중심점(centroid)이라는 특정한 임의의 지점을 선택해 해당 중심에 가장 가까운 포인트들을 선택하는 군집화 기법
    - 군집중심점: 선택된 포인트의 평균 지점으로 이동하고 이동된 중심점에서 다시 가까운 포인트를 선택, 다시 중심점을 평균 지점으로 이동하는 프로세스를 반복적으로 수행
    - 더이상 중심점의 이동이 없을 시, 반복을 멈추고 해당 중심점에 속하는 데이터 포인트들을 군집화하는 기법
    - **사진**

1. 먼저 군집화의 기준이 되는 중심을 구성하려는 군집화 개수만큼 임의의 위치(적합한 위치)에 가져다 놓음. (ex. 전체 데이터를 2개로 군집화→ 2개의 중심을 임의의 위치에 가져다 놓음
2. 각 데이터는 가장 가까운 곳에 위치한 중심점에 소속됨. (그림에서 A,B 데이터가 같은 중심점/C,E,F 데이터가 같은 중심점)
3. 소속이 결정되면 군집 중심점을 소속된 데이터의 평균 중심으로 이동.
4. 중심점이 이동한 후, 각 데이터는 기존에 속한 중심점보다 더 가까운 중심점이 있다면 해당 중심점으로 다시 소속을 변경. 위 그림에서는 C데이터가 기존의 중심점보다 더 가까운 중심점으로 변경됨

5. 다시 중심을 소속된 데이터의 평균 중심으로 이동.
6. 중심점을 이동했는데 데이터의 중심점 소속 변경이 없으면 군집화를 종료. 그렇지 않을 시, 4번~.. 과정 반복

- K-평균의 장점
  - 일반적인 군집화에서 가장 많이 활용되는 알고리즘
  - 알고리즘이 쉽고 간결
- K-평균의 단점
  - 거리 기반의 알고리즘으로 속성의 개수가 매우 많을 경우 군집화 정확도가 떨어짐(PCA로 차원 감소를 적용해야할 수 있음)
  - 반복 횟수가 많을 경우, 수행 시간이 매우 느림
  - 몇 개의 군집을 선택해야 할지 가이드하기 어려움

02. 군집 평가(Cluster Evaluation)

대부분의 군집화 데이터 세트는 비교할 만한 타깃 레이블을 가지고 있지 않음.

또한, 군집화 ≠ 분류. 데이터 내에 숨어있는 별도의 그룹을 찾아서 의미를 부여하거나 동일한 분류 값에 속하더라도 그 안에서 더 세분화된 군집화를 추구하거나 서로 다른 분류 값의 데이터도 더 넓은 군집화 레벨화 등의 영역을 가지고 있음.

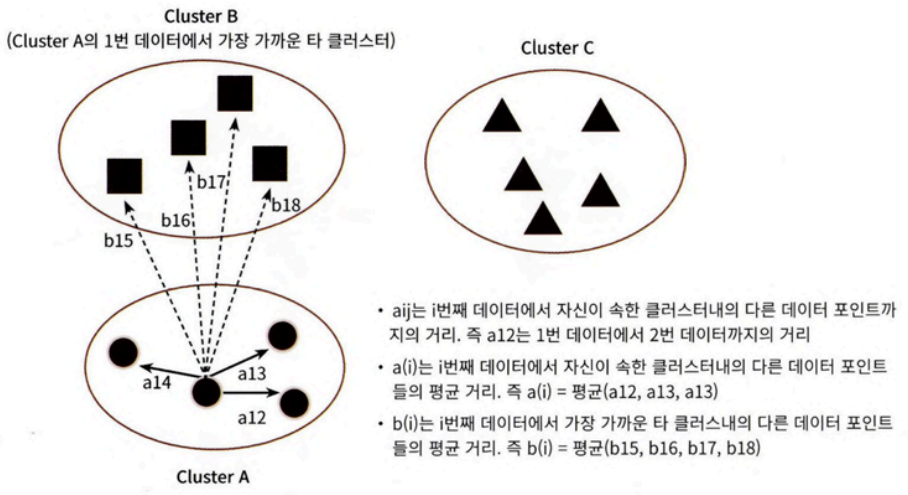
→군집화가 효율적으로 이뤄졌는지 평가하는 지표: 실루엣 분석

실루엣 분석의 개요

실루엣 분석

: 각 군집 간의 거리가 얼마나 효율적으로 분리(다른 군집과의 거리는 떨어져있고 동일 군집끼리의 데이터는 가깝게 뭉쳐있는 상태) 돼 있는지를 나타냄

- 실루엣 계수(silhouette coefficient)를 기반으로 함
  - 개별 데이터가 가지는 군집화 지표
  - 해당 데이터가 같은 군집 내의 데이터와 얼마나 가깝게 군집화돼 있고, 다른 군집에 있는 데이터와는 얼마나 멀리 분리돼 있는지를 나타내는 지표



- a(i) = 같은 군집 내 다른 데이터 포인트와의 거리 평균
- b(i) = 다른 군집 내 가장 가까운 군집과의 평균 거리
- a(i)와 b(i) 기반으로 계산.
  - 두 군집 간의 거리 b(i) - a(i) ==정규화⇒ MAX(a(i), b(i))
  - i번째 데이터 포인트의 실루엣 계수 s(i)

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

  - -1~1사이의 값을 가짐.
    - 1에 가까울수록 근처의 군집과 더 멀리 떨어져 있다는 것.
    - 0에 가까울수록 근처의 군집과 가까워진다는 것
    - - 값은 아예 다른 군집에 데이터 포인트가 할당됐음을 뜻함
- 사이킷런에서의 메서드
  - sklearn.metrics.silhouette_samples(X, labels, metric = ‘euclidean’, **kwds): 인자로 X feature 데이터 세트와 각 피처 데이터 세트가 속한 군집 레이블 값인 labels 데이터를 입력해주면 각 데이터 포인트의 실루엣 계수를 계산해 반환함
  - sklearn.metrics.silhouette_score(X, labels, metric = ‘euclidean’, sample_size = None, **kwds): 인자로 X feature 데이터 세트와 각 피처 데이터 세트가 속한 군집 레이블 값인 labels 데이터를 입력해주면 전체 데이터의 실루엣 계수 값을 평균해 반환.
    np.mean(silhouette_samples()) ⇒이 값이 높을 수록 군집화가 잘됐다고 판단할 수 있지만 이 값이 높다고 해서 무조건 군집과가 잘 됐다고 판단할순 X

- 좋은 군집화를 위한 기준 조건
  1. 전체 실루엣 계수의 평균값 silhouette_score() 값은 0~1사이의 값을 가짐. 1에 가까울수록 좋음
  2. 전체 실루엣 계수의 평균값과 더불어 개별 군집의 평균값의 편차가 크지 않아야함. 즉, 개별 군집의 실루엣 계수 평균값이 전체 실루엣 계수의 평균값에서 크게 벗어나지 않는 것이 중요. if, 전체 실루엣 계수의 평균값이 높지만, 특정 군집의 실루엣 계수 평균값만 유난히 높고 다른 군집들의 실루엣 계수 평균값은 낮으면 좋은 군집화 조건이 아님

군집별 평균 실루엣 계수의 시각화를 통한 군집 개수 최적화 방법

데이터의 평균 실루엣 계수값이 높다고 해서 군집화가 잘된게X

⇒ 개별 군집별로 적당히 분리된 거리를 유지하면서도 군집 내의 데이터가 뭉쳐 있는 경우 K-평균의 적절한 군집 계수 설정됨

[여러 개의 군집 개수가 주어졌을 때 평균 실루엣 계수로 군집 개수를 최적화하는 방법]

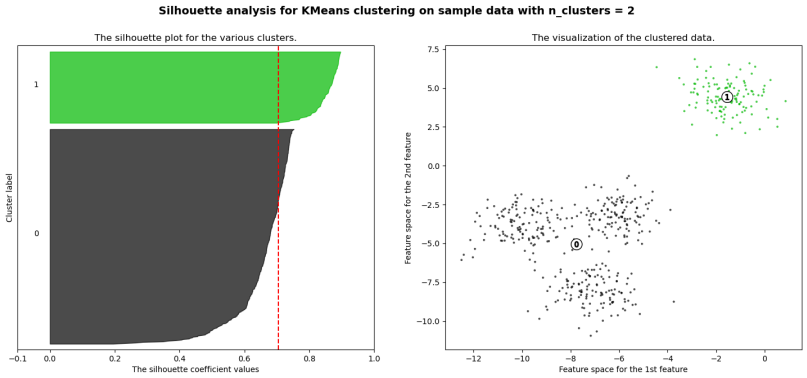
Selecting the number of clusters with silhouette analysis on KMeans clustering

Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the ne...

http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

:해당 링크에 시각화 코드 존재

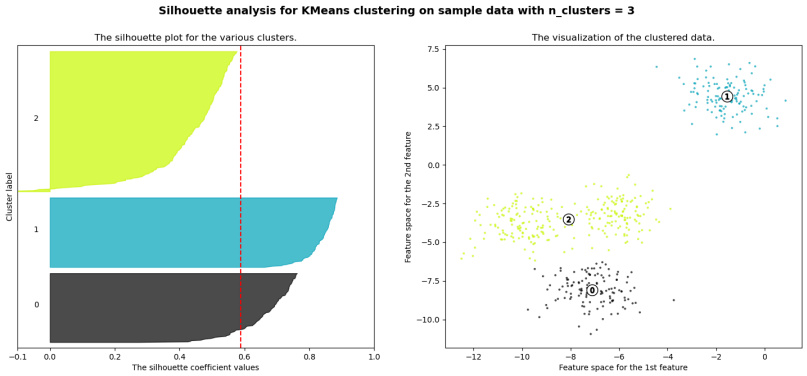
1. 주어진 데이터에 대해서 군집의 개수 2개를 정했을 때



- 평균 실루엣 계수, silhouette_score는 약 0.704로 매우 높게 나타남
- 왼쪽 그림: 개별 군집에 속하는 데이터의 실루엣 계수를 2차원을 나타낸 것

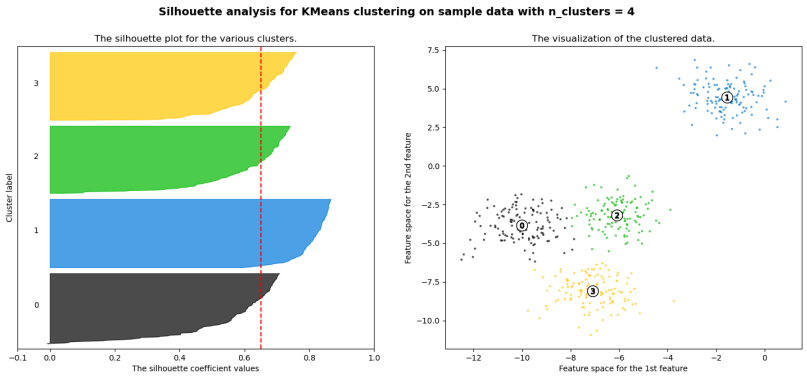
- x축: 실루엣 계수 값/ y축: 개별 군집(0/1)과 이에 속하는 데이터(높이로 추측)
- 점선: 전체 평균 실루엣 계수 값
- 1번 군집의 모든 데이터는 평균 실루엣 계수 값 이상, 2번 군집의 경우 평균보다 적은 데이터 값이 많음

- 오른쪽 그림:
  - 1번 군집-0번 군집과 멀리 떨어져있고, 내부 데이터끼리도 잘뭉쳐짐
  - 0번 군집-내부 데이터끼리 많이 떨어져있음



2. 군집 개수가 3개일 경우

- 전체 데이터의 평균 실루엣 계수 값은 약 0.588
  - 1번, 2번 군집의 경우 평균보다 높은 실루엣 계수 값.
  - 0번의 경우 모든 평균보다 낮음
  - 오른쪽 그림: 0번의 경우 내부 데이터 간의 거리도 멀지만, 2번 군집과도 가깝게 위치



3. 군집이 4개인 경우

- 평균 실루엣 계수 값은 0.65
- 개별 군집의 평균 실루엣 계수 값이 비교적 균일하게 위치
    - 1번 군집의 경우 모든 데이터가 평균보다 높은 계수 값을 가짐
    - 0번, 2번의 경우 절반 이상이 평균보다 높은 계수 값을 가짐
    - 3번 군집의 경우만 약 1/3 정도가 평균보다 높은 계수 값을 가지고 있음
    - 군집이 2개인 경우보다는 평균 실루엣 계수 값이 작지만 4개인 경우가 가장 이상적인 군집화 개수로 판단할 수 있음

*실루엣 계수를 통한 K-평균 군집 평가 방법은 직관적으로 이해하기 쉽지만, 각 데이터 별로 다른 데이터와의 거리를 반복적으로 계산해야하므로 데이터양이 늘어나면 수행시간이 크게 늘어남

-몇 만건 이상의 데이터에 대해 사이킷런의 실루엣 계수 평가 API를 개인용 PC에서 수행할 경우 메모리 부족 등의 에러가 발생하기 쉬움

03.평균 이동

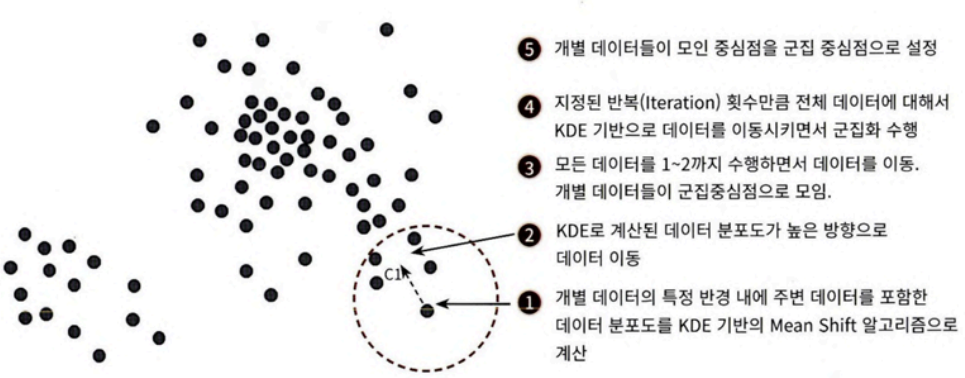
평균 이동(Mean Shift)의 개요

평균이동

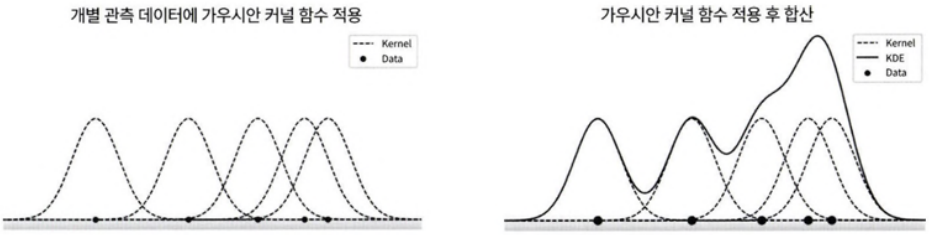
: K-평균과 유사하게 중심을 군집의 중심으로 지속적으로 움직이면서 군집화 수행

but, 평균 이동은 중심을 데이터가 모여 있는 밀도가 가장 높은 곳으로 이동시킴

- 데이터의 분포도를 이용해 군집 중심점을 찾음
    - 군집 중심점: 데이터 포인트가 모여있는 곳. 확률 밀도 함수 이용→확률밀도 함수가 피크인 점을 군집 중심점으로 선정하며 일반적으로 KDE(Kernel Density Estimation)를 이용



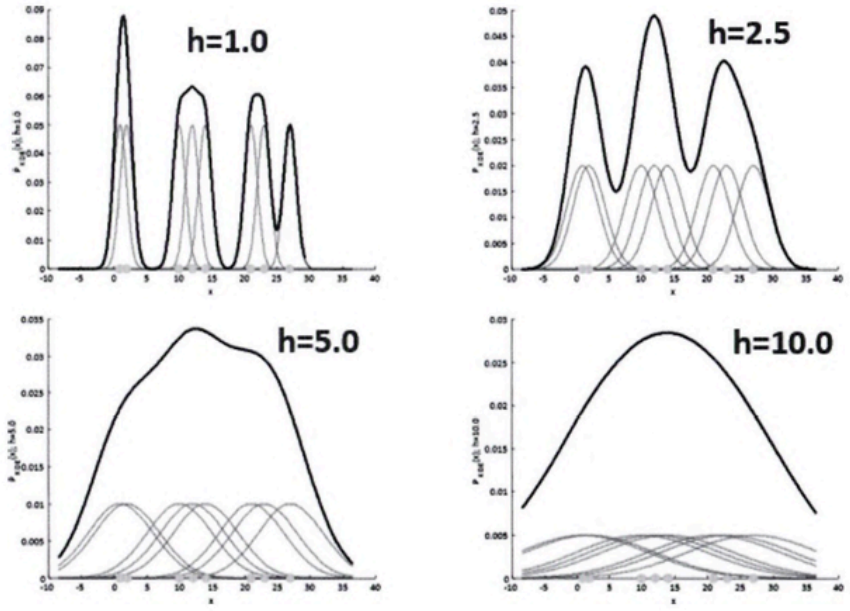
- KDE: 커널 함수를 통해 어떤 변수의 확률 밀도 함수를 측정하는 대표적인 방법
  - 관측된 데이터 각각에 커널 함수를 적용한 값을 모두 더한 뒤 데이터 건수로 나눠 확률 밀도 함수를 측정
  - 대표적인 커널 함수: 가우시안 분포 함수



- 확률밀도 함수 PDF(Probability Density Function)는 확률 변수의 분포를 나타내는 함수
  - 정규분포 함수, 감마 분포, t-분포 등 존재
  - 특정 변수가 어떤 값을 갖게 될지에 대한 확률을 알게 되므로 변수의 특성(평균, 분산 등), 확률 분포 등 변수의 많은 요소를 얻을 수 있음
- KDE 커널함수식
  - K = 커널 함수, x = 확률 변숫값, xi = 관측값, h = 대역폭(bandwidth)

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

- h : KDE 형태를 부드러운(or 뾰족한) 형태로 평활화(Smoothing)하는데 적용
  - 확률 밀도 추정 성능 크게 좌우
  - 작은 h값(h = 1.0)은 좁고 뾰족한 KDE. 변동성이 큰 방식으로 확률 밀도 함수를 추정→ 과적합 쉬움
  - 큰 h값(h=10)은 과도하게 평활화된 KDE.→과소적합 쉬움



- 평균 이동 군집화는 대역폭 클수록→적은 수의 군집 중심점//대역폭 적을수록 → 많은 수의 군집 중심점
- 군집의 개수 지정X. 오직 대역폭의 크기에 따라 군집화 수
- 평균 이동의 장점
  - 데이터 세트의 형태를 특정 형태로 가정한다든가, 특정 분포도 기반의 모델로 가정하지 않기 때문에 좀 더 유연한 군집화 가능
  - 이상치의 영향력이 크지 않음
  - 미리 군집의 개수 정할 필요X
- 평균 이동의 단점
  - 알고리즘의 수행 시간 오래 걸림
  - bandwidth의 크기에 따른 군집화 영향도 매우 큼

⇒ 컴퓨터 비전 영역에서 더 많이 사용됨. 이미지/영상 데이터에서 특정 개체 구분하거나 움직임을 추적하는데 뛰어난 역할을 수행하는 알고리즘

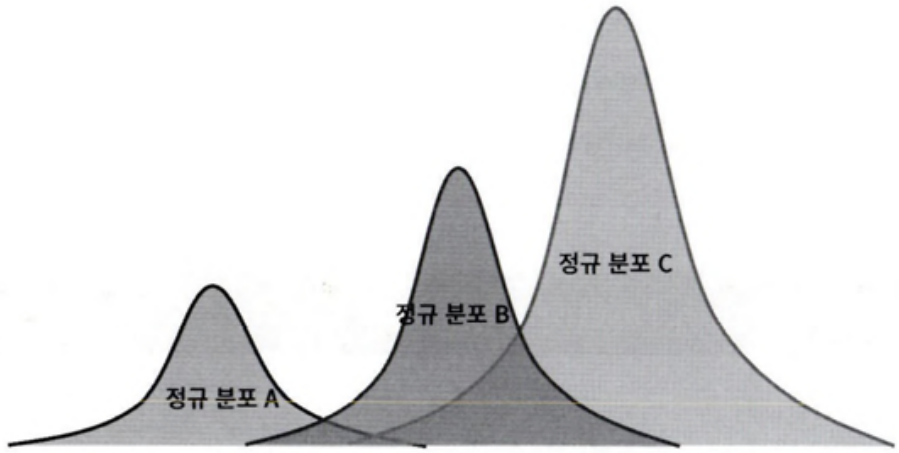
04.GMM(Gaussian Mixture Model)

GMM(Gaussian Mixture Model) 소개

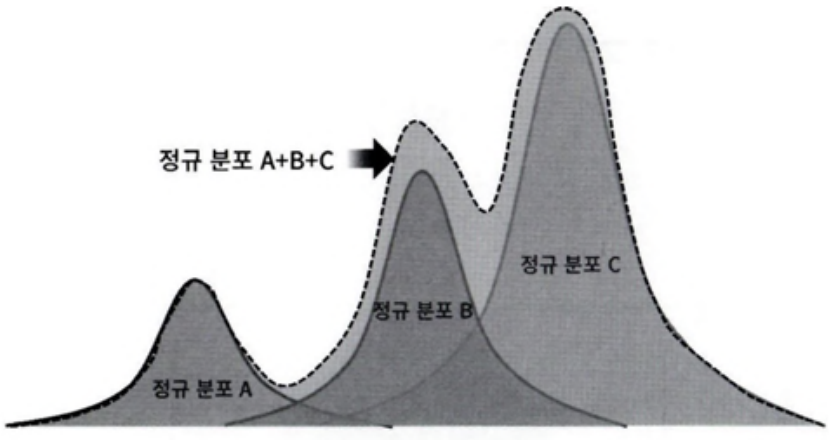
GMM 군집화

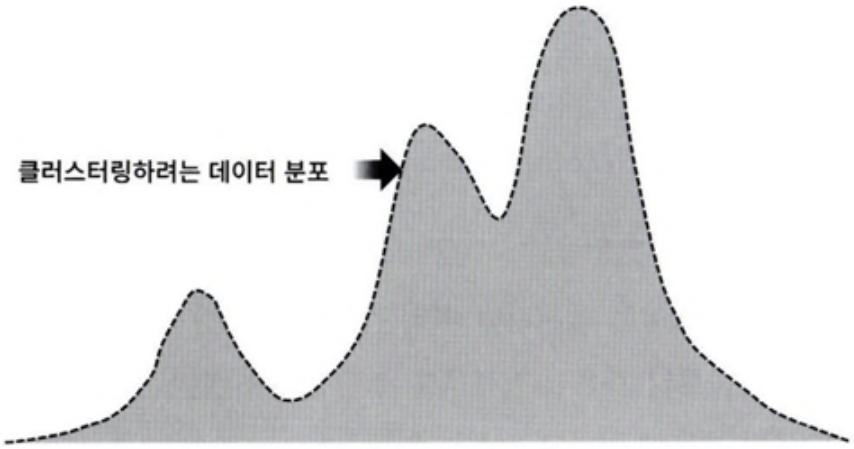
: 군집화를 적용하고자 하는 데이터가 여러 개의 가우시안 분포를 가진 데이터 집합들이 섞여서 생성된 것이라는 가정하에 군집화를 수행하는 방식

- 가우시안 분포: 좌우 대칭형의 종 형태 연속 확률 함수
  - 정규 분포는 평균을 중심으로 높은 데이터 분포도. 좌우 표준편차 1에 전체 데이터의 68.27%, 좌우 표준편차 2에 전체 데이터의 95.45%를 가지고 있음
  - 평균 = 0, 표준편차 = 1인 정규 분포를 표준 정규 분포
- 세 개의 가우시안 분포 A,B,C를 가진 데이터 세트가 있다고 가정

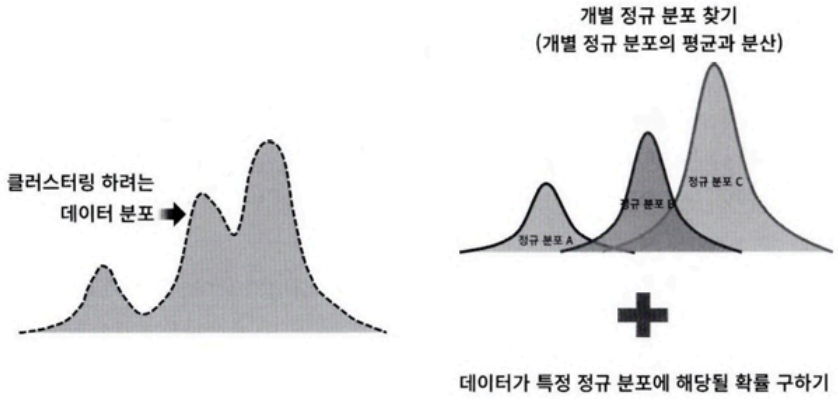


이 세 개의 정규분포 합치면





- 서로 다른 정규 분포에 기반해 군집화를 수행하는 것이 GMM 군집화 방식

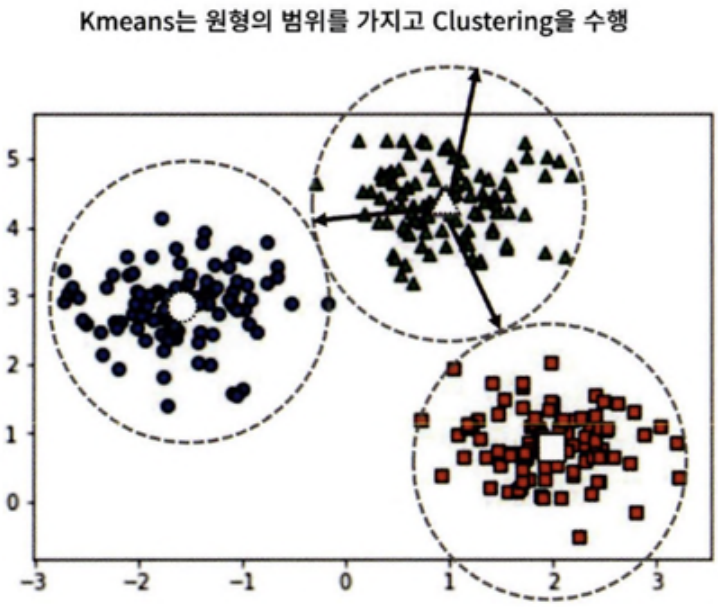


- 모수 측정: 데이터 세트를 구성하는 여러 개의 정규 분포 곡선을 추출하고, 개별 데이터가 이 중 어떤 정규분포에 속하는지 결정하는 방식
  - 추정 방식
    - 개별 정규 분포의 평균과 분산
    - 각 데이터가 어떤 정규 분포에 해당되는지의 확률
  - 모수 측정을 위해 EM(Expectation and Maximization) 방법을 적용
    - 사이킷런은 GaussianMixture 클래스를 지원

GMM과 K-평균의 비교

KMeans는 원형의 범위에서 군집화를 수행

→ 데이터 세트가 원형의 범위를 가질수록 KMeans의 군집화 효율을 더욱 높아짐



- make_blobs()의 군집 수를 3개, cluster_std = 0.5로 설정해 군집 내 데이터를 뭉치게 유도한 데이터셋에 KMeans를 적용한 결과

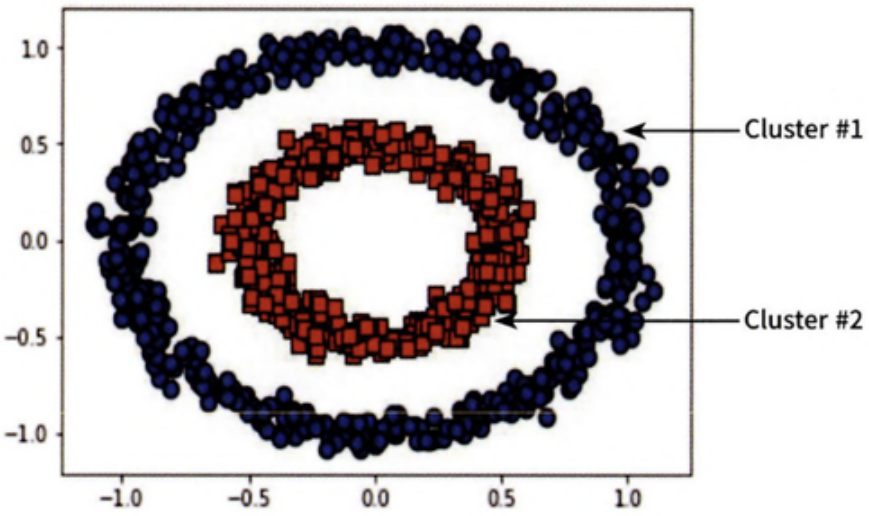
→데이터가 원형 형태로 퍼져있지 않을 때 군집화 수행 어려움 존

05.DBSCAN

DBSCAN 개요

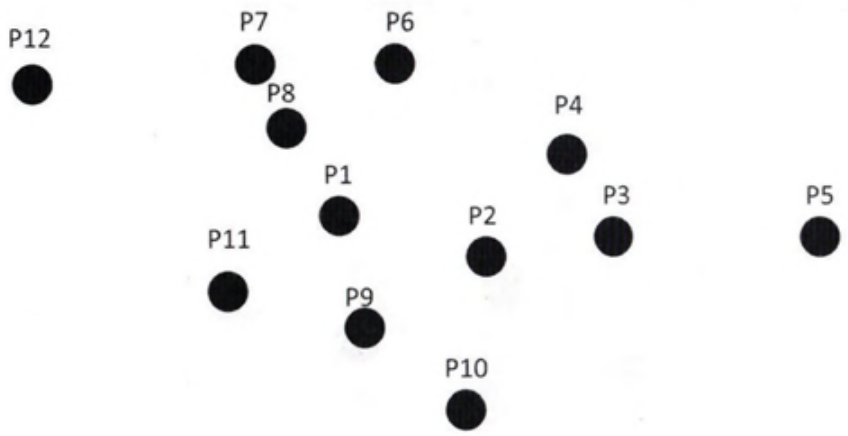
DBSCAN(Density Based Spatial Clustering of Application with Noise)

: 특정 공간 내에 데이터 밀도 차이를 기반 알고리즘으로 하고 있어서 데이터 분포가 기하학적으로 복잡한 데이터 세트에도 효과적인 군집화가 가능

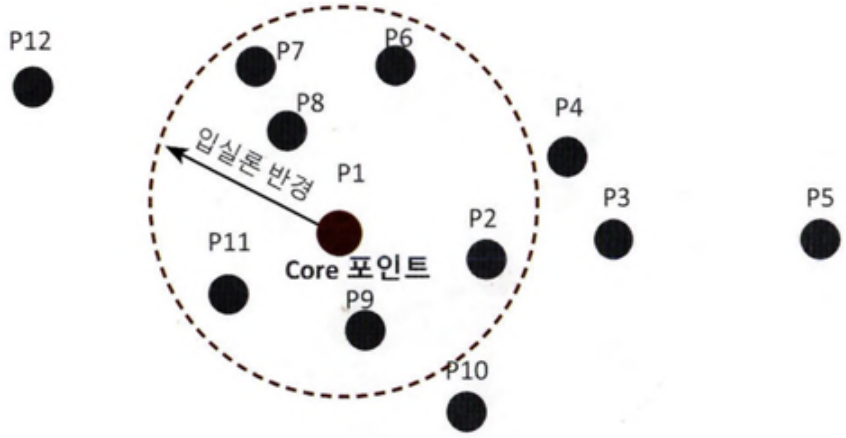


- 주요 파라미터
  - 입실론 주변 영역(epsilon): 개별 데이터를 중심으로 입실론 반경을 가지는 원형의 영역
  - 최소 데이터 개수(min points): 개별 데이터의 입실론 주변 영역에 포함되는 타 데이터의 개수
    - 최소 데이터 개수 충족 여부 확인
      - 핵심 포인트(Core Point): 주변 영역 내에서 최소 데이터 개수 이상의 타 데이터를 가지고 있을 경우 해당 데이터
      - 이웃 포인트(Neighbor Point): 주변 영역 내에 위치한 타 데이터를
      - 경계 포인트(Border Point): 주변 영역 내에 최소 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 않는 데이터
- DBSCAN 군집화 개념 설명

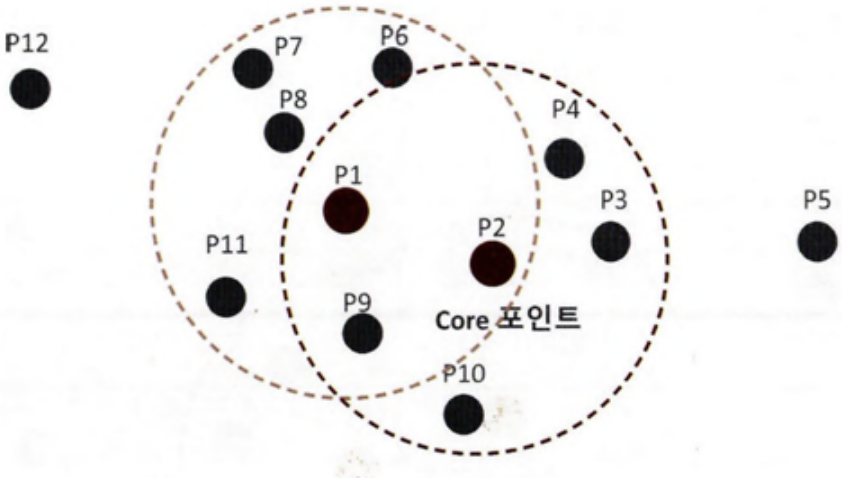
1. 특정 입실론 반경 내에 포함될 최소 데이터 세트를 6개(자기 자신 포함) 가정



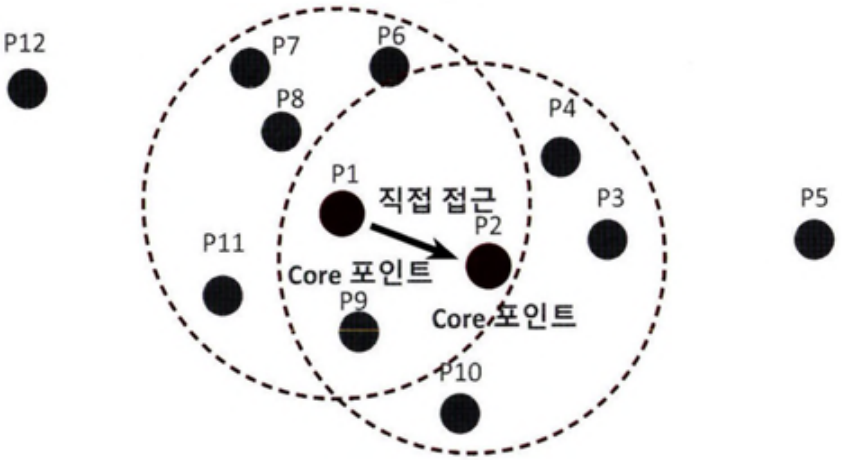
2. P1 데이터를 기준으로 입실론 반경 내에 포함된 데이터 7개(자신-P1, 이웃-P2, P6, P7, P8, P9, P11)로 최소 데이터 5개 이상을 만족하므로 P1 데이터는 핵심 포인트



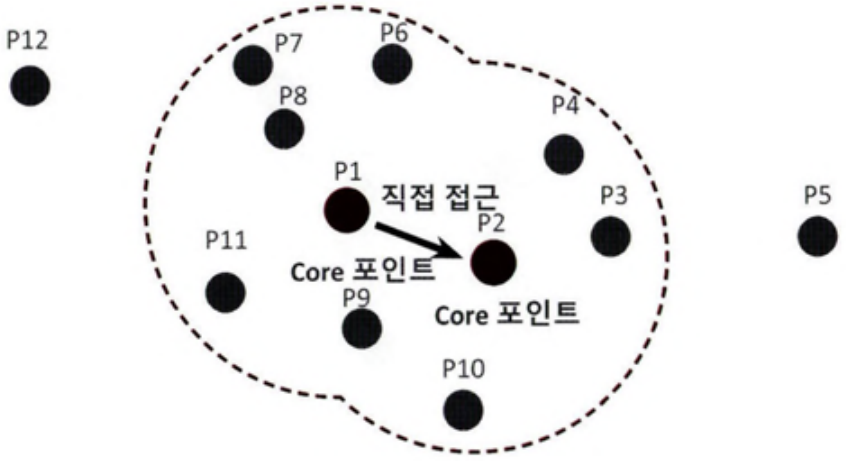
3. P2 역시 반경 내에 6개의 데이터(자신-P2, 이웃-P1, P3, P4, P9, P10)를 가지고 있으므로 핵심 포인트



4. 핵심 포인트 P1의 이웃데이터 P2 역시 핵심 포인트일 경우 P1에서 P2로 연결해 직접 접근이 가능함

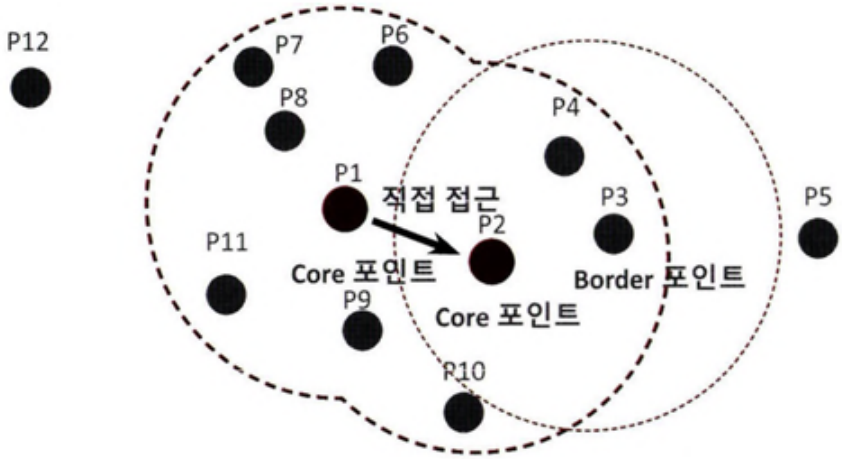


5. 서로 연결하면서 군집화를 구성→ 점차적으로 군집 영역을 확장해 나가는 것: DBSCAN 군집화 방식

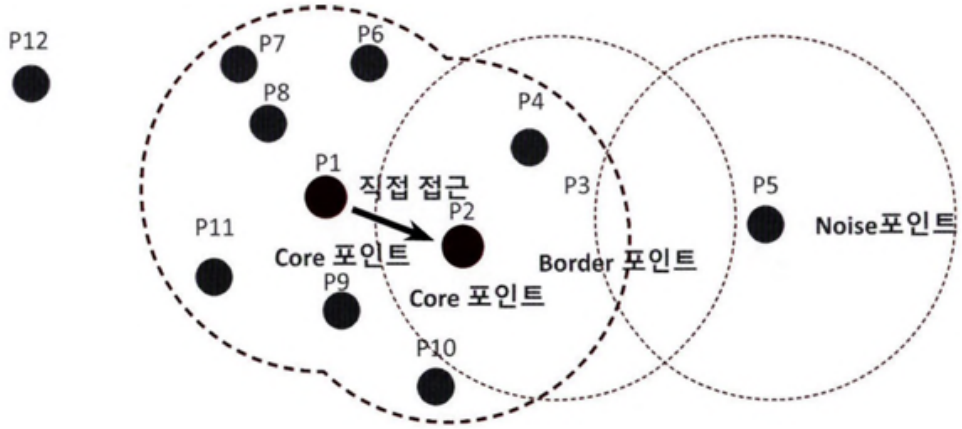


6. P3 데이터의 경우, 이웃 데이터-P2, P4 → 핵심 포인트X. but 핵심 포인트인 P2 소유 ⇒ 이웃데이터로 핵심 포인터를 가지고 있으므로 P3는 경계 포인트

   경계 포인트는 군집의 외곽을 형성



7. P5-반경 내에 최소 데이터X, 이웃데이터X ⇒ 잡음 포인트



06. 군집화 시습-고객 세그멘테이션

고객 세그멘테이션의 정의와 기법

고객 세그멘테이션

: 다양한 기준으로 고객을 분류하는 기법. CRM이나 마케팅의 중요 기반 요소

- 고객 분류 요소→ 어떤 상품을 얼마나 많은 비용을 써서 얼마나 자주 사용하는가
- 주요 목표: 타깃 마케팅(고객을 여러 특성에 맞게 세분화해서 그 유형에 따라 맞춤형 마케팅이나 서비스 제공)
- 어떤 요소 기반으로 군집화 할건지가 관건
- 기본적인 고객 분석 요소: RFM 기법
    - Recency(R): 가장 최근 상품 구입일에서 오늘까지의 기간
    - Frequency(F): 상품 구매 횟수
    - Monetary Value(M) : 총 구매 금액

지원 파일

https://colab.research.google.com/drive/1gjiP2cQeRbSzycfTnhge0oOhn-jRcSS2?usp=sharing