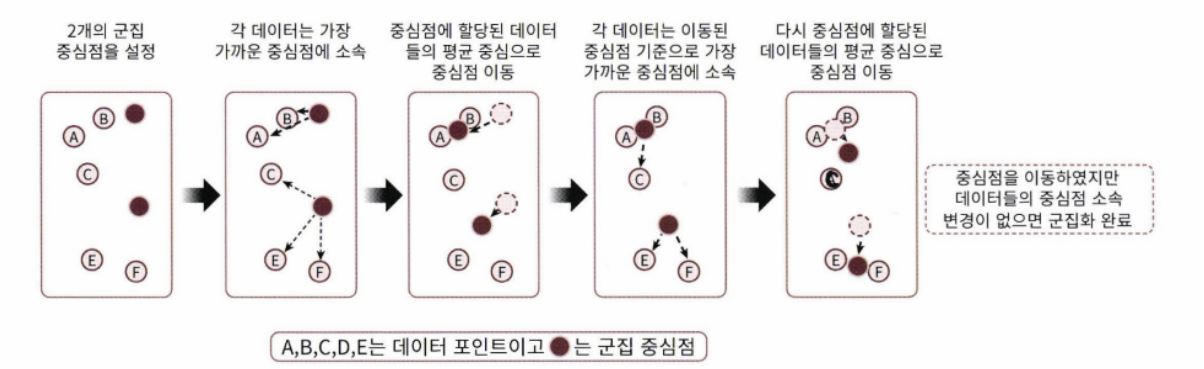
7장 군집화(Clustering)

01. K-평균 알고리즘 이해

- **K -평균 알고리즘** : 군집 중심점(centroid)라는 특정한 임의의 지점을 선택해 해당 중심에 가장 가까운 포인트들을 선택하는 군집화 기법
- **K-평균 알고리즘의 동작 원리**

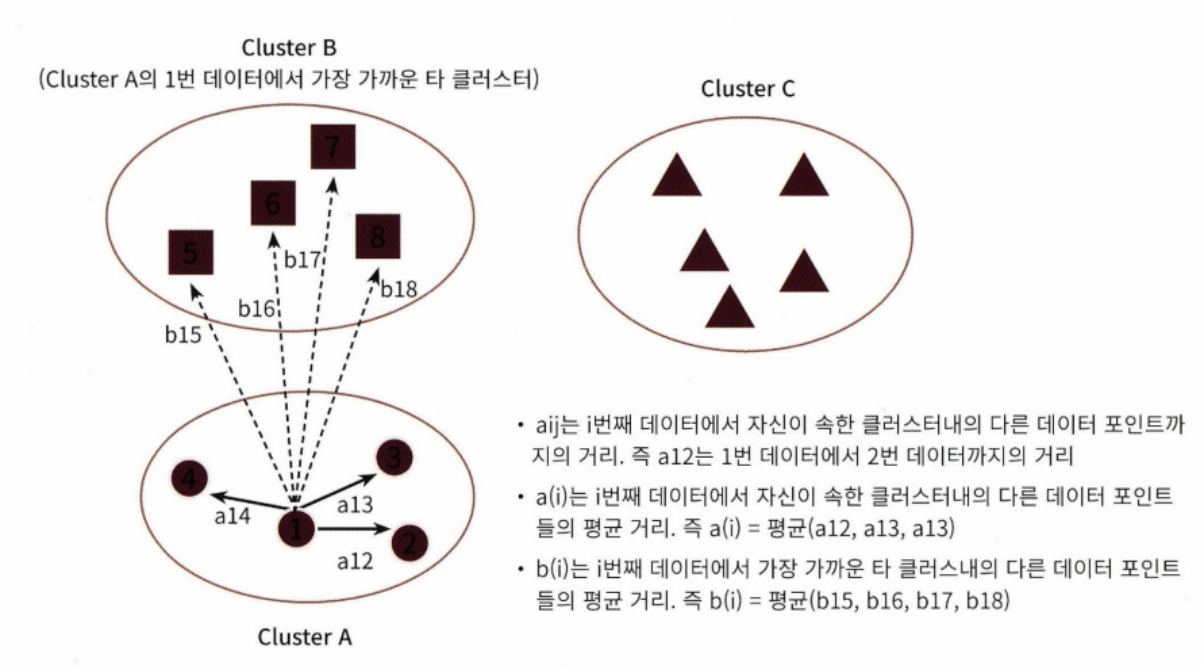


- **K-평균의 장점**
  1. 일반적인 군집화에서 가장 많이 활용되는 알고리즘입니다.
  2. 알고리즘이 쉽고 간결합니다.
  3. 데이터가 원형으로 분포되어 있을 경우 추정을 잘한다.(거리기반이기떄문)
- **K-평균의 단점**
  1. 거리 기반 알고리즘으로 속성의 개수가 매우 많을 경우 군집화의 정확도가 떨어집니다.(이를 위해 PCA로 차원 감소를 적용해야 할 수도 있습니다.)
  2. 반복을 수행하는데, 반복 횟수가 많을 경우 수행 시간이 매우 느려집니다.
  3. 몇 개의 군집(cluster)을 선택해야 할지 가이드하기 어렵다.
- **사이킷런 Kmenas 클래스 소개 (주요 파라미터 위주로)**
  - n_cluster : 군집 중심점의 개수
  - init : 초기 군집 중심점의 좌표를 설정할 방식(k-menas++)
  - max_iter : 최대 반복 횟수, 이 횟수 이전에 모든 데이터 중심점 이동이 없으면 종료
  - lables_ : 각 데이터 포인트가 속한 군집 중심점 레이블
  - cluster_centers_ : 각 군집 중심점 좌표(Shape[군집 개수, 피처 개수]), 이를 이용하면 군집 중심점 좌표가 어디인지 시각화가 가능하다.
- **K-평균을 이용한 붓꽃 데이터 세트 군집화**

- 코드를 통해 리뷰!

02. 군집 평가(Cluster Evaluation)

- 앞의 붓꽃 같은 경우 라벨링된 타겟 데이터가 있어 군집화의 결과를 비교하기 용이하였다. 하지만 그렇지 않은 경우는 어떻게 군집화의 효율성을 판단할까?
- 비지도학습의 특성상 어떠한 지표라도 정확하게 성능을 평가하기 어렵다 하지만 대표적인 방법으로 실루엣 분석을 이용한다.
- **실루엣 분석** : 각 군집 간의 거리가 얼마나 효율적으로 분리돼 있는지를 나타낸다.
- **실루엣 계수** : 해당 데이터가 **같은 군집** 내의 데이터와 얼마나 가깝게 군집화돼 있고, **다른 군집**에 있는 데이터와는 얼마나 멀리 분리돼 있는지를 나타내는 지표
- 실루엣 계수는 -1에서 1사이의 값을 가지며, 1로 가까워질수록 근처의 군집과 더 멀리 떨어져 있다는 것이고 0에 가까울수록 근처의 근집과 가까워진다는 것이다. -값은 아예 다른 군집에 데이터 포인트가 할당됐음을 의미한다.



$$\text{실루엣계수} : S(i) = \frac{(b(i) - a(i))}{(max(a(i), b(i))}$$

- **좋은 군집화의 기준 조건**
  1. 전체 실루엣 계수의 평균값이 1에 가까울수록 좋다.
  2. 전체 실루엣 계수의 평균값과 더불어 개별 군집의 평균값의 편차가 크지 않아야 한다. 즉 특정 군집의 실루엣 계수만 높은 그런 경우가 없어야 한다는 의미다.

군집별 평균 실루엣 계수의 시각화를 통한 군집 개수 최적화 방법

- 코드 참고
- 사이트 참고 :

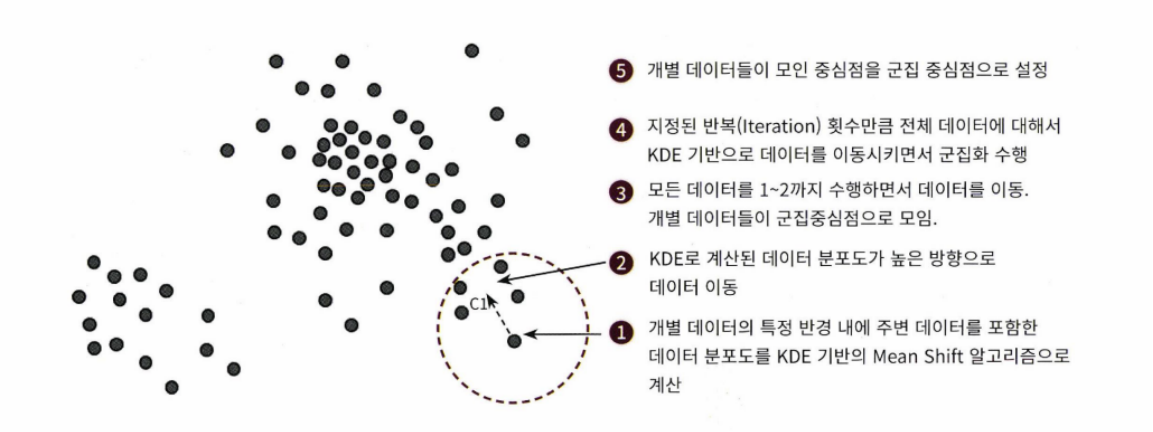
Selecting the number of clusters with silhouette analysis on KMeans clustering

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...

https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html#sphx-glr-auto-examples-cluster-plot-kmeans-silhouette-analysis-py

03. 평균 이동(Mena Shift)

- k-평균과 유사하게 중심을 군집의 중심으로 지속적으로 이동하면서 군집화를 수행한다. 하지만 차이점은 k-평균이 중심에 소속된 데이터의 평균 거리 중심으로 이동하는 데 반해, 평균 이동은 중심을 데이터가 모여 있는 밀도가 가장 높은 곳으로 이동시킨다.
- **아이디어** : 확률 밀도 함수를 KDE(Kernel Density Estimation)을 이용해 찾고 가장 피크인 점을 군집 중심으로 선정하여 이동하며 업데이트 한다.



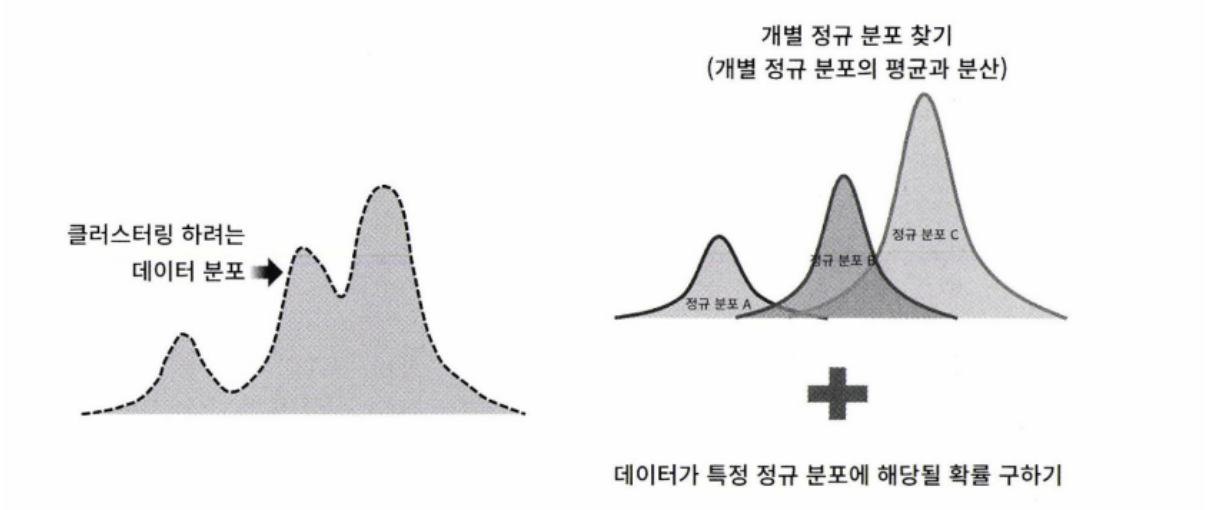
- **KDE(Kernel Density Estimation)** : 개별 관측 데이터에 커널 함수를 적용한 뒤, 이 적용 값을 모두 더한 후 개별 관측 데이터의 건수로 나눠 확률 밀도 함수(PDF)를 추정하며, 대표적인 커널 함수로는 가우시안 분포 함수가 사용된다.
- 위와 같은 방식으로 확률밀도함수를 추정하고 이를 바탕으로 군집 중심점을 이동시키며 학습이 진행되는데 특이하게 **군집의 개수를 지정하지 않으며, 오직 대역폭의 크기에 따라 군집화를 수행**한다.
- **평균 이동의 장점**
  1. 데이터 세트의 형태를 특정 형태로 가정하지 않고, 특정 분포도 기반의 모델로 가정하지 않으므로 좀 더 유연한 군집화가 가능하다.
  2. 이상치의 영향력도 크지 않으며, 미리 군집의 개수를 정할 필요가 없다.
- **평균 이동의 단점**

1. 알고리즘의 수행 시간이 오래 걸린다.
2. band-width의 크기에 따른 군집화 영향도가 매우 크다.

위와 같은 단점들로 인해 분석 업무 보다는 컴퓨터 비전 의 tracking 혹은 특정 개체 구분에 뛰어난 역활을 수행하는 알고리즘이다.

04. GMM(Gaussian Mixture Model)

- **GMM 군집화** : 군집화를 적용하고자 하는 데이터가 **여러 개의 가우시안 분포 (GaussaianDistribution)를 가진 데이터 집합들이 섞여서 생성**된 것이라는 가정하에 군집화를 수행하는 방식이다.
- 여기서 가우시안 분포가 생소하시다면 정규분포라 생각하시면 이해가 빠를것입니다.



왼쪽 그림처럼 데이터 분포가 존재한다면, 이를 오른쪽 그럼처럼 개별 정규분포를 추정(찾기)하는 것입니다.(즉 데이터가 특정 정규 분포에 해당될 확률을 구하는 것)

- 위와 같은 방식을 GMM의 모수 추정이라 부르고 2가지 방식이 존재합니다.
  1. 개별 정규 분포의 평균과 분산
  2. 각 데이터가 어떤 정규 분포에 해당되는지의 확률
- EM알고리즘을 통해 이러한 모수추정을 진행
- **GMM을 이용한 붓꽃 데이터 세트 군집화**

  코드 참고
- **GMM과 K-평균의 비교**

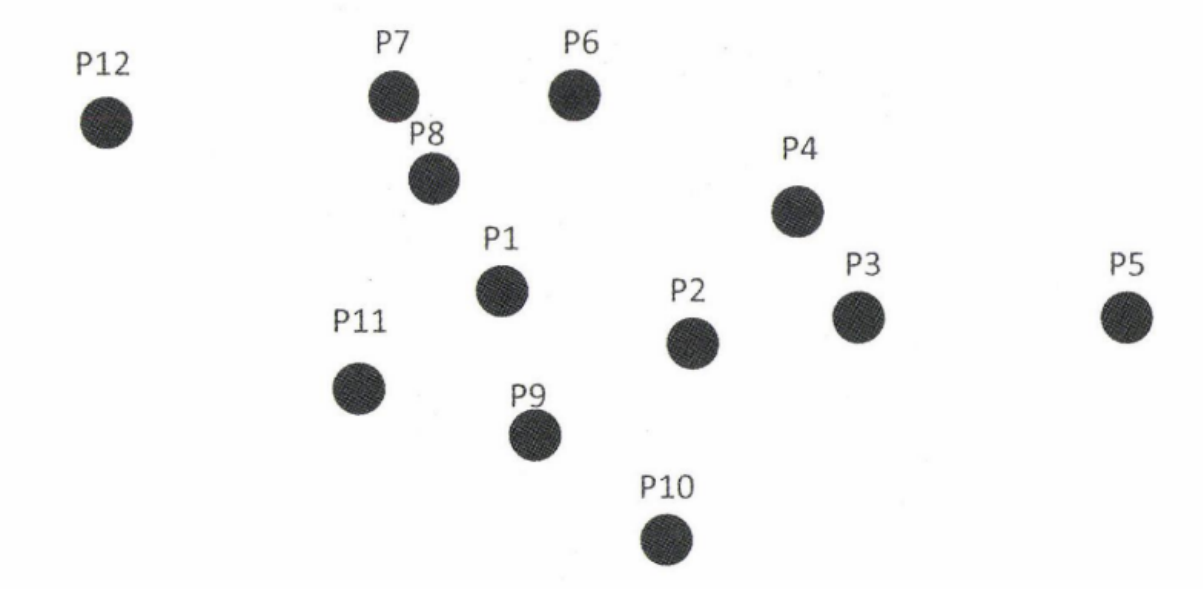
  코드 참고

05. DBSCAN

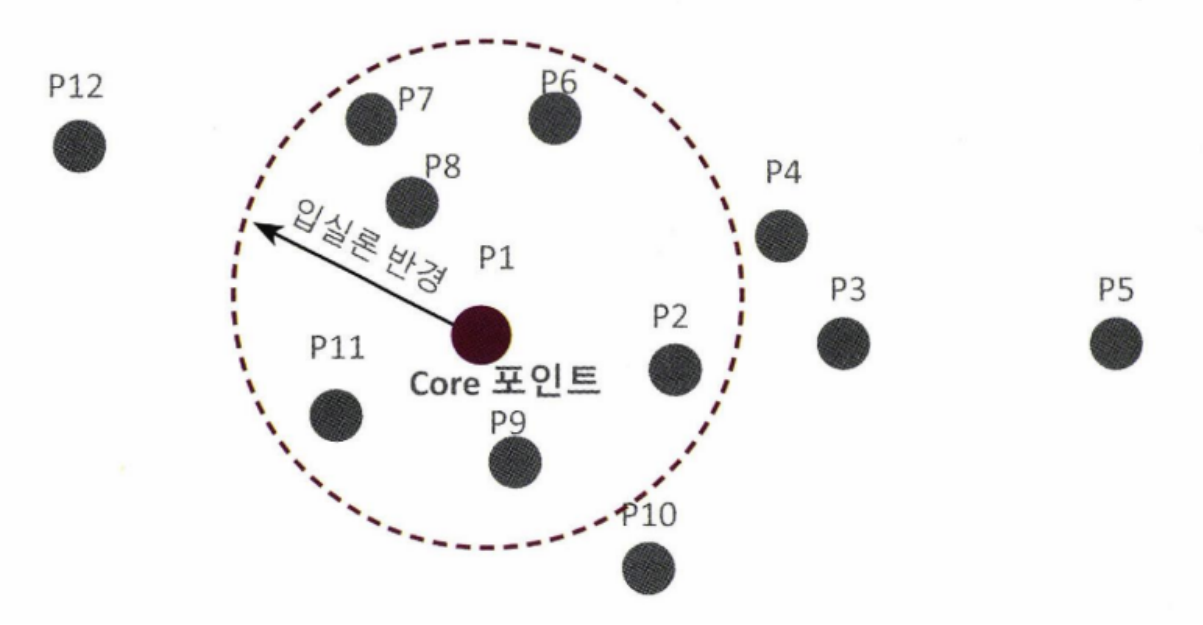
- 밀도 기반 군집화의 대표적인 알고리즘이고 특정 공간 내에 데이터 밀도 차이를 기반으로 **기하학적 분포도**를 가진 데이터 세트에 대해서도 군집화를 잘 수행한다.
- DBSCAN을 구성하는 2가지 핵심 파라미터
  - **입실론 주변 영역(epsilon)** : 개별 데이터를 중심으로 입실론 반경을 가지는 원형의 영역입니다.
  - **최소 데이터 개수(min points)** : 개별 데이터의 입실론 주변 영역에 포함되는 타 데이터의 개수입니다.
  - 즉 입실론 주변 영역 내에 포함되는 최소 데이터 개수를 충족하냐 안하냐에 따라 데이터 포인트를 다르게 정의한다.
  - **핵심 포인트(Core Point)** : 주변 영역 내에 최소 데이터 개수 이상의 타 데이터를 가지고 있을 경우
  - **이웃 포인트(Neighbor Point)** : 주변 영역 내에 위치한 타 데이터
  - **경계 포인트(Border Point)** : 주변 영역 내에 최소 데이터 개수 이상의 이웃 포인트를 가지고 있지 않지만 핵심 포인트로 가지 있는 데이터
  - **잡음 포인트(Noise Point)** : 최소 데이터 개수 이상의 이웃 포인트를 가지고 있지 않으며, 핵심 포인트도 이웃 포인트로 가지고 있지 않는 데이터

  이렇게만 말하면 이해가 쉽지 않으므로 아래의 예시를 기준으로 해당 포인트의 의미를 설명해보겠습니다.

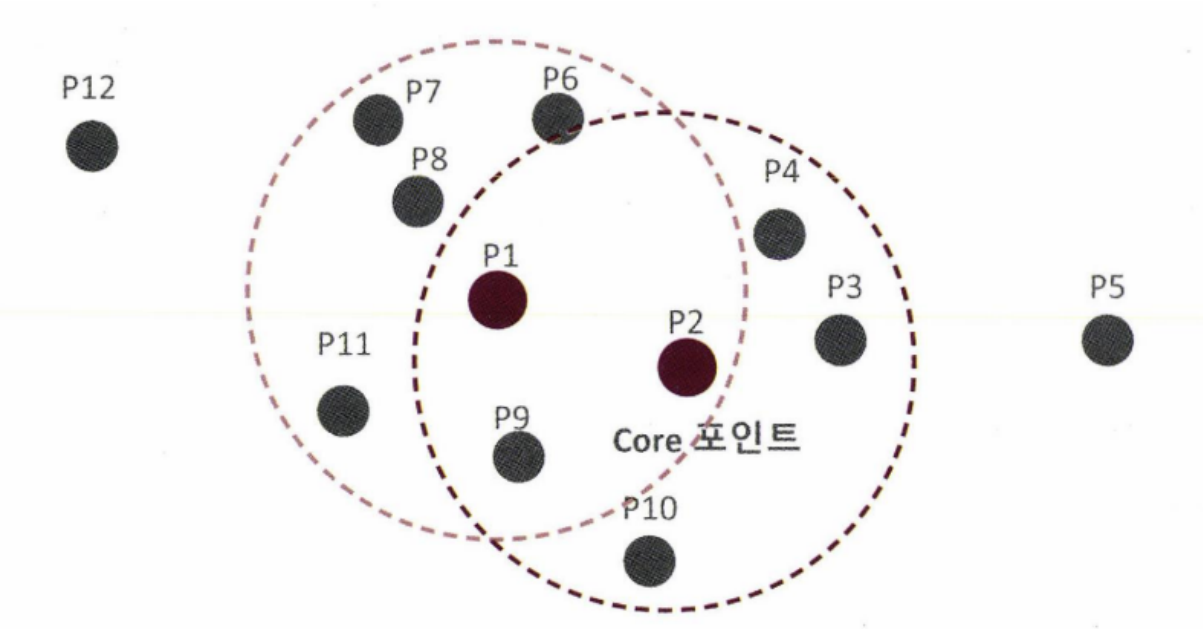
  1. P1~P12까지 12개의 데이터 세트에 대해 특정 입실론 반경 내에 포함될 최소 데이터 세트를 5개로(자기 자신의 데이터 포함) 가정하겠습니다.



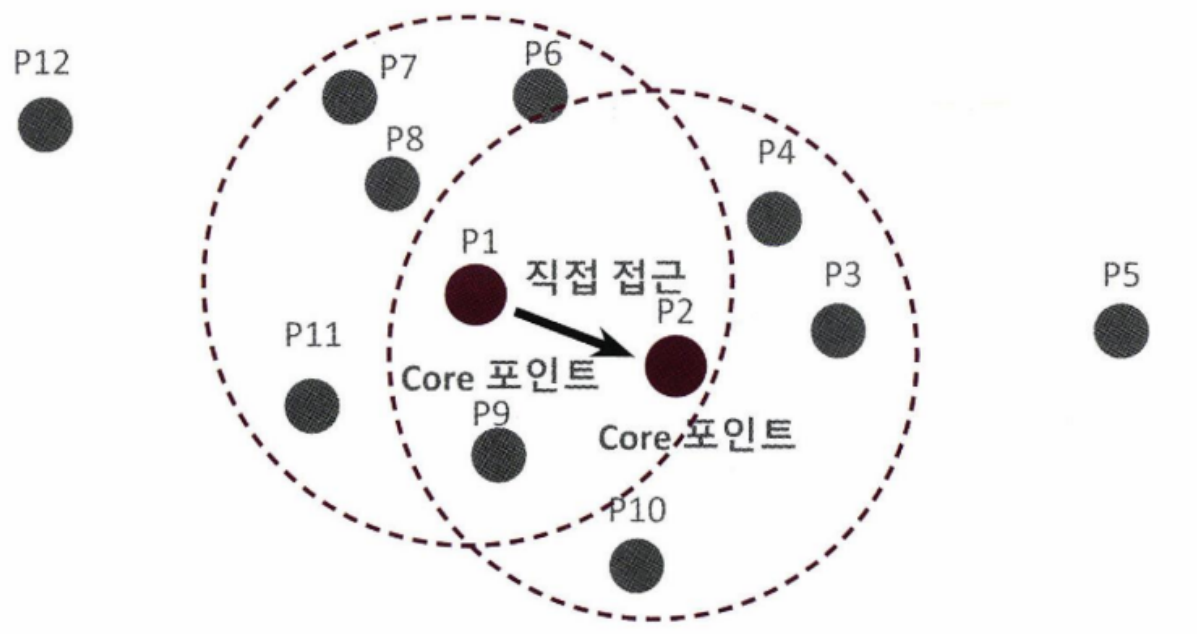
  2. P1 데이터를 기준으로 입실론 반경 내에 포함된 데이터의 개수가 7개 이므로 최소 데이터 5개를 만족하므로 P1 데이터는 핵심 포인트(Core Point)이다.



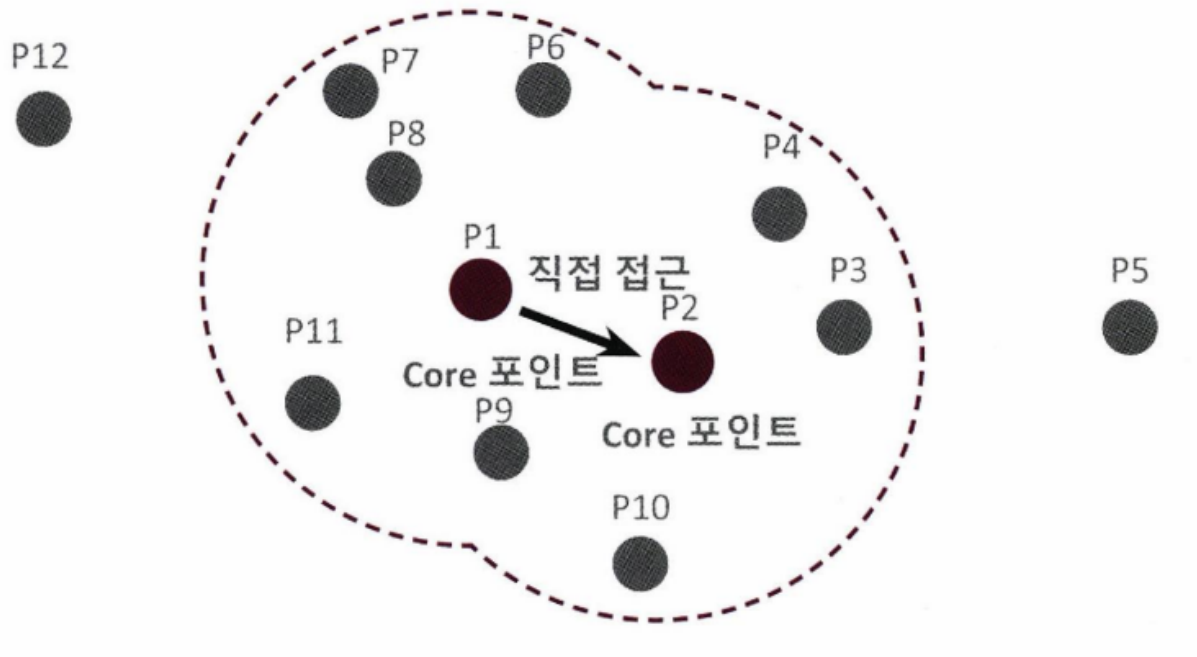
3. 다음으로 P2 데이터 포인트를 살펴보면 이 역시 6개의 데이터(P2, P1, P3, P4, P9, P10)을 가지고 있으므로 마찬가지로 핵심 포인트(Core Point)이다.



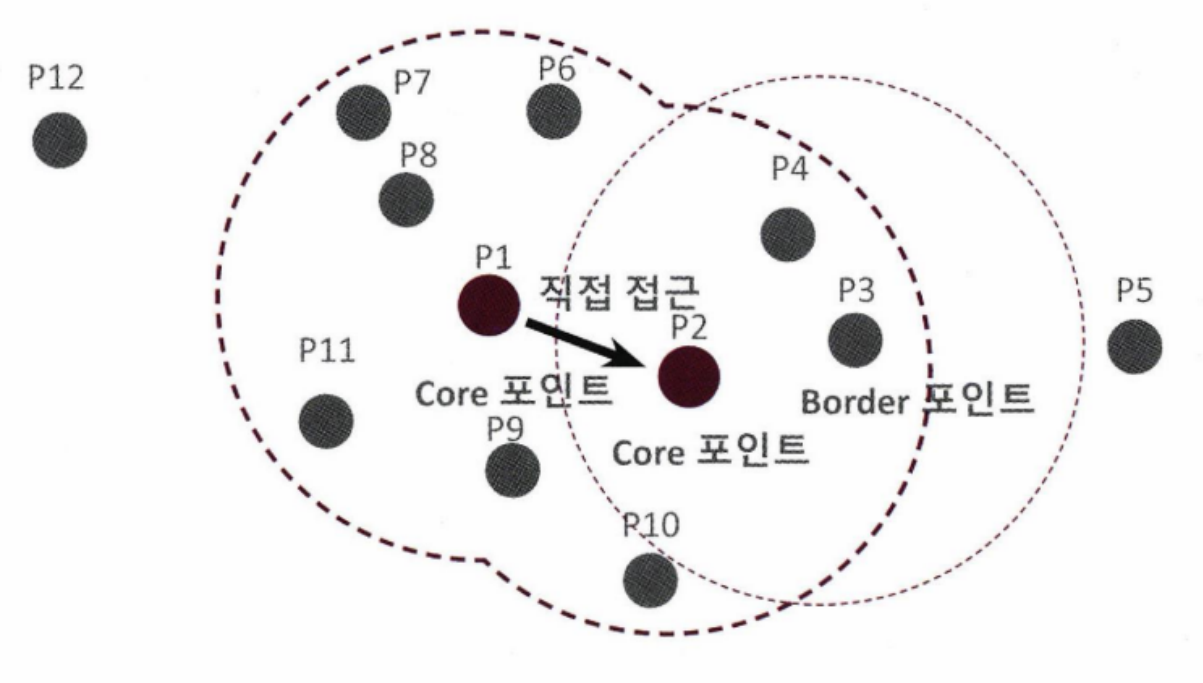
4. 핵심 포인트 P1의 이웃 데이터 포인트 P2 역시 핵심 포인트일경우 P1에서P2로 연결해 직접 접근이 가능합니다.



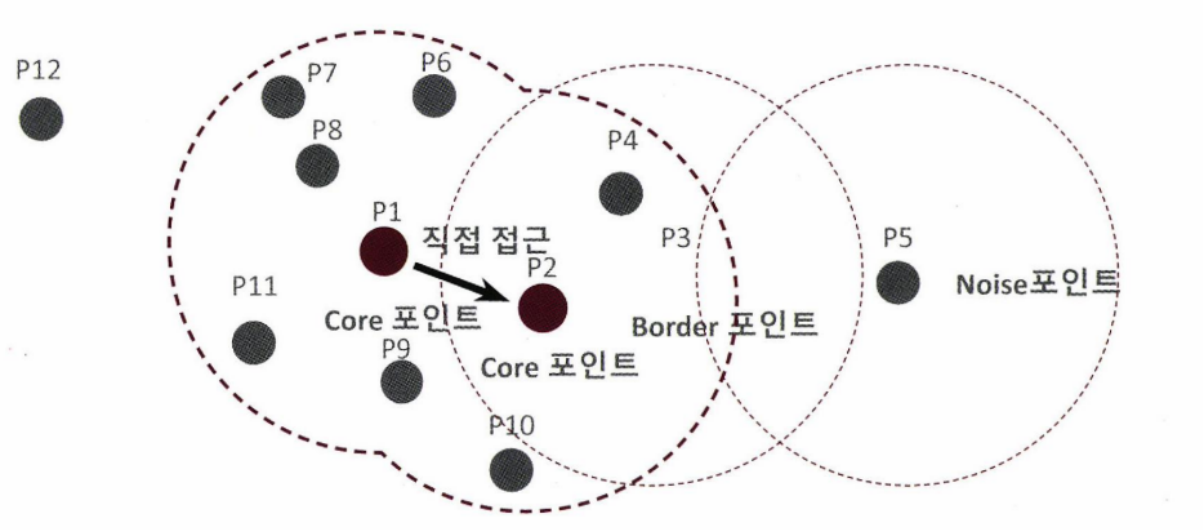
5. 특정 핵심 포인트에서 직접 접근이 가능한 다른 핵심 포인트를 서로 연결하면서 군집화를 구성합니다. 이러한 방식으로 점차 군집(Cluster) 영역을 확장해 나가는 방식으로 DBSCAN은 군집화를 해나간다.



6. P3 데이터의 경우 반경 내에 포함되는 이웃 데이터가 P2,P4 2개이므로 핵심 포인트가 될 수 없지만 이웃 데이터 중에 핵심 포인트인 P2를 가지고 있다. 이처럼 자신은 핵심 포인트가 아니지만, 이웃 데이터로 핵심 포인트를 가지고 있는 데이터를 경계 포인트(Border Point)라고 합니다. 이 경계 포인트가 군집의 외곽을 형성합니다.



7. 그림의 P5처럼 반경 내에 최소 데이터를 가지고 있지도 않고, 핵심 포인트 또한 이웃 데이터로 가지고 있지 않은 데이터를 잡음 포인트(Noise Point)라고 한다.



- DBSCAN을 적용할 때는 특정 군집 개수로 군집을 강제하지 않는 것이 좋다
- 적절한 eps와 min_samples 파라미ㅓ를 통해 최적의 군집을 찾는게 중요하다 일반적으로 eps의 값을 크게 하면 반경이 커져 노이즈 데이터 개수가 적어진다. min_samples를 크게 하면 반경 내에 더 많은 데이터를 포함해야 하므로 노이즈 데이터 개수가 커지게 된다.

06. 군집화 실습 - 고객 세그멘테이션

07. 정리