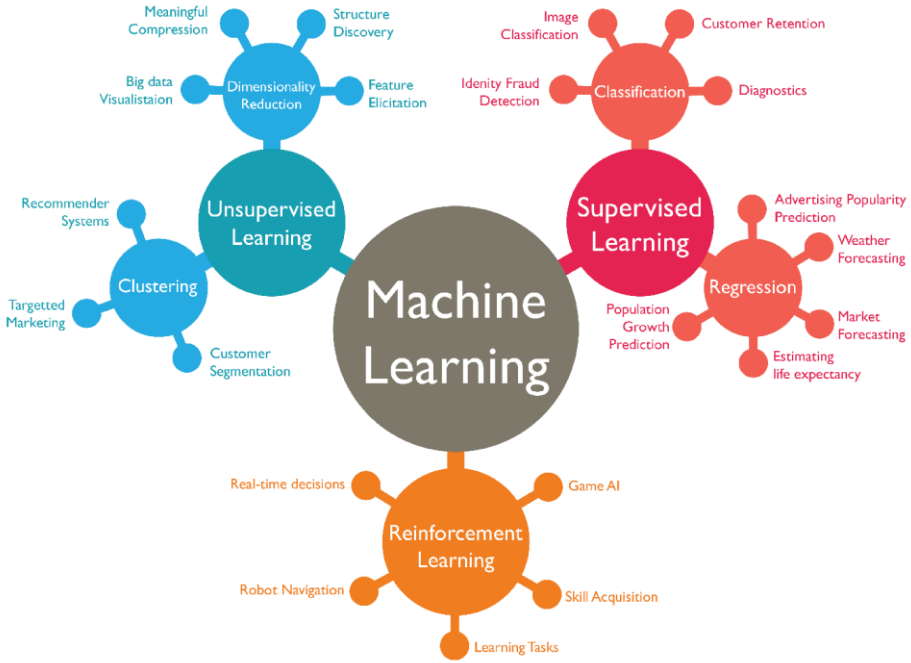
# 군집 분석
# (Clustering Analysis)

학부연구생 이현재

# 비지도 학습 (Unsupervised Learning)

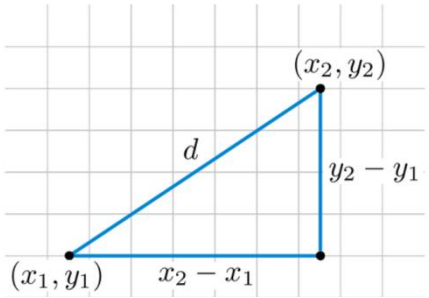
- 정의 : 지도학습과 다르게 정답 레이블이 없는 데이터를 비슷한 특징끼리 군집화하여 새로운 데이터에 대한 결과를 예측
- 비지도 학습 알고리즘
  - 군집화 : K 평균, 계층군집 분석(HCA)
  - 시각화와 차원 축소 : 주성분 분석(PCA),
  - 연관 규칙 학습
- 군집 분석 : 개체들을 유사성에 기초하여 n개의 군집으로 집단화하여 집단의 특성을 분석하는 다변량 분석



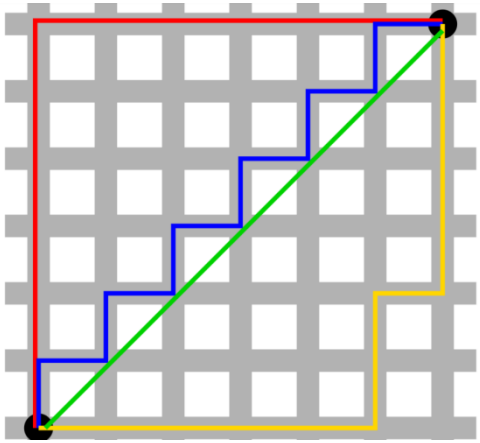
# 데이터 간 거리 측정 척도

- 유클리드 거리(Euclidean Distance)
  - L2 Distance
  - 두 점 사이의 최단 거리
- 맨해튼 거리(Manhattan Distance)
  - L1 Distance
  - 두 점 p, q가 있을 때, 가상 체스보드가 있다고 생각하고 오직 수평, 수직 이동만 하여 p에서 q까지 이동할 때 최단으로 걸리는 거리
- 코사인 유사도(Cosine Similarity) $D(x,y) = cos(\theta) = \frac{x \cdot y}{\|x\| \, \|y\|}$
  - 내적 공간의 두 벡터간 각도와 코사인 값을 이용하여 측정된 벡터간 유사한 정도. 코사인 유사도는 '방향'에 대한 유사도. 즉, '거리'는 고려하지 않는다
  - 거리 기반 (ex.유클리디안 거리)은 좌표를 기준으로, 가까운 좌표에 있는 점들이 유사도가 높다고 측정되는 반면, 각도 기반 (ex.코사인 유사도)은 기울기와 방향이 같은 벡터가 유사도가 높다고
  - 코사인 유사도는 2차원보다 높은 차원의 데이터, 또 벡터의 크기가 중요하지 않은 데이터에 주로 사용

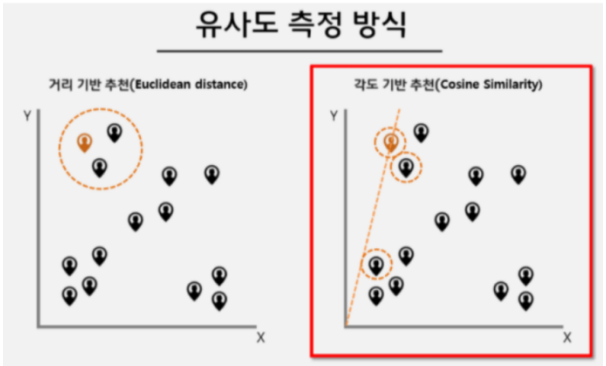
$d(A,B) = \sqrt{(x2-x1)^2 + (y2-y1)^2}$




<유클리드 거리>



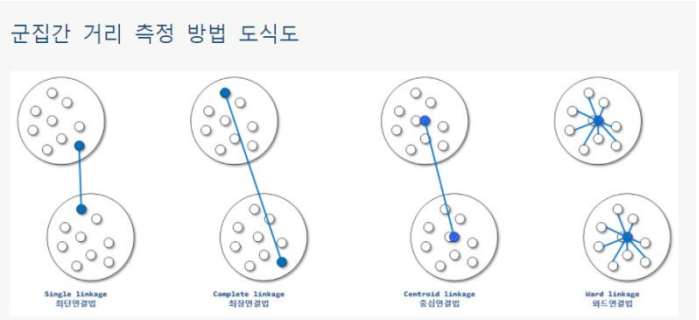
초록색은 유클리드
그 이외는 맨해튼




Prof. Hyerim Bae (hrbae@pusan.ac.kr)

# 군집 간 거리 측정 척도

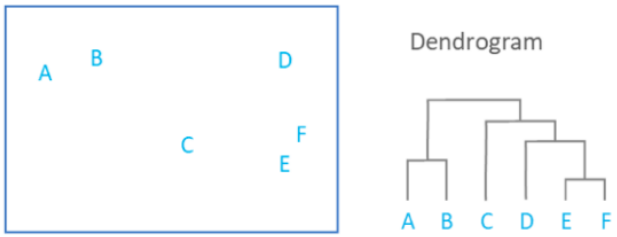
- 최단연결법(단일연결법)
  - 각 군집에서 관측값을 뽑았을 때 나타날 수 있는 거리의 최소값을 군집간 거리로 측정
  - 사슬 모양
  - 고립된 군집을 찾는데 중점을 둔 방법
- 최장연결법(완전연결법)
  - 각 군집에서 관측값을 뽑았을 때 나타날 수 있는 거리의 최대값을 군집간 거리로 측정
  - 군집들의 내부 응집성에 중점을 둔 방법
- 중심연결법
  - 두 군집의 중심간 거리를 군집간 거리로 측정
  - 군집이 결합될 때, 새로운 군집의 평균은 가중평균을 통해 측정
- 평균연결법
  - 모든 항목에 대한 거리 평균을 구하면서 군집화를 수행
  - 계산량이 불필요하게 많아질 수 있다
- 와드연결법
  - 군집 내의 오차제곱합에 기초하여 군집을 수행
  - 군집이 병합되면 오차제곱합은 증가하는데, 증가량이 가장 적어지도록 군집을 형성
  - 크기가 비슷한 군집끼리 병합하게 되는 경향



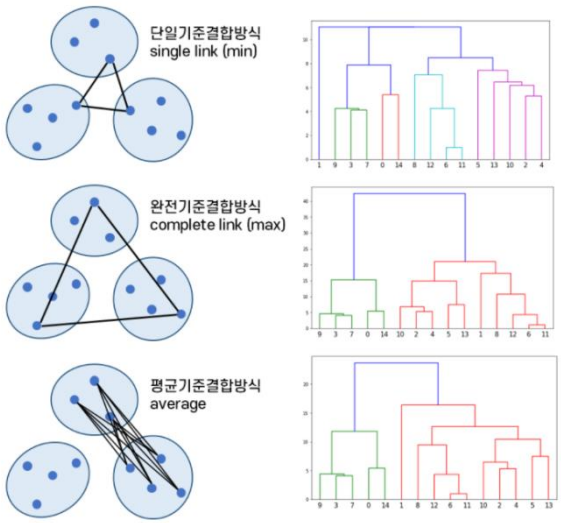

Prof. Hyerim Bae (hrbae@pusan.ac.kr)

# 계층적 군집 분석 (Hierarchial Clustering)

- 유클리드 거리를 이용한 군집 분석 방법
- 계층적으로 군집 결과 도출
- 탐색적 군집 분석
- 계층적 군집 분석 결과 => 덴드로그램 : 표본들이 군을 형성하는 과정을 나타내는 나무 형식 그림



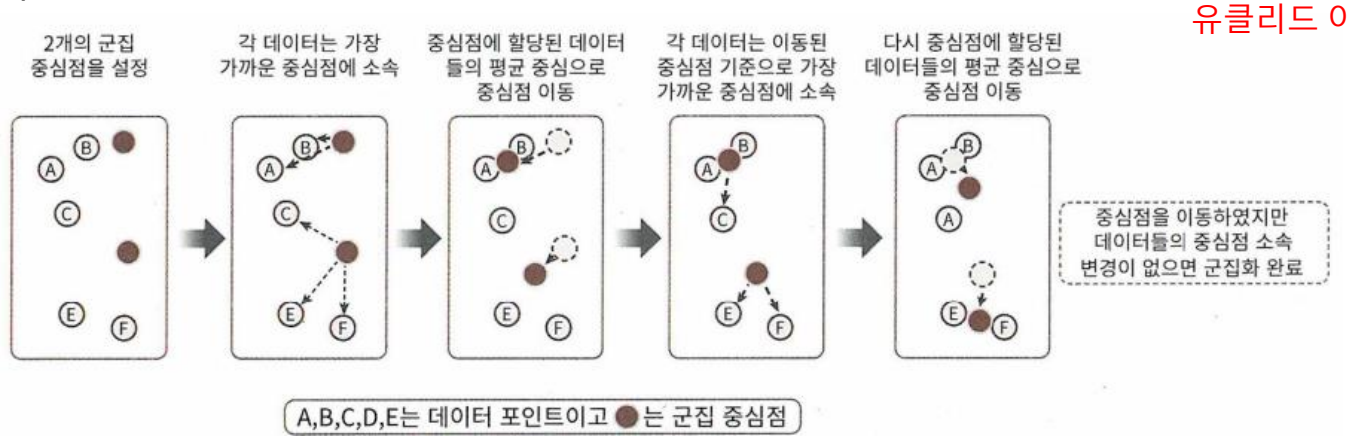

- 군집화 방식



# K-means 알고리즘 (비계층적 군집 분석)

- 군집 중심점(Centroid)이라는 특정한 임의의 지점을 선택해 해당 중심에 가장 가까운 포인트들을 선택하는 군집화 기법 => N개의 데이터를 K개의 클러스트 중 하나에 할당
- 목표 : 클러스터 내의 차이를 최소화하고 클러스터 간의 차이를 최대화
- 진행 과정

유클리드 이용

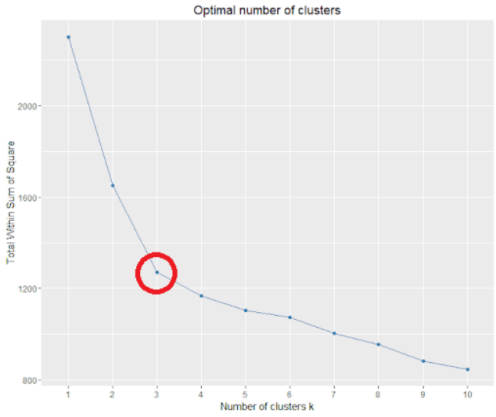



1. 먼저 군집화의 기준이 되는 중심을 구성하려는 군집화 개수만큼 임의의 위치에 중심 지정. 만약 전체 데이터를 2개로 군집화하려면 2개의 중심을 임의의 위치에 지정
2. 각 데이터는 가장 가까운 곳에 위치한 중심점에 소속. 위 그림에서는 그림 2에서 A,B데이터가, C,E,F 데이터가 같은 소속
3. 이렇게 소속이 결정되면 군집 중심점을 소속된 데이터의 평균 중심으로 이동. 그림 3에서는 A,B의 평균 위치로 중심점이 이동했고 다른 중심점도 마찬가지
4. 중심점이 이동했기 때문에 각 데이터는 기존에 속한 중심점보다 더 가까운 중심점이 있다면 해당 중심점으로 다시 소속을 변경. 그림 4에서는 c가 이동
5. 다시 중심을 소속된 데이터의 평균 중심으로 이동.
6. 중심점을 이동했는데 데이터의 중심점 소속 변경이 없으면 군집화 종료

Prof. Hyerim Bae (hrbae@pusan.ac.kr)

# K-means 알고리즘 (비계층적 군집 분석)

- 장점
  - 일반적인 군집화에서 가장 많이 활용되는 알고리즘
  - 알고리즘이 쉽고 간결
- 단점
  - 거리 기반 알고리즘으로 속성의 개수가 매우 많을 경우 군집화 정확도가 떨어짐(이를 위해 PCA로 차원 감소를 적용해야할 수도 있음)
  - 반복을 수행하는데, 반복 횟수가 많을 경우 수행 시간이 매우 느려짐
  - 몇 개의 군집을 선택할지(k를 뭐로할지) 정하기 어려움
- 엘보우 : 최적의 k 찾기
  - Intertia
    - 클러스터 중심과 클러스터에 속한 샘플 사이의 거리 제곱 합
    - 클러스터에 속한 샘플이 얼마나 가깝게 모여 있는지를 나타내는 값
    - 일반적으로 클러스터의 개수가 늘어나면 클러스터 개개의 크기는 줄어들기 때문에 Intertia도 함께 줄어듬
  - 클러스터 개수(k)를 늘려가면서 intertia의 변화를 관찰하여 최적의 클러스터 개수를 찾는 방법

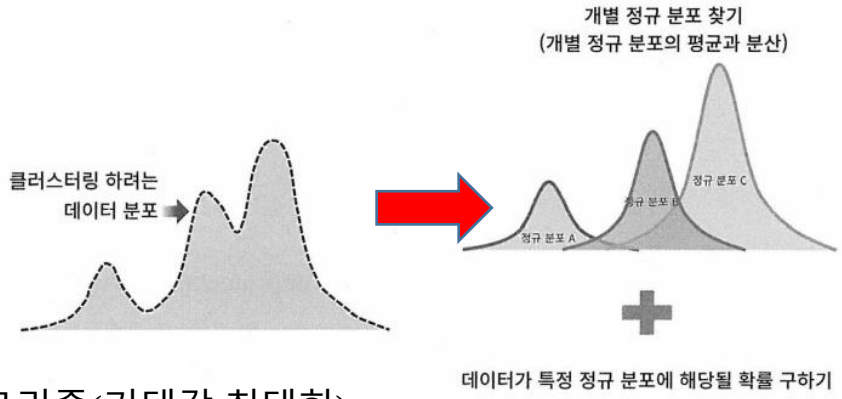



클러스터 개수 k에 따른 이너셔(inertia)의 감소 그래프

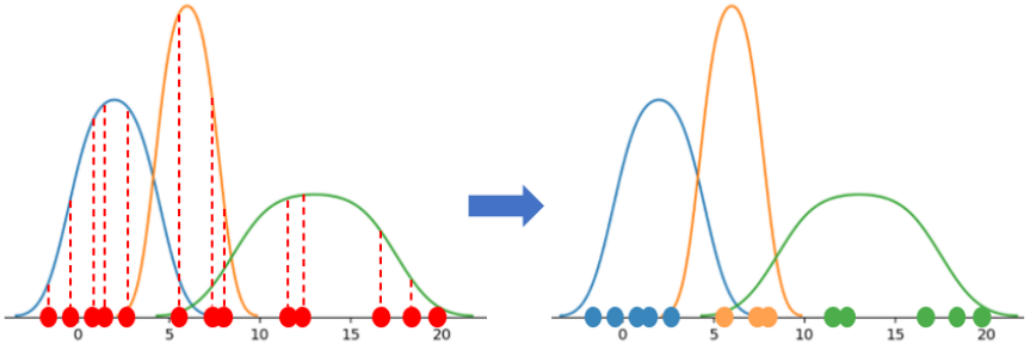
속도가 꺾이는 지점이 최적 개수

# GMM (Gaussian Mixture Model) (비계층적 군집 분석)

- 군집화를 적용하고자 하는 데이터가 여러 개의 가우시안 분포를 가진 데이터 집합들이 섞여서 생성된 것이라는 가정하에 군집화를 수행하는 방식
- **GMM은 데이터를 여러 개의 가우시안 분포가 섞인 것으로 간주**




- 진행 방법 – EM 알고리즘(기대값 최대화)
  - 1. 예측
    - 개별 데이터 포인트가 각 정규 분포로부터 생성되었을 가능도를 계산하여, 가장 높은 확률을 가진 정규 분포에 해당

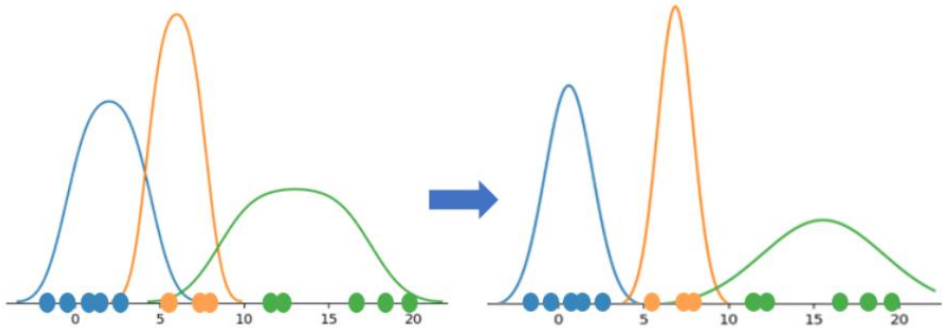



위 그림처럼 빨간 색으로 표시된 각 데이터 포인트에 대해 각 정규 분포에 속할 확률을 계산한 후, 가장 큰 확률을 갖는 정규 분포로 할당한다. 파란색 정규 분포에 5개의 데이터 포인트가, 주황색 정규 분포에 3개의 데이터 포인트가, 초록색 정규 분포에 5개의 데이터 포인트가 할당되는 것을 확인할 수 있다.

Prof. Hyerim Bae (hrbae@pusan.ac.kr)

# GMM (Gaussian Mixture Model) (비계층적 군집 분석)

- 진행 방법(계속)
  - 2. 최대화
    - 위 1번 단계에서 개별 데이터 포인트를 모두 할당한 후, 각 그룹의 데이터 포인트를 이용하여 Maximum Likelihood Estimation(최대 우도 추정)으로 모분포의 모평균과 모분산을 추정




    - 개별 데이터 포인트로 각 정규 분포의 모수를 추정한다. 위 그림에서 정규 분포의 평균과 분산이 변경된 것을 확인할 수 있다. 개별 데이터들의 소속과 정규 분포의 모수(평균과 분산)가 변하지 않을 때까지 1,2 단계를 반복 수행
- GMM과 K-means 비교
  - K-means는 거리 기반이기 때문에 같은 거리상 원형으로 구성된 데이터 구조에 군집화하기 알맞고 GMM에서는 다양한 기하학적 모양에 대한 군집화가 가능하다
  - 두 모델 모두 적절한 군집 개수 혹은 절절한 가우시안 분포 개수를 지정해줘야 한다
  - 가우시안 분포마다 충분한 데이터 포인터들이 있어야 모수 추정(우리가 확인한 진행방법)이 잘 이뤄진다.

Prof. Hyerim Bae (hrbae@pusan.ac.kr)

# DBSCAN (비계층적 군집 분석)

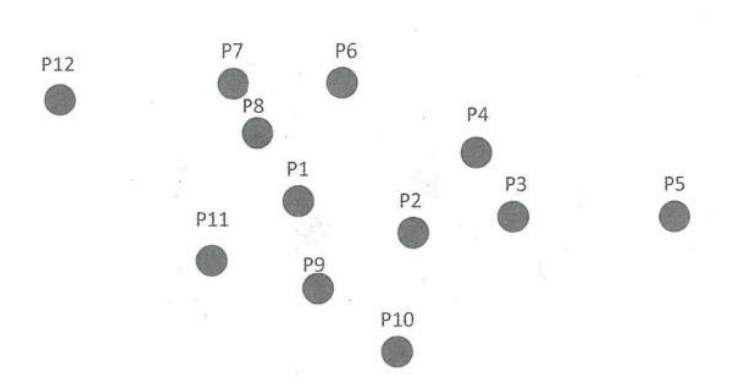
- 밀도 기반 군집화
- 특정 공간 내에 데이터 밀도 차이를 기반 알고리즘으로 하고 있어서 복잡한 기하학적 분포도를 가진 데이터 세트에 대해서도 군집화를 잘 수행
- 입실론 주변 영역(epsilon) : 개별 데이터를 중심으로 입실론 반경을 가지는 원형의 영역
- 최소 데이터 개수(min points) : 개별 데이터의 입실론 주변 영역에 포함되는 타 데이터 개수
- 핵심 포인트(Core Point) : 주변 영역 내에 최소 데이터 개수 이상의 타 데이터를 가지고 있을 경우
- 이웃 포인트(Neighbor Point) : 주변 영역 내에 위치한 타 데이터
- 경계 포인트(Border Point) : 주변 영역 내에 최소 데이터 개수 이상의 이웃 포인트를 가지고 있지 않지만 핵심 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
- 장점
  - 다른 기존 알고리즘처럼 클러스터의 개수를 지정해주지 않아도 됨
  - 클러스터의 밀도에 따라서 클러스터를 서로 연결하기 때문에 기하학적인 모양을 갖는 군집도 잘 찾을 수 있다
  - Noise point를 통하여, outlier 검출이 가능

Prof. Hyerim Bae (hrbae@pusan.ac.kr)

# DBSCAN (비계층적 군집 분석)

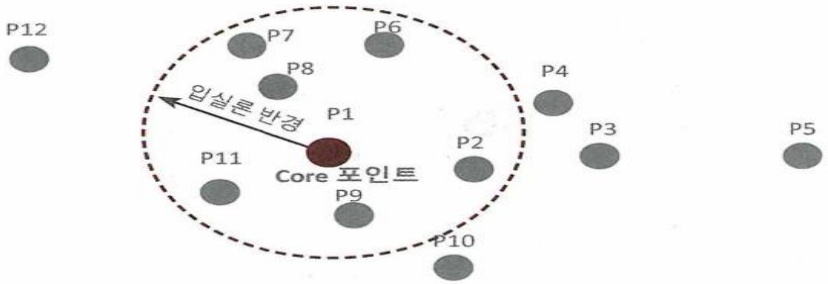
- 진행 방법

1. 다음 그림과 같이 P1에서 P12까지 12개의 데이터 세트에 대해서 DBSCAN 군집화를 적용하면서 주요 개념을 설명하겠습니다 특정 입실론 반경 내에 포함될 최소 데이터 세트를 6개로(자기 자신의 데이터를 포함) 가정하겠습니다.




이 경우에는 '최소 데이터 개수'가 6

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



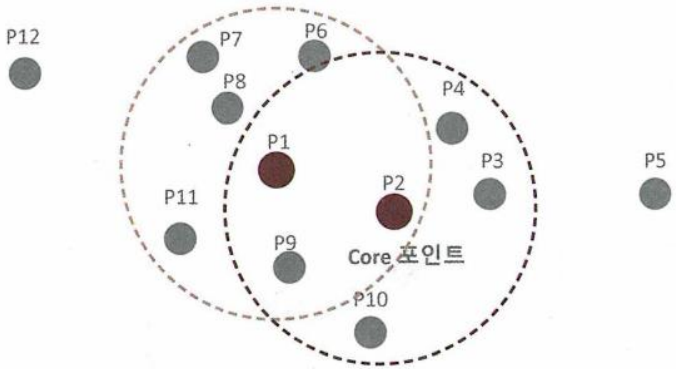

입실론 == 범위

그리고 최소 데이터를 만족했으면 그 점은 Core Point가 된다.

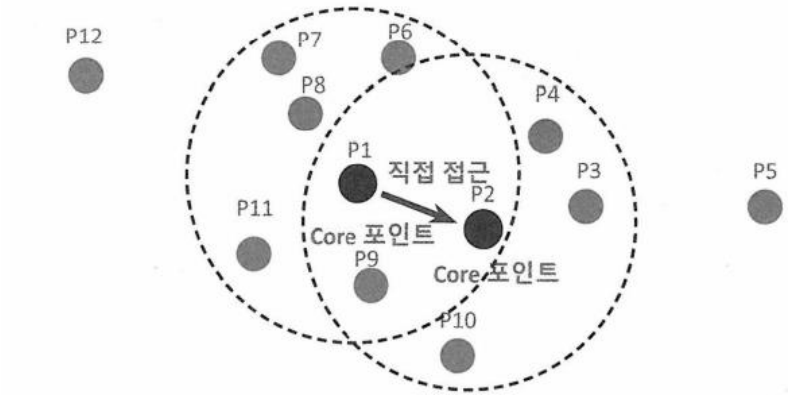
# DBSCAN (비계층적 군집 분석)

- 진행 방법

3. 다음으로 P2 데이터 포인트를 살펴보겠습니다. P2 역시 반경 내에 6개의 데이터(자신은 P2, 이웃 데이터 P1, P3, P4, P9, P10)를 가지고 있으므로 핵심 포인트입니다.




4. 핵심 포인트 P1의 이웃 데이터 포인트 P2 역시 핵심 포인트일 경우 P1에서 P2로 연결해 직접 접근이 가능합니다.



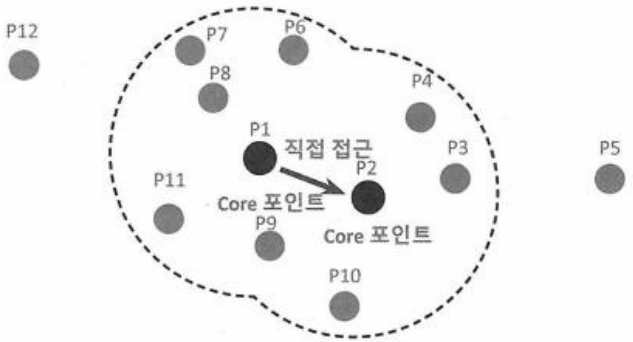

Core Point 끼리는 직접 접근 가능

Prof. Hyerim Bae (hrbae@pusan.ac.kr)

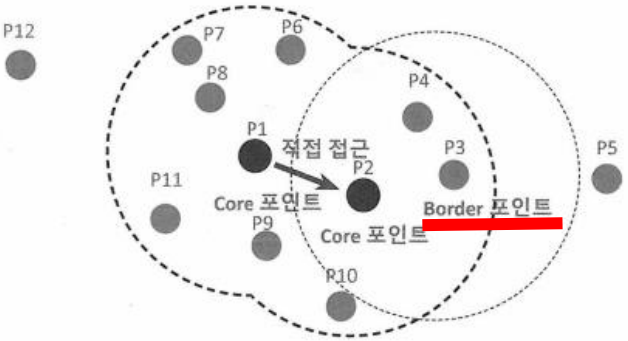
# DBSCAN (비계층적 군집 분석)

- 진행 방법

**5.** 특정 핵심 포인트에서 직접 접근이 가능한 다른 핵심 포인트를 서로 연결하면서 군집화를 구성합니다. 이러한 방식으로 점차적으로 군집(Cluster) 영역을 확장해 나가는 것이 DBSCAN 군집화 방식입니다.




**6.** P3 데이터의 경우 반경 내에 포함되는 이웃 데이터는 P2, P4로 2개이므로 군집으로 구분할 수 있는 핵심 포인트가 될 수 없습니다. 하지만 이웃 데이터 중에 핵심 포인트인 P2를 가지고 있습니다. 이처럼 자신은 핵심 포인트가 아니지만, 이웃 데이터로 핵심 포인트를 가지고 있는 데이터를 경계 포인트(Border Point)라고 합니다. 경계 포인트는 군집의 외곽을 형성합니다.

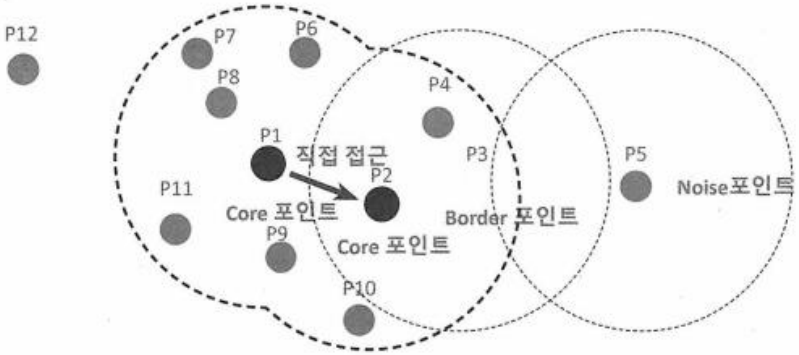



직접 접근을 통하여 군집을 형성

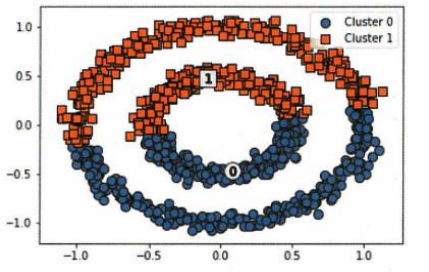
# DBSCAN (비계층적 군집 분석)

- 진행 방법

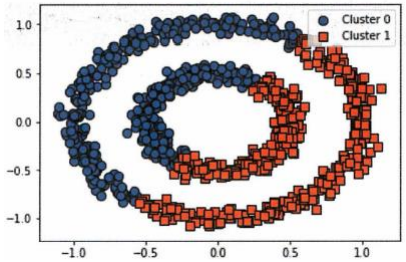
**7.** 다음 그림의 P5와 같이 반경 내에 최소 데이터를 가지고 있지도 않고, 핵심 포인트 또한 이웃 데이터로 가지고 있지 않는 데이터를 잡음 포인트(Noise Point)라고 합니다.



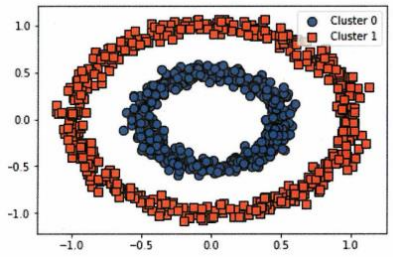

- 장점




K-means




GMM




DBSCAN

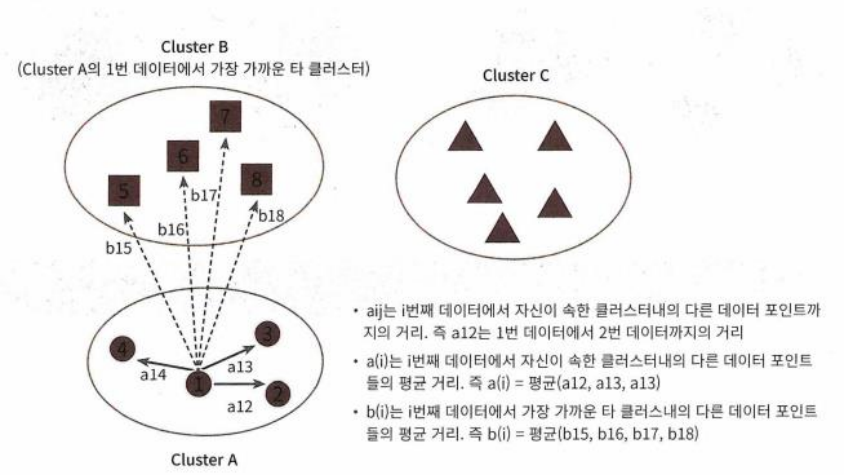
- 밀도 기반 군집화이기 때문에 내부와 외부의 원형으로 구성된 더 복잡한 형태의 데이터 세트에서도 군집화가 잘 수행된다.

Prof. Hyerim Bae (hrbae@pusan.ac.kr)

# 군집화 평가 – 실루엣 분석

- 각 군집 가의 거리가 얼마나 효율적으로 분리돼 있는지

  = 다른 군집과의 거리는 떨어져 있고 동일 군집끼리의 데이터는 서로 가깝게 잘 뭉쳐 있다.
- 실루엣 분석은 실루엣 계수를 기반(실루엣 계수는 개별 데이터가 가지는 군집화 지표)
- 실루엣 계수는 해당 데이터가 같은 군집 내의 데이터와 얼마나 가깝게 군집화돼 있고, 다른 군집에 있는 데이터와는 얼마나 멀리 분리돼 있는지를 나타내는 지표




실루엣 계수

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

같은 군집

다른 군집

- 실루엣 계수는 -1~1 사이의 값을 가지며, 1로 가까워질수록 근처의 군집과 더 멀리 떨어져 있다는 것이고 0에 가까울수록 근처의 군집과 가까워진다는 것
- - 값은 아예 다른 군집에 데이터 포인트가 할당됐다는 것

Prof. Hyerim Bae (hrbae@pusan.ac.kr)

# 군집화 모델 코드 구현

# 참고 자료

- Pytorch로 시작하는 딥러닝 입문
- 귀퉁이 서재 블로그(https://bkshin.tistory.com/)