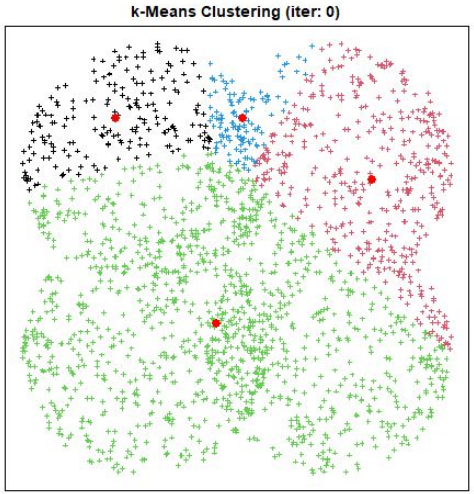
군집 (Clustering)



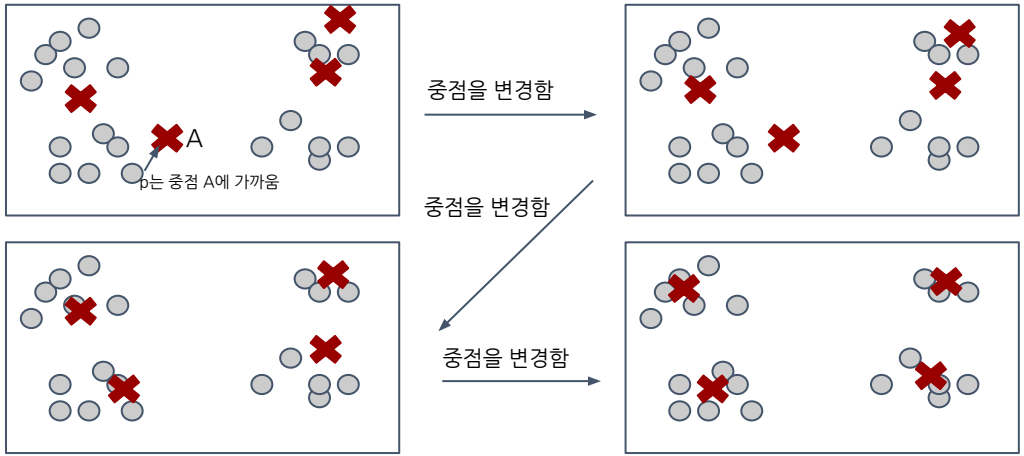
군집(Clustering)

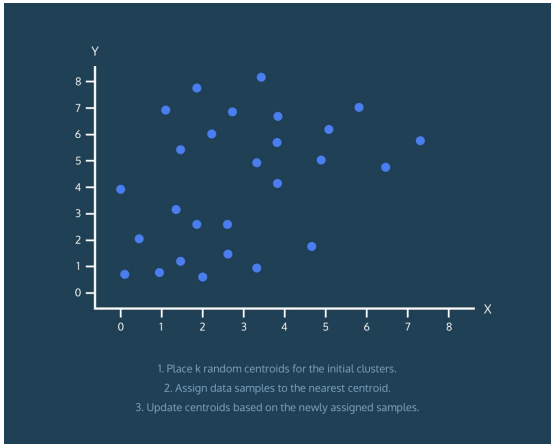
- 데이터 포인트들을 별개의 군집으로 그룹화 하는 것
- 유사성이 높은 데이터들을 동일한 그룹으로 분류하고 서로 다른 군집들이 상이성을 가지도록 그룹화 한다
- 활용 분야 :
  - 고객, 마켓, 브랜드, 사회 경제 활동 세분화(Segmentation)
  - Image 검출, 세분화, 트래킹
  - 이상 검출(Anomaly detection)

- 어떻게 유사성을 정의할 것인가?

K-Means clustering

- 비지도학습(Unsupervised Learning) 으로 학습 데이터들을 K개 그룹(Cluster)으로 나눈다
- K개 클러스터(ex: K=4) 의 중점 (Centroid)을 임의로 부여한 후 각 중점과 가까운 점들을 찾는다.(ex : 점 p는 중점 A에 가까움)
- 중점과 가까운 점들의 평균 지점(무게 중심)을 계산하여 각 클러스터의 새로운 중점으로 사용한다
- 새로운 중점과 가까운 점들을 다시 찾고, 평균 지점을 계산하여 또 새로운 클러스터의 중점으로 사용한다. 클러스터의 중점이 변하지 않을 때까지 반복한다.



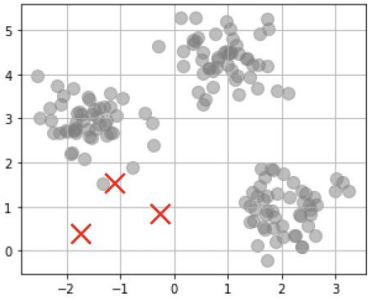


출처: https://i.imgur.com/WL1tlZ4.gif

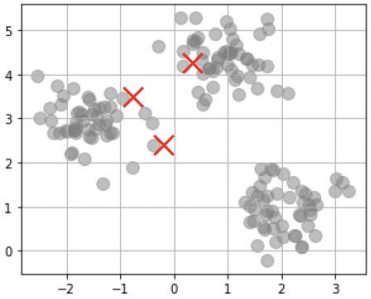
K-Means++

- K-Means++ 알고리즘은 초기 중점을 좀 더 합리적으로 설정하는 방법이다
- 처음에는 임의의 데이터를 선택해서 첫번째 중점으로 설정한다. 그리고 이 중점과 거리가 먼 데이터를 선택해서 두번째 중점으로 설정한다. 이런 방식으로 k개의 초기 중점을 설정한다.

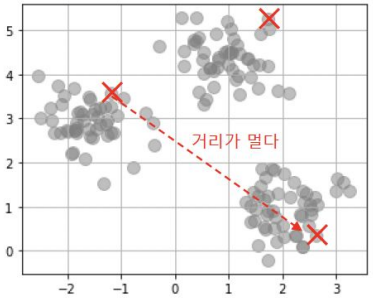
A) 초기 중점이 데이터 영역 밖에 모여 있다.



B) 초기 중점이 데이터 영역에 있지만 가까이 모여 있다.



C) 초기 중점이 데이터 영역에 있으며 서로 멀리 떨어져 있다.



K-Means clustering

실습

K Means

최적의 K를 찾는 방법

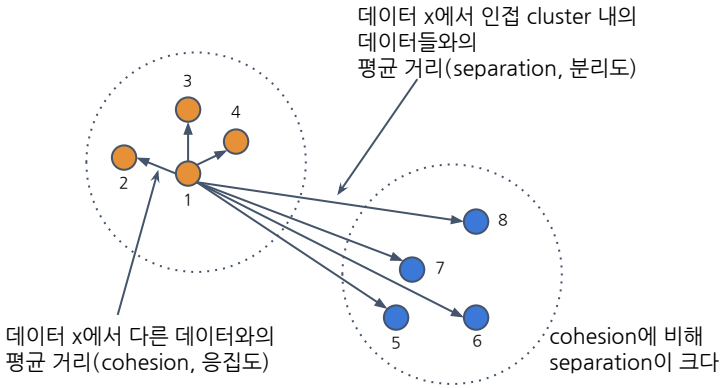
- K-Means Elbow Method
  - K-Means는 unsupervised learning으로 정답에 해당하는 label이나 target이 없으므로 error라는 개념이 없다
  - 군집화가 잘 된 경우라면 각 중점과 해당 cluster내의 데이터들 간의 거리 합이 작을 것이다. 따라서 이 거리의 합을 error(SSE)의 대용치로 쓸 수 있다.
  - Cluster 개수인 k가 증가할 수록 SSE는 줄어든다. 그러나 K가 증가할수록 줄어드는 폭이 작아진다
  - K가 증가할 때 SSE가 줄어들긴 하지만, 줄어드는 폭이 갑자기 작아지는 지점의 K 값을 최적 cluster 개수라 할 수 있다.

실습

K-Means Elbow

군집 평가

- 실루엣(Silhouette) 방법에 의한 K-Means 품질 평가
  - 군집화가 잘 됐는지 여부는 아래와 같이 응집력(cohesion)과 분리도(separation)을 사용한 실루엣 (s) 계수로 평가할 수 있다. (0<s<1)
  - 분리가 잘 된 경우는 b≫a 이므로 실루엣 계수가 크지만(1에 가까움), cluster가 겹치는 경우는 b≅a 이므로 계수가 작아진다(0에 가까워짐)

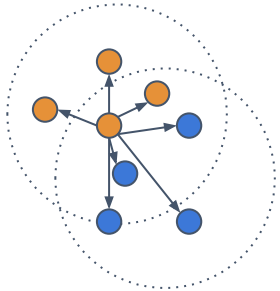


분리가 잘 된 경우

$$S(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)}$$

i = 1일때

ai = 평균(a12, a13, a14)
bi = 평균(b15, b16, b17, b18)



cohesion에 비해
separation이 크다

클러스터가 겹친 경우

K-Means clustering

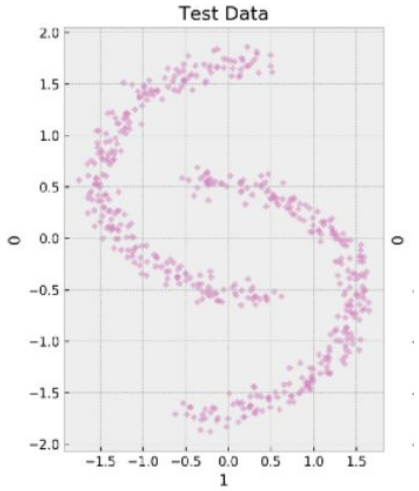
실습

K-Means Elbow
& Silhouette

DBSCAN

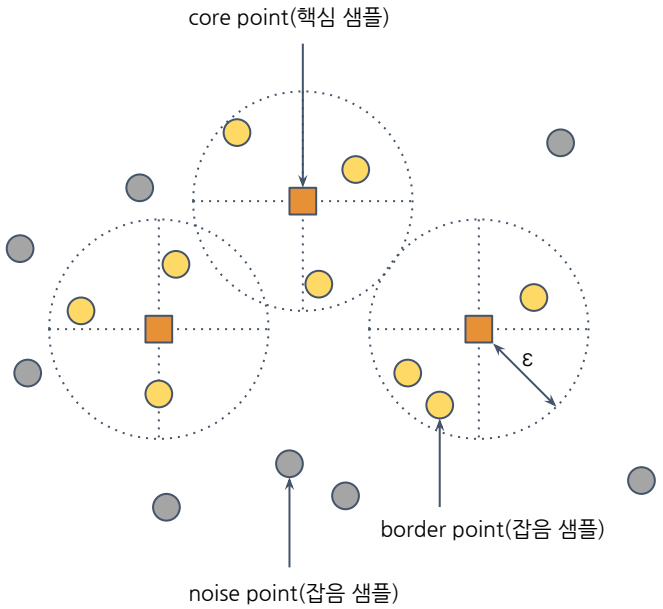
- Density-Based Spatial Clustering of Applications with Noise
- 다음과 같은 데이터는 어떻게 clustering을 할까?

  ➡️ DBSCAN

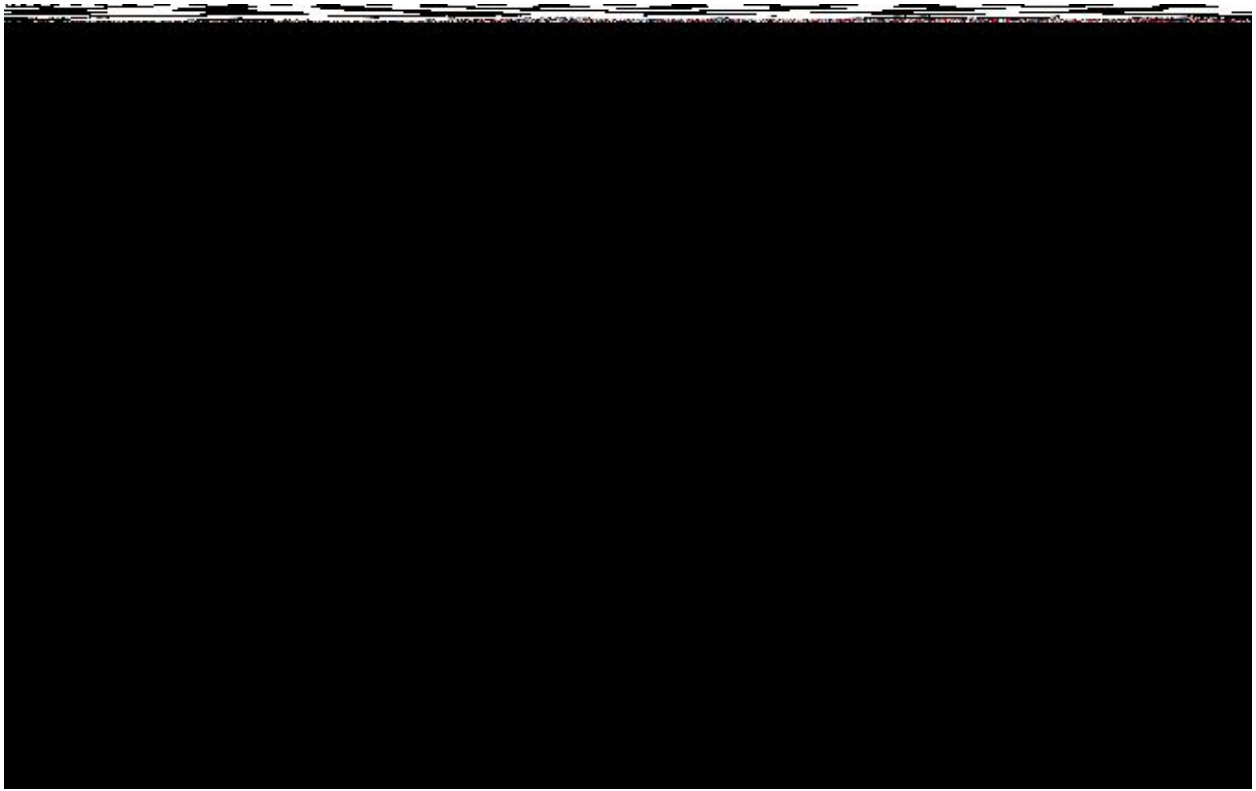


DBSCAN(Density-Based Spatial Clustering of Applications with Noise)

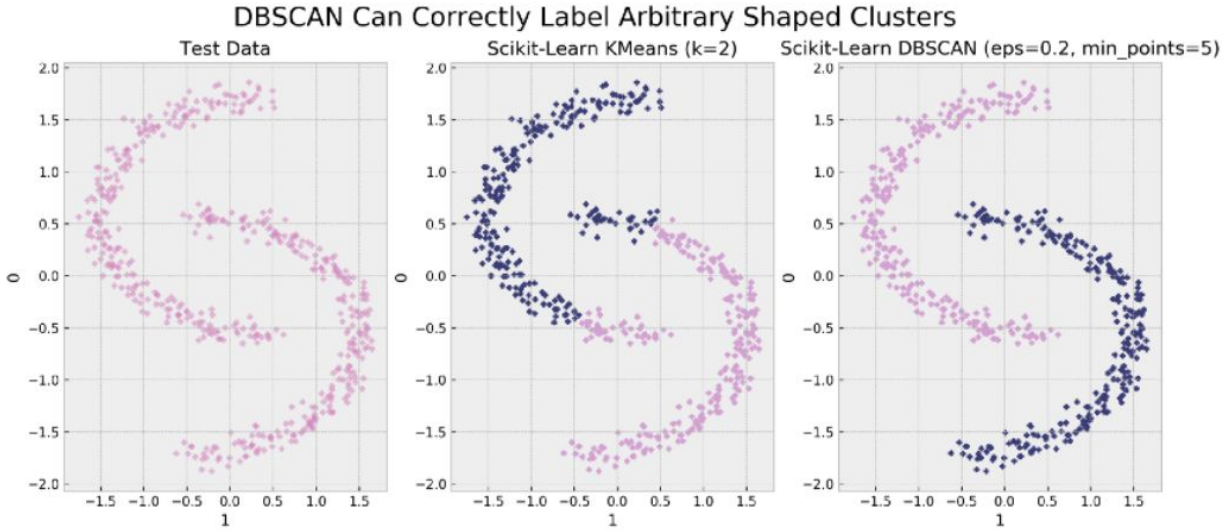
- DBSCAN은 밀집도 기반의 군집 알고리즘
- 밀집도는 특정 반경 ε안에 있는 샘플 개수로 정의
- 반경 ε안에 있어야 할 최소 데이터 개수 (min points)를 정의 (ex, 3개)
- 데이터 마다 반경 ε인 원을 그린 후 원 안에 있는 데이터 개수를 카운트
- 카운트 결과가 사전에 정의한 min points 이상이면 (3개 이상이면)
  해당 데이터를 core point (핵심 샘플)로 표시
- 카운트 결과가 min point 이하이면 (2개 라면) border point (경계 샘플)나
  noise point(잡음 샘플)이 되며 나중에 원안의 샘플 중(2개 중 하나라도)
  core point가 생기면 이 데이터는 border point로 표시, 아니면 noise point
- 모든 데이터에 대해 위의 절차가 끝나면 core point 에 연결된 데이터들을
  하나의 cluster로 설정
- DBSCAN은 K-Means나 계층 군집과는 달리 모든 샘플을
  클러스터에 할당하지 않고 잡음 샘플을 구분



DBSCAN(Density-Based Spatial Clustering of Applications with Noise)



DBSCAN



DBSCAN

실습

DBSCAN