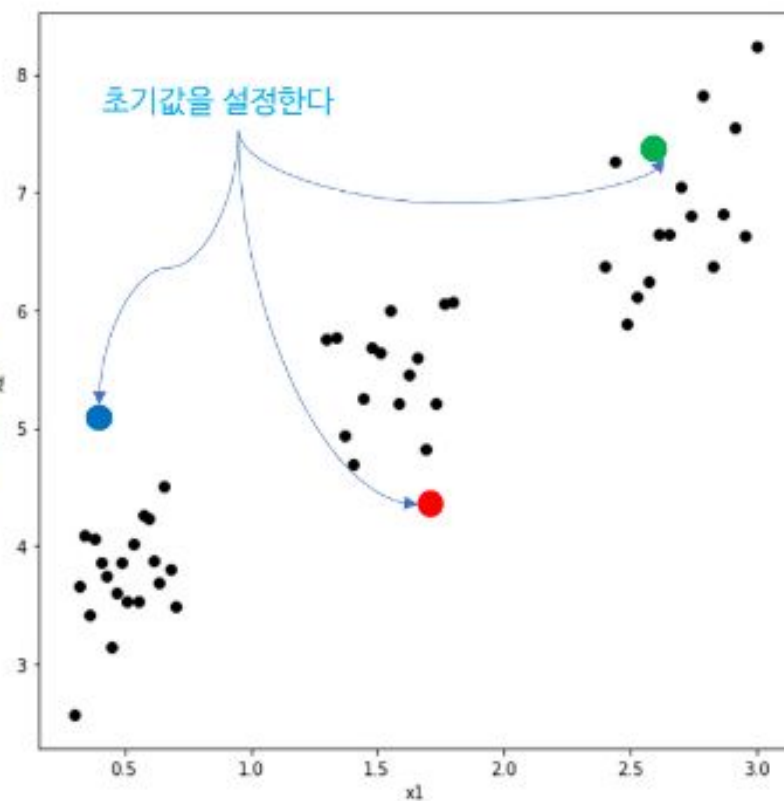
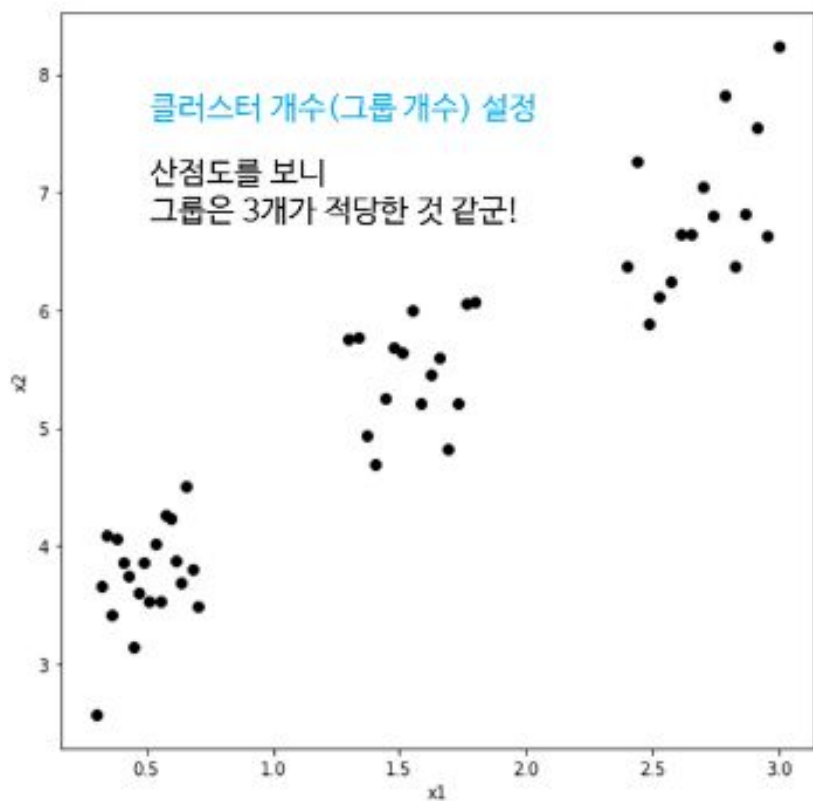


군집알고리즘

K-Means 알고리즘

- K-Means 알고리즘 과정

- 1. 데이터셋에서 K 개의 centroids를 임의로 지정.
- 2. 각 데이터들을 가장 가까운 centroids가 속한 그룹에 할당.
- 3. 2번 과정에서 할당된 결과를 바탕으로 centroids를 새롭게 지정합니다.
- 4. 2 ~ 3번 과정을 centroids가 더 이상 변하지 않을 때 까지 반복합니다.



- 장점

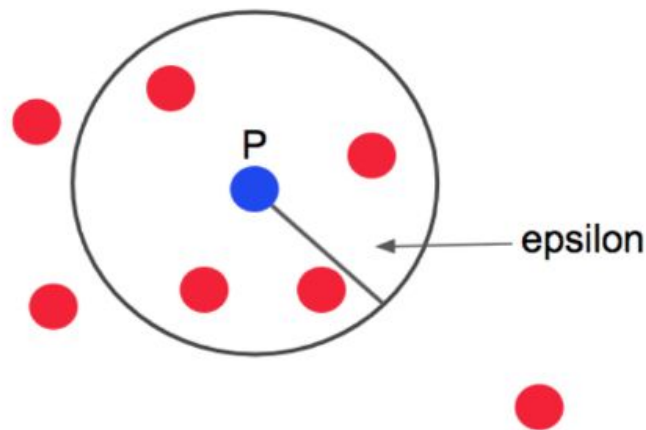
- 직관적이고 구현이 쉽다.
- 대용량 데이터 적용 가능

- 단점

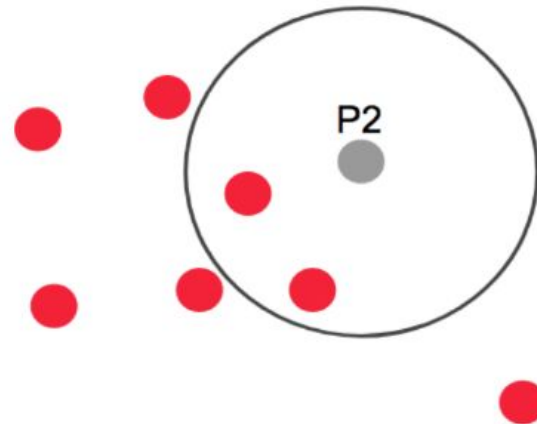
- 초기값에 민감
- 이상치에 영향
- 그룹 내 분산 구조 반영 불가
- 클러스터링 개수를 정해야 함

DBSCAN (밀도 기반 클러스터링)

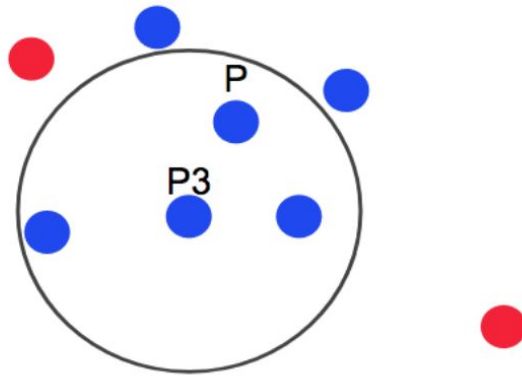
- 아래 그림에서 $\text{minPts} = 4$ 라고 하면, 파란점 P를 중심으로 반경 epsilon 내에 점이 4개 이상 있으면 하나의 군집으로 판단할 수 있는데, 아래 그림은 점이 5개가 있기 때문에 하나의 군집으로 판단이 되고, P는 core point가 된다.



- P2의 경우 점 P2를 기반으로 epsilon 반경내의 점이 3개 이기 때문에, minPts=4에 미치지 못하기 때문에, 군집의 중심이 되는 core point는 되지 못하지만, 앞의 점 P를 core point로 하는 군집에는 속하기 때문에 이를 border point (경계점)이라고 한다.



- P3를 중심으로 하는 반경내에 다른 core point P가 포함이 되어 있는데, 이 경우 core point P와 P3는 연결되어 있다고 하고 하나의 군집으로 묶이게 된다.



- P4는 어떤 점을 중심으로 하더라도 $\text{minPts}=4$ 를 만족하는 범위에 포함이 되지 않는다. 즉 어느 군집에도 속하지 않는 outlier가 되는데, 이를 noise point라고 한다.

