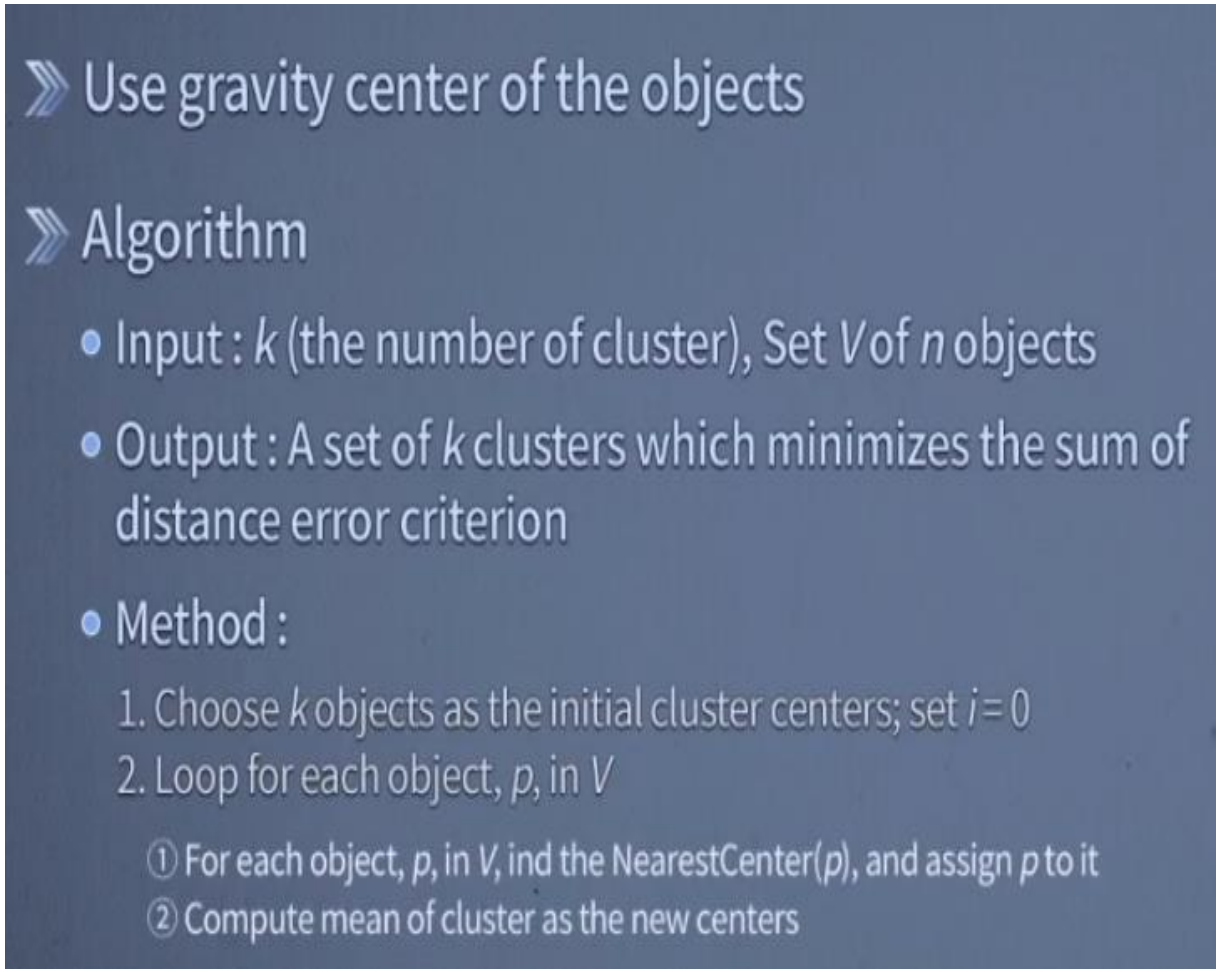
K-means

K-means, Clustering

- 데이터 중에서 유사한 데이터 그룹을 찾는 것
- 그룹 내 similarity 는 높고,
- 그룹 간 similarity 는 낮을수록 좋음
- 정답은 없음

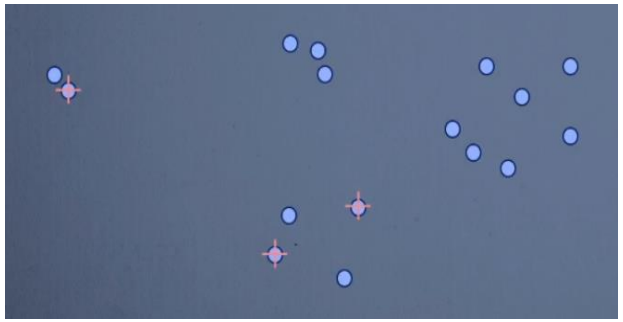
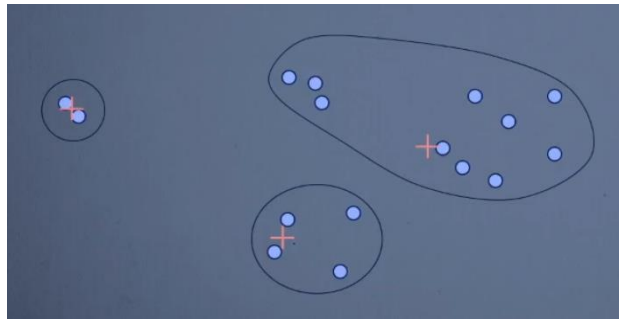
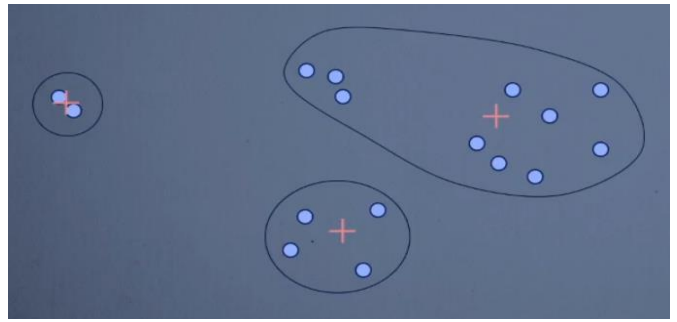
K-means



Mean : 데이터의 Center Point
K : K개(Grouping 개수)

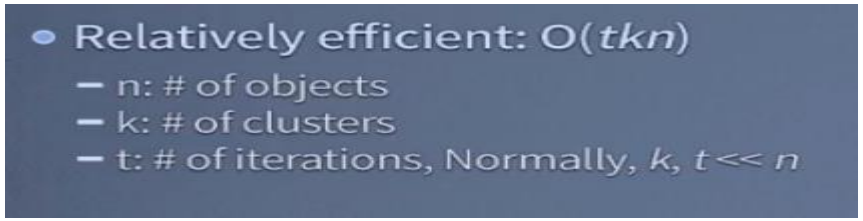
K-means

- 랜덤하게 3개의 데이터를 Random으로 선정(3개의 Cluster의 중심점을 의미)
- 각각의 데이터를 Cluster Center에 할당(가까운 쪽)
- Cluster Center를 Update, 다시 데이터 할당을 반복

K-means 장점 단점

- 장점
  - 상대적으로 Efficient한 Algorithm이기 때문에 일반적으로 사용



- 단점
  - 시작점을 다르게 잡으면 최종적으로 Cluster 결과가 변화함
  - Real Number 또는 Integer과 같이 평균 구할 수 있는 데이터 타입만 적용 가능
  - 몇 개의 Cluster로 묶을지 미리 선정해야함
  - Noisy Data나 Outlier에 민감함
  - 볼록하지 않은 Non-Convex한 Shape에 Cluster를 찾아낼 수 없음