Clustering Method

➢ 통계적 학습 모델의 목적

: 데이터를 통해 값 예측 / 분류 / 데이터 구조 파악

➢ 통계적 학습 모델의 종류

- ✓ 지도학습(Supervised learning) : 예측과 분류를 진행하는 반응값(response)이 존재
  - ❖ 회귀(Regression) : 반응값이 연속형인 경우
  - ❖ 분류(Classification) : 반응값이 범주형인 경우
  - ❖ Parametric model : 반응값과 독립변수 사이에 특별한 모양을 가정 ex) 선형 회귀 모델
  - ❖ Non-Parametric model : 반응값과 독립변수 사이에 특별한 가정 x
- ✓ 비지도학습(Unsupervised learning) : 반응값이 없고 독립변수만 존재하는 학습 모델, 데이터 구조 파악이 주 목적

➢ 클러스터링의 종류

- ➢ 클러스터 중심, 평균 기반 알고리즘

  : K-means

- ➢ 클러스터 빈도 수 기반 알고리즘

  : K-medoids

- ➢ 계층적 클러스터링

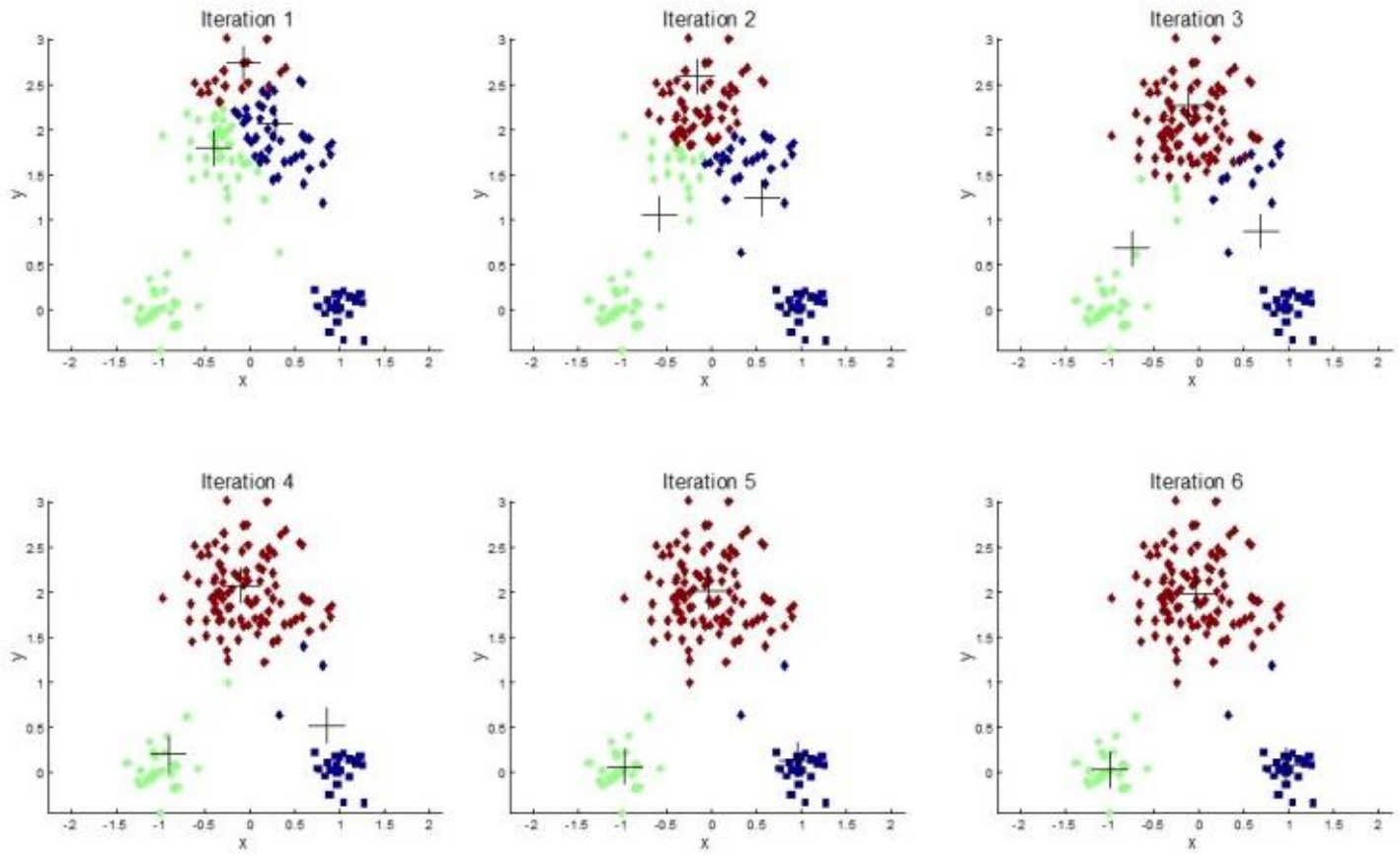
  : 응집형 알고리즘, 분리형 알고리즘

- ➢ 밀도기반 클러스터링

  : DBSCAN

➢ K-Means, K-medoids 알고리즘

: 군집의 중심을 정하고 군집의 중심과 데이터 사이의 거리를 측정하여 군집 형성, 다시 군집의 중심을 정하는 과정을 수렴할 때 까지 반복하는 알고리즘



➢ 장점

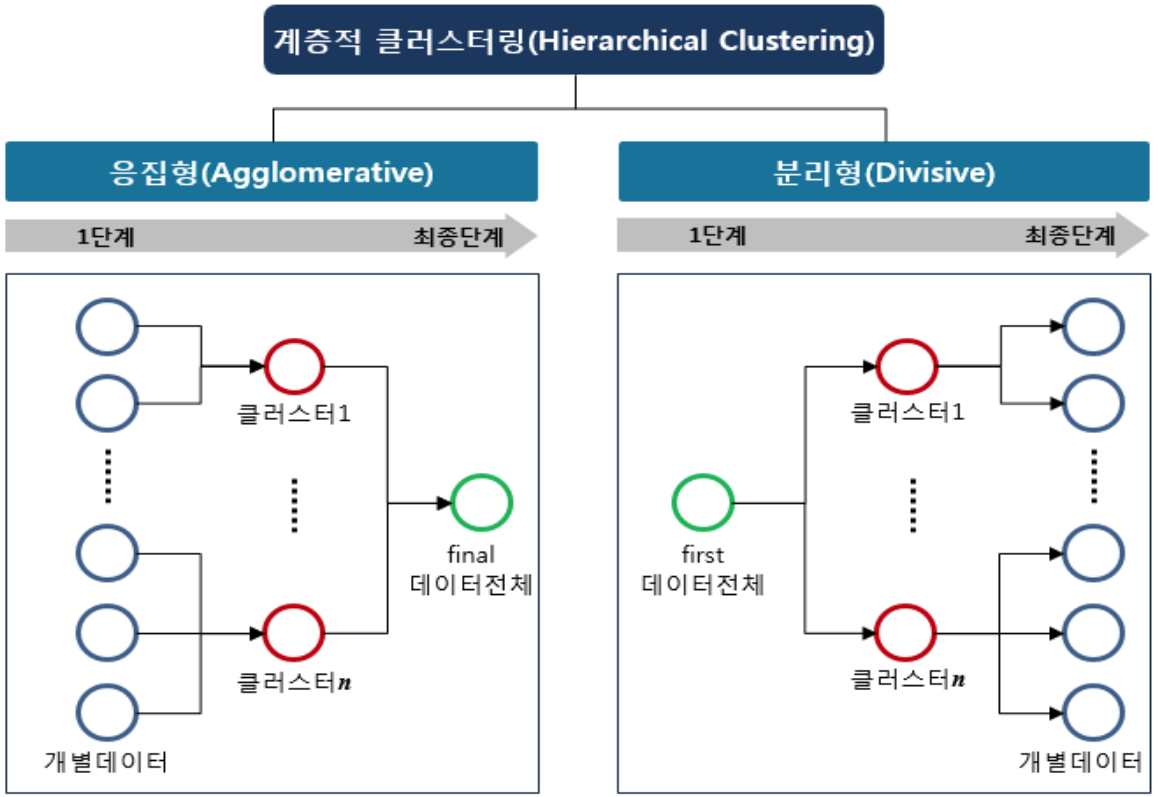
- ✓ 계산이 다른 클러스터링 기법에 비해 빠른 편
- ✓ 데이터가 많을 수록 정확도가 높은 알고리즘
- ✓ 다양한 데이터에 활용 가능(알고리즘 자체에 큰 제한조건이 없다)

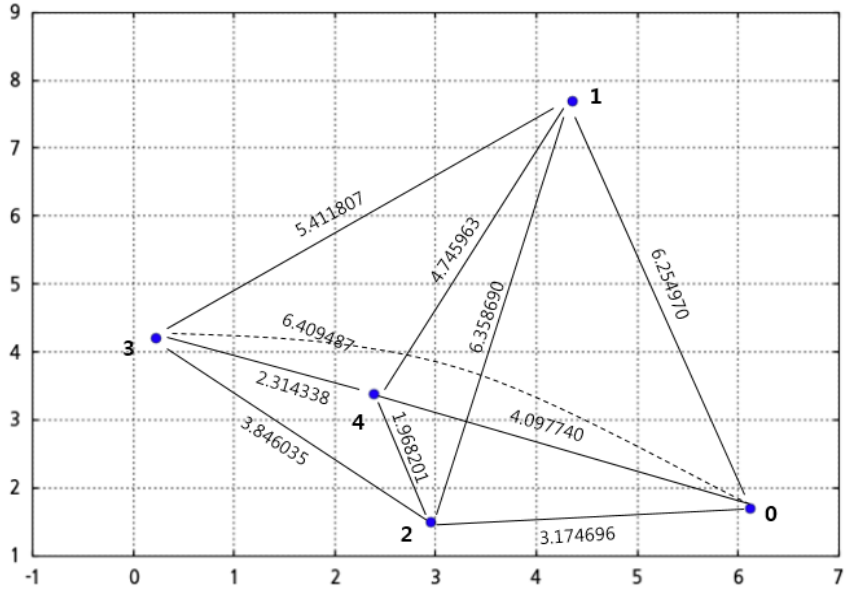
➢ 단점

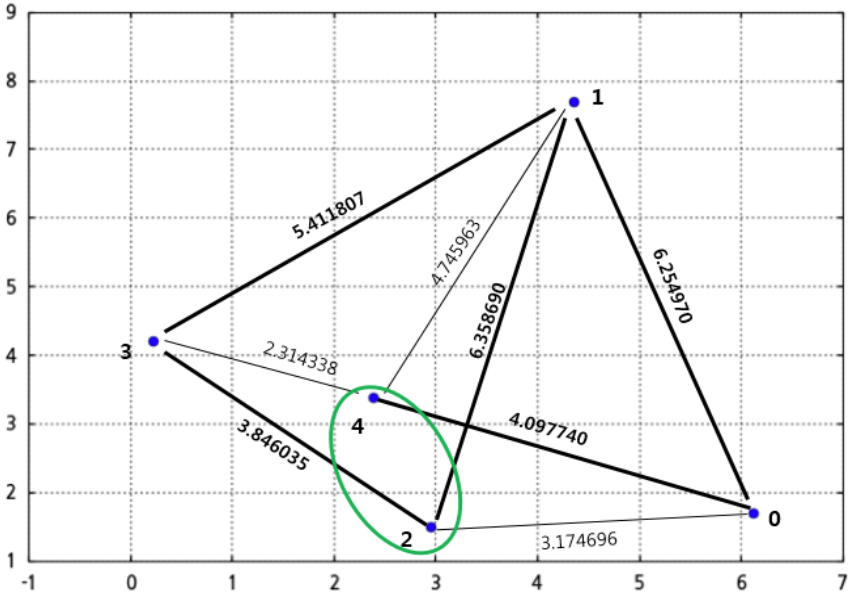
- ✓ HyperParameter인 K에 대한 최적값 선택이 불분명
- ✓ 최초 군집 위치 및 거리 계산 방식에 따라 성능 차이가 심함(모델의 분산이 크다)
- ✓ 클러스터의 중심과 데이터의 거리를 기반으로 클러스터링이 진행되기 때문에 군집 자체가 원형으로 형성됨(원형이 아닌 군집형태를 잡아내기 어려움)

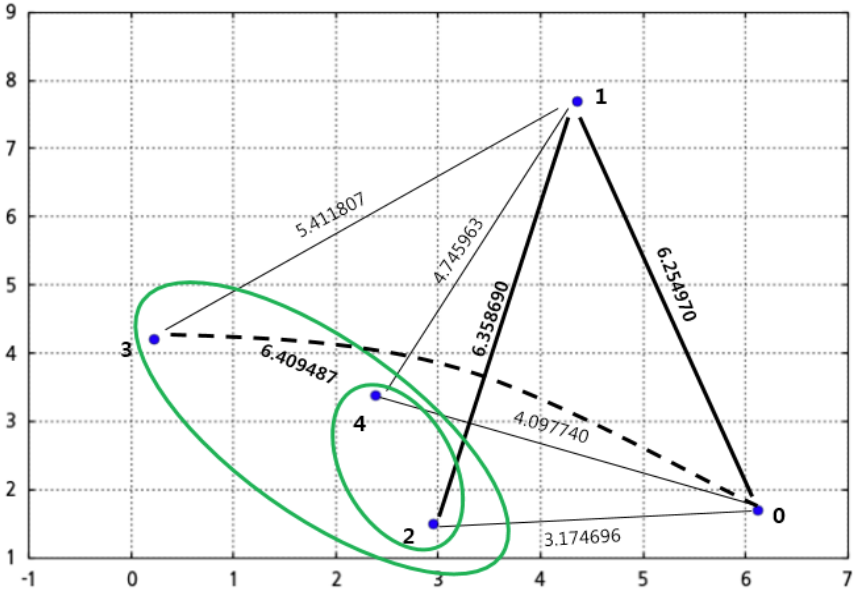
➢ 계층적 클러스터링

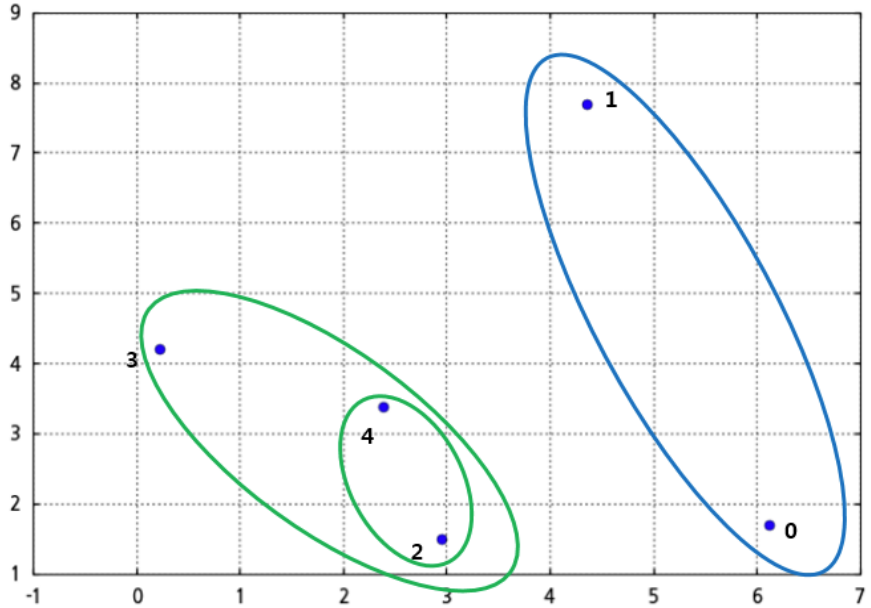
: 데이터 하나하나를 각각의 클러스터로 가정하고 이를 병합 혹은 분리 하여 클러스터를 구성해 나가는 방식의 알고리즘

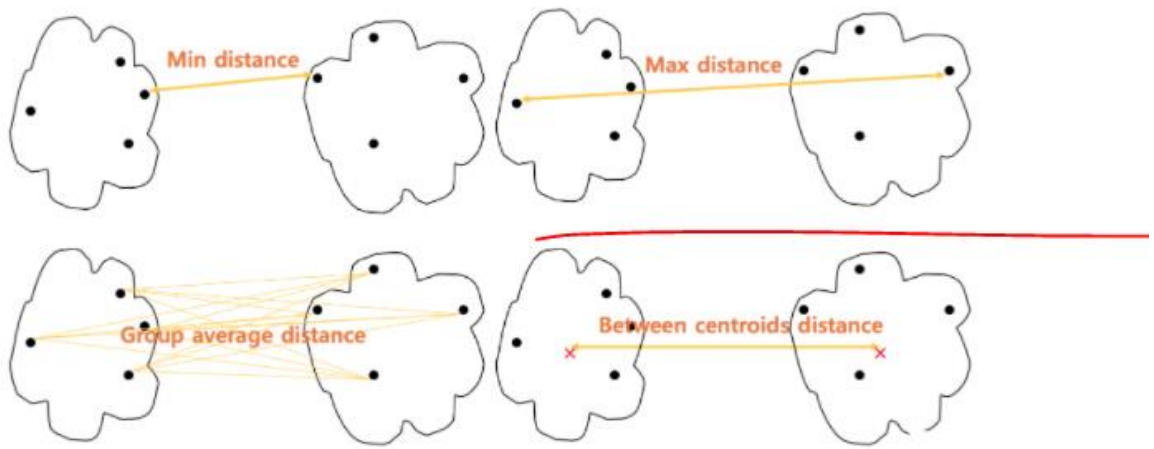


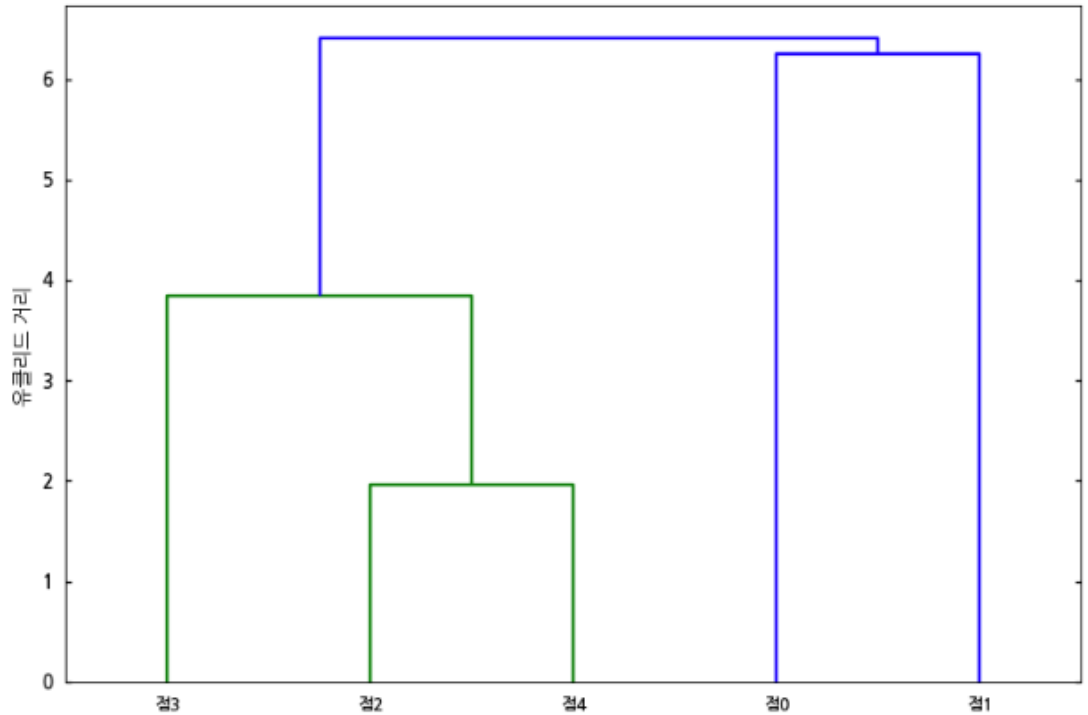

1
2
3
4
0
5.411807
4.745963
6.358690
6.254970
6.409487
2.314338
3.846035
1.968201
4.097740
3.174696


1
2
3
4
0
5.411807
4.745963
6.358690
6.254970
2.314338
3.846035
4.097740
3.174696


1
2
3
4
0
5.411807
4.745963
6.358690
6.254970
6.409487
4.097740
3.174696


1
2
3
4
0


Min distance
Max distance
Group average distance
Between centroids distance


유클리드 거리
0
1
2
3
4
5
6
점3
점2
점4
점0
점1

➢ 장점

- ✓ K-means처럼 클러스터 개수에 대한 가정이 필요 없음
- ✓ 결과의 시각화가 직관적이어서 데이터 구조를 파악하는데 도움이 됨
- ✓ 다양한 혼합 모델의 모수 결정용으로 많이 사용

➢ 단점

- ✓ 시간이 너무 오래 걸림....(진짜 많이 걸림)