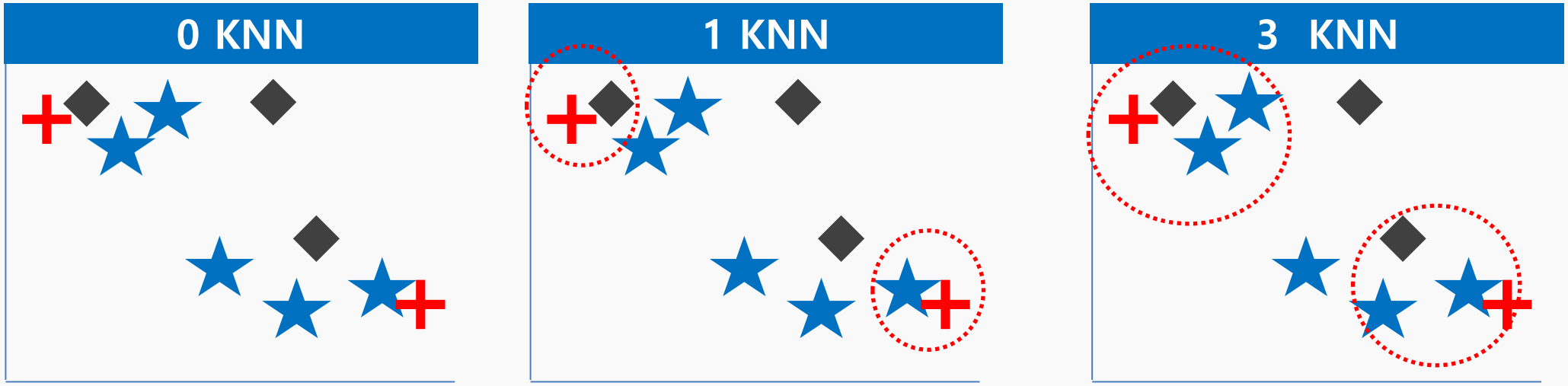
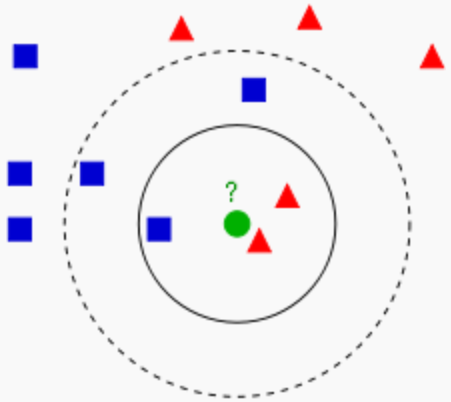
1) KNN(K-Nearest Neighbor: 최근접 이웃 알고리즘) 개요

- 데이터를 묶는 모델링-묶은 데이터 그룹을 하나의 데이터로 인식
  - ✓ 매개변수 K값이 커질수록 정확성이 감소하지만 대그룹으로 묶을 수 있으나
  - ✓ 항목간 경계가 불분명해짐



1) KNN(K-Nearest Neighbor: 최근접 이웃 알고리즘) 개요



- k=3일때

  ■ 은 1개 , 확률은 1/3

  ▲ 은 2개, 확률은 2/3

  ==> ● 은 ▲ 로

  분류됨

- k=5일때

  ■ 은 3개 , 확률은 3/5

  ▲ 은 2개, 확률은 2/5

  ==> ● 은 ■ 로

  분류됨

2) KNN(K-Nearest Neighbor: 최근접 이웃 알고리즘) 알고리즘

- 분류와 회귀에 사용됨
- 지도학습의 일종으로 레이블이 있는 데이터 사용
- 최근접 이웃을 찾아가는 알고리즘으로 K는 최근접 이웃의 갯수임
  - 이웃을 찾을때는 주로 유클리디안 거리를 사용함
  - featur들이 numerical할때, 데이터를 표준화 시켜주는것이 좋음 (R scale함수)

분류와 회귀	• K-NN 분류 출력: 소속된 항목 • K-NN 회귀 출력: 객체의 특성값(k개의 최근접 이웃이 가진 값의 평균)

1) 개요

- 각 개체(대상)의 유사성이 높으며(분산이 작은), 집단간 차이가 큰 대상집단을 분류
- 군집에 속한 객체들의 유사성과 서로 다른 군집에 속한 객체간의 상이성을 규명하는 분석방법
  - 군집의 개수나 구조에 대한 가정없이 데이터로부터 거리를 기준으로 군집화를 유도
    - ✓ 마케팅 조사에서 소비자들의 상품구매행동
    - ✓ life stye에 따른 소비자군을 분류하여 시장 전략수립등에 활용
- 독립적으로 사용되지 않음(데이터 탐색을 위한 방법임)

분류와 회귀	• 임의로 나누는 방법: 신규고객과 기존고객, 고객등급등 • 통계적 기법: clustering, k-means등

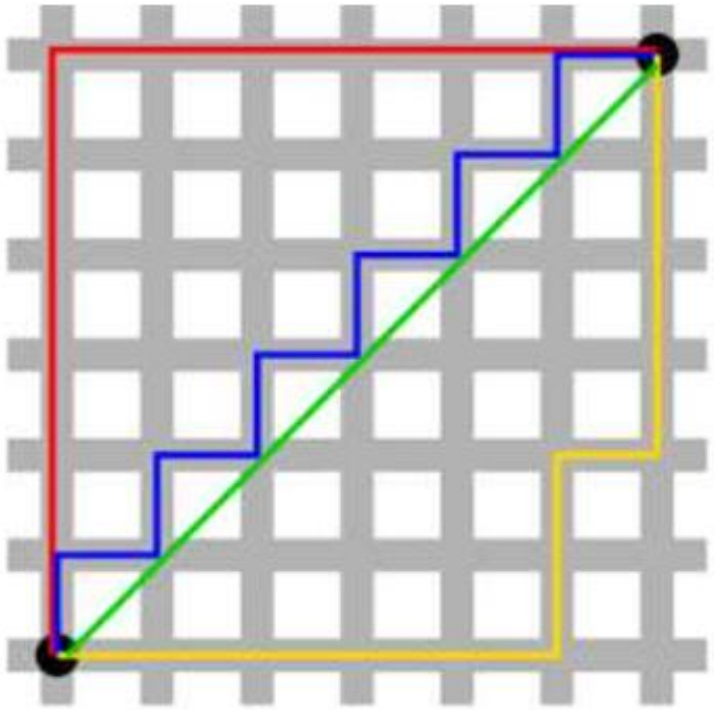
2) 군집분석 방법

- 군집분석은 비지도 학습으로 목표변수가 없음 (테스트와 트레이닝 데이터를 나눌 필요가 없음)

구분	내용
계층적	순차적으로 데이터를 군집화함
비계층적	랜덤으로 데이터를 군집화하고 군집과정에서 중앙값의 변화에 따라 각 데이터드를 적적한 군집으로 이동시켜줌

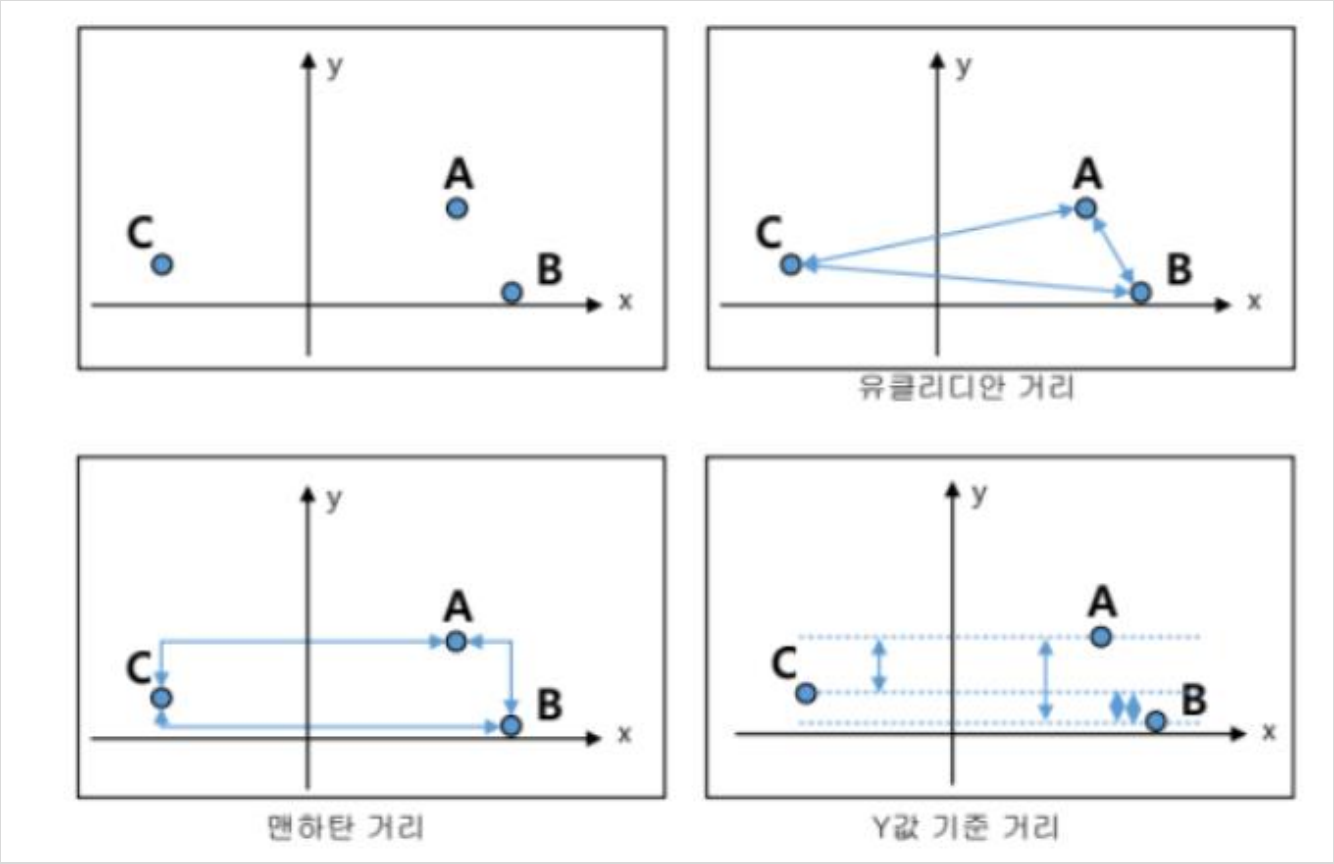
3) 거리: 군집을 묶는 기준

- **Euclidean Distance: 가장 흔히 사용되는 거리**
  척도로 두 관측치 사이의 직선 최단 거리를 의미.
  (우측 그림에서 초록선)
- **Menhatten Distance: 사각형 격자로 이루어진**
  지도에서 출발점부터 도착점 까지 건물들 (사각형)을
  가로지르지 않고 갈 수 있는 최단거리.
  (우측 그림에서 빨간선, 파란선, 노란선)
- **Y값 기준거리**



[출처] http://en.wikipedia.org/wiki/Taxicab_geometry

4) 연속형 변수 거리



Key Point

- 가장 가까운 군집
  - 유클리디안: A와B
  - 맨하탄: A와 B
  - Y값기준: A와 C 또는 C와B

5) 범주형변수 거리

- 두 개체가 서로 다른 범주에 속한 횟수

  데이터=(과정, 프로그램명, 지역)

  A=(통계,R,국내)

  B=(통계,SPSS,국내)

  C=(딥러닝,파이썬,국외)

Key Point

- A와 B의 거리=2
  (프로그램명과 지역이 다름)
- A와 C의 거리=1
  (프로그램명이 다름)
- B와 C의 거리=3
  (과정명, 프로그램명, 지역이 모두 다름)

6) 군집분석시 유의사항

- 차원저주문제에 걸릴수 있음
  - 저차원으로 돌려야함 (대표적인 변수만 선택함)
  - 주성분분석 사용함(PCA)

- 변수별로 단위가 다를때 비교가 어려움
  - 표준화 또는 정규화 과정 거침
    - ✓ 표준화(Standardization): 표준정규분포표를 구하는 식으로 변화
      표준편차를 통해 평균으로부터 얼마나 떨어져 있는지 의미
    - ✓ 정규화(normalization): 0~1사이의 값을 가지도록 변환

7) 군집분석시 평가

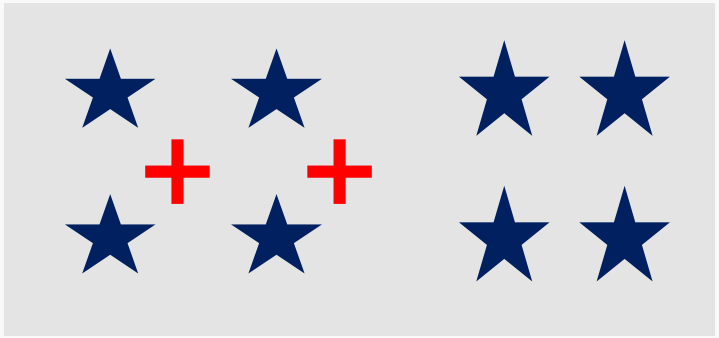
- 유사성이 높을수록 좋음(분산 작음, 거리의 합이 작음)
- 집단간 차이가 크수록 좋음(집단간 거리가 멈)

Davies-Bouldin Index	• 집단간 거리가 클수록, 군집내 중심과 모든 점들의 거리 평균이 클수록 좋음. 결과가 낮을수록 좋음
Dunn Index	• 밀도가 높은 군집이 잘 나뉘어진 클러스터링 결과로 봄. 결과가 높을 수록 좋음.
실루엣(silhouette) 기법	• -1 (bad) ~ 1(good) 까지의 점수를 줌

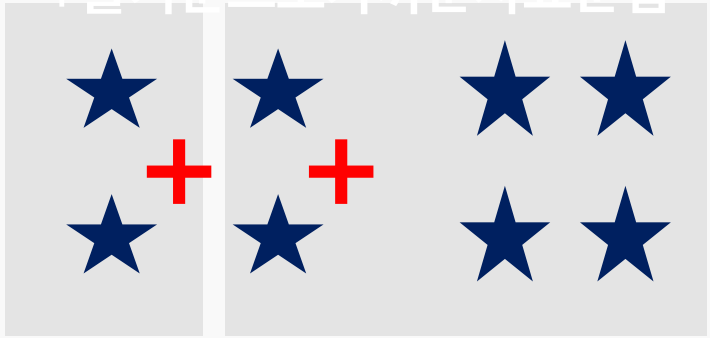
1) 알고리즘 개요

- 예: K-means 군집수 k 2일때 아래 과정을 변화없을때까지 반복함

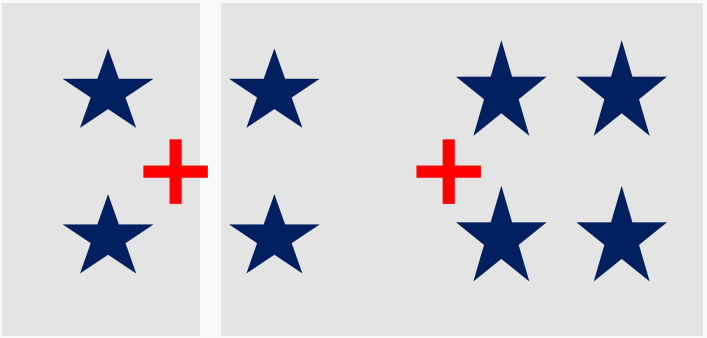
1. 무작위 seed 지정(+위치)



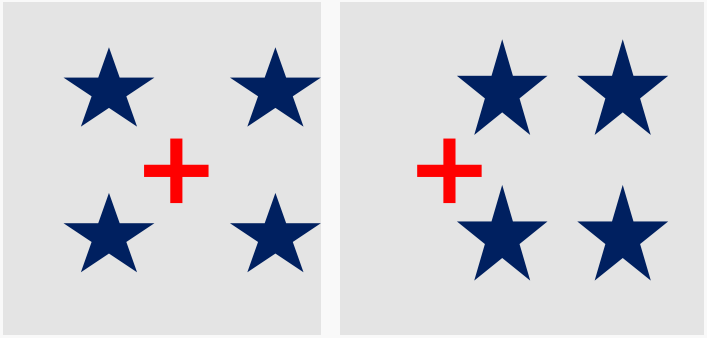
2. +를 기준으로 가까운 자료 군집



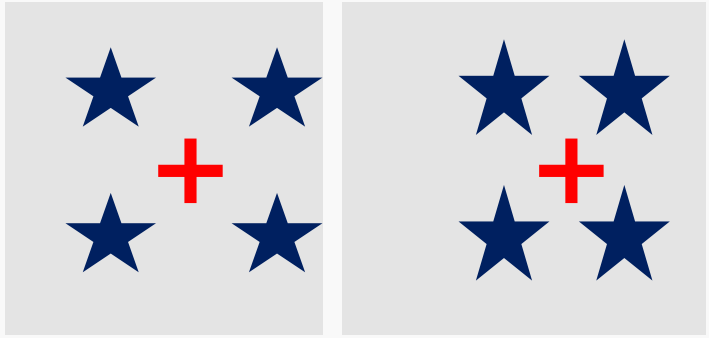
3. 군집내에서 중심값 이동



4. +를 기준으로 다시 군집



5. 군집내에서 중심값 이동



2) K-means 클러스터링 알고리즘(비계층적 군집분석, 분할법)

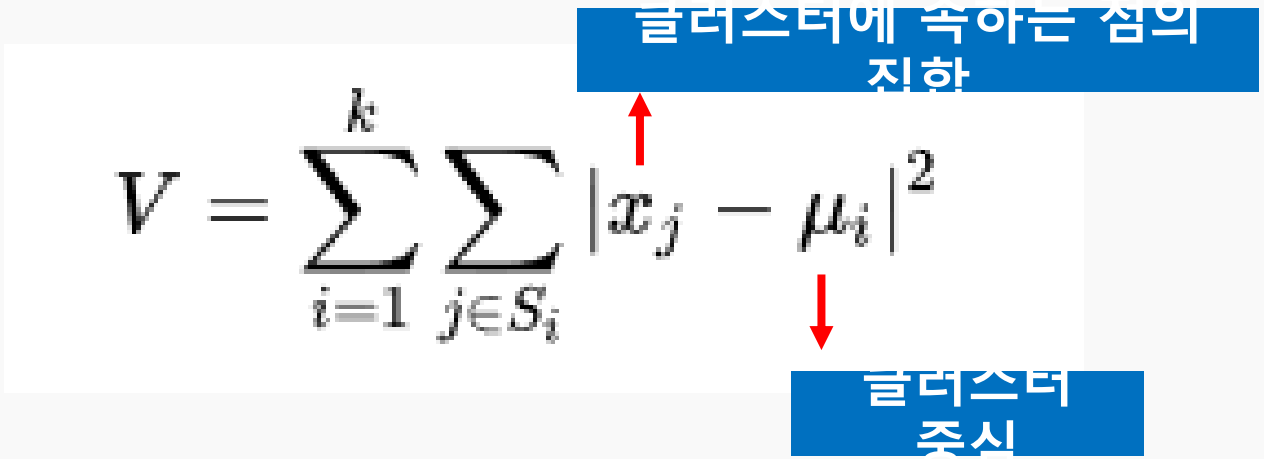
- 비지도 학습(Unsupervised)으로 Y값을 제시해주지 않음.
- 응집도는 최대, 분산도는 최소해야함
  - 주어진 데이터를 k개의 클러스터(군집)으로 묶는 알고리즘
  - 가까운 데이터를 하나의 그룹으로 (KNN에서 활용함)
  - 각 클러스터간 분산 최소화

사용처

- 이상 패턴 추적 / 장바구니 분석
- IOT 기계설비, Computer Vision
- 벡터양자화(컴퓨터 그래픽스에서 색의 종류를 K개로 줄임) 등
- 데이터 전처리 과정에서 많이 사용됨

3) 알고리즘 개요

- 거리 기반의 그룹간 비유사도 비용함수 최소화함
  - 같은 그룹내에서 오브젝트간 유사도는 증가
  - 다른 그룹에 있는 데이터 오브젝트 와의 유사도는 감소
- 비용함수: 각 그룹의 중심과 그룹내의 오브젝트와의 거리의 제곱합
  - 비용함수를 최소화하는 방향으로 각 데이터 오브젝트의 소속 그룹 업데이트



$$V = \sum_{i=1}^{k} \sum_{j \in S_i} |x_j - \mu_i|^2$$

4) 알고리즘 계산 스텝

- [1] K값(그룹의 갯수)를 임의로 지정
- [2] 임의의 중간값에 따라 그룹 지정
  - 방법1: 주어진 k값에 따라 K개의 중간값을 Random하게 고른후, 주변 데이터값과 거리의 제곱 계산
  - 방법2: 최초 k개 데이터 포인트에서 시작해서 가까운 데이터를 바꿔넣으며 거리의 제곱 계산 trial-and-error 반복
- [3] 다른 중간값 지정
  - 방법1: Random하게 뽑은 다른 중간값에서 다시 계산
  - 방법2: 최초 k개중 1개를 바꿔서 다시 계산
- [2], [3] 작업을 계속 반복해서 중간값이 더 이상 움직이지 않을때 까지 진행

5) 초기값설정 μ_i

- 무작위 분할(Random Partition) 사용
  - 각 데이터들을 임의의 클러스터에 배당한후
  - 각 클러스터에 배당된 점들의 평균값을 초기값으로 설정함.

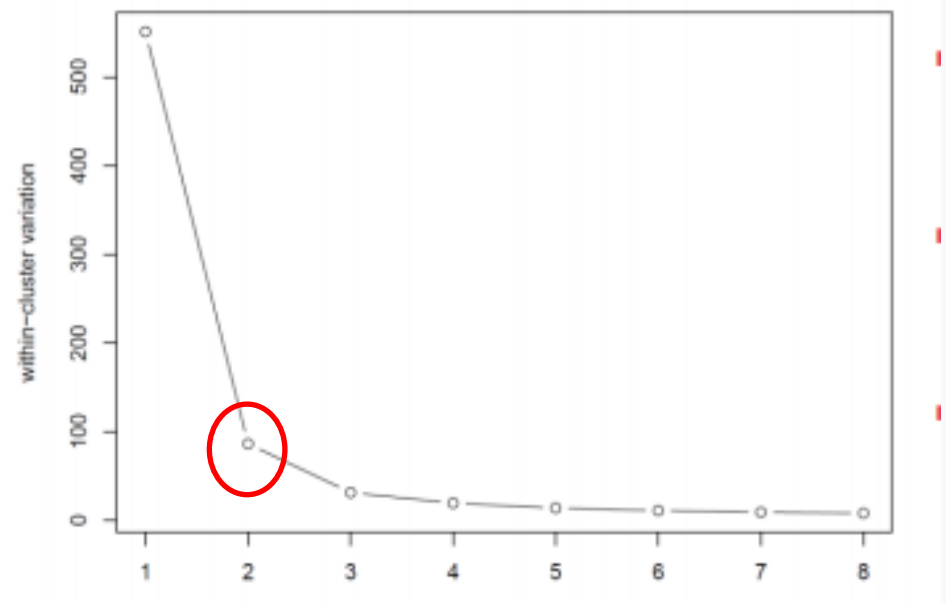
- 무작위 분할의 경우 초기 클러스터가 각 데이터들에 대해 고르게 분포되어 있
- 초기 클러스터의 무게중심들이 데이터 집합의 중심에 ㅏ깝게 위치하는 경향을 띔
- K-조화평균이나 퍼지 K-평균에서 많이 사용됨

주의	• k<m (데이터 갯수보다 그룹 숫자가 작아야함)

1) 클러스터의 수(k) 숫자 설정 방법론

구분	내용
Rule of thumb	가장 간단한 방법으로 데이터의 수가 n이라 할때 필요한 클러스터의 수는 $k \approx \sqrt{n/2}$ 로 계산
Elbow Method	클러스터의 수를 순차적으로 늘려가면서 결과를 모니터링 함 클러스가 추가시 이전보다 결과가 좋지않을때 중단함
정보기준 접근법	클러스터링 모델에 대해 가능도를 계산하는 것이 가능할때 사용하는 방법, k-means 인경우 가우시안 혼합모델에 대한 가능도를 만들어 정보기준값을 설정할 수 있음.

2) Elbow Method를 이용한 최적의 K값 찾기



- Elbow method 꺽이는 포인트가 최적 k값 (K=2에서 elbow가 나타남)
- 소형화(Compactness): 그룹안에 있는 점들이 얼마나 더 가까운가? (짧을수록 좋음) 중간점에서의 거리의 합-최대값
- 분리(Separation): 다른 그룹이 얼마나 멀리 떨어져 있는가? (길수록 좋음) 중간점끼리 거리의 합-최소값을 씀

1) 비계층적 군집방법의 장단점

- n개의 개체를 g개의 군집으로 나눌수 있는 모든 가능한 방법을 점검해 최적화한 군집을 형성

장점

- 주어진 데이터의 내부구조에 대한 사전 정보없이 의미 있는 자료구조를 찾을수 있다.
- 다양한 형태의 데이터에 적용 가능하다.
- 분석방법 적용이 용이하다.

단점

- 가중치와 거리정의가 어렵다
- 초기 군집수를 결정하기 어렵다
- 사전에 주어진 목적이 없으므로 결과 해석이 어렵다.