

Ⅲ. 비지도학습 (Unsupervised Learning)

- Cluster Analysis

① K-means Clustering

Unsupervised Learning

Cluster Analysis

K-Means Clustering

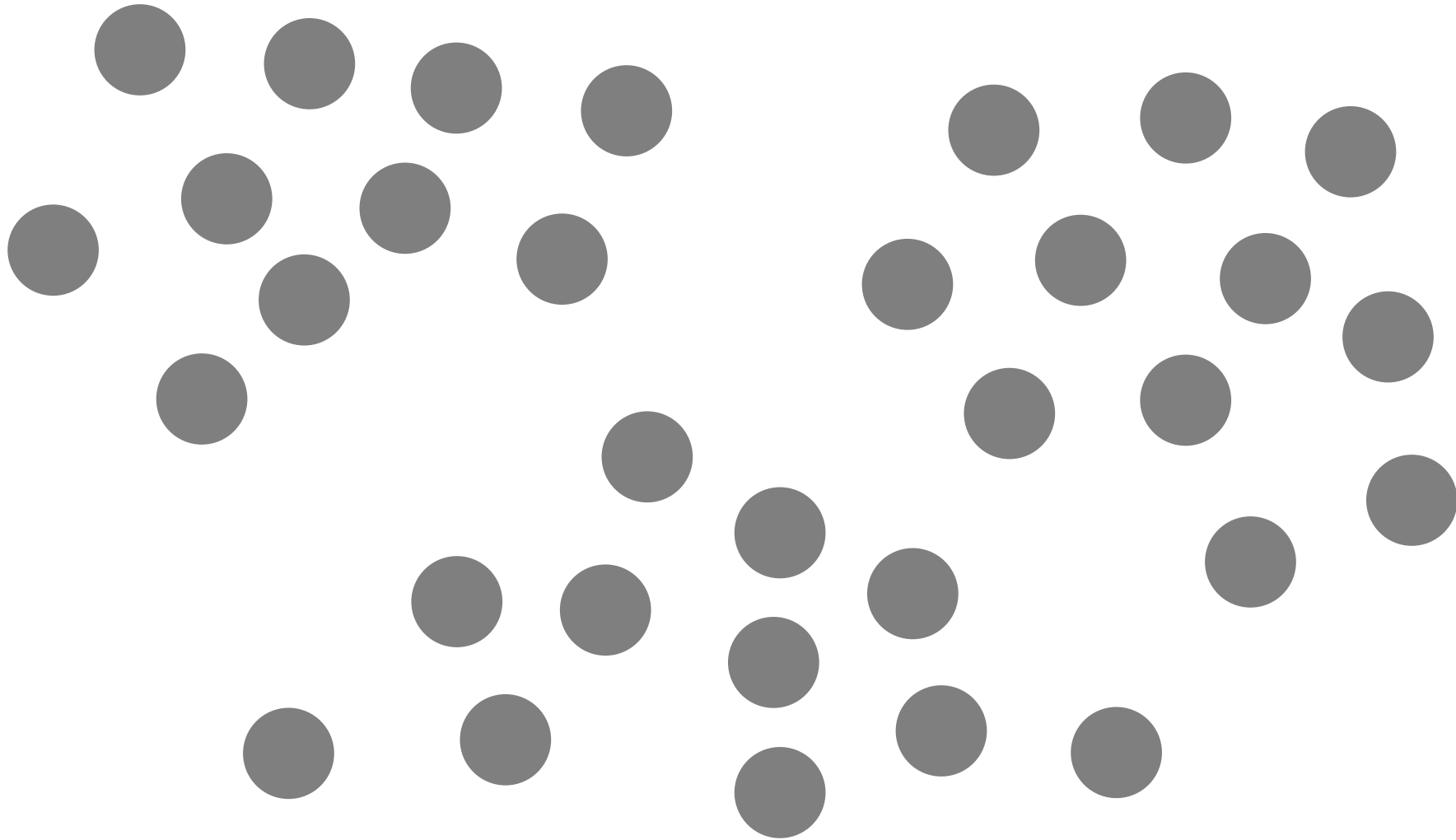
Hierarchical Clustering

Dimension Reduction

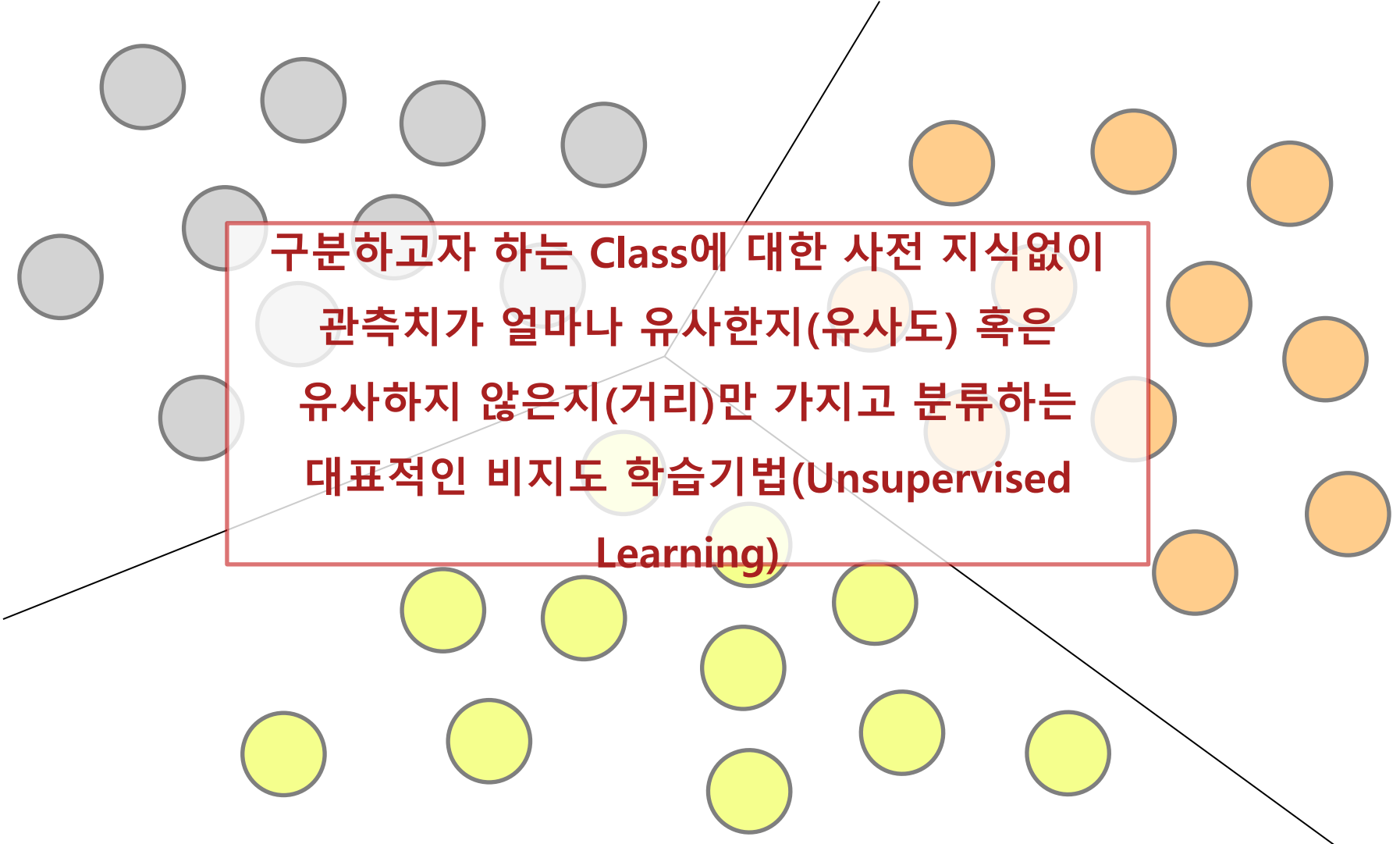
Principal Component Analysis (PCA)

Linear Discriminant Analysis (LDA)

클러스터링(Clustering)이란?



클러스터링(Clustering)이란?

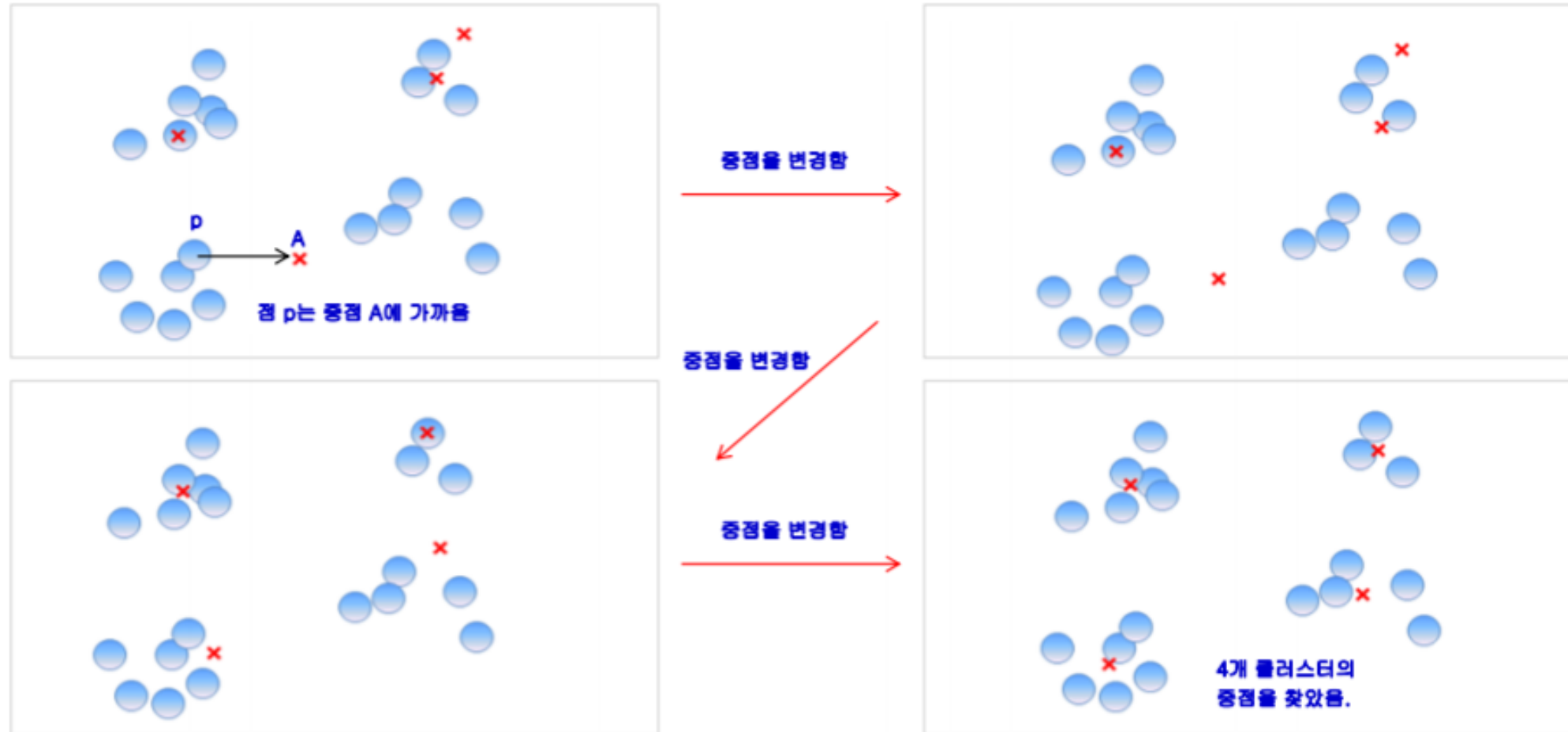


구분하고자 하는 Class에 대한 사전 지식없이
관측치가 얼마나 유사한지(유사도) 혹은
유사하지 않은지(거리)만 가지고 분류하는
대표적인 비지도 학습기법(Unsupervised
Learning)

K-Means Clustering 알고리즘이란?

1. 비 지도학습 (Unsupervised Learning)으로 훈련 데이터들을 K 개 그룹 (Cluster)으로 나눔.
2. K개 클러스터 (ex : K = 4)의 중점 (Centroid)을 임의로 부여한 후 각 중점과 가까운 점들을 찾음. (ex : 점 p는 중점 A에 가까움)
3. 중점과 가까운 점들의 평균 지점 (무게 중심)을 계산하여 각 클러스터의 새로운 중점으로 사용함.
4. 새로운 중점과 가까운 점들을 다시 찾고, 평균 지점을 계산하여 또 새로운 클러스터의 중점으로 사용함. 클러스터의 중점이 변하지 않을 때까지 반복함.

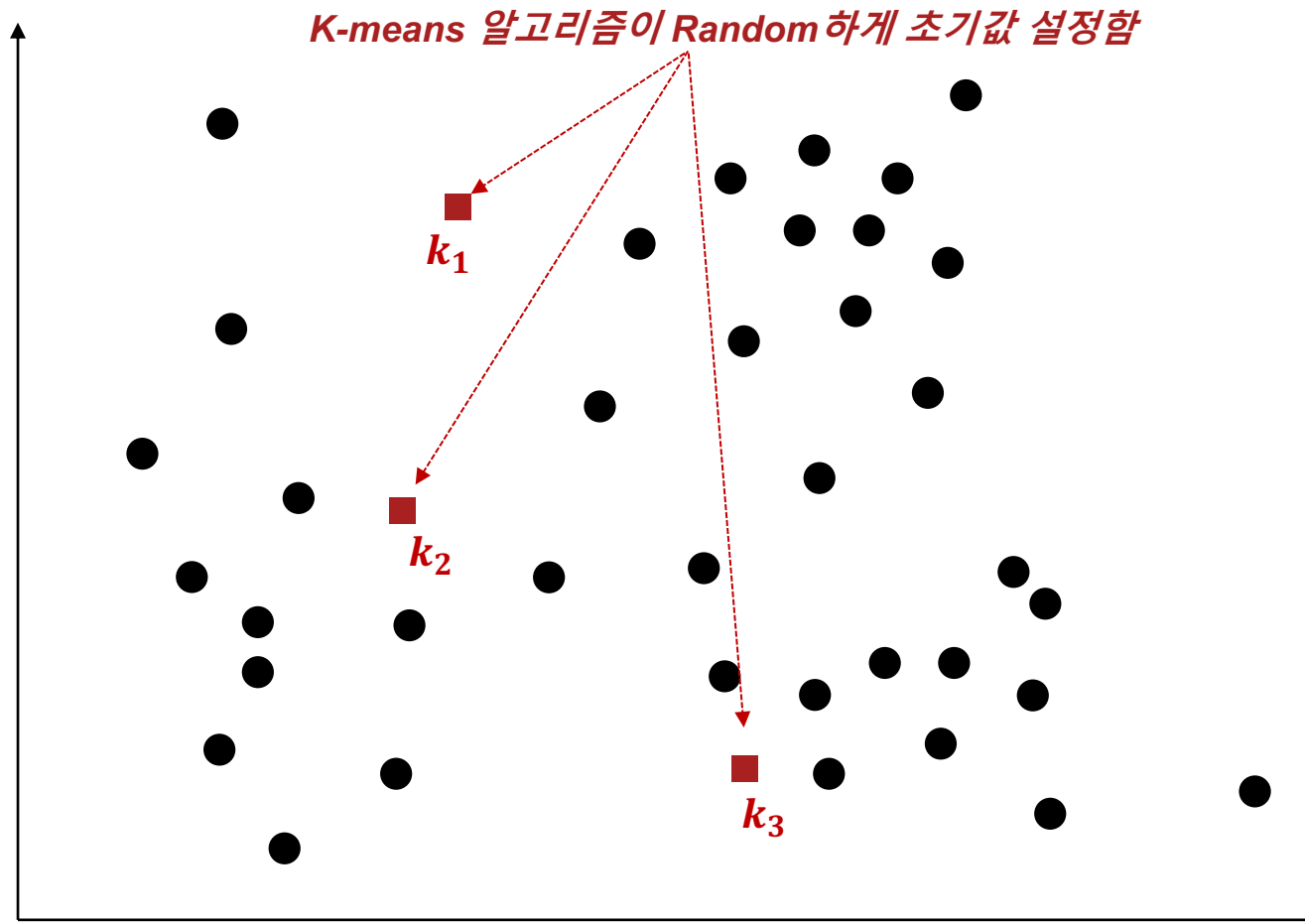
K-Means Clustering 알고리즘이란?



K-means 클러스터분석 (Cluster Analysis)

예 : $k = 3$ 으로 설정

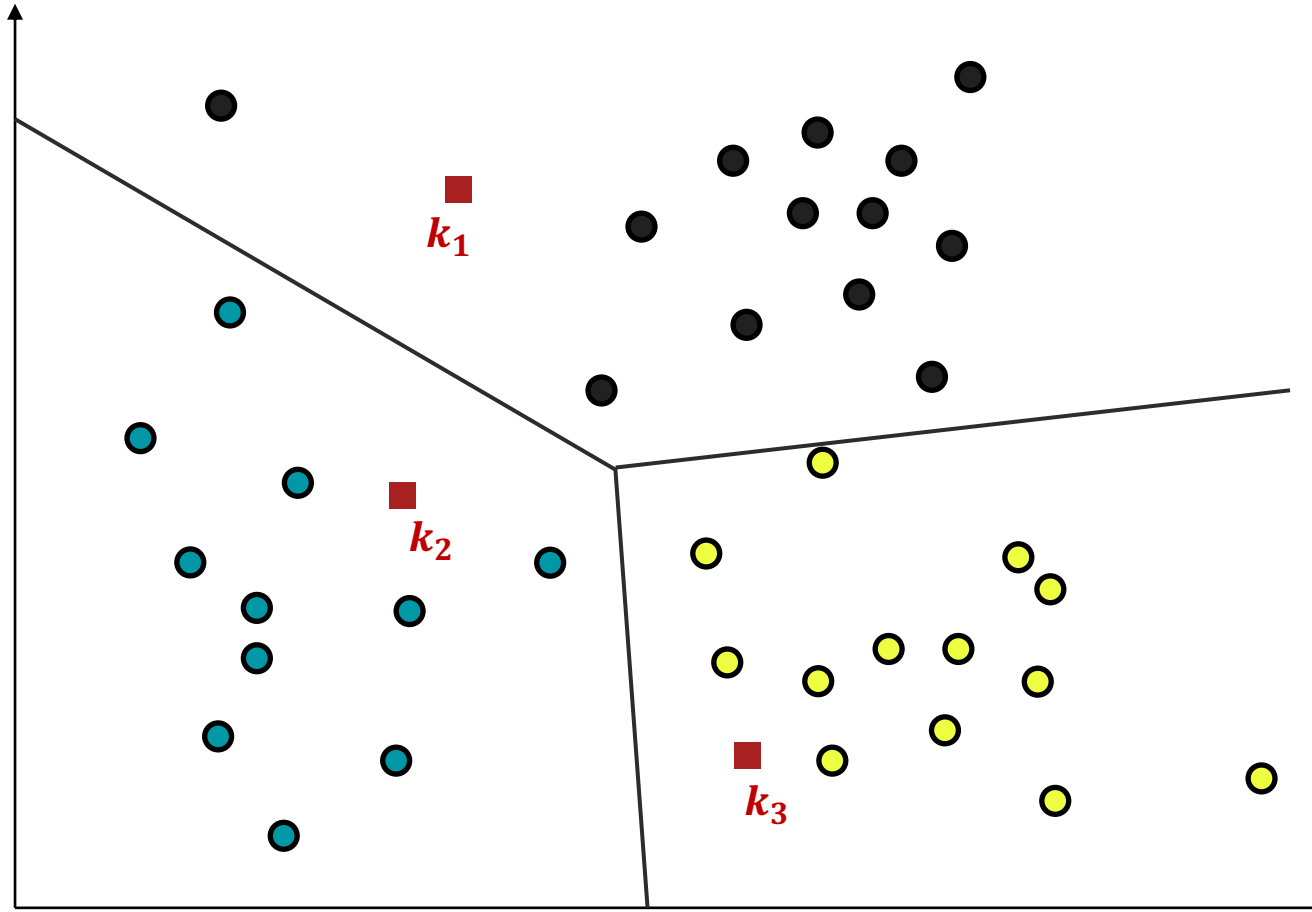
1) 초기 K 개의 중심(Centroid)이 random하게 선택됨



K-means 클러스터분석 (Cluster Analysis)

예 : $k = 3$ 으로 설정

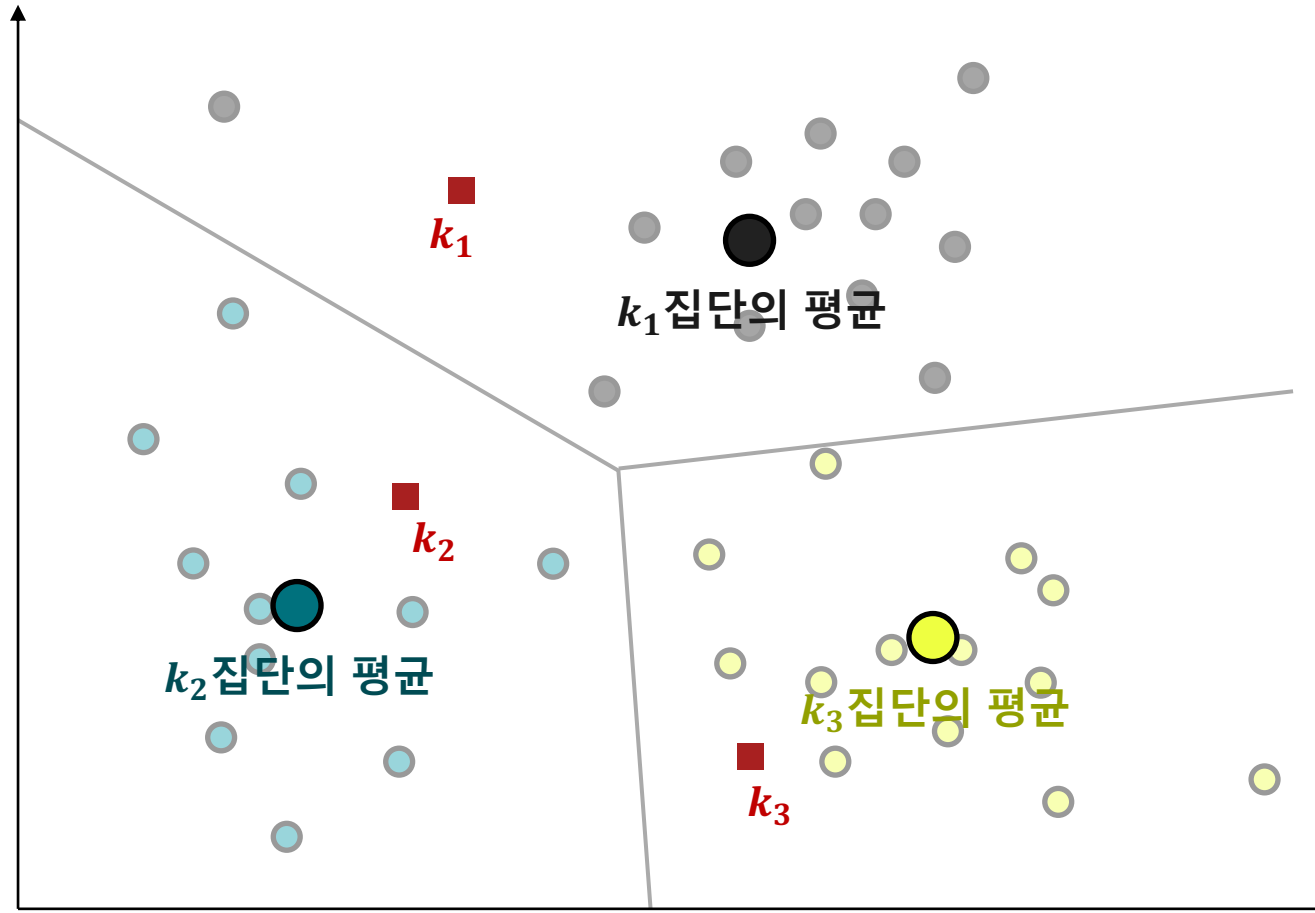
2) 각 개체(Observation)는 자신에게 가장 가까운 초기 중심(centroid) k 에 할당됨



K-means 클러스터분석 (Cluster Analysis)

예 : $k = 3$ 으로 설정

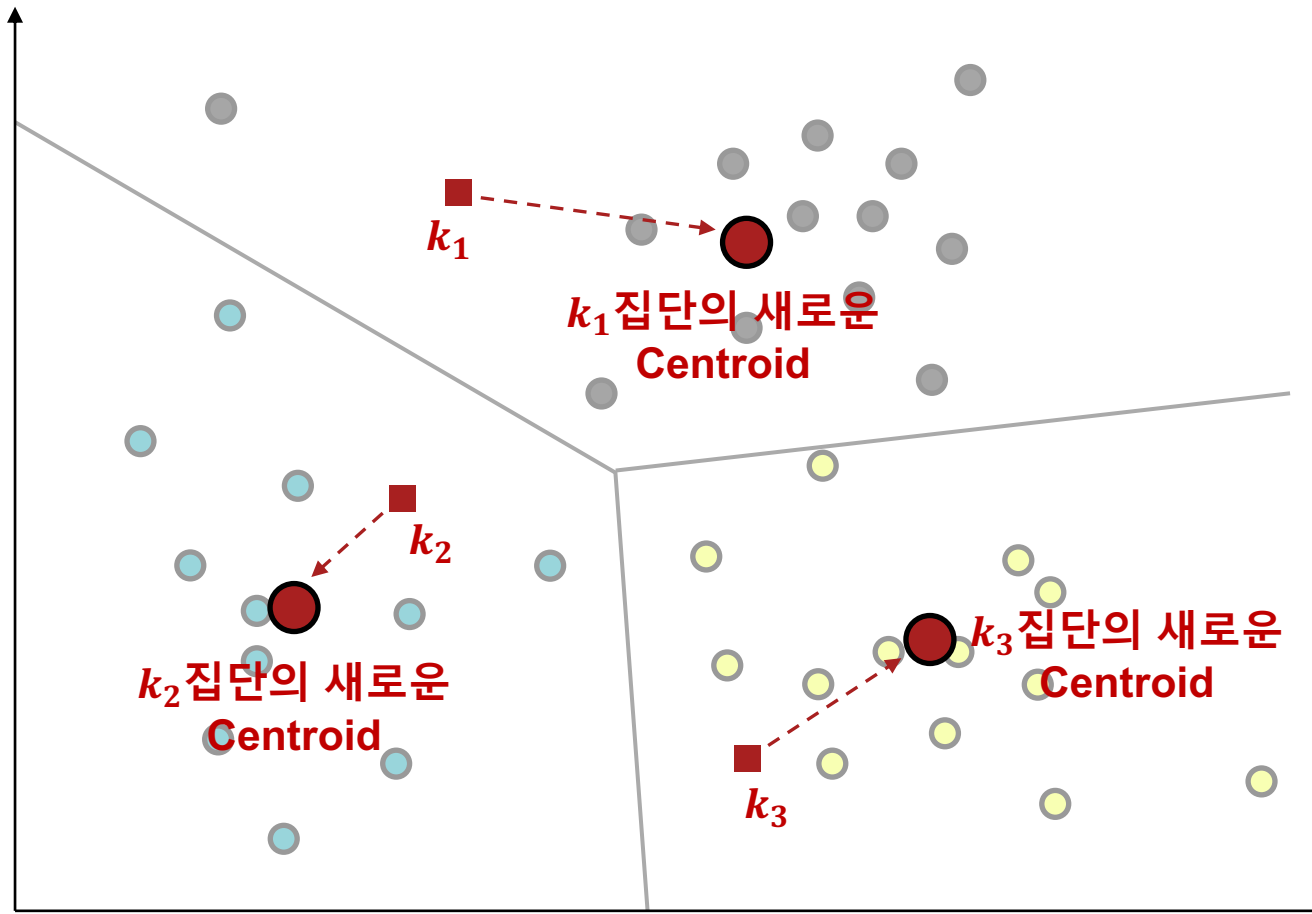
3) 같은 Centroid에 할당된 개체들의 평균(Mean)을 구한다



K-means 클러스터분석 (Cluster Analysis)

예 : $k = 3$ 으로 설정

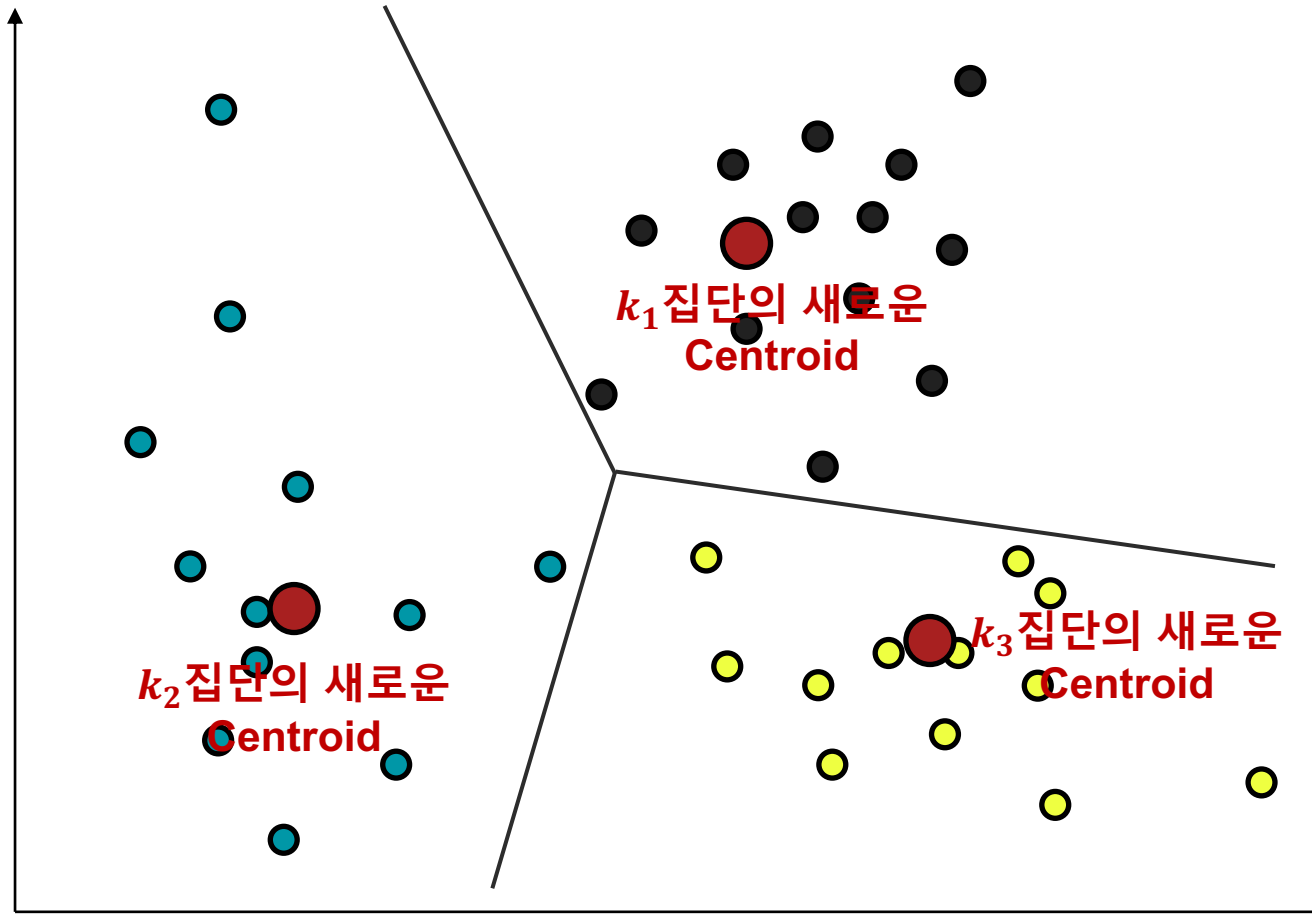
4) 초기 설정된 Centroid는 없어지고, 새롭게 계산한 각 집단의 평균으로 Centroid가 이동함



K-means 클러스터분석 (Cluster Analysis)

예 : $k = 3$ 으로 설정

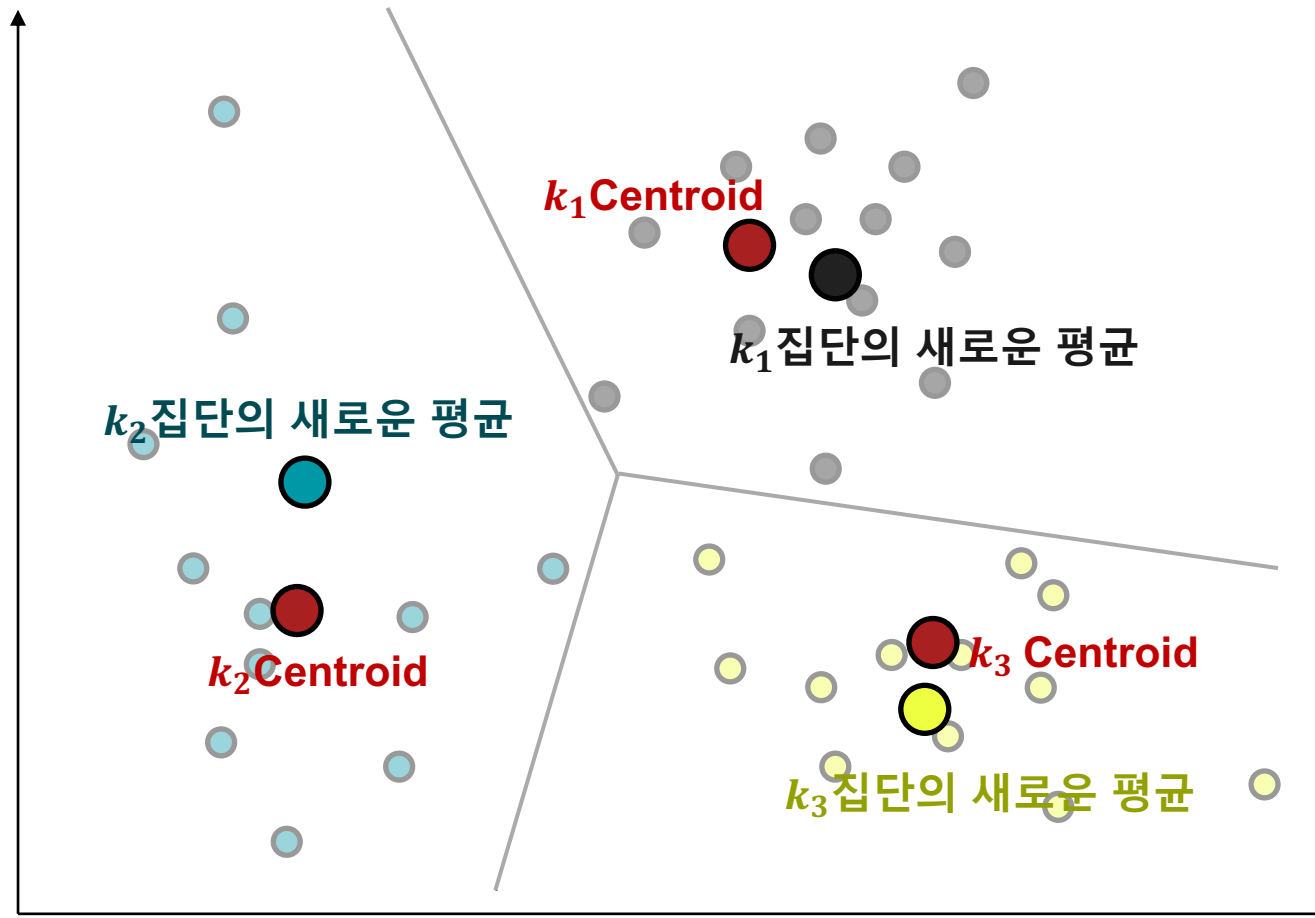
5) 새롭게 정의된 Centroid를 기준으로 각 개체는 가장 가까이에 있는 Centroid에 다시 속하도록 함



K-means 클러스터분석 (Cluster Analysis)

예 : $k = 3$ 으로 설정

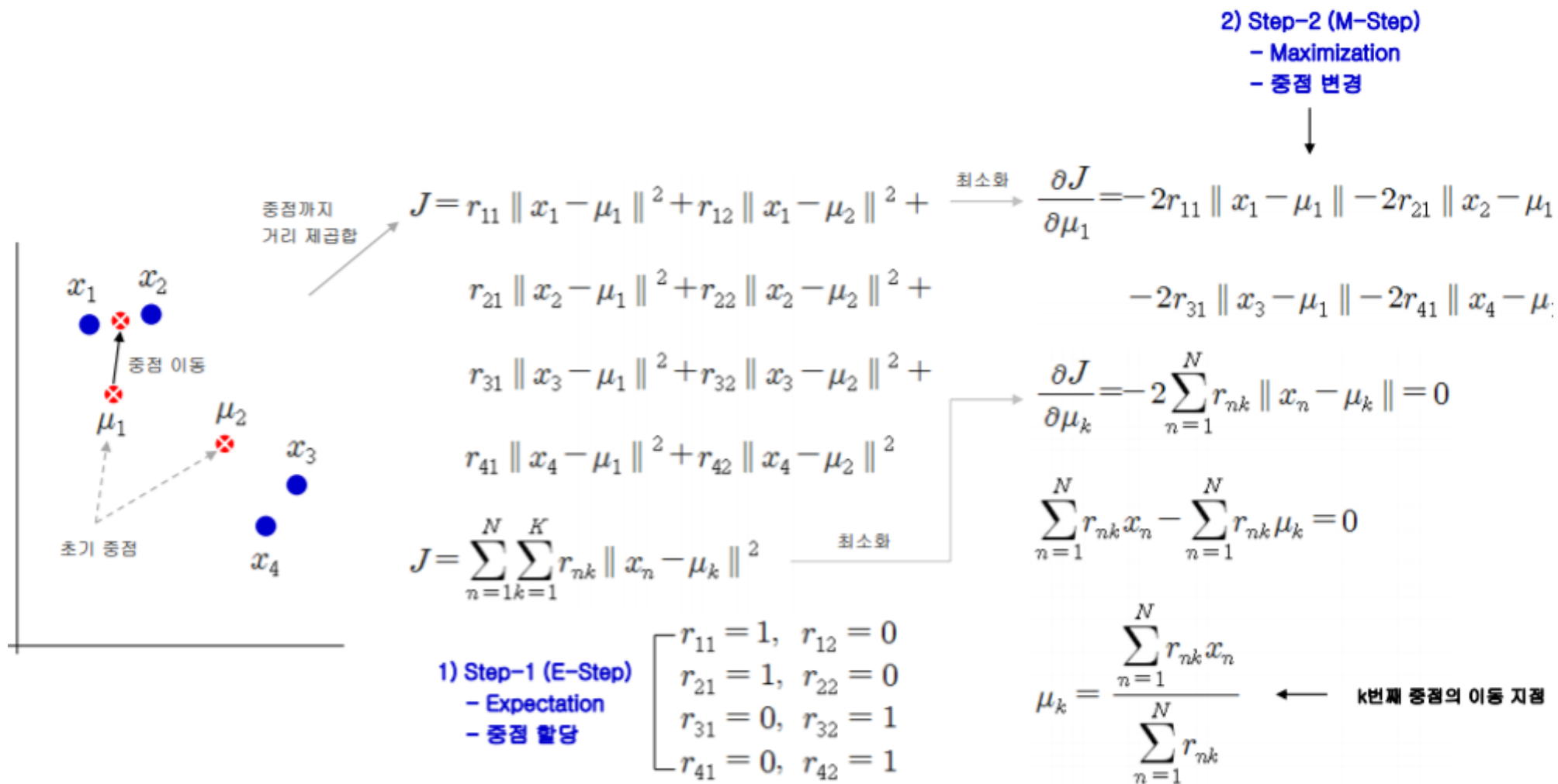
6) 1)부터 5)까지의 과정을 반복적으로 시행하되, 더 이상 집단이 바뀌지 않을 때, 반복을 멈춤



K-Means Clustering 알고리즘 – EM 알고리즘

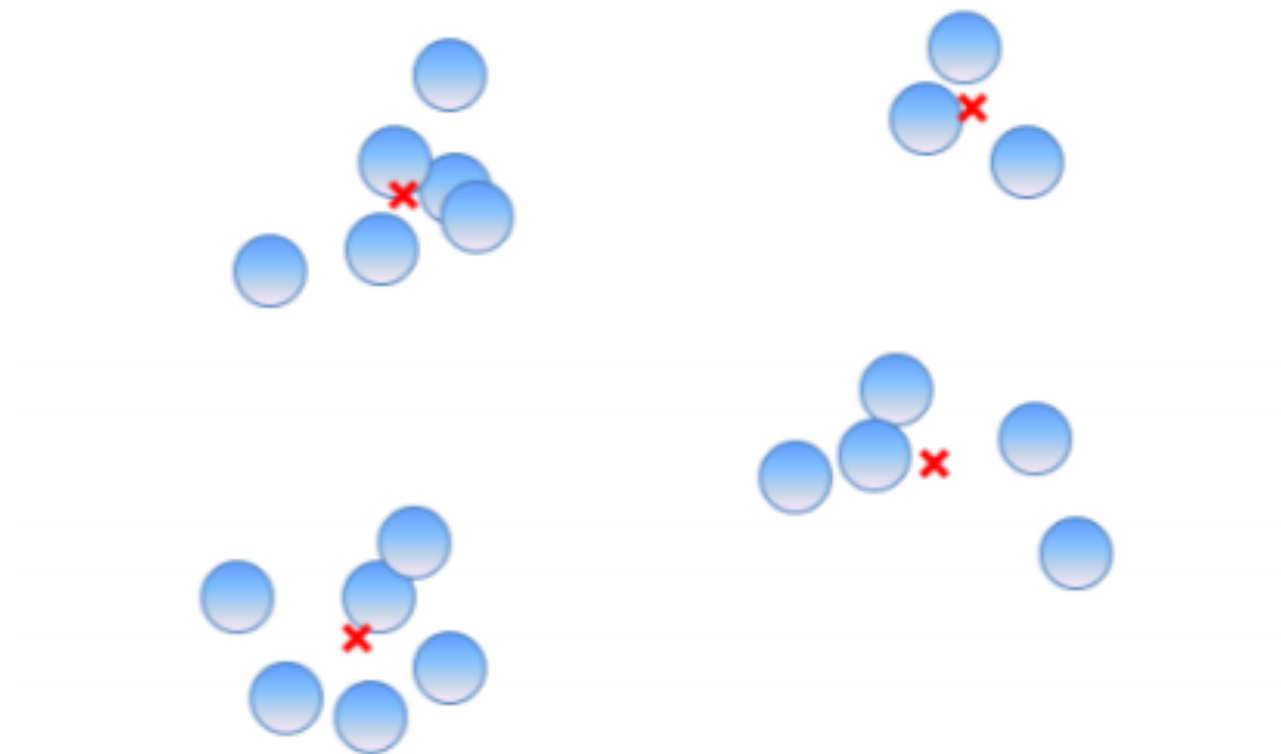
- **EM 알고리즘** : 중점을 할당한 후 할당된 중점까지의 거리의 합을 최소화하는 알고리즘
- 1) E-Step에서는 중점을 할당하고, 2) M-Step에서는 할당된 중점까지의 거리를 최소화함.
- r (Assignment)이 $\{0, 1\}$ 이므로 Hard Clustering 형태임. r 이 연속형 (확률)이면 GMM 모형으로 확장이 필요함.
- 파라미터 벡터가 두 개 (r 과 u) 이므로 MLE로 최적화하기 어려움. -> EM 알고리즘이 필요함.

K-Means Clustering 알고리즘 - EM 알고리즘



K-Means Clustering 알고리즘 구현 예시

- 20개의 훈련 데이터를 4개의 클러스터로 분류함.
- 최초 4개의 랜덤 좌표를 생성하여 초기 클러스터의 중심으로 사용하고, 한 스텝씩 진행하면서 중심을 업데이트함.



K-Means Clustering 알고리즘 구현 예시 : 주가의 캔들스틱 차트 군집화

- 캔들 스틱의 형태를 4개로 분류하여 4가지 유형으로 군집화 함. 예 : 군집-(1) = 큰 양봉 (장대 양봉), 군집-(2) = 큰 음봉 등
- K-Means 알고리즘으로 4개의 초기 중심 좌표를 설정하고, 각 캔들에서 중심 까지의 거리가 최소가 되도록 중심을 조절함.

주가 캔들스틱 군집화 (K-means Clustering)

클러스터 중심 좌표				
	시가	고가	저가	종가
1	0.907	0.320	0.048	0.401
2	0.420	0.060	0.237	0.857
3	0.229	0.356	0.434	0.264
4	0.265	0.724	0.150	0.898

초기화

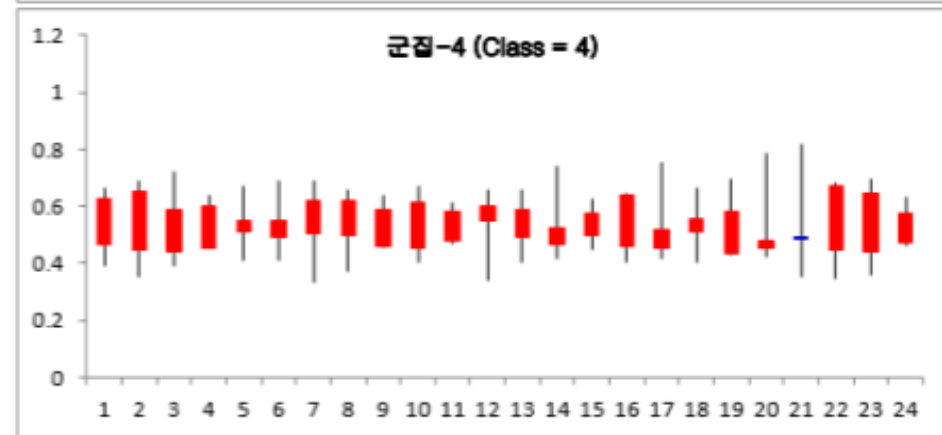
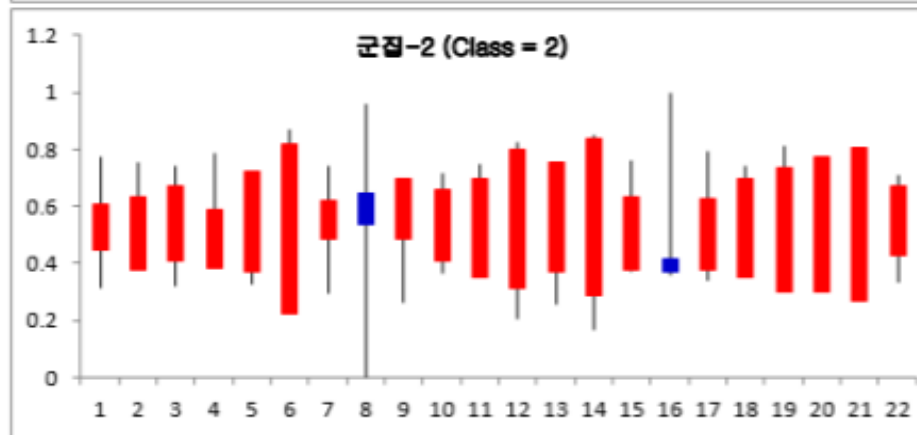
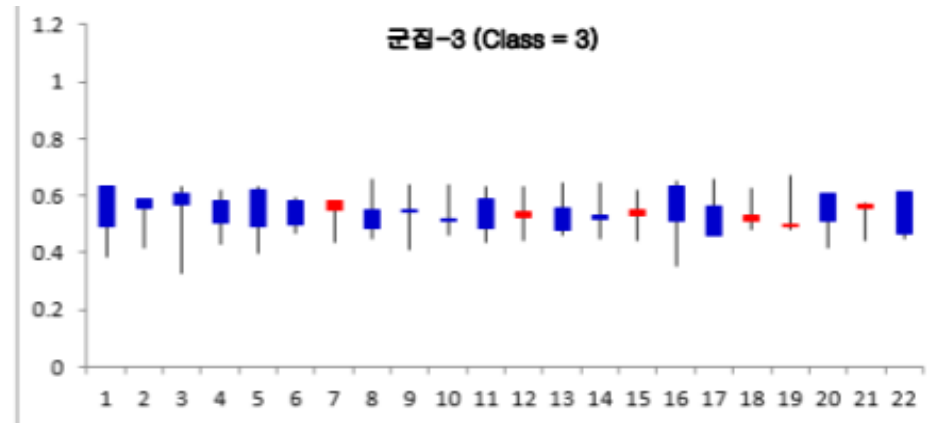
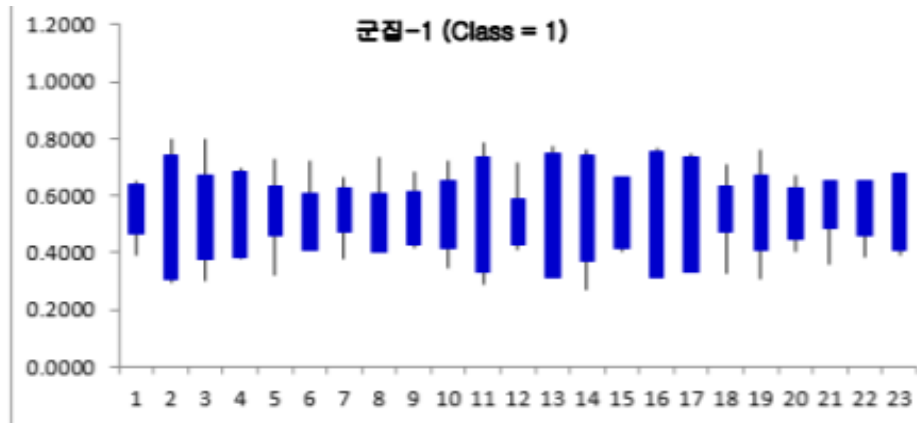
스텝

Error 19.5752



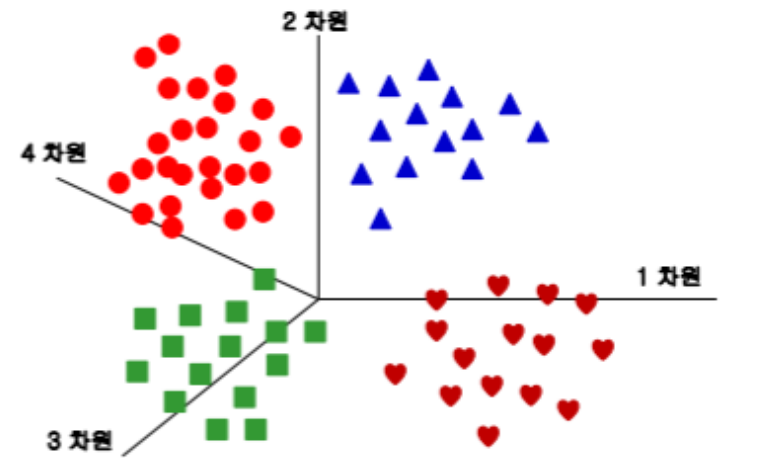
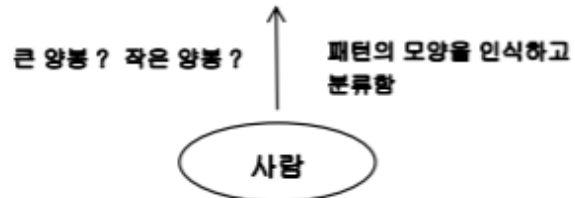
K-Means Clustering 알고리즘 구현 예시 : 주가의 캔들스틱 차트 군집화

- K-Means 알고리즘은 아래와 같이 4개로 구분하였음.
- 큰 음봉과 큰 양봉은 군집-(1) 과 군집-(2) 로 할당하고, 작은 음봉 과 작은 양봉은 군집-(3) 과 군집-(4)로 할당하였음.



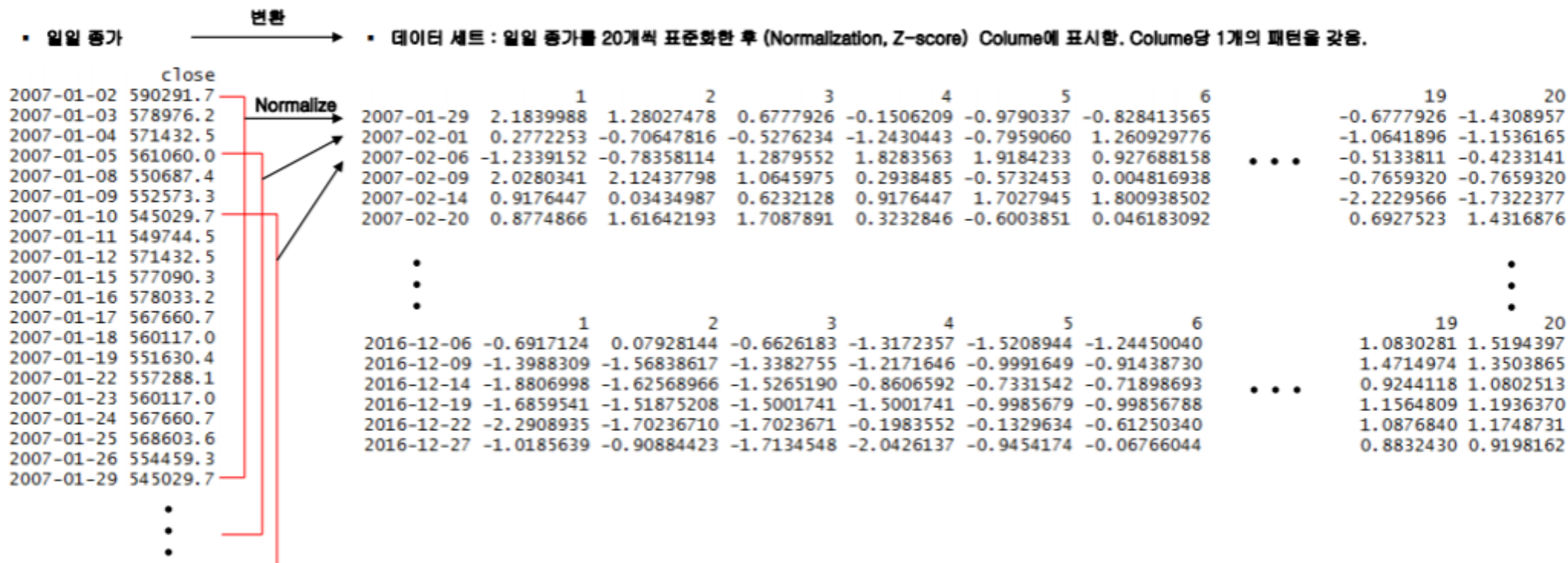
K-Means Clustering 알고리즘 구현 예시 : 주가의 캔들스틱 차트 군집화 – 사람과 K-Means 알고리즘의 인식의 차이

- 사람은 캔들 스틱의 모양을 인식하고, 모양 별로 분류하려고 할 것임.
- K-Means Clustering (기계) 알고리즘은 캔들의 정보를 4차원 공간 (시가, 고가, 저가, 종가)에 뿌려 놓고, 공간 상의 위치를 인식하여, 각각의 중심까지의 거리가 최소가 되도록 분류함. (우리는 4차원 이상의 공간을 인식하지 못함)
- 4차원 공간에서 각 점들은 유사도가 높은 것들끼리 모여 있음. (유사도 척도는 유클리디언 거리를 이용할 수도 있고, 코사인 거리를 이용할 수도 있음.)



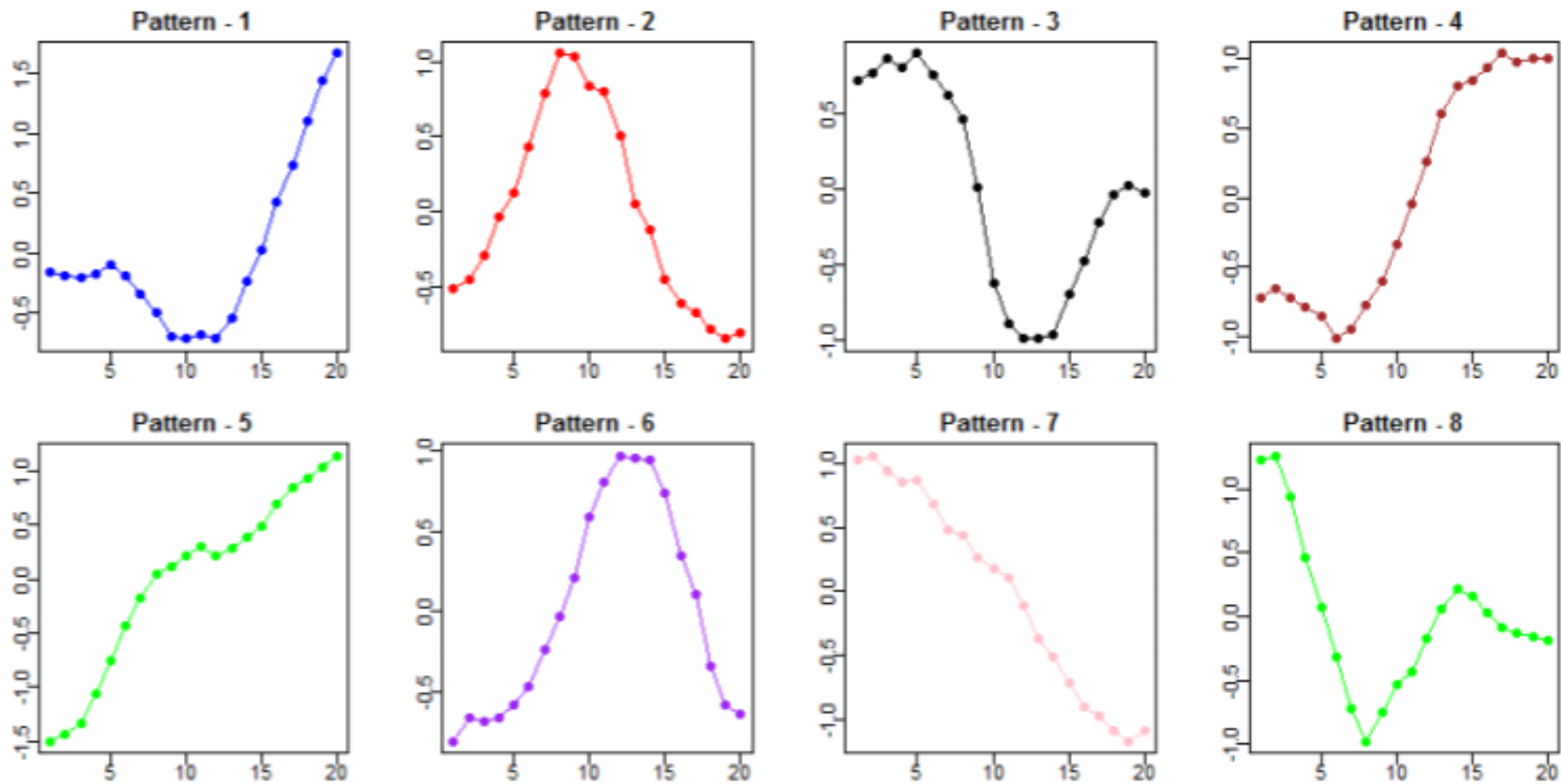
K-Means Clustering 알고리즘 구현 예시 : 주가의 패턴 분석 실습 (학습 데이터 세트 생성)

- Yahoo 사이트에서 삼성전자 주가를 읽어와서 종가를 기준으로 패턴을 8개의 그룹 (클러스터)로 분류함.
- 아래와 같이 종가 20개를 한 개의 패턴으로 (약 1개월 패턴) 정의하고, 각 패턴을 8개의 그룹으로 분류함. 각 그룹에는 유사한 패턴끼리 모여 있음.
- 종가를 3일씩 건너뛰면서 20개씩 벡터로 모아서 각 벡터마다 한 개의 패턴이 되도록 함.



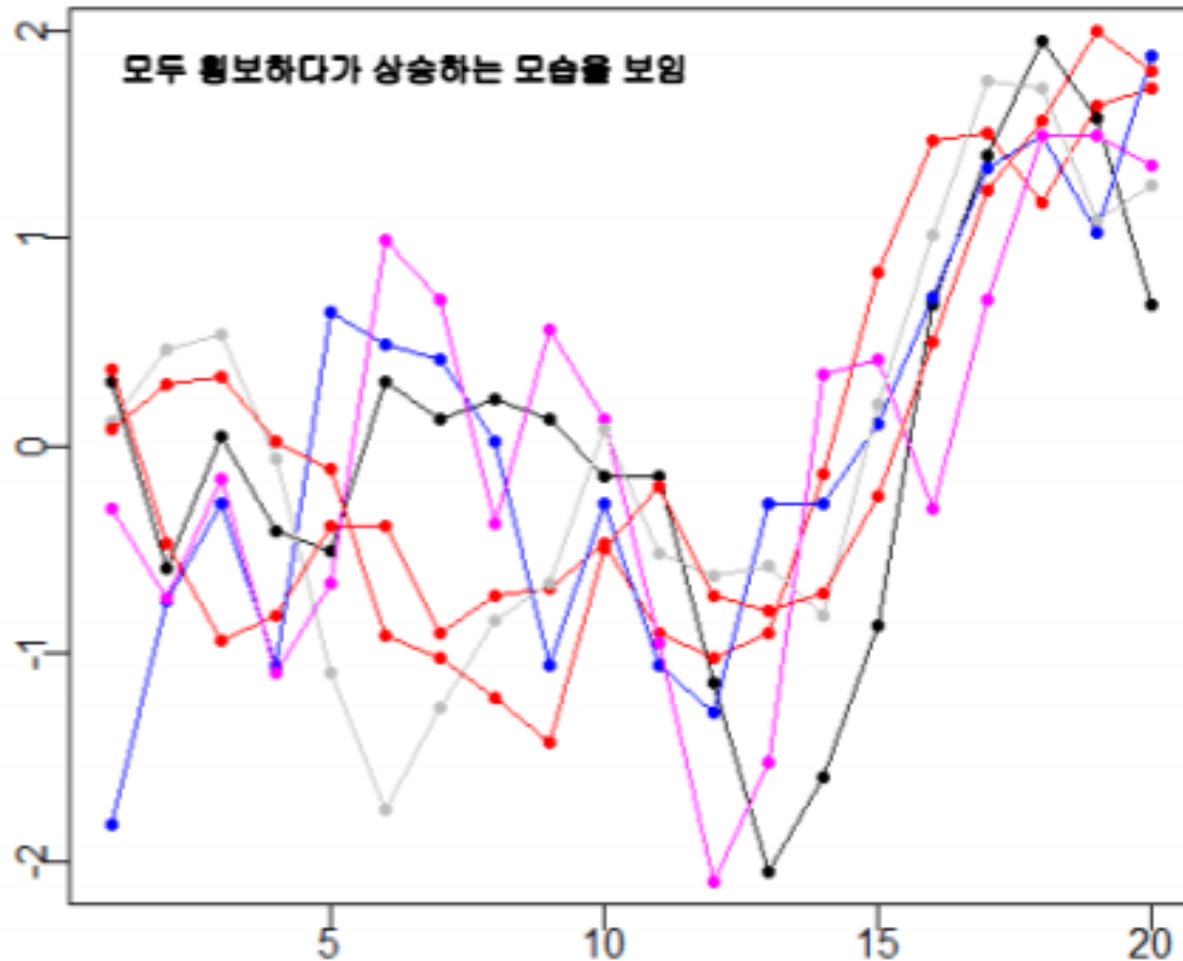
K-Means Clustering 알고리즘 구현 예시 : 주가의 패턴 분석 (20일 주가를 8개의 유사 패턴으로 분류함)

- K-Means Clustering 알고리즘 구현 예시 : 주가의 패턴 분석 (20일 주가를 8개의 유사 패턴으로 분류함)



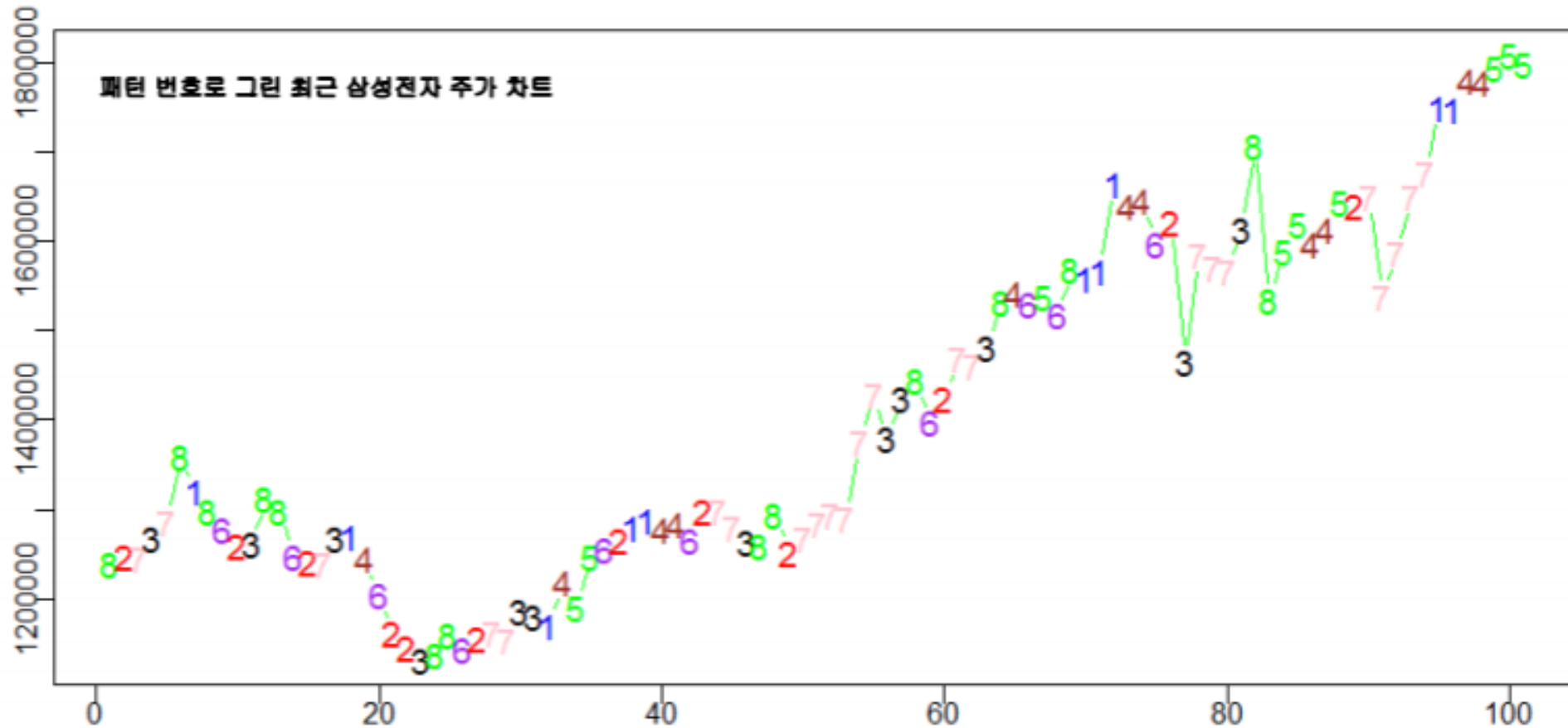
K-Means Clustering 알고리즘 구현 예시 : 주가의 패턴 분석

- 1 번 패턴 그룹에 속한 몇 개의 패턴을 확인함. 오차는 있지만 모두 횡보하다가 상승하는 경향을 보임.



K-Means Clustering 알고리즘 구현 예시 : 주가의 패턴 분석

- 최근 100 기간의 종가 차트 위에 패턴 번호를 표시함. 최근에는 주로 1, 4, 5 번 패턴이 발생했음.



K-Means Clustering 알고리즘 구현 예시 : 주가의 패턴 분석 (최근 패턴 테이블 작성)

- 최근 100 기간 동안 발생한 패턴의 비율을 확인함. 4, 5 번 패턴이 주로 발생하였음 (Multinomial 분포). 패턴 번호는 실행할 때마다 달라짐.
- 군집 분석 (Unsupervised)의 결과물은 Supervised Learning의 입력이 될 수도 있고, Bayesian Inference 같은 방법으로 Next likelihood pattern 분포 등을 추론해 볼 수도 있음. -> 패턴 분석의 과제임.

