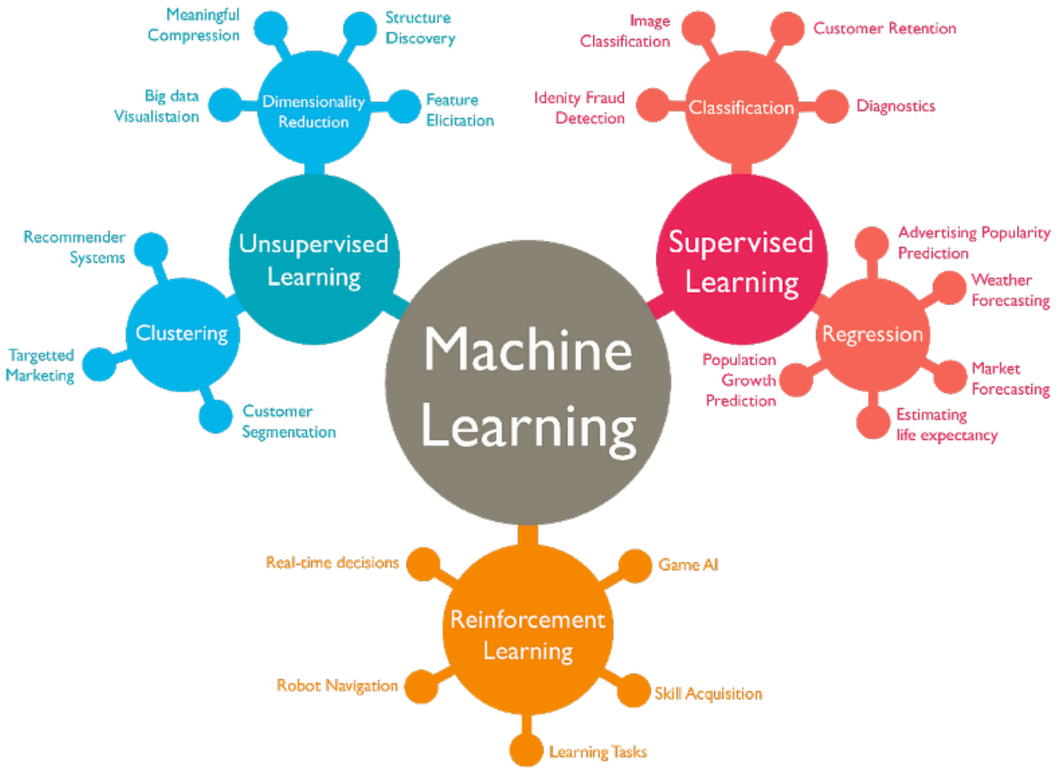
KNN and K-means

Jin Hyun Kim



https://wendys.tistory.com/169

In this class

- k-Nearest Neighbor - A Lazy Learning
- k-means - A Unsupervised Learning

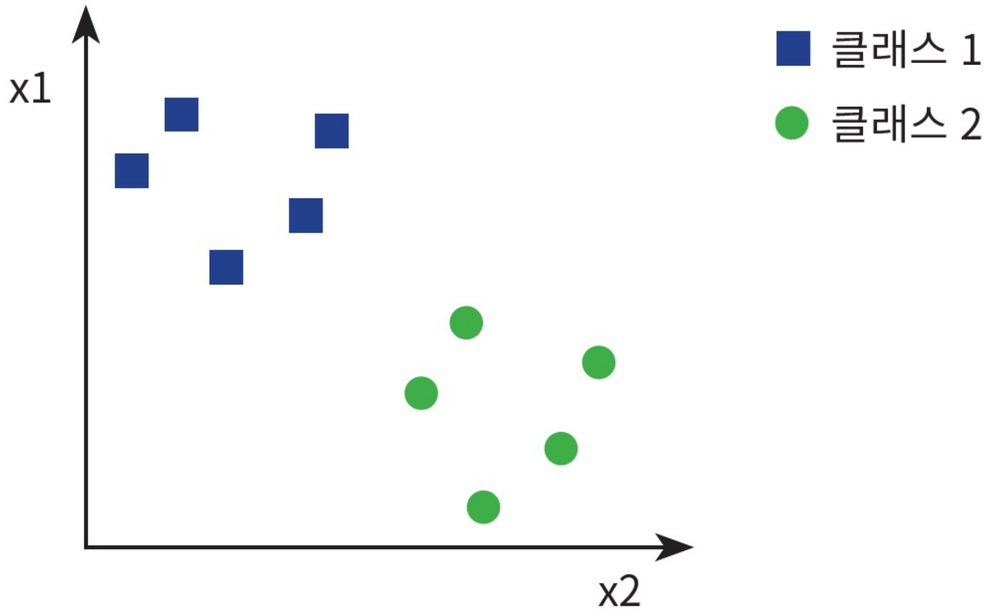
References

- 인공지능 - 튜링테스트에서 딥러닝까지
- 인공지능 - 파이선으로 배우는 머신러닝과 딥러닝

k-Nearest Neighbor (kNN) 알고리즘

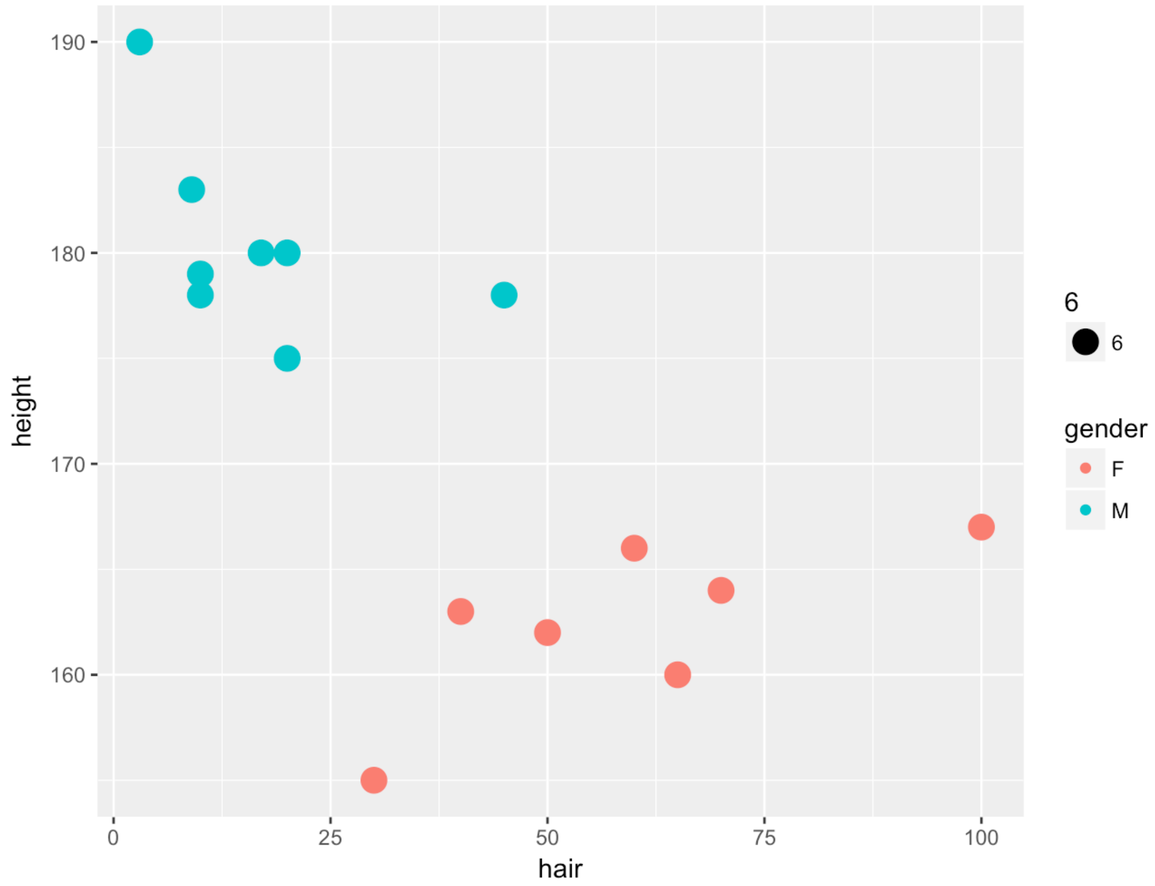
Lazy Learning

- 분류 알고리즘
- 특징
  - 학습단계에서는 실질적인 학습이 일어나지 않고 데이터만 저장
    - 학습데이터가 크면 메모리 문제
    - 게으른 학습(lazy learning)
- 새로운 데이터가 주어지면 저장된 데이터를 이용하여 학습
  - 시간이 많이 걸릴 수 있음



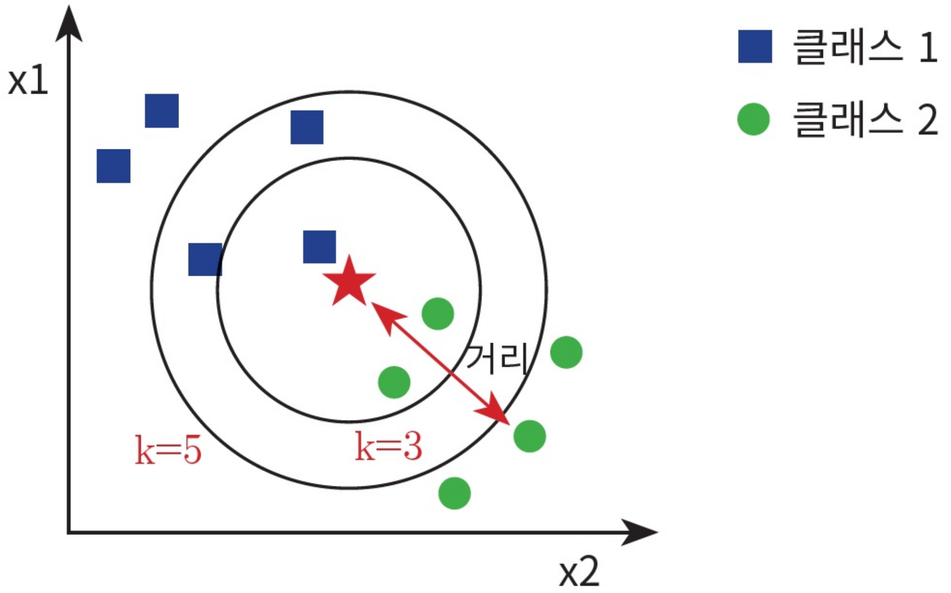
kNN 알고리즘

적용이 예

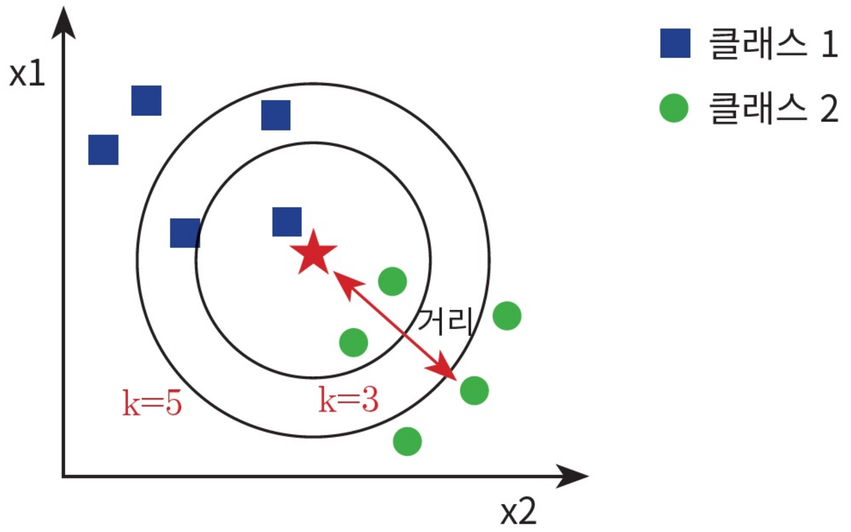


kNN 알고리즘

- k-Nearest Neighbor(kNN)
  - (입력, 결과)가 있는 데이터들이 주어진 상황에서, 새로운 입력에 대한 결과를 추정할 때
    - 결과를 아는 최근접한 k개의 데이터에 대한 결과정보를 이용하는 방법
  - 질의(query)와 데이터간의 거리 계산
  - 효율적으로 근접이웃 탐색이 핵심
  - 근접 이웃 k개로 부터 결과를 추정



kNN 알고리즘



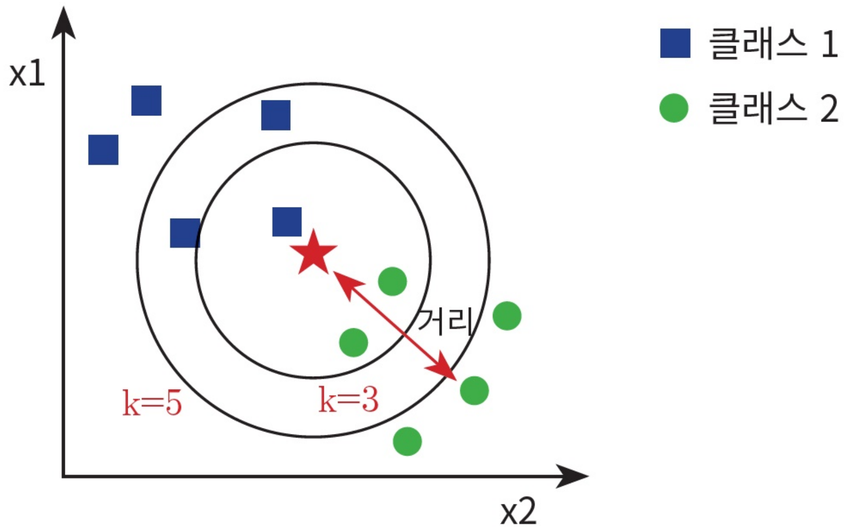
- 데이터간의 거리 계산
  - 수치 데이터의 경우
    - 유클리디언 거리(Euclidian distance)

$$X = (x_1, x_2, \ldots, x_n), Y = (y_1, y_2, \ldots, y_n)$$

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2 +}$$

    - 응용분야의 특성에 맞춰 개발
- 범주형 데이터가 포함된 경우
  - 응용분야의 특성에 맞춰 개발

kNN 알고리즘

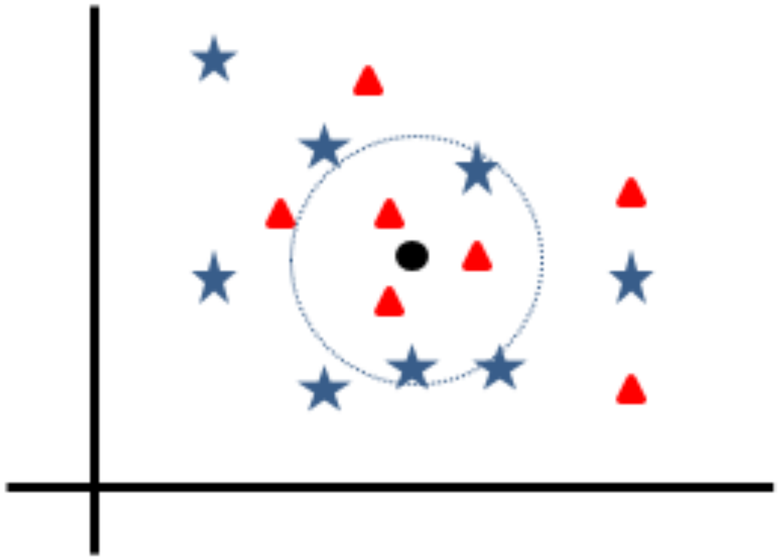


- 효율적인 근접 이웃 탐색
- 데이터의 개수가 많아지면 계산시간 증가 문제
- 색인(indexing) 자료구조 사용
  - R-트리, k-d 트리 등

kNN 알고리즘

최근접 k개로 부터 결과를 추정하는 방법

- 분류 (Classification)
  - 출력이 범주(Class)형 값
  - 다수결 투표(majority voting) : 개수가 많은 범주 선택



kNN 알고리즘

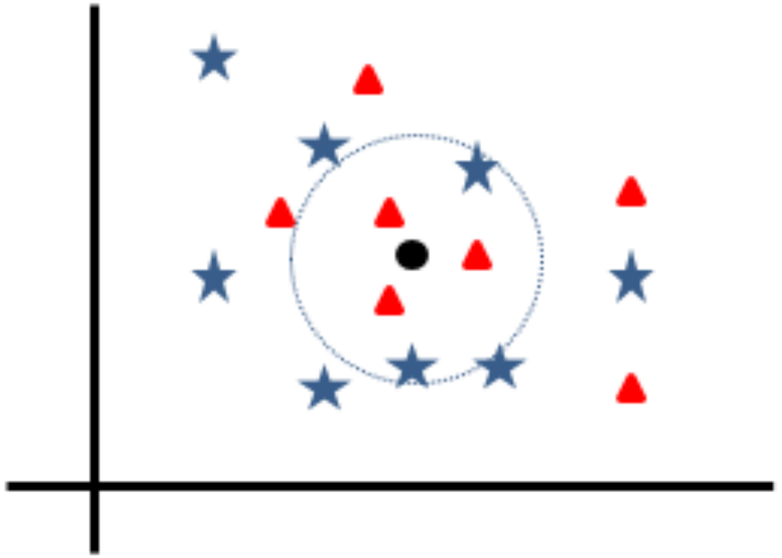
최근접 k개로 부터 결과를 추정하는 방법

- 회귀분석
- 출력이 수치형 값
  - k-근접이웃 KNN = $\{(X_1, y_1),((X_2, y_2),\ldots,((X_k, y_k)\}$
  - 평균 : 최근접 k개의 평균값

$$y\ means = \frac{1}{k}\sum_{i=1}^{k} y_i$$

  - 가중합(weighted sum) : 거리에 반비례하는 가중치 사용

Weighted sum: $y = \sum_{i=1}^{k} \frac{w_i y_i}{\sum_{k=1}^{k} w_i}$ where $w_i = \frac{1}{d(X, X_i)}$



kNN 알고리즘

Example

- 영화 x 등급을 예측하기 위해서 가장 가까운 이웃 3개의 영화를 찾았다고 하자.
  - 영화 : A / 등급: 5.0 / X까지의 거리: 3.2
  - 영화 : B / 등급: 6.8 / X까지의 거리: 11.5
  - 영화 : C / 등급: 9.0 / X까지의 거리: 1.1

- 평균 : 최근접 k개의 평균값:

$$\frac{5.0 + 6.8 + 9.0}{3} = 6.93$$

- 가중치의 합:

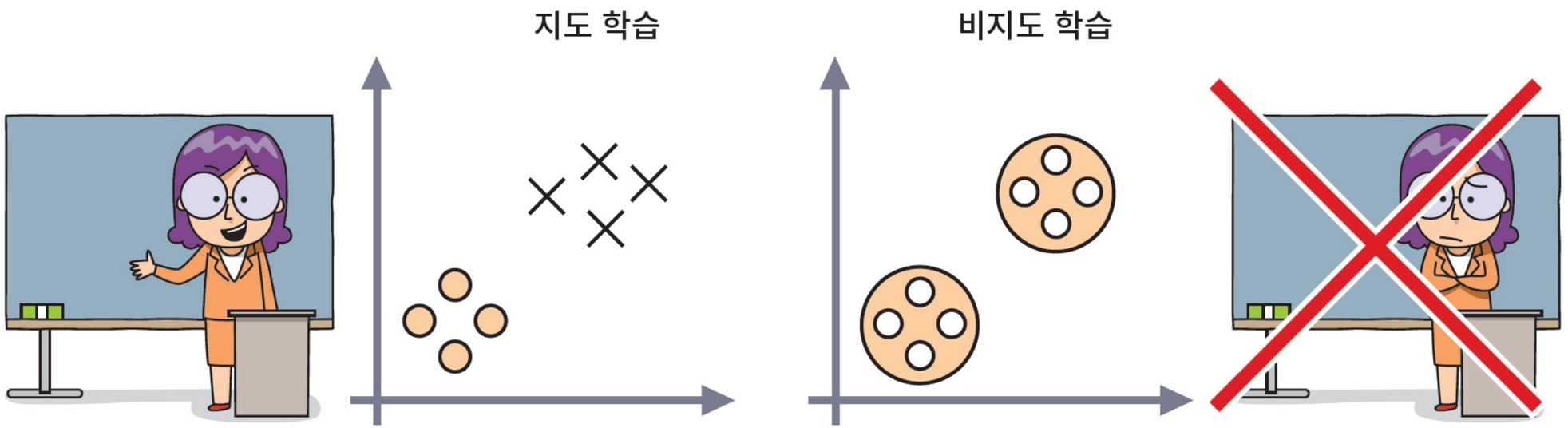
$$\frac{5.0/3.2 + 6.8/11.5 + 9.0/1.1}{1/3.2 + 1/11.5 + 1/1.1} = 7.9$$

kNN Lab

- 유방암 관련 데이터 분석
  - https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/
  - https://github.com/stedy/Machine-Learning-with-R-datasets

비지도 학습 - K-means

비지도학습 (Unsupervised Learning)



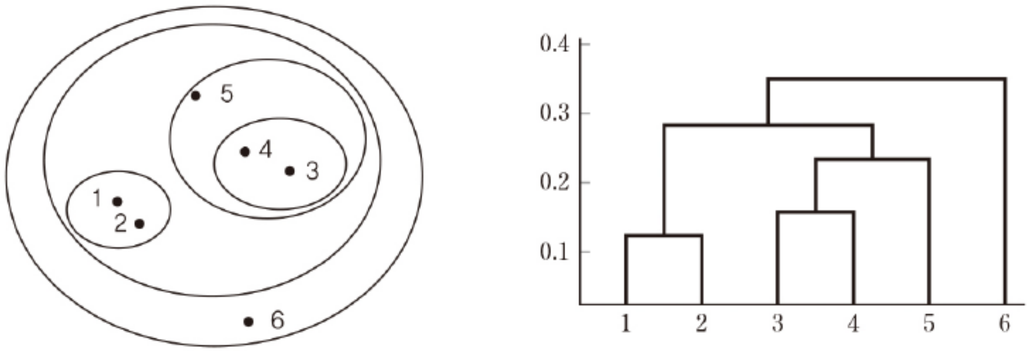
Clustering Algorithm

- 군집화(clustering) 알고리즘
  - 데이터를 유사한 것들끼리 모으는 것
  - 군집 간의 유사도(similarity)는 작게, 군집 내의 유사도는 크게
  - 종류
    - 계층적 군집화 (hierarchical clustering)
    - 분할 군집화 (partitioning clustering)

Clustering Algorithm

- 계층적 군집화 (hierarchical clustering)
  - 군집화의 결과가 군집들이 계층적인 구조를 갖도록 하는 것
  - 병합형(agglomerative) 계층적 군집화
    - 각 데이터가 하나의 군집을 구성하는 상태에서 시작하여, 가까이에 있는 군집들을 결합하는 과정을 반복하여 계층적인 군집 형성
  - 분리형(divisive) 계층적 군집화
    - 모든 데이터를 포함한 군집에서 시작하여 유사성을 바탕으로 군집을 분리하여 점차 계층적인 구조를 갖도록 구성

계층적 군집화와 덴드로그램(dendrogram)



Clustering Algorithm

- 분할 군집화 (partitioning clustering)
  - 계층적 구조를 만들지 않고 전체 데이터를 유사한 것들끼리 나누어서 묶는 것
  - 예. k-means 알고리즘

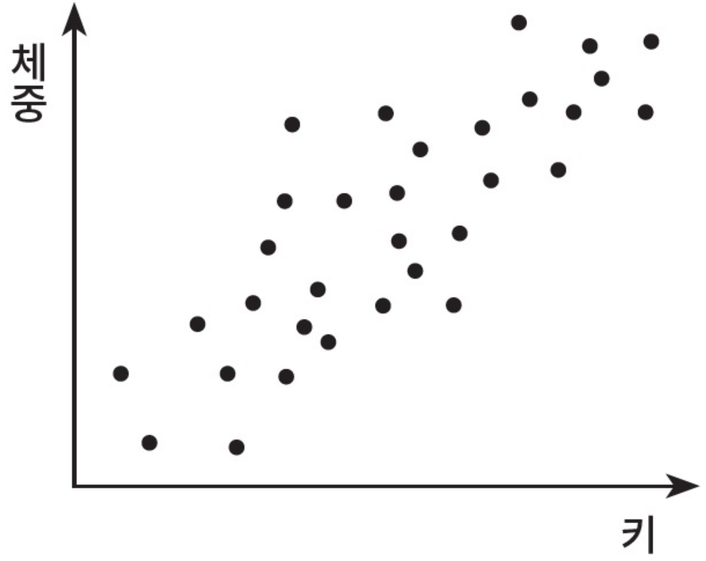
K-means Clustering

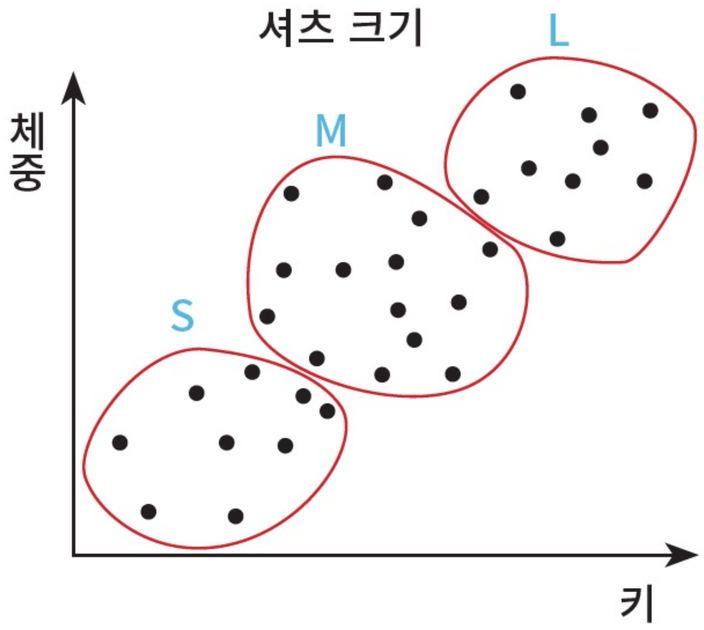
- 비지도 학습 중에서 가장 대표적
- K-means 알고리즘(K-means algorithm)은 주어진 n개의 관측값을 k개의 클러스터로 분할하는 알고리즘
  - 관측값들은 거리가 최소인 클러스터로 분류

K means Clustering

Example

- 수집된 키와 체중과의 관계를 나타낸 데이터를 통해, 셔츠의 사이즈를 클러스터링





K-means Algorithms

- i 번째 클러스터의 중심을 μ_i, 클러스터에 속하는 점의 집합 S_i을 라고 할 때, 전체 분산

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

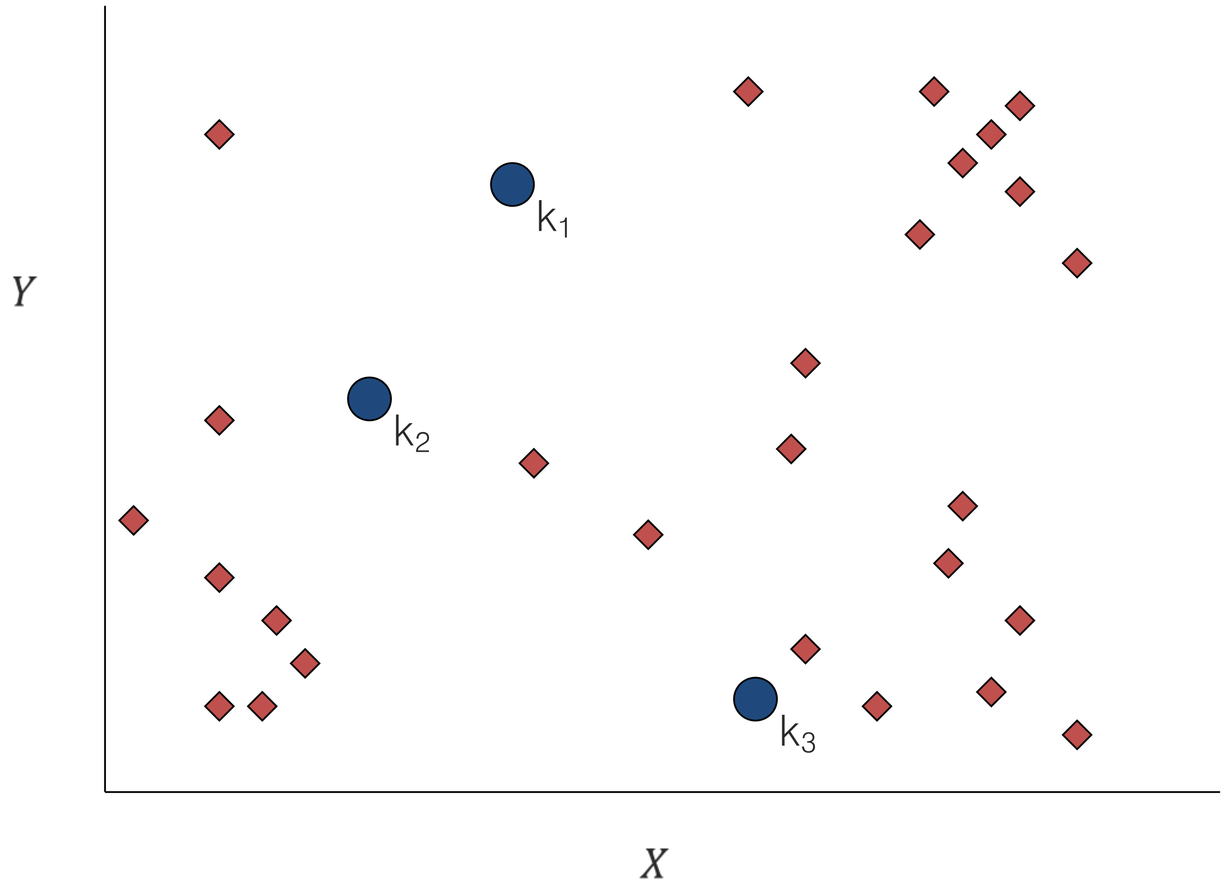
  - 분산값 V을 최소화하는 S_i를 찾는 것이 알고리즘의 목표
- 특성
  - 군집의 개수 k는 미리 지정
  - 초기 군집 위치에 민감

1. 우선 초기의 μ_i 를 임의로 설정
2. 다음 두 단계를 클러스터가 변하지 않을 때까지 반복
   A. 클러스터 설정: 각 점에 대해, 그 점에서 가장 가까운 클러스터를 찾아 배당한다.
   B. 클러스터 중심 재조정: mi를 각 클러스터에 있는 점들의 평균값으로 재설정해준다.

K-means Clustering

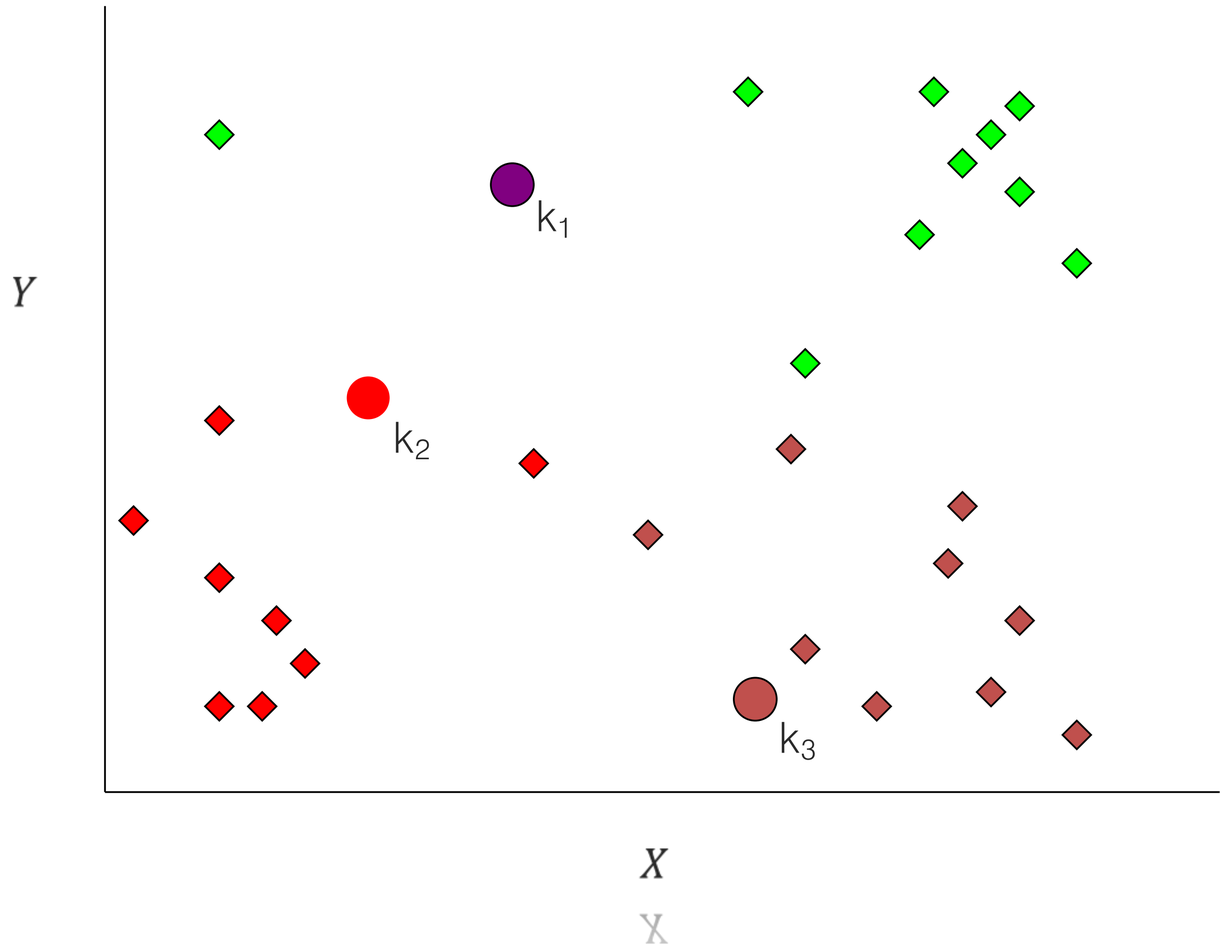
STEP 1

- 무작위로 군집 중심위치 3개를 선택



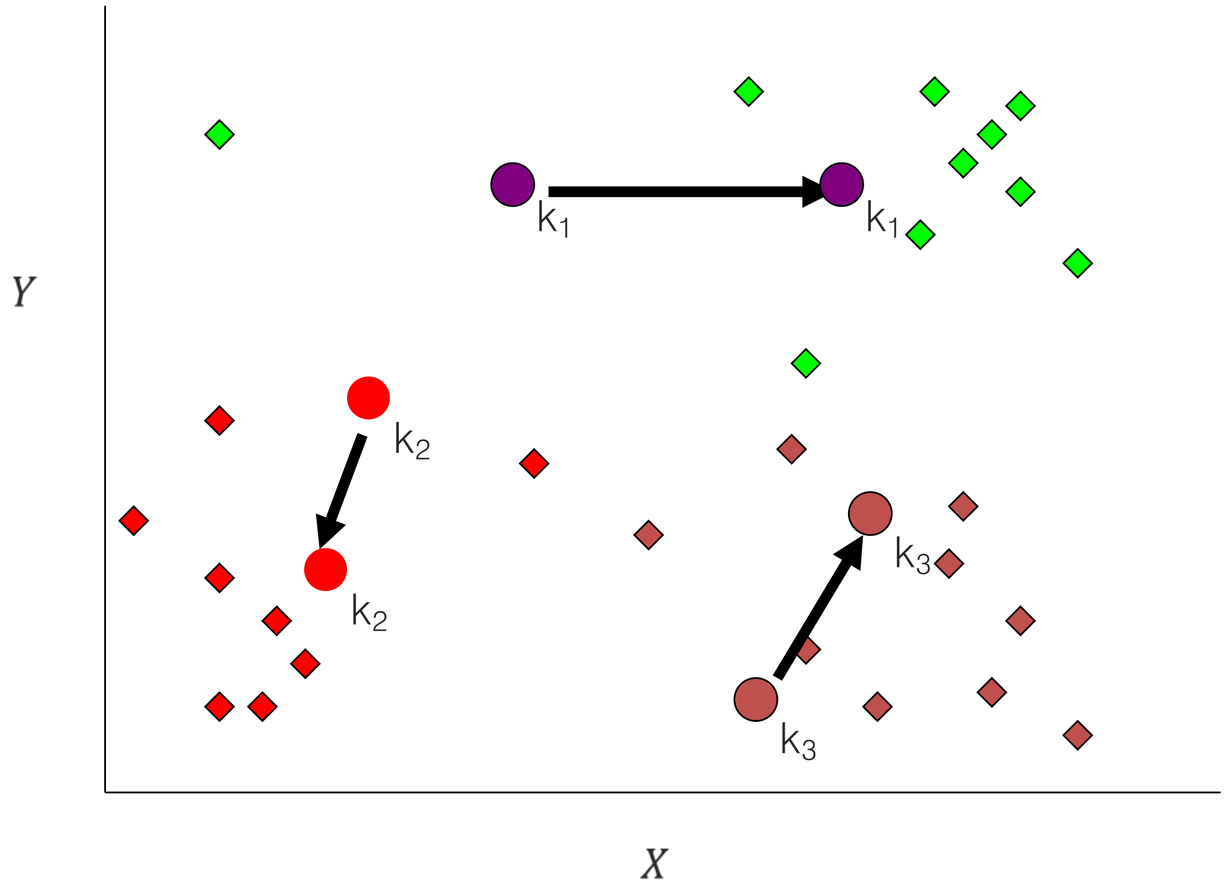
K-means Clustering
STEP 2

- 각 점을 최근접 군집 중심위치에 할당



K-means Clustering
STEP 3

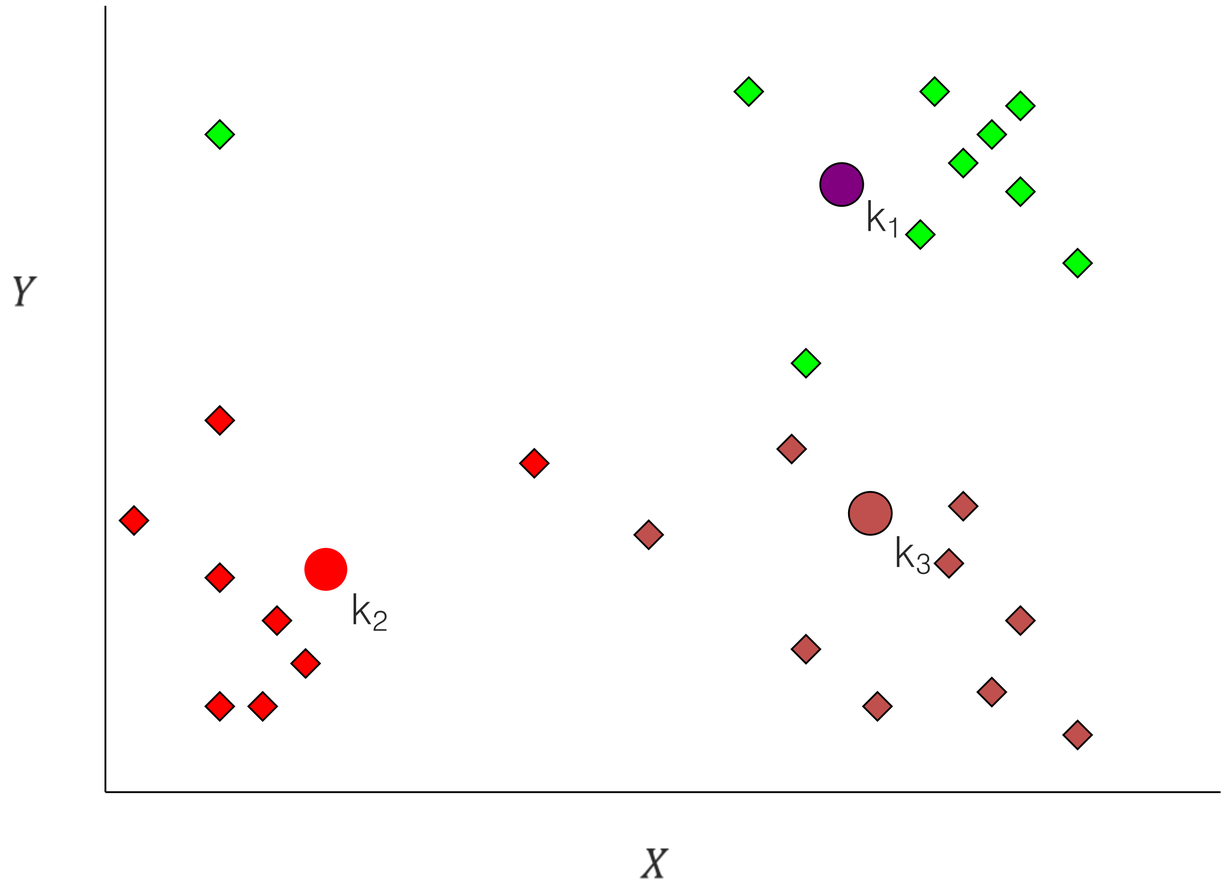
- 군집 중심 위치를 군집의 평균 위치로 이동



K-means Clustering

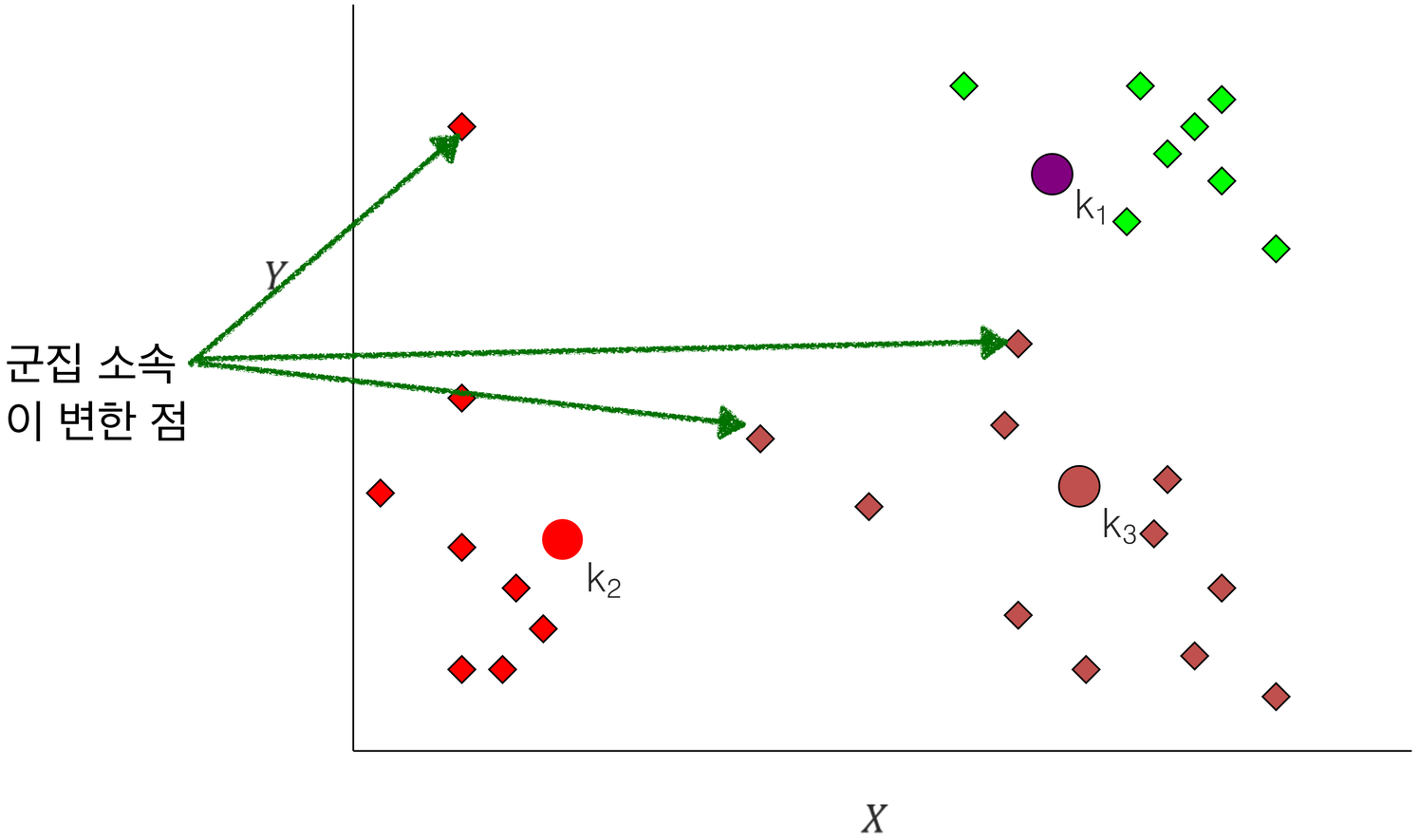
STEP 4

- 새로운 군집 중심을 기준으로 각 점의 소속을 재할당



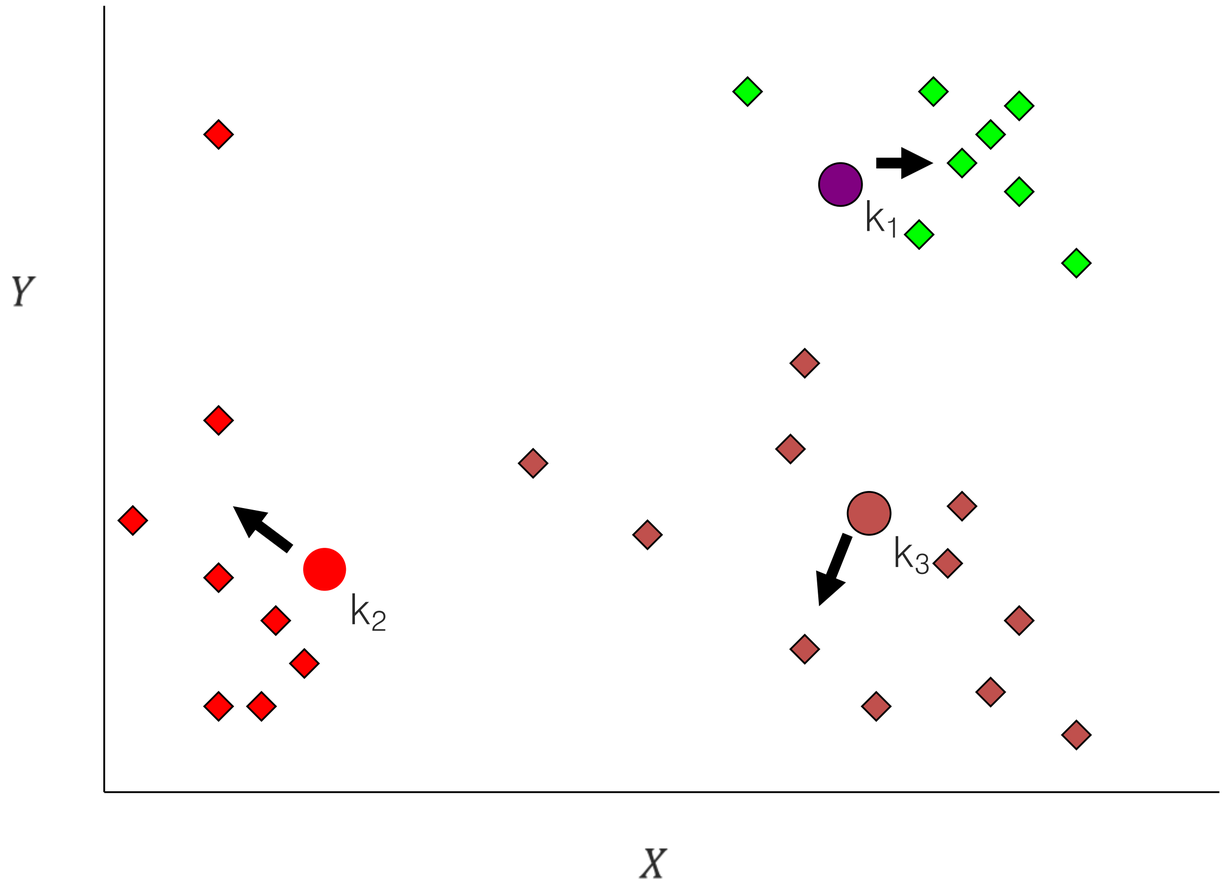
K-means Clustering

STEP 5



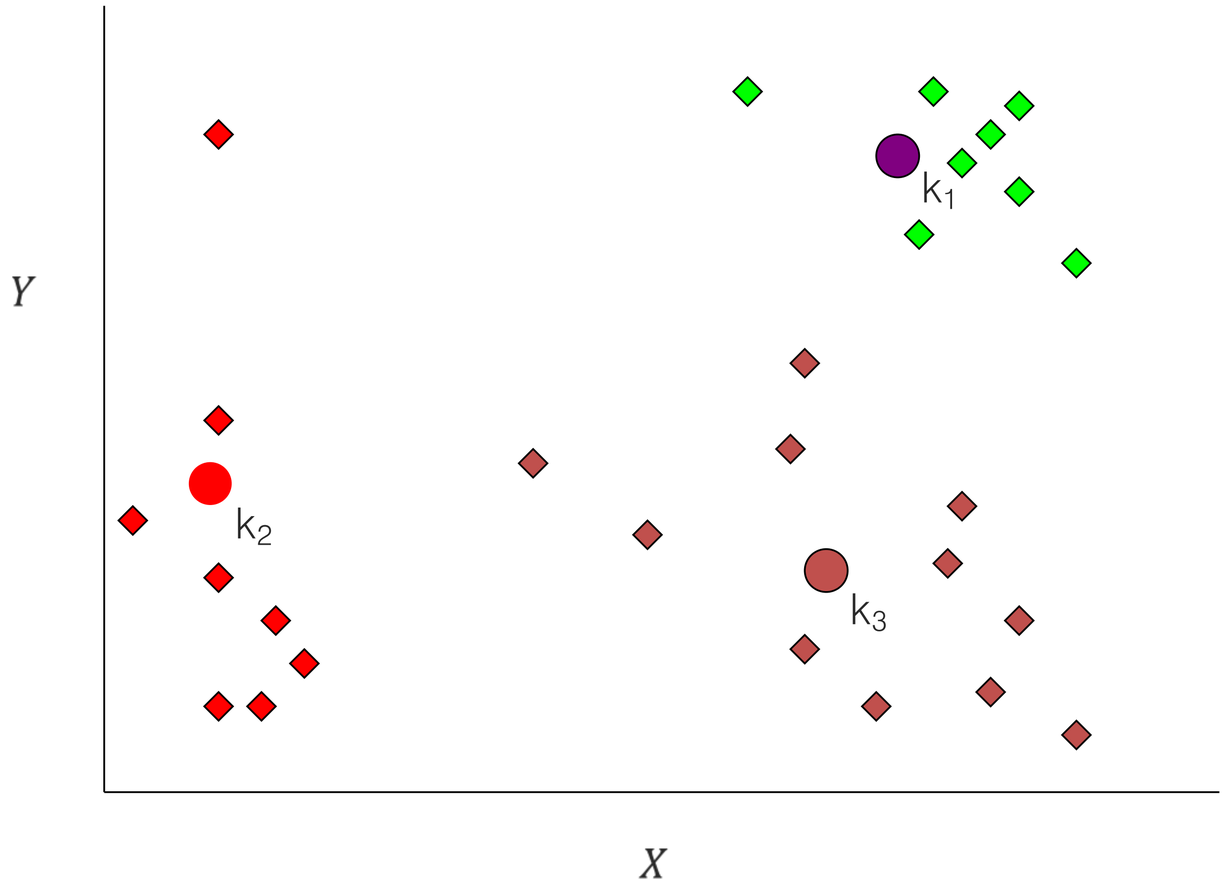
K-means Clustering
STEP 6

- 군집 평균 재계산



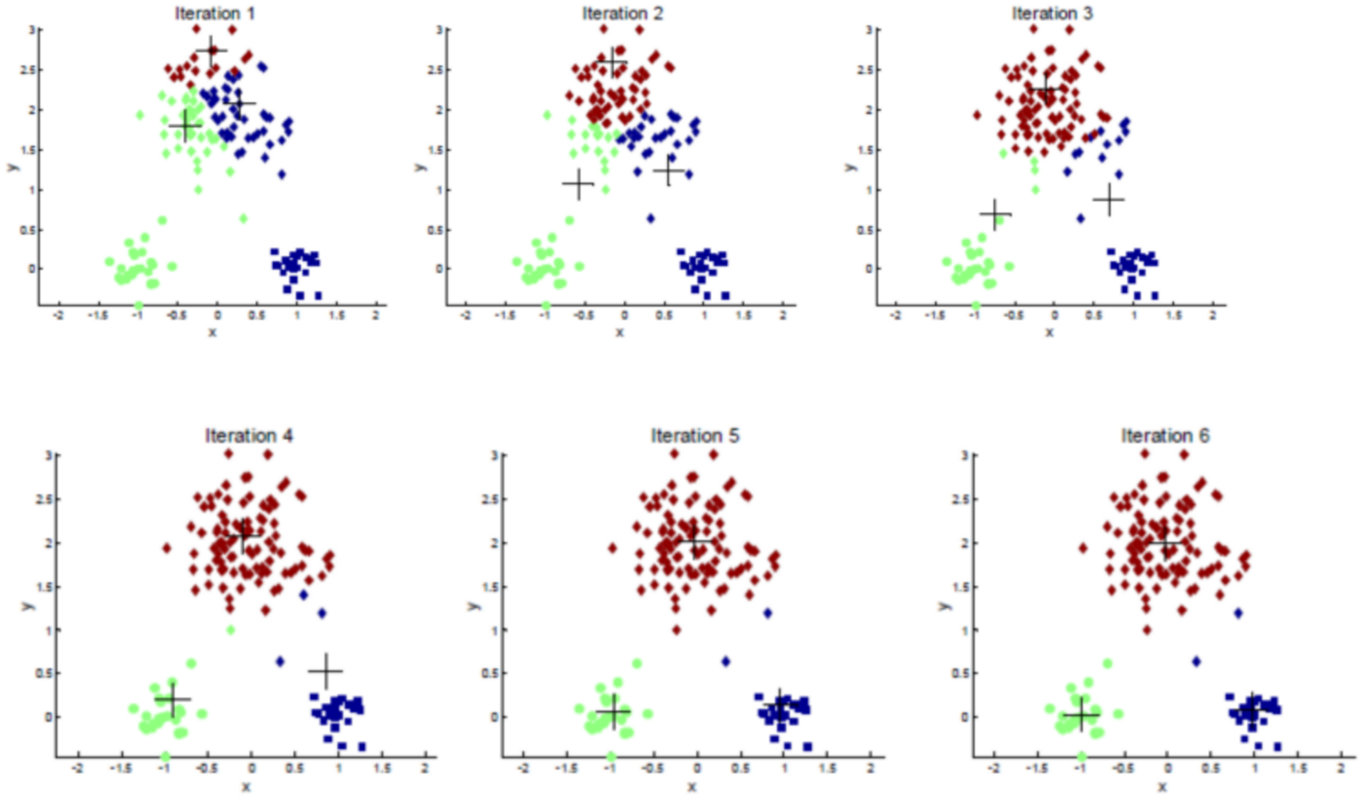
K-means Clustering
STEP 7

- 군집 중심을 군집 평균위치로 변경



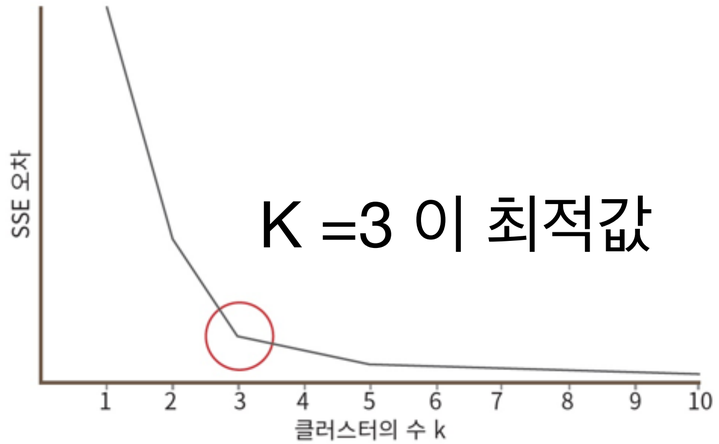
K-means Clustering

- 초기 중심값에 대해 민감한 군집화 결과



K를 결정하는 방법

- "팔꿈치" 방법(elbow method)에서는 k를 1부터 증가시키면서 K-means 클러스터링을 수행
  - 각 k의 값에 대하여 SSE(sum of squared errors)의 값을 계산



K-means Lab

Next

- Decision Tree 결정나무

Conclusions

- KNN - Lazy learning
  - No use of NN
  - 분류 Classification
- K-means - Unsupervised learning