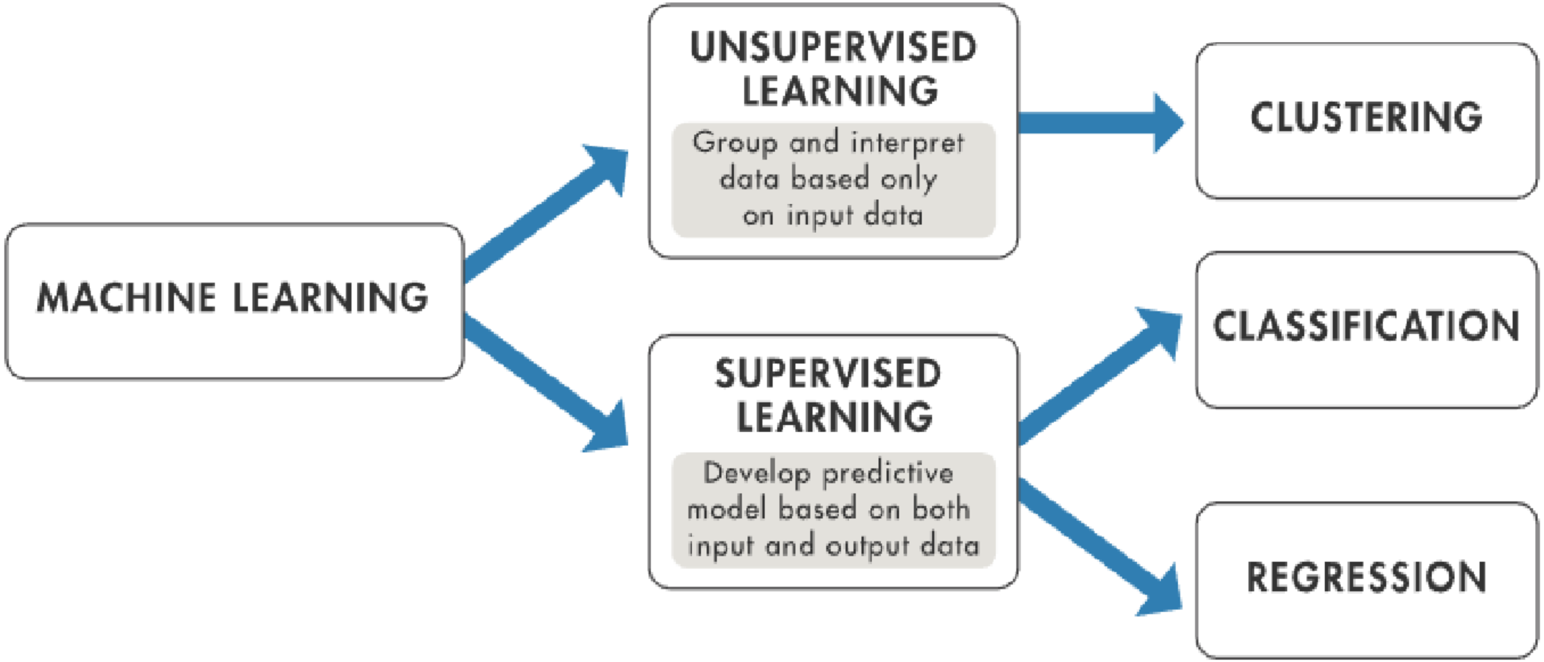
Distance Measure
&
K-means Clustering

KUBIG
박소현

INDEX

Ⅰ. 지도학습과 비지도학습



Ⅰ. 지도학습과 비지도학습

데이터에 대한 Label이 주어진 상태에서 컴퓨터를 학습시키는 방법

Feature : "개별적이고 측정 가능한 Heuristic한 속성"

NO.	SIZE	COLOR	SHAPE	FRUIT NAME
1	Big	Red	Rounded shape with a depression at the top	Apple
2	Small	Red	Heart-shaped to nearly globular	Cherry
3	Big	Green	Long curving cylinder	Banana
4	Small	Green	Round to oval,Bunch shape Cylindrical	Grape

(Training) Data

Label : "명시적인 정답"

- ✓ 수박? -> (Big, Green, Round shape)
- ✓ Feature가 다양할수록 복잡한 문제를 풀 수 있다

Ⅰ. 지도학습과 비지도학습

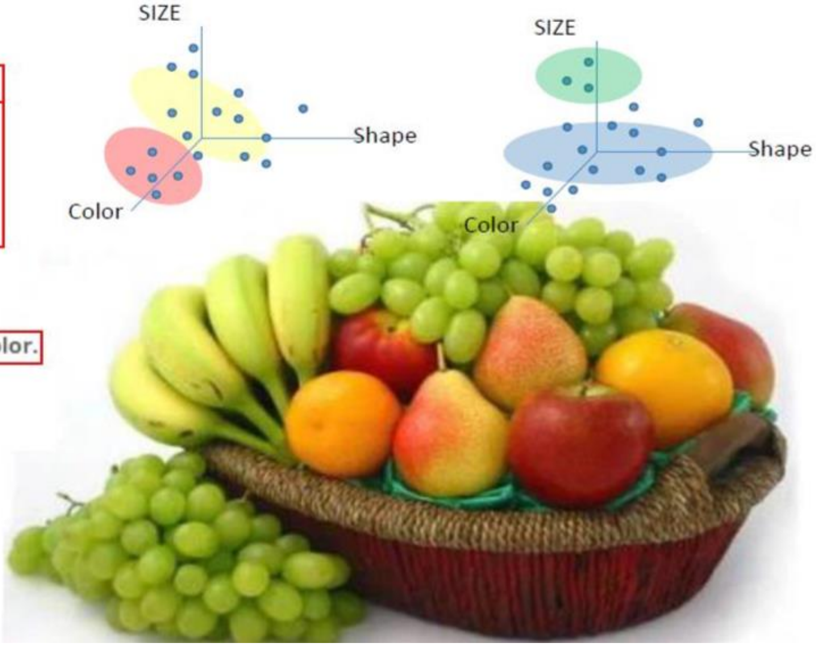
데이터에 대한 Label이 주어지지 않은 상태에서 컴퓨터를 학습시키는 방법

NO	SIZE	COLOR	SHAPE
1	Big	Red	Rounded shape with a depression at the top
2	Small	Red	Heart-shaped to nearly globular
3	Big	Green	Long curving cylinder
4	Small	Green	Round to oval,Bunch shape Cylindrical

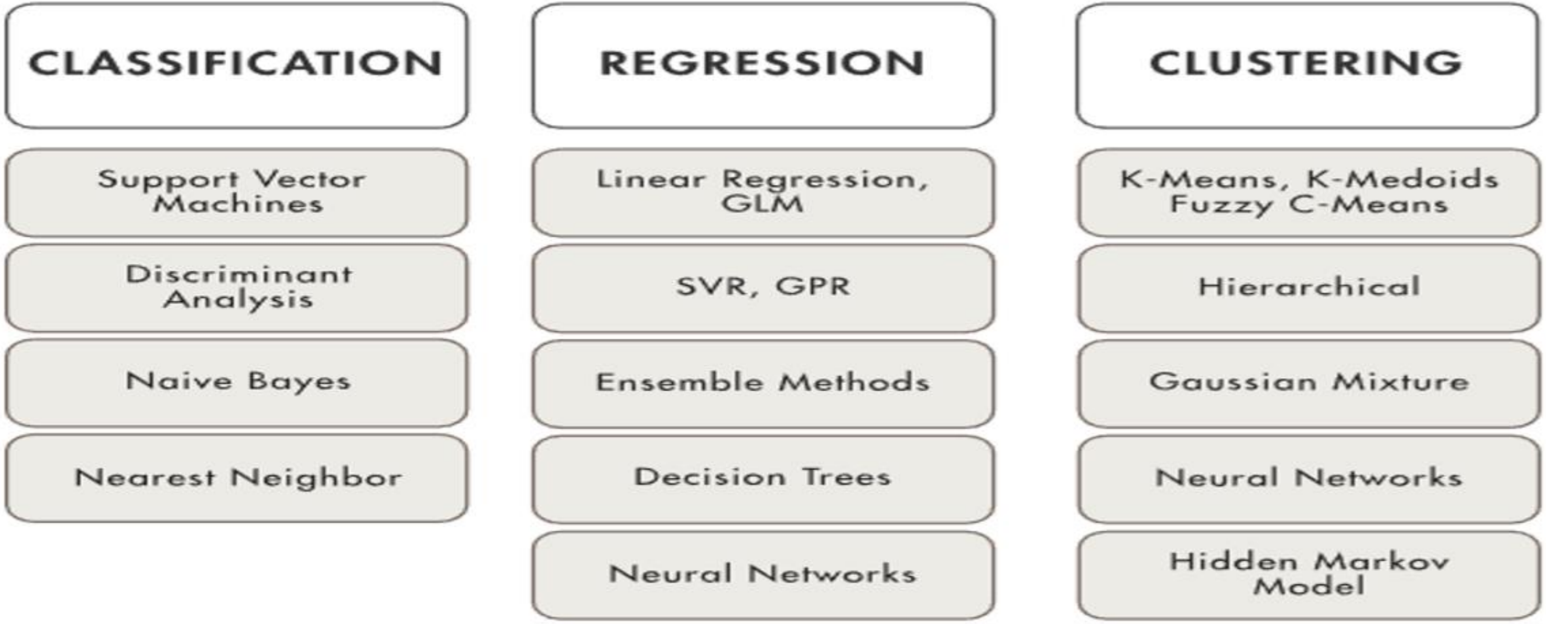
(Training) Data

- Then you will arrange them on considering base condition as **color.**
- Then the groups will be some thing like this.
- RED COLOR GROUP: apples & cherry fruits.
- GREEN COLOR GROUP: bananas & grapes.

- so now you will take another physical character such as **size**
- RED COLOR AND BIG SIZE: apple.
- RED COLOR AND SMALL SIZE: cherry fruits.
- GREEN COLOR AND BIG SIZE: bananas.
- GREEN COLOR AND SMALL SIZE: grapes.

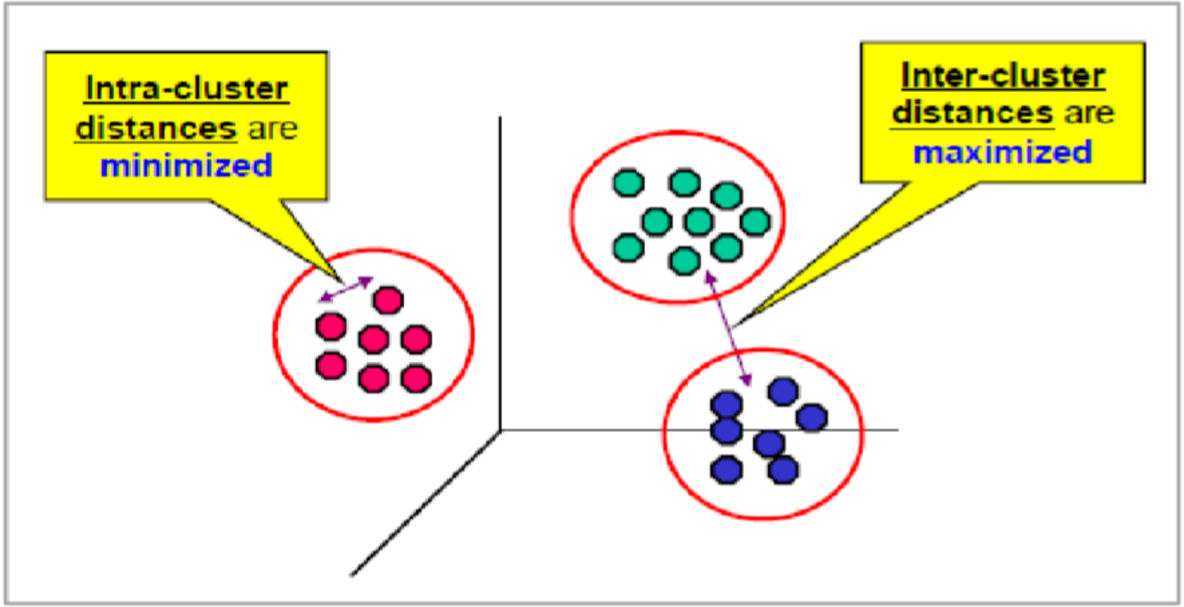


Ⅰ. 지도학습과 비지도학습

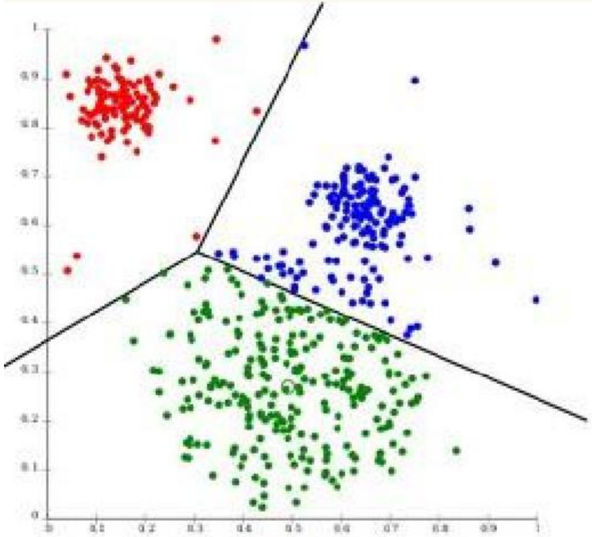


Ⅱ. 군집분석

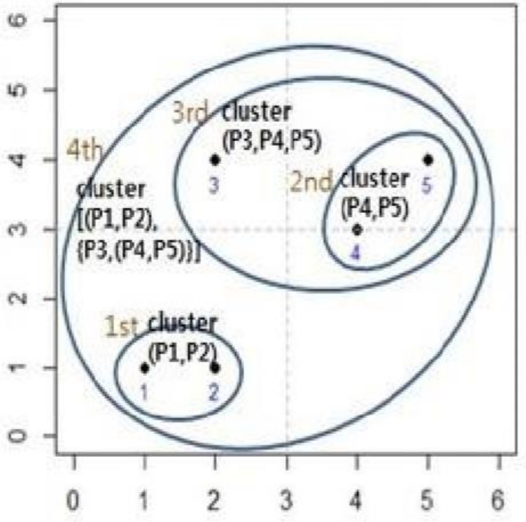
그룹 속 개체: 서로 비슷하거나 관련되어 있음/ 그룹 간 개체: 다르거나 관련이 없음



Ⅱ. 군집분석



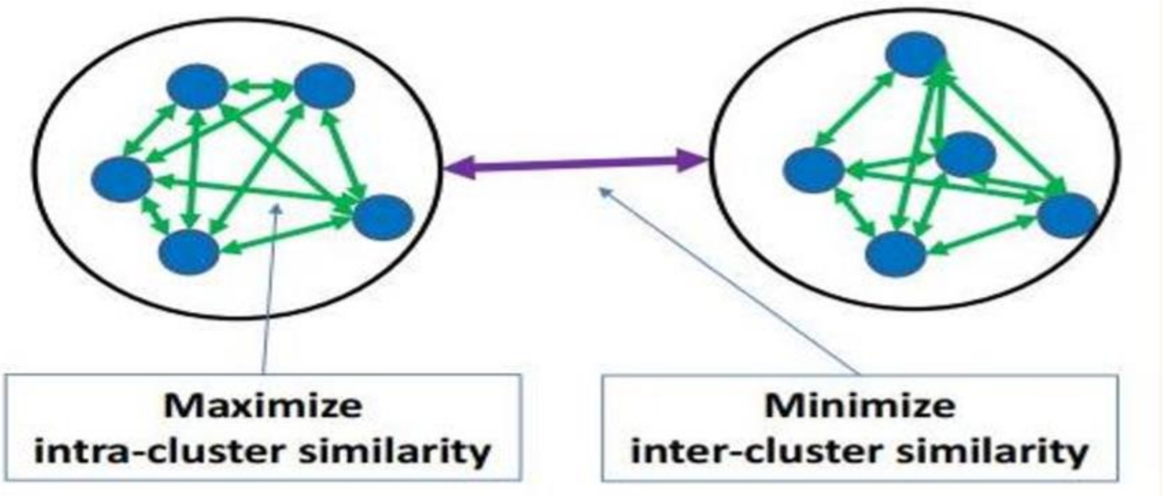
비계층적 군집분석



계층적 군집분석

KU-BIG

Ⅲ. 거리측정



동일한 군집에 속한 데이터는 서로 유사할수록 좋고, 다른 군집에 속한 데이터들은 서로 다를수록 좋다

High intra-class similarity & Low inter-class similarity

➡ **여기서 유사하다(Similarity) or 유사하지 않다(Dis-similarity)를 어떻게 측정할까?**

Ⅲ. 거리측정

유클리디안 거리
(Euclidean distance)

$$dist = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2}$$

where n is **the number of dimensions (attributes)** and

p_k and q_k are **the value of k^{th} attribute** of data objects p and q.

- 유클리드 거리는 가장 직관적이고 일반적으로 생각하는 거리 개념에 부합한다.
(점과 점 사이의 거리)

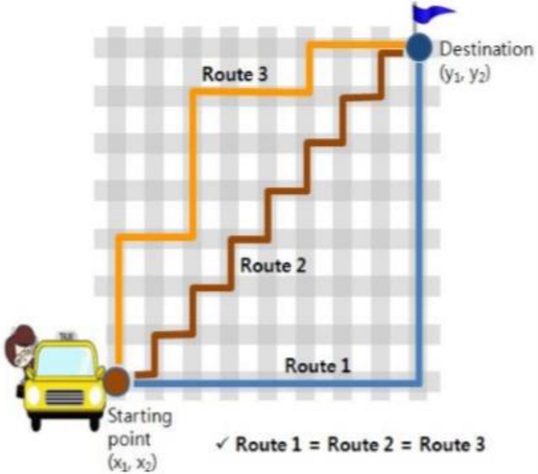
KU-BIG

Ⅲ. 거리측정

맨하탄 거리 (Manhattan distance)

$$d_M(x,y)=\sum_{j=1}^{m}|x_j-y_j|$$

- Manhattan 거리 측정법은 두 점의 좌표 간의 절대값 차이를 구함.
- Manhattan 격자 무늬 도로를 가진 Manhattan 에서 유래.
- 변수 간 상관성이 없고, 데이터 변수가 많을 때(차원이 클 때) 이용하는 것이 좋음



Ⅲ. 거리측정

민코스키 거리
(Minkowski Distance)

$$dist = \left(\sum_{k=1}^{n} | p_k - q_k |^r \right)^{\frac{1}{r}}$$

where r is a **parameter** (r **= 2 → Euclidean Distance**),

n is **the number of dimensions (attributes)** and

p_k and q_k are **the value of** k^{th} **attribute** of data objects p and q.

Ⅲ. 거리측정

민코우스키 거리 (Minkowski Distance)

$$d_{Minkowski}(x,y)=\left(\sum_{j=1}^{m}|x_j-y_j|^r\right)^{1/r}$$

✓ 1-norm distance

$$r=1 \Rightarrow d(x,y)=\sum_{j=1}^{m}|x_j-y_j|$$

✓ 맨하탄 거리 (Manhattan Distance)

✓ 2-norm distance

$$r=2 \Rightarrow d(x,y)=\left(\sum_{j=1}^{m}|x_j-y_j|^2\right)^{1/2}$$

✓ 유클리드 거리 (Euclidean Distance)

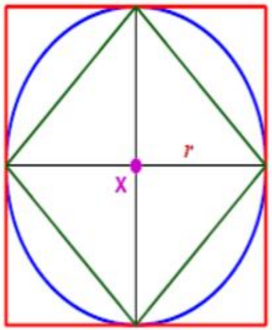
(In the Euclidean space R^n, the distance between two points)

✓ p-norm distance

$$r=p \Rightarrow d(x,y)=\left(\sum_{j=1}^{m}|x_j-y_j|^p\right)^{1/p}$$

✓ Infinity norm distance

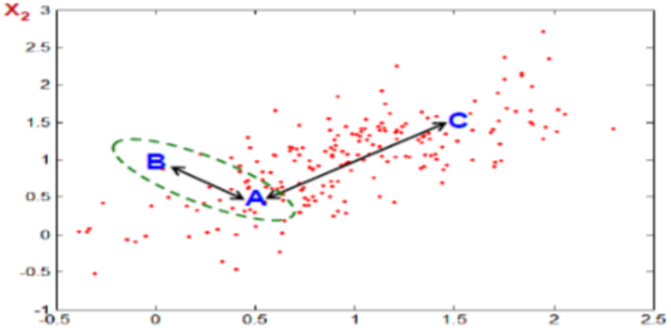
$$r\to\infty \Rightarrow d(x,y)=\lim_{r\to\infty}\left(\sum_{j=1}^{m}|x_j-y_j|^r\right)^{1/r}=\max|x_j-y_j|$$



- Green: All points y at distance $L_1(x, y) = r$ from point x
- Blue: All points y at distance $L_2(x, y) = r$ from point x
- Red: All points y at distance $L_\infty(x, y) = r$ from point x

Ⅲ. 거리측정

마할라노비스 거리
(Mahalanobis distance)



$$d(x,y) = \sqrt{(x-y)\Sigma^{-1}(x-y)^T}$$

where $x = (x_1, x_2, \dots, x_n), y = (y_1, y_2, \dots, y_n)$

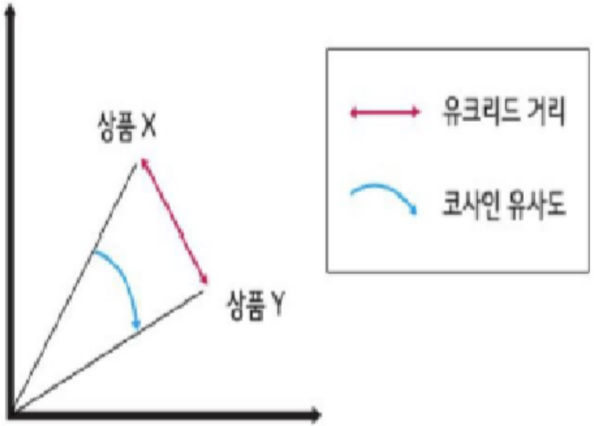
$$\Sigma = \begin{pmatrix} Cov(X_1,X_1) & \cdots & Cov(X_1,X_n) \\ \vdots & \ddots & \vdots \\ Cov(X_n,X_1) & \cdots & Cov(X_n,X_n) \end{pmatrix}$$

- 데이터의 속성들의 공분산을 반영하여 거리를 계산
- 변수 간의 상관 관계가 존재할 때 사용
- 계산 값이 0에 가까울수록 유사함

Ⅲ. 거리측정

코사인 유사도 (Cosine Similarity)

$$\underbrace{\frac{(a_1b_1 + a_2b_2 + \cdots a_n - b_n)}{\left(\sqrt{(a_1{}^2 + a_2{}^2 + \cdots + a_1{}^2)}\right)\left(\sqrt{(b_1{}^2 + b_2{}^2 + \cdots + b_1{}^2)}\right)}}_{\text{cosine similarity}} \quad D_C(A,B) = 1 - S_C(A,B$$



$$sim(x,y) = \frac{\langle x, y \rangle}{\|x\|\|y\|}$$

where $x = (x_1, x_2, \ldots, x_n), y = (y_1, y_2, \ldots, y_n)$

$$\|x\| = \sqrt{\sum_{i=1}^{n}(x_i)^2}, \langle x, y \rangle = \sum_{i=1}^{n} x_i y_i$$

Ⅲ. 거리측정

자카드 유사도 (Jaccard Similarity)

Simple Matching and Jaccard Similarity Coefficients

- SMC = number of matches / number of attributes

  $= (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

- J = # of 11 matches / # of not-both-zero attributes values

  $= (M_{11}) / (M_{01} + M_{10} + M_{11})$

Compute **similarities** using the following **quantities**

- M_{01} = the number of attributes where **p was 0** and **q was 1**
- M_{10} = the number of attributes where **p was 1** and **q was 0**
- M_{00} = the number of attributes where **p was 0** and **q was 0**
- M_{11} = the number of attributes where **p was 1** and **q was 1**

Ⅲ. 거리측정

SMC, Jaccard 예제

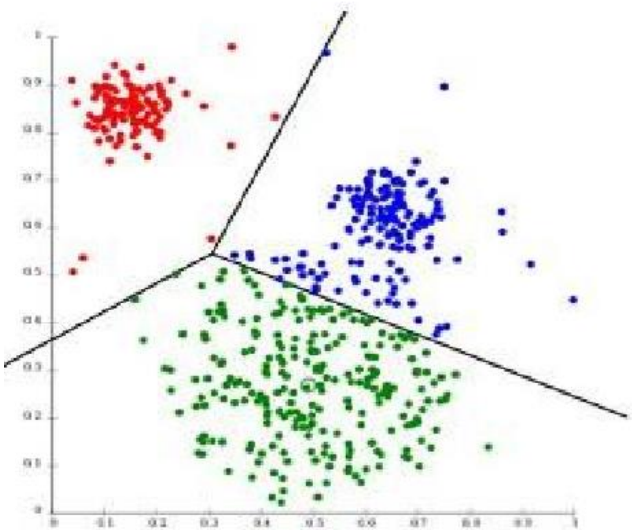
- p = 1 0 0 0 0 0 0 0 0 0
- q = 0 0 0 0 0 0 1 0 0 1

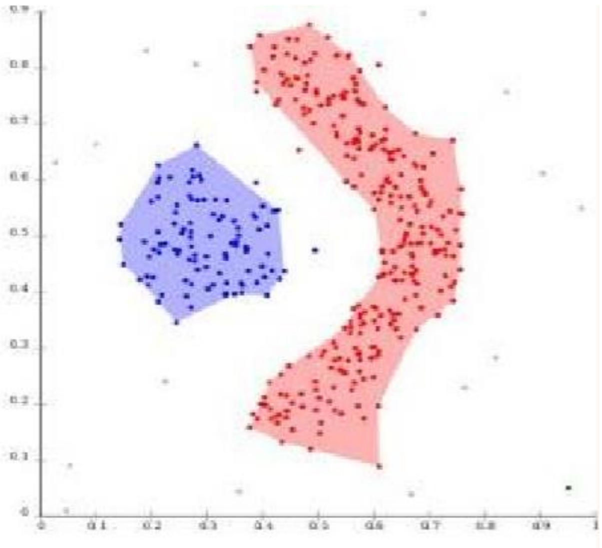
- $M_{01} = 2$ (the number of attributes where p was 0 and q was 1)
- $M_{10} = 1$ (the number of attributes where p was 1 and q was 0)
- $M_{00} = 7$ (the number of attributes where p was 0 and q was 0)
- $M_{11} = 0$ (the number of attributes where p was 1 and q was 1)

- SMC = $(M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$ = (7/10) = 0.7
- J = $(M_{11}) / (M_{01} + M_{10} + M_{11})$ = 0/3 = 0

KU-BIG

IV. 비계층적군집분석





군집끼리 포함관계를 이루지 않고 서로 독립적인 한 군집으로 만드는 기법
거리-기반 군집화: K-means Clustering
밀도-기반 군집화: DBSCAN

Ⅴ. K-means클러스터링

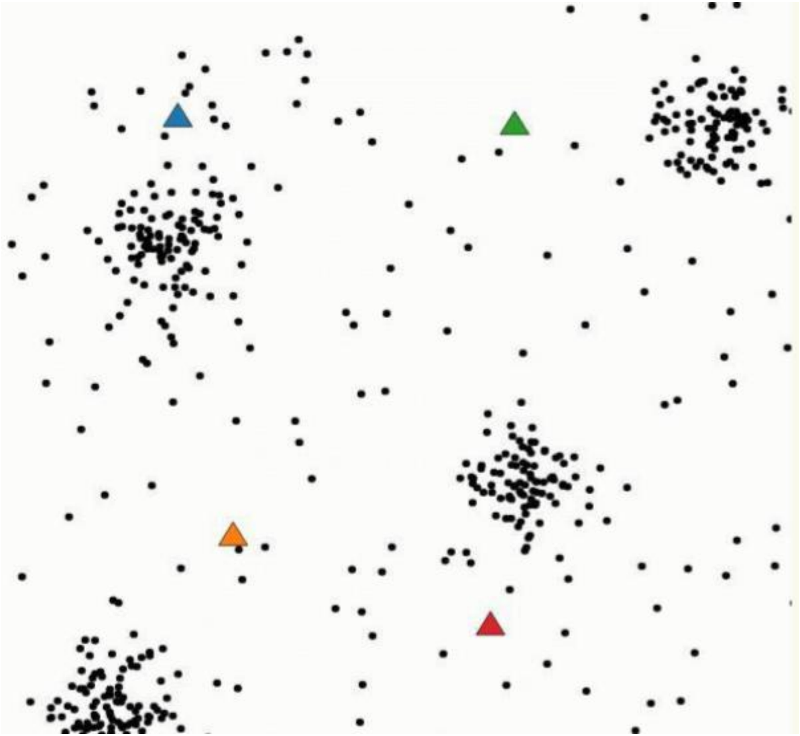
- 비계층적 군집방법 중 가장 널리 사용
- 미리 주어진 클러스터의 개수 k를 바탕으로 클러스터의 중심으로부터 가까운 데이터들을 찾아서 묶어주는 알고리즘

K-means Algorithm

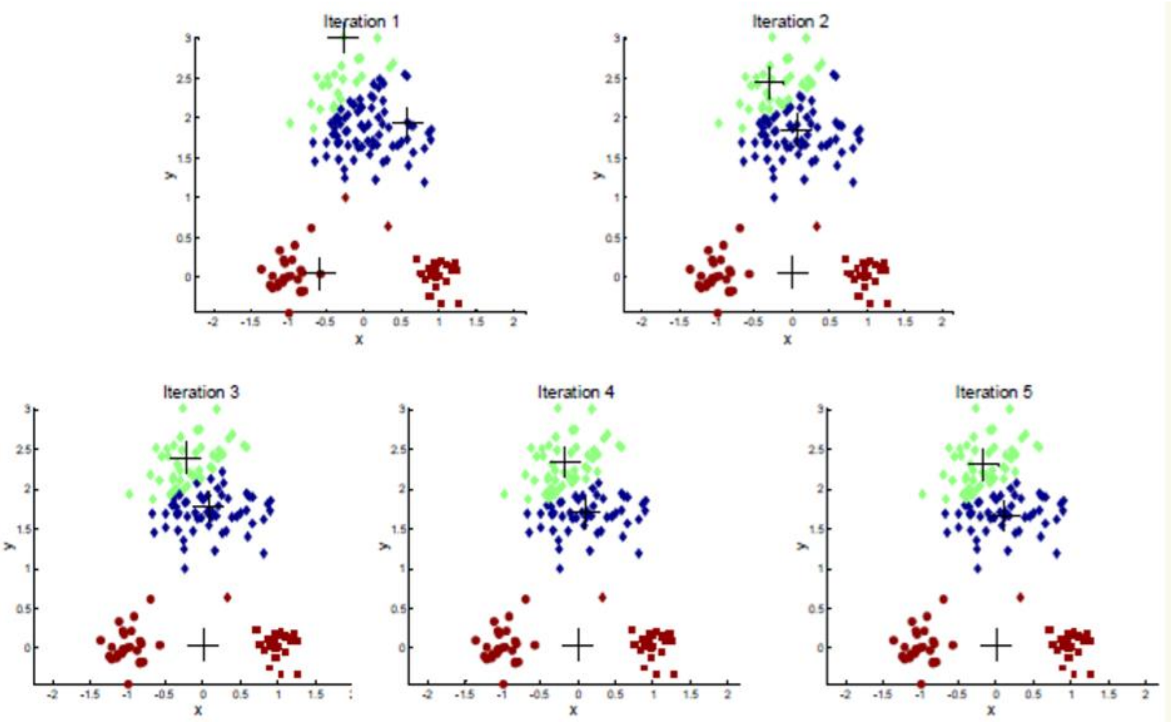
Select K points as the initial centroids.
repeat
    Form K clusters by assigning all points to the closest centroid.
    Recompute the centroid of each cluster.
until The centroids don't change (stopping condition)

KU-BIG

Ⅴ. K-means클러스터링



Ⅴ. K-means클러스터링



Thank you