K-means Clustering

Jeonghun Yoon

Clustering이란?

2000년과 2004년도 미국 대통령 선거는 정말 치열했다.

가장 많은 표를 얻은 후보는 전체 득표의 50.7%를 차지했고, 가장 적은 표를 얻은 후보는 전체 득표의 47.9%를 차지했다. 만약 1.4%의 투표자가 자신의 표를 바꾸었다면, 투표 결과는 완전히 달라졌을 것이다. 전체 인구에서 특정 후보를 강력하게 지지하지 않고, 단지 선거 유세를 통해 마음을 바꿀 수 있는 1.4%의 사람들만 찾았으면 말이다...

만약 이 소수의 투표자 그룹의 마음을 돌렸다면, 특히 이렇게 치열한 레이스에서, 미국 대통령은 달라졌을 수 있다.

여기서, 우리는 전체 집단을 몇 개의 그룹을 나누는 문제를 생각해 볼 수 있다.

즉, 투표자들 전체 집단을 어떤 기준에 따라 그룹을 나누는 것이다.
그리고 각각의 그룹에 특성화 된 선거 전략을 세우는 것이다.

Clustering이란?

그러면 어떻게 그룹을 나눌까?

먼저, 미국 투표자들 개개인의 정보를 얻는다. 이 정보는 그들이 투표를 하는 데에 영향을 미칠 수 있는 요소가 무엇인지에 대한 단서를 얻을 수 있는 중요한 것 들일 수 있다.

feature

다음으로 이 정보들을 가지고 그룹(clusters)을 나눈다.

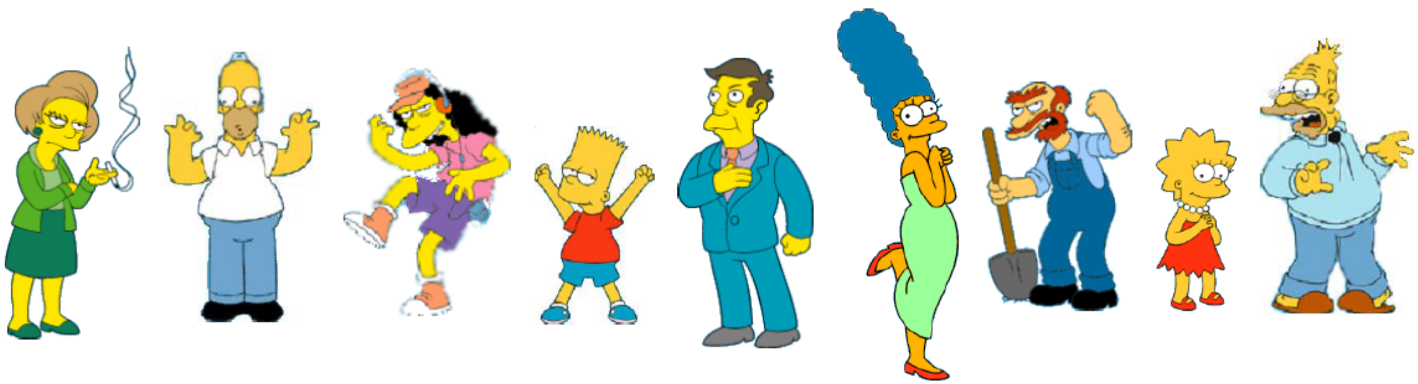
clustering

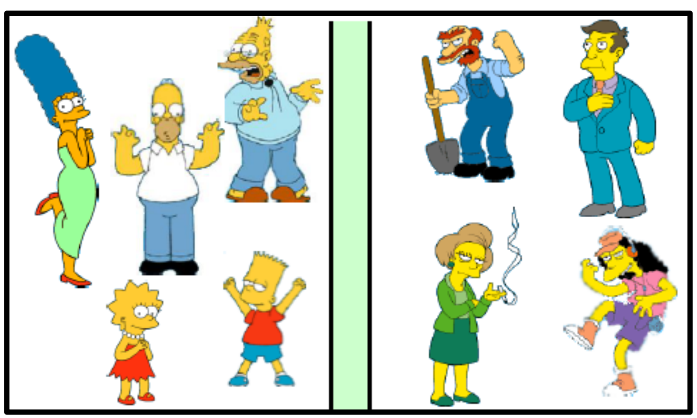
그러면, 특정 정당과 후보를 지지하는 그룹, 정당과 후보 보다는 공약에 좌우되는 그룹, 주변의 분위기에 편승되는 그룹 등등 그룹들을 찾을 수 있다.

그리고 각 그룹에 맞는, 즉 각 그룹의 유권자들에 특화된, 선거 유세를 그룹별로 진행한다.

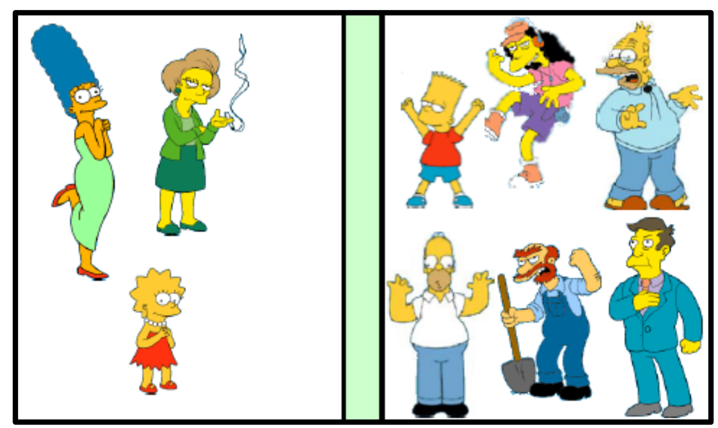
Clustering이란?

What is natural grouping among these objects?





Simpson's Family   None Simpson's Family (School Employees)



Females   males

Clustering이란?

Original Image

2 colors

4 colors

8 colors

Clustering이란?

What is Clustering?

- Organizing data into clusters such that there is
  - high intra-cluster similarity
  - low inter-cluster similarity
- Informally, finding natural grouping among object.

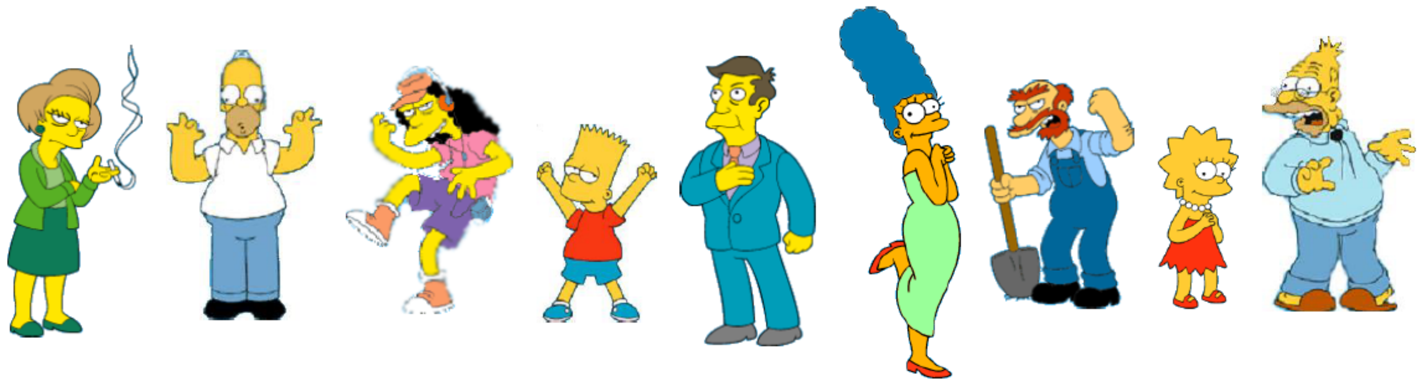
Clustering 이란?

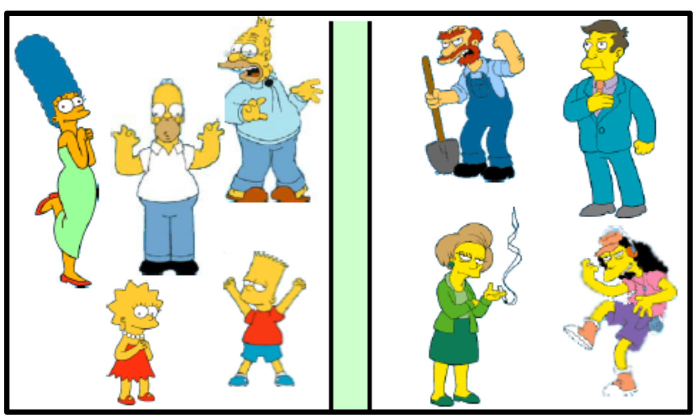
- data를 다음의 두 가지 조건을 만족하는 cluster(군집, 집단)로 조직(organizing)
  - cluster 내부의 data들 간의 <u>similarity</u>(유사성)이 높다.
  - cluster 외부의 data들 간의 <u>similarity</u>(유사성)이 낮다.

    ※ 유사성이 무엇일까?

Clustering이란?

What is natural grouping among these objects?

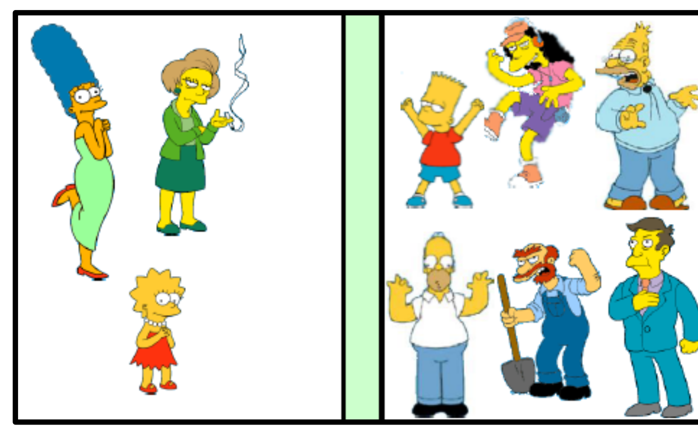




Simpson's Family

None Simpson's Family
(School Employees)

similarity : 이름



Females

males

similarity : 하의(치마, 바지)
눈썹

유사도Similarity

What is **Similarity**?

- The quality or state of being similar; likeness; resemblance; as a similarity of feature. (Webster's Dictionary)
- similarity(유사성)은 정의하기 힘드나, 우리는 그것을 눈으로 보면 직관적으로 안다.
- similarity의 실제 의미는 철학적인 문제이다.

비슷한가요??

유사한가요??

유사도$_{\text{Similarity}}$

두 개의 object사이의 similarity를 측정하기 위해서 두 object 사이의 거리 (distance , dissimilarity)를 측정한다.

아래 그림의 두 object 사이의 거리는 어떻게 측정할까?

Patty  *Selma*

- object 사이의 거리를 재기 위해서는, objects을 거리를 잴 수 있는 공간의 데이터들로 mapping 시켜야 된다.
- Patty와 Selma의 거리를 재기 위해서는, 예를 들어, 사람이라는 objects을 (입은 옷, 귀고리, 머리 모양, 몸무게, 키, 흡연)의 특성 벡터로 표현 될 수 있는 데이터로 mapping 하는 것이다.

유사도Similarity

거리(Distance)란?

- Edit Distance (두 개의 objects 사이의 similarity를 측정하는 일반적인 기법)
  - 한 개의 object에서, 다른 한 개의 object으로 변화되기 위하여 필요한 노력(effort)를 측정한다.

The distance between Patty and Selma

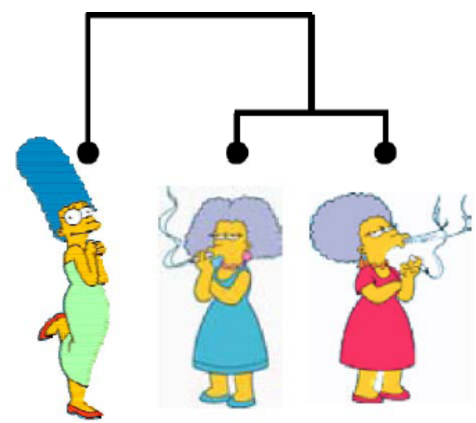
- Change dress color, 1 point
- Change earning shape, 1 point
- Change hair part, 1 Point

D(Patty, Selma)=3

The distance between Marge and Selma

- Change dress color, 1 point
- Add earning, 1 point
- Decrease height, 1 point
- Take up smoking, 1 point
- Lose weight, 1 point

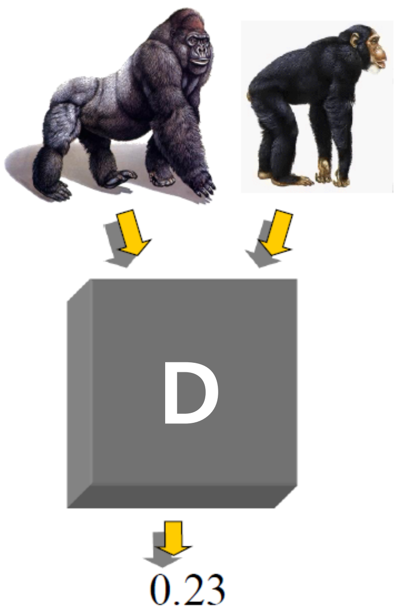
D(Patty, Selma)=5

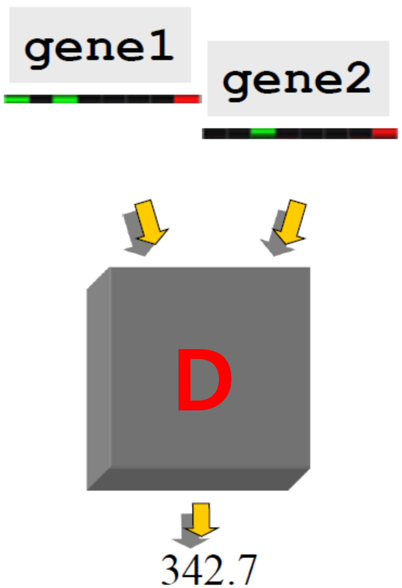


This is called the Edit distance or the Transformation distance

Similarity

- 거리(Distance)의 수학적인 개념
  - O_1과 O_2를 두 개의 object이라고 할 때, O_1과 O_2 사이의 distance는 실수(real number)이고 $\boldsymbol{D(O_1, O_2)}$라고 표기
  - 주의할 것! distance는 우리가 고등학교 때 배워서 알고 있는 유클리드 거리만 의미하는 것이 아니다. distance function은 다양하게 설정할 수 있다.





Similarity

- Distance function 과 Similarity function의 예

  $\mathbb{x} = (x_1, x_2, \ldots), \mathbb{y} = (y_1, y_2, \ldots)$ 라고 하자.

  - Euclidian distance (dissimilarity)

$$D(\mathbb{x}, \mathbb{y}) = d(\mathbb{x}, \mathbb{y}) = \sqrt{\sum_i (x_i - y_i)^2}$$

② Manhattan distance (dissimilarity)

$$D(\mathbb{x}, \mathbb{y}) = d(\mathbb{x}, \mathbb{y}) = \sum_i |x_i - y_i|$$

③ "sup" distance (dissimilarity)

$$D(\mathbb{x}, \mathbb{y}) = d(\mathbb{x}, \mathbb{y}) = \max_i |x_i - y_i|$$

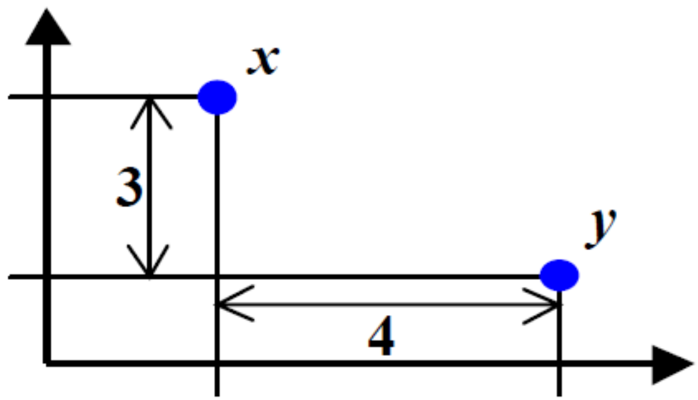
④ Correlation coefficient (similarity)

$$D(\mathbb{x}, \mathbb{y}) = s(\mathbb{x}, \mathbb{y}) = \frac{\sum_i (x_i - \mu_{\mathbb{x}})(y_i - \mu_{\mathbb{y}})}{\sigma_x \sigma_y}$$

⑤ Cosine similarity (similarity)

$$D(\mathbb{x}, \mathbb{y}) = \cos(\mathbb{x}, \mathbb{y}) = \frac{\mathbb{x} \cdot \mathbb{y}}{\|\mathbb{x}\| \|\mathbb{y}\|} = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2} \sqrt{\sum_i y_i^2}}$$

Similarity



1 : Euclidean distance : $\sqrt[2]{4^2+3^2}=5.$

2 : Manhattan distance : $4+3=7.$

3 : "sup" distance : $\max\{4,3\}=4.$

Why Clustering

data를 cluster(군집)들로 조직하면(organizing), data의 내부 구조(internal structure)에 대한 정보를 얻을 수 있음

data를 분할하는 것(partitioning) 자체가 목적이 될 수 있음

- 이미지 분할(image segmentation)

data에서 지식을 발견하는 것이 목적이 될 수 있음 (knowledge discovery in data)

- Underlying rules
- topic

Clustering problem

Input

- Training set $\boldsymbol{S_n} = \{\mathbb{x}^{(i)}, \boldsymbol{i} = \mathbf{1}, \dots, \boldsymbol{n}\}$, where $\mathbb{x}^{(i)} \in R^d$
- Integer $\boldsymbol{k}$

Output

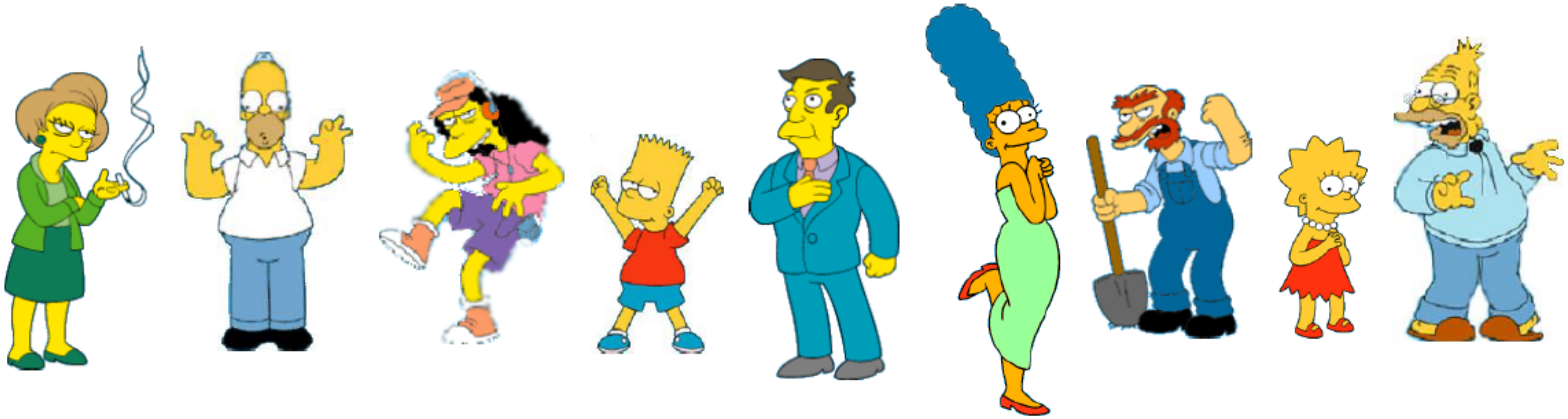
- A set of clusters $\boldsymbol{C_1}, \boldsymbol{C_2}, \dots, \boldsymbol{C_k}$

$$
\begin{aligned}
\mathbb{x}^{(1)} &= \left\{x_1^{(1)}, x_2^{(1)}, \dots, x_d^{(1)}\right\} \\
\mathbb{x}^{(2)} &= \left\{x_1^{(2)}, x_2^{(2)}, \dots, x_d^{(2)}\right\} \\
&\dots \\
\mathbb{x}^{(n)} &= \left\{x_1^{(n)}, x_2^{(n)}, \dots, x_d^{(n)}\right\}
\end{aligned}
\quad \xrightarrow{\text{Clustering}} \quad
\begin{aligned}
C_1 &= \{\mathbb{x}^{(1)}, \mathbb{x}^{(4)} \dots\} \\
C_2 &= \{\mathbb{x}^{(2)}, \mathbb{x}^{(6)} \dots\} \\
&\dots \\
C_k &= \{\mathbb{x}^{(3)}, \mathbb{x}^{(8)} \dots\}
\end{aligned}
$$

K-Means Clustering

같은 cluster에 속하는 데이터들의 inner similarity를 증가시키는 방향으로 cluster를 형성

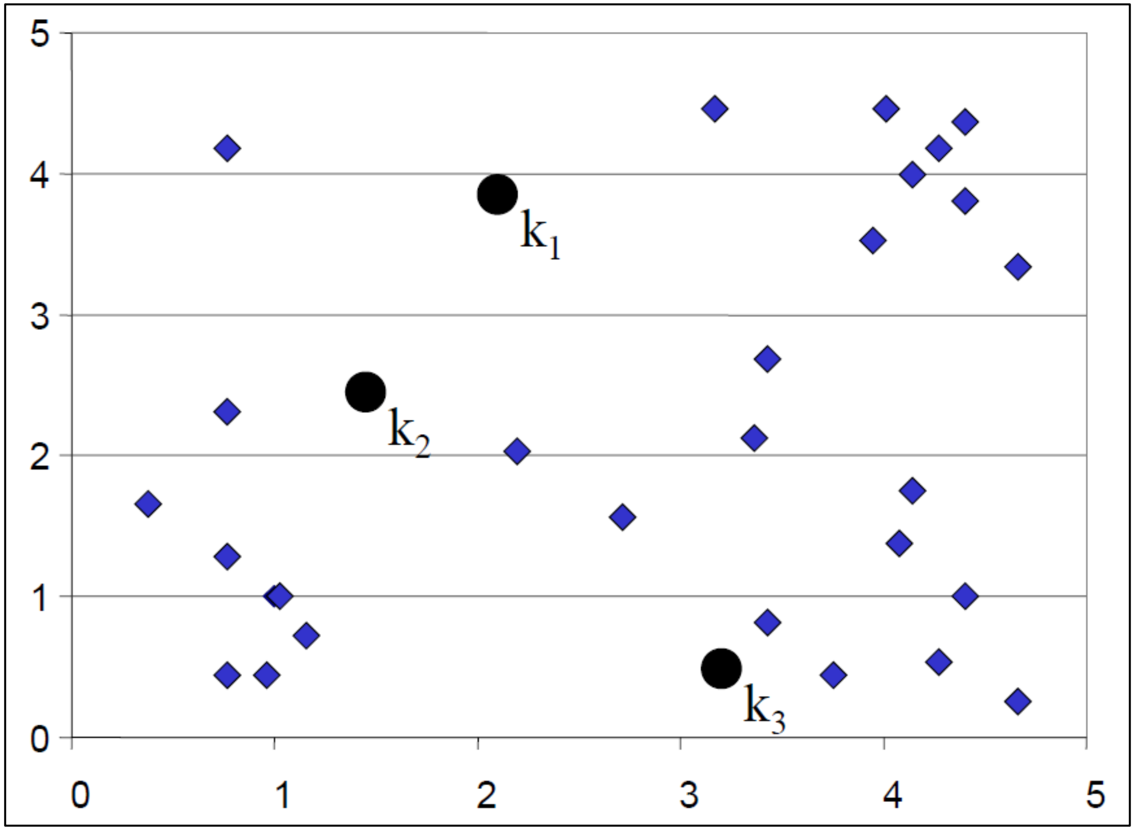
가정 : 데이터가 유클리디안 공간위에 있어야 한다.
(평균을 구할 수 있도록, 실수의 좌표를 가져야 한다.)



(1,2) (2,2) (4,10) ....

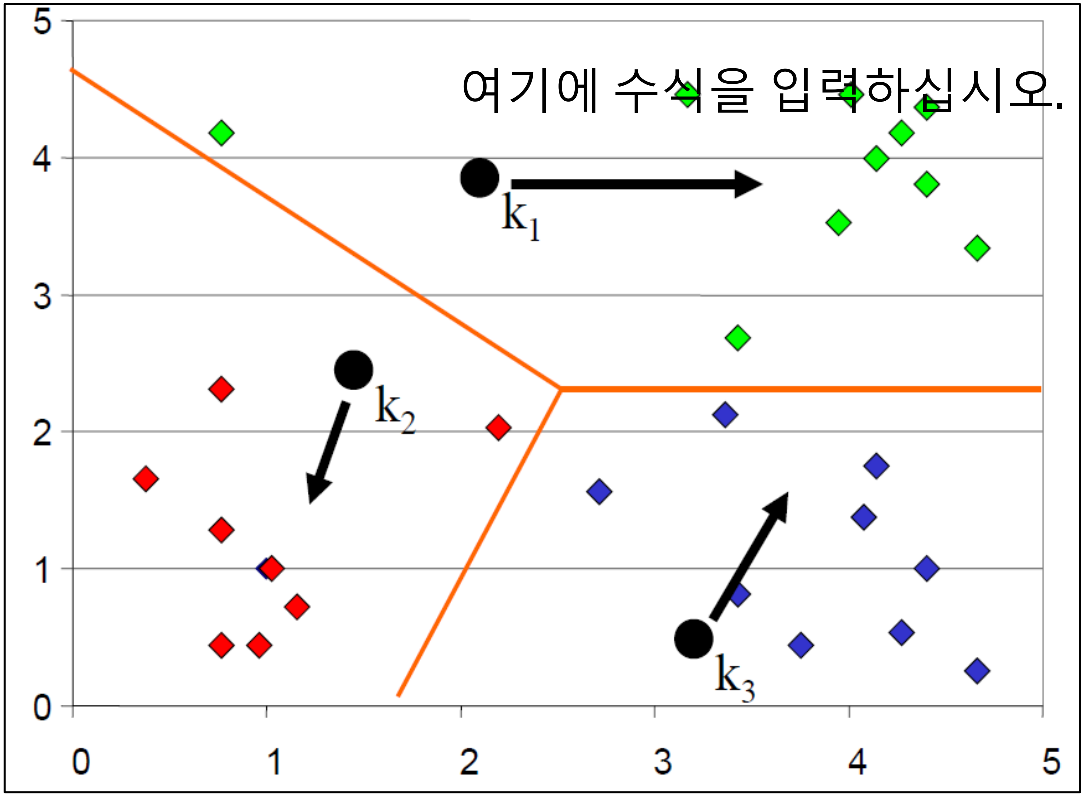
K-means clustering algorithm

- Cluster의 개수를 3개로 정함
- K개의 초기 centroid(클러스터의 중심이 될)를 random하게 선택한다.



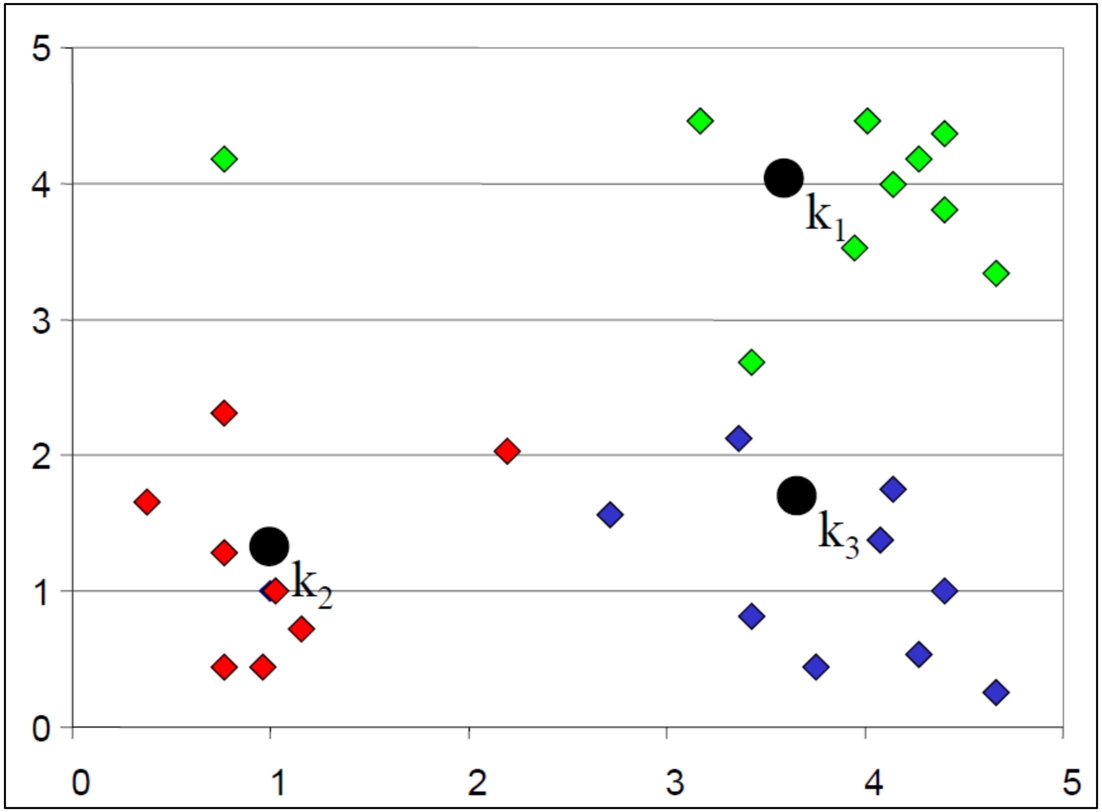
K-means clustering algorithm

- 각각의 object을, 자기자신에게서 가장 가까운 centroid k_i에 할당한다.
- 같은 centroid에 할당된 object들의 평균을 구한다.



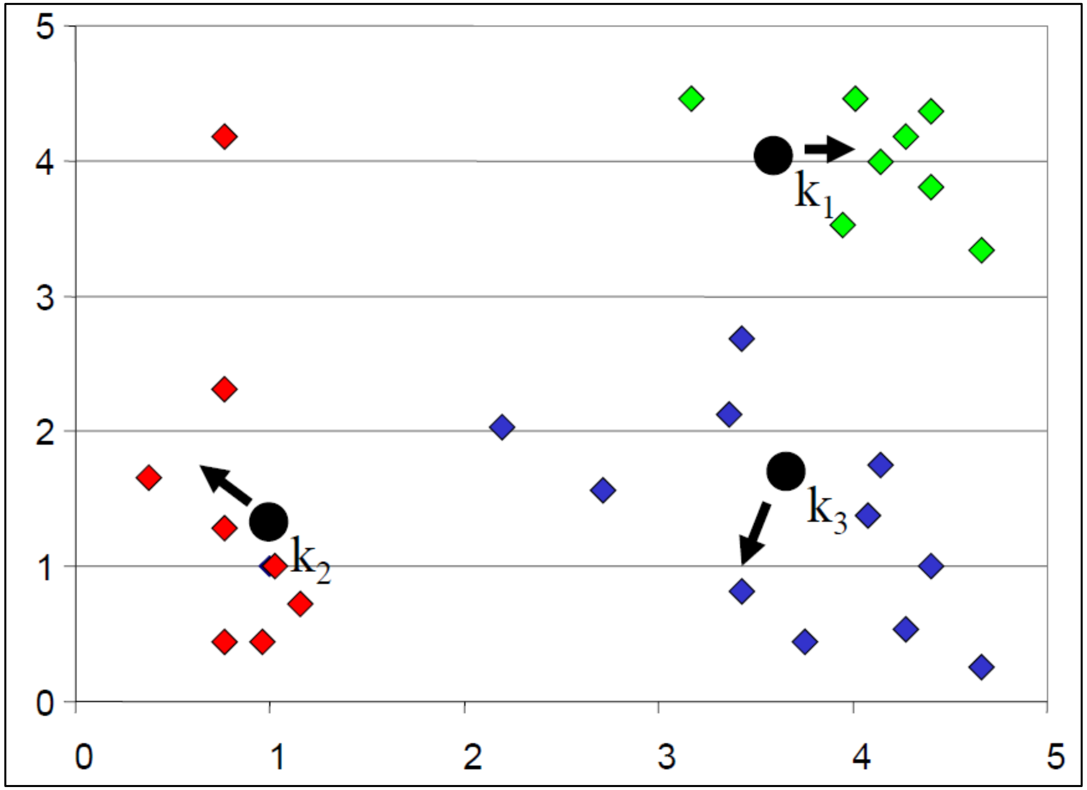
K-means clustering algorithm

- 구한 평균값을 2번째 centroid 값으로(클러스터의 중심으로) 설정한다.



K-means clustering algorithm

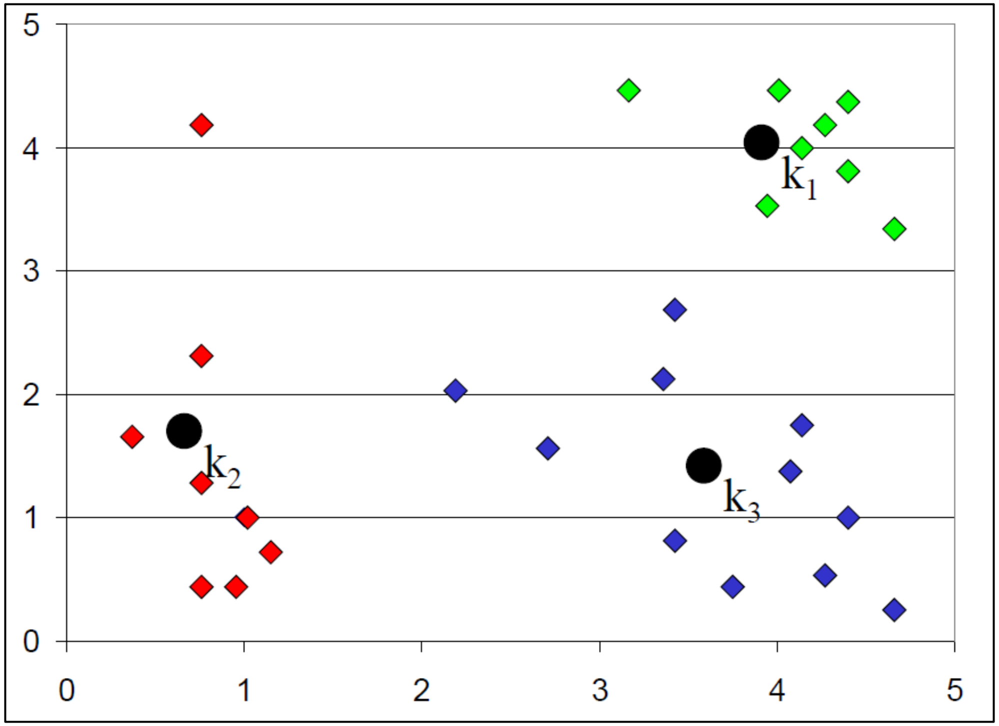
- 각각의 object을, 자기자신에게서 가장 가까운 centroid k_i에 할당한다. (2번째 centroid)
- 같은 centroid에 할당된 object들의 평균을 구한다.



K-means clustering algorithm

- 구한 평균값을 3번째 centroid 값으로(클러스터의 중심으로) 설정한다.
- 각각의 object을, 자기자신에서 가장 가까운 centroid k_i에 할당한다. (3번째 centroid)
- 각각의 centroid에 할당된 object들의 멤버쉽이 변화가 없을 때까지 반복한다.

더 이상, 데이터 자신이 속한 클러스터가 변하지 않는 것을 의미한다.



K-means clustering algorithm

① K의 값을 결정한다. K는 cluster의 개수이다.

② K cluster의 초기 centroid $\mathbb{C} = \left\{\mathbb{c}^{(1)}, \mathbb{c}^{(2)}, \dots, \mathbb{c}^{(K)}\right)$를 random하게 설정한다.

③ 각각의 object을, 그 object과 가장 가까운 centroid로 할당한다. 이것을 통해 전체 object의 class membership을 결정한다.

④ 같은 centroid에 할당된 object들의 평균값 $\boldsymbol{\mu} = \{\mathbb{m}^{(1)}, \mathbb{m}^{(2)}, \dots, \mathbb{m}^{(K)}\}$을 구한다.

⑤ 구한 평균값을 새로운 centroid로 설정. $\mathbb{C} = \left\{\mathbb{c}^{(1)}, \mathbb{c}^{(2)}, .., \mathbb{c}^{(K)}\right) \leftarrow \boldsymbol{\mu} = \{\mathbb{m}^{(1)}, \mathbb{m}^{(2)}, .., \mathbb{m}^{(K)}\}$

⑥ 3 ~ 5 단계를, 더 이상 어떠한 object도 자신이 속한 cluster가 변하지 않을 때까지, 즉 cluster의 membership이 바뀌지 않을 때까지 반복한다.

더 이상, 데이터 자신이 속한 클러스터가 변하지 않는 것을 의미한다.

K-means clustering 수학적인 표기

Clusters based on centroid

$$C_j = \{i \mid i \in \{1, \dots, n\}\ s.t.\ the\ closest\ centroid\ from\ \mathbb{x}^{(i)}\ is\ \mathbb{c}^{(j)}\}$$

- C_j는 인덱스의 집합

Cost function based on centroids

$$cost\left(C_1, C_2, \dots, C_K, \mathbb{c}^{(1)}, \mathbb{c}^{(2)}, \dots, \mathbb{c}^{(K)}\right) = \sum_{j=1,\dots,K} \sum_{i \in C_j} \left\| \mathbb{x}^{(i)} - \mathbb{c}^{(j)} \right\|$$

① Initialize centroids $\mathbb{c}^{(1)}, \mathbb{c}^{(2)}, \dots, \mathbb{c}^{(K)}$

② Repeat until there is no further change in $cost\left(C_1, C_2, \dots, C_K, \mathbb{c}^{(1)}, \mathbb{c}^{(2)}, \dots, \mathbb{c}^{(K)}\right)$

1) for each $j = 1, \dots, K$: $C_j = \{i \mid i\ s.t.\ \mathbb{x}^{(i)}\ is\ closest\ to\ \mathbb{c}^{(j)}\}$

2) for each $j = 1, \dots, K$: $\mathbb{c}^{(j)} = \frac{1}{|C_j|} \sum_{i \in C_j} \mathbb{x}^{(i)}$

Convergence of K-means clustering

K-means clustering 알고리즘은 수렴하는가?

- 그렇다. 알고리즘의 ①, ②번 스텝에서 cost function의 값이 줄어든다. 따라서 전체 알고리즘에서의 cost 값은 점진적으로(monotonically) 감소한다.
- Step ① : reassigning clusters based on distance

  Old clusters : $C_1^o, C_2^o, \ldots, C_K^o$

  New clusters : $C_1^N, C_2^N, \ldots, C_K^N$

$$cost(C_1^o, C_2^o, \ldots, C_K^o, \mathbb{c}_1, \mathbb{c}_2, \ldots, \mathbb{c}_K) \geq \min_{C_1,\ldots,C_K} cost(C_1, C_2, \ldots, C_K, \mathbb{c}_1, \mathbb{c}_2, \ldots, \mathbb{c}_K)$$
$$= cost(C_1^N, C_2^N, \ldots, C_K^N, \mathbb{c}_1, \mathbb{c}_2, \ldots, \mathbb{c}_K)$$

- Step ② : reassigning centroids based on clusters

  Old centroid : $\mathbb{c}_1, \mathbb{c}_2, \ldots, \mathbb{c}_K$

  New centroid : $\mathbb{c}_1^N, \mathbb{c}_2^N, \ldots, \mathbb{c}_K^N$

$$cost(C_1^N, C_2^N, \ldots, C_K^N, \mathbb{c}_1, \mathbb{c}_2, \ldots, \mathbb{c}_K) \geq \min_{\mathbb{c},\ldots,\mathbb{c}_K} cost(C_1^N, C_2^N, \ldots, C_K^N, \mathbb{c}_1, \mathbb{c}_2, \ldots, \mathbb{c}_K)$$
$$= cost(C_1^N, C_2^N, \ldots, C_K^N, \mathbb{c}_1^N, \mathbb{c}_2^N, \ldots, \mathbb{c}_K^N)$$

K-means clustering의 장/단점

Strength

- Simple, easy to implement and debug
- Intuitive objective function : optimize intra-cluster similarity
- Relatively efficient : $O(tkn)$, where n is number of objects, k is number of clusters and n is number of iterations. Normally, k, $t \ll n$.

Weakness

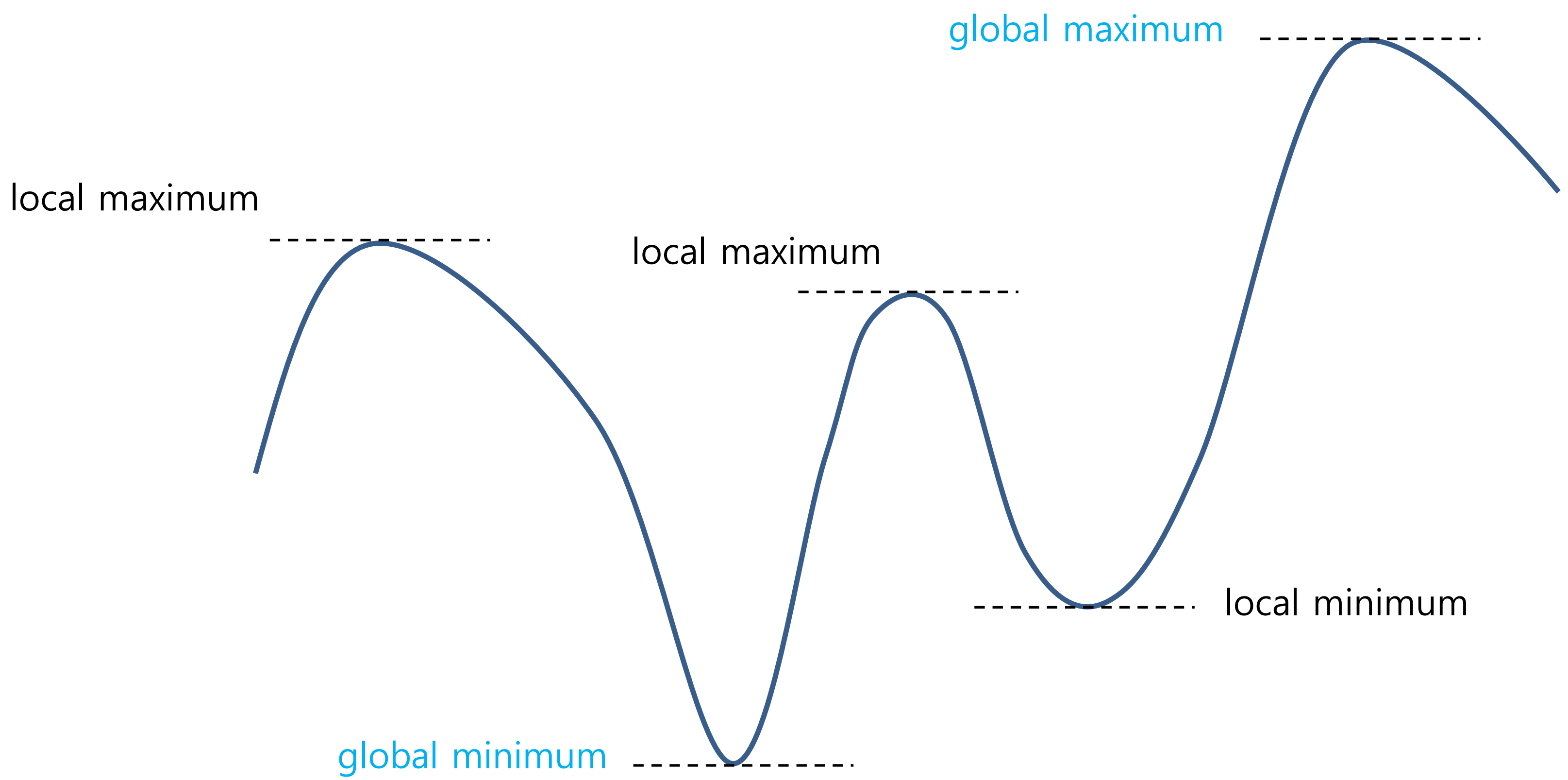
- Applicable only when mean is defined.
- Often terminates at a local optimum. Initialization is important.

  → Not suitable to discover clusters with non-convex shapes
- Need to specify K, the number of clusters, in advance.
- Unable to handle noisy data and outliers

Summary

- Assign members based on current centers
- Re-estimate centers based on current assignment

Global/ Local optimum



Convex function : local optimum = global optimum
ex) 2차 함수

그러면 K-means는 local optimum에 빠질 수 있는데 어떻게 해야 하나?
초기값을 바꾸면서 여러 번 반복한다. 그리고 반복 된 값 중에서 비용함수가 가장 작은 값을 선택한다.