TFIP-AI – Advanced Machine Learning

Unit 2 Clustering (K-Means)

Clustering

- Cluster
  - Collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters
- Clustering Analysis
  - Birds of a feather flock together

Byname →

Unsupervised learning
Learning without a teacher
Numerical taxonomy
Typology
Partition

What is Clustering

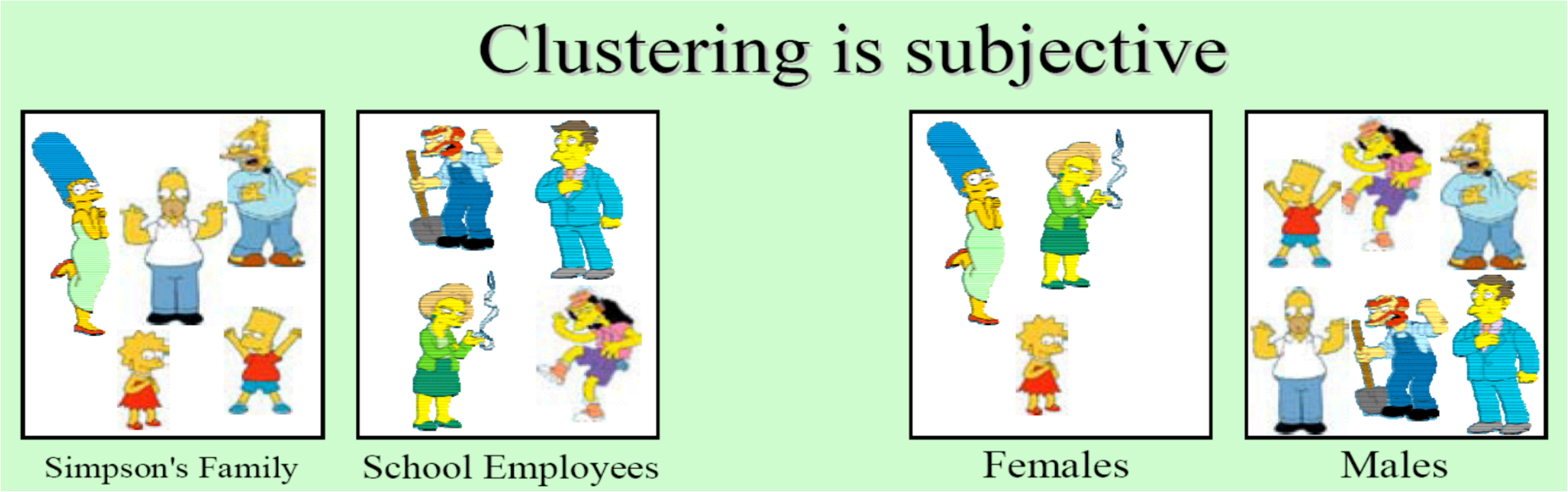
Clustering:

The process of grouping a set of objects into classes of similar objects

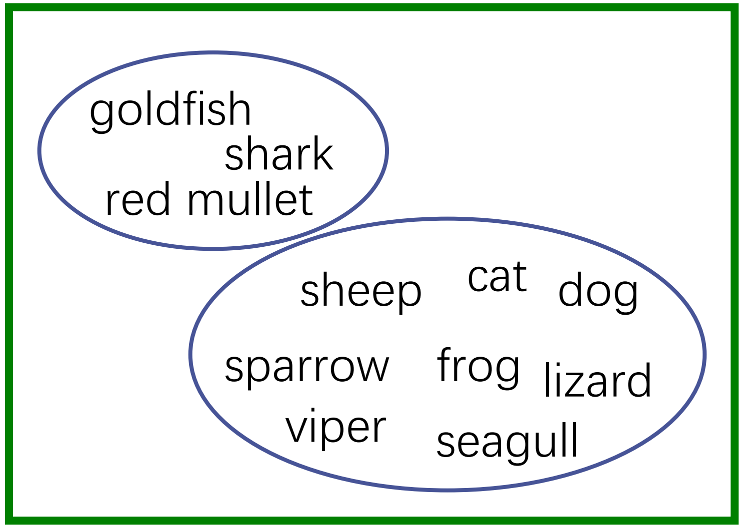
–high intra-class similarity

–low inter-class similarity

–It is the commonest form of unsupervised learning

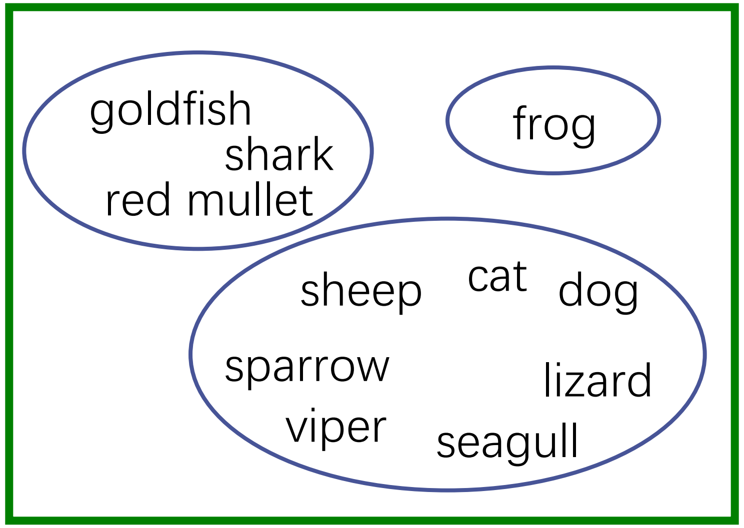


Clustering Criterion



The existence of lungs

The environment to live



What is Similarity?

Hard to define! *But we know it when we see it*

The real meaning of similarity is a philosophical question. We will take a more pragmatic approach: think in terms of a **distance** (rather than similarity) between random variables.

Clustering Similarity

- Numerical
  - Euclidean distance
  - Manhattan distance
  - Minkowski distance
  - …
- Binary, Nominal, Ordinal etc.
  - Jaccard coefficient
    - $sim(p_i, p_j) = | p_i \cap p_j | / | p_i \cup p_j |$
- Mixed

Typical Application

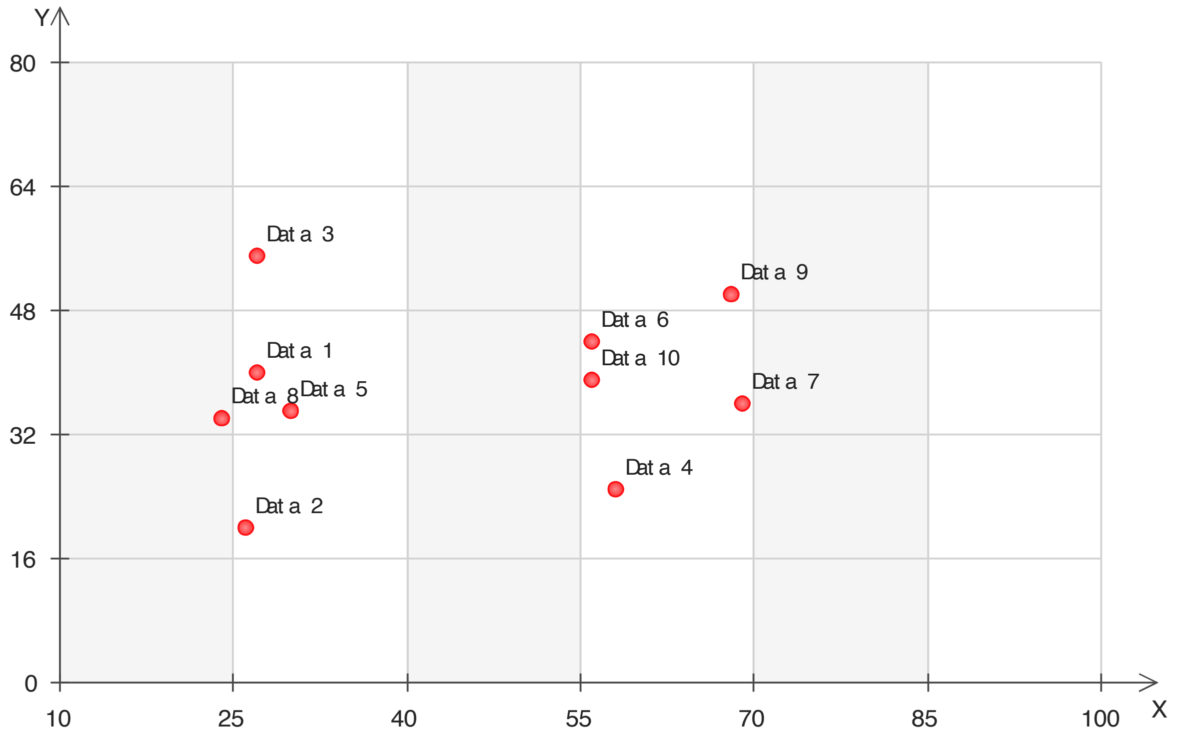
- Business: CRM
- Biology: Gene
- Identification of groups of ...

- Image processing
- Gain distribution of data
- Web for information discovery
- Preprocessing step

Clustering – input and result

- To find structure from the training data set

$$\begin{bmatrix} x_{11}x_{12}\ldots x_{1n} \\ x_{21}x_{22}\ldots x_{2n} \\ \ldots \\ x_{m1}x_{m2}\ldots x_{mn} \end{bmatrix}$$



Objective function

- Given
  - n objects
  - k represents number of clusters
  - *objective function*
- Gain
  - n objects are organized into k cluster
  - the formed clusters optimize the *objective function*

$$E = \frac{Total\ Distance(intraCluster)}{Total\ Distance(interCluster)}$$

Collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters

Clustering – 1-d example

Collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters

D = {o1, o2, o3, o4, o5}={3, 1, 9, 10, 2 }, K=2

Clustering1: {3,1,9}, {10,2}

$$E1 = \frac{[d(3,1)+d(3,9)+d(1,9)]+[d(10,2)]}{d(3,10)+d(3,2)+d(1,10)+d(1,2)+d(9,10)+d(9,2)}$$

Clustering2: {3,1,2}, {9,10}

$$E2 = \frac{[d(3,1)+d(3,2)+d(1,2)]+[d(9,10)]}{d(3,10)+d(3,9)+d(1,10)+d(1,9)+d(2,10)+d(2,9)}$$

…

ClusteringN: ⋯

$$EN = \cdots$$

$$\boxed{E = \frac{\sum_{m=1}^{K}\sum_{Oi,Oj\in C_m} d(O_i,O_j)}{\sum_{m=1}^{K}\sum_{n=1}^{K}\sum_{O_i\in C_m,O_j\in C_n} d(O_i,O_j)}}$$

When the size of D grows → combination explosion

Clustering – 1-d example

D = {o1, o2, o3, o4, o5}={3, 1, 9, 10, 2 }, K=2

K-centroid

- centroid: an actual object, representative object centrally located in a cluster

$$E = \sum_{i=1}^{K} \sum_{O \in C_i} d(o, centroid_i)$$

$$E = \frac{1}{n} \sum_{i=1}^{K} \sum_{O \in C_i} d(o, centroid_i)$$

- Groups n objects into k clusters by minimizing the E
- Find k centroids that minimize E
  - Brute-force algorithm – exhaustive search

K-medoid – exhaustive search

D = {o1, o2, o3, o4, o5}={3, 1, 9, 10, 2 }, K=2

Iteration	Centroids	Clustering	E
1	**3**, **1**	C1={**3**,9,10} C2={**1**,2}	13+1=14
2	**3**, **9**	C1={**3**,1,2} C2={**9**,10}	3+1=4
3	**3**,**10**	C1={**3**,1,2} C2={**10**,9}	3+1=4
4	**3**, **2**	…	…
…			
10			

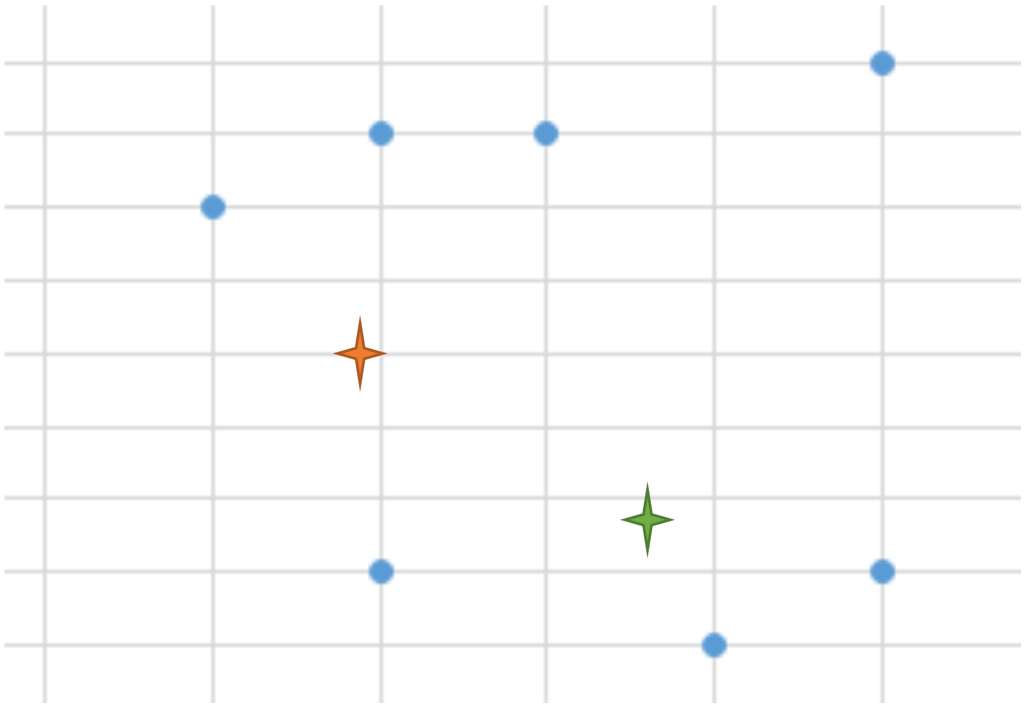
$O(C_n^k k(n-k))$
Global minimum

K-means (k=2)

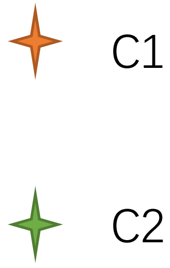
- Step1: Randomly select 2 centroids
- Step2: Assign each sample into the class represented by the closest centroids
- Step3：Update centroids as mean of the cluster
- Step4：Repeat step2 and step3 until convergence

K-means (k=2)

5	8
4	7
8	9
6	8
8	2
7	1
5	2

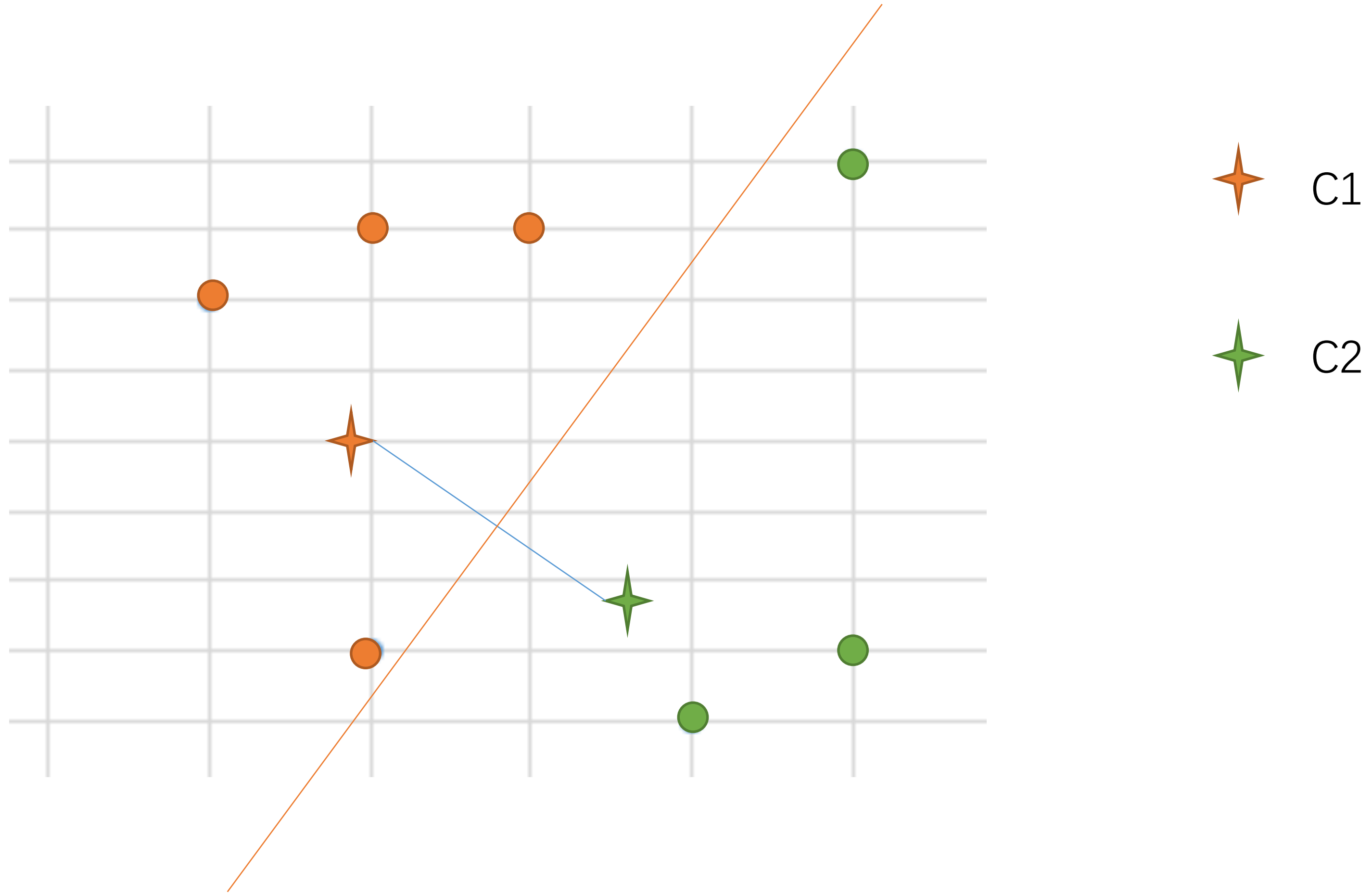


Randomly generate two medoids



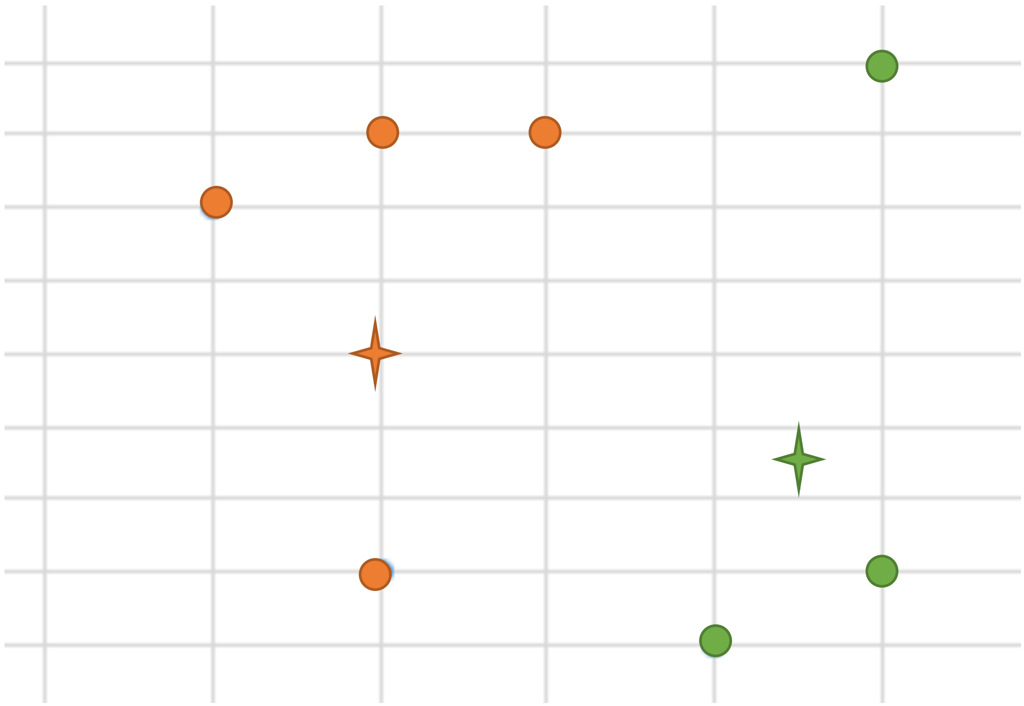
K-means (k=2)

5	8
4	7
8	9
6	8
8	2
7	1
5	2



K-means (k=2)

5	8
4	7
8	9
6	8
8	2
7	1
5	2

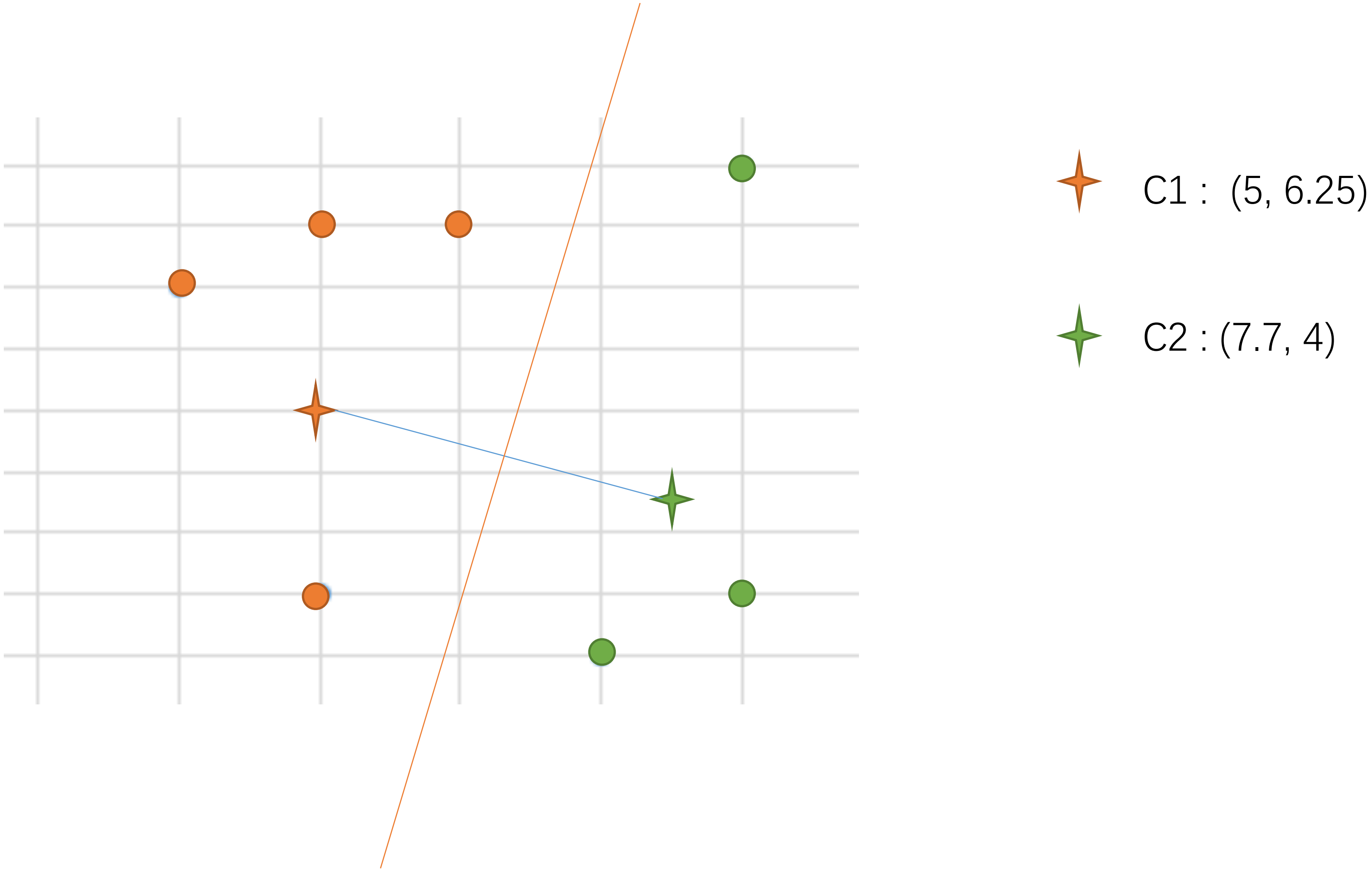


C1 : (5, 6.25)

C2 : (7.7, 4)

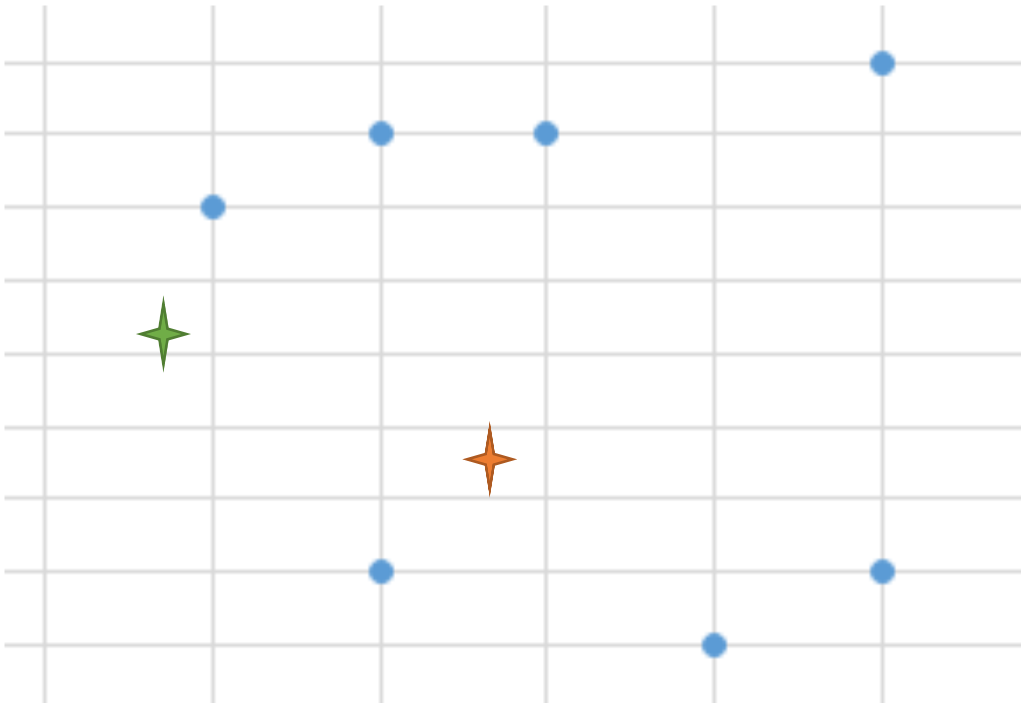
K-means (k=2)

5	8
4	7
8	9
6	8
8	2
7	1
5	2

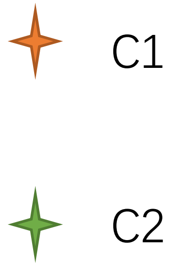


Different initial values K-means (k=2)

5	8
4	7
8	9
6	8
8	2
7	1
5	2

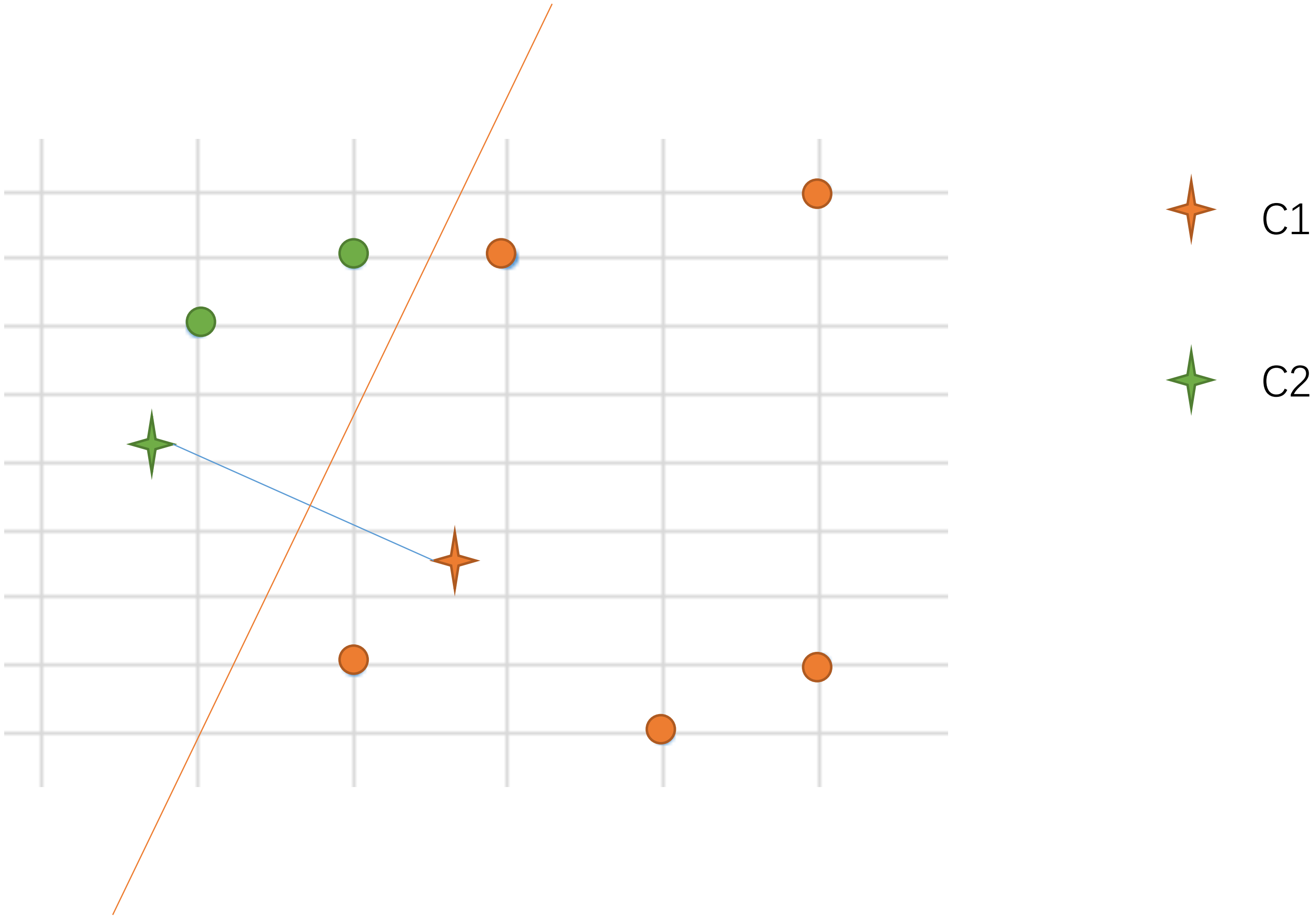


Randomly generate two medoids



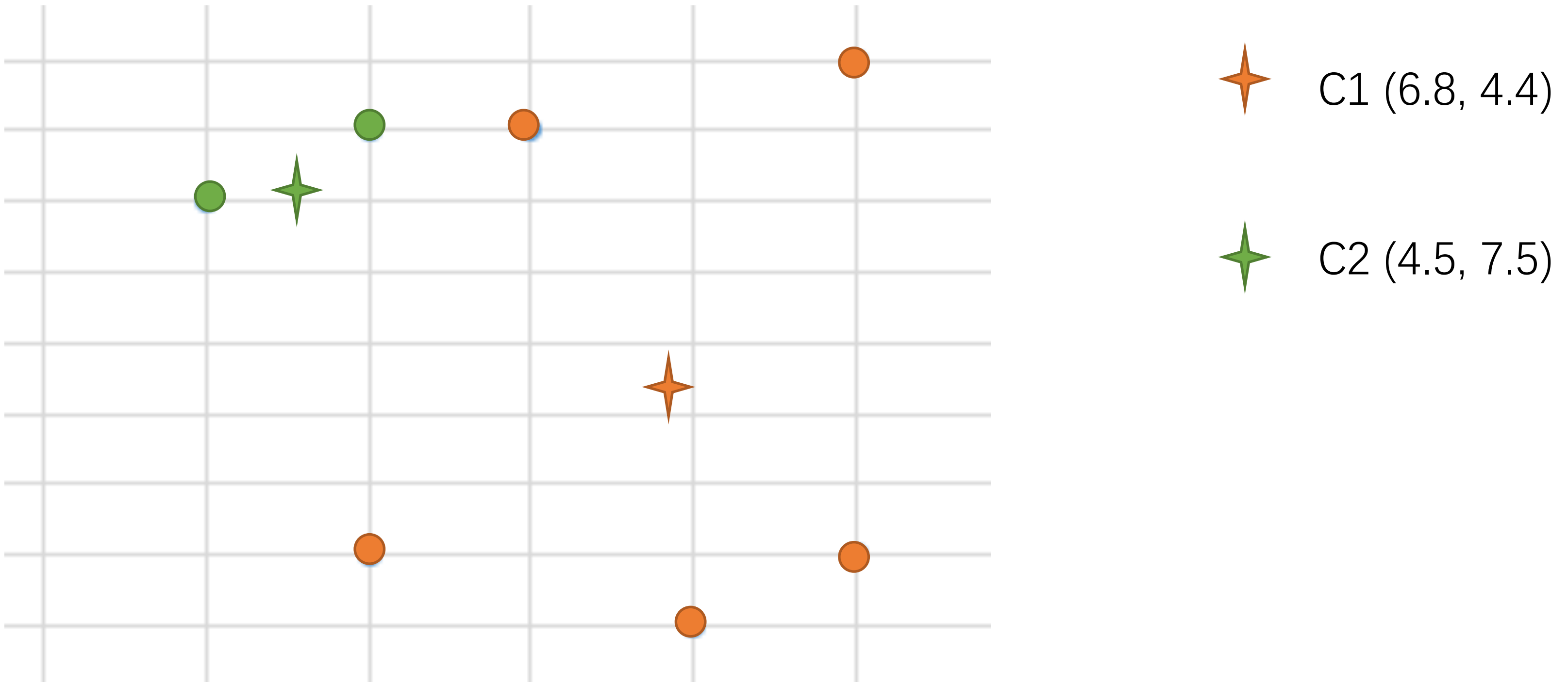
K-means (k=2)

5	8
4	7
8	9
6	8
8	2
7	1
5	2



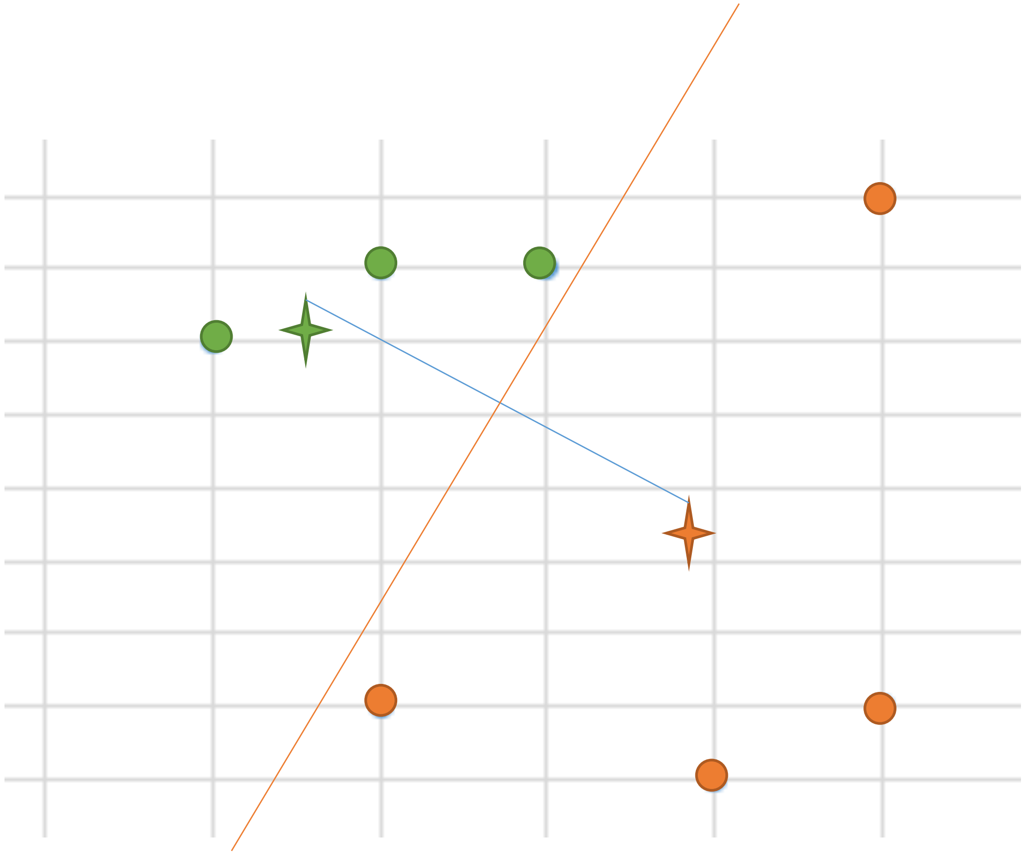
K-means (k=2)

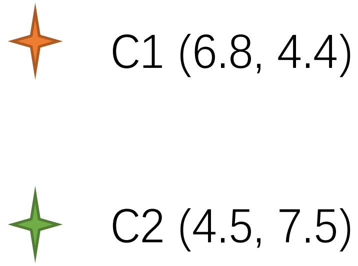
5	8
4	7
8	9
6	8
8	2
7	1
5	2



K-means (k=2)

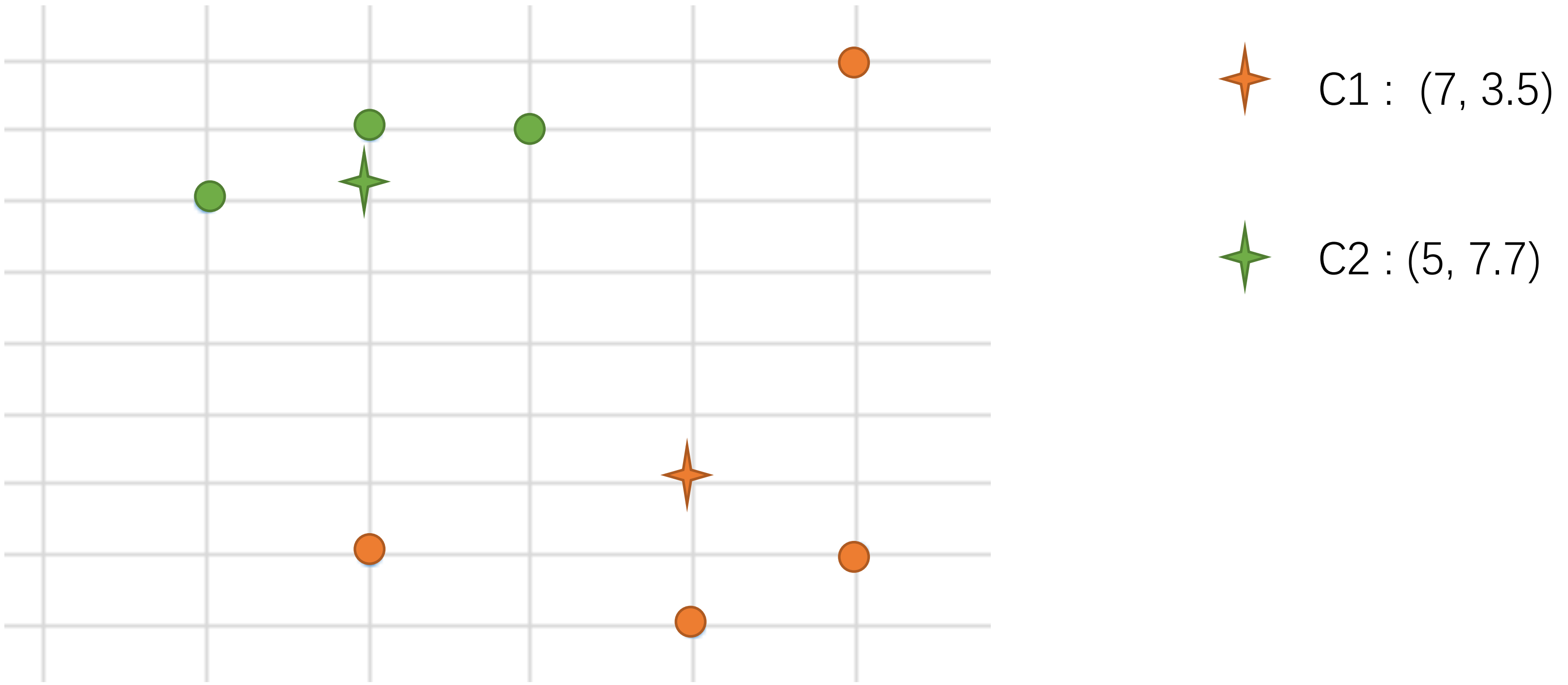
5	8
4	7
8	9
6	8
8	2
7	1
5	2





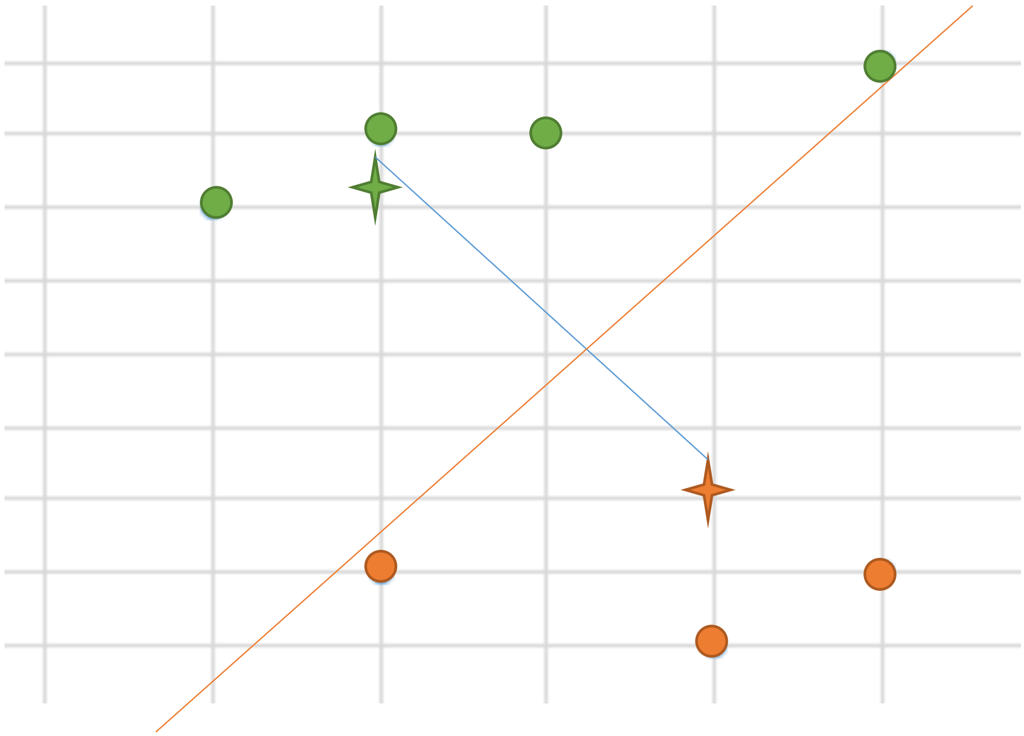
K-means (k=2)

5	8
4	7
8	9
6	8
8	2
7	1
5	2



K-means (k=2)

5	8
4	7
8	9
6	8
8	2
7	1
5	2

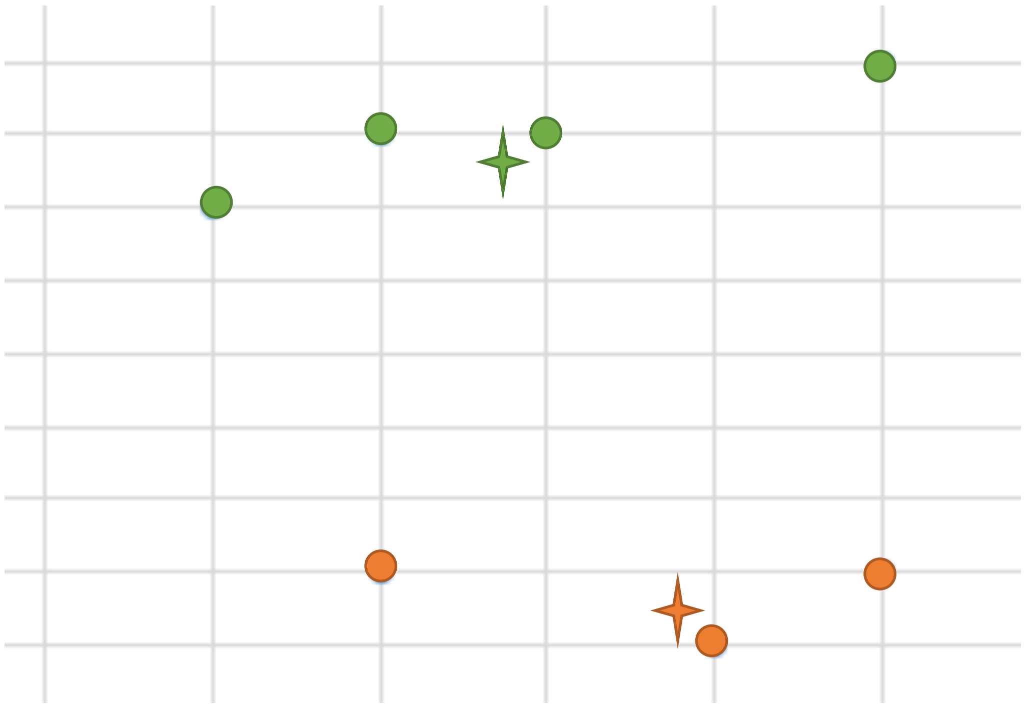


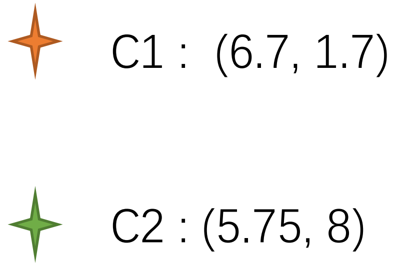
C1 : (7, 3.5)

C2 : (5, 7.7)

K-means (k=2)

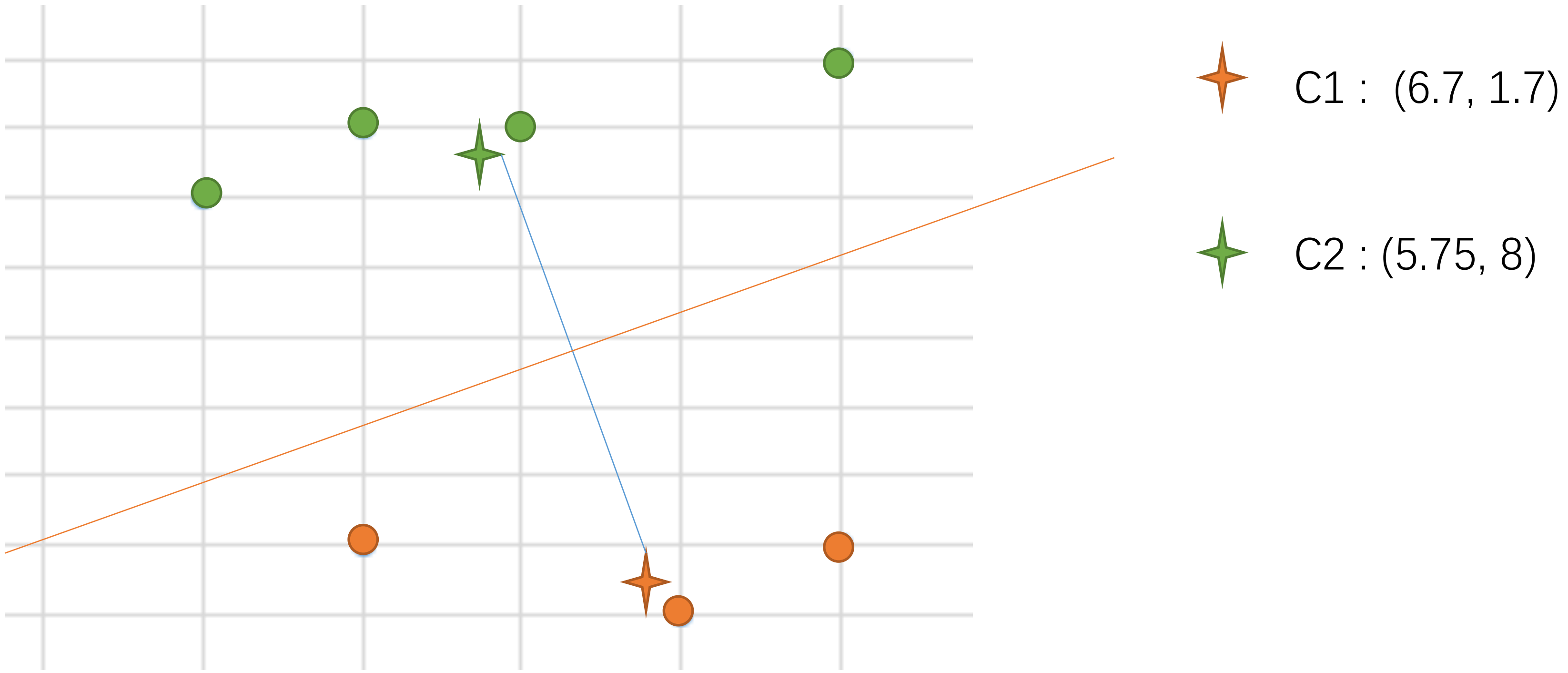
5	8
4	7
8	9
6	8
8	2
7	1
5	2





K-means (k=2)

5	8
4	7
8	9
6	8
8	2
7	1
5	2



K-Means Clustering Algorithm

Algorithm

Input

– Data + Desired number of clusters, K

Initialize

– the K cluster centers (randomly if necessary)

Iterate

1. Decide the class memberships of the n objects by assigning them to the nearest cluster centers

2. Re-estimate the K cluster centers (aka the centroid or mean), by assuming the memberships found above are correct.

Termination

– If none of the n objects changed membership in the last iteration, exit.

Otherwise go to 1.

K-means - summary

- K is user-defined
- Clustering is to find optimal solution. It is a NP hard problem
- K-means finds local optimal solution.
- Visualization:
- https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

Seed Selection

The results of the K- means Algorithm can vary based on random seed selection.

- Some seeds can result in **poor convergence rate**, or convergence to **sub-optimal** clustering.
- K-means algorithm can get stuck easily in **local minima.**
  – Select good seeds using a heuristic (e.g., object least similar to any existing mean)
  – Try out **multiple** starting points (very important!!!)
  – Initialize with the results of another method.

K-means Algorithm (more formally)

- **Randomly initialize k centers**

$$\mu^0 = (\mu_1^0, \ldots, \mu_K^0)$$

- **Classify**: At iteration t, assign each point j $\in$ {1,…,n} to nearest center:

$$C^t(j) \leftarrow \arg\min_i \|\mu_i^t - x_j\|^2$$

Classification at iteration *t*

- **Recenter**: μ_i is the centroid of the new sets:

$$\mu_i^{(t+1)} \leftarrow \arg\min_{\mu} \sum_{j:C^t(j)=i} \|\mu - x_j\|^2$$

Re-assign new cluster centers at iteration *t*

21

What is K-means optimizing?

- Define the following potential function F of centers μ and point allocation C

$$\mu = (\mu_1, \ldots, \mu_K)$$

$$C = (C(1), \ldots, C(n))$$

$$\left.\begin{aligned} F(\mu, C) &= \sum_{j=1}^{n} \|\mu_{C(j)} - x_j\|^2 \\ &= \sum_{i=1}^{K} \sum_{j:C(j)=i} \|\mu_i - x_j\|^2 \end{aligned}\right\} \text{Two equivalent versions}$$

- Optimal solution of the K-means problem:

$$\boxed{\min_{\mu, C} F(\mu, C)}$$

K-means Algorithm

Optimize the potential function:

$$\min_{\mu,C} F(\mu, C) = \min_{\mu,C} \sum_{j=1}^{n} \|\mu_{C(j)} - x_j\|^2 = \min_{\mu,C} \sum_{i=1}^{K} \sum_{j:C(j)=i} \|\mu_i - x_j\|^2$$

K-means algorithm:

(1) Fix μ, Optimize C

$$\min_{C(1),C(2),\ldots,C(n)} \sum_{j=1}^{n} \|\mu_{C(j)} - x_j\|^2 = \sum_{j=1}^{n} \underbrace{\min_{C(j)} \|\mu_{C(j)} - x_j\|^2}_{\text{Exactly first step}}$$

Assign each point to the nearest cluster center

(2) Fix C, Optimize μ

$$\min_{\mu_1,\ldots,\mu_K} \sum_{i=1}^{K} \sum_{j:C(j)=i} \|\mu_i - x_j\|^2 = \sum_{i=1}^{K} \underbrace{\min_{\mu_i} \sum_{j:C(j)=i} \|\mu_i - x_j\|^2}_{\text{Exactly 2nd step (re-center)}}$$

K-means Algorithm cont...

Optimize the potential function:

$$\min_{\mu,C} F(\mu, C) = \min_{\mu,C} \sum_{j=1}^{n} \|\mu_{C(j)} - x_j\|^2$$

K-means algorithm: (coordinate descent on F)

(1) Fix μ, Optimize C **Expectation step**

(2) Fix C, Optimize μ **Maximization step**

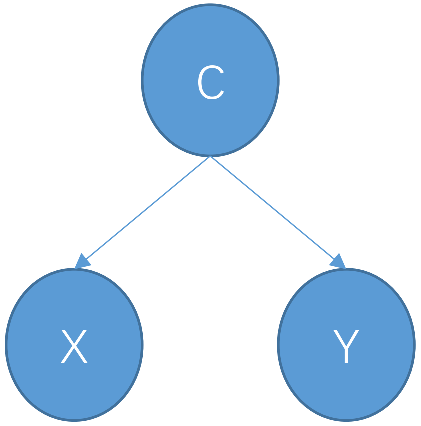
Today, we will see a generalization of this approach:

EM algorithm

From a Bayes Network perspective

X	Y	C
5	8	?
4	7	?
8	9	?
6	8	?
8	2	?
7	1	?
5	2	?

- Clustering is a Bayes Network Problem
- Known structure, partly observed data

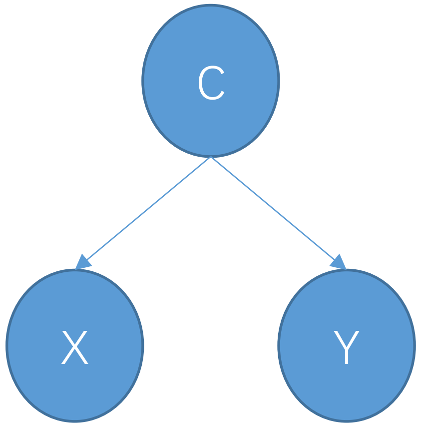


K-means under Bayes Network

X	Y	C
5	8	1
4	7	1
8	9	1
6	8	0
8	2	0
7	1	0
5	2	0

- Randomly generate two centroids
- The two centroids give a label to each sample
- Each iteration, it gives a 0 or 1 to variable C

$$p(C = 1|X = 5, Y = 8) = 1$$



To measure probability of C value

X	Y	C
5	8	?
4	7	?
8	9	?
6	8	?
8	2	?
7	1	?
5	2	?

- It's natural that measuring it using a real number [0..1]
- Major Problem, to compute

$$p(C = 1|X = 5, Y = 8)$$

- Need to know

$$p(X = 5, Y = 8|C = 1)$$

- But C values are missing in training set
- Minor Problem, features are continuous number