جامعة الإسكندرية
ALEXANDRIA UNIVERSITY
كلية الحاسبات وعلوم البيانات

LEC.5.

Introduction to Data Science

Advanced Analytical Theory and Methods

Cluster Analysis using K-Means

What is Cluster ?

A collection of data objects:

- They are similar (or related) to one another within the same group.
- And dissimilar (or unrelated) to the objects in other groups.

What is Clustering ?

- Clustering is the process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters.

- Dissimilarities and similarities are assessed based on the attribute values describing the objects and often involve distance measures.

What is Cluster Analysis?

- Cluster analysis (or *clustering*, *data segmentation*, …) is a method often used for exploratory analysis of the data.

- In clustering, there are no predictions made.

- Clustering methods find the similarities between objects according to the object attributes and group the similar objects into clusters.

What is Cluster Analysis? …… con.

- Clustering is known as *unsupervised learning* because the class label information is not present. For this reason, clustering is a form of learning by observation.

- Clustering techniques are utilized in marketing, economics, and various branches of science.

- A popular clustering method is k-means

What is Unsupervised Learning?

Learning useful structure *without* labeled classes, optimization criterion, feedback signal, or any other information beyond the raw data

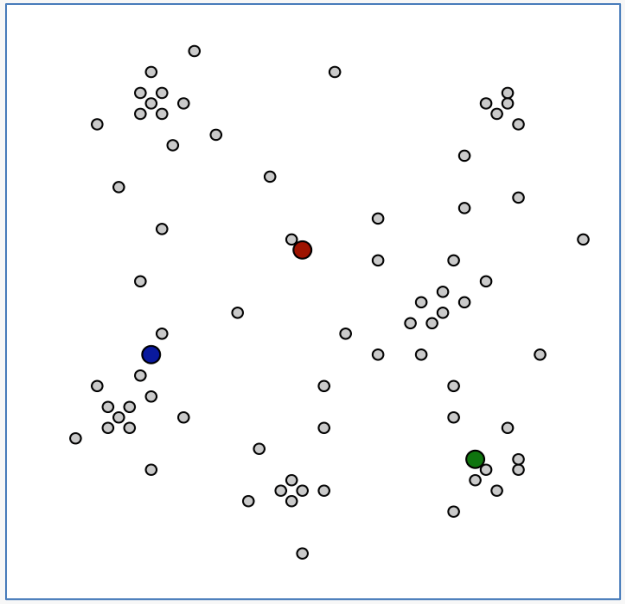
K-Means Clustering - What is it?

- Used for clustering numerical data, usually a set of measurements about objects of interest.
- Input: numerical. There must be a distance metric defined over the variable space.
  - Euclidian distance
- Output: The centers of each discovered cluster, and the assignment of each input datum to a cluster.
  - Centroid

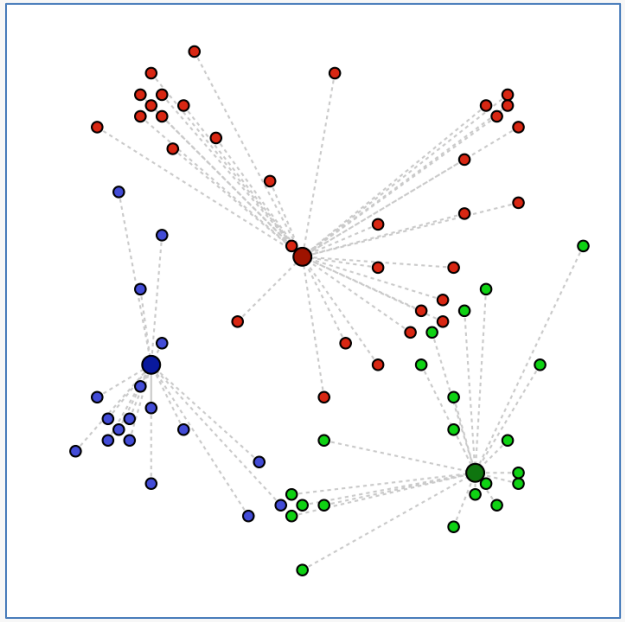
The Algorithm

1. Choose *K*; then select K random "centroids"

   In our example, K=3



2. Assign records to the cluster with the closest centroid



The Algorithm (Continued)

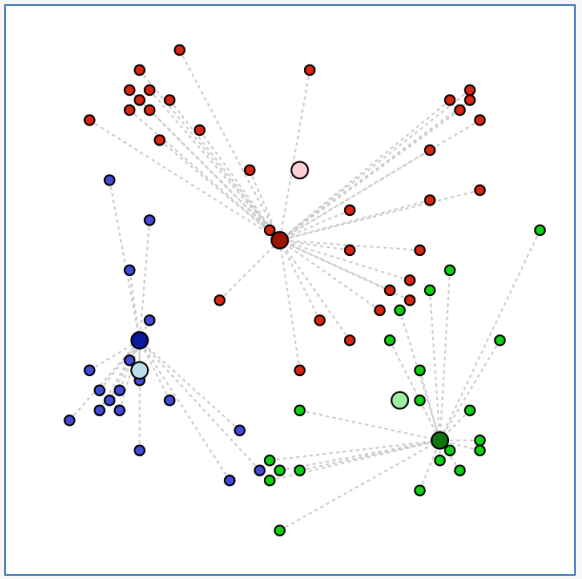
3. Recalculate the resulting centroids

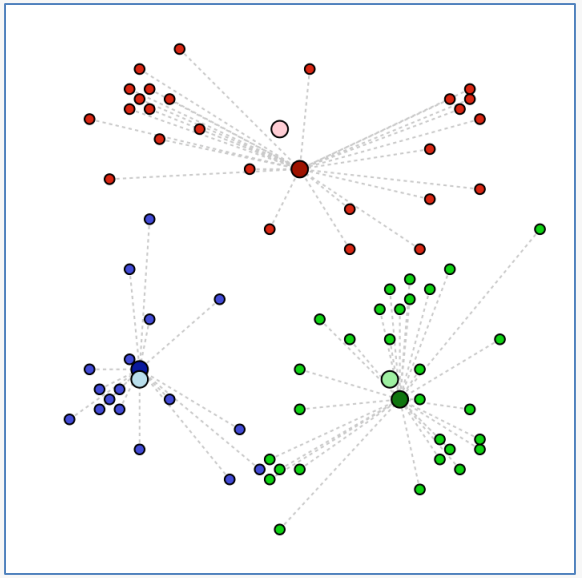
   Centroid: the mean value of all the records in the cluster

4. Repeat steps 2 & 3 until record assignments no longer change

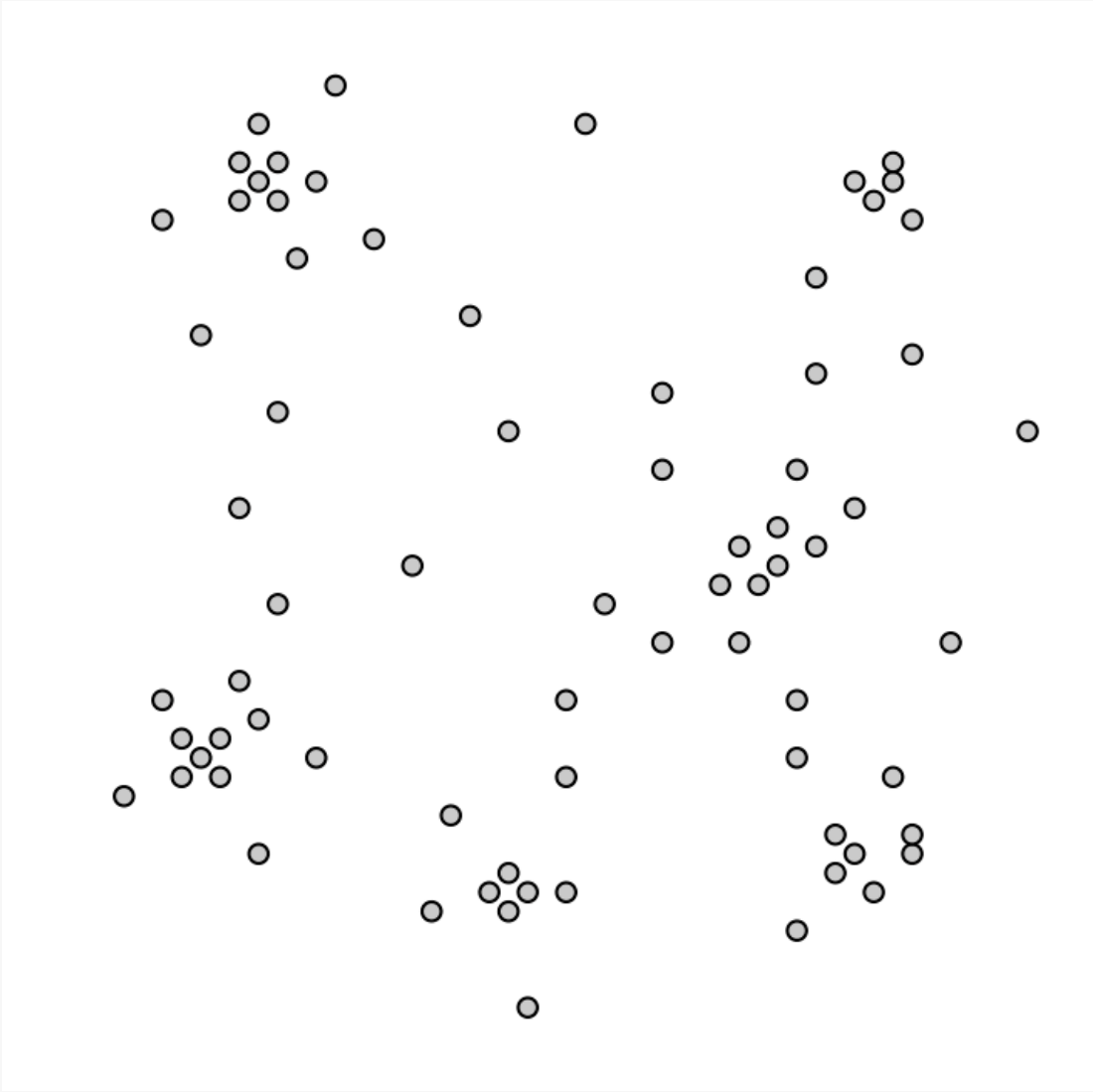
Model Output:

- The final cluster centers
- The final cluster assignments of the training data

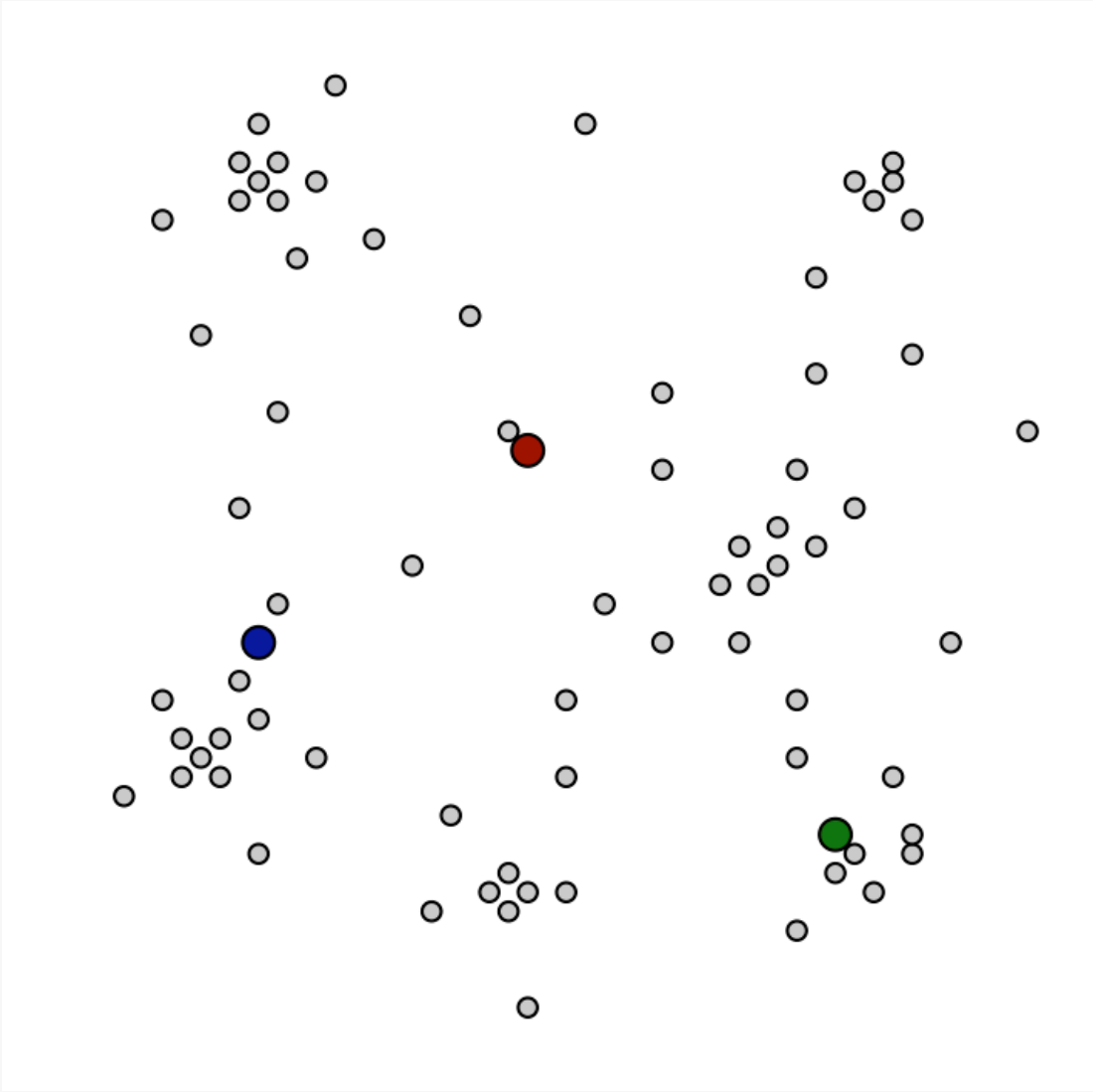




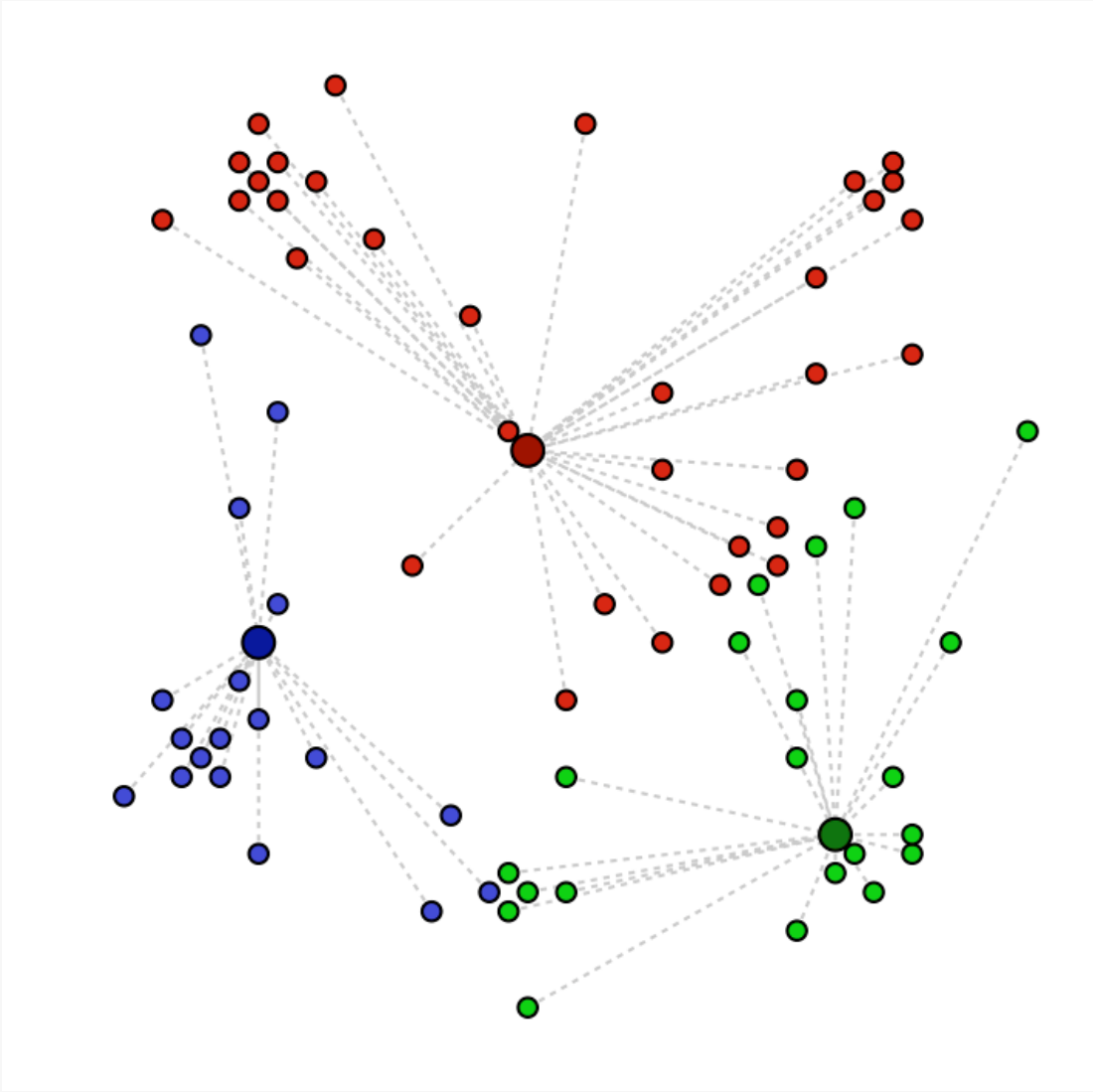
The Algorithm



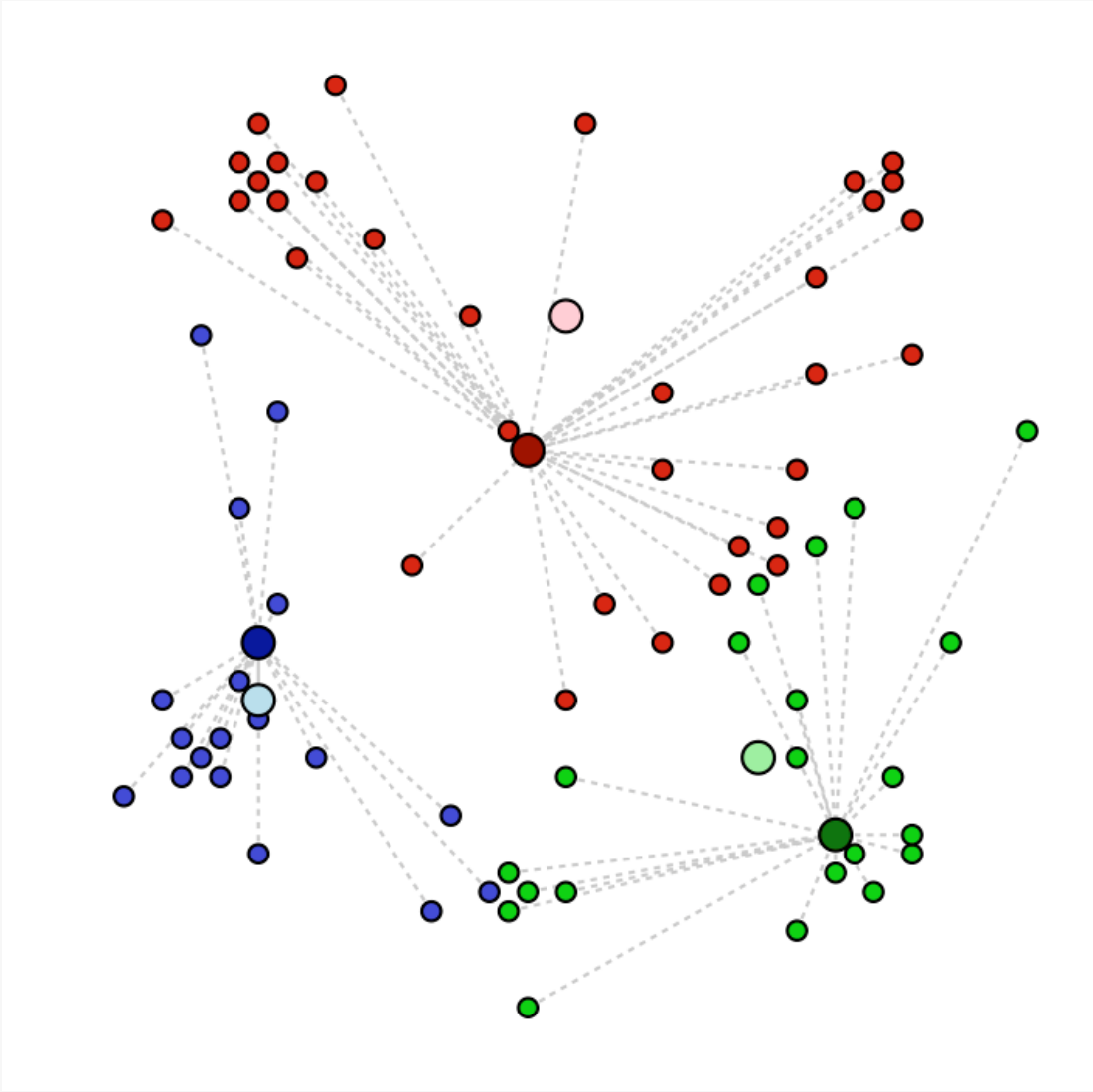
The Algorithm



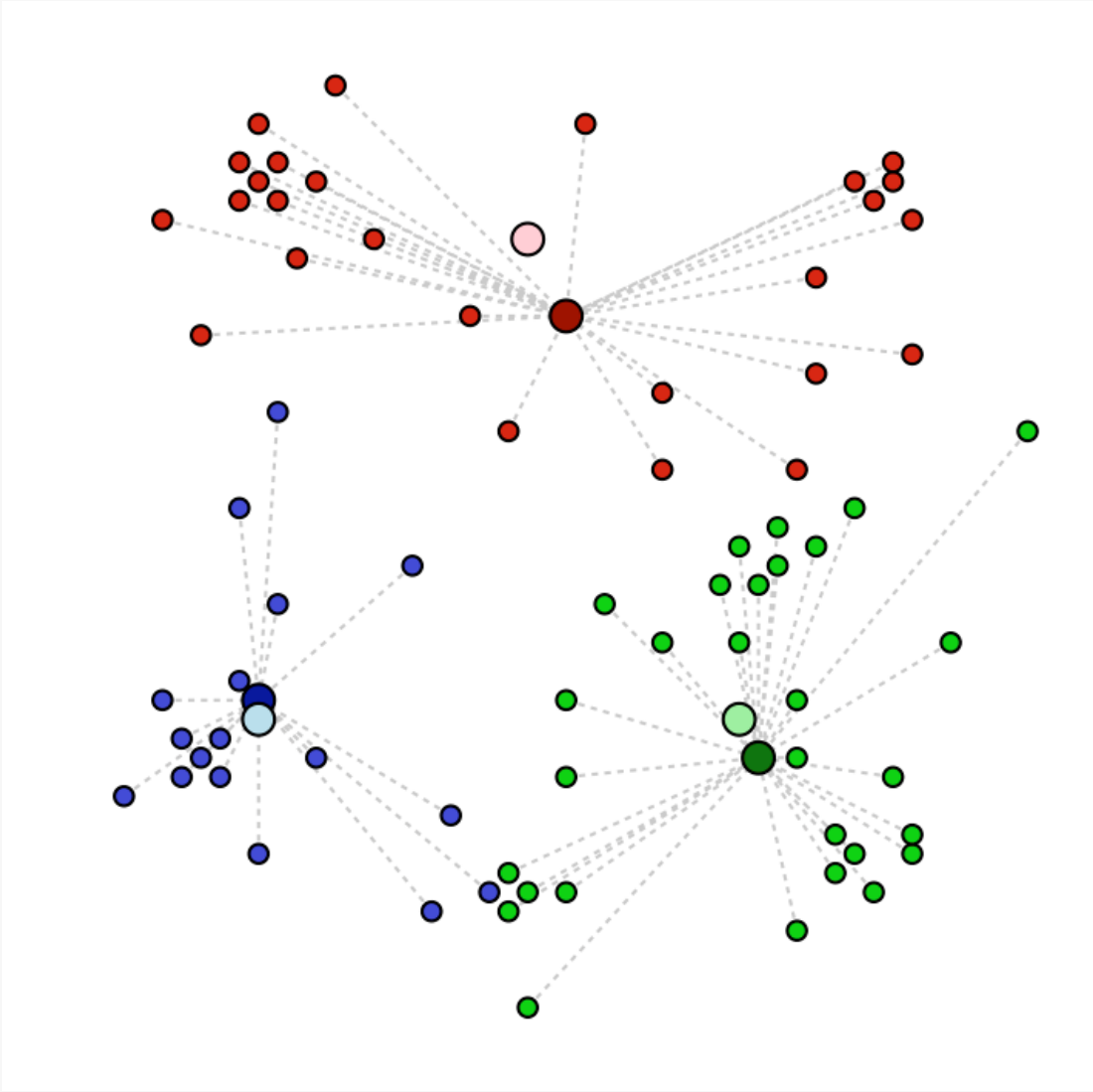
The Algorithm



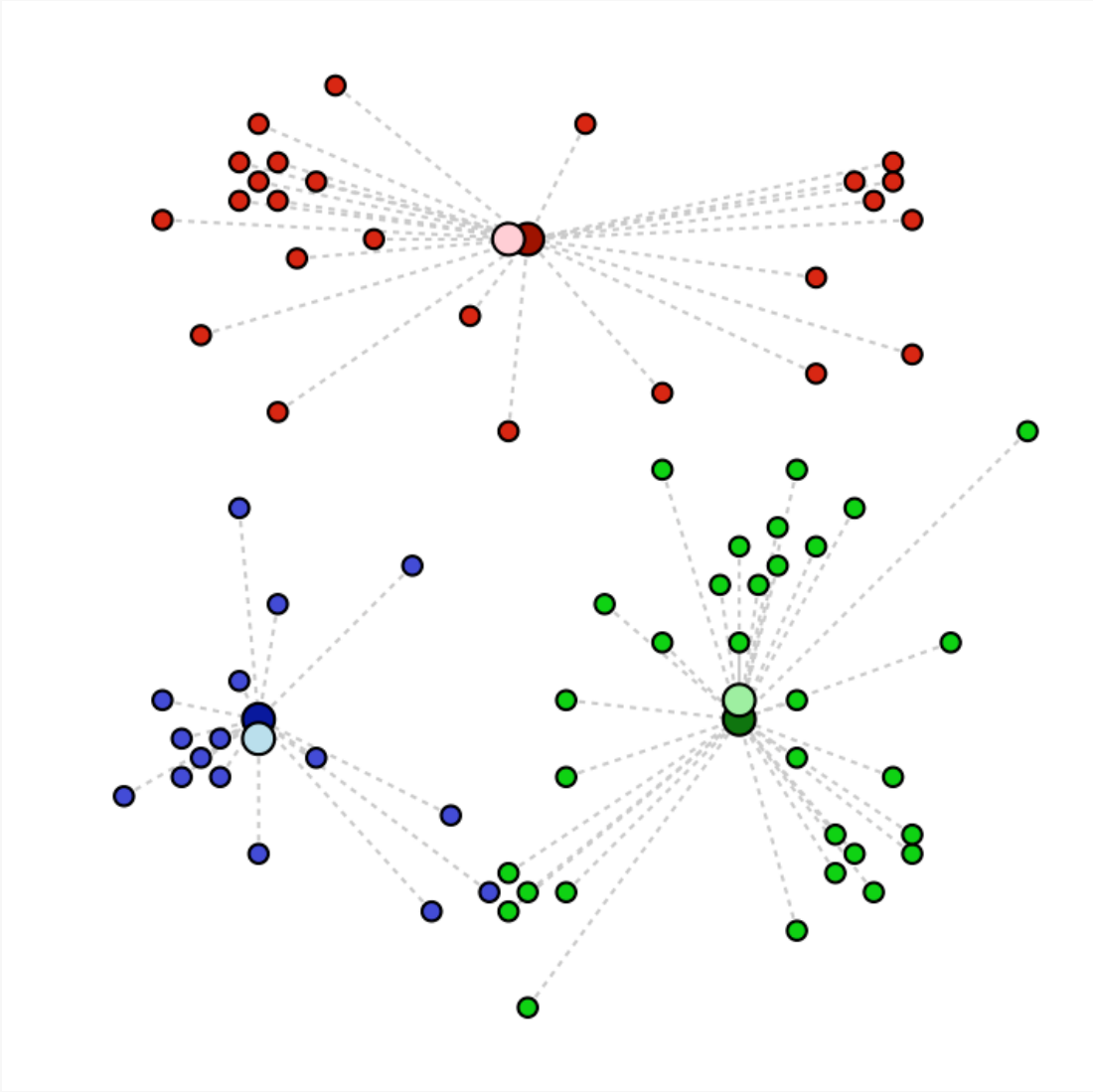
The Algorithm



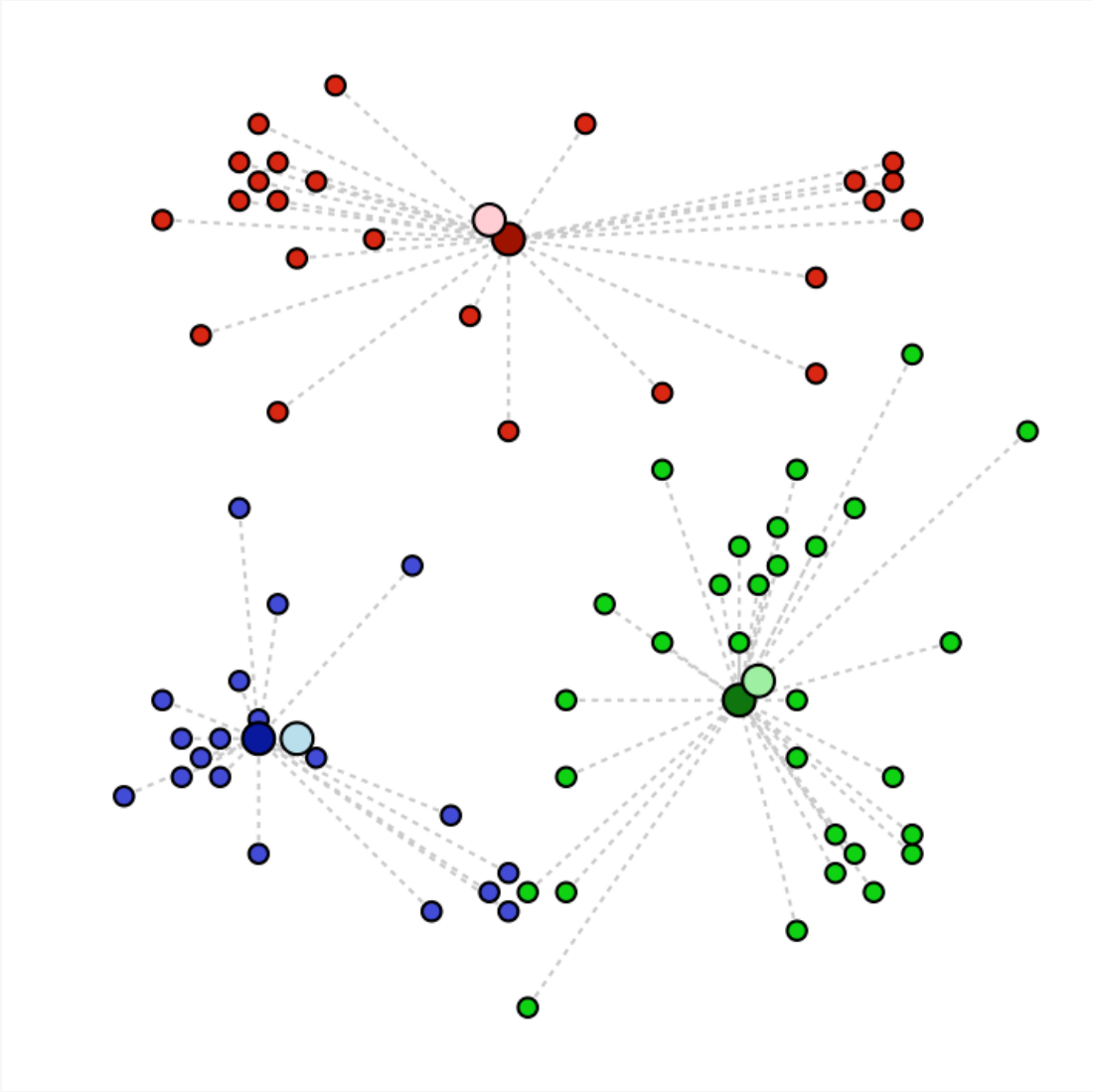
The Algorithm



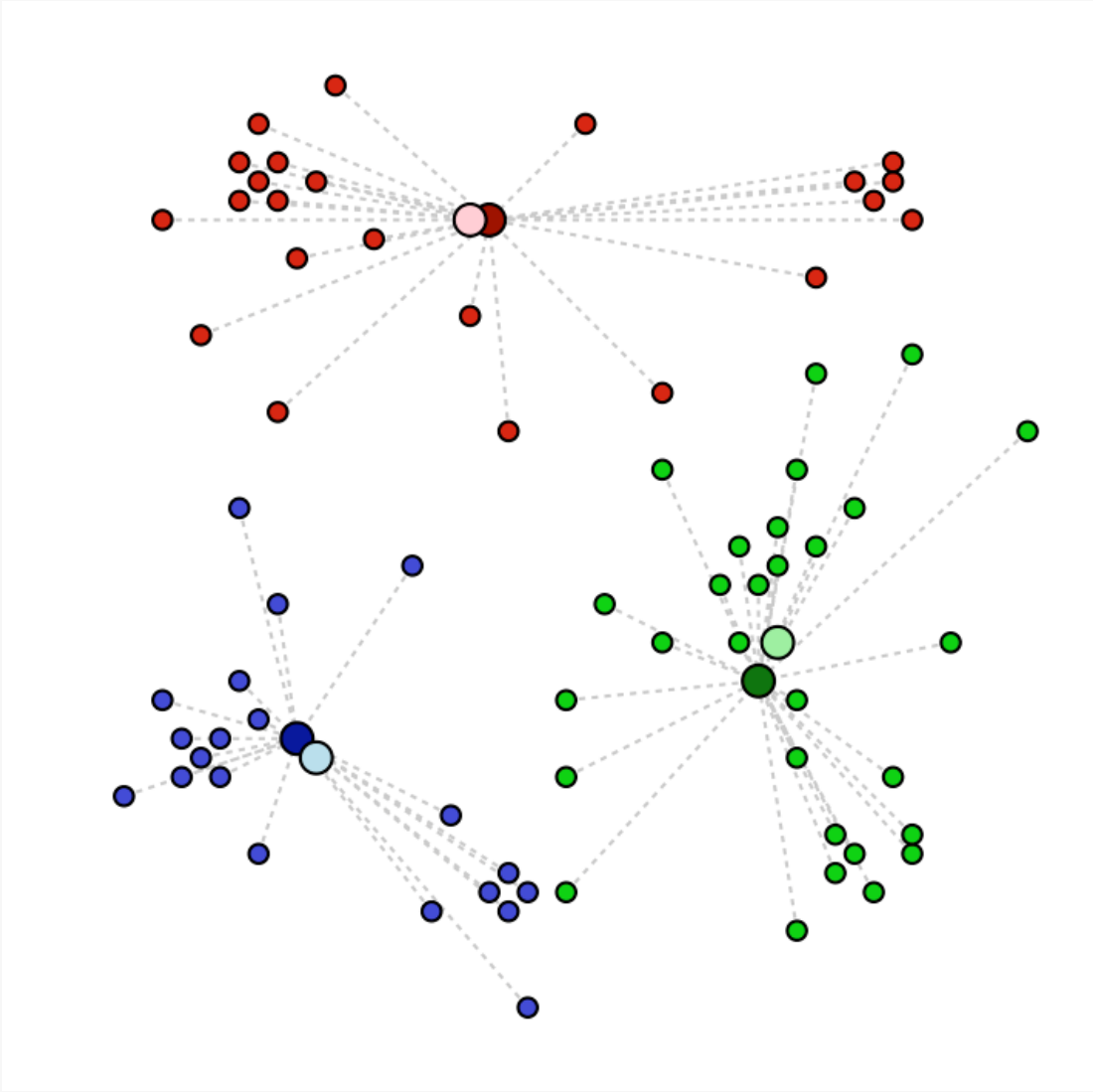
The Algorithm



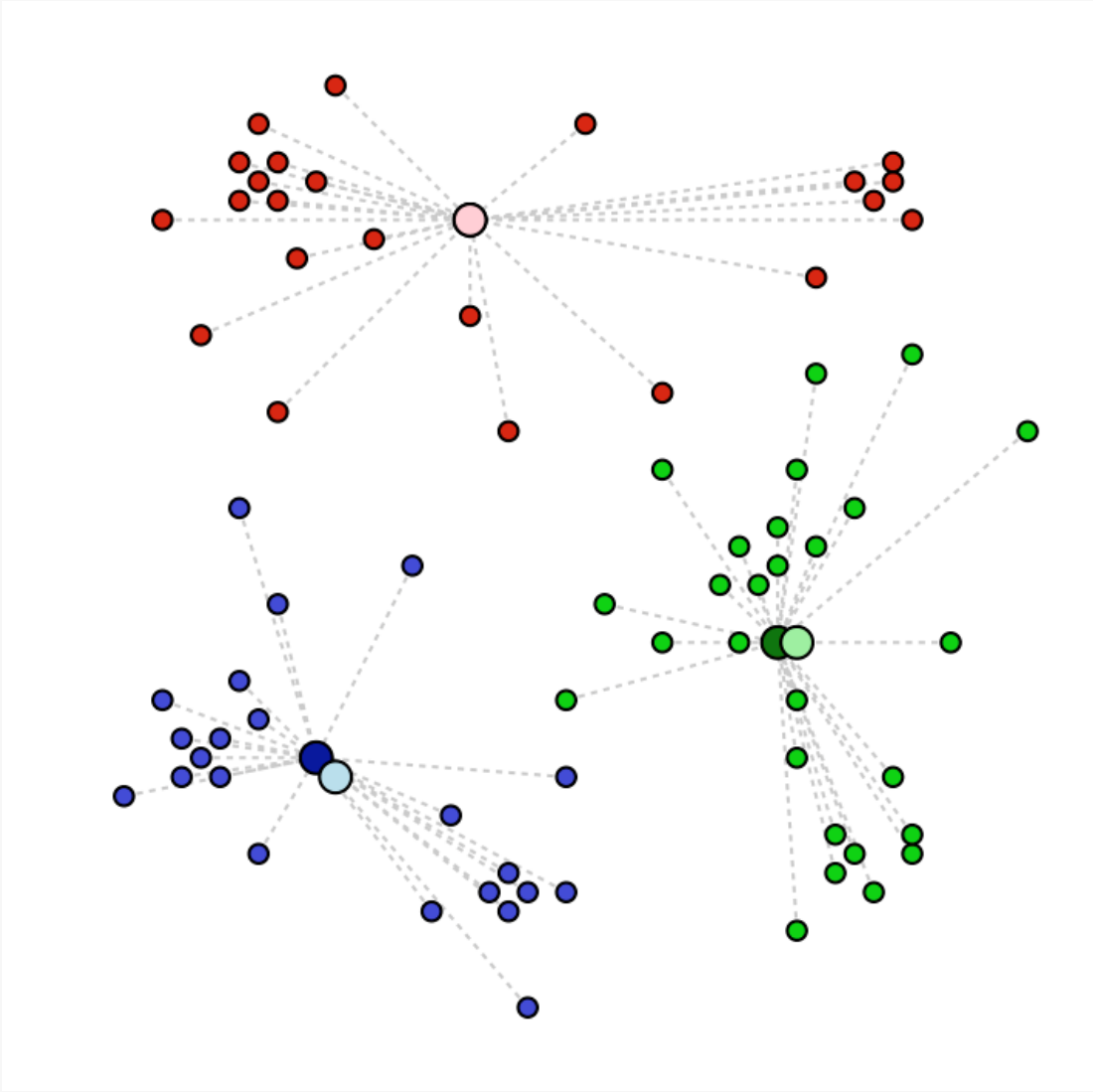
The Algorithm



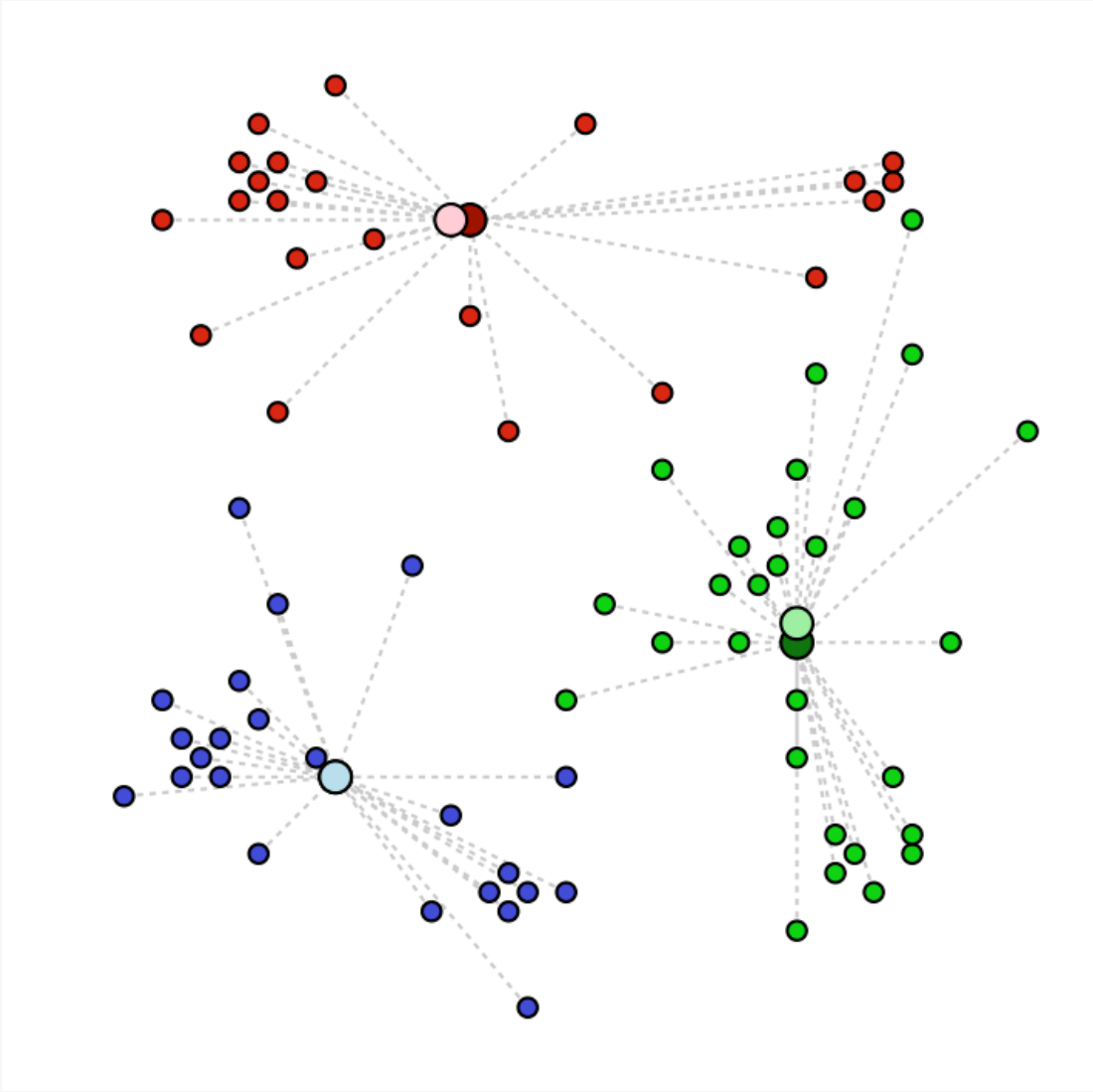
The Algorithm



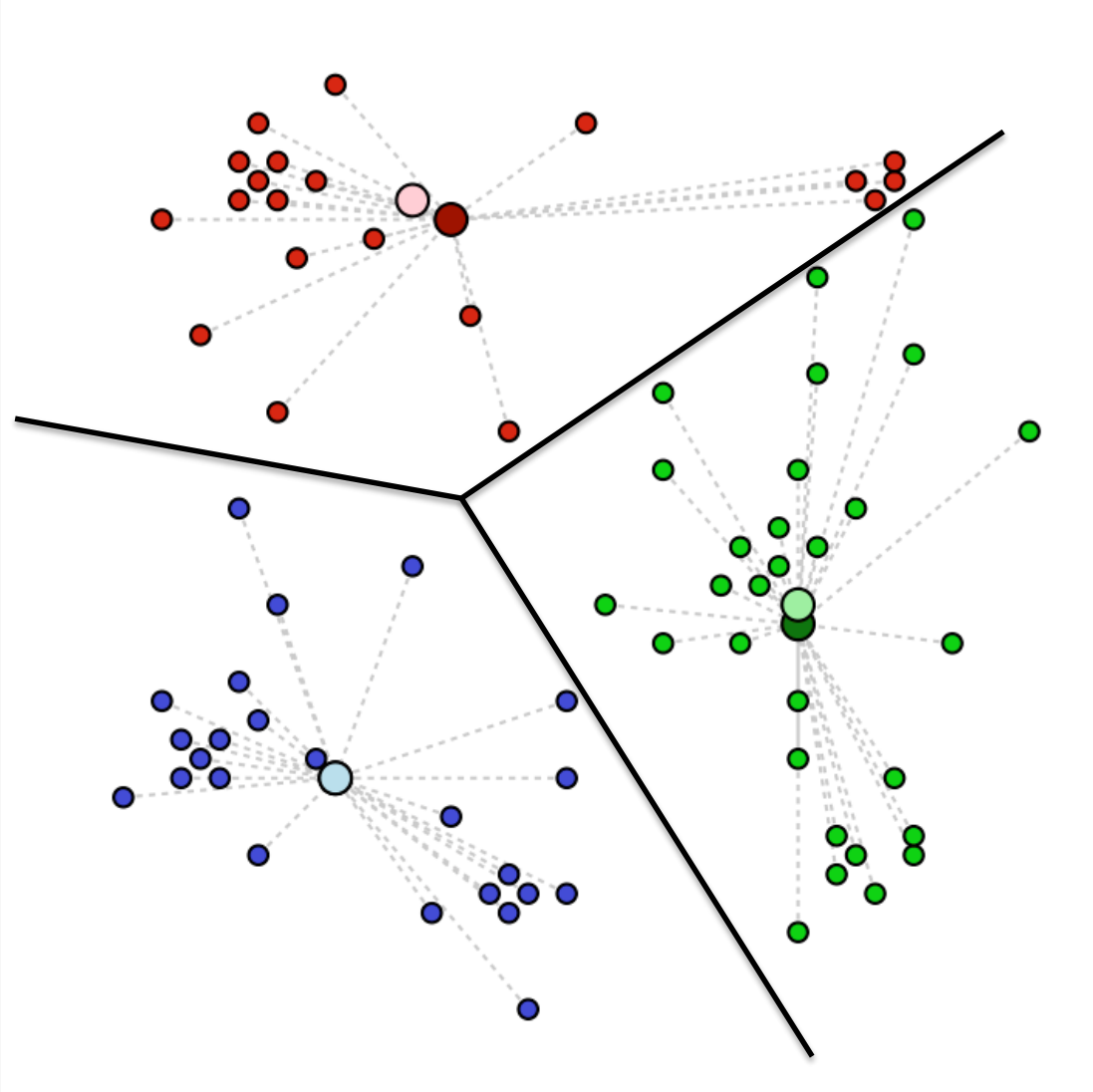
The Algorithm



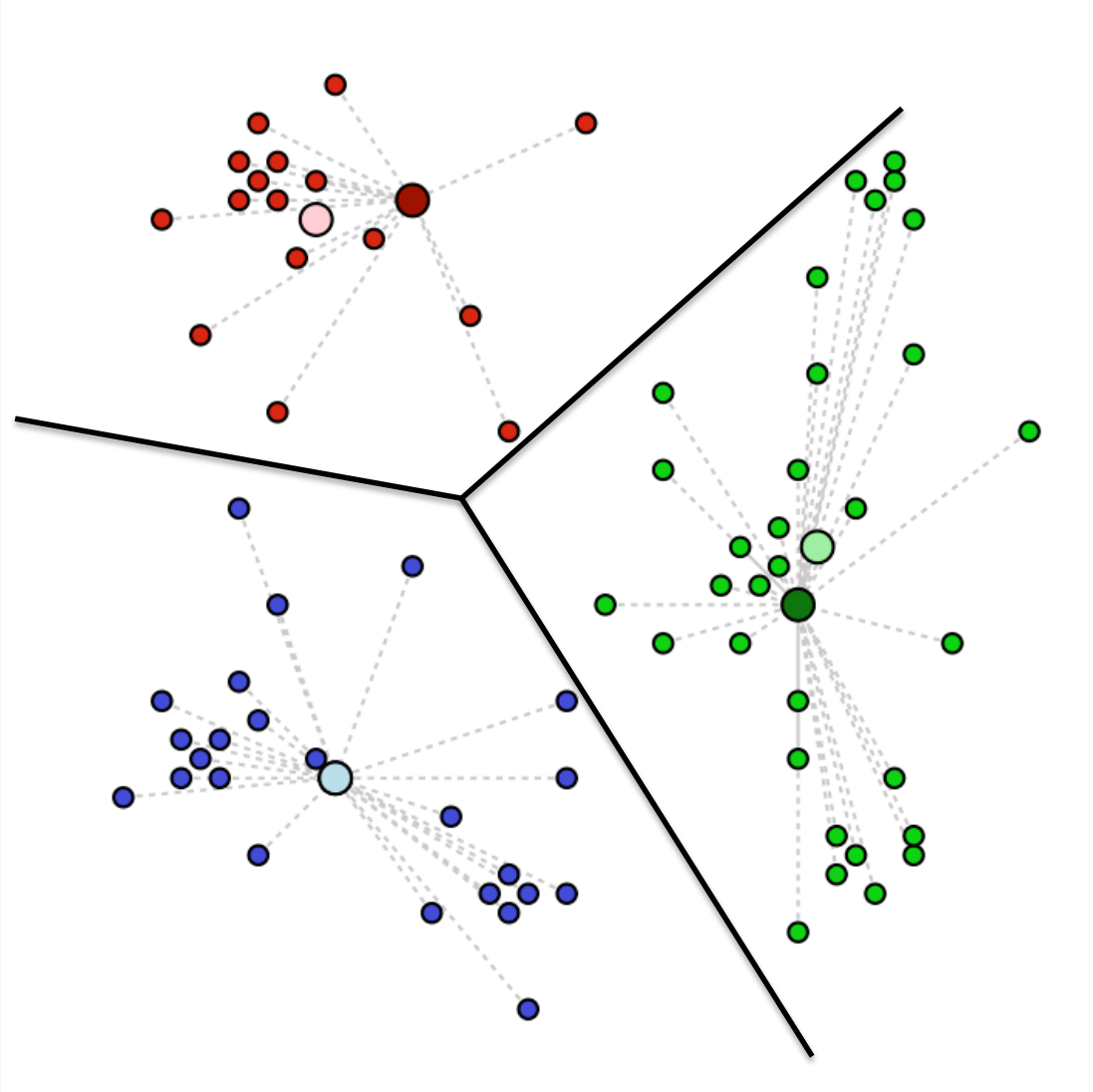
The Algorithm



The Algorithm



The Algorithm



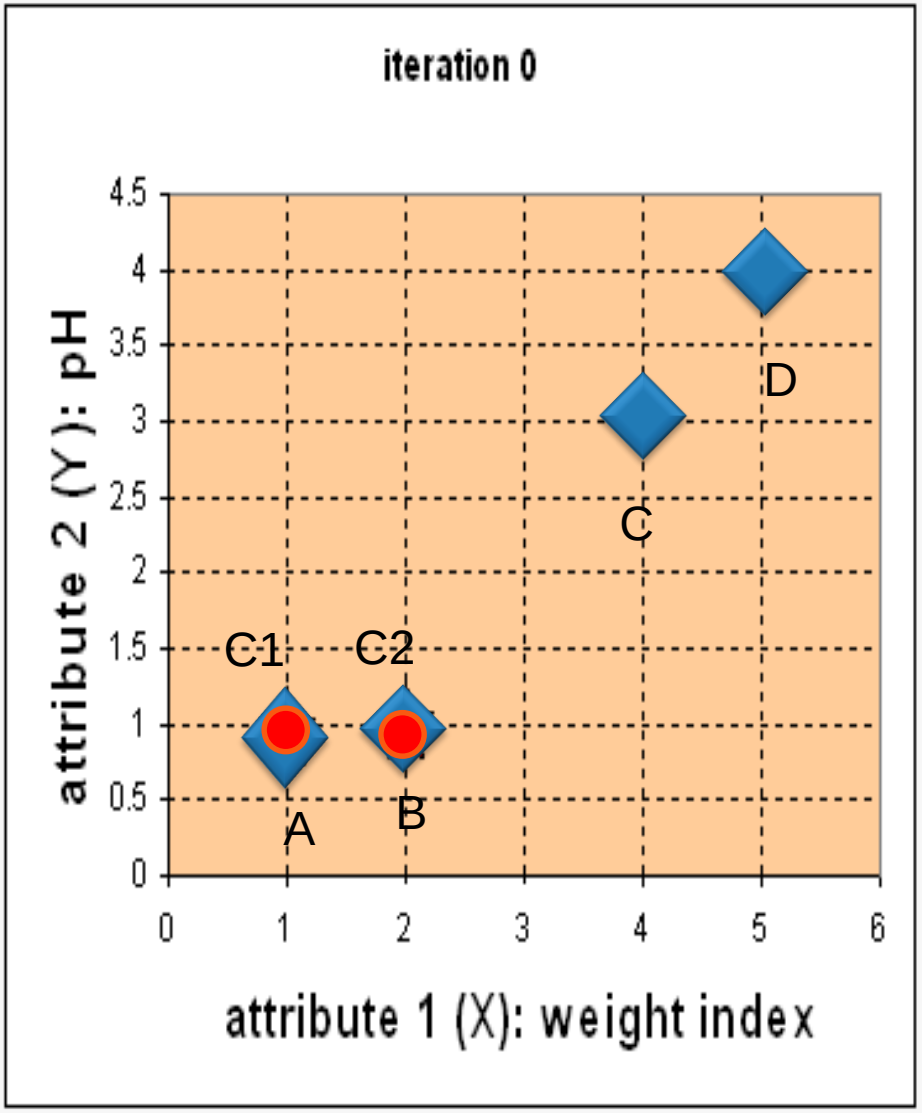
Real-Life Numerical Example of K-Means Clustering

We have 4 medicines as our training data points object and each medicine has 2 attributes. Each attribute represents coordinate of the object. We have to determine which medicines belong to cluster 1 and which medicines belong to the other cluster.

Object	Attribute1 (X): weight index	Attribute 2 (Y): pH
Medicine A	1	1
Medicine B	2	1
Medicine C	4	3
Medicine D	5	4

Step 1:

- **Initial value of centroids** : Suppose we use medicine A and medicine B as the first centroids.
- Let and c_1 and c_2 denote the coordinate of the centroids, then c_1=(1,1) and c_2=(2,1)



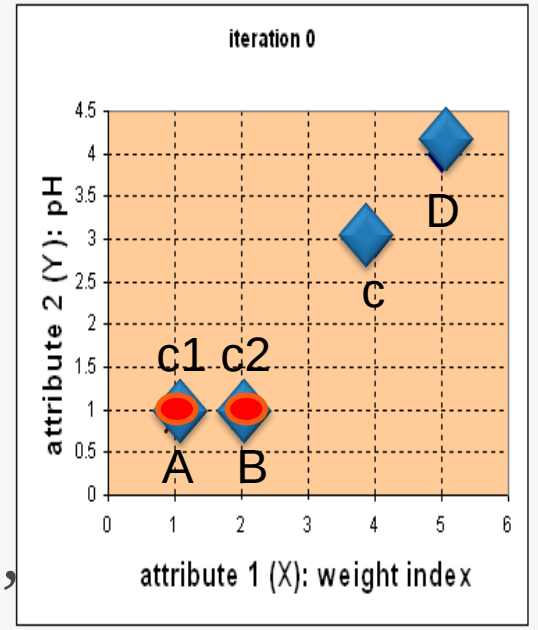
- **Objects-Centroids distance** :
- we calculate the distance between cluster centroid to each object.

Let us use Euclidean distance,

Distsnce = squr. $((x1-x2)^2 + (y1-y2)^2)$

- Then we calculate distance matrix
- Each column in the distance matrix symbolizes the object.
- The first row of the distance matrix corresponds to the distance of each object to the first centroid and the second row is the distance of each object to the second centroid.

- For example:
- distance from medicine C = (4, 3) to the first centroid $\mathbf{c}_1 = (1,1)$ is , $\sqrt{(4-1)^2+(3-1)^2} = 3.61$
- and distance to the second centroid $\mathbf{c}_2 = (2,1)$,
- Is $\sqrt{(4-2)^2+(3-1)^2} = 2.83$ etc.



Distance matrix →

$$\mathbf{D}^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \quad \begin{matrix} \mathbf{c}_1 = (1,1) & group-1 \\ \mathbf{c}_2 = (2,1) & group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

Original matrix →

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad \begin{matrix} X \\ Y \end{matrix}$$

Distance matrix →

$$D^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \quad \begin{matrix} c_1 = (1,1) & group-1 \\ c_2 = (2,1) & group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

Original matrix →

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad \begin{matrix} X \\ Y \end{matrix}$$

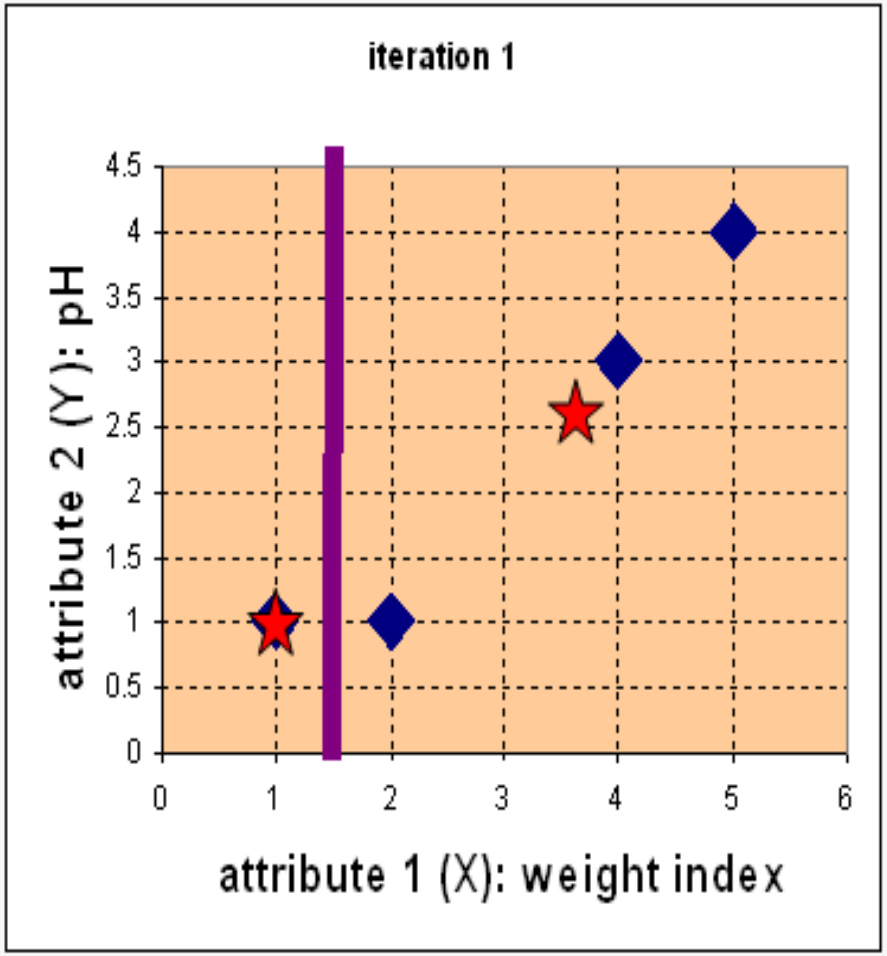
New Groups

C1 group	C2 Group
A	B
	C
	D

Step 2:

- **Objects clustering** : We assign each object based on the minimum distance.
- Medicine A is assigned to group 1, medicine B to group 2, medicine C to group 2 and medicine D to group 2.
- The elements of Group matrix below is 1 if and only if the object is assigned to that group.

$$\mathbf{G}^0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \begin{matrix} group-1 \\ group-2 \end{matrix}$$
$$A \quad B \quad C \quad D$$



- **<u>Iteration-1, Objects-Centroids distances</u>** :
- The next step is to compute the distance of all objects to the new centroids.

- Now we calculate the new centroids coordinate based on the clustering of previous iteration. Group1 has one member and group 2 has three members, thus the new centroids are :
- C1= (1,1) the same point
- C2 = ((2+4+5)/3),((1+3+4)/3)

$$c_1 = (1,1) \quad group-1$$
$$c_2 = \left(\frac{11}{3}, \frac{8}{3}\right) \quad group-2$$

- **Iteration-1, Objects-Centroids distances** :

- Similar to step 2, we have distance matrix at iteration 1 is

$$D^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} \quad \begin{matrix} c_1 = (1,1) & group-1 \\ c_2 = (\frac{11}{3}, \frac{8}{3}) & group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

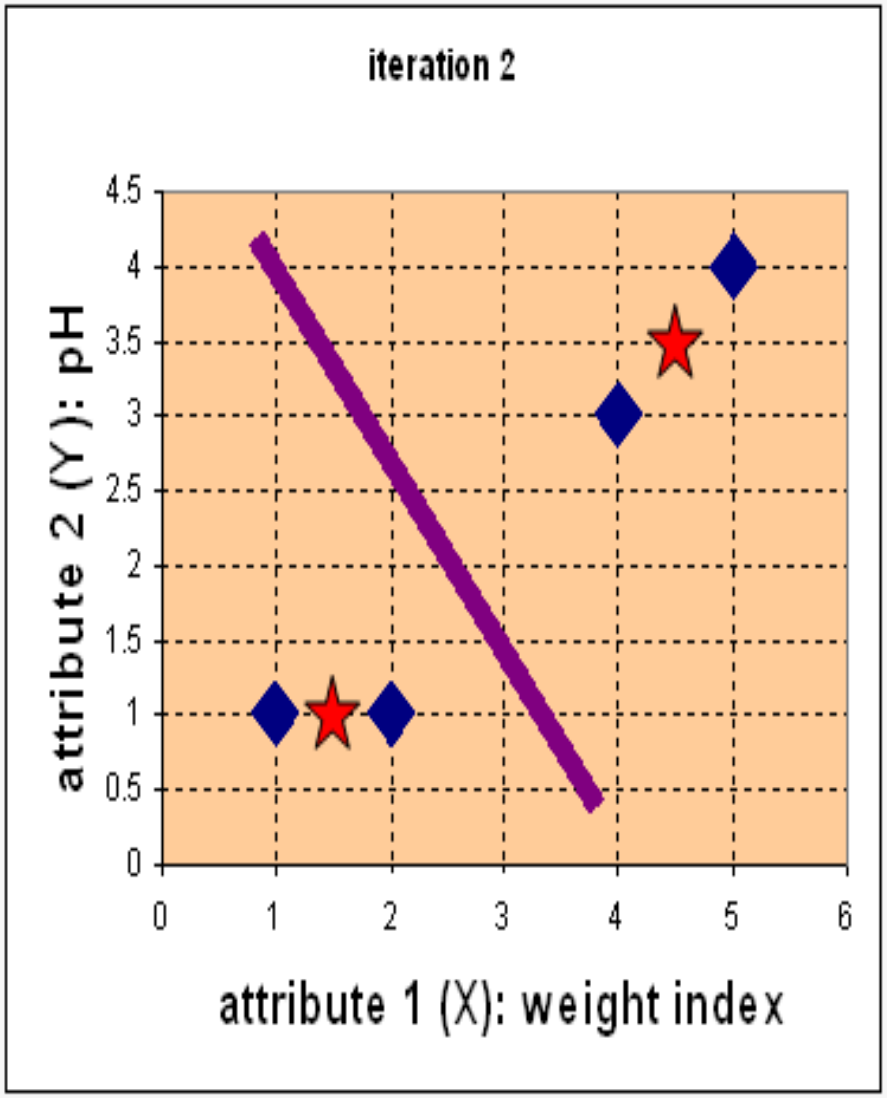
$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \begin{matrix} X \\ Y \end{matrix}$$

- **Iteration-1, Objects clustering:** Based on the new distance matrix, we move the medicine B to Group 1 while all the other objects remain. The Group matrix is shown below

$$G^1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad \begin{matrix} group-1 \\ group-2 \end{matrix}$$
$$A \quad B \quad C \quad D$$

- **Iteration 2, determine centroids:** Now we repeat step 4 to calculate the new centroids coordinate based on the clustering of previous iteration. Group1 and group 2 both has two members, thus the new centroids are $c_1 = (\frac{1+2}{2}, \frac{1+1}{2}) = (1\frac{1}{2}, 1)$

and $c_2 = (\frac{4+5}{2}, \frac{3+4}{2}) = (4\frac{1}{2}, 3\frac{1}{2})$



- **Iteration-2, Objects-Centroids distances** :
  - Repeat step 2 again, we have new distance matrix at iteration 2 as

$$\mathbf{D}^2 = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix} \quad \begin{matrix} \mathbf{c}_1 = (1\frac{1}{2}, 1) \quad group-1 \\ \mathbf{c}_2 = (4\frac{1}{2}, 3\frac{1}{2}) \quad group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \quad \begin{matrix} X \\ Y \end{matrix}$$

- **<u>Iteration-2, Objects clustering:</u>** Again, we assign each object based on the minimum distance.

$$\mathbf{G}^2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{matrix} group-1 \\ group-2 \end{matrix}$$
$$A \quad B \quad C \quad D$$

- We obtain result that $\mathbf{G}^2 = \mathbf{G}^1$. Comparing the grouping of last iteration and this iteration reveals that the objects does not move group anymore.

- Thus, the computation of the k-mean clustering has reached its stability and no more iteration is needed..

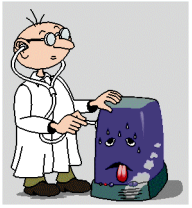
We get the final grouping as the results as:

Object	Feature1(X): weight index	Feature2 (Y): pH	Group (result)
Medicine A	1	1	1
Medicine B	2	1	1
Medicine C	4	3	2
Medicine D	5	4	2

Weaknesses of K-Mean Clustering

1. When the numbers of data are not so many, initial grouping will determine the cluster significantly.
2. The number of cluster, K, must be determined before hand. Its disadvantage is that it does not yield the same result with each run, since the resulting clusters depend on the initial random assignments.
3. We never know the real cluster, using the same data, because if it is inputted in a different order it may produce different cluster if the number of data is few.
4. It is sensitive to initial condition. Different initial condition may produce different result of cluster. The algorithm may be trapped in the *local optimum*.

Diagnostics – Evaluating the Model



- Do the clusters look separated in at least some of the plots when you do pair-wise plots of the clusters?
  - Pair-wise plots can be used when there are not many variables
- Do you have any clusters with few data points?
  - Try decreasing the value of K
- Are there splits on variables that you would expect, but don't see?
  - Try increasing the value K
- Do any of the centroids seem too close to each other?
  - Try decreasing the value of K

K-Means Clustering - Reasons to Choose (+) and Cautions (-)

Reasons to Choose (+)	Cautions (-)
Easy to implement	Doesn't handle categorical variables
Easy to assign new data to existing clusters Which is the nearest cluster center?	Sensitive to initialization (first guess)
Concise output Coordinates the K cluster centers	Variables should all be measured on similar or compatible scales Not scale-invariant!
	K (the number of clusters) must be known or decided a priori Wrong guess: possibly poor results
	Tends to produce "round" equi-sized clusters. Not always desirable

Check Your Knowledge

1. Why do we consider K-means clustering as a unsupervised machine learning algorithm?
2. Detail the four steps in the K-means clustering algorithm.
3. What is the most common measure of distance used with K-means clustering algorithms?