University of Roehampton London

Machine Learning

Dr Changjiang He, Dr Kuo-Ming Chao
Computer Science| School of Art
University of Roehampton

University of Roehampton London

Lesson 7.2

K-Means

Contents

What is K-means

University of Roehampton London

- Let us understand the K-means clustering algorithm with its simple definition.

- *A K-means clustering algorithm tries to group similar items in the form of clusters. The number of groups is represented by K.*

What is K-means

University of Roehampton London

- Let’s take an example. Suppose you went to a vegetable shop to buy some vegetables. There you will see different kinds of vegetables.
- The one thing you will notice there that the vegetables will be arranged in a group of their types. Like all the carrots will be kept in one place, potatoes will be kept with their kinds and so on.
- If you will notice here then you will find that they are forming a group or cluster, where each of the vegetables is kept within their kind of group forming the clusters.

What is K-means

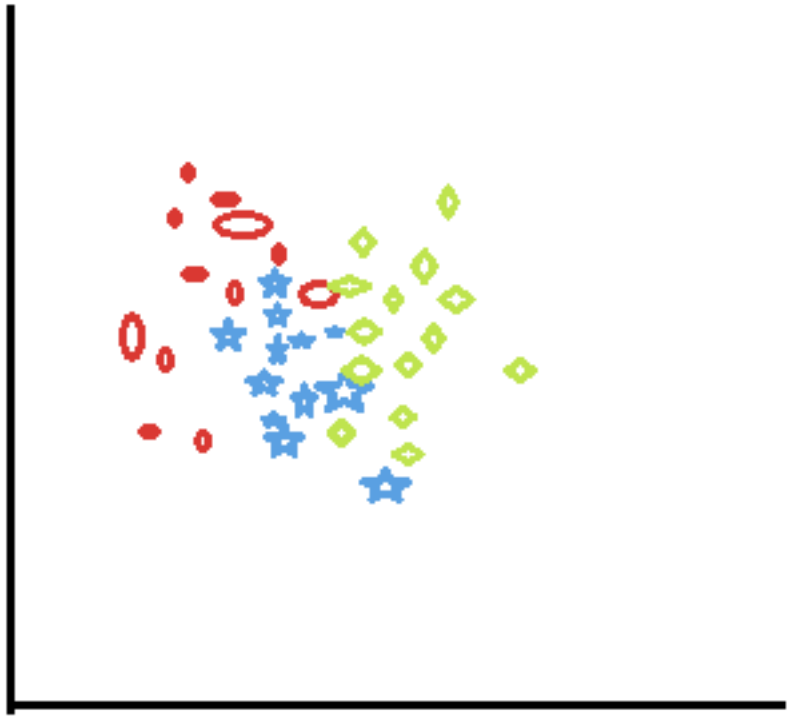


fig 1: before applying k-means clustering

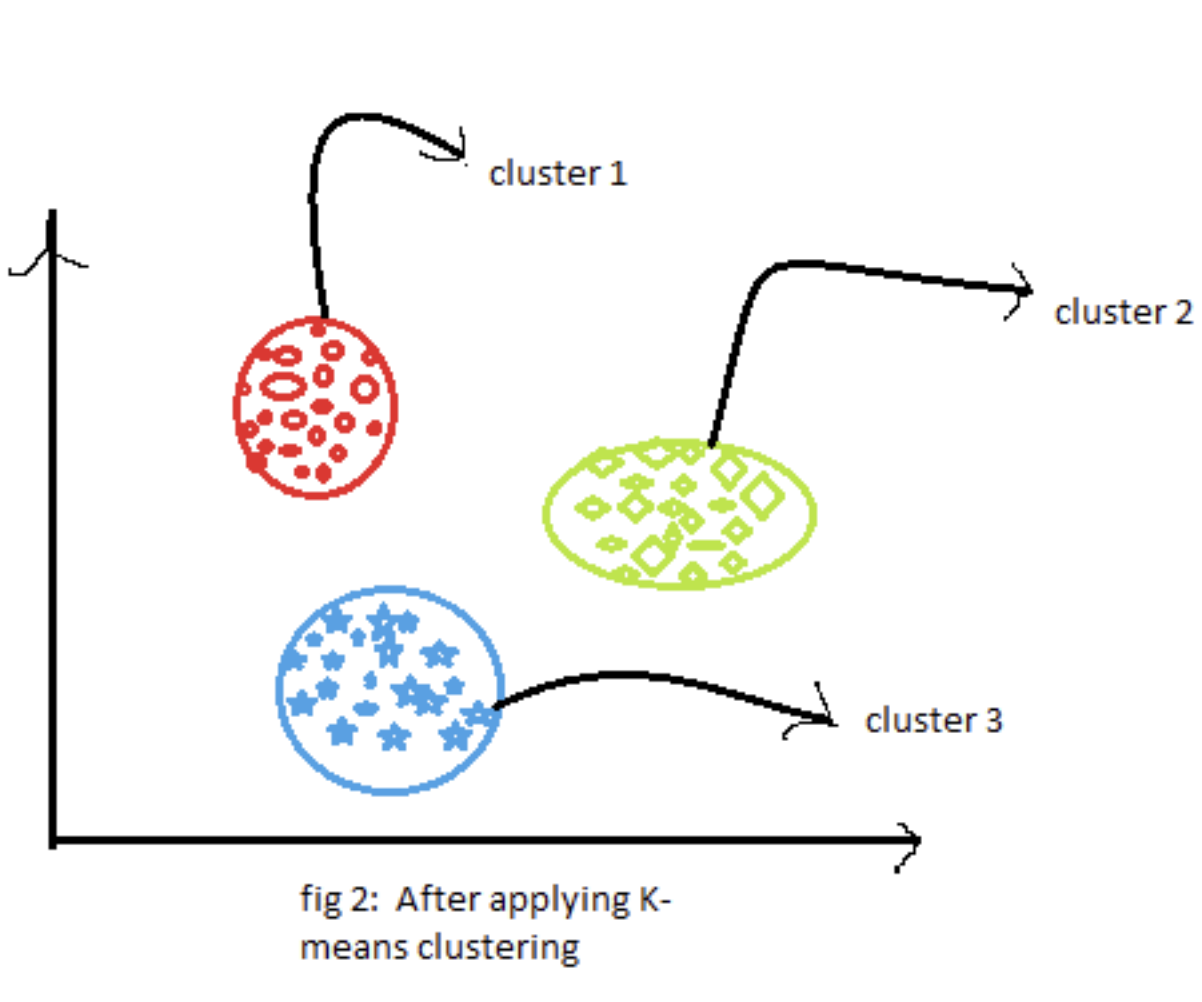


fig 2: After applying K-means clustering

How k-means works

K-means clustering algorithm works in three steps.

- Select the k values.
- Initialize the centroids.
- Select the group and find the average.

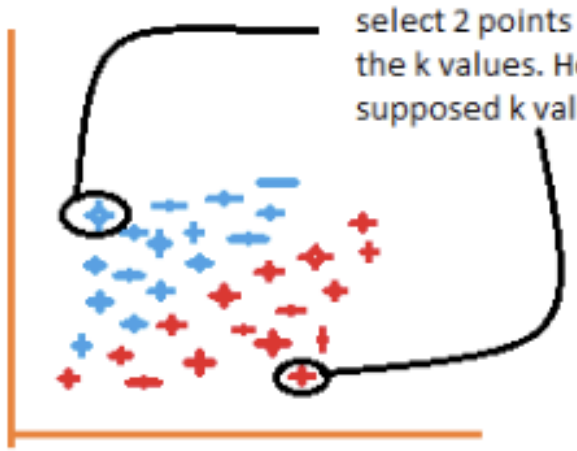
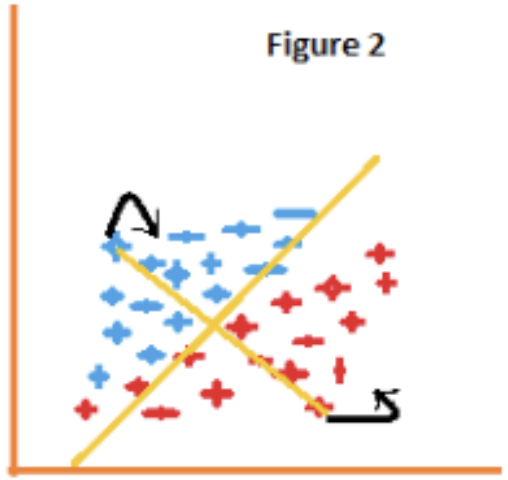


Figure 1

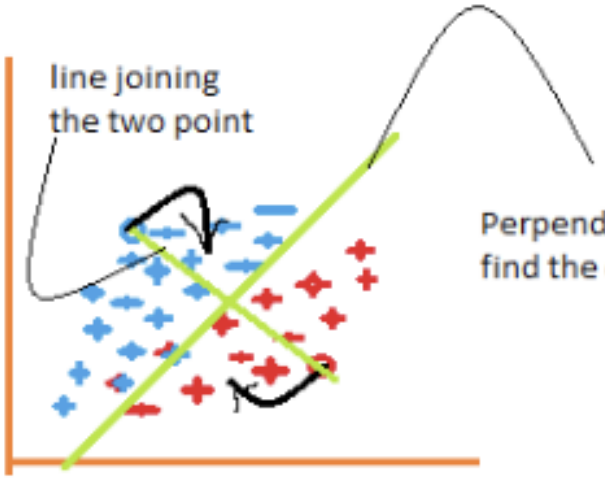
Figure 2



= original points

= the original points moved to centroid.

F2: Find the average of all the blue points and red points and move the selected points to **centroid.**



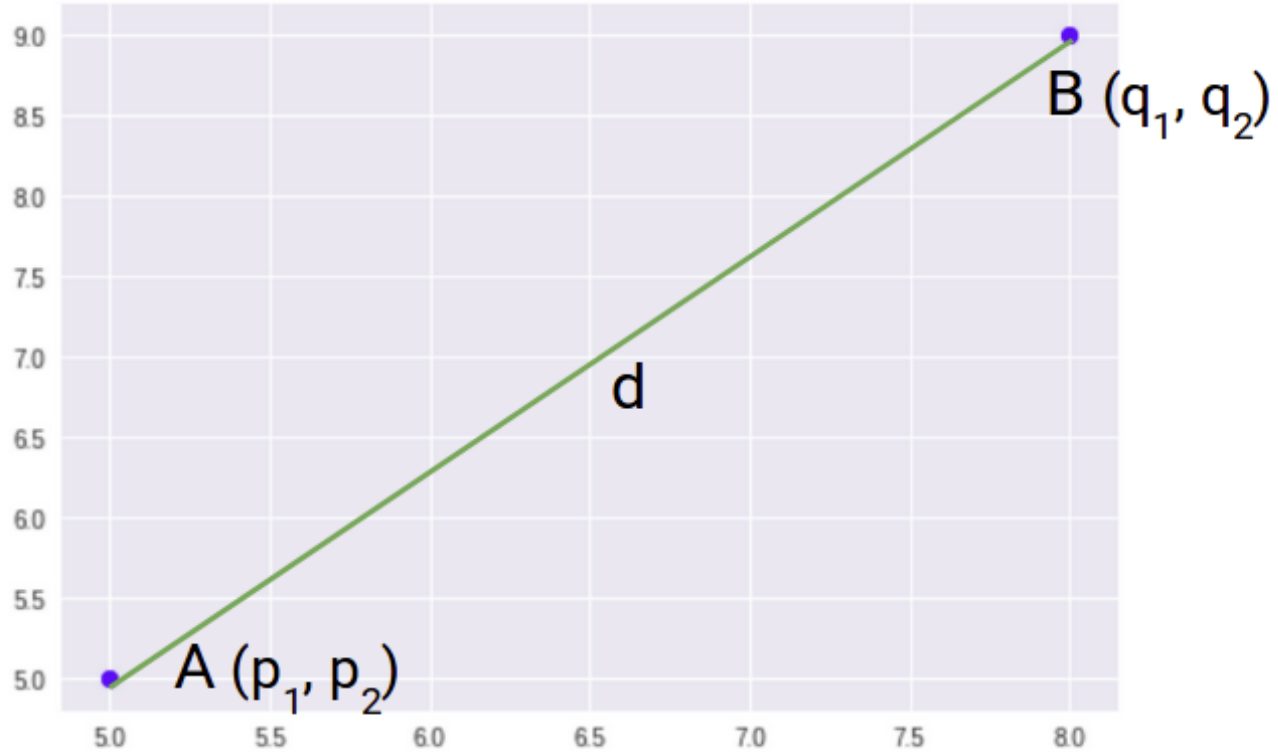
F3: Some of the red points changed to blue points, that means they belong to the group blue now. Again the repeat the same process.

F4: The same process has been applied here. This process will be continued until we get the **two complete different cluster.**

How k-means works

University of Roehampton London

- Please note that the K-means clustering uses the Euclidean distance method to find out the distance between the points.



Here's the formula for Euclidean Distance:

$$d = ((p_1 - q_1)^2 + (p_2 - q_2)^2)^{1/2}$$

How k-means works

University of Roehampton London

We use this formula when we are dealing with 2 dimensions. We can generalize this for an n-dimensional space as:

$$D_e = \left(\sum_{i=1}^{n} (p_i - q_i)^2 \right)^{1/2}$$

Where,

- n = number of dimensions
- pi, qi = data points

In seminar and lab session, we will practice how to use K-means

Advantages & Disadvantages

University of Roehampton London

Advantage

- It is very simple to implement.
- It is scalable to a huge data set and also faster to large datasets.
- it adapts the new examples very frequently.
- Generalization of clusters for different shapes and sizes.

Disadvantage

- It is sensitive to the outliers.
- Choosing the k values manually is a tough job.
- As the number of dimensions increases its scalability decreases.