K-Means

By Arunkumar Nair
arunkg99@gmail.com
91-9890652675
https://www.linkedin.com/in/arunkumarnair/

K Means

- **k-means** is one of the simplest unsupervised learning algorithms that solve the clustering problems.
- The procedure follows a **simple** and easy way to classify a given data set through a certain number of clusters (assume **k** clusters). The main idea is to define **k** centers, one for each cluster.

K-Means

- The k-means algorithm is an **unsupervised** clustering algorithm. It takes a bunch of unlabeled points and tries to group them into "K" number of clusters.

- Used unlabeled data (Data without defined categories or Group)

- Data points are clustered based on similarities

- The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable *K*.

K-Means Uses Cases

K-Means is applied in

The *K*-means clustering algorithm is used to find groups which have not been explicitly labeled in the data and to find patterns and make better decisions.. Once the algorithm has been run and the groups are defined, any new data can be easily assigned to the most relevant group.

- Customer Profiling:
- market segmentation,
- computer vision
- Geo-statistics
- Astronomy

Common business cases where K-means is used

- Behavioral segmentation:
  - Segment by purchase history
  - Segment by activities on application, website, or platform
  - Define personas based on interests
  - Create profiles based on activity monitoring
- Inventory categorization:
  - Group inventory by sales activity
  - Group inventory by manufacturing metrics
- Sorting sensor measurements:
  - Detect activity types in motion sensors
  - Group images
  - Separate audio
  - Identify groups in health monitoring
- Detecting bots or anomalies:
  - Separate valid activity groups from bots
  - Group valid activity to clean up outlier detection

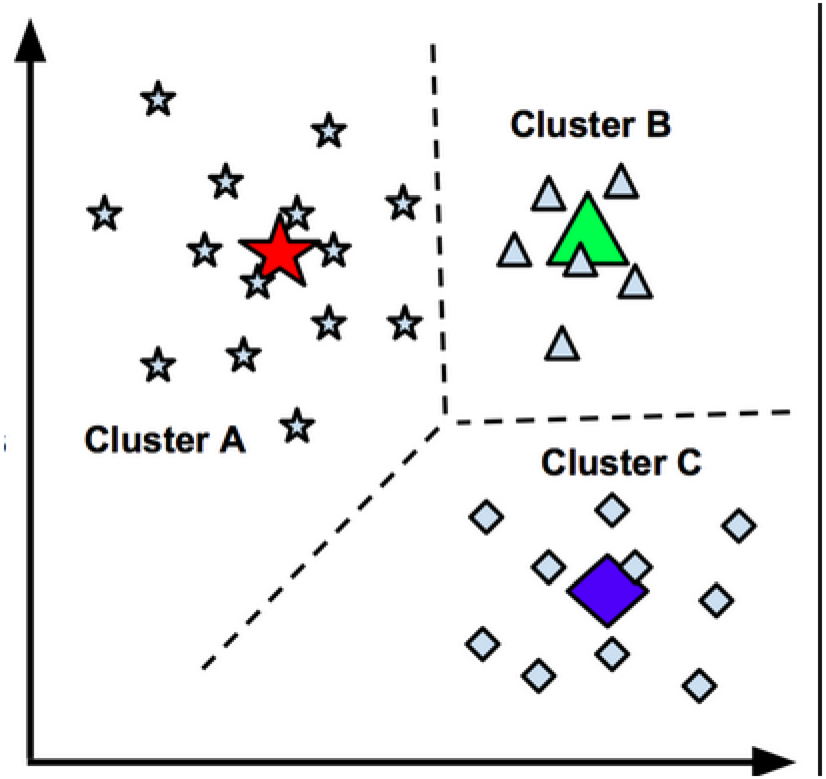
K-Means-How is works

K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids

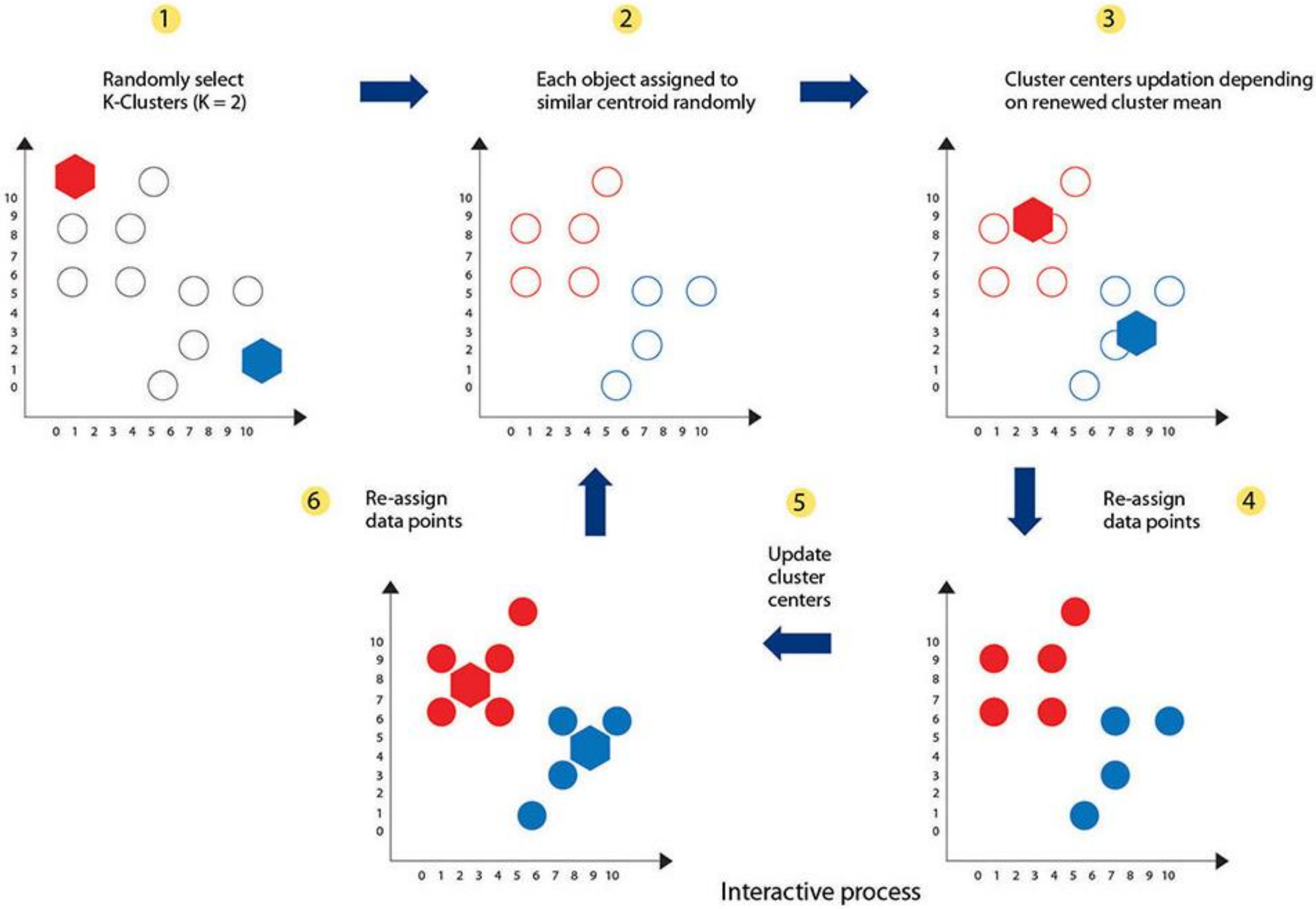
K-means is an **iterative algorithm** and it does two main steps:

1. **Cluster assignment**
2. **Move centroid step**.

K=3 Clusters



K=2 Clusters



1 Randomly select K-Clusters (K = 2)

2 Each object assigned to similar centroid randomly

3 Cluster centers updation depending on renewed cluster mean

4 Re-assign data points

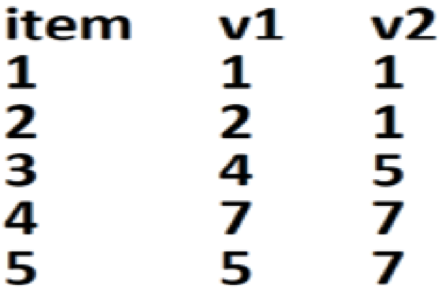
5 Update cluster centers

6 Re-assign data points

Interactive process

K-Means-How is works

Data with No Answers



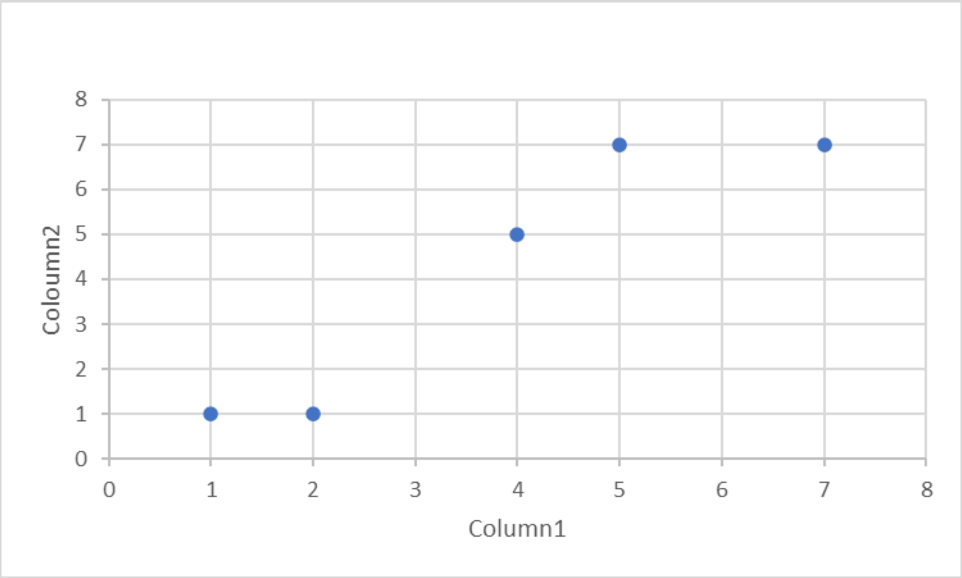
	Age	Weight
item	v1	v2
1	1	1
2	2	1
3	4	5
4	7	7
5	5	7

Data with No Answers

Sr no	Column1	Column2
1	1	1
2	2	1
3	4	5
4	7	7
5	5	7

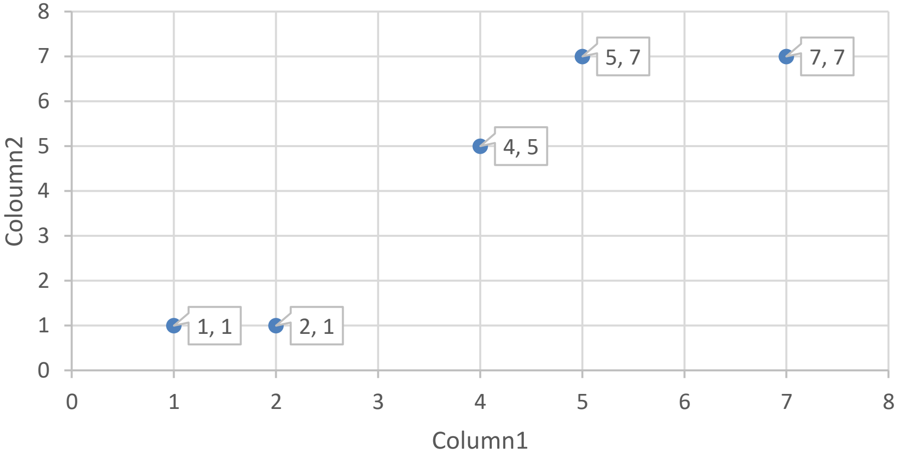
Plot a Graph

Sr no	Column1	Column2
1	1	1
2	2	1
3	4	5
4	7	7
5	5	7



Plot a Graph with the Data

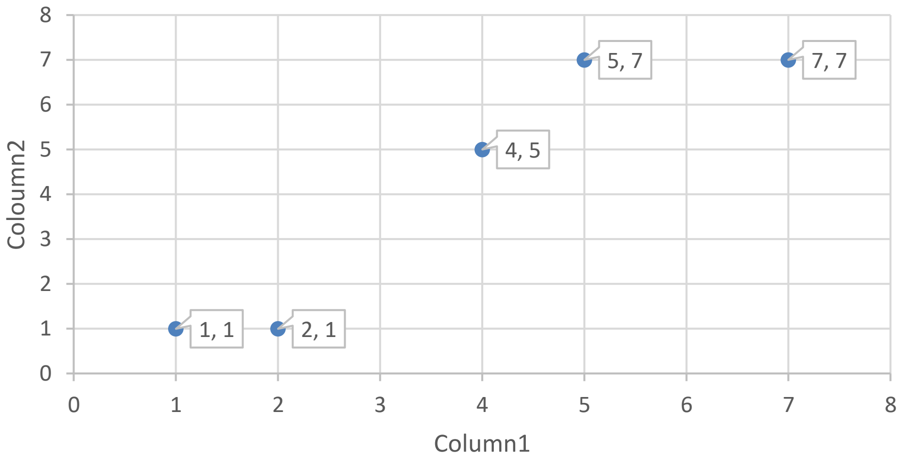
Sr no	Column1	Column2
1	1	1
2	2	1
3	4	5
4	7	7
5	5	7



1. Cluster assignment

When K =2, which means let us create 2 clusters of data points.

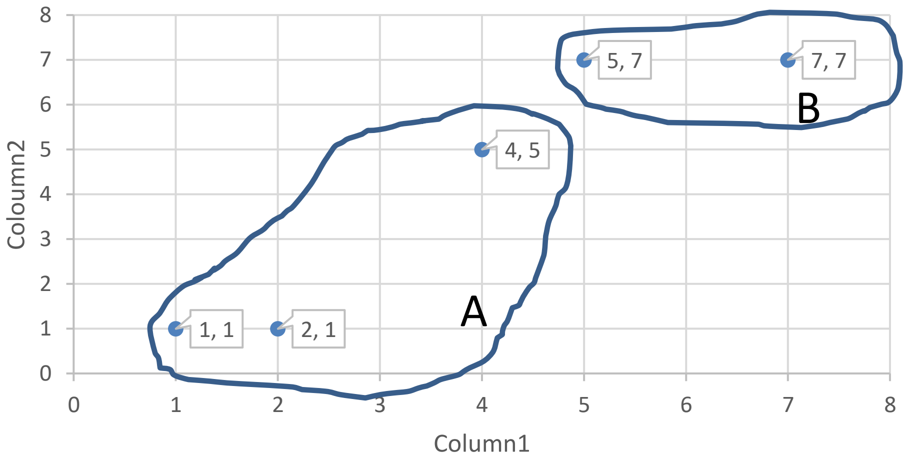
Sr no	Column1	Column2
1	1	1
2	2	1
3	4	5
4	7	7
5	5	7



1. Cluster assignment

First time the Algorithm creates 2 Random Clusters – A and B

Cluster	Sr no	Column1	Column2
A	1	1	1
	2	2	1
	3	4	5
B	4	7	7
	5	5	7

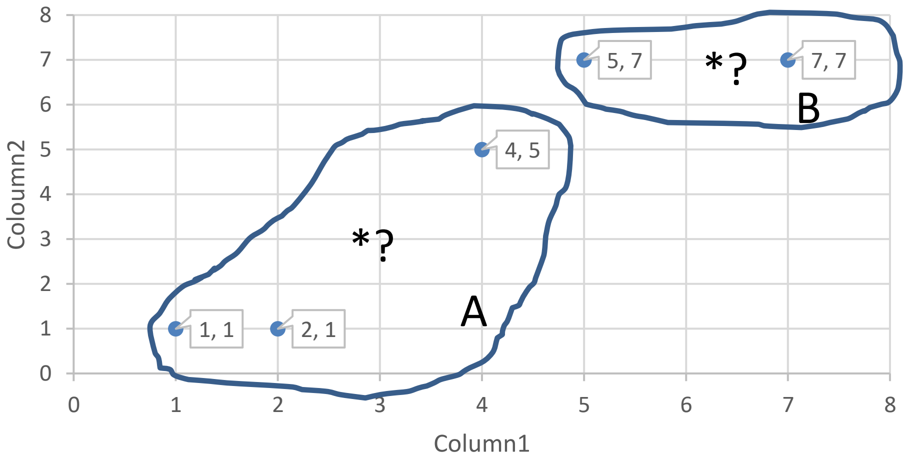


Start with a **Random Cluster, let us assign**
Row 1,2,3 to Cluster A
Row 4,5 to Cluster B

2. Compute the Centroid

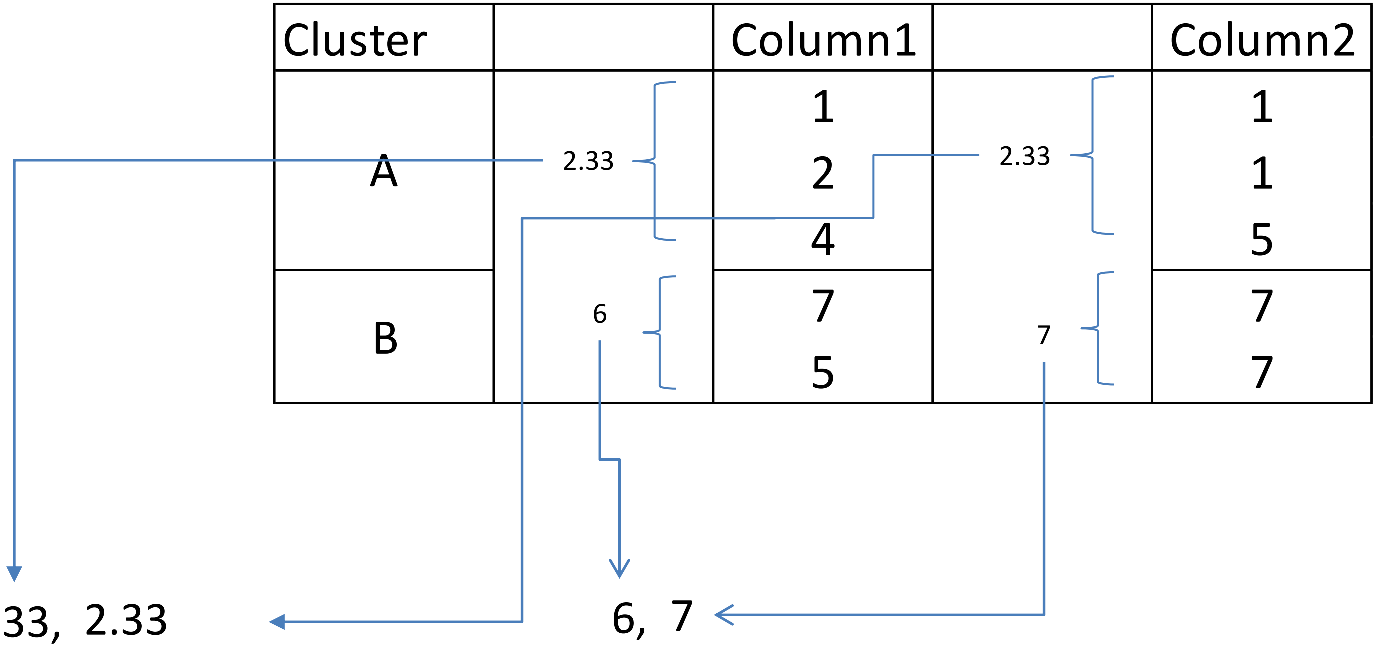
The algorithm goes through the clusters and will compute the Centroid for Cluster A and Cluster B

Cluster	Sr no	Column1	Column2
A	1	1	1
	2	2	1
	3	4	5
B	4	7	7
	5	5	7



What is the centroid of Cluster A and Cluster B?

2. How to Compute the centroid for A, B



Cluster		Column1		Column2
A	2.33	1 2 4	2.33	1 1 5
B	6	7 5	7	7 7

2.33, 2.33

6, 7

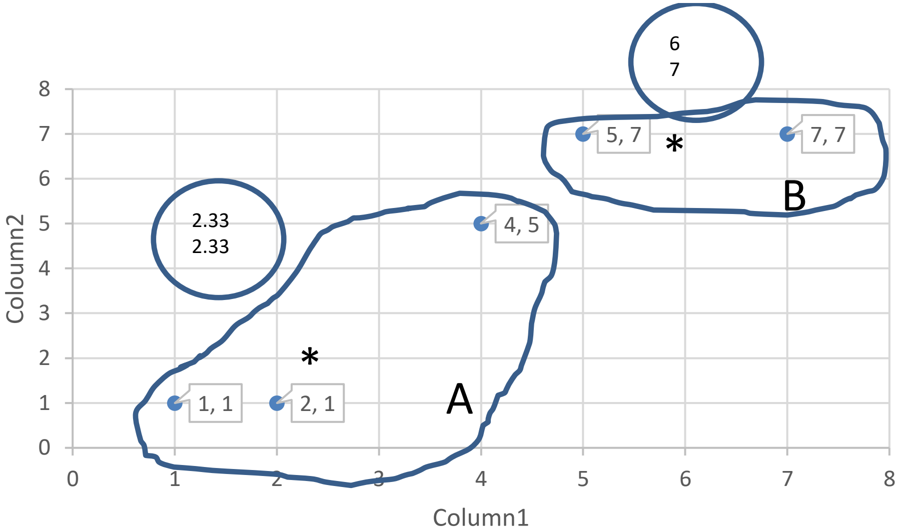
A= (1+2+4)/3=2.33, (1+1+5)/3=2.33

B= (7+5)/2=6, (7+7)/2=7

2. Compute the centroid and plot it on the Graph

Centroid values for Cluster A and Cluster B

Cluster	Sr no	Column1	Column2
A	1	1	1
	2	2	1
	3	4	5
B	4	7	7
	5	5	7



(2.33, 2.33)

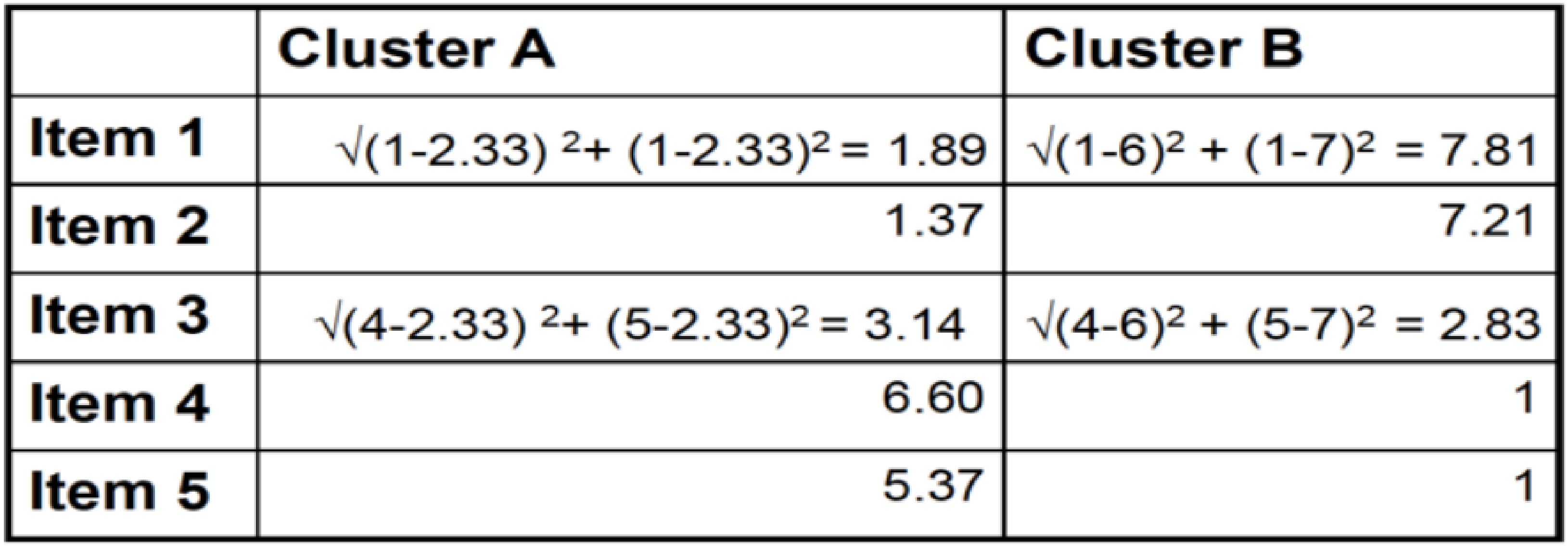
A= (1+2+4)/3=2.33, (1+1+5)/3=2.33

(6, 7)

B= (7+5)/2=6, (7+7)/2=7

3. Measure the Euclidean Distances

Compute Euclidean distance of each record from each centroid, and re-assign to closest cluster.

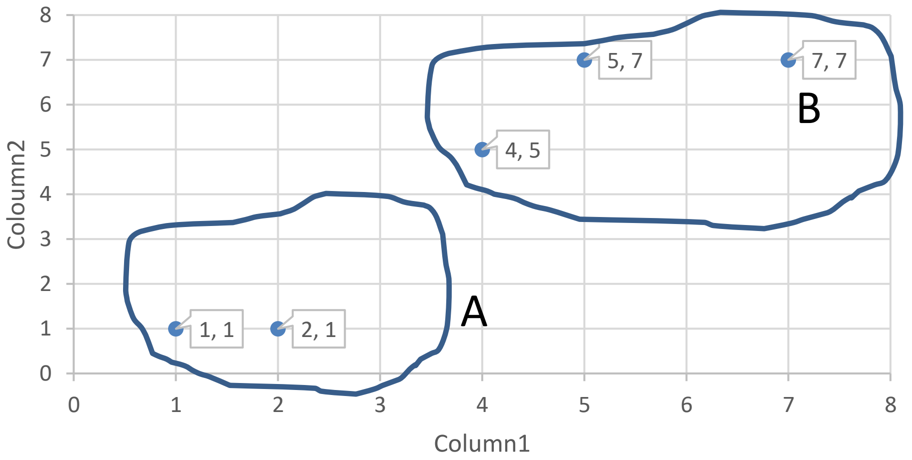


	Cluster A	Cluster B
Item 1	$\sqrt{(1-2.33)^{2}+(1-2.33)^{2}} = 1.89$	$\sqrt{(1-6)^{2}+(1-7)^{2}} = 7.81$
Item 2	1.37	7.21
Item 3	$\sqrt{(4-2.33)^{2}+(5-2.33)^{2}} = 3.14$	$\sqrt{(4-6)^{2}+(5-7)^{2}} = 2.83$
Item 4	6.60	1
Item 5	5.37	1

4. Reassign the data point to closest centroid. Redraw or Rearrange the data and the Clusters

After calculating the Euclidean distance, the data point (3) is reassigned to the closest centroid, which is in cluster (B).

Cluster	Sr No	Column1		Column2
A	1	1		1
	2	2		1
B	3	4		5
	4	7		7
	5	5		7

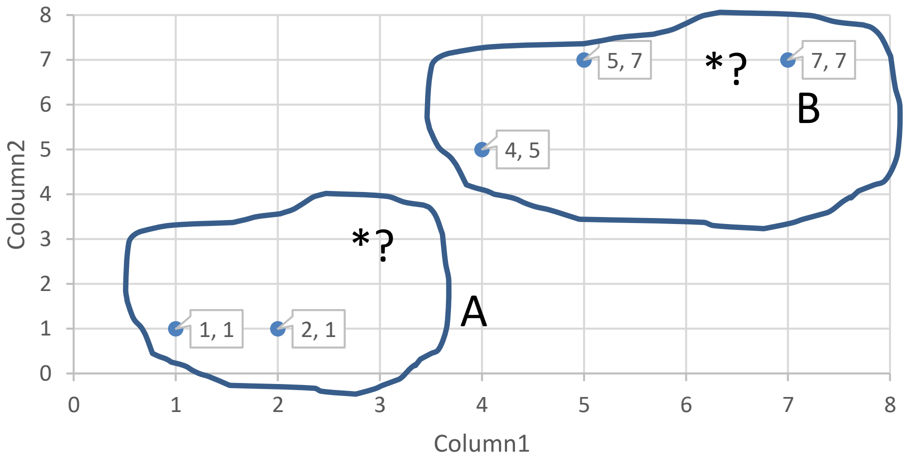


Recreate the **Cluster, let us assign 3 to Cluster B**
Row 1,2 to Cluster A
Row 3,4,5 to Cluster B

5. Re Compute the Centroid for the rearranged clusters

Recompute the centroid of the clusters with the modified data points

Cluster	Sr No	Column1		Column2
A	1	1		1
	2	2		1
B	3	4		5
	4	7		7
	5	5		7



6. How to Compute the centroid for A, B

Cluster		Column1		Column2
A	1.5	1 2	1	1 1
B	5.33	4 7 5	6.33	5 7 7

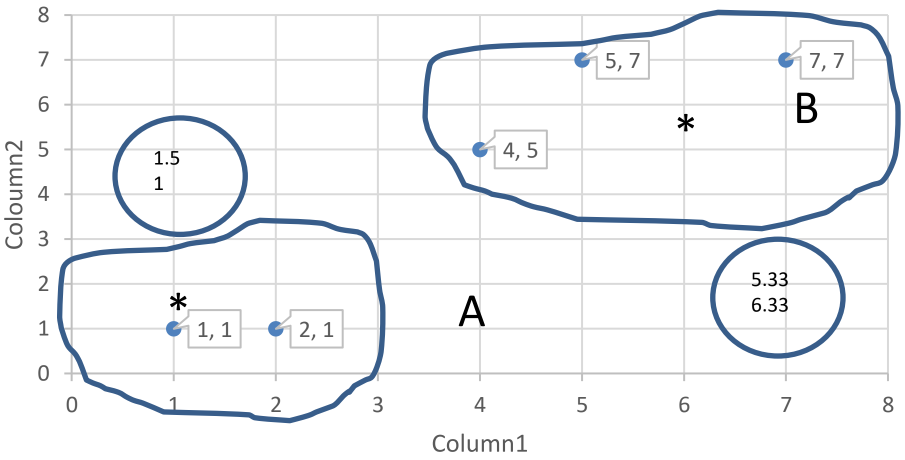
A= (1+2)/2=1.5, (1+1)/2=1

B= (4+7+5)/3=5.33, (5+7+7)/3=6.33

7. Plot the new Centroids on the graph

Recompute the centroid of the clusters with the modified data points

Cluster	Sr No	Column1		Column2
A	1	1		1
	2	2		1
B	3	4		5
	4	7		7
	5	5		7



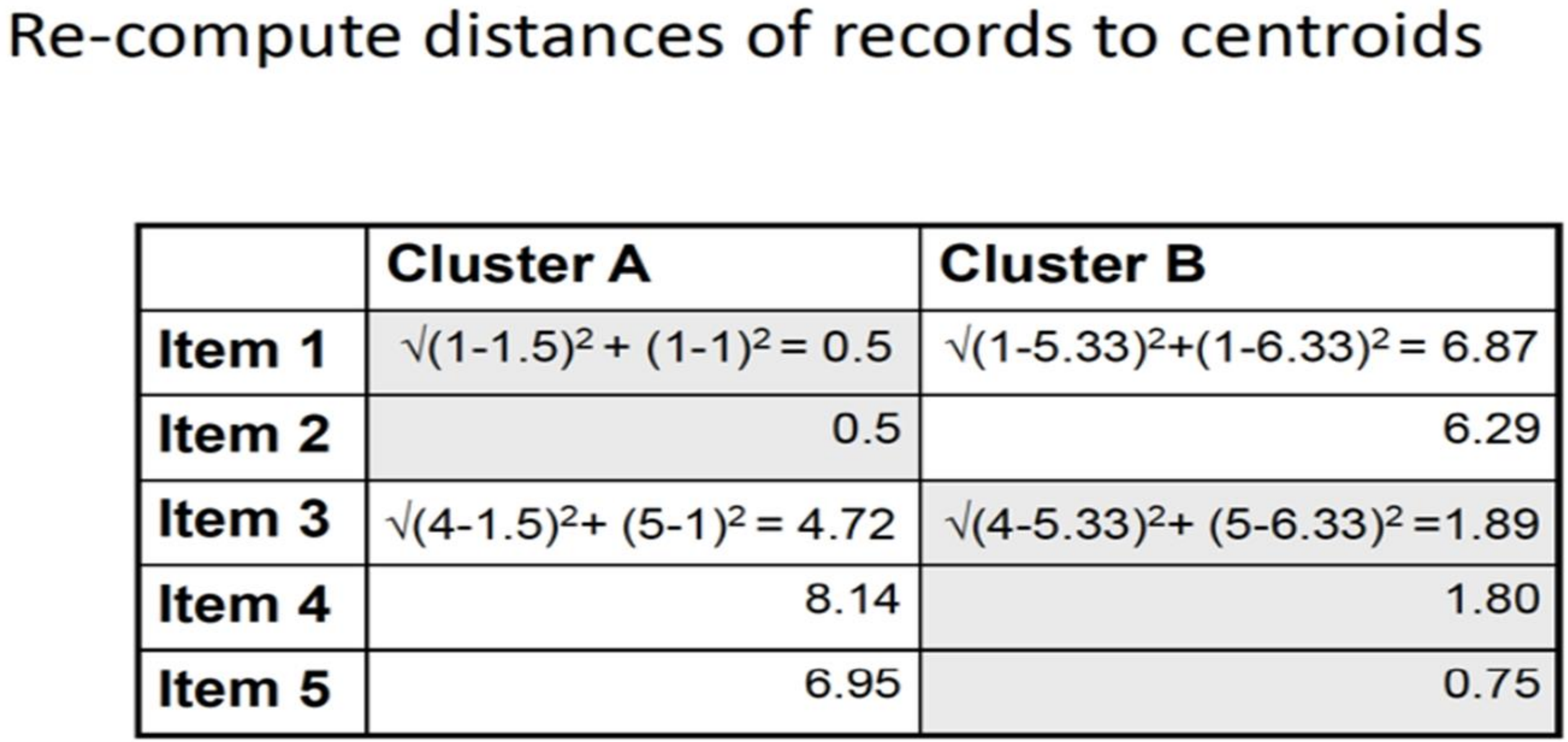
(1.5, 1)

A= (1+2)/2=1.5, (1+1)/2=1

(5.33, 6.33)

B= (4+7+5)/3=5.33, (5+7+7)/3=6.33

Re-Compute the Euclidean Distances from the centroids



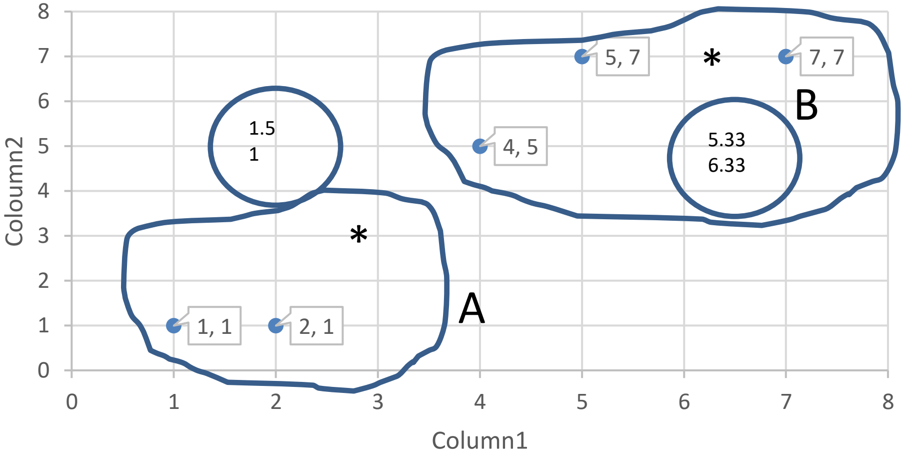
Re-compute distances of records to centroids

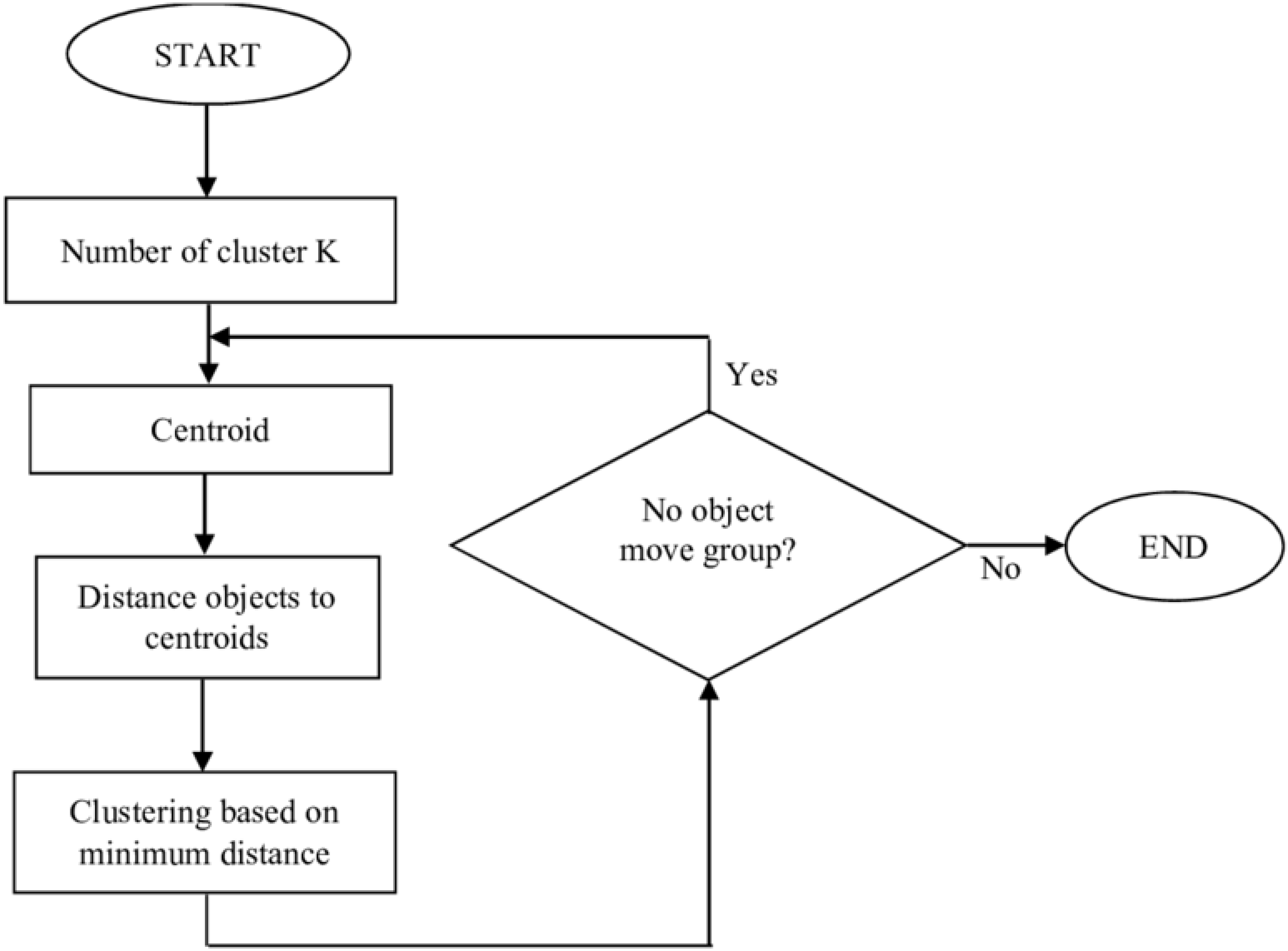
	Cluster A	Cluster B
Item 1	$\sqrt{(1-1.5)^2 + (1-1)^2} = 0.5$	$\sqrt{(1-5.33)^2+(1-6.33)^2} = 6.87$
Item 2	0.5	6.29
Item 3	$\sqrt{(4-1.5)^2+ (5-1)^2} = 4.72$	$\sqrt{(4-5.33)^2+ (5-6.33)^2} =1.89$
Item 4	8.14	1.80
Item 5	6.95	0.75

Repeat

This process is repeated until there is no change in the clusters distance. K is chosen randomly or by giving specific initial starting points by the user.

Cluster	Sr No	Column1		Column2
A	1	1		1
	2	2		1
B	3	4		5
	4	7		7
	5	5		7




START
Number of cluster K
Centroid
Distance objects to centroids
Clustering based on minimum distance
No object move group?
Yes
No
END

Thanks