K-Means:Clustering

K-means is an algorithm that trains a model that groups similar objects together.

DONE BY:SAHITI EMANI
STUDENT ID:19556
GUIDED BY:PROF.HENRY CHANG

TABLE OF CONTENTS

- Introduction
- How does k-means work?
- Example and explanation based on clustering concept
- Applications of k-means clustering
- Steps explained in detail
- Conclusion
- Bibliography

INTRODUCTION:

- K-means is an algorithm that trains a model that groups similar objects together.
- The k-means algorithm accomplishes this by mapping each observation in the input dataset to a point in the *n*-dimensional space here *n* is the number of attributes of the observation.
- Clustering algorithms are unsupervised.
- In unsupervised learning, which labels that might be associated with the objects in the training dataset aren't used.
- For example, your dataset might contain observations of temperature and humidity in a particular location, which are mapped to points (t, h) in 2-dimensional space.

TYPES OF CLUSTERING:

Clustering is a type of unsupervised learning wherein data points are grouped into different sets based on their degree of similarity.

The various types of clustering are:

- Hierarchical clustering
- Partitioning clustering

Hierarchical clustering is further subdivided into:

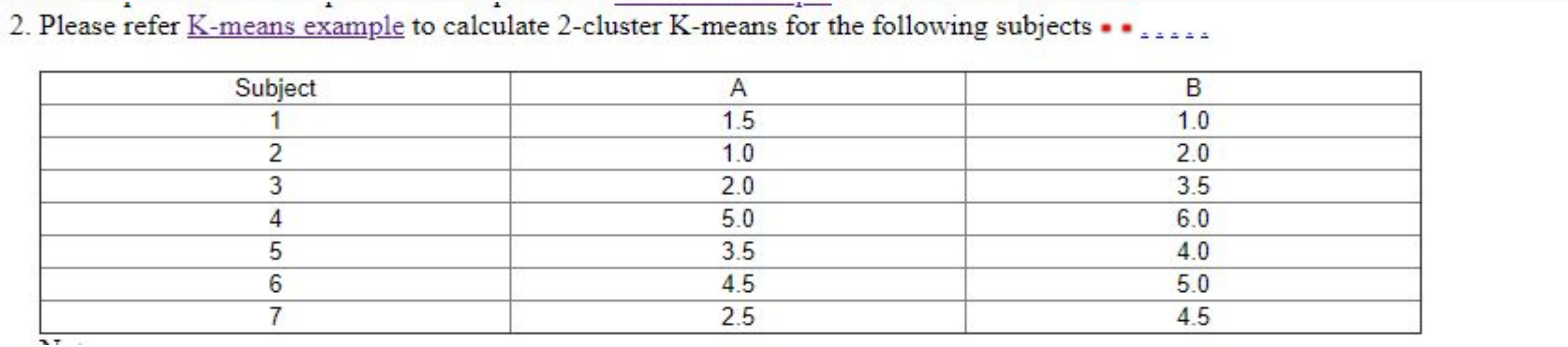
- Agglomerative clustering
- Divisive clustering

Partitioning clustering is further subdivided into:

- **K-Means clustering**
- Fuzzy C-Means clustering

EXAMPLE ON CLUSTERING CONCEPT:

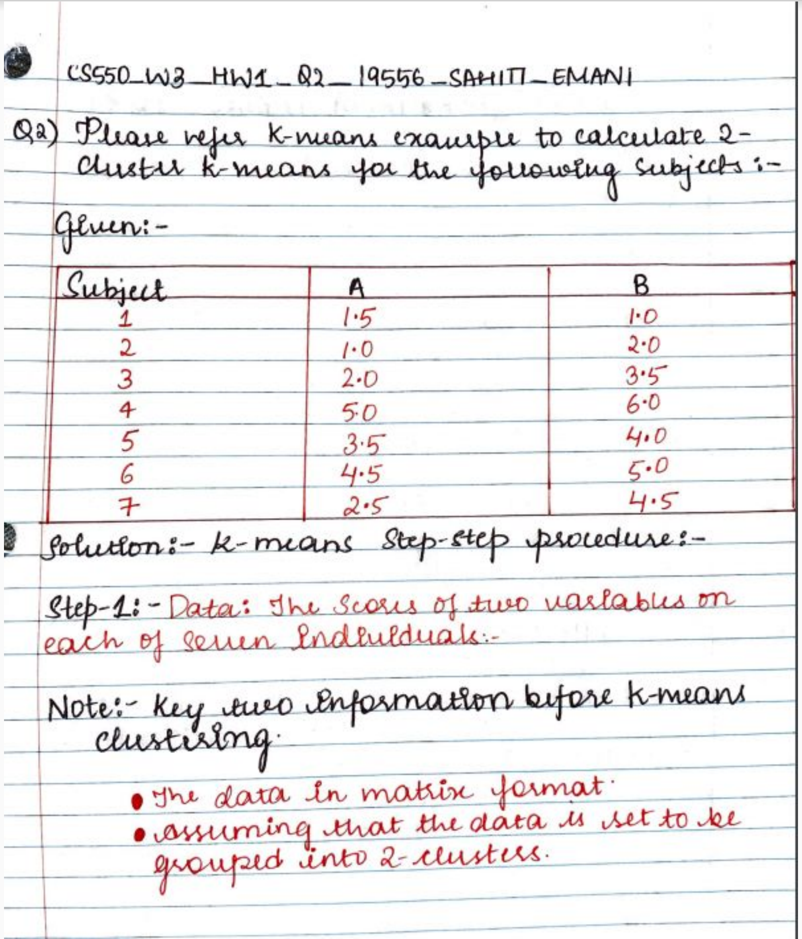
EXAMPLE:https://npu85.npu.edu/~henry/npu/classes/data_science/algorithm/slide/exercise_algorithm.html(Link for further reference)



2. Please refer K-means example to calculate 2-cluster K-means for the following subjects

Subject	A	B
1	1.5	1.0
2	1.0	2.0
3	2.0	3.5
4	5.0	6.0
5	3.5	4.0
6	4.5	5.0
7	2.5	4.5

EXAMPLE:CONT'D



CS550_W3_HW1_Q2_19556_SAHITI_EMANI

Q2) Please refer K-means example to calculate 2-cluster k-means for the following subjects :-

Given:-

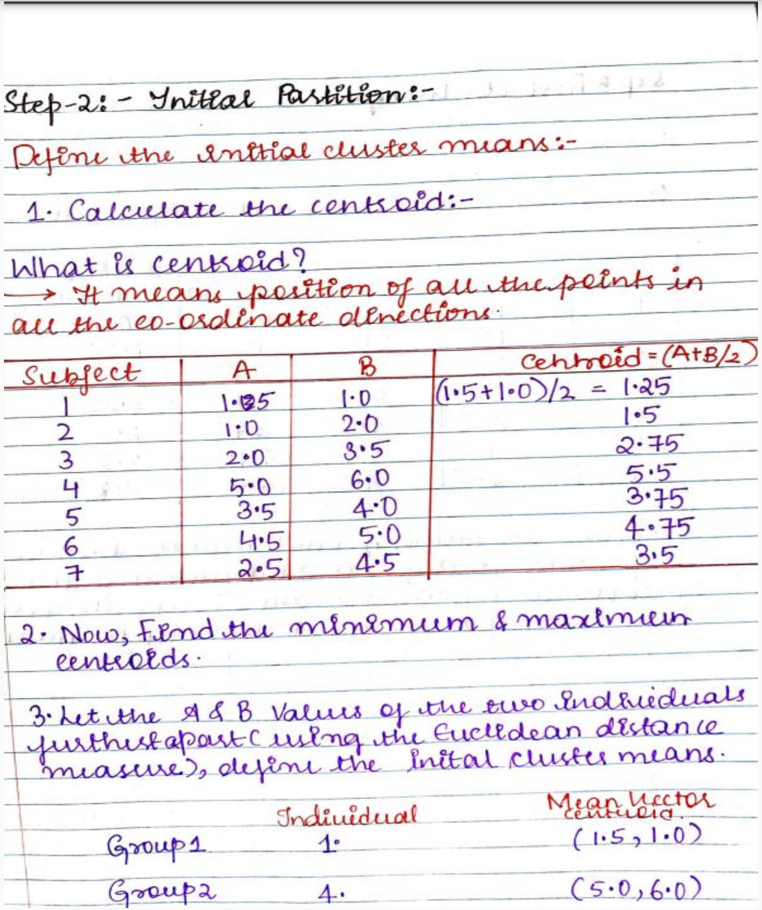
Subject	A	B
1	1·5	1·0
2	1·0	2·0
3	2·0	3·5
4	5·0	6·0
5	3·5	4·0
6	4·5	5·0
7	2·5	4·5

Solution:- k-means Step-step procedure:-

Step-1:- Data: The Scores of two variables on each of seven Individuals:-

Note:- Key two information before k-means clustering.

- The data in matrix format.
- assuming that the data is set to be grouped into 2-clusters.



Step-2:- Initial Partition:-

Define the initial cluster means:-

1. Calculate the centroid:-

What is centroid?
→ It means position of all the points in all the co-ordinate directions.

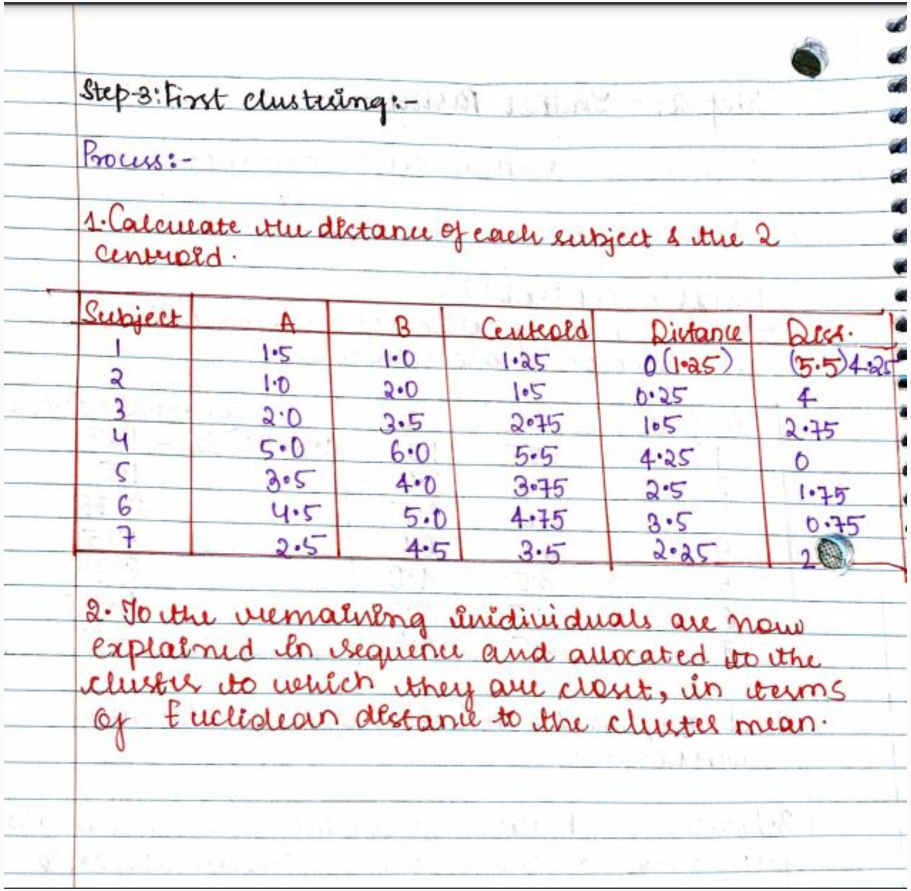
Subject	A	B	Centroid = (A+B/2)
1	1·5	1·0	(1·5+1·0)/2 = 1·25
2	1·0	2·0	1·5
3	2·0	3·5	2·75
4	5·0	6·0	5·5
5	3·5	4·0	3·75
6	4·5	5·0	4·75
7	2·5	4·5	3·5

2. Now, Find the minimum & maximum centroids.

3. Let the A & B Values of the two Individuals furthest apart (using the Euclidean distance measure), define the initial cluster means.

	Individual	Mean Vector centroid.
Group 1	1.	(1·5, 1·0)
Group 2	4.	(5·0, 6·0)

EXAMPLE:CONT'D



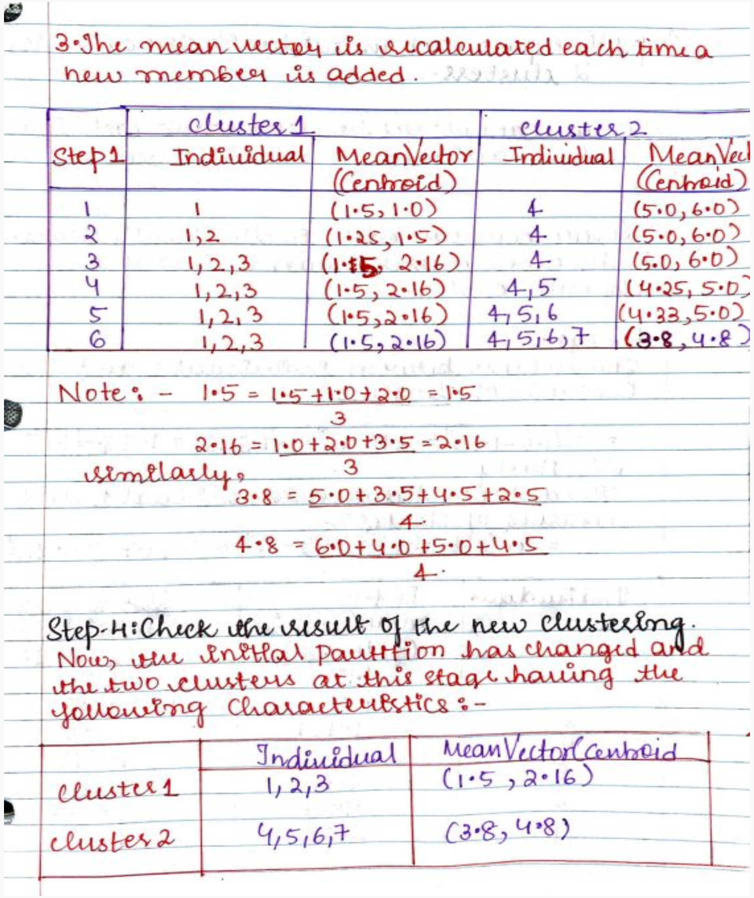
Step-3: First clustering:-

Process:-

1. Calculate the distance of each subject & the 2 centroid.

Subject	A	B	Centroid	Distance	Dist.
1	1.5	1.0	1.25	0 (1.25)	(5.5) 4.25
2	1.0	2.0	1.5	0.25	4
3	2.0	3.5	2.75	1.5	2.75
4	5.0	6.0	5.5	4.25	0
5	3.5	4.0	3.75	2.5	1.75
6	4.5	5.0	4.75	3.5	0.75
7	2.5	4.5	3.5	2.25	2

2. To the remaining individuals are now explained in sequence and allocated to the cluster to which they are closest, in terms of Euclidean distance to the cluster mean.



3. The mean vector is recalculated each time a new member is added.

	cluster 1		cluster 2	
Step 1	Individual	Mean Vector (Centroid)	Individual	Mean Vec (Centroid)
1	1	(1.5, 1.0)	4	(5.0, 6.0)
2	1,2	(1.25, 1.5)	4	(5.0, 6.0)
3	1,2,3	(1.5, 2.16)	4	(5.0, 6.0)
4	1,2,3	(1.5, 2.16)	4,5	(4.25, 5.0
5	1,2,3	(1.5, 2.16)	4,5,6	(4.33, 5.0)
6	1,2,3	(1.5, 2.16)	4,5,6,7	(3.8, 4.8)

Note:- $1.5 = \frac{1.5+1.0+2.0}{3} = 1.5$

$$2.16 = \frac{1.0+2.0+3.5}{3} = 2.16$$

similarly,

$$3.8 = \frac{5.0+3.5+4.5+2.5}{4}$$

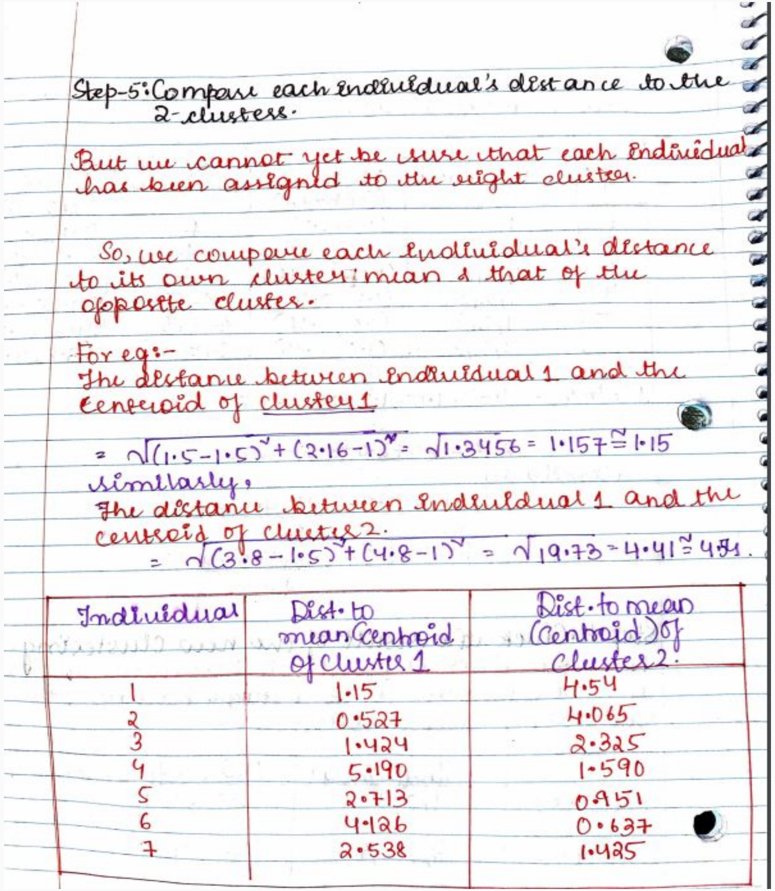
$$4.8 = \frac{6.0+4.0+5.0+4.5}{4}.$$

Step-4: Check the result of the new clustering.

Now, the initial partition has changed and the two clusters at this stage having the following characteristics:-

	Individual	Mean Vector (Centroid)
cluster 1	1,2,3	(1.5, 2.16)
cluster 2	4,5,6,7	(3.8, 4.8)

EXAMPLE:CONT’D



Step-5: Compare each individual's distance to the 2-clusters.

But we cannot yet be sure that each individual has been assigned to the right cluster.

So, we compare each individual's distance to its own cluster mean & that of the opposite cluster.

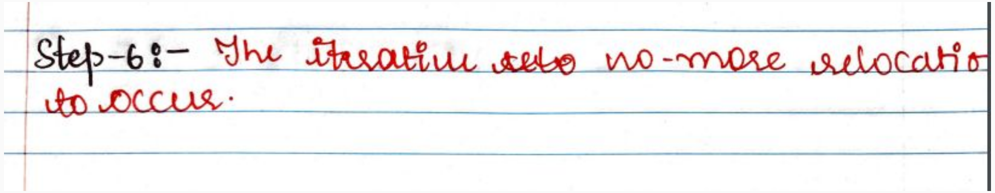
For eg:-
The distance between individual 1 and the centroid of cluster 1

$= \sqrt{(1.5-1.5)^2+(2.16-1)^2} = \sqrt{1.3456} = 1.157 \cong 1.15$

similarly,
The distance between individual 1 and the centroid of cluster 2.

$= \sqrt{(3.8-1.5)^2+(4.8-1)^2} = \sqrt{19.73} = 4.41 \cong 4.54.$

Individual	Dist. to mean (Centroid of Cluster 1	Dist. to mean (Centroid) of Cluster 2.
1	1.15	4.54
2	0.527	4.065
3	1.424	2.325
4	5.190	1.590
5	2.713	0.951
6	4.126	0.637
7	2.538	1.425



Step-6:- The iteration sets no-more relocatio to occur.

APPLICATIONS OF K-MEANS:

k-means algorithm is very popular and used in a variety of applications such as market segmentation, document clustering, image segmentation and image compression, etc.

It’s applications are:

- Geyser eruptions segmentation (2D dataset).
- Image compression.
- Academic performance.
- Search Engines

CONCLUSION:

- K-means gives more weight to the bigger clusters.
- K-means assumes spherical shapes of clusters with radius equal to the distance between the centroid and the furthest data point and doesn't work well when clusters are in different shapes such as elliptical clusters.
- If there is overlapping between clusters, kmeans doesn't have an intrinsic measure for uncertainty for the examples belong to the overlapping region in order to determine for which cluster to assign each data point.
- K-means may still cluster the data even if it can't be clustered such as data that comes from uniform distributions.

BIBLIOGRAPHY:

https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a#:~:text=kmeans%20algorithm%20is%20very%20popular,data%20we're%20dealing%20with.

https://docs.aws.amazon.com/sagemaker/latest/dg/algo-kmeans-tech-notes.html

https://npu85.npu.edu/~henry/npu/classes/data_science/algorithm/slide/k-means_example.html