Redesigning Common Mind & Business Towards Excellence

3Purpose Process People Culture | sns INSTITUTIONS www.snsgroups.com | 1st GenAI Powered Design Thinking FrameWork

Build an Entrepreneurial Mindset Through Our Design Thinking FrameWork

SNS COLLEGE OF TECHNOLOGY

Coimbatore-35
An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A++' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

23AMB201 - MACHINE LEARNING

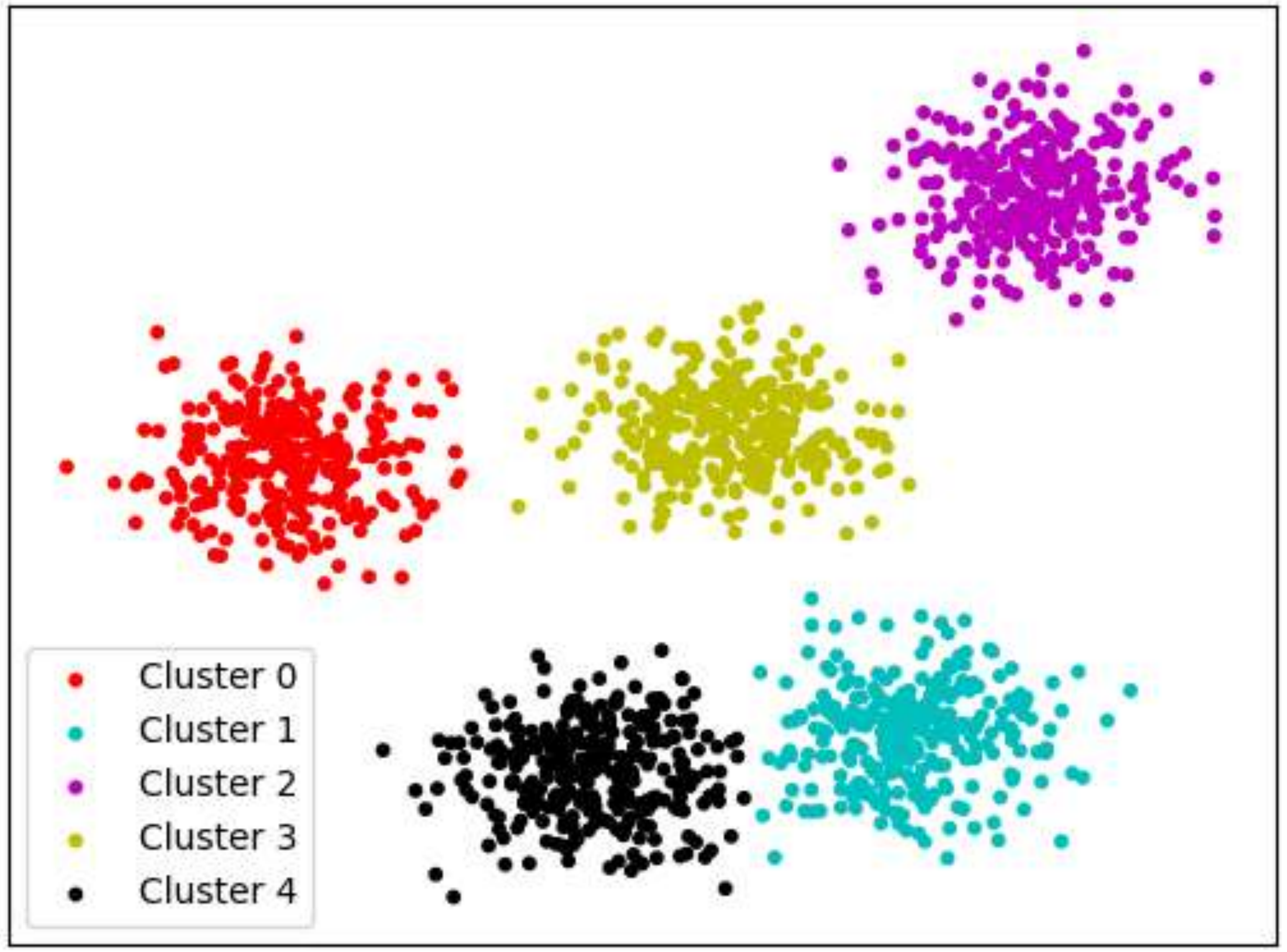
II YEAR IV SEM

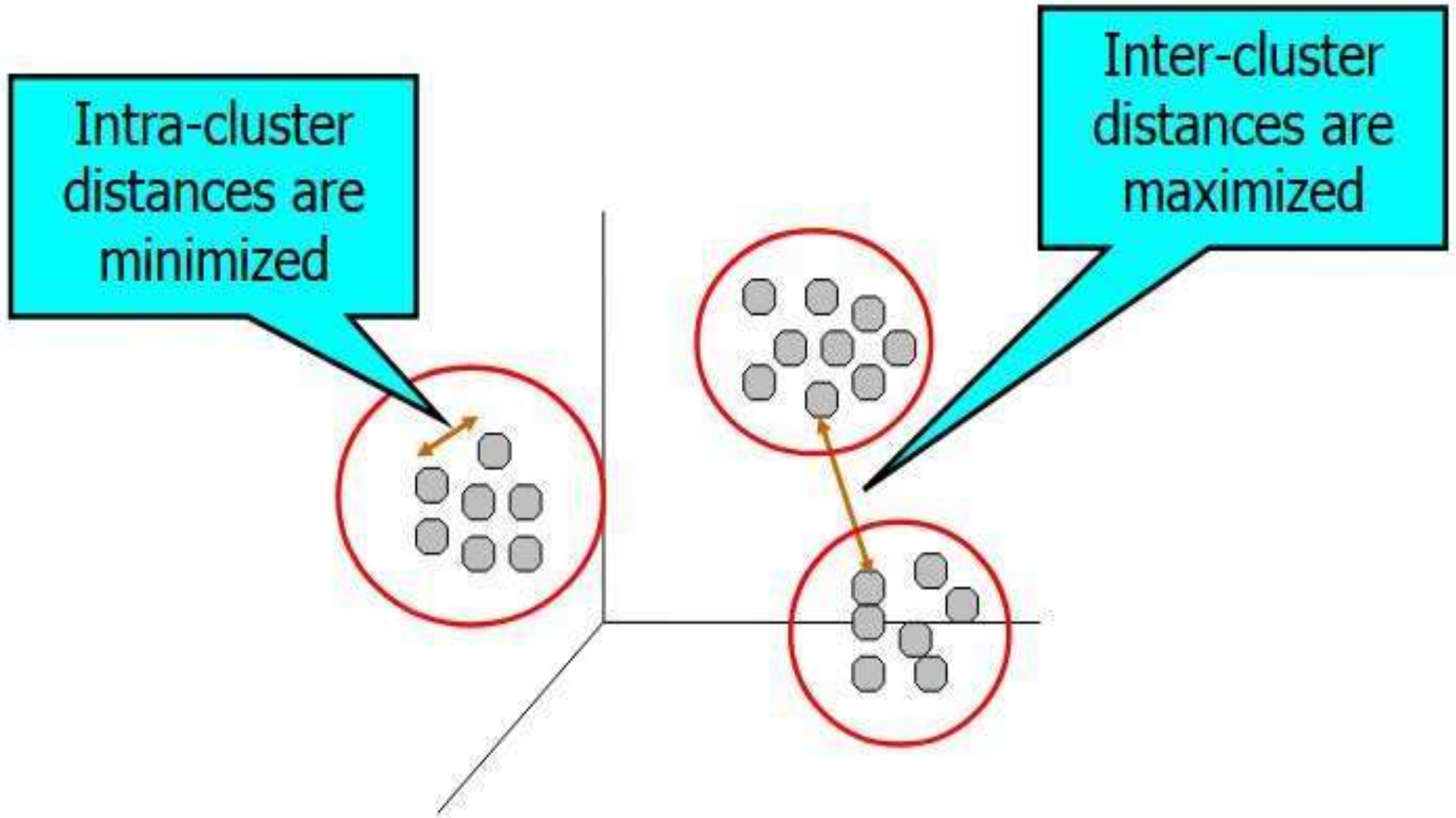
UNIT IV – UNSUPERVISED LEARNING ALGORITHM

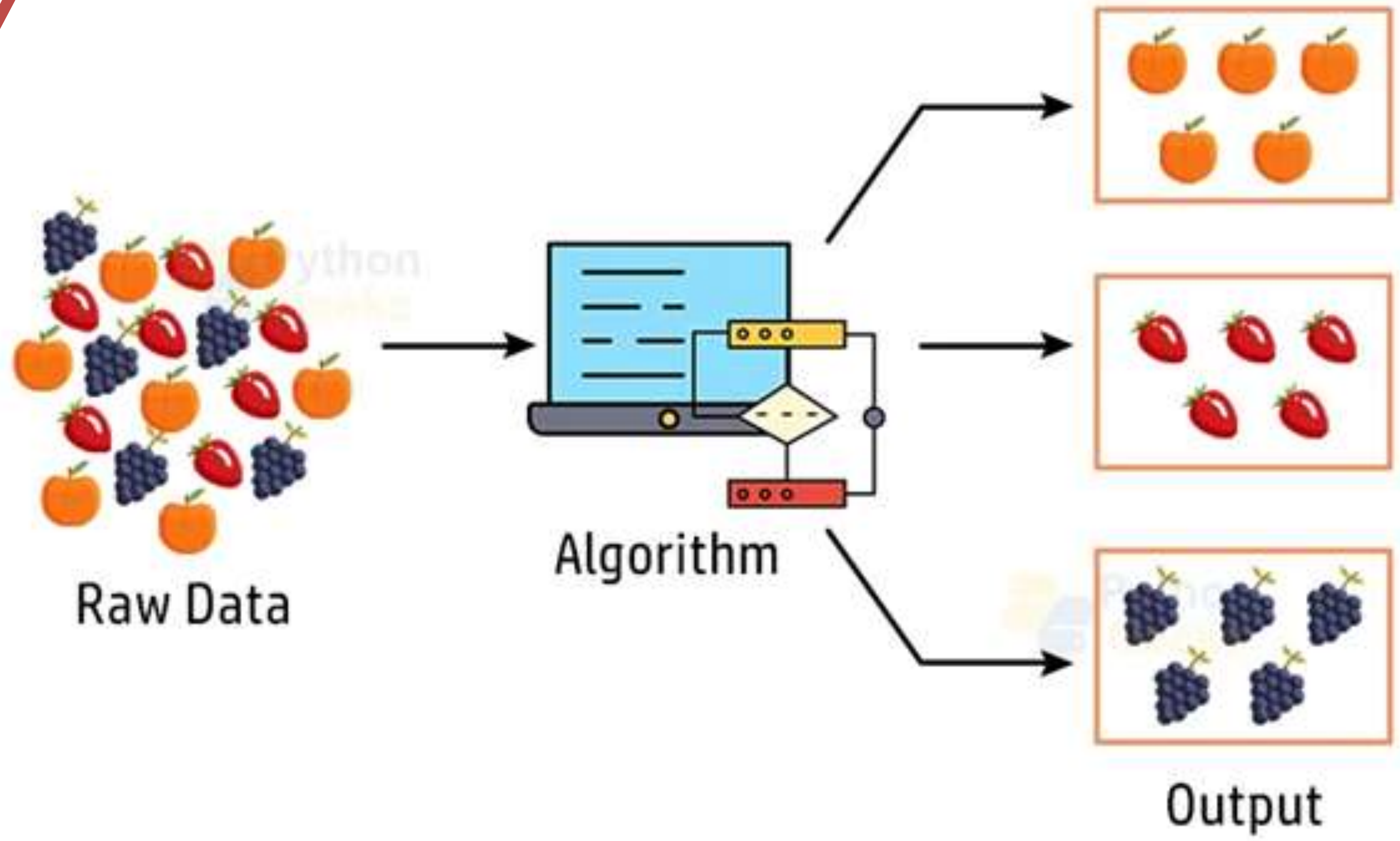
TOPIC 22 – Clustering – K-Means

Clustering

Grouping the same items together depends







What is Clustering?

- **Clustering** is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often according to some defined distance measure.
- Applications:
    1. Market Segmentation
    2. Statistical data analysis
    3. Social network analysis
    4. Image segmentation
    5. Amazon and Netflix

Types of Clustering

1. **Hierarchical algorithms**: - Find successive clusters

    1.**Agglomerative** ("bottom-up"): Begins with each element as a separate cluster and merge them into successively larger clusters.

    2.**Divisive** ("top-down"): Begins with the whole set and proceed to divide it into successively smaller clusters.

2. **Partitional clustering:** Partitional algorithms determine all clusters at once.They include:

    ***K*-means and derivatives**

    Fuzzy *c*-means clustering

3. **Density based clustering**

4. **Fuzzy clustering**

Common Distance measures

- *Distance measure* will determine how the *similarity* of two elements is calculated and it will influence the shape of the clusters.

  They include:

1. The Euclidean distance (also called 2-norm distance) is given by:

$$d(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

2. The Manhattan distance (also called taxicab norm or 1-norm) is given by:

$$d(x,y) = \sum_{i=1}^{m}|x_i - y_i|$$

Common Distance measures

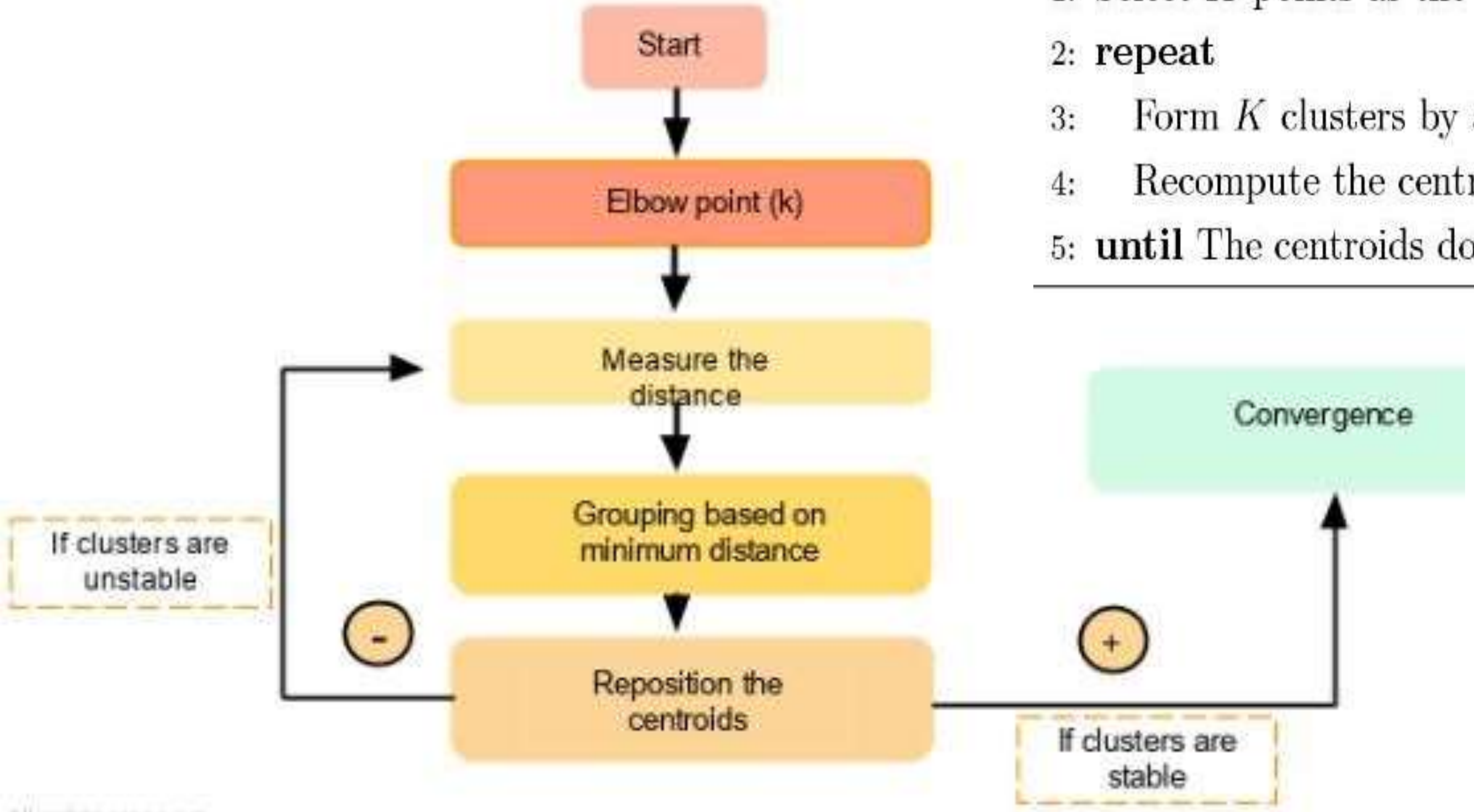
3. The maximum norm is given by: $d(x, y) = \max_{1 \le i \le p} | x_i - y_i |$

4. Hamming distance (sometimes edit distance) measures the minimum number of substitutions required to change one member into another.

$$D_H = \sum_{i=1}^{k} |x_i - y_i|$$

- K-Means Clustering is an Unsupervised Machine Learning algorithm which groups the unlabeled dataset into different clusters.
- The **k-means algorithm** is an algorithm to cluster *n* objects based on attributes into *k* partitions, where *k* < *n*.
- K-means clustering is a technique used to organize data into groups based on their similarity.
- For example online store uses K-Means to group customers based on purchase frequency and spending creating segments like:
    - Budget Shoppers
    - Frequent Buyers
    - Big Spenders for personalised marketing.

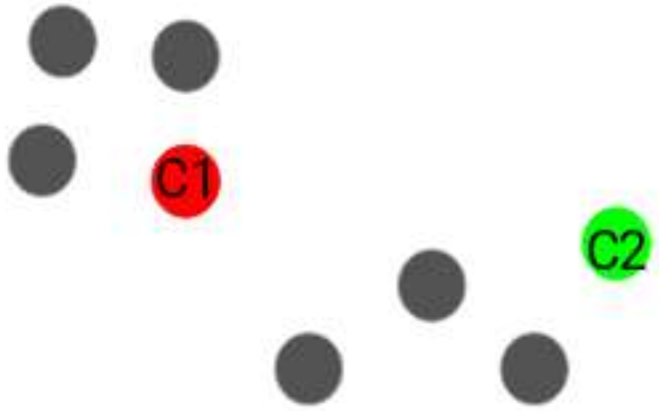
1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change



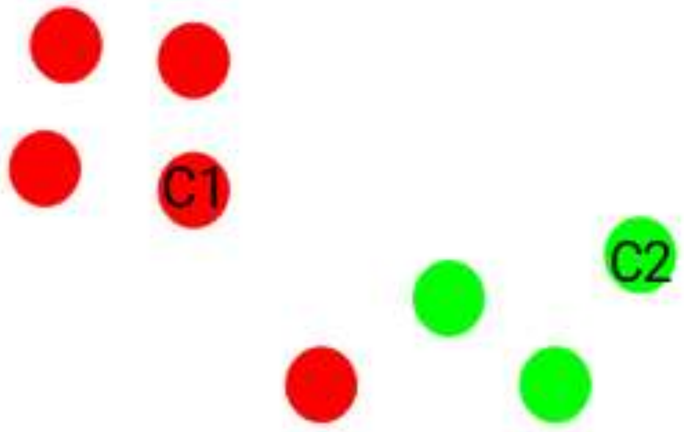
Dataset

Choose the number of clusters k

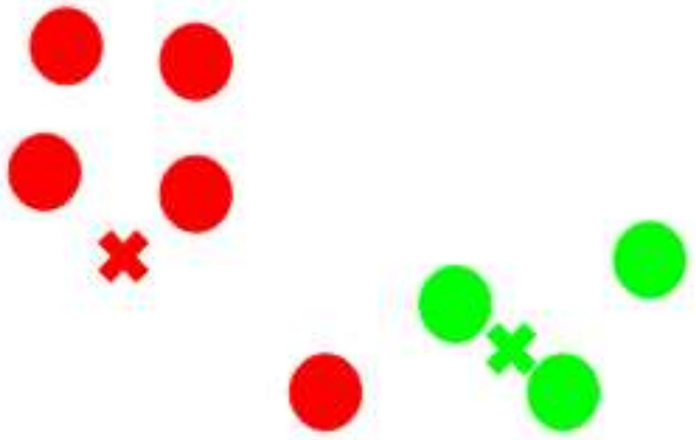
SQRT(N) = SQRT(8) = 2.8 = 2



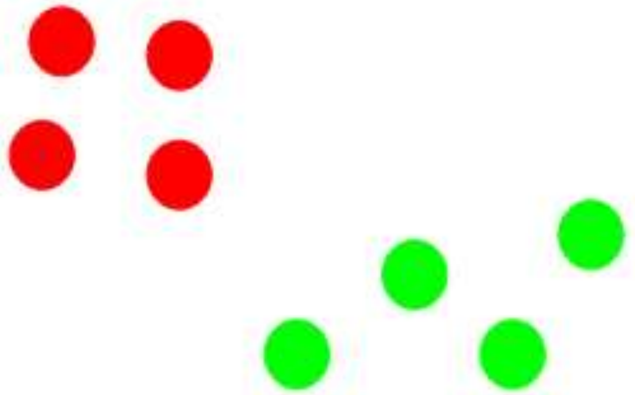
Assign all the points to the closest cluster Centroid



Recompute the centroids of newly formed clusters



Repeat steps 3 and 4



Stopping Criteria for K-Means Clustering

1. Centroids of newly formed clusters do not change
2. Points remain in the same cluster
3. Maximum number of iterations is reached

A Simple example showing the implementation of k-means algorithm (using K=2)

Data set {2, 4, 10, 12, 3, 20, 30, 11, 25}

Iteration 1

M1, M2 are the two randomly selected centroids/means where

M1= 4, M2=11

and the initial clusters are

C1= {4}, C2= {11}

Calculate the Euclidean distance as

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

Datapoint	D1	D2	Cluster
2	2	9	C1
4	0	7	C1
10	6	1	C2
12	8	1	C2
3	1	8	C1
20	16	9	C2
30	26	19	C2
11	7	0	C2
25	21	14	C2

Therefore

C1= {2, 4, 3}

C2= {10, 12, 20, 30, 11, 25}

A Simple example showing the implementation of k-means algorithm (using K=2)

Data set {2, 4, 10, 12, 3, 20, 30, 11, 25}

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

Datapoint	D1	D2	Cluster
2	2	9	C1
4	0	7	C1
10	6	1	C2
12	8	1	C2
3	1	8	C1
20	16	9	C2
30	26	19	C2
11	7	0	C2
25	21	14	C2

Iteration 1

Therefore

C1= {2, 4, 3}

C2= {10, 12, 20, 30, 11, 25}

New Clusters

M1= (2+3+4)/3= 3

M2= (10+12+20+30+11+25)/6= 18

Datapoint	D1	D2	Cluster
2	1	16	C1
4	1	14	C1
3	0	15	C1
10	7	8	C1
12	9	6	C2
20	17	2	C2
30	27	12	C2
11	8	7	C2
25	22	7	C2

Iteration 2

New Clusters

C1= {2, 3, 4, 10}

C2= {12, 20, 30, 11, 25}

A Simple example showing the implementation of k-means algorithm (using K=2)

Data set {2, 4, 10, 12, 3, 20, 30, 11, 25}

$$d(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

Datapoint	D1	D2	Cluster
2	1	16	C1
4	1	14	C1
3	0	15	C1
10	7	8	C1
12	9	6	C2
20	17	2	C2
30	27	12	C2
11	8	7	C2
25	22	7	C2

Iteration 2

New Clusters

C1= {2, 3, 4, 10}

C2= {12, 20, 30, 11, 25}

New Clusters

M1= (2+3+4+10)/4= 4.75

M2= (12+20+30+11+25)/5= 19.6

Datapoint	D1	D2	Cluster
2	2.75	17.6	C1
4	0.75	15.6	C1
3	1.75	16.6	C1
10	5.25	9.6	C1
12	7.25	7.6	C1
20	15.25	0.4	C2
30	25.25	10.4	C2
11	6.25	8.6	C1
25	20.25	5.4	C2

Iteration 3

New Clusters

C1= {2, 3, 4, 10, 12, 11}

C2= {20, 30, 25}

A Simple example showing the implementation of k-means algorithm (using K=2)

Data set {2, 4, 10, 12, 3, 20, 30, 11, 25}

$$d(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

Datapoint	D1	D2	Cluster
2	2.75	17.6	C1
4	0.75	15.6	C1
3	1.75	16.6	C1
10	5.25	9.6	C1
12	7.25	7.6	C1
20	15.25	0.4	C2
30	25.25	10.4	C2
11	6.25	8.6	C1
25	20.25	5.4	C2

Iteration 3

New Clusters

C1= {2, 3, 4, 10, 12, 11}

C2= {20, 30, 25}

New Clusters

M1= (2+3+4+10+12+11)/6=7

M2= (20+30+25)/3= 25

Datapoint	D1	D2	Cluster
2	5	23	C1
4	3	21	C1
3	4	22	C1
10	3	15	C1
12	5	13	C1
11	4	14	C1
20	13	5	C2
30	23	5	C2
25	18	0	C2

Iteration 4

New Clusters

C1= {2, 3, 4, 10, 12, 11}

C2= {20, 30, 25}

No Change between Iteration 3 and 4

References

Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, —Learning from Data, AML Book Publishers, 2012.

P. Flach, —Machine Learning: The art and science of algorithms that make sense of data‖, Cambridge University Press, 2012.

W3school.com

https://discourse.opengenus.org/t/using-id3-algorithm-to-build-a-decision-tree-to-predict-the-weather/3343