Machine Learning & Deep Learning

Week-10

Instructor: *Najam Aziz*

Clustering Techniques in Machine Learning

What is Clustering?

- Grouping unlabeled examples is called clustering.

  As the examples are unlabeled, clustering relies on unsupervised machine learning. If the examples are labeled, then clustering becomes classification.

  A cluster refers to a collection of data points aggregated together because of certain similarities.

- Before you can group similar examples, you first need to find similar examples. You can measure similarity between examples by combining the examples' feature data into a metric, called a similarity measure. When each example is defined by one or two features, it's easy to measure similarity. For example, you can find similar books by their authors. As the number of features increases, creating a similarity measure becomes more complex.

Classification vs Clustering

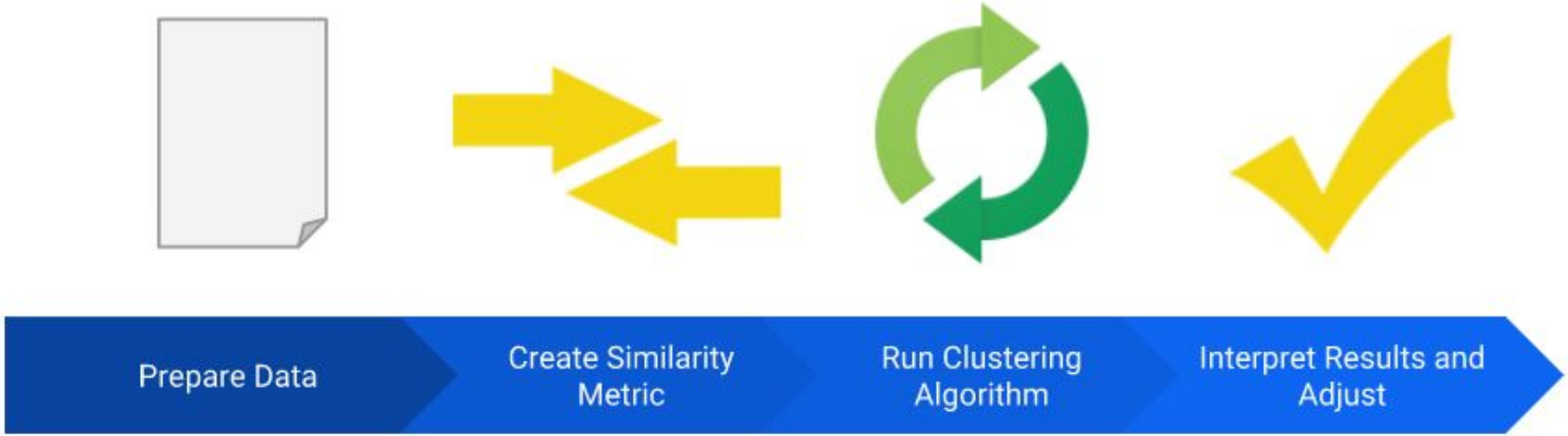
Two methods of pattern identification used in machine learning. Although both techniques have certain similarities, the difference lies in the fact that

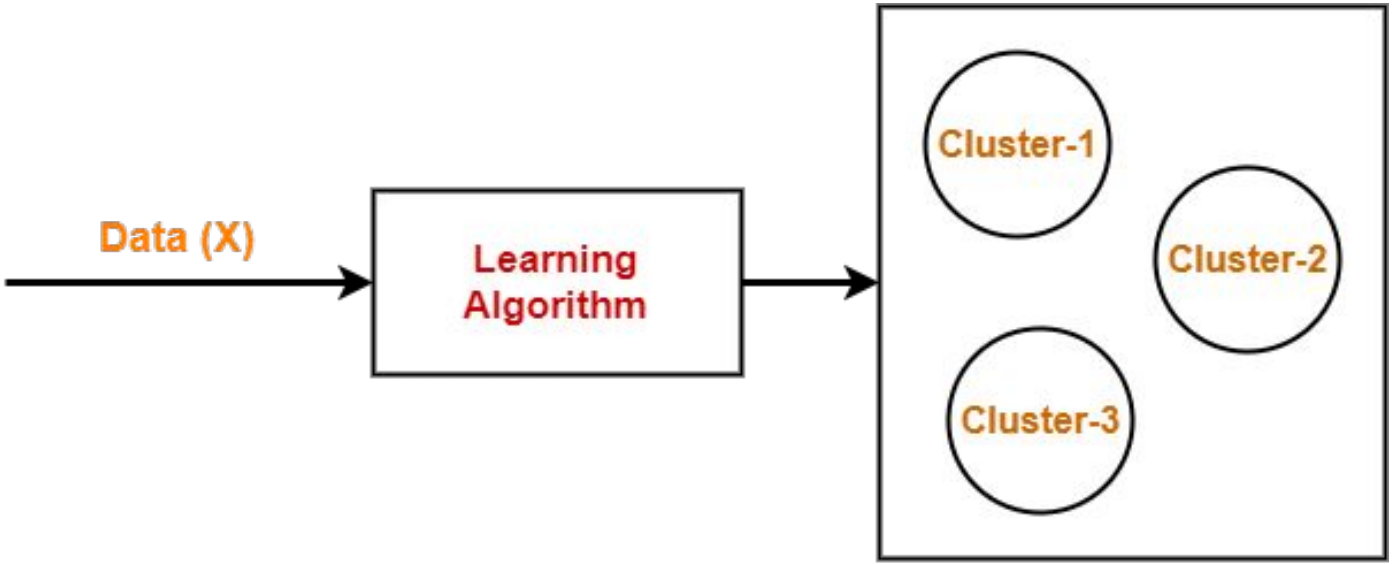
- **Classification** uses predefined classes in which objects are assigned(labeled), while

- **Clustering** identifies similarities between objects, which it groups according to those characteristics in common and which differentiate them from other groups of objects. These groups are known as "clusters".

Clustering Workflow

To cluster your data, you'll follow these steps:





Unsupervised Learning

Examples

Group stars by brightness.

Group organisms by genetic information into a taxonomy.

Group documents by topic.

Organize music by genre.

Organize music by decade.

Application areas of Clustering

Clustering has a myriad of uses in a variety of industries. Some common applications for clustering include the following:

market segmentation

social network analysis

search result grouping

medical imaging

image segmentation

anomaly detection

Types of Clustering

Several approaches to clustering exist. Each approach is best suited to a particular data distribution.

Centroid-based Clustering **(***K-Mean Clustering***)**

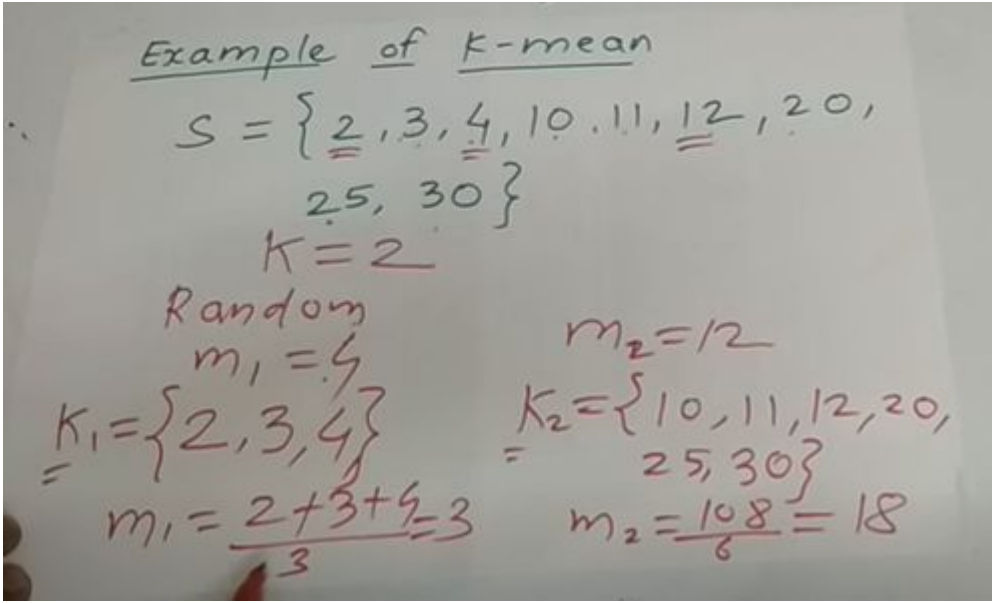
Density-based Clustering

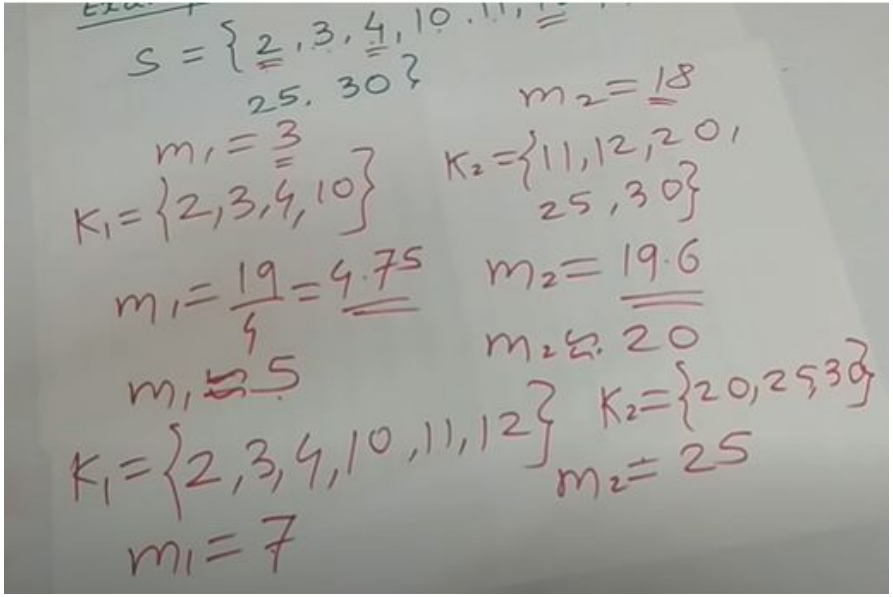
Distribution-based Clustering

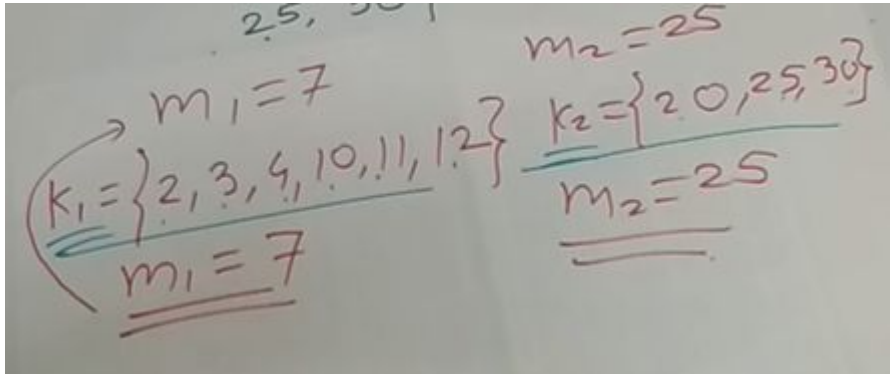
Hierarchical Clustering (*Agglomerative, Divisive****)***

K-Mean Clustering

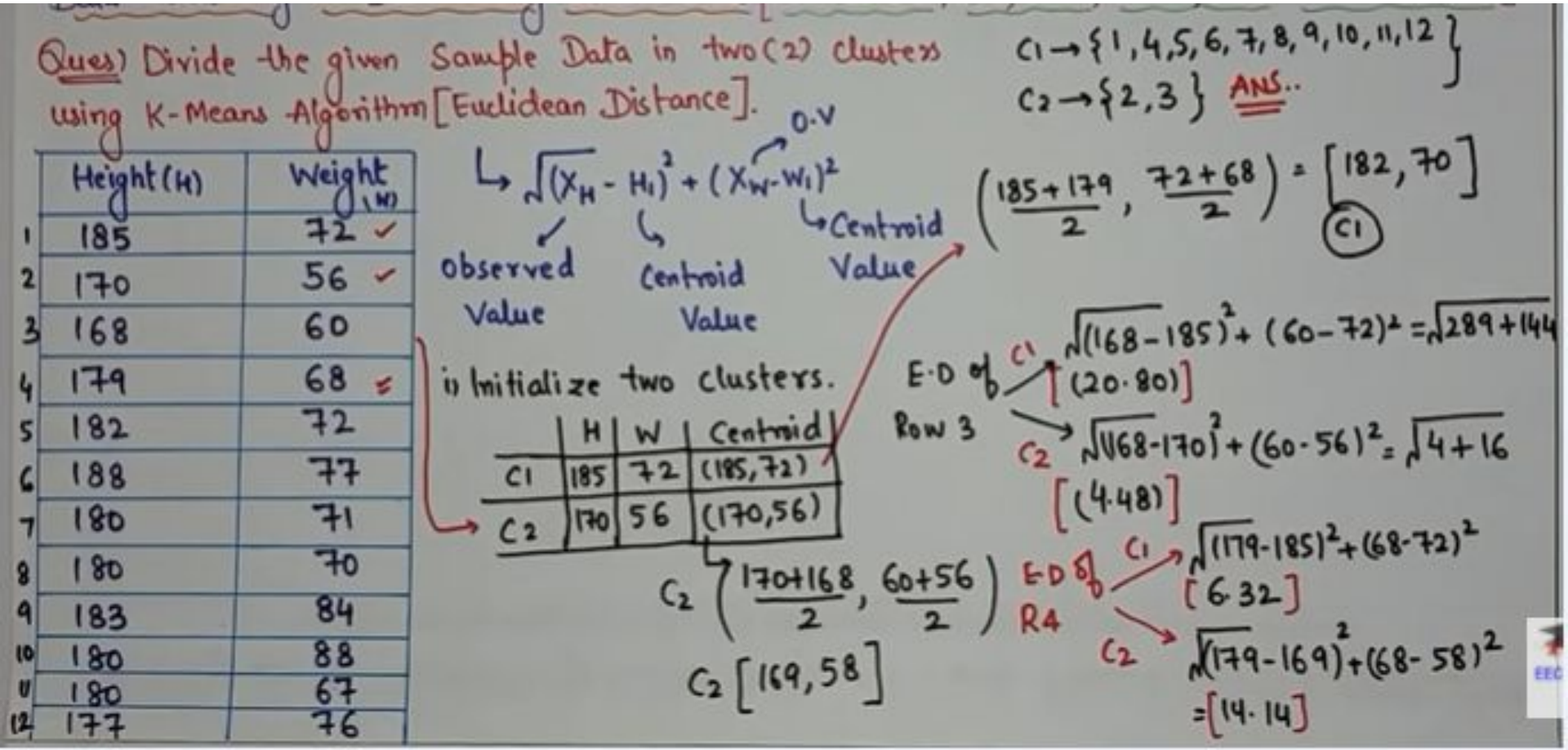
Example Problem-1





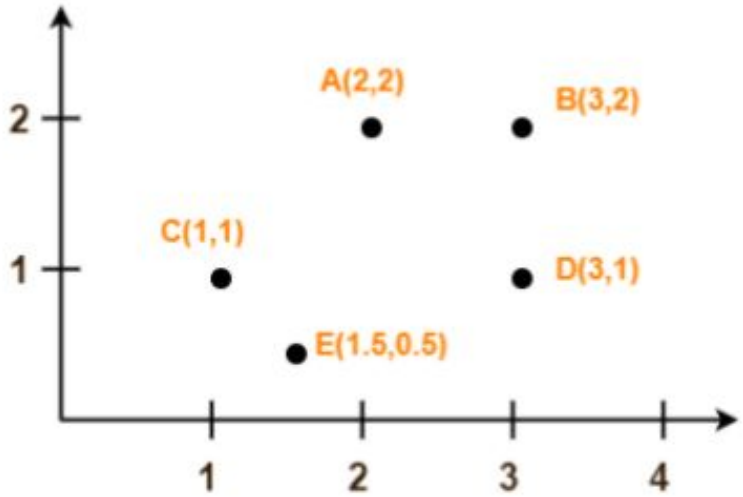


Example Problem-2



Example Problem-3

Use K-Means Algorithm to create two clusters-



Solution-

We follow the above discussed K-Means Clustering Algorithm.

Assume A(2, 2) and C(1, 1) are centers of the two clusters.

Given Points	Distance from center (2, 2) of Cluster-01	Distance from center (1, 1) of Cluster-02	Point belongs to Cluster
A(2, 2)	0	1.41	C1
B(3, 2)	1	2.24	C1
C(1, 1)	1.41	0	C2
D(3, 1)	1.41	2	C1
E(1.5, 0.5)	1.58	0.71	C2

Now,

- We re-compute the new cluster clusters.
- The new cluster center is computed by taking mean of all the points contained in that cluster.

For Cluster-01:

Center of Cluster-01

= ((2 + 3 + 3)/3, (2 + 2 + 1)/3)

= (2.67, 1.67)

For Cluster-02:

Center of Cluster-02

= ((1 + 1.5)/2, (1 + 0.5)/2)

= (1.25, 0.75)

This is completion of Iteration-01.

Next, we go to iteration-02, iteration-03 and so on until the centers do not change anymore.

K-Means Clustering

- K-Means Clustering-
  - K-Means clustering is an unsupervised iterative clustering technique.
  - It partitions the given data set into k predefined distinct clusters.
  - A cluster is defined as a collection of data points exhibiting certain similarities.
- It partitions the data set such that
  - Each data point belongs to a cluster with the nearest mean.
  - Data points belonging to one cluster have high degree of similarity.
  - Data points belonging to different clusters have high degree of dissimilarity.

How K-Means Clustering Works?

To process the learning data, the 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

It halts creating and optimizing clusters when either:

The centroids have stabilized — there is no change in their values because the clustering has been successful.

The defined number of iterations has been achieved.

K-Means Clustering Algorithm Process

K-Means Clustering Algorithm involves the following steps-

Step-01:

- Choose the number of clusters K.

Step-02:

- Randomly select any K data points as cluster centers.
- Select cluster centers in such a way that they are as farther as possible from each other.

Step-03:

- Calculate the distance between each data point and each cluster center.
- The distance may be calculated either by using given distance function or by using euclidean distance formula.

Step-04:

- Assign each data point to some cluster.
- A data point is assigned to that cluster whose center is nearest to that data point.

Step-05:

- Re-compute the center of newly formed clusters.
- The center of a cluster is computed by taking mean of all the data points contained in that cluster.

Step-06:

Keep repeating the procedure from Step-03 to Step-05 until any of the following stopping criteria is met-

- Center of newly formed clusters do not change
- Data points remain present in the same cluster
- Maximum number of iterations are reached

https://www.gatevidyalay.com/tag/k-means-clustering-numerical-example-pdf/

Examples

- K Means Clustering Algorithm - Solved Numerical Example Big Data Analytics Tutorial by Mahesh Huddar

  https://www.youtube.com/watch?v=FllcPjvztTI

- Solved numerical example 1 - K-mean clustering
  https://www.youtube.com/watch?v=P2KZisgs4A4
  Solved numerical example 1 - K-mean clustering
  https://www.youtube.com/watch?v=K2sBRVCXZqs
- K means algorithm explained with example(Very Easy)
  https://www.youtube.com/watch?v=U4MfJAmDH8s&t=12s

Pros & Cons

Pros: K-Means Clustering Algorithm offers the following advantages-

- It is relatively efficient with time complexity O(nkt) where-

  n = number of instances
  k = number of clusters
  t = number of iterations
- It often terminates at local optimum.
- Techniques such as Simulated Annealing or Genetic Algorithms may be used to find the global optimum.

Cons: K-Means Clustering Algorithm has the following disadvantages-

- It requires to specify the number of clusters (k) in advance.
- It can not handle noisy data and outliers.
- It is not suitable to identify clusters with non-convex shapes.

Task

Problem-01:

Cluster the following eight points (with (x, y) representing locations) into three clusters:

A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)

Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2).

The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as-

$$P(a, b) = |x2 – x1| + |y2 – y1|$$

Use K-Means Algorithm to find the three cluster centers after the second iteration.

Hierarchical Clustering

Example Problem-1

Ques.) Perform Agglomerative Algorithm on the following Data and Plot a dendogram using Single link Approach. The given data indicates the distance b/w elements.

Item	E	A	C	B	D
E	0	1	2	2	3
A	1	0	2	5	3
C	2	2	0	1	6
B	2	5	1	0	3
D	3	3	6	3	0

$((E,A) \rightarrow C) = \min[(E,C),(A,C)]$
$= \min[2, 2] = 2$

Pair(E,A) (1)

	(E,A)	C	B	D
(E,A)	0			
C	2	0		
B	2	1	0	
D	3	6	3	0

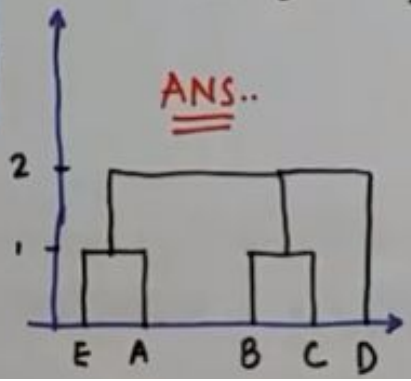
Pair(B,C) (1)

	(E,A)	(B,C)	D
(E,A)	0		
(B,C)	2	0	
D	3	3	0

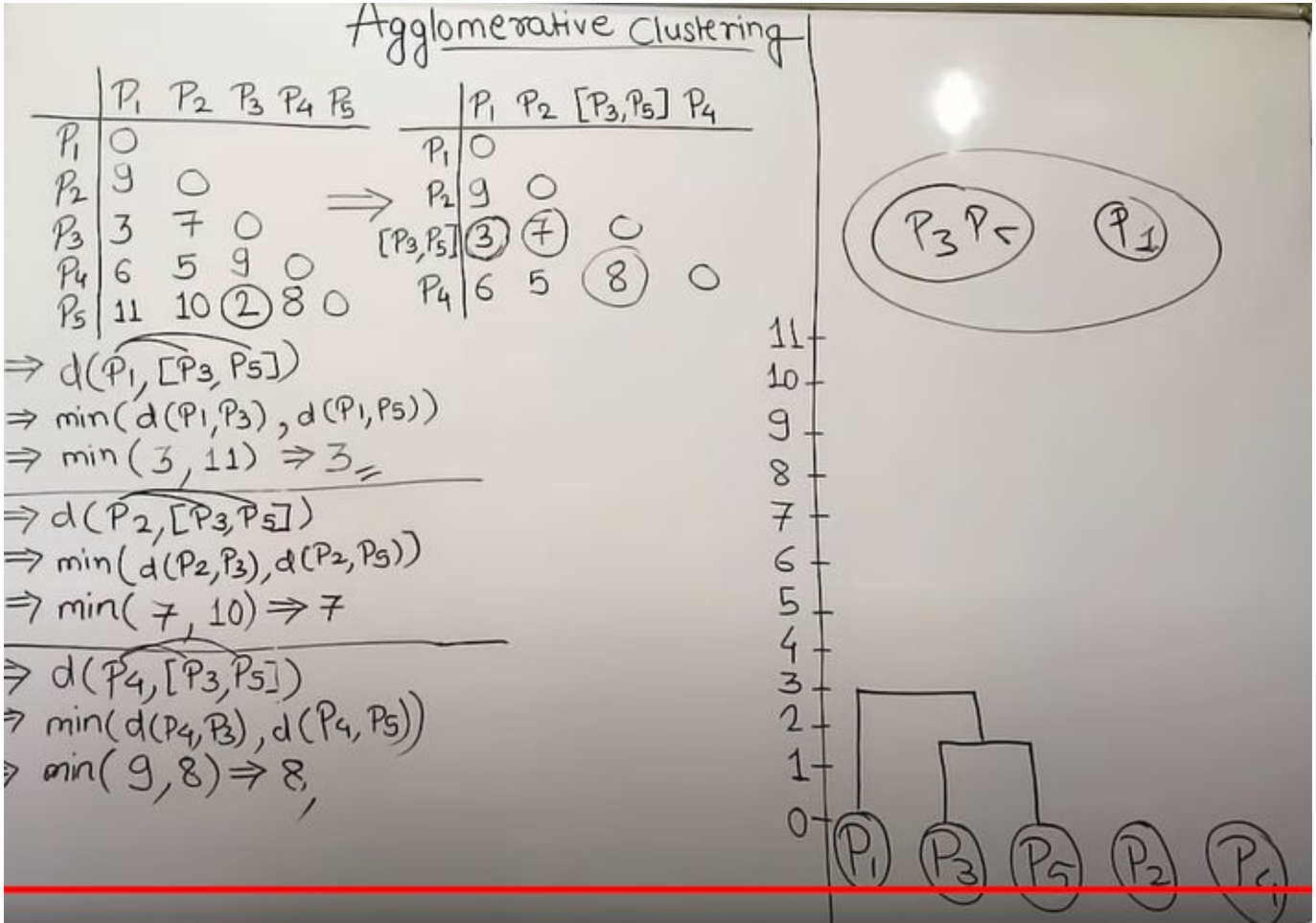
Pair (E,A) and (B,C) (2)

	((E,A),(B,C))	D
((E,A),(B,C))	0	
D	2	0

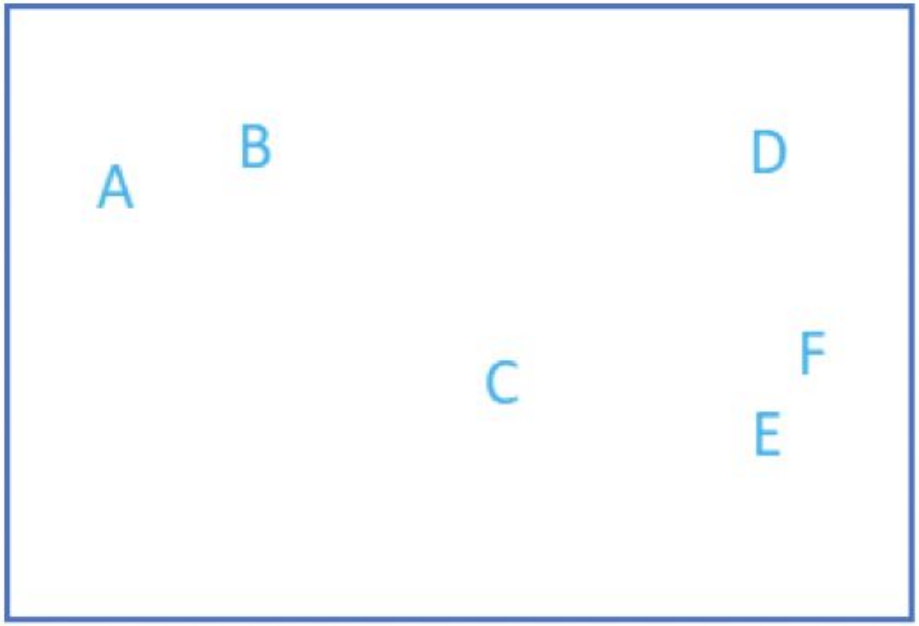
ANS..

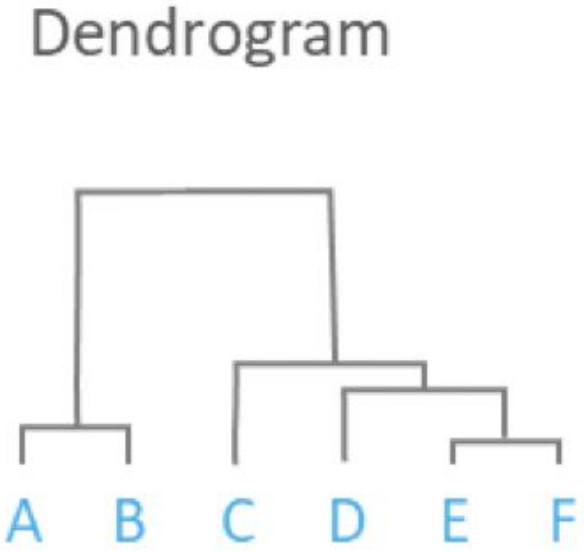


Example Problem-2



Dendogram





Hierarchical Clustering

- Hierarchical clustering, also known as hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters.
- The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

Types of Hierarchical Clustering

There are two types of hierarchical clustering:

- Divisive (top-down)

  Divisive hierarchical clustering works by starting with 1 cluster containing the entire data set, until each observation is its own cluster.

- Agglomerative (bottom-up).

  Agglomerative clustering starts with each observation as its own cluster. The two **closest** clusters are joined into one cluster. The next closest clusters are grouped together and this process continues until there is only one cluster containing the entire data set.

What does it mean to be close?

There are a variety of possible metrics: 4 most popular are:

Single-linkage: Single-linkage (nearest neighbor) is the shortest distance between a pair of observations in two clusters.

Complete-linkage: Complete-linkage (farthest neighbor) is where distance is measured between the farthest pair of observations in two clusters.

Average-linkage:
Centroid-linkage:

https://towardsdatascience.com/introduction-hierarchical-clustering-d3066c6b560e

The choice of distance

The choice of distance metric should be made based on theoretical concerns from the domain of study. That is, a distance metric needs to define similarity in a way that is sensible for the field of study.

For example, if clustering crime sites in a city, city block distance may be appropriate. Or, better yet, the time taken to travel between each location. Where there is no theoretical justification for an alternative, the Euclidean should generally be preferred, as it is usually the appropriate measure of distance in the physical world.

Linkage Criteria

After selecting a distance metric, it is necessary to determine from where distance is computed. For example, it can be computed between the two most similar parts of a cluster (single-linkage), the two least similar bits of a cluster (complete-linkage), the center of the clusters (mean or average-linkage), or some other criterion. Many linkage criteria have been developed.

As with distance metrics, the choice of linkage criteria should be made based on theoretical considerations from the domain of application. A key theoretical issue is what causes variation. For example, in archeology, we expect variation to occur through innovation and natural resources, so working out if two groups of artifacts are similar may make sense based on identifying the most similar members of the cluster.Where there are no clear theoretical justifications for the choice of linkage criteria, Ward's method is the sensible default. This method works out which observations to group based on reducing the sum of squared distances of each observation from the average observation in a cluster. This is often appropriate as this concept of distance matches the standard assumptions of how to compute differences between groups in statistics (e.g., ANOVA, MANOVA).

https://www.displayr.com/what-is-hierarchical-clustering/

Example

Agglomerative Clustering Algorithm - Plot Dendogram Solved Numerical Question 1(Hindi)

https://www.youtube.com/watch?v=Griyhs5Pjbc

Difference between K-Mean & Hierarchical clustering

- K- means clustering is simply a division of the set of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset).
- A hierarchical clustering is a set of nested clusters that are arranged as a tree.

?¿?