# APPLIED MACHINE LEARNING

## UNIT-V: CLUSTERING (UNSUPERVISED CLASSIFICATION)

# Unit 5: Applied Machine Learning

**Syllabus Unit-V: Clustering**

**Unsupervised Classification:**

- ❖ k-Means Algorithm
- ❖ Hierarchical Agglomerative Clustering
- ❖ Self Organizing Maps (SOMs)

# APPLIED MACHINE LEARNING

## UNIT-V: CLUSTERING

1

# Unit 5: Clustering (k-Means)

**k-Means (and k-Medoids) Clustering:**

- k-Means clustering method is one of the most important commonly used unsupervised classification algorithm.
- The other one being the k-Medoids clustering, a special case of k-Means classifier.
- In k-Means, the centroid of the prototype (i.e. not final) cluster is identified for clustering, which is usually mean position (computed) of all points in the m-D  feature space.
- In k-Medoids method, the most representative data point for a cluster is identified.
- In most cases of classification, the centroid may NOT be represented by a data point.
- But, in k-Medoids, it has to be a data point itself, preferably nearest to the computed centroid point.

# Unit 5: Clustering (k-Means)

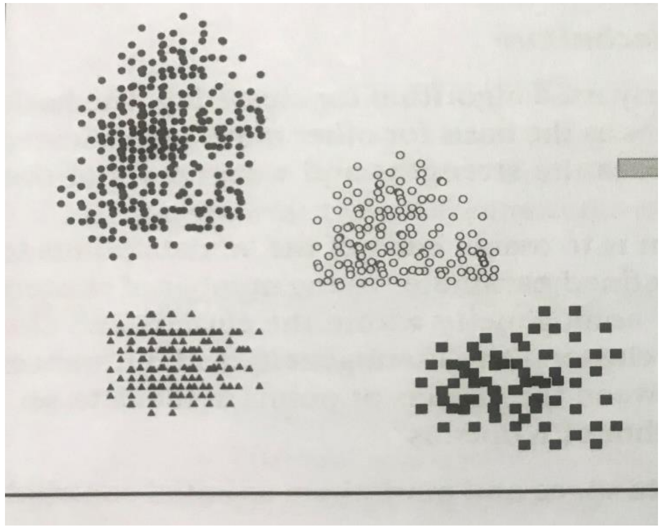
**k-Means Clustering:**

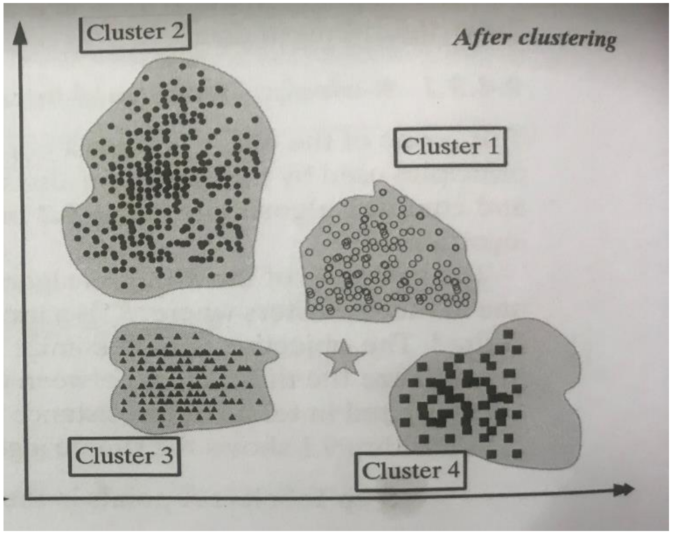
- The principle of k-Means algorithm is to assign each of the n data/sample points to one of the k-number of clusters.
- Here, k is the user defined number of clusters. The k number has to be judiciously selected by a user. How, we see later.
- There are some criteria to start with some k-value and to finally settle on best set of k-values.
- Within cluster distances determine homogeneity and inter cluster distances to determine heterogeneity or separability.
- The end goal is to maximise the homogeneity within clusters in terms of distance between objects.
- The above end goal is to be achieved by minimising within cluster (intra cluster) distances at the same time maximising the inter cluster distances at the same time.

# Unit 5: Clustering (k-Means)

## k-Means Clustering:

- **Steps in K-Means Algorithm:**
  - Select k-points in the data hyperspace (i.e. m-D feature space) and mark them as initial centroids.
  - Then assign each of the n data points/samples in the feature space to nearest centroid cluster to form initial k-clusters.
  - Compute the distance of each point in the cluster (i.e. intra cluster) from the centroid of that cluster.
  - This Computed SSD is a measure of compactness of clustering.
  - Also compute inter cluster SSD using centroid points of each cluster.

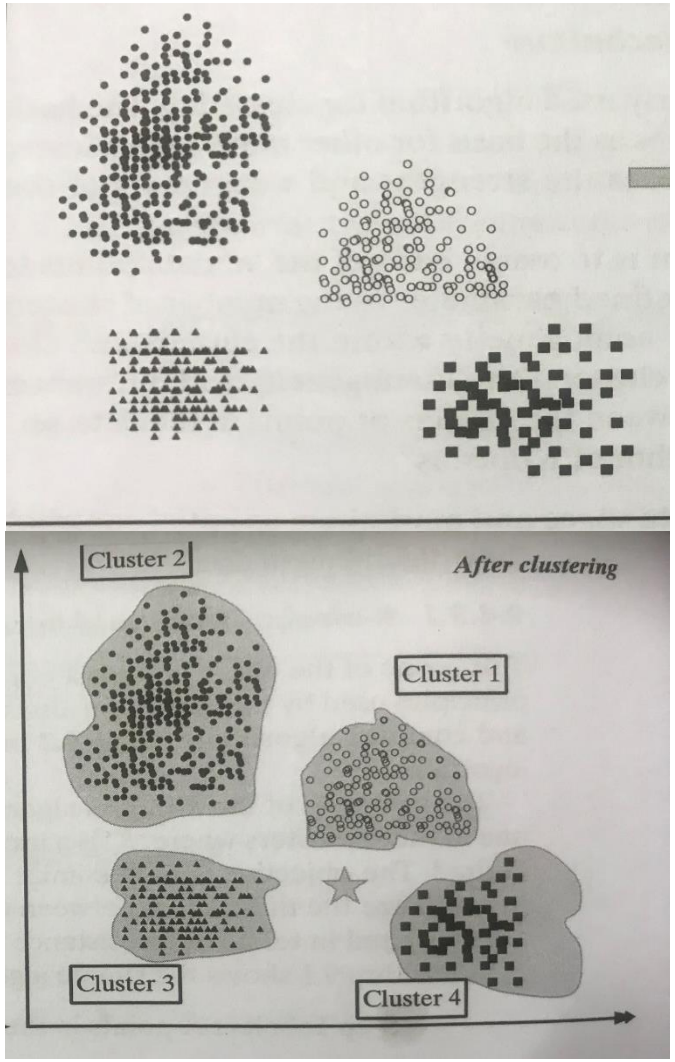




# Unit 5: Clustering (k-Means)

## k-Means Clustering: Steps in K-Means Algorithm:

- Determine the co-ordinates in feature space of the new centroid based on the previous clustering done.
- Use these as new centroids. Usually they are not the same as the previous centroids.
- Repeat (i.e. iterate) the above steps until there is no change or change less than a pre-set or pre-defined value by the user in shift of centre points.
- Identify the maximum change in value in the centroid shifts.
- In this way, finally all the data points are assigned to one of the k-clusters.
- At the end, the SSD of data points inside (i.e. intra) assigned the final cluster is minimum and that based on inter cluster distances is maximum for the given value of k.
- The ratio of inter to intra sum of square of distances has to be maximum or intra to inter SSD has to be minimum.



# Unit 5: Clustering (k-Means)

## K-Means Clustering:

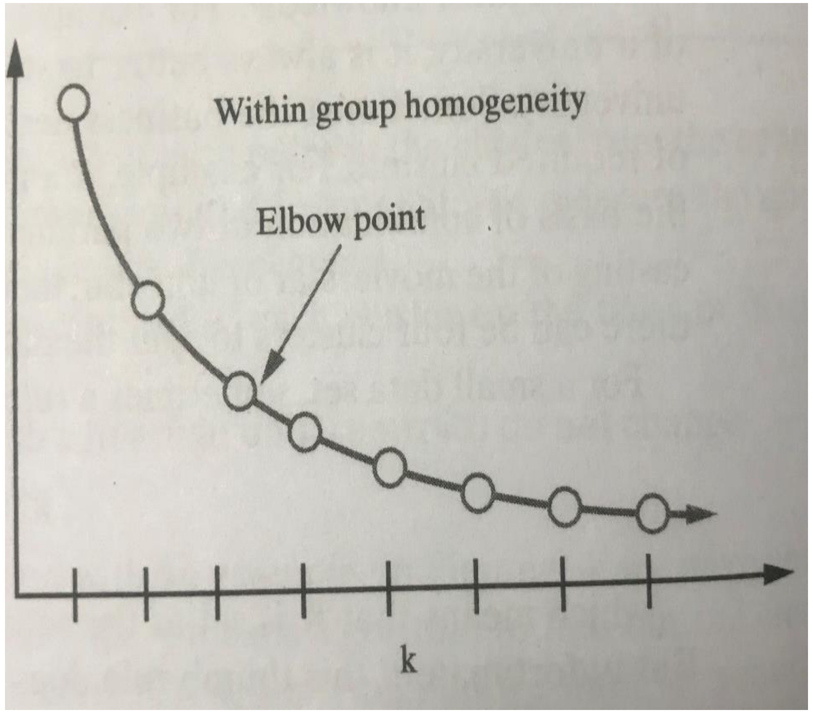
### Optimum number of classes:

- The number of clusters existing in data is not usually known a priori.
- In such a situation, which is true very often in practice, run the k-Means algorithm with varying k-max within some reasonable range.
- In every run of the algorithm, determining the ratio of inter and intra cluster distances for each value of k.
- Compute $SSD_{inter}$ and $SSD_{intra}$ of Euclidean distances for all k-Clusters after classification by applying criterion of maximum shift in cluster centres.
- Inter refers to sum of squares of Inter-cluster distances and also intra (inside) cluster distances.
- **Important Point:** These quantities are summed over all the clusters.
- Then, Compute the ratio of above SSDs for varying number of clusters.
- Plot the ratio as a function of k values for determining optimum number of classes.

# Unit 5: Clustering (k-Means)

**K-Means Clustering:**

- Plot ratio of (SSD$_{inter}$ /SSD$_{intra}$) vs k.
- Such a plot shows decreasing ratio with increasing k.
- Choose that value of k where the curve just starts flattening out.
- At this point the rate of change of slope is maximum.
- The best value of k (i.e. appropriate for the given data set) is known as the Elbow point.
- This method is known as Elbow method of deciding k-value.



# Unit 5: Clustering (k-Means)

## K-Means Clustering:

### Choosing initial centroids:

- This is another important key step in k-Means clustering based classification.
- Usually, k-random points (having m co-ordinate values) in m-D feature space are chosen as initial centroids by specifying their co-ordinates in all m-D space.
- Then, the co-ordinates are refined further in each iteration until the convergence of centroid locations.
- But, it is found this above method of choosing randomly co-ordinates of centroid leads to higher SSD in final clustering.
- Hence it is not an optimal solution.
- One effective alternative method would be:

  (i) Choose a sample points from total data set

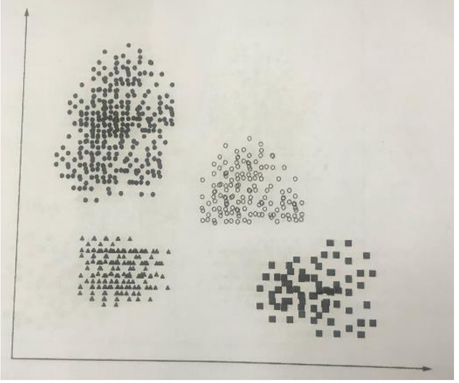
  (ii) Apply hierarchical clustering technique on it and

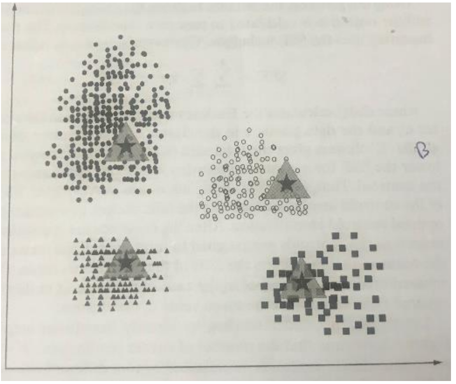
  (iii) Arrive at k-initial cluster centroids based on sample data set.

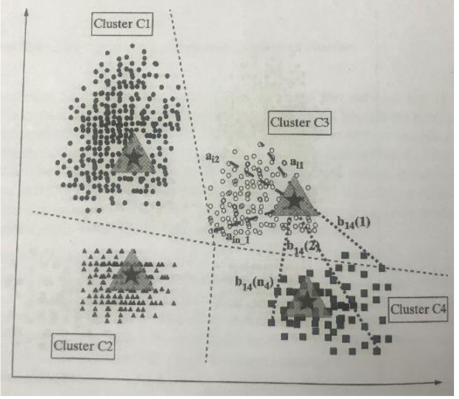
# Unit 5: Clustering (k-Means)

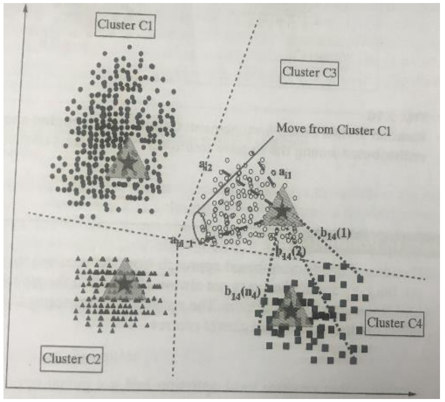
## K-Means Clustering:

- Example with 4 clusters:
  - Four clusters are shown in top-left figure.
  - Top-right figure shows initial centroids for the four clusters, which are much off from cluster centres.
  - On the basis of proximity, these sets are partitioned  into four segments known.
  - The figure showing portioning boundaries is known Voronoi diagram.





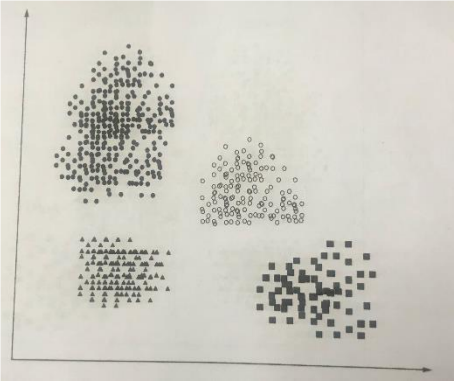


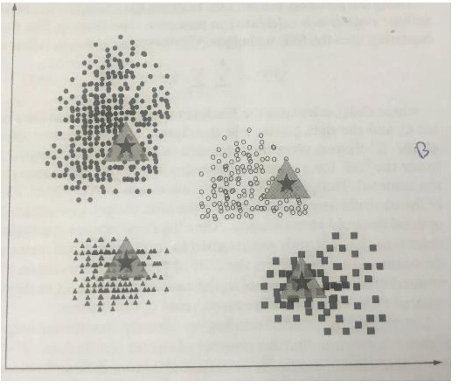


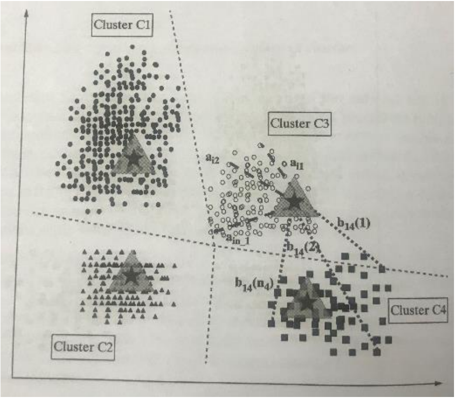
# Unit 5: Clustering (k-Means)

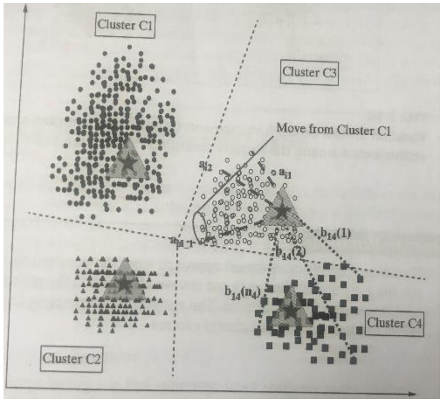
**K-Means Clustering:** Example with 4 clusters …

- Next, compute the SSE of this clustering and update the partitions of the centroids.
- The new centroids would be the means of the data points in the respective clusters.
- Within cluster distances determine homogeneity and inter cluster distances to determine heterogeneity, thus their separability.
- The objective is to finally maximise inter cluster distances and minimize intra cluster distances.
- These quantities are summed over all the clusters.









# Unit 5: Clustering (k-Means)

**K-Means Clustering:**

- There other methods such as bisecting k-Means and use of post processing to fix initial clustering locations.
- These above alternate methods are found to produce better initial centroids and better SSD for the final clusters.
- However, so far there is no final working solution or method for this problem.

# APPLIED MACHINE LEARNING

## UNIT-V: CLUSTERING

2

# Unit 5: Clustering (HAC)

## Hierarchical Agglomerative Clustering (HAC):

### Agglomerative clustering:

- Agglomerative Clustering is a type of group wise hierarchical clustering algorithm.
- It is an unsupervised machine learning technique that divides the population into several clusters.
- In this method, the data points in the same cluster are having more similarity and data points in different clusters are having more dissimilarity.
- Agglomerative clustering uses a bottom-up approach, wherein each data point starts in its own pre-undefined cluster.
- These clusters are then joined greedily, by taking the two most similar clusters together and merging them.
- Divisive clustering uses a top-down approach, wherein all data points start from a main cluster.

# Unit 5: Clustering (HAC)

**Hierarchical Agglomerative Clustering (HAC) …… :**

**Type of Agglomerative Clustering:**

- Agglomerative clustering is a strategy of hierarchical clustering.
- Hierarchical clustering is also known as Connectivity based clustering.
- Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters.

## Step of Agglomerative Clustering:

- Each data point is assigned to a single cluster in each step.
- This is based on determining the distance measures and calculated distance matrix.
- Determine the linkage criteria to merge the clusters.
- Then update the distance matrix.
- Repeat the process until every data point becomes member of any one cluster.

# Unit 5: Clustering (HAC)

## Hierarchical Agglomerative Clustering (HAC) …… :

### Main basis of Agglomerative Clustering method:

- The Agglomerative clustering can be used as long as we have pairwise distances between any two objects can be computed.
- The mathematical representation of the objects is irrelevant when the pairwise distances are given.
- Hence agglomerative clustering can be applied to non-vector data.

### Complexity of Agglomerative Clustering method:

- The time complexity of a naive agglomerative clustering is $O(n^3)$.
- It is because we exhaustively scan the N x N distance matrix for the lowest distance in each of N-1 iterations.
- Using priority queue data structure, we can reduce this complexity to $O(n^2 logn)$.
- By using some more optimizations it can be brought down to $O(n^2)$.

# Unit 5: Clustering (HAC)

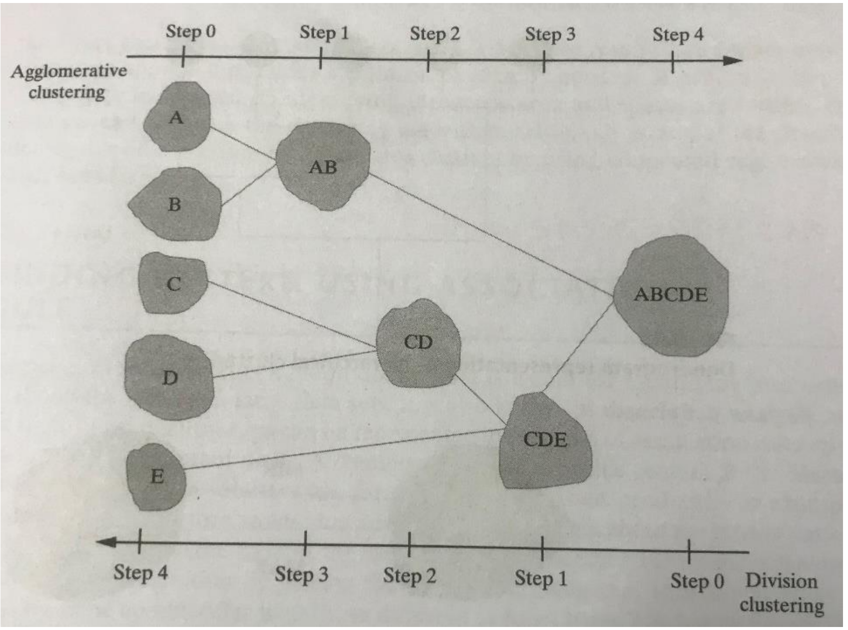
**Hierarchical Agglomerative Clustering (HAC):**

- The Hierarchical Agglomerative clustering (HAC) methods deals with partitioning data into groups at different levels, such as in case of different levels of hierarchy.
- These methods are used to group data into hierarchy or tree-like structure.
- Example: Organizing employees of a university in different departments and then within each department.
- They are grouped according to their roles, such as Assistant Professors, Associate Professors and full Professors, supervisors and lab assistants.
- This creates a hierarchical structure and make easy to visualize and analysis.

# Unit 5: Clustering (HAC)

## Hierarchical Agglomerative Clustering (HAC):

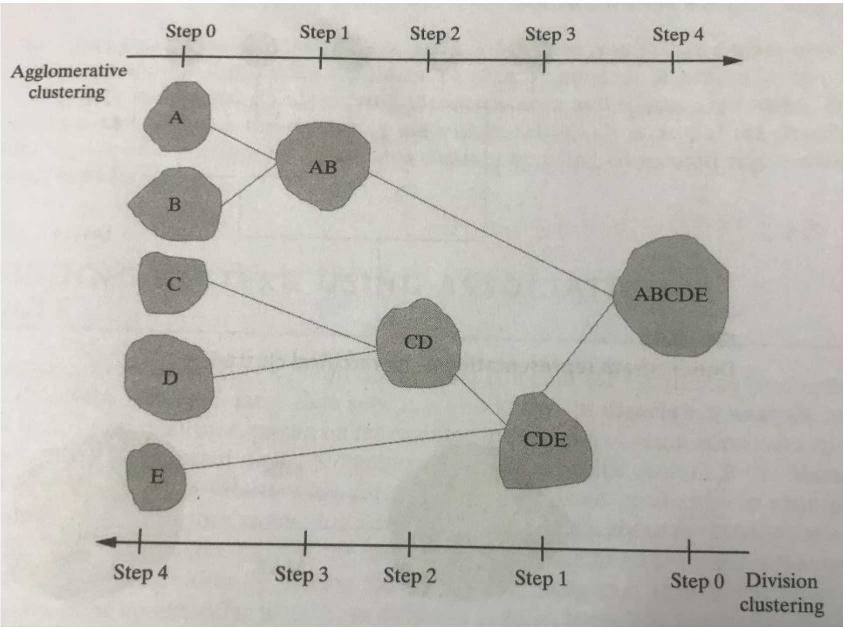
- This method is used also for discovering the structures or clusters in the data.
- Two main methods are: (i) Agglomerative Clustering, and (ii) Divisive Clustering.
- The Agglomerative clustering (HAC) is a bottom up technique and the Divisive clustering (HDC) is top-down technique.
- The HDC starts with one cluster with all given objects and the splits it iteratively to form smaller clusters.



# Unit 5: Clustering (HAC)

## Hierarchical Divisive Clustering (HDC):

- In HDC, the end of iterations is reached when the objects in the final clusters are sufficiently similar to each other or final cluster contains only one object.
- In both the methods, it is important to choose split and merge points carefully.
- This is because the next splitting or merging is the result of the previous steps of merging or splitting.

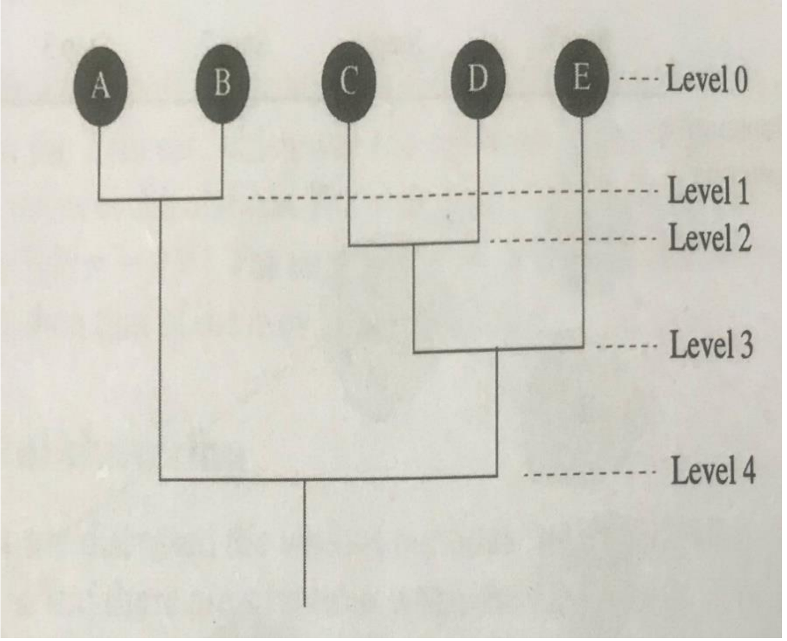


# Unit 5: Clustering (HAC)

**Hierarchical Agglomerative Clustering (HAC):**

**Dendrogram:**

- A dendrogram is commonly used tree structure for representing step-by-step creation of hierarchical clustering.
- It shows how the clusters are merged or split iteratively.
- One of the core measures of proximity between clusters is the distance between the clusters.
- In these clustering algorithms, the absolute distances are used rather than Euclidean distances.



# Unit 5: Clustering (HAC)

## Hierarchical Agglomerative Clustering (HAC):

- If $C_i$ and $C_j$ are two clusters within $n_i$ and $n_j$ objects and, if $P_i$ and $P_j$ represent objects in the cluster i and j, then following four distances can be defined:

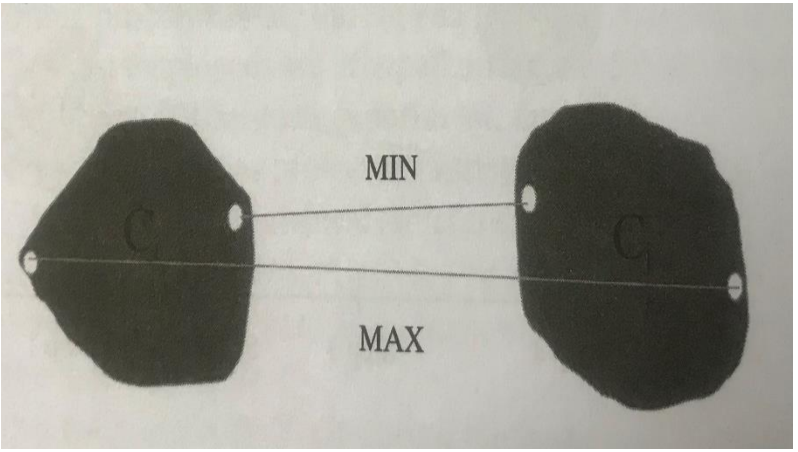
  (i) Minimum Distance $D_{min}(c_i, c_j) = \min\{|P_i\text{-}P_j|\}$

  (ii) Maximum Distance $D_{max}(c_i, c_j) = \max\{|P_i\text{-}P_j|\}$

  (iii) Mean Distance $D_{mean} = \{\{|m_i - m_j|\}$

  (iv) Average Distance $D_{avg} = \frac{\sum\{|Pi - Pj|\}}{n_i * nj}$

- The $m_i$ and $m_j$ are the mean of all points in the ith and jth clusters, respectively.



# Unit 5: Clustering (HAC)

**Hierarchical Agglomerative Clustering (HAC):**

- Usually, distance measure is used to exit the iterations.
- In HAC, the merging iterations may be stopped when minimum distance between two neighbouring clusters becomes less than the user defined threshold ($D_{min}$).
- This method of stopping criterion is called Nearest Neighbour Clustering algorithm (NNC algorithm).
- If $D_{max}$ is used, then it is called, then it called Farthest Neighbour Clustering algorithm.

# Unit 5: Clustering (HAC)

**Hierarchical Agglomerative Clustering (HAC):**

- If a user defined limit on $D_{max}$ is used, then it is called Complete Linkage Algorithm.
- Use of min or max distances between clusters are prone to the outliers and noisy data, as they may give false min and max distances.
- Therefore, the use of mean and average distances helps in avoiding such above problems and provide more consistent results.

# Unit 5: Clustering (HAC)

## Hierarchical Agglomerative Clustering (HAC):

### Advantages

- The agglomerative technique is easy to implement.
- It can produce an ordering of objects, which may be informative for the display.
- In agglomerative Clustering, there is no need to pre-specify the number of clusters.
- Agglomerative approach: This method is also called a bottom-up approach.
- In this method, each node represents a single cluster at the beginning.
- Eventually, nodes start merging based on their similarities and all nodes belong to the same cluster.
- Ultimately, optimal number of clusters are formed.

# Unit 5: Clustering (HAC)

**Hierarchical Agglomerative Clustering (HAC):**

**Disadvantages:**

- One drawback is that groups with close pairs can merge sooner before the clusters being optimal
- Above stated merger can be even if those groups have overall dissimilarity.
- Complete Linkage: Calculates similarity based on the farthest away pair.
- One disadvantage of this method is that outliers can cause less-than-optimal merging.