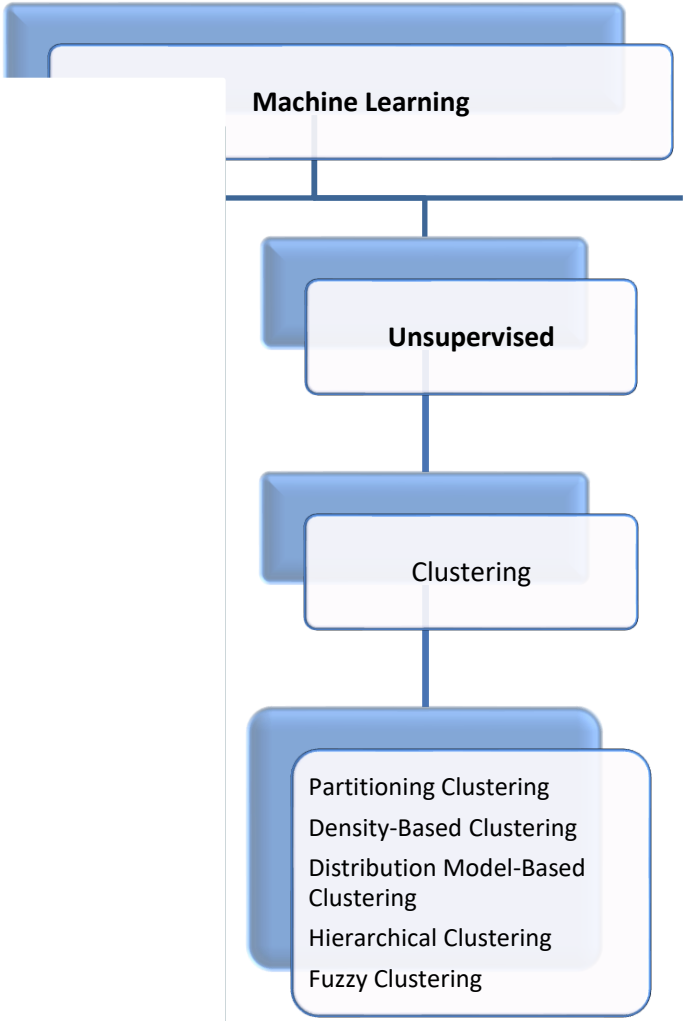
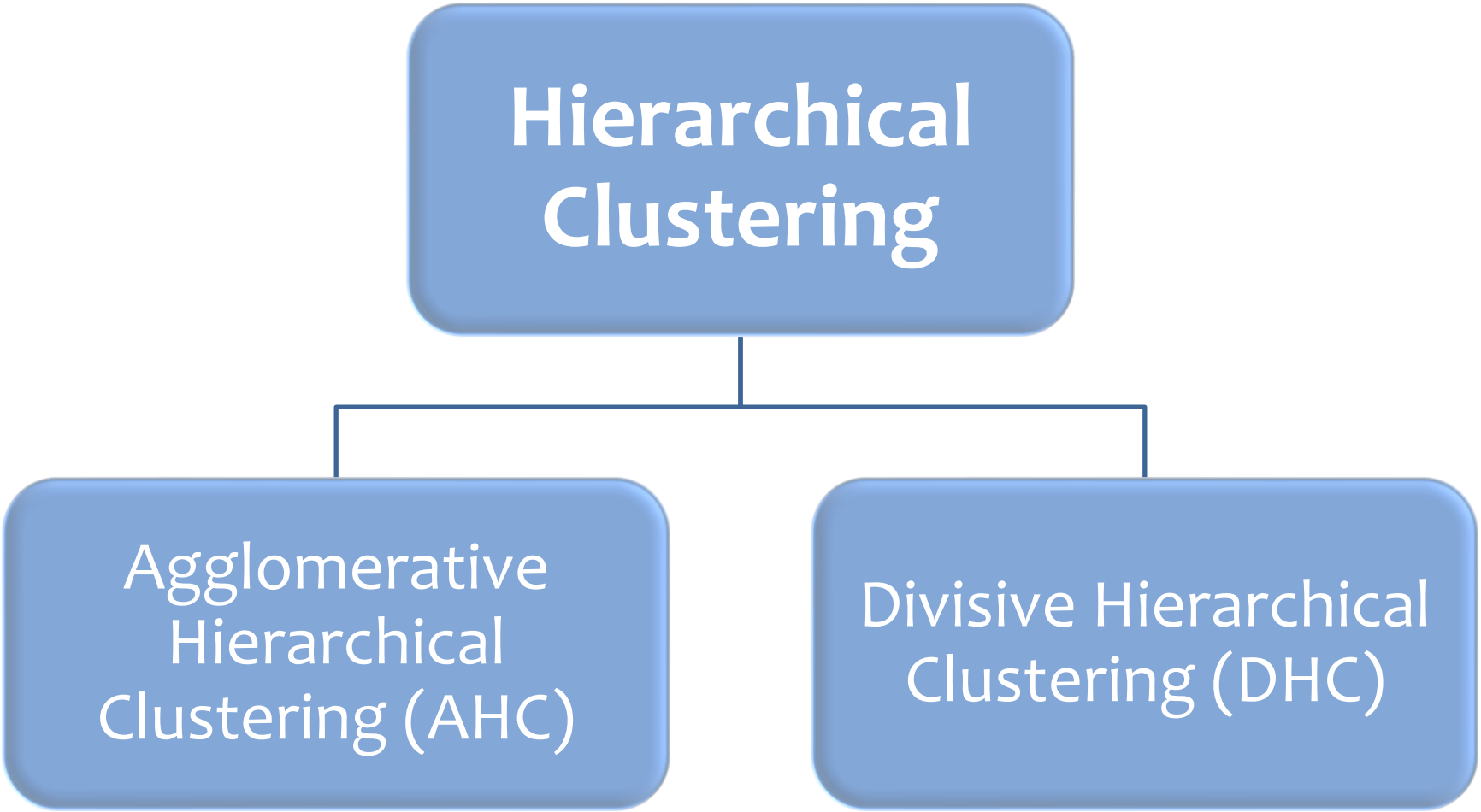
Hierarchical Clustering

MUKESH KUMAR


Machine Learning
Unsupervised
Clustering
Partitioning Clustering
Density-Based Clustering
Distribution Model-Based Clustering
Hierarchical Clustering
Fuzzy Clustering

- Hierarchical clustering is a type of clustering algorithm that groups similar data points into clusters based on their similarity. There are two main types of hierarchical clustering:

Types


Hierarchical Clustering
Agglomerative Hierarchical Clustering (AHC)
Divisive Hierarchical Clustering (DHC)

Types

- **Agglomerative clustering:** Divide the data points into different clusters and then aggregate them as the distance decreases.

- **Divisive clustering:** Combine all the data points as a single cluster and divide them as the distance between them increases.

Agglomerative Clustering

- **Agglomerative** refers to the process of clustering or grouping things together based on their similarities.

- **Bottom-Up Approach**: Starts with individual data points as separate clusters and merges them into larger clusters.

Algorithm

- Initialize the Proximity Matrix
- Make each point a cluster
  - Loop this:
    - Merge the closest two clusters (Linkage methods are used to merge clusters)
    - Update the proximity matrix
- Above loop is repeated until eventually there is one big cluster

Proximity Matrix

	x1	x2	x3	X4	x5
x1	0	D(x1,x2)	D(x1,x3)	D(x1,x4)	D(x1,x5)
x2	D(x2,x1)	0	D(x2,x3)	D(x2,x4)	D(x2,x5)
x3	D(x3,x1)	D(x3,x2)	0	D(x3,x4)	D(x3,x5)
X4	D(x4,x1)	D(x4,x2)	D(x4,x3)	0	D(x4,x5)
x5	D(x5,x1)	D(x5,x2)	D(x5,x3)	D(x5,x4)	0

	p1	P2	P3	P4	P5
P1	0				2
P2		0			
P3			0	1	
P4			1	0	
p5	2				0

	p1	P2	c1	P5
P1	0			2
P2		0		
c1			0	
p5	2			0

Single Linkage/ Minimum Linkage

- Single linkage, also known as nearest neighbor linkage, determines the distance between two clusters as the shortest distance between any two points in the two clusters.

- This method tends to produce long, chain-like clusters that are sensitive to outliers and noise in the data

Complete Linkage/Maximum Linkage

- Complete linkage, also known as farthest neighbor linkage, determines the distance between two clusters as the longest distance between any two points in the two clusters.

- This method tends to produce compact, spherical clusters that are less sensitive to outliers and noise in the data.

Average Linkage

- Average linkage determines the distance between two clusters as the average distance between all pairs of points in the two clusters.

- This method tends to produce clusters that are somewhere between the long, chain-like clusters produced by single linkage and the compact, spherical clusters produced by complete linkage.

Centroid Linkage

- First the centroid of the clusters is calculated
- Clusters with least distance between centroids are merged

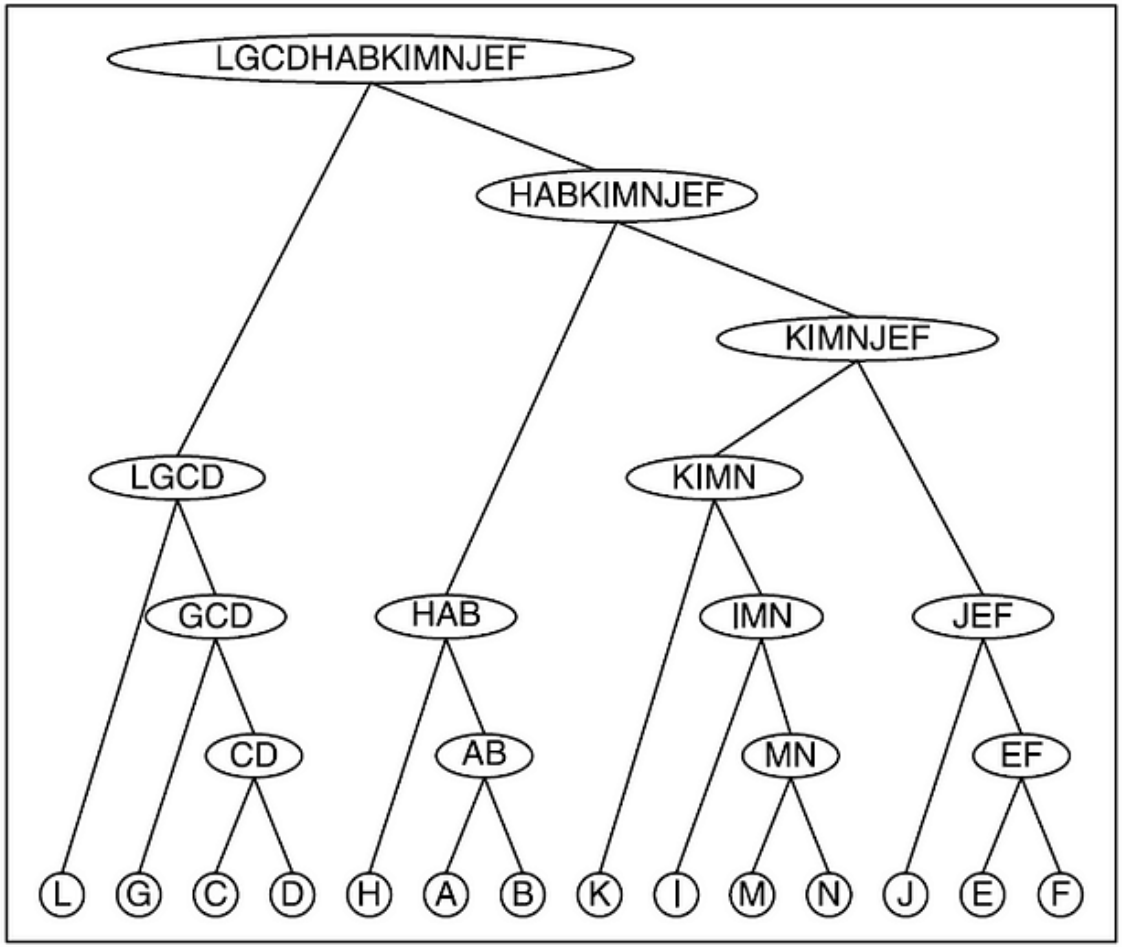
Ward Linkage

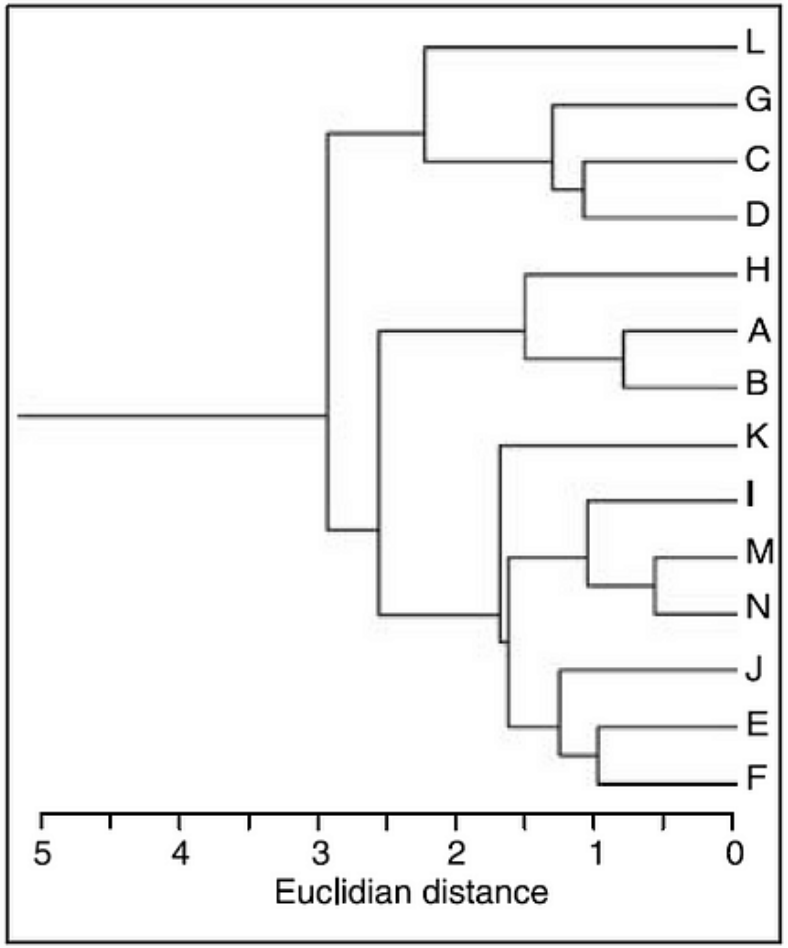
- Ward linkage, also known as minimum variance linkage.
- determines the distance between two clusters by minimising the increase in variance when the two clusters are merged.
- This method tends to produce clusters that have similar variances and sizes.

Dendrogram

- Results are often visualized using a dendrogram, which illustrates the hierarchical structure of the data and the sequence of merges.

Dendogram

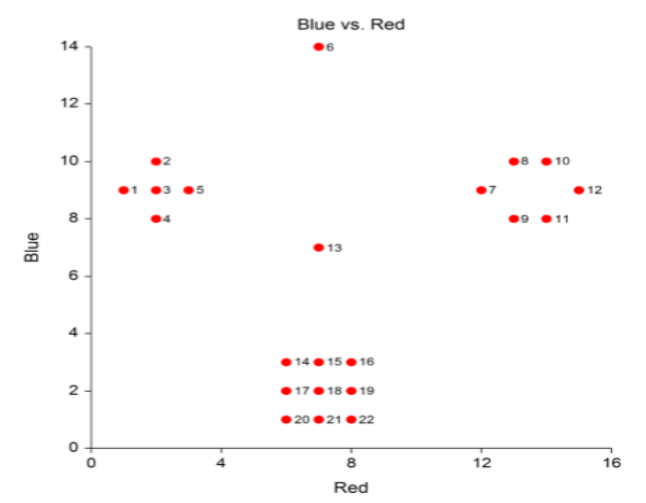

LGCDHABKIMNJEF
HABKIMNJEF
KIMNJEF
LGCD
KIMN
GCD
HAB
IMN
JEF
CD
AB
MN
EF
L
G
C
D
H
A
B
K
I
M
N
J
E
F

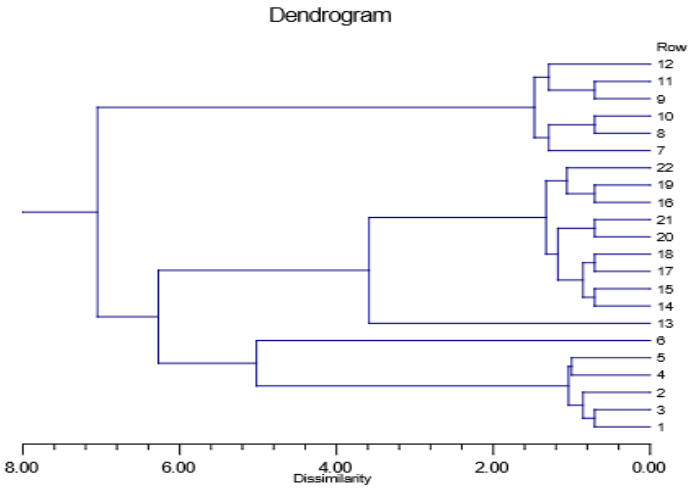

L
G
C
D
H
A
B
K
I
M
N
J
E
F
5
4
3
2
1
0
Euclidian distance

Reading Dendogram

1. Similar records are joined by lines who's vertical length reflects the relative distance between the data points

2. When viewed bottom up, the tree posses a monotonicity property. Dissimilarity between the merged clusters is monotone increasing with the level of merger





Machine Learning (Hierarchical (Agglomerative) Clustering)

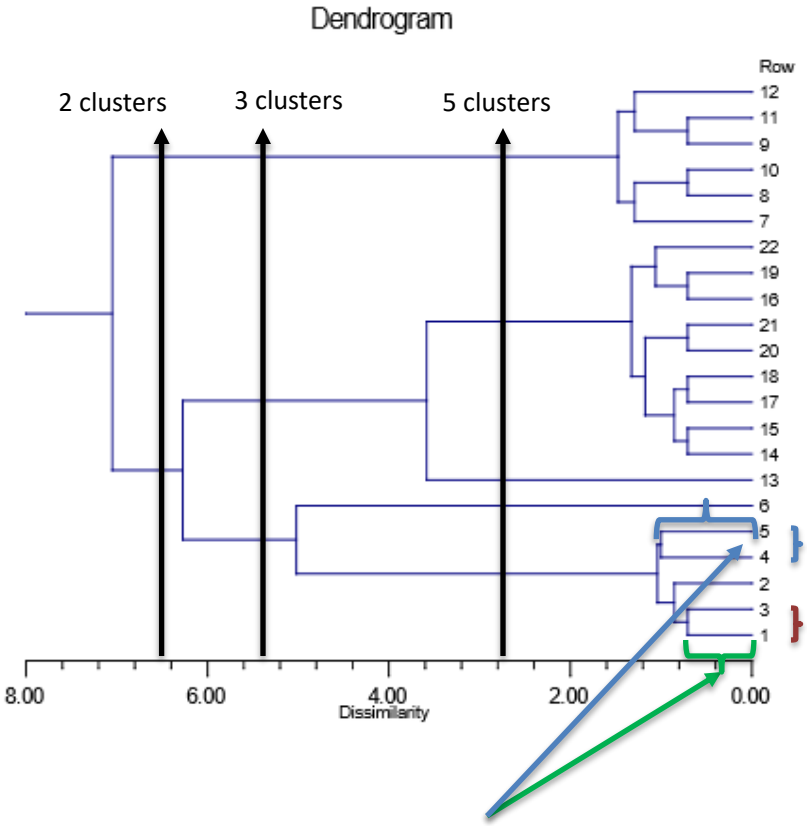
1. The horizontal axis of the dendrogram represents the distance or dissimilarity between clusters (the scale is in reverse order)

2. The vertical axis represents the objects and clusters.

3. Each fusion of two clusters is represented on the graph by the splitting of a horizontal line

4. The horizontal position of the split, shown by the short vertical bar, gives the distance (dissimilarity) between the two clusters

5. When we draw a vertical at any point on the X axis, the number of lines it cuts indicates number of clusters at that value of dissimilarity



Distance/ dissimilarity between 1,3 is less than between 4 and 5. This is reflected in the length of the horizontal bar which is longer for 4,5 compared to 1,3

Source: https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Hierarchical_Clustering-Dendrograms.pdf

- Step by step Process to find best clustering using Agglo:

- Take the actual dataset
- Build dendogram using various linkage methods
- Find the cophenet coeff for each linkage method
- Whichever linkage method give hightest coeff - choose that clustering
- Review the dendogram and decide the number of cluster
- Build the final algo with above k