University of Windsor

GENG-8900 MACHINE LEARNING

Instructor: Dr. Yasser Alginahi

Hierarchical Clustering

Meet Pachchigar

Dhruv Sutaria

Ayush Patel

Date: 10-Nov-2023

University of Windsor

Contents

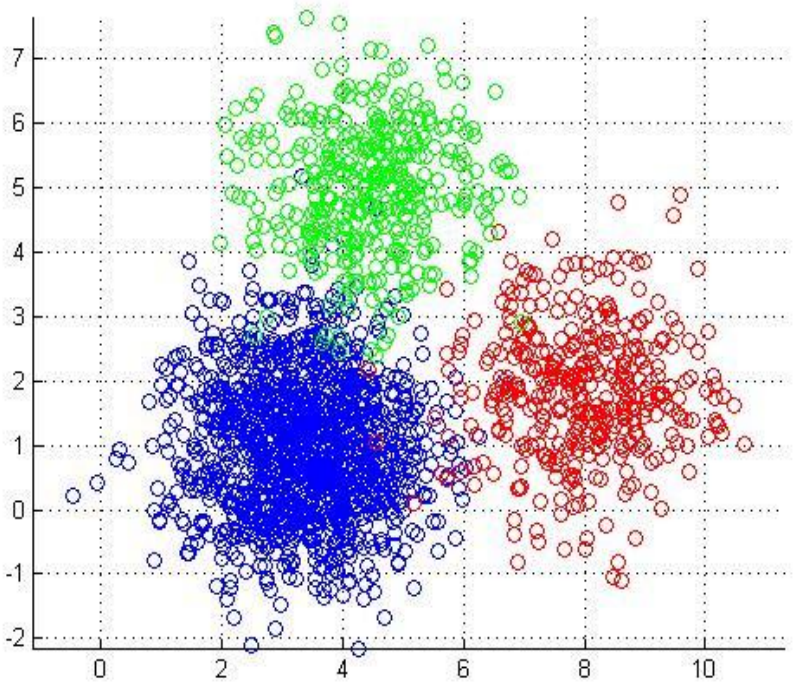
- Unsupervised learning
- Clustering
- Hierarchical clustering
- Terminologies
- Types of Hierarchical clustering
- Code implementation
- Solved Example
- Applications
- Advantages and Disadvantages
- Summary

Unsupervised learning

- Unsupervised learning is a machine learning approach that uses algorithms to analyze and discover patterns in a dataset without explicit labels.
- Unsupervised learning aims to extract meaningful information based on the inherent characteristics and similarities of data points, unlike supervised learning which relies on labeled data.
- Unsupervised learning is a paradigm in machine learning where, in contrast to supervised learning and semi-supervised learning, algorithms learn patterns exclusively from unlabeled data.

Clustering

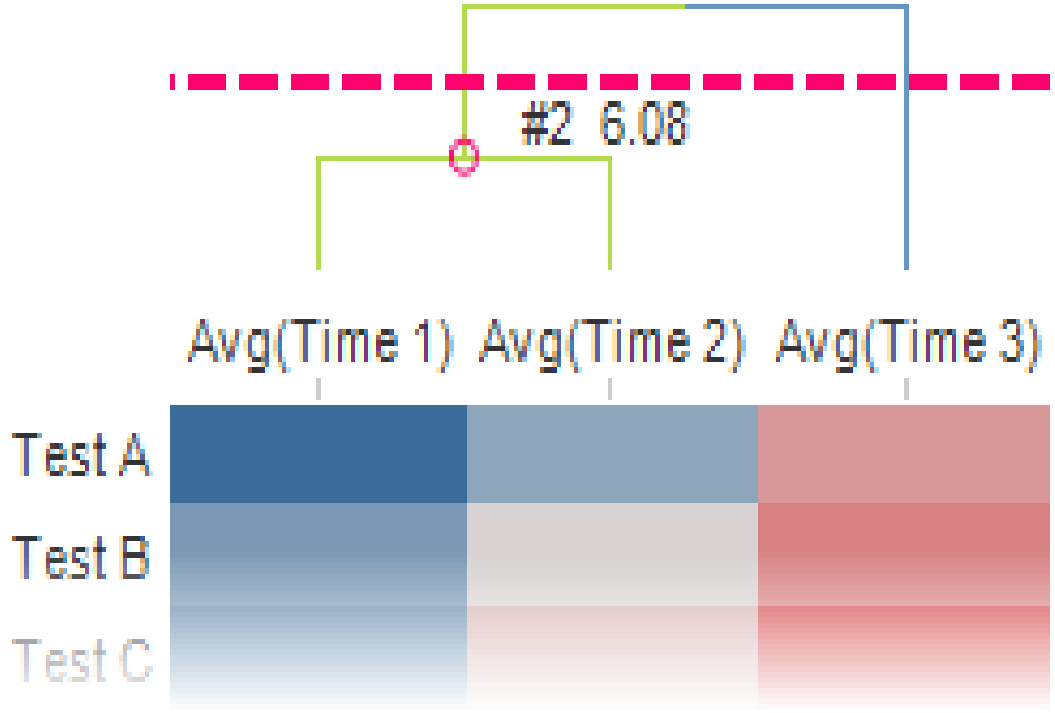
- Clustering is a technique used in machine learning and data analysis to group similar data points together based on certain criteria.
- The goal of clustering is to identify inherent patterns in the data and organize it into meaningful subgroups or clusters.
- Clustering can help us in understanding the structure of the data, discovering relationships, and making it more manageable for further analysis [1].



Source: www.kdnuggets.com/2019/09/hierarchical-clustering.html

Dendrogram

- A dendrogram is a tree-structured graph that shows the outcome of a hierarchical clustering process.
- Depending on the chosen distance metric, the clustered rows or columns' distance or similarity is displayed as the clustering's outcome [2].



Source: docs.tibco.com/pub/spotfire/6.5.0/doc/html/heat/heat_dendrograms_and_clustering.htm

Distance Metric

- The **Euclidean distance metric** measures the straight-line distance between two points in Euclidean space. It is suitable for continuous data and is given by the formula: $d(p, q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2}$
- The **Manhattan distance**, also known as L1 norm or taxicab distance, is the sum of absolute differences between coordinates, measured mathematically between two points : $d(A, B) = |x_2 - x_1| + |y_2 - y_1|$
- The **Maximum distance is distance** between two points (A, x, y) and (B, x, y) is calculated by calculating the maximum absolute difference between coordinates: $d(A, B) = \max(|x_2 - x_1|, |y_2 - y_1|)$

Hierarchical clustering

- Hierarchical clustering is a method of cluster analysis used in data science and statistics.
- It is a technique that builds a hierarchy of clusters by iteratively grouping data points or items based on their similarity or dissimilarity. The result is typically visualized as a tree-like structure called a dendrogram.
- In this method, 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 [3].

- The method known as hierarchical clustering begins with each data point as its own cluster and repeatedly combines clusters according to similarity or dissimilarity metrics. This method keeps on until every data point belongs to a single cluster or until a predefined halting condition is met.
- Hierarchical clustering is a data mining and statistics technique that iteratively groups data points based on similarity or dissimilarity [3].

Types of hierarchical clustering

Agglomerative

Initially, we consider each object to be a single cluster. Using certain procedures we start merging the clusters one by one until a single cluster remains containing all the elements.

Divisive

This is the opposite of Agglomerative. Here, we consider all objects to be a single cluster. Using the division process, we start the cluster division until each object forms a different cluster [4].

Proximity matrix

- This is the matrix consisting of the distance between each pair of data points.
- The matrix in the diagram consists of *n* points named *x*, and the $d(x_i,x_j)$ represents the distance between the points.

	x_1	x_2	x_3	...	x_n
x_1	$d(x_1,x_1)$	$d(x_1,x_2)$	$d(x_1,x_3)$	...	$d(x_1,x_n)$
x_2	$d(x_2,x_1)$	$d(x_2,x_2)$	$d(x_2,x_3)$	...	$d(x_2,x_n)$
x_3	$d(x_3,x_1)$	$d(x_3,x_2)$	$d(x_3,x_3)$	...	$d(x_3,x_n)$
...	...	...	...	...	...
x_n	$d(x_n,x_1)$	$d(x_n,x_2)$	$d(x_n,x_3)$		$d(x_n,x_n)$

Source: https://tinyurl.com/va9zr4nr

- After forming the matrix, we use linkage function to group the data points in a cluster.

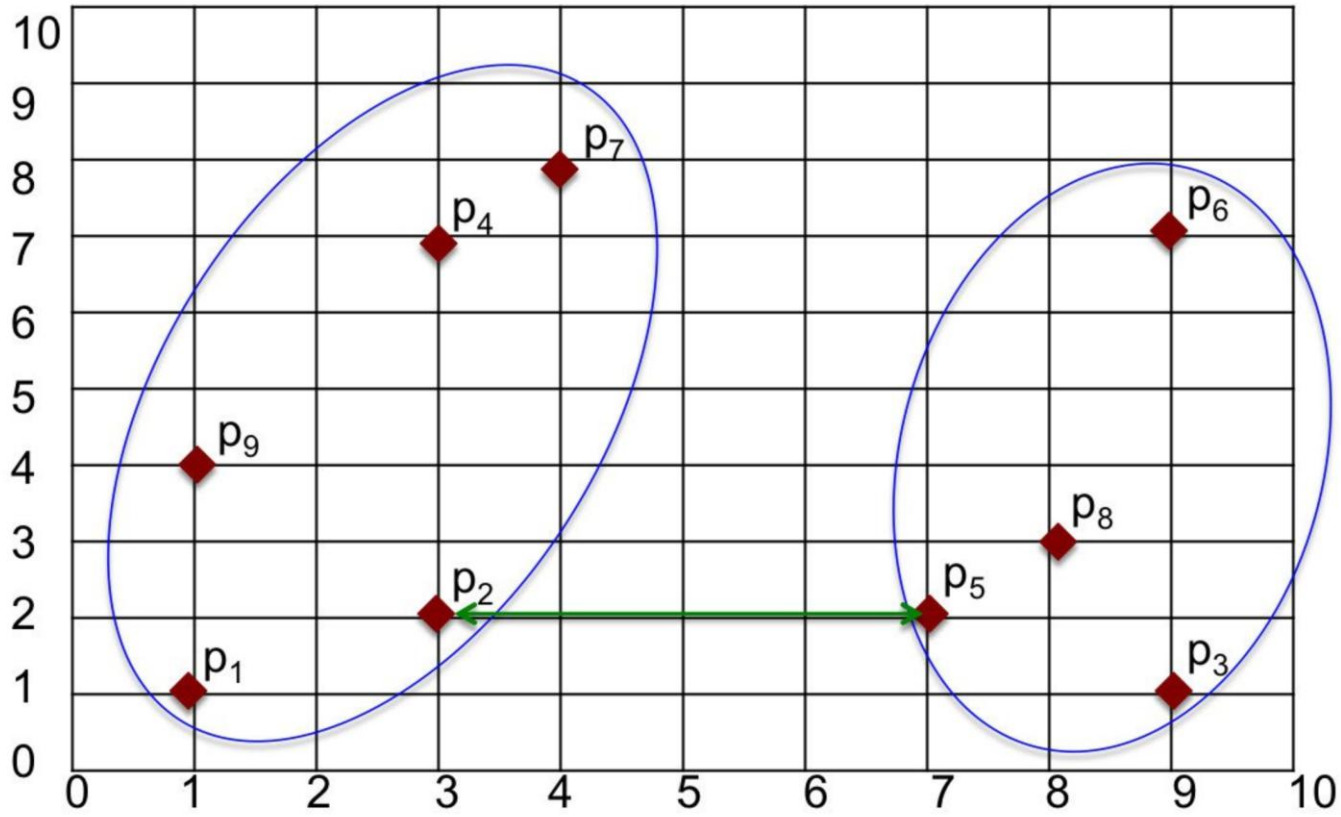
Linkage Functions

- These are the functions that take in the distance information and returns groups pairs of objects into clusters based on their similarity.
- The most common Linkage methods are

1. Single Linkage
2. Complete Linkage
3. Centroid Linkage
4. Average Linkage
5. Ward Linkage

Single Linkage

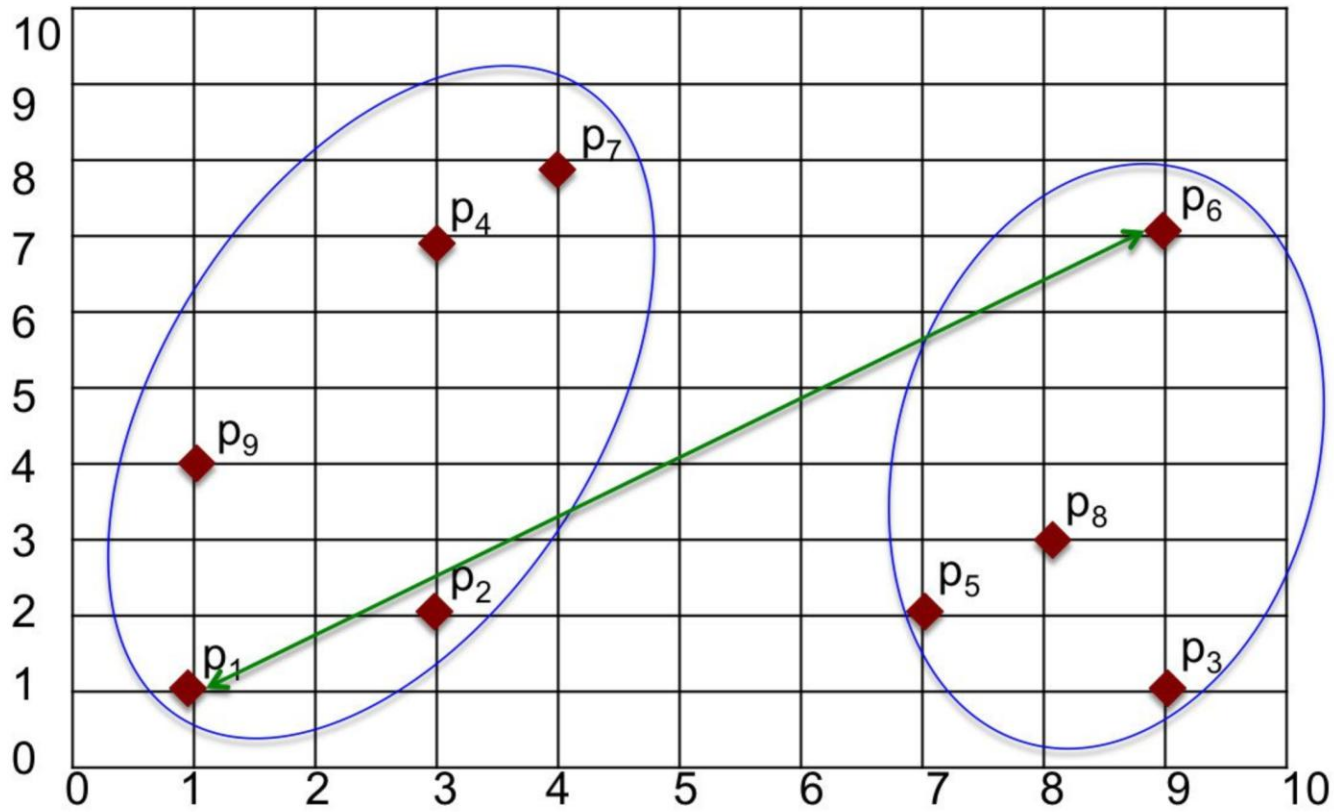
- It uses minimum distance between points in the cluster as a way to measure distance between cluster
- It is computed as D(r,s) = Min { d(x,y) : Where object x is in cluster a and object y is cluster b }



Source: https://tinyurl.com/y2b98xb4

Complete Linkage

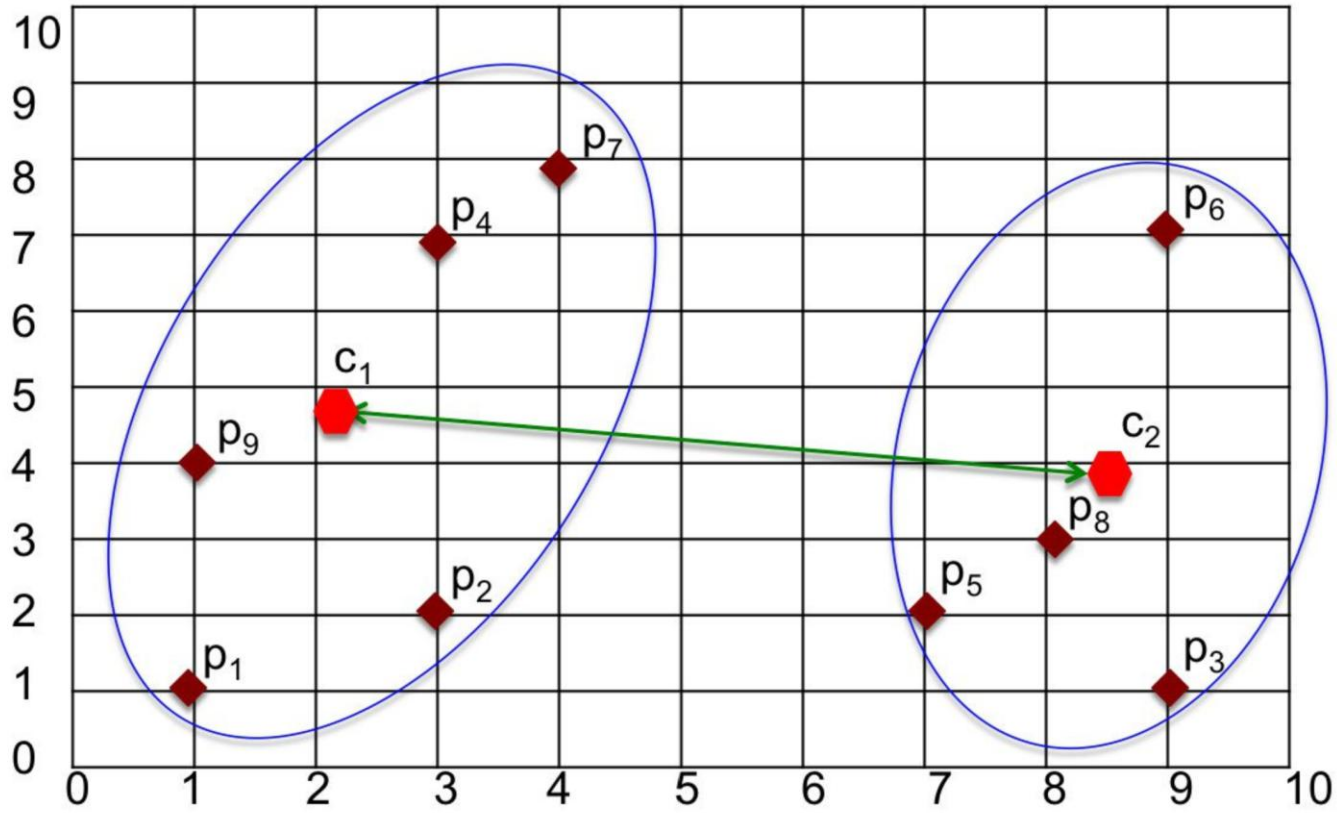
- It uses maximum distance between points in the cluster as a way to measure distance between cluster
- It is computed as D(r,s) = Max { d(x,y) : Where object x is in cluster a and object y is cluster b }



Source: https://tinyurl.com/y2b98xb4

Centroid Linkage

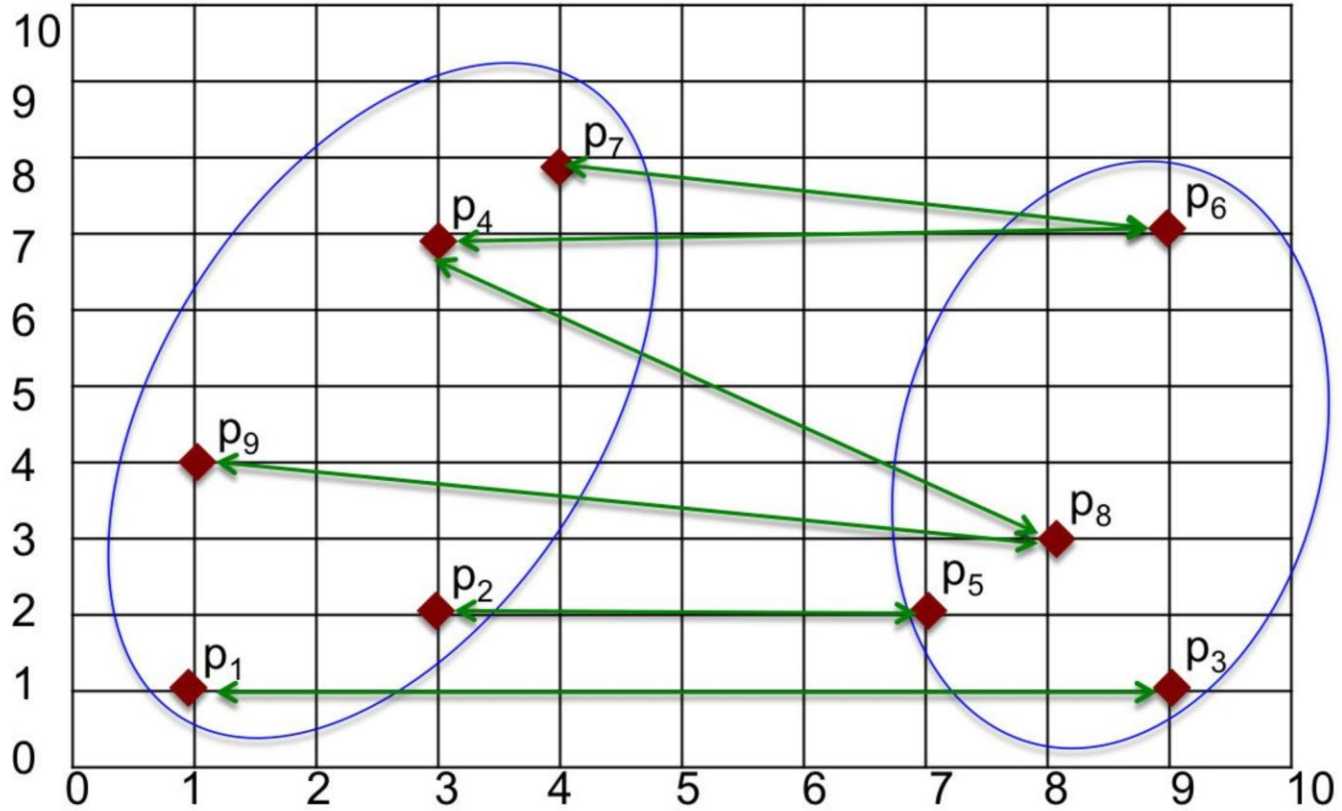
- It uses distance between centroids of the cluster as a way to measure the distance between clusters.
- It is computed as D(r,s) ={ d($\bar{x}$, $\bar{y}$) : Where $\bar{x}$ is the mean vector of cluster a and $\bar{y}$ is the mean of cluster b }



Source: https://tinyurl.com/y2b98xb4

Average Linkage

- It uses distance between all pair of objects where each pair is made up of one object from each group.
- It is computed as D(r,s) = T(rs) / { N(r) * N(s)}
- Here T(rs) is the sum of all pairwise distances between cluster r and cluster s. N(r) and N(s) are the sizes of the clusters r and s, respectively.



Source: https://tinyurl.com/y2b98xb4

Ward Linkage

- The distance between two clusters is related to how much the variance will increase when combined.
- This is computed as $$\Delta(A,B) = \sum_{i \epsilon AUB} \|\vec{x_i} - \vec{m_{AUB}}\|^2 - \sum_{i \epsilon A} \|\vec{x_i} - \vec{m_A}\|^2 - \sum_{i \epsilon B} \|\vec{x_i} - \vec{m_b}\|^2$$
$$= \frac{n_A n_B}{n_A + n_B} \|\vec{m_A} - \vec{m_B}\|^2$$

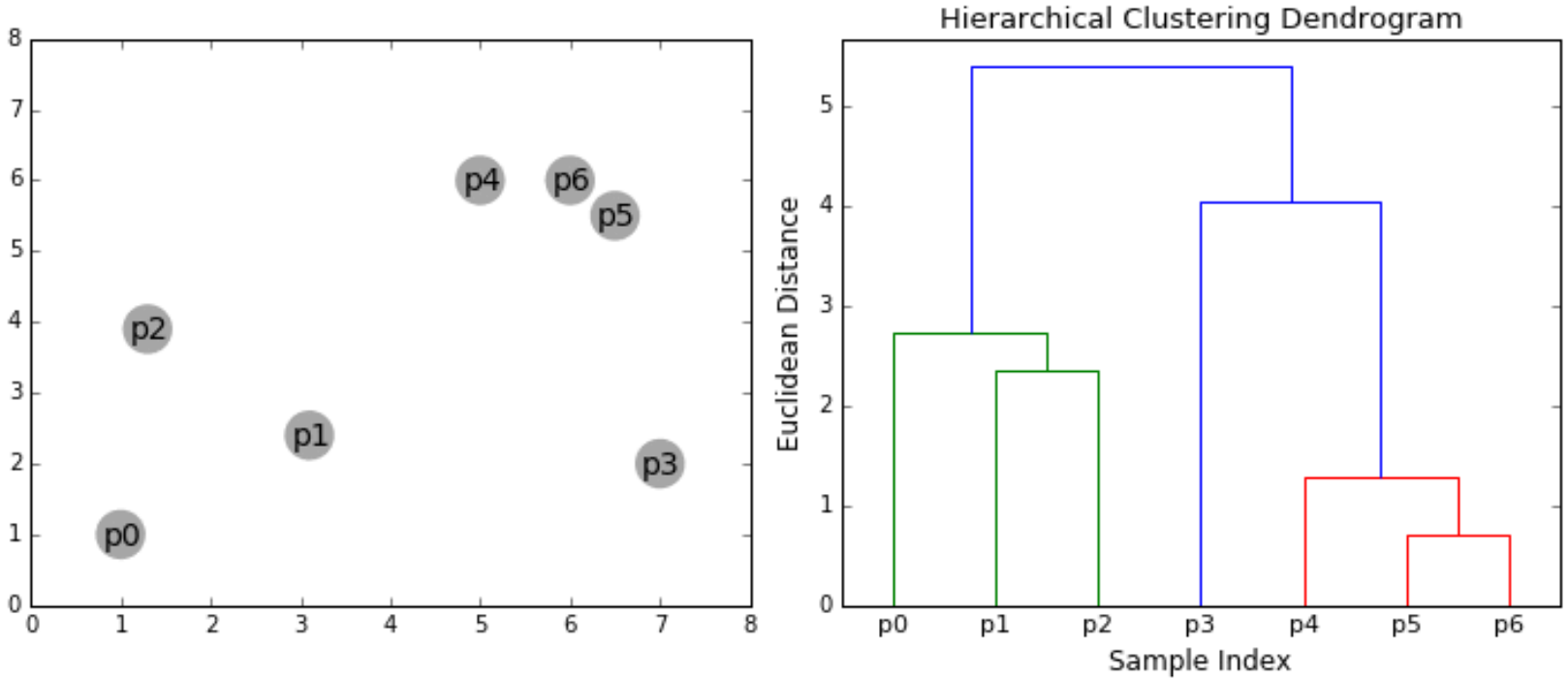
- Here, x(i) is individual points in a cluster, n(A) and n(B) are the number of objects in cluster A and B, m(A) and m(B) are mean of cluster A and cluster B respectively, and m(AUB) is the new mean created after combining cluster A and B
- After calculation, the one with lowest Δ(A,B) will be clustered together

Agglomerative hierarchical clustering

- Agglomerative clustering is an approach that begins with individual data points as separate clusters and gradually combines them based on their similarity until they eventually form a single large cluster containing all the objects [5]. Here are the steps to do it:

1. Compute the proximity matrix using a distance metric.
2. Use a linkage function to group objects into a hierarchical cluster tree based on the computed distance matrix from the above step.
3. Data points with close proximity are merged together to form a cluster.
4. Repeat steps 2 and 3 until a single cluster remains.

Agglomerative hierarchical clustering



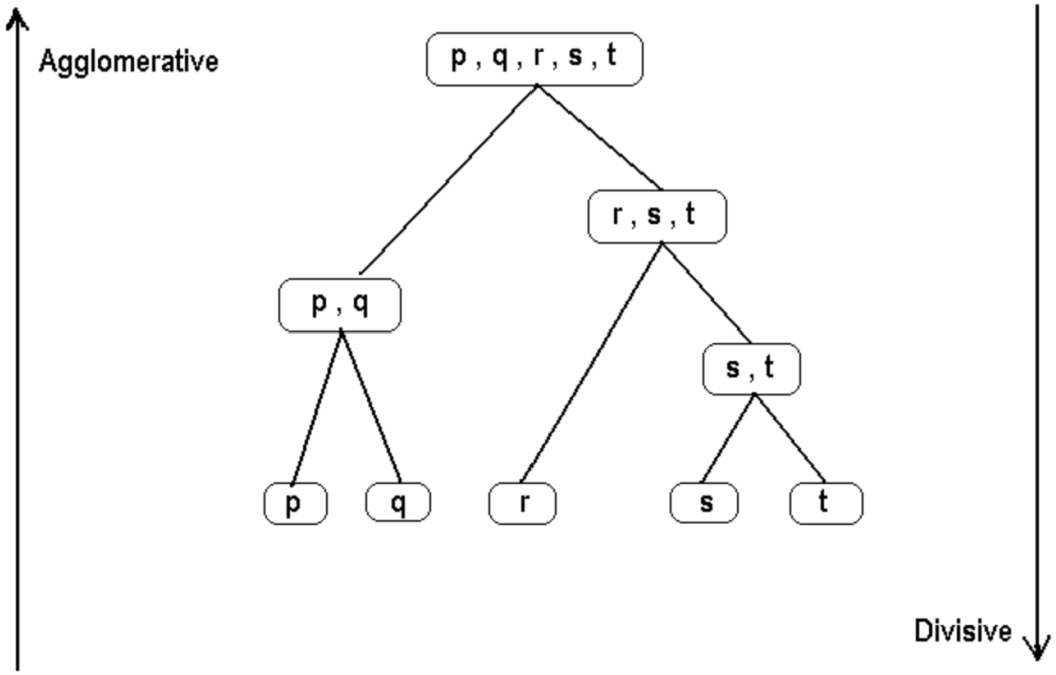
Source: https://tinyurl.com/8xepea52

Divisive Hierarchical Clustering

- Divisive clustering, in contrast to agglomerative clustering, begins with all data points grouped into a single large cluster and then proceeds to split this cluster into smaller, more diverse clusters one step at a time until each data point resides in its own individual cluster. It's a top-down approach that gradually divides the data into finer subgroups.
- Here are the steps:

1. Split into clusters using any flat-clustering method, say K-Means.
2. Choose the best cluster among the clusters to split further, choose the one that has the largest Sum of Squared Error (SSE).
3. Repeat steps 2 and 3 until a single cluster is formed .

Difference between Agglomerative and Divisive clustering



Source: https://tinyurl.com/5n6c4k8e

Code Example

https://colab.research.google.com/drive/1YGfQm4wIuoTzCojGXxG2l5x7nt2oQhYX#scrollTo=0V3wryWmwW9v

Hierarchical clustering step by step:

Step-1

	18	22	25	27	42	43
18	0	4	7	9	24	25
22	4	0	3	5	20	21
25	7	3	0	2	17	18
27	9	5	2	0	15	16
42	24	20	17	15	0	1
43	25	21	18	16	1	0

	18	22	25	27	42	43
18	0	4	7	9	24	25
22	4	0	3	5	20	21
25	7	3	0	2	17	18
27	9	5	2	0	15	16
42	24	20	17	15	0	1
43	25	21	18	16	1	0

(42,43)

Step-2

	18	22	25	27	42,43
18	0	4	7	9	24
22	4	0	3	5	20
25	7	3	0	2	17
27	9	5	2	0	15
42,43	24	20	17	15	0

	18	22	25	27	42,43
18	0	4	7	9	24
22	4	0	3	5	20
25	7	3	0	2	17
27	9	5	2	0	15
42,43	24	20	17	15	0

(42,43), (25,27)

Step-3

	18	22	25,27	42,43
18	0	4	7	24
22	4	0	3	20
25,27	7	3	0	15
42,43	24	20	15	0

	18	22	25,27	42,43
18	0	4	7	24
22	4	0	3	20
25,27	7	3	0	15
42,43	24	20	15	0

(42,43), ((25,27),22)

Final step

Step-4

	18	22,25,27	42,43
18	0	4	24
22,25,27	4	0	15
42,43	24	15	0

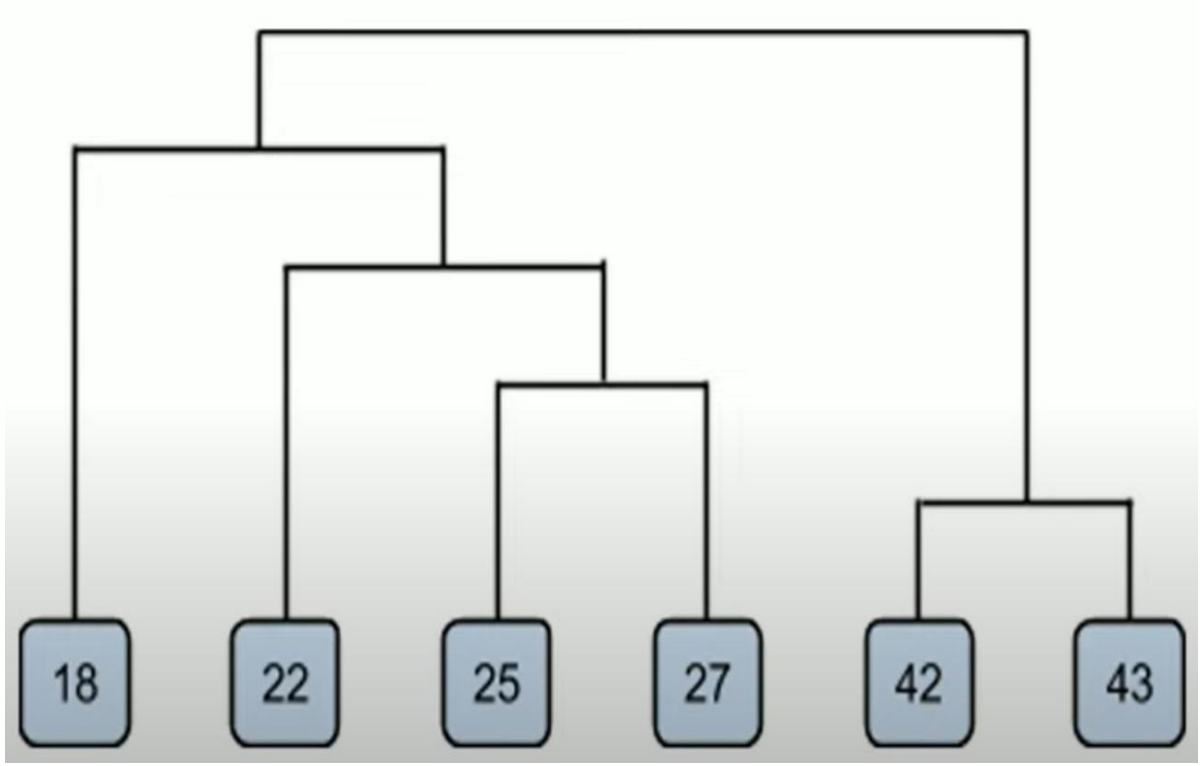
	18	22,25,27	42,43
18	0	4	24
22,25,27	4	0	15
42,43	24	15	0

(42,43), (((25,27),22),18)

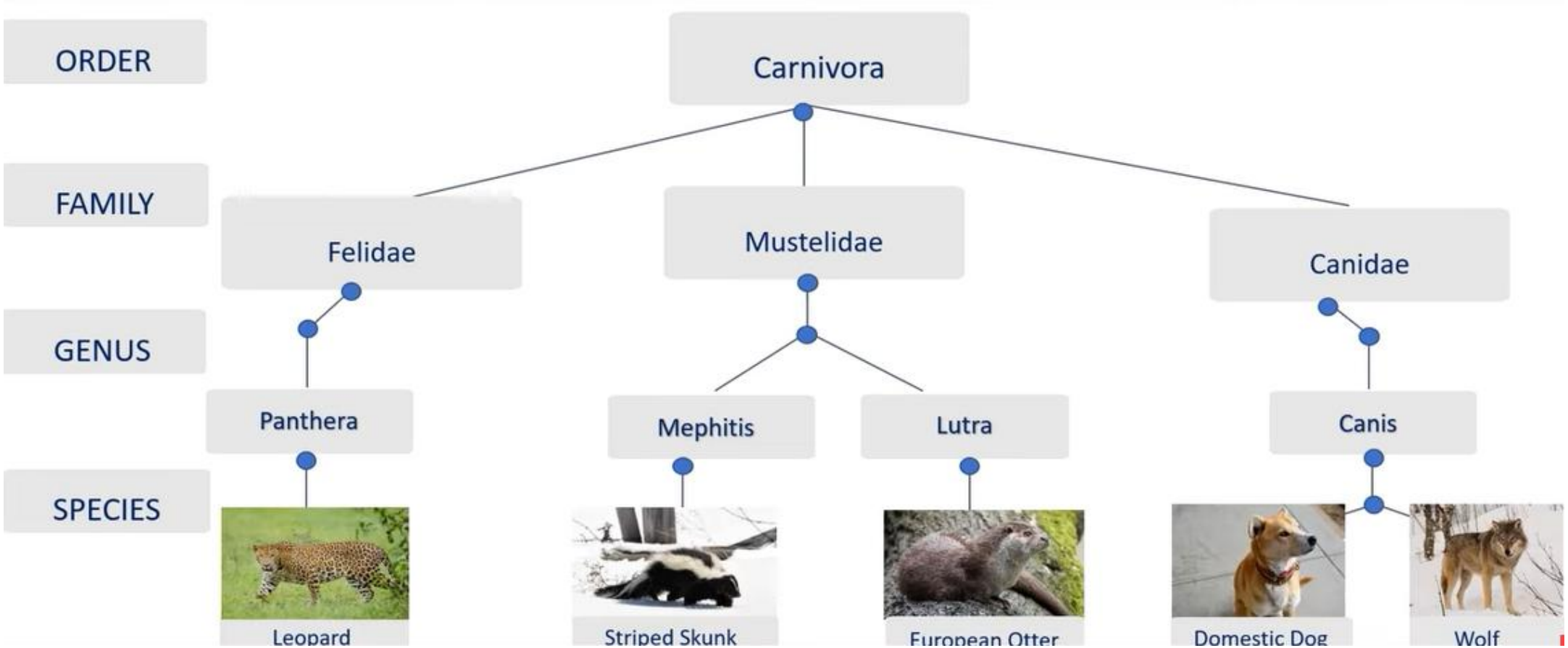
	18,22,25,27,42,43
18,22,25,27,42,43	0

((42,43), (((25,27),22),18))

Dendrogram of the example:



Application of Hierarchical Clustering:



Applications:

- There are many real-life applications of Hierarchical clustering. They include:
- Bioinformatics: grouping animals according to their biological features to reconstruct phylogeny trees
- Business: dividing customers into segments or forming a hierarchy of employees based on salary.
- Image processing: grouping handwritten characters in text recognition based on the similarity of the character shapes.
- Information Retrieval: categorizing search results based on the query.

Advantages

- Interpretability: Dendrograms provide an intuitive visualization of cluster relationships.
- No Predefined Clusters: Hierarchical clustering doesn't require specifying the number of clusters beforehand.
- Flexibility: You can choose the granularity of clusters by cutting the dendrogram.
- Handling Non-Globular Shapes: Suitable for complex data distributions.
- Agglomerative and Divisive: Supports both bottom-up and top-down approaches.
- Incremental Clustering: Accommodates new data points without re-running the entire process.
- Hierarchical Relationships: Reveals not only clusters but their hierarchical relationships.

Disadvantages

- Computationally Intensive: High time complexity for large datasets.
- Memory Usage: Can be memory-intensive, less suitable for very large datasets.
- Lack of Scalability: Challenging for high-dimensional data.
- Noise Sensitivity: Sensitive to noise, which can impact clustering results.
- Subjective Cutting: Deciding where to cut the dendrogram can be subjective.
- Inefficiency with Large Datasets: Complex dendrograms in large datasets can be hard to interpret.
- Outlier Handling: Hierarchical clustering doesn't handle outliers well.

Summary

Hierarchical Structure: Hierarchical clustering organizes data into a tree-like structure, representing relationships between data points through iterative merging or division of clusters.

Flexibility and Interpretability: This method is flexible and interpretable, providing a visual hierarchy of clusters at different levels of granularity, without the need to predefine the number of clusters.

References

[1] Dimitrios Gunopulos, “Clustering Overview and Applications,” Springer eBooks, pp. 383–387, Jan. 2009, doi: https://doi.org/10.1007/978-0-387-39940-9_602.

[2] “Dendrograms and Clustering,” docs.tibco.com. https://docs.tibco.com/pub/spotfire/6.5.0/doc/html/heat/heat_dendrograms_and_clustering.htm#:~:text=A%20dendrogram%20is%20a%20tree (accessed Nov. 10, 2023).

[3] Chaitanya Reddy, “Understanding the concept of Hierarchical clustering Technique,” Medium, Dec. 10, 2018. https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec

[4] Author: Fatih Karabiber Ph.D. in Computer Engineering, Fatih Karabiber Ph.D. in Computer Engineering, E. R. Psychometrician, and E. B. F. of LearnDataSci, “Hierarchical clustering,” Learn Data Science - Tutorials, Books, Courses, and More, https://www.learndatasci.com/glossary/hierarchical-clustering/

[5] “Hierarchical clustering: Agglomerative + divisive clustering,” Built In, https://builtin.com/machine-learning/agglomerative-clustering.

[6] van der W.M.P. Aalst, “Workflow model analysis,” Eindhoven University of Technology research portal, https://research.tue.nl/en/publications/workflow-model-analysis (accessed Nov. 10, 2023).