Unsupervised Machine Learning

Presented By
Rateesh Babu

Pragathi
Tech & Socio-Cultural Group

Livares Technologies Pvt Ltd

What is Machine Learning?

ML is a subfield of AI that involves the creation of Algorithms and statistical models, which enables computers to perform a specific task or to make a prediction by learning from data.

- ➔ Subfield of AI.
- ➔ Involves the creation of Algorithms and Models.
- ➔ Enables computer to perform specific tasks/predictions

Key feature of ML is it's ability to automatically learn and improve from the experience.

Why we need ML?

- Predictive Analytics: *Forecasting future based on historical data.*
- Natural Language Processing: *Understand human languages*
- Computer vision: *Recognize objects in Images or Videos*
- Recommender Systems: *Build personalized recommendations*
- Fraud detection: *Detect fraudulent activities or anomalies*

Types of ML

1. **Supervised learning**: *Model is trained on a labeled dataset.*
2. **Unsupervised learning**: *Model is trained on a unlabeled dataset.*
3. **Reinforcement learning**: *Learns to make decisions based on feedbacks*

Unsupervised ML- What?

A technique in which models are not supervised using training dataset
Models finds the patterns itself and insights from given dataset

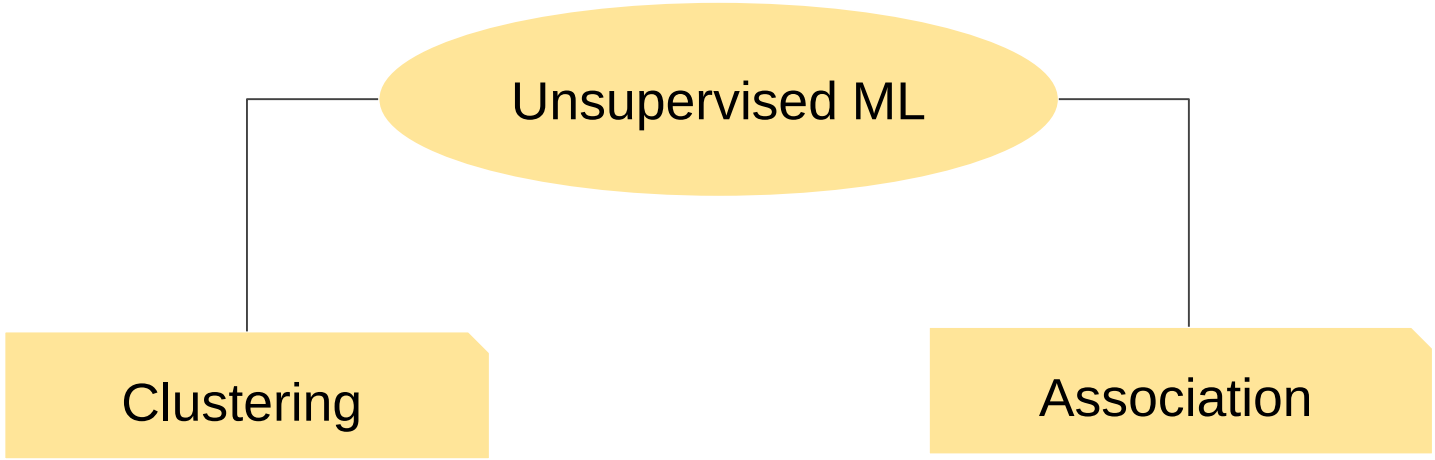
Real-LIFE Example:

Humans

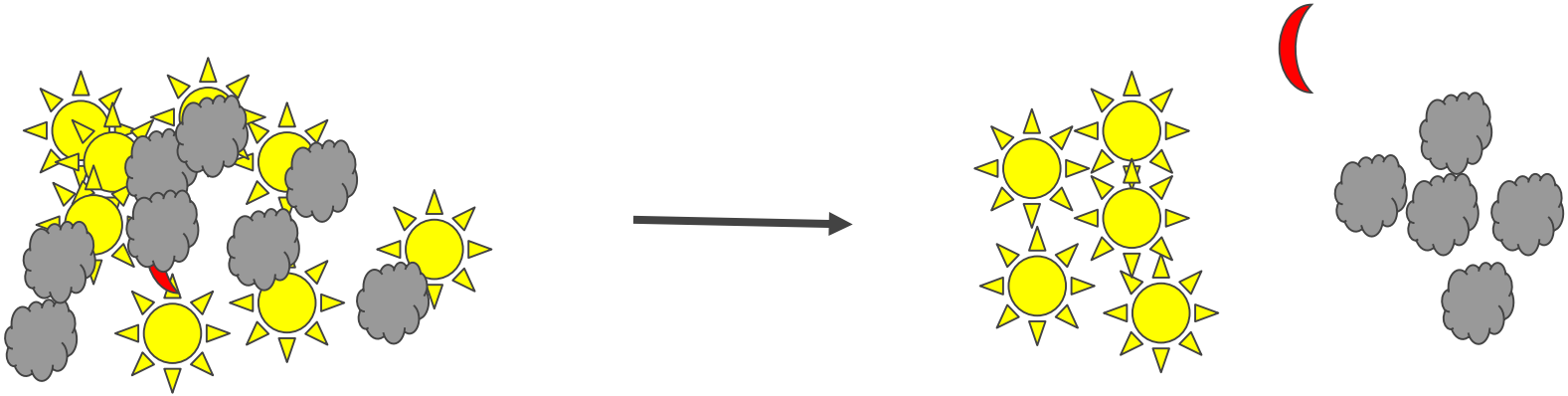
Unsupervised ML - Why?

- When 0 manual labour is desired in training.
- When it is difficult to find patterns.
- When there is a need to find the hidden pattern

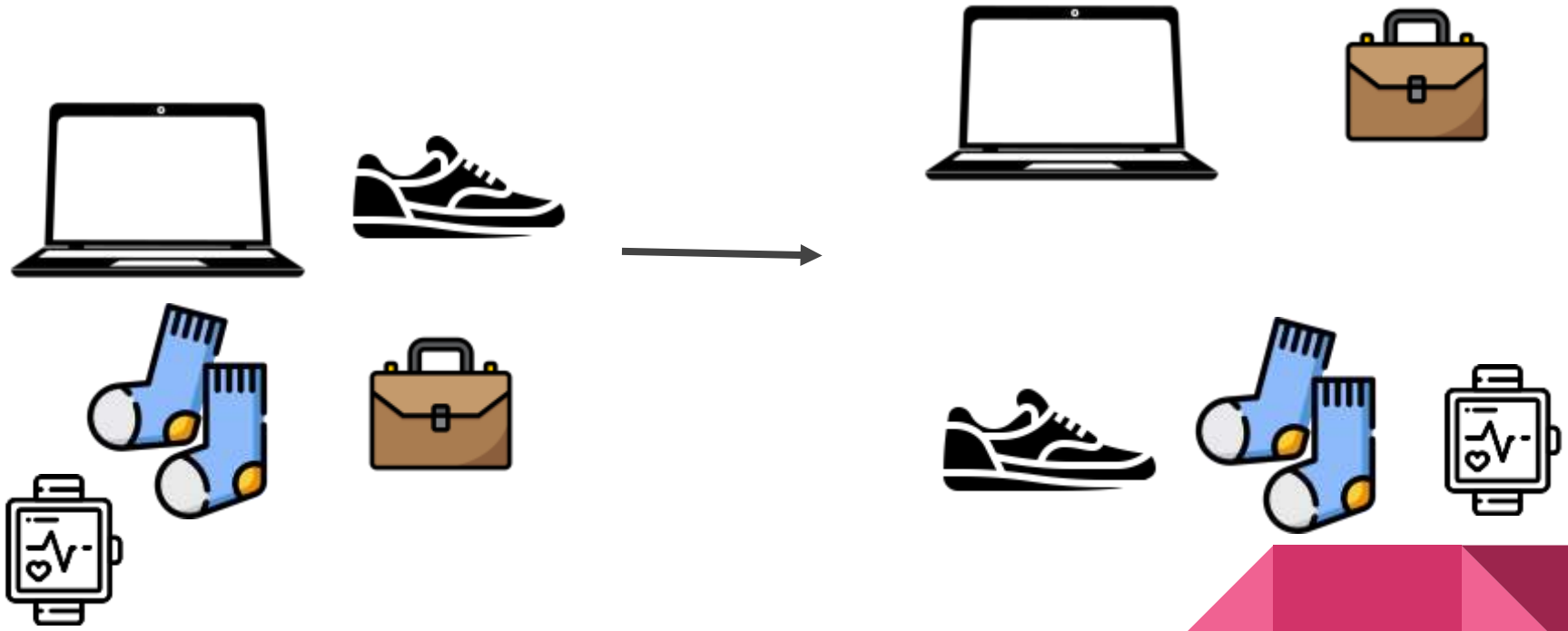
Unsupervised ML - How?



Clustering



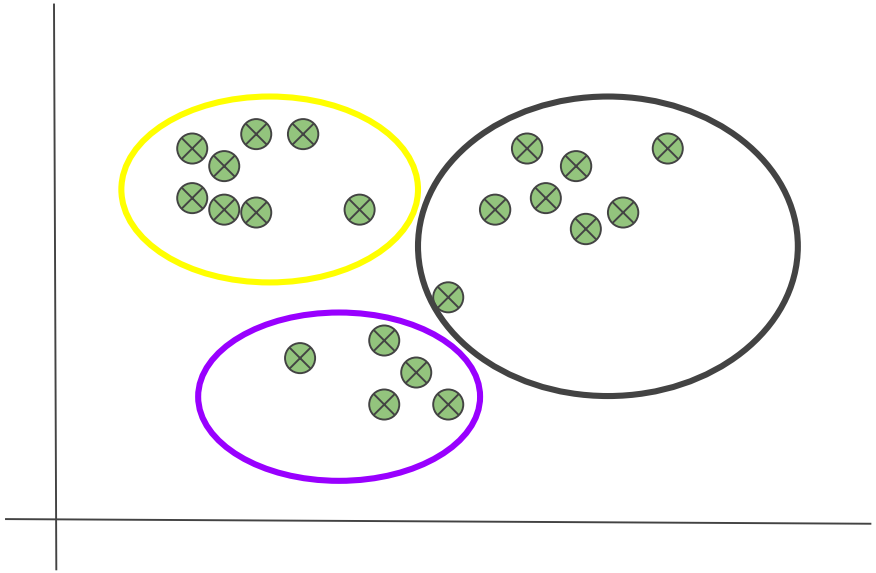
Association



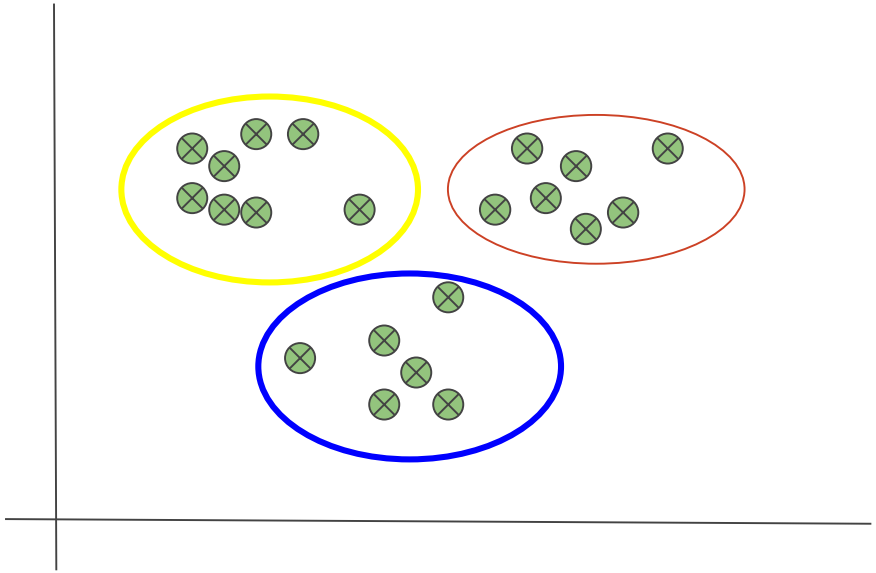
Popular Algorithms

- ★ K-means Clustering: *K clusters and centroids*
- ★ Hierarchical Clustering: *Dendrogram shaped*
- ★ Apriori Association Algorithm
- ★ Eclat Algorithm
- ★ F-P Growth Algorithm
- ★ Anomaly Detection Algorithm

K-Means Algorithm(K=3)



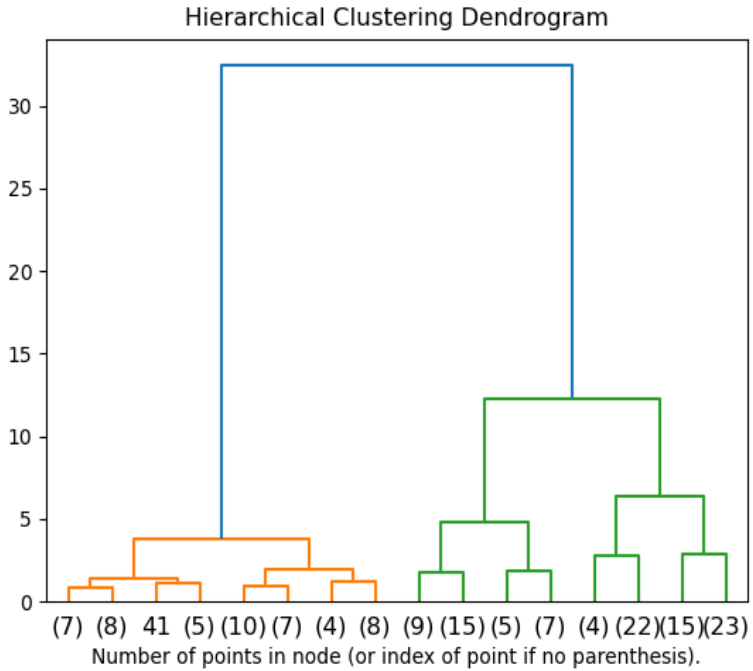
K-Means Algorithm(K=3)



Pros and cons of K-mean

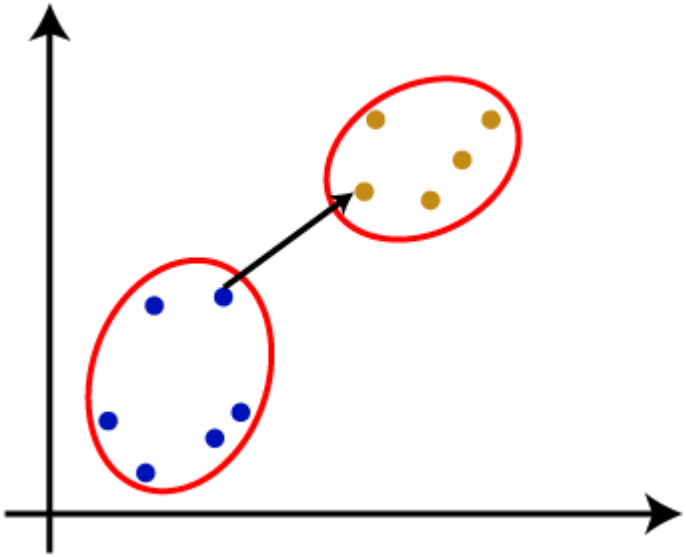
Pros	Cons
Various Applications like Customer segmentation, image segmentation, etc.	Sensitive to initial choice of centroids
Scalable (Suitable for large datasets)	Prior knowledge of number of clusters
Fast	Limited to linearly separable data
Simple and easy to understand	Prone to converge into local optima

Hierarchical Clustering



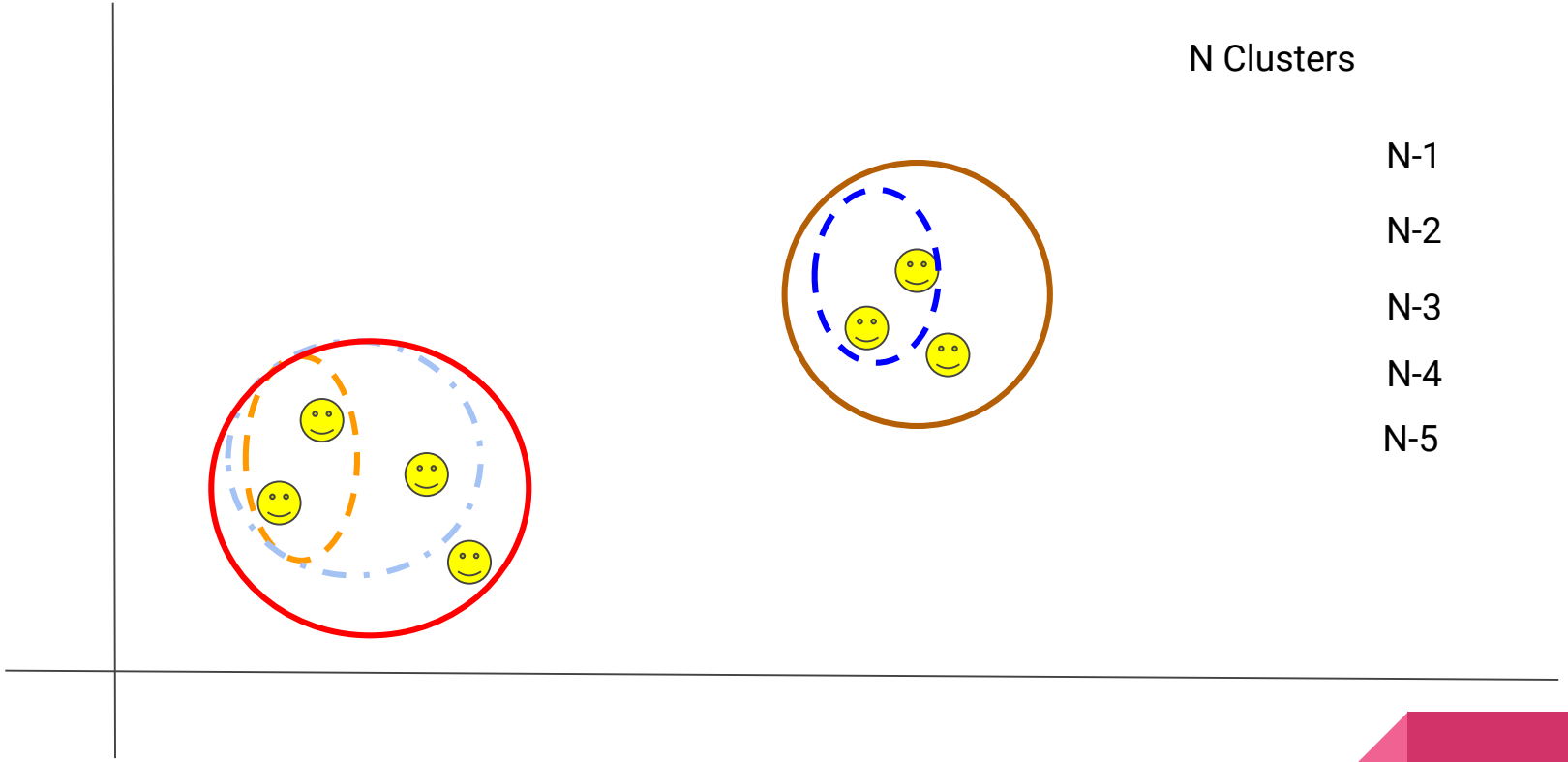
- Hierarchy of clusters in Dendrogram shaped
- **Agglomerative Approach**: ***Bottom-up***
  i.e. each data set is a cluster
- **Divisive Approach**: ***Top-down*** approach
  i.e. whole data set is a single cluster.

Hierarchical Clustering - Linkage methods



Single Linkage

- Single(Shortest)
- Complete
- Average
- Centroid



N Clusters

N-1

N-2

N-3

N-4

N-5

Pros and cons of Hierarchical Clustering

Pros	Cons
Easy to understand	Computationally expensive(time complexity is O(n^3))
Suitable for small datasets	Different methods gives different results
Detection of overlapping clusters	Greedy Algorithm
Handles different types of distance measure and linkage	Not suitable for large dataset and when clusters are not well separated

Apriori Association Algorithm

- Uses Frequent item sets to generate association rules.
- Designed to work on Databases containing transactions.
- Determines the strength(Strong or weak) connection between two objects.
- Iterative in nature.
- Mainly used for market-basket analysis and drug-reaction analysis on patients.

Apriori Association Algorithm

Terminologies:

- ➔ Antecedent ***A***
- ➔ Consequent ***B***
- ➔ Confidence: *measure of how often B appear in a transaction that also contains the A*

Confidence (A -> B) = Support (A ∪ B) / Support (A)

Apriori Association Algorithm

Confidence (A -> B) = Support (A ∪ B) / Support (A)

Support(X) = number of times X appeared in any transaction / total number of transactions.

∪ = Union i.e., Transactions that have A and B together

Apriori Association Algorithm

Sales data of a shop

1. 1000 overall Transactions.
2. 200 transactions included the purchase of milk
3. 150 transactions included the purchase of bread along with milk

Confidence(Milk->Bread) = Support (Milk U Bread) / Support (Milk)

Confidence = (150/1000) / (200/1000)

Confidence = 0.15 / 0.2

Confidence = 0.75

Frequent Itemset

TID	ITEMSETS
T1	A, B
T2	B, D
T3	B, C
T4	A, B, D
T5	A, C
T6	B, C
T7	A, C
T8	A, B, C, E
T9	A, B, C

Given: Minimum Support= 2, Minimum Confidence= 50%

Frequent Itemset- Candidate 1 and Itemset 1

Itemset	Support_Count
A	6
B	7
C	5
D	2
E	1

Itemset	Support_Count
A	6
B	7
C	5
D	2

Frequent Itemset- Candidate 2 and Itemset 2

Itemset	Support_Count
{A, B}	4
{A,C}	4
{A, D}	1
{B, C}	4
{B, D}	2
{C, D}	0

Itemset	Support_Count
{A, B}	4
{A, C}	4
{B, C}	4
{B, D}	2

A, B, C, D

Frequent Itemset- Candidate 3 and Itemset 3

Itemset	Support_Count
{A, B, C}	2
{B, C, D}	1
{A, C, D}	0
{A, B, D}	0

Itemset	Support count
{A, B, C}	2

Confidence

Confidence((A∪B)->C) = Support ((A∪B)∪C)/Support (A∪B)

Pros and cons of Apriori Algorithm

Pros	Cons
Scalable	Computationally expensive for large dataset.
Simple	Only suitable for binary data
Widely used	Need high memory to store all the itemsets generated in between

Unsupervised - Pros and cons

Pros	Cons
No need for labeled data	Challenging to know the performance of algorithm
Discovers hidden patterns	Computationally expensive
Useful in preprocessing of data	Difficult to interpret results

What’s next?

KNN Algorithm

OUR CONTACT DETAILS

Livares Technologies Pvt Ltd
5th Floor, Yamuna Building
Technopark Phase III Campus
Trivandrum, Kerala, India-695581

contact@livares.com
+91-471-2710003 | +91-471-2710004
www.facebook.com/livaresofficial
@livaresofficial

Our helpline is always open to receive any inquiry or feedback.**Please feel free to contact us**

THANK YOU

Pragathi
Tech&Socio-Cultural Group