K Means Clustering

AGENDA

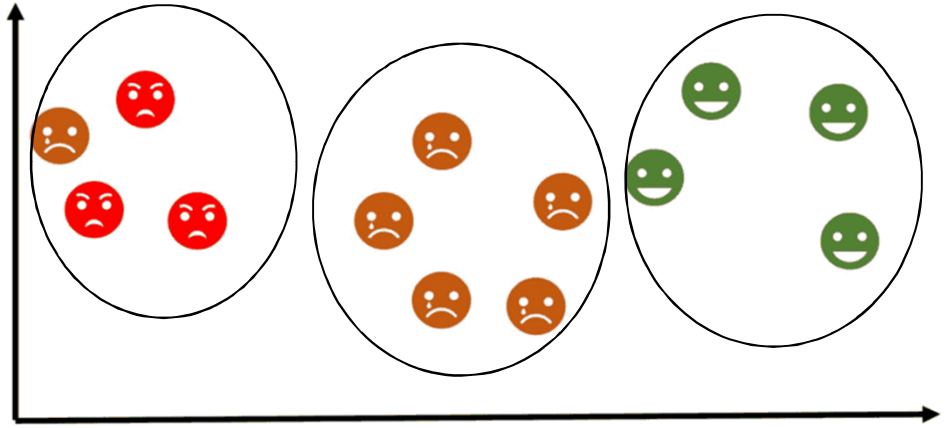
- **What is Clustering?**
- Unsupervised Learning
- Why Clustering?
- Types of Clustering
  - Partitioning Clustering
- K Means Clustering
- Challenges in K Means Clustering
- Elbow Method
- Euclidean Distance
- Illustration of K Means algorithm
- Applications of K Means

References

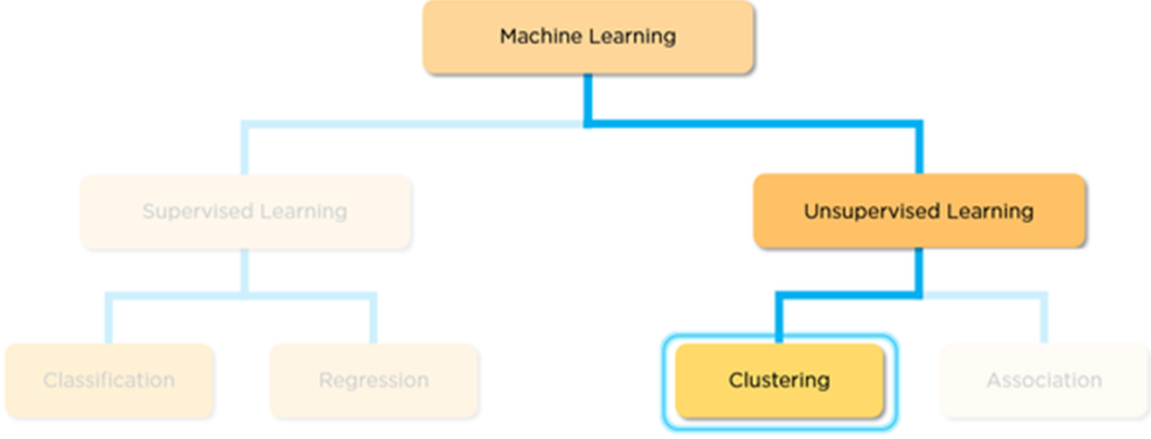
- Hierarchical Clustering
  - Agglomerative Clustering
  - Divisive Clustering
- Applications
- Density Based Clustering
- Distance metrics
  - Manhattan
  - Minkowski
  - Mahalanobis

What is Clustering?

- Clustering is the process of **grouping** similar entities together.
- The goal of this machine learning technique is to **find similarities** in the data point and group similar data points together.



- Clustering, falling under the category of **unsupervised machine learning**, is a one of the problems that machine learning algorithms solve.
- An unsupervised learning method is a method in which we **draw references from datasets** consisting of input data without labeled responses.



MARVEL®

Cluster 1

DC

Cluster 2

AGENDA

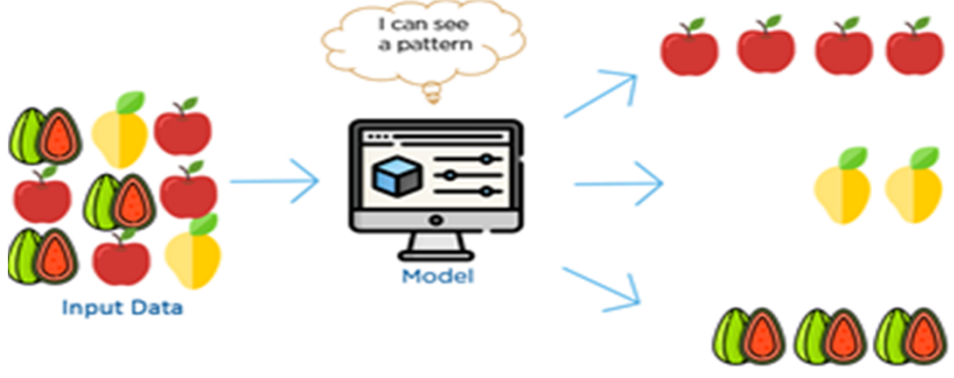
- What is Clustering?
- **Unsupervised Learning**
- Why Clustering?
- Types of Clustering
  - Partitioning Clustering
- K Means Clustering
- Challenges in K Means Clustering
- Elbow Method
- Euclidean Distance
- Illustration of K Means algorithm
- Applications of K Means

References

- Hierarchical Clustering
  - Agglomerative Clustering
  - Divisive Clustering
- Applications
- Density Based Clustering
- Distance metrics
  - Manhattan
  - Minkowski
  - Mahalanobis

What is Unsupervised Learning?

- It is a class of Machine Learning techniques to find **the patterns in data**.
- Input variables(X) are given with no **corresponding output variables**



AGENDA

- What is Clustering?
- Unsupervised Learning
- **Why Clustering?**
- Types of Clustering
  - Partitioning Clustering
- K Means Clustering
- Challenges in K Means Clustering
- Elbow Method
- Euclidean Distance
- Illustration of K Means algorithm
- Applications of K Means

References

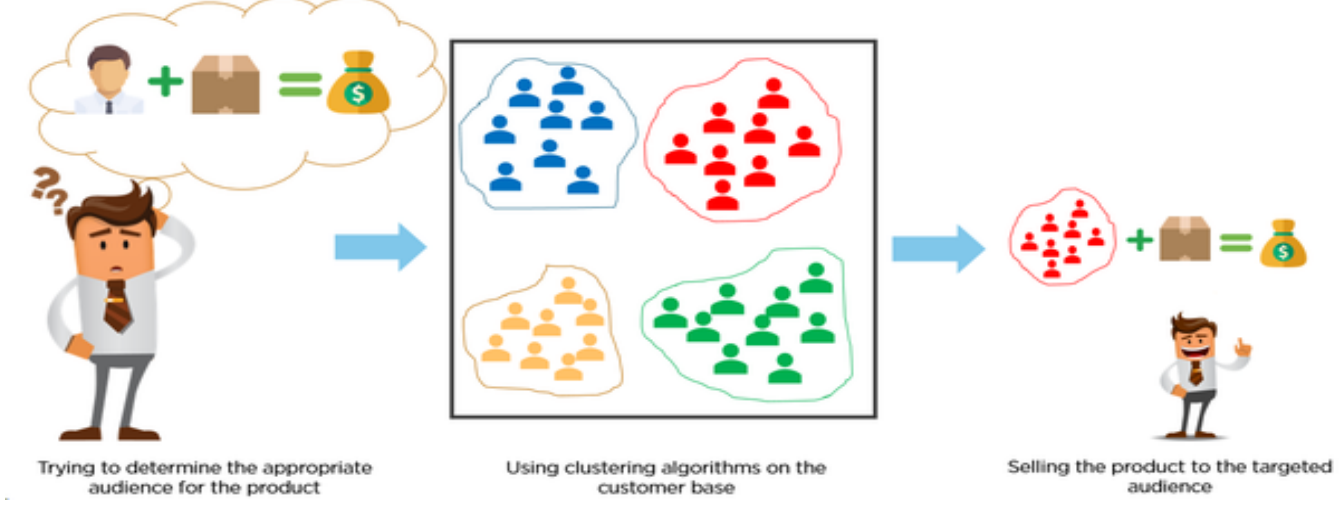
- Hierarchical Clustering
  - Agglomerative Clustering
  - Divisive Clustering
- Applications
- Density Based Clustering
- Distance metrics
  - Manhattan
  - Minkowski
  - Mahalanobis

Why Clustering?

- Imagine you're a marketing manager, and you have **a new product to sell.**
- You're sure the product would **bring a huge profit**, as long as it is sold to the **right people.**
- So, how can you tell who's best suited for the product from your company's large customer base?

Why Clustering?

- That's where clustering comes to your aid.
- Now that the data from your customer base is divided into clusters.
- You can take an informed decision on who you think is best suited for this product.



AGENDA

- What is Clustering?
- Unsupervised Learning
- Why Clustering?
- **Types of Clustering**
  - **Partitioning Clustering**
- K Means Clustering
- Challenges in K Means Clustering
- Elbow Method
- Euclidean Distance
- Illustration of K Means algorithm
- Applications of K Means

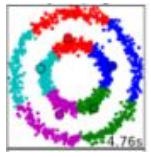
References

- Hierarchical Clustering
  - Agglomerative Clustering
  - Divisive Clustering
- Applications
- Density Based Clustering
- Distance metrics
  - Manhattan
  - Minkowski
  - Mahalanobis

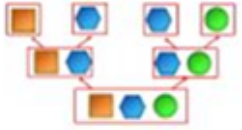
Types of clustering

- Partitioning Clustering
  - K means

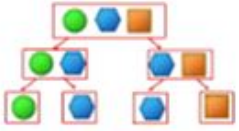
- Density Based Clustering
  - DBSCAN



- Hierarchical Clustering
  - Agglomerative



  - Divisive

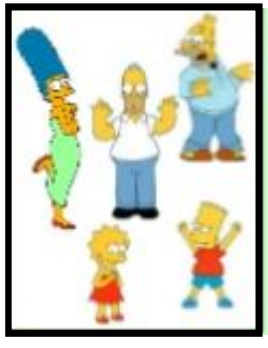


When to use Partitioning clustering?

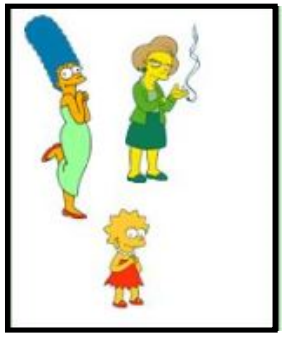
- The main objective of partition clustering algorithm is to **divide the data points** into K partitions.
- Each **partition** will reflect one cluster.
- Weakness is whenever a point is closer to the **wrong cluster**.
- The result becomes **poor or misleading** due to **overlapping** of the data points.

Let’s consider an example -

How do we partition them based on school employees, family , females and males?



Family



Females

School Employees

Males

AGENDA

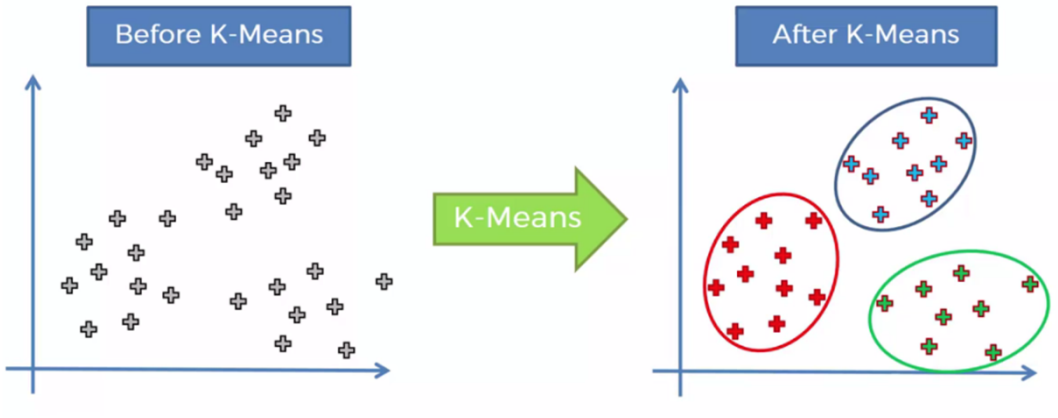
- What is Clustering?
- Unsupervised Learning
- Why Clustering?
- Types of Clustering
  - Partitioning Clustering
- 👉 **K Means Clustering**
- Challenges in K Means Clustering
- Elbow Method
- Euclidean Distance
- Illustration of K Means algorithm
- Applications of K Means

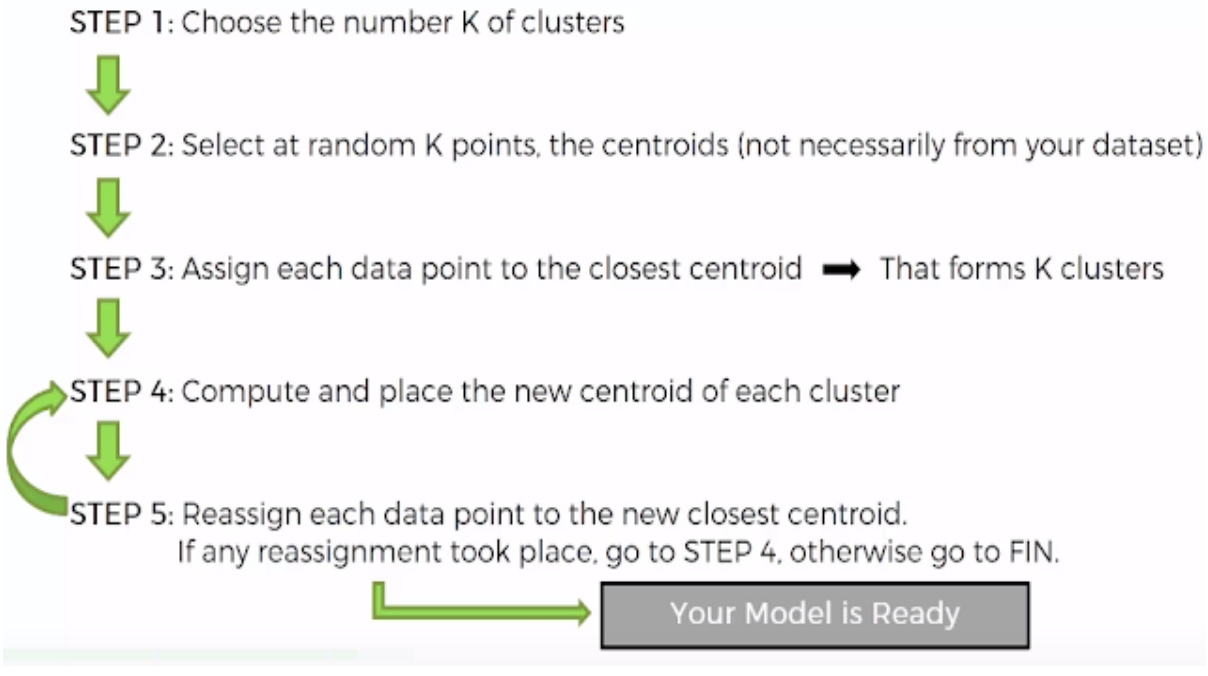
References

- Hierarchical Clustering
  - Agglomerative Clustering
  - Divisive Clustering
- Applications
- Density Based Clustering
- Distance metrics
  - Manhattan
  - Minkowski
  - Mahalanobis

K Means Clustering

- It is the most popularly used **unsupervised learning algorithm** that solves clustering problem.
- It aims to **partition n observations into k clusters** where each observation belongs to the cluster with the **nearest mean** serving as a prototype of the cluster .




STEP 1: Choose the number K of clusters
STEP 2: Select at random K points, the centroids (not necessarily from your dataset)
STEP 3: Assign each data point to the closest centroid → That forms K clusters
STEP 4: Compute and place the new centroid of each cluster
STEP 5: Reassign each data point to the new closest centroid.
If any reassignment took place, go to STEP 4, otherwise go to FIN.
Your Model is Ready

STEP 1

- Choose the number K of clusters : k = 2 (We will see later how to choose the right value of '**k'**)

STEP 2

- Select random K centroids (not necessarily from your dataset)