Deepesh

# K Means Clustering

# TIME TABLE

ML 2

| | | | |
|---|---|---|---|
| 1 | Sat 27 July 19 | PCA | |
| | Sun 28 July 19 | PCA + practical | |
| 2 | Sat 03 Aug 19 | KNN | KNN |
| | Sun 04 Aug 19 | KNN | K means |
| 3 | Sat 10 Aug 19 | K means | K means |
| | Sun 11 Aug 2019 | | |
| 4 | Sat 17 Aug 2019 | | |
| | Sun 18 Aug 2019 | | |
| 5 | Sat 24 Aug 2019 | | |
| | Sun 25 Aug 2019 | | |
| 6 | Sat 31 Aug 2019 | | |
| | Sun 01 Sep 2019 | | |

2 → 1. Principal Component Analysis (PCA)

1½ 2. K Nearest Neighbors (KNN)

3. Naive Bayes

4. Support Vector machines (SVM)

½ → 5. K means (clustering)

6. Time series Analysis

≥20% practical

INSAID

# 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

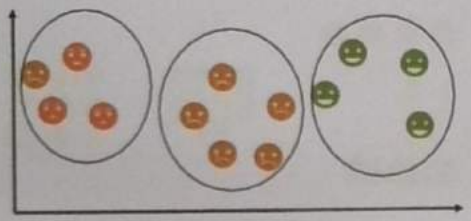
- Hierarchical Clustering
  - Agglomerative Clustering
  - Divisive Clustering
- Applications
- Density Based Clustering
- Distance metrics
  - Manhattan
  - Minkowski
  - Mahalanobis

3

INSAID

## What is Clustering?

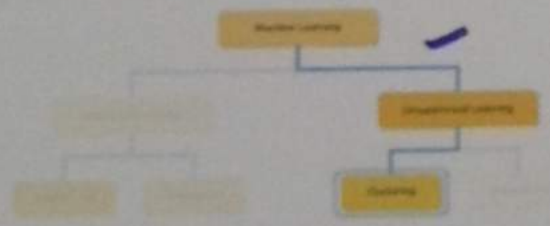
- Clustering is the process of grouping similar entities together.
- The goal of this machine learning technique is to find similarities in the data point and group similar data points together.



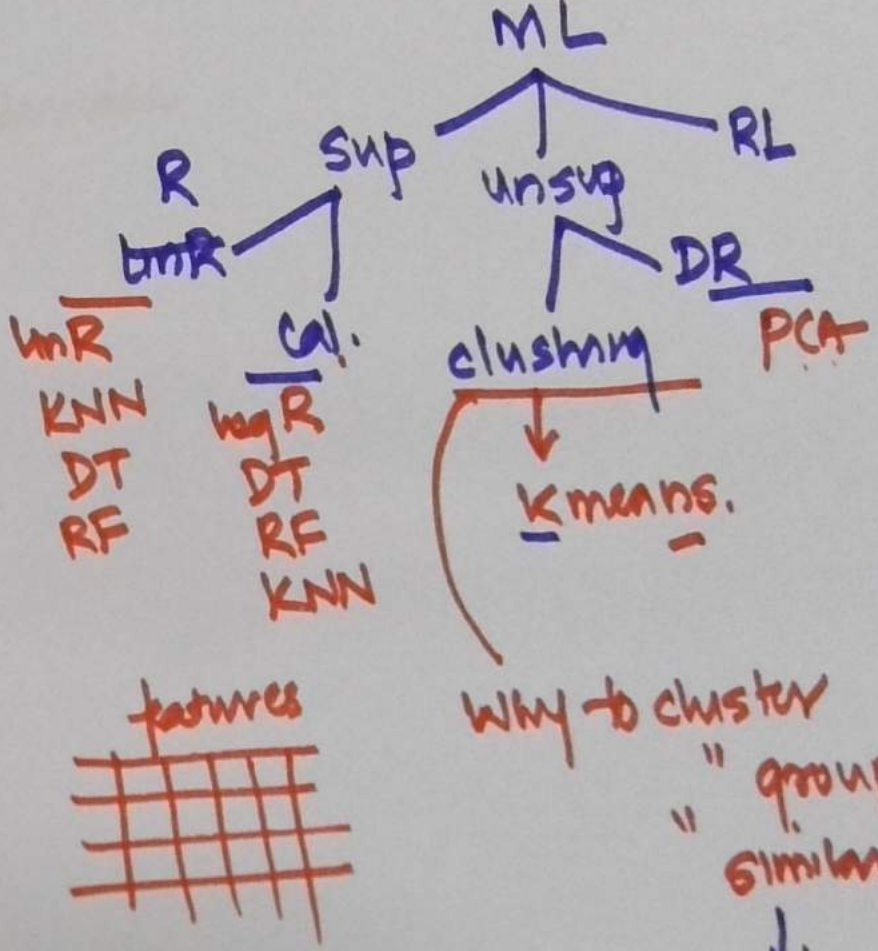


INSAID

- 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.







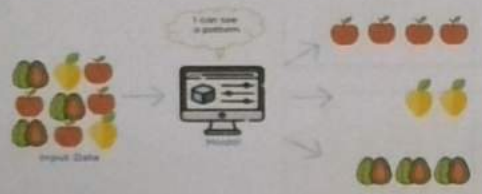
4

INSAID

## What is Unsupervised Learning?

No Target Variable

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



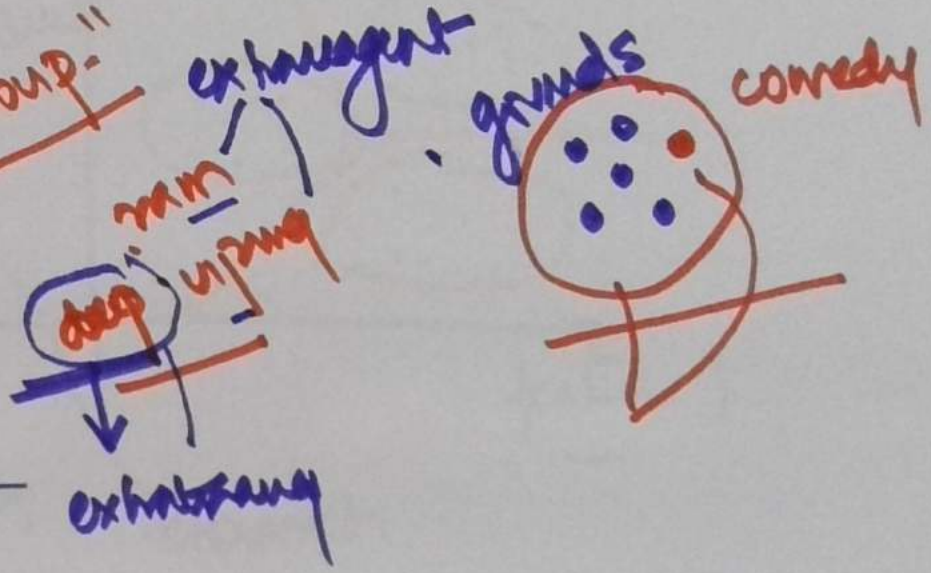
5

## 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?

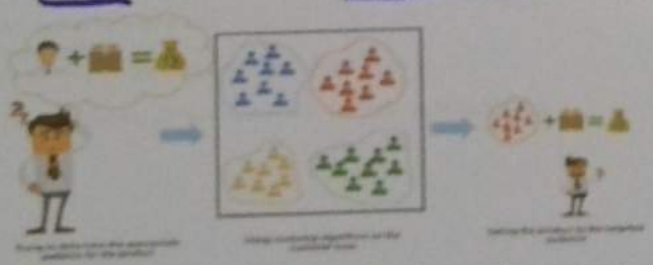




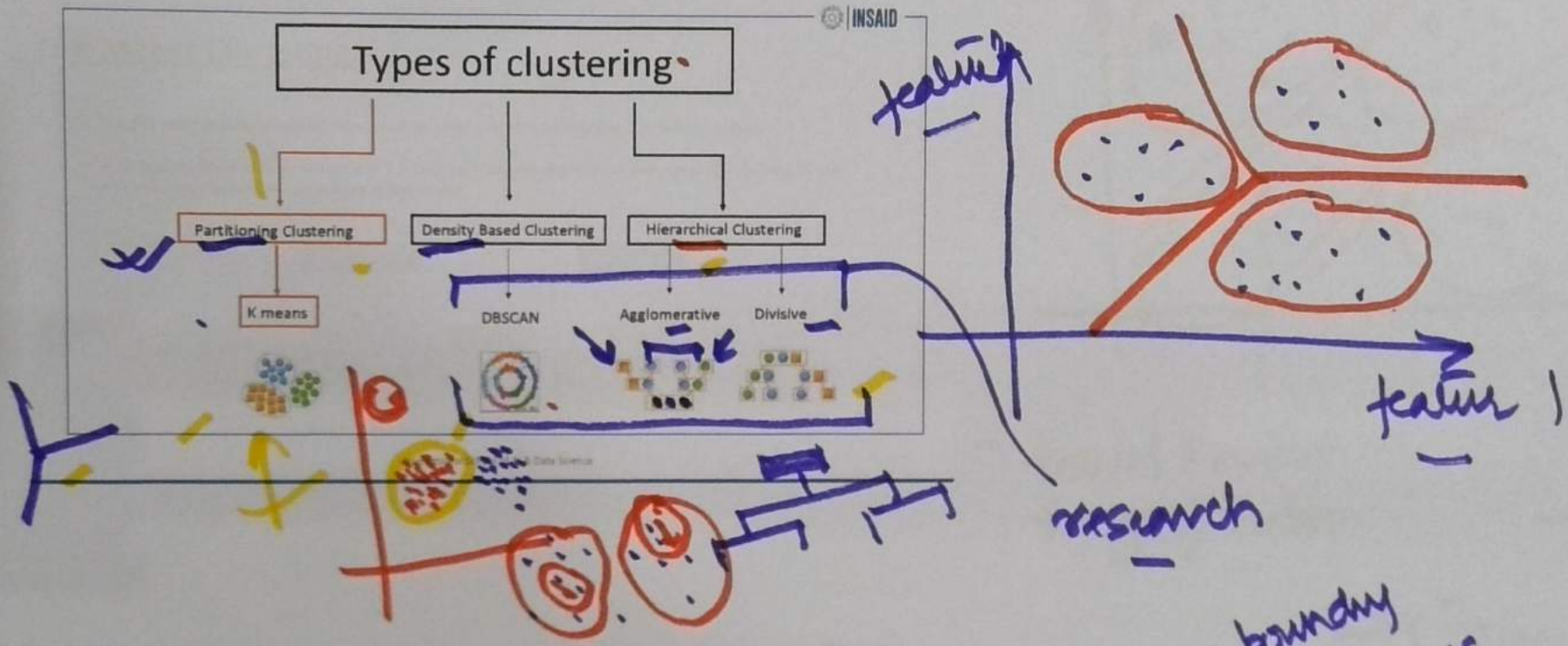



## Why Clustering?

- 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.



6




PARTITIONING

feature 2

feature 1

research

## 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.

decision boundry or partitions

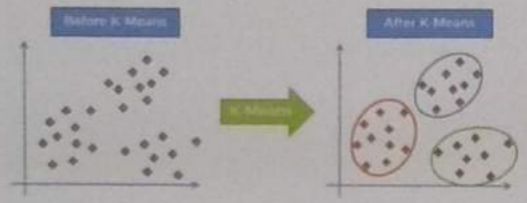
clusters

7

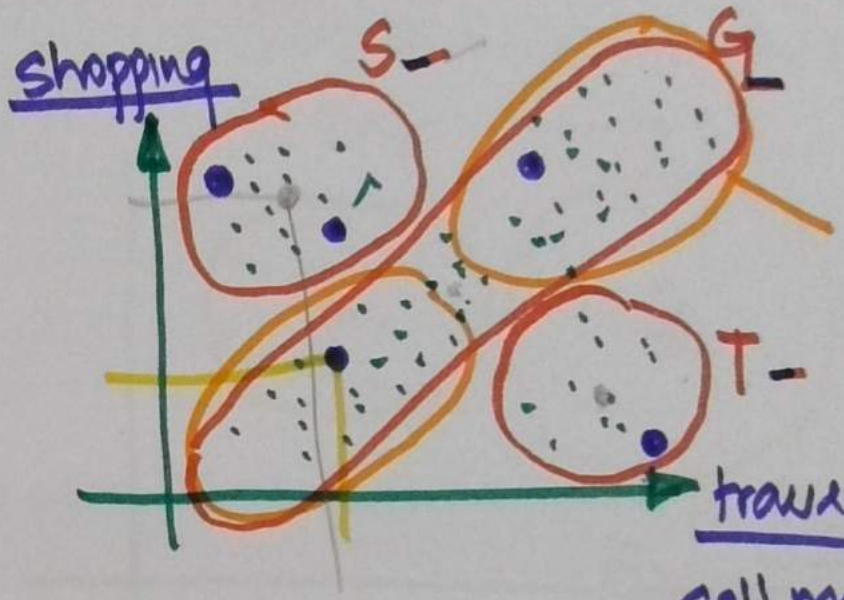
INSAID

## 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 .






- travel Factor
  shopping Factor

sell more CCs

- ↓ more profits

INSAID

- K

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.



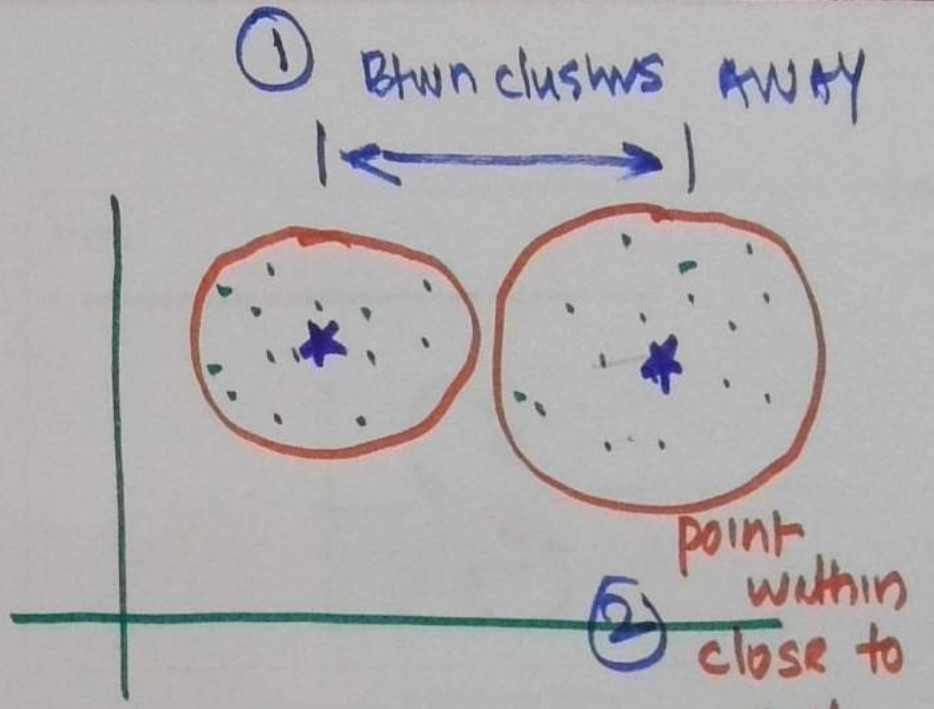
- TRAVELER'S CARD
- SHOPPER'S CARD

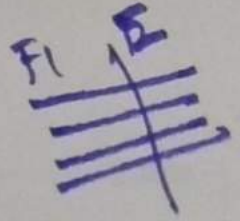
How many types of CARDS?

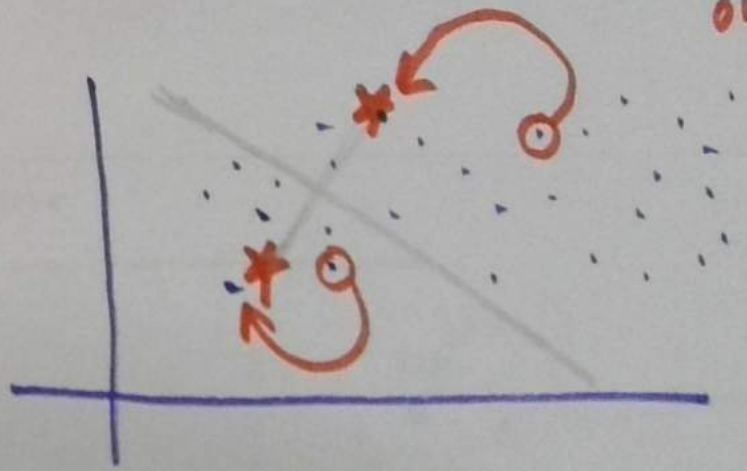
≡ K = no. of clusters

8

decision boundry



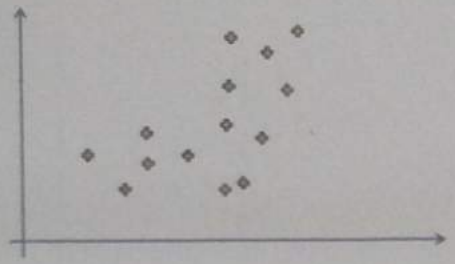



9

INSAID

## STEP 1

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

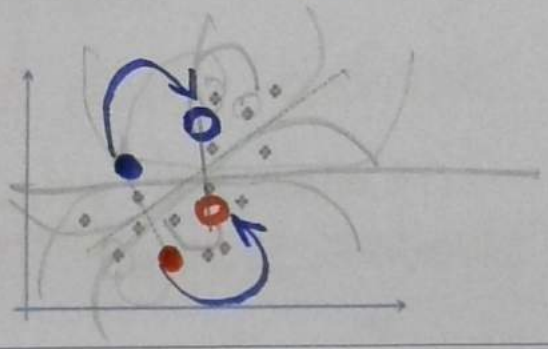




INSAID

## STEP 3

- Assign each data point to the closest centroid ➡ That forms K clusters.

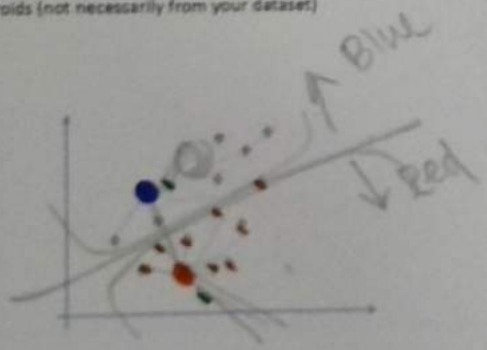




INSAID

## STEP 2

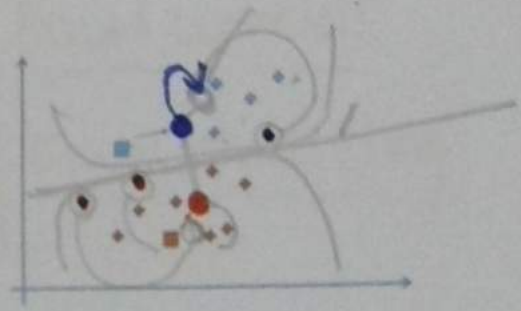
- Select random K centroids (not necessarily from your dataset)




INSAID

## STEP 4

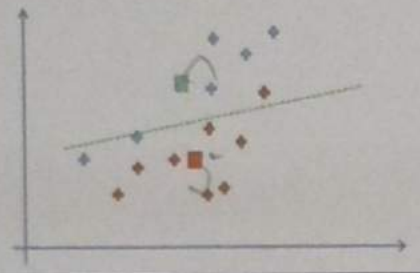
- Compute and place the new centroid of each cluster.



INSAID

STEP 5

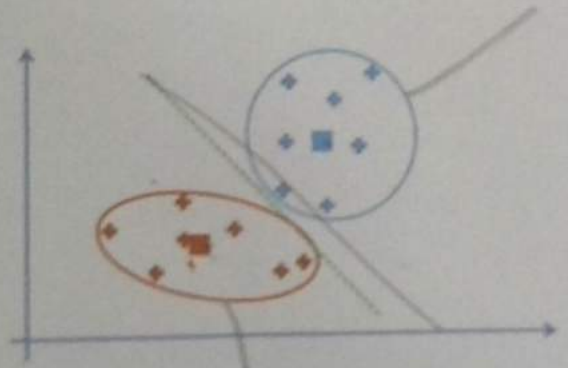
- Reassign each data point to the new closest centroid.
- If any reassignment took place, go to STEP 4
- Otherwise FINISH





INSAID

Final Model





K MEANS

1 — K

2 — plotting data.

3 — random cluster centroids

4 — a cluster the data on basis of centroids

b recalculate centroids

Iterate until centroids do not move

Final centroids

Final clustering