Unsupervised learning Algorithm

— K-means clustering and Association rule mining

Unsupervised learning is a type of Machine Learning algorithm used to draw inferences from datasets consisting of input data without labelled responses.

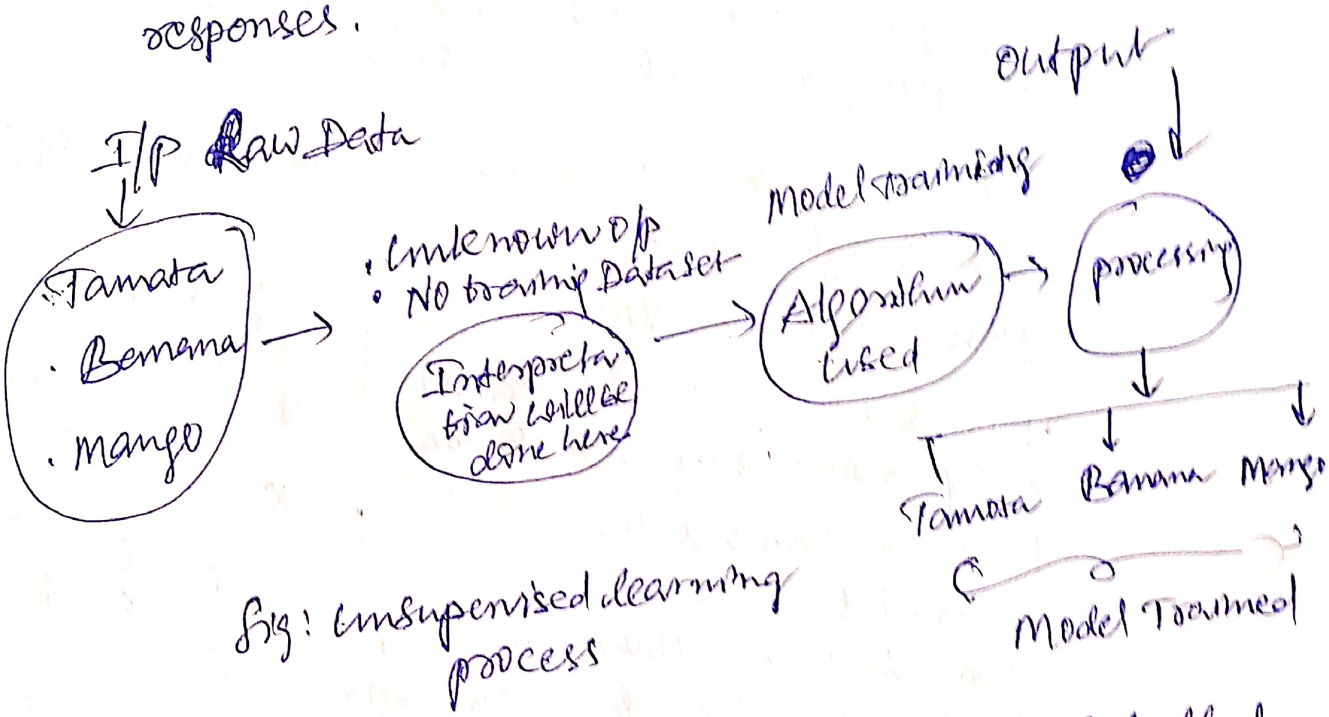


Fig: Unsupervised learning process

→ When the given data is unstructured and unlabelled it is difficult to classify the data into different categories then unsupervised learning helps to solve this type of problems.

→ This learning is used to cluster the input data and classes on the basis of their statistical properties.

→ Ex: We can cluster different bikes based upon the speed limit, their acceleration or the average that they are giving

→ Unsupervised learning is a type of ML algorithm used to draw inferences from the data sets consisting of I/P data without labelled responses

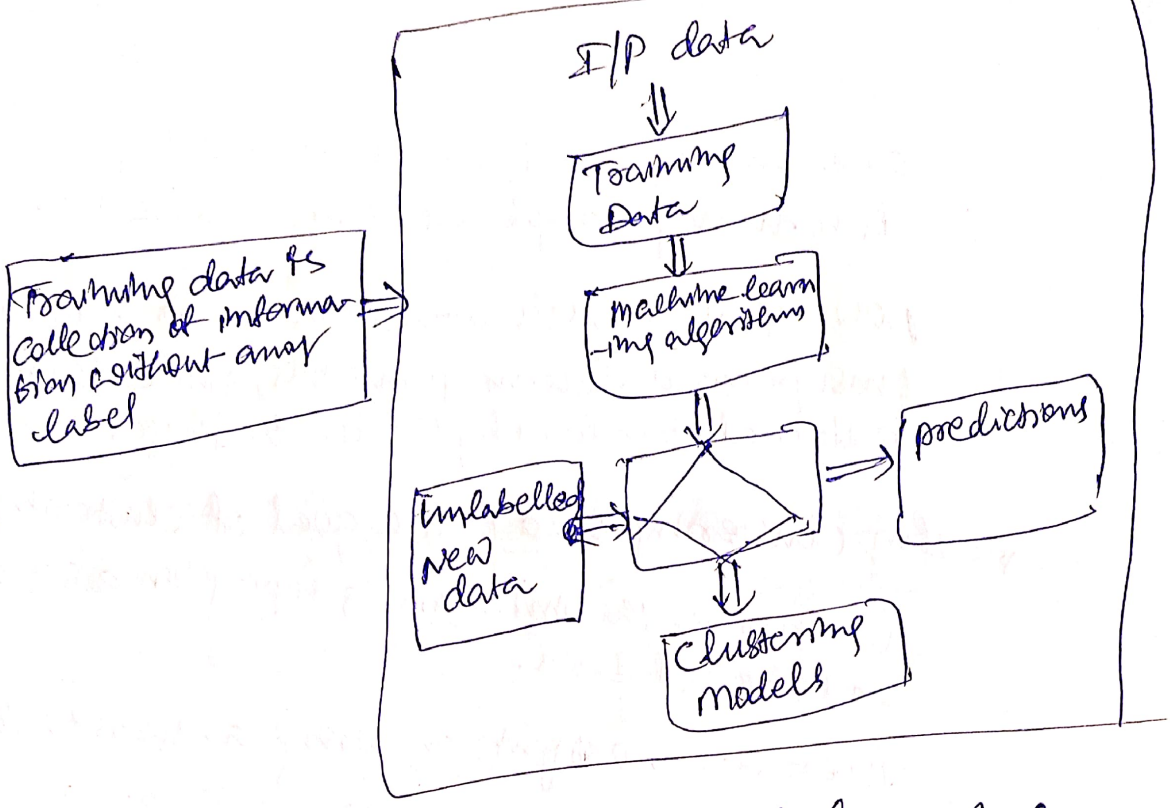


fig: Work flow of unsupervised learning

The trained data is collection of information without any label.
We have the machine learning algorithm and then we have the clustering models
the data is distributed into different classes and algorithm if any new data is provided it will make a prediction and find out to which cluster that particular data or the dataset belongs to or particular data points belongs to.
one of the most important algorithm in unsupervised ~~algorithm~~ learning is Clustering.

"Clustering is the process of dividing the datasets into groups, consisting of similar data-points"
It means grouping of objects based on the information found in the data, describing the objects or their relationship.
Clustering models focus on groups of similar records and labelling records according to the group to which they belong, this is done without the benefit of prior knowledge about the groups and their characteristics.

and we may not even know exactly how many groups are there to look for.

Now these models are often referred to as unsupervised learning models, since there is no external standard by which to judge.

→ Why clustering used? The goal of clustering is to determine the intrinsic grouping in a set of unlabelled Data.

There are no right or wrong answers to these models.

→ Where clustering is used?

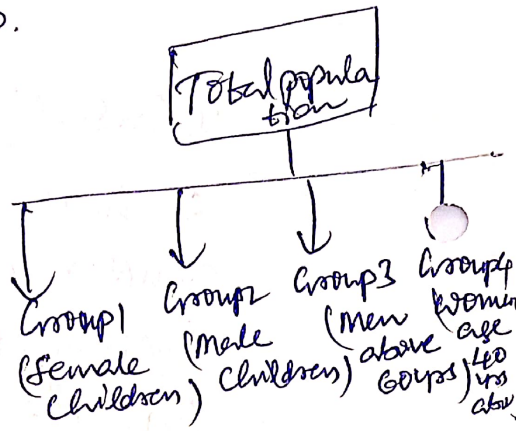
Sometimes the passion is the goal or the purpose of clustering algorithms to make sense of and extract value from the last set of structured and unstructured data.

So clustering is used in the industry and if you have a look at the videos these use cases of clustering in the industry

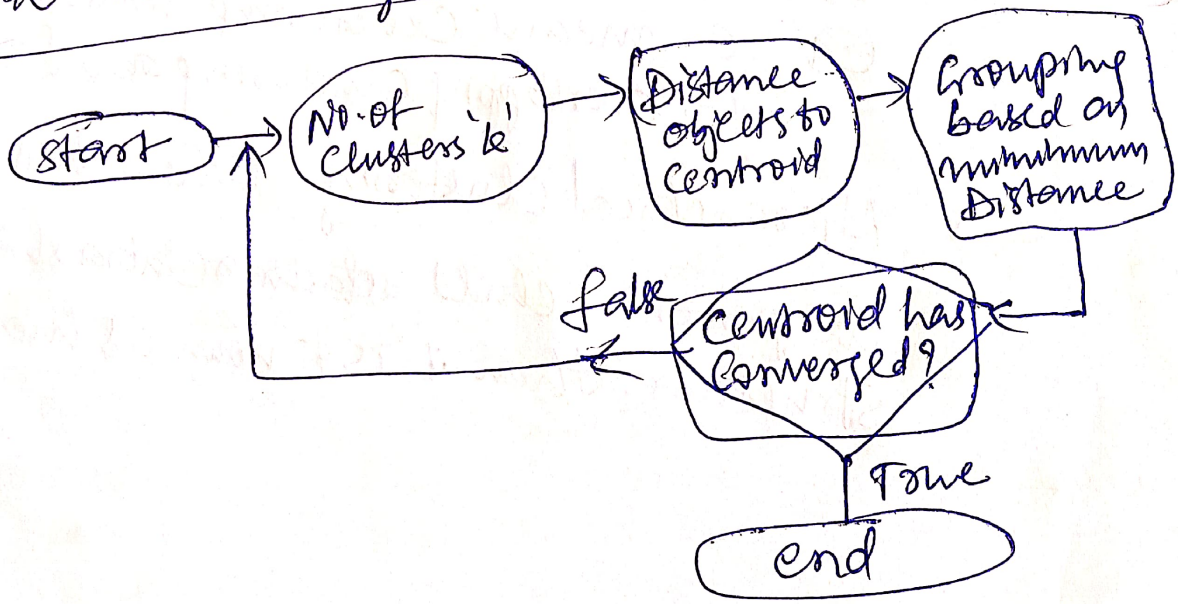
K-means Clustering:

The process by which objects are classified into a predefined number of groups so that they are as much as dissimilar as possible from group to another group, but as much similar as possible within each group.

(K-means clustering is an algorithm whose main goal is to group similar elements of data points into a cluster and it is a process of by which objects are classified into a predefined number of groups so that they are as much as dissimilar as possible from one group to another group but as much as similar as possible within each group



→ K-means Algorithm working:



K-Means Algorithm working: It starts with and identifying the no. of clusters which is 'k' then we can find the centroid, find distance, of objects from centroid minimum distance.

(1) we need to decide the no. of clusters to be made (guessing)

(2) Then we provide centroids of all the clusters (guessing)

(3) The algorithm calculates Euclidean distances of the points from each centroid and assigns the point to the closest closer.

(4) Next centroids are calculated again, when we have our new closer

(5) The distance of the points from the centre of clusters are calculated.

(6) And then again the new centroid for the cluster is calculated.

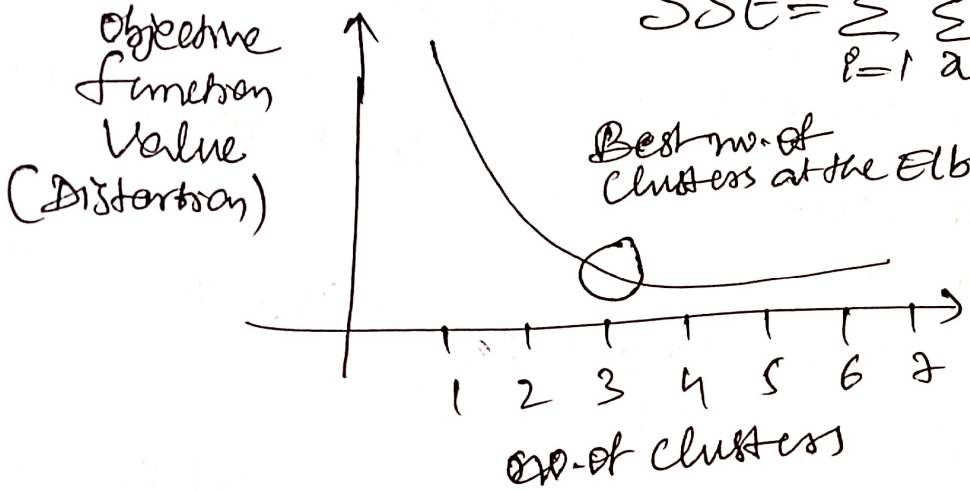
(7) These steps are repeated until we have a repetition in centroids or new centroids are very close to the previous ones.

How to Decide the no. of Clusters:

The Elbow method:

First of all, compute the sum of squared error (SSE) for some values of k (for ex: 2, 4, 6, 8, etc).
The SSE is defined as the sum of the squared distance between each member of the cluster and its centroid.
Mathematically:

$$SSE = \sum_{i=1}^{k} \sum_{a \in C_i} dist(a, c_i)^2$$



SSE is defined which is the sum squared is sum of the squared distance between each number of the cluster and its centroid.
If k against SSE the error decreases as 'k' gets large. now this is because the no. of cluster increases they should be smaller. the distortion is also smaller
the idea of Elbow method is to choose the 'k' at which the SSE decreases abruptly.
Eg: the best no. of cluster is at the elbow the graph changes abruptly after the number 4.
4 is the number of clusters.

While working with K-means clustering two key points (1) Where to start Choosing the first centre at random and the second centre that is far away from the first centre. Similarly choosing the 3rd centre as far away as possible from the closest of the all the other centres

Second Idea is to do as many runs of K-means each with different random starting points, so that we get an idea of where exactly and how many clusters we need to make and where exactly the centroid lies, and how the data is conversed.

Pros and Cons: K-Means clustering:

Pros: (1) Simple, understandable
(2) Items automatically assigned to clusters

Cons: (1) Must define no. of clusters - It is a heavy task
(2) All items forced into clusters
(3) Unable to handle noisy data and outliers

KNN step by step:

(4) **Response**: Generate a response from a set of data instances.

(1) **Handling the data**: Open the dataset from CSV and split into test/train datasets.

(2) **Similarity**: Calculate the distance between two data instances

once calculate the distances near look for the neighbour

(3) **Neighbours**: locate k most similar data instances which are having least distances to the new point. once you get a neighbour then

→ we will generate a response from a set of data instances. So this will decide whether the new point belongs to class A or class B.

(5) **Accuracy**: Summarising the accuracy of predictions.

(6) **Main**: Tie all together in the main function.

Note: **CSV file**: A Comma-Seperated Values file is a delimited text file that uses a comma to seperate values. Each line of the file is a data record. Each record consists of one or more fields seperated by commas. The use of the comma as a field seperator is the source of the name for this file format.

if the new item is:
Sepal Length = 5.2; } then it belongs to
Sepal width = 3.1; } which species?

KNN Classifier Solved Example - 1

Sepal Length	Sepal Width	Species	Distance	Rank
5.3	3.7	Setosa	0.608	3
5.1	3.8	Setosa	0.707	6
7.2	3.0	Virginica	2.002	13
5.4	3.4	Setosa	0.36	2
5.1	3.3	Setosa	0.22	1
5.4	3.9	Setosa	0.82	8
7.4	2.8	Virginica	2.22	15
6.1	2.8	Verscicolor	0.94	10
7.3	2.9	Virginica	2.1	14
6.0	2.7	Verscicolor	0.89	9
5.8	2.8	Virginica	0.67	5
6.3	2.3	Verscicolor	1.36	12
5.1	2.5	Verscicolor	0.60	4
6.3	2.5	Verscicolor	1.25	11
5.5	2.4	Verscicolor	0.75	7

Step 3: Find the Nearest Neighbor

If k = 1 – Setosa

If k = 2 – Setosa

If k = 5 – Setosa

Here the new data item belongs to Setosa species if K = 1, 2, 5 else