**Unsupervised Learning : Clustering**

Two most commonly used types of clustering algorithms - **K-Means Clustering** and **Hierarchical Clustering**, as well as their application in Python. Then, you will also look at what segmentation is and how it is different from clustering

Supervised learning techniques such as regression and classification use a training set to make the algorithm learn, and then apply what is learnt to new, unseen data points. unsupervised learning technique, where you try to find patterns based on similarities in the data.

**Understanding Clustering**

In the previous modules, you saw various supervised machine learning algorithms. Supervised machine learning algorithms make use of labelled data to make predictions.

For example, an email will be classified as spam or ham, or a bank’s customer will be predicted as ‘good’ or ‘bad’. You have a target variable Y which needs to be predicted.

On the other hand, in unsupervised learning, you are not interested in prediction because you do not have a target or outcome variable. The objective is to discover interesting patterns in the data, e.g. are there any subgroups or ‘clusters’ among the bank’s customers?

**PRACTICAL APPLICATIONS OF CLUSTERING**

1. **Customer Insight:** Say, a retail chain with so many stores across locations wants to manage stores at best and increase the sales and performance. Cluster analysis can help the retail chain to get desired insights on customer demographics, purchase behaviour and demand patterns across locations. This will help the retail chain for assortment planning, planning promotional activities and store benchmarking for better performance and higher returns.
2. **Marketing:** Cluster Analysis can help with In the field of marketing, Cluster Analysis can help in market segmentation and positioning, and to identify test markets for new product development.
3. **Social Media:** In the areas of social networking and social media, Cluster Analysis is used to identify similar communities within larger groups.
4. **Medical**: Cluster Analysis has also been widely used in the field of biology and medical science like human genetic clustering, sequencing into gene families, building groups of genes, and clustering of organisms at species.

Customer segmentation for targeted marketing is one of the most vital applications of the clustering algorithm. Here, as a manager of the online store, you would want to group the customers into different clusters, so that you can make a customised marketing campaign for each of the group. You do not have any label in mind, such as good customer or bad customer. You want to just look at patterns in customer data and then try and find segments. This is where clustering techniques can help you with segmenting the customers. Clustering techniques use raw data to form clusters based on common factors among various data points. This is exactly what will also be done in segmentation, where various people or products will be grouped together on the basis of similarities and differences between them.

For successful segmentation, the segments formed must be stable. This means that the same person should not fall under different segments upon segmenting the data on the same criteria. You also saw that segments should have **intra-segment homogeneity** and **inter-segment heterogeneity**. You will see in later sessions how this can be defined mathematically.

Behaviour Segmentation: Segmentation is based on the actual patterns displayed by the consumer

Attitudinal Segmentation: Segmentation is based on the beliefs or the intents of people, which may not translate into similar action

Demographic Segmentation: Segmentation is based on a person’s profile and uses information like Gender, Age, Location, Income, Family members, House Type.

Top Three take aways from Unsupervised Learning – Clustering:

1. In unsupervised learning we do not have predefined label to categorise or classify or segment but by running algorithms we can cluster data into similar groups or labels. We do not have a target variable for us to predict but to identify similar groups from the data.
2. Segments formed should be stable and there should be intrasegment homogeneity and inter segment heterogeneity for a successful clustering.
3. Two most common algorithms for clustering the data are k-means clustering and Hierarchial clustering

**First common algorithm to achieve unsupervised clustering — the K-Means algorithm**

**Euclidean Distance**

How clustering works - it groups the objects on the basis of their similarity or closeness to each other

the Euclidean Distance between the 2 points is measured as follows: If there are 2 points X and Y having n dimensions

X=(X1,X2,X3,...Xn)

Y=(Y1,Y2,Y3,....Yn)

Then the **Euclidean Distance D**is given as

D=√(X1−Y1)2+(X2−Y2)2+...(Xn−Yn)2

The idea of distance measure is quite intuitive. Essentially, the observations which are closer or more similar to each other would have a low Euclidean distance and the observations which are farther or less similar to each other would have a higher Euclidean distance. **So can you now guess how the Clustering process would work based on the Euclidean distance?**

**Centroid**

The next concept that is crucial for understanding how clustering generally works is the idea of centroids. centroids are essentially the centre points of triangles. Similarly, in the case of clustering, centroids are the**centre points of the clusters**that are being formed.

the Centroids are essentially**the cluster centres** of a group of observations that help us in **summarising the cluster's properties**. The centroid value in the case of clustering is essentially the mean of all the observations that belong to a particular cluster

**Steps of the Algorithm:**

1. **Assignment step**
2. **Optimization step**

So the cost function for the K-Means algorithm is given as:

J=∑ni=1||Xi−μk(i)||2=∑Kk=1∑i?Ck||Xi−μk||2

#### K-Means

What is the significance of "argmin" in the assignment step equation?

My Answer : The euclidean distance between a datapoint and all the cluster centers, whichever distance is min we can assign the data point to that cluster center. so argmin signifies the min value so that we can assign the datapoint to that cluster

**Both are same**

Suggesterd Answer : For a ith data point which is a 2d object and μ which is again a 2d object, we compute the distance between these two, this is given by d(xi,μk) where k is the number of clusters and then from these k different results we will choose the minimum of all.

In the assignment step, we assign every data point to K clusters. The algorithm goes through each of the data points and depending on which cluster is closer, in our case, whether the green cluster centroid or the blue cluster centroid; It assigns the data points to one of the 2 cluster centroids.

The equation for the assignment step is as follows:

Zi=argmin||Xi−μk||2

In the optimisation step, the algorithm calculates the average of all the points in a cluster and moves the centroid to that average location.

The equation for optimisation is as follows:

μk=1nk∑i:zi=kXi

The process of assignment and optimisation is repeated until there is no change in the clusters or possibly until the algorithm converges.

# K Means++ Algorithm

We looked in the previous segment that for K-Means optimisation problem, the algorithm iterates between two steps and tries to minimise the objective function given as,

Zi=argmin||Xi−μk||2

To choose the cluster centres smartly, we will learn about K-Mean++ algorithm. K-means++ is just an initialisation procedure for K-means. In K-means++ you pick the initial centroids using an algorithm that tries to initialise centroids that are far apart from each other.

To summarise, In K-Means++ algorithm,

1. We choose one data point as the cluster centre at random.
2. For each data point Xi, We compute the distance between Xi and the nearest centre that had already been chosen.
3. Now, we choose the next cluster centre using the weighted probability distribution where a point X is chosen with probability proportional to d(X)2 .
4. Repeat Steps 2 and 3 until K centres have been chosen.

The major practical considerations involved in K-Means clustering are:

* The number of clusters that you want to divide your data points into, i.e. the value of K has to be pre-determined.
* The choice of the initial cluster centres can have an impact on the final cluster formation.
* The clustering process is very sensitive to the presence of outliers in the data.
* Since the distance metric used in the clustering process is the Euclidean distance, you need to bring all your attributes on the same scale. This can be achieved through standardisation.
* The K-Means algorithm does not work with categorical data.
* The process may not converge in the given number of iterations. You should always check for convergence.

Silhouette coefficient is a measure of how similar a data point is to its own cluster (cohesion) compared to other clusters (separation).

# Cluster Tendency

Before we apply any clustering algorithm to the given data, it's important to check whether the given data has some meaningful clusters or not? which in general means the given data is not random. The process to evaluate the data to check if the data is feasible for clustering or not is known as the clustering tendency.

One of the important considerations while doing K-means clustering is:

1. Choosing initial clusters
2. Choosing k
3. Check whether data is such that clustering is possible

To check cluster tendency, we use Hopkins test. Hopkins test examines whether data points differ significantly from uniformly distributed data in the multidimensional space.

A value close to 1 tends to indicate the data is highly clustered, random data will tend to result in values around 0.5, and uniformly distributed data will tend to result in values close to 0

**Top Three take aways from K-means Clustering:**

1. we can check for cluster tendency to determine if we can form a specific number of clusters, we can perform hopkins test to see if there potential for high cluster tendency.

2. we use silhouette measure to determine how similar the data points in a same cluster compared to other cluster.

3. There are few practical considerations that we need to consider while using k-means: choice of initial clusters can have an impact on final cluster formation, all data points need to be brought to same scale for standardization, clustering is sensitive to outliers.

**Implementation of the K-Means algorithm in Python on the Online Retail case study:** will learn about

* Data preparation
* How to make the clusters
* Decide the optimal number of clusters
* How to interpret the results

**Optimal Number of Clusters:** To find the optimum number of clusters, we use two techniques - the elbow curve method and the silhouette score method.

How to make clusters using the K-Means algorithm. Let's use that knowledge to play around with clustering using K-Means

# Behavioural Segmentation Types

You have seen what RFM segmentation is. Now, you will look at other segmentation types commonly used in the industry. RPI segmentation, CDJ segmentation.

Summary :

create clusters using the K-means algorithm in Python with the analysis of the Online Store data set. We wanted to group the customers of the store into different clusters based on their purchasing habits. The different steps involved were:

* Missing values treatment
* Data transformation
* Outlier treatment
* Data standardisation
* Finding the optimal value of K
* Implementing K Means algorithm
* Analysing the clusters of customers to obtain business insights

Once we are through with the data preparation, the K-means algorithm is quite easy to implement. All it takes is running the KMeans() function. The only ambiguous point you may notice here is that you need to decide the number of required clusters beforehand and in fact run the algorithm multiple times with a different number K before you can figure out the most optimal number of clusters.

This is also what happens in the industry practices that we run the algorithm multiple times with different values of K and then pick the clusters which make the most business sense. In fact, the k-means algorithm finds a large application in the industry. For example, it can be used to find out the most optimal centre to install mobile towers by clustering the customers geographically. Similarly, it has wide application in medical science, where say the patients can be clustered together on the basis of their symptoms, and then analysed to figure out the cause of their illness.

**Hierarchical Clustering:**

learn about another algorithm to achieve unsupervised clustering. This is called **Hierarchical Clustering**. Here, instead of pre-defining the number of clusters, you first have to visually describe the similarity or dissimilarity between the different data points and then decide the appropriate number of clusters on the basis of these similarities or dissimilarities

set of N items to be clustered, the steps in hierarchical clustering are:

1. Calculate the NxN distance (similarity) matrix, which calculates the distance of each data point from the other
2. Each item is first assigned to its own cluster, i.e. N clusters are formed
3. The clusters which are closest to each other are merged to form a single cluster
4. The same step of computing the distance and merging the closest clusters is repeated till all the points become part of a single cluster

#### Hierarchical vs K-Means

What are the benefits of Hierarchical Clustering over K-Means clustering? What are the disadvantages?

**1. In Hierarchial Clustering the algorithm gives you number of cluster to be formed whereas in K-means we need to provide a number of clusters to be formed**

**2. Hierarchial clustering may need more computing power as compared to K-means clustering.**

**Suggested Answer**

Hierarchical clustering generally produces better clusters, but is more computationally intensive.

#### Hierarchical Clustering

Can you use the dendrogram to make meaningful clusters? (By looking at which elements leave and join at what height)

**Dendrogram helps us in finding meaning number of clusters by understanding how the data points leave and join a different cluster after breaking and will that cluster make sense from the business context.**

**Suggested Answer:**

Yes. It is a great tool. You can look at what stage an element is joining a cluster and hence see how similar or dissimilar it is to the rest of the cluster. If it joins at the higher height, it is quite different from the rest of the group. You can also see which elements are joining which cluster at what stage and can thus use business understanding to cut the dendrogram more accurately.

#### Hierarchical Clustering

Compare the different linkages. Which one do you think gives a well-separated dendrogram? Are there any advantages of that?

Single linkages dendrograms are not well separated dendrograms whereas complete linkage and average linkage are well separated dendrograms. Single linkage is based on min euclidean distance the cluster formed are loose cluster and can be far apart. Complete linkage is based on maximum distance between the data points that can be formed , so this gives rise to compact cluster but not far apart. Average linkage is kind of overcomes the shortcomings of single and complete.

**Suggested Answer**

Average and Complete linkage methods give a well-separated dendrogram, whereas single linkage gives us dendrograms which are not very well separated. We generally want well separated clusters.

#### Hierarchical Clustering

Play around with various linkages and number of clusters. You will be able to see the number of natural clusters from the dendrogram itself. If you want, you can change the scale as well. Which group of parameters give you the best result - For the best result, you can use your general knowledge about various Indian states. (Basically which clusters make logical sense)

**Average linkage -3 clusters-(scale both mean and variance) parameters gives the best dendrograms.**

**Basically 3 clusters makes logical sense because with 4 clusters Bihar is singled out and with 5 clusters lakshadweep & tripura give rise to another cluster but from the policy point of view and implementation of education policies across the country, the clusters formed should be uniform and more clusters may lead to dissimilarities among the states but at the same time implementation of policies should be easy.**