Understanding Clustering

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?

Types of Clustering

Clustering is a type of unsupervised learning wherein data points are grouped into different sets based on their degree of similarity.

The various types of clustering are:

Hierarchical clustering

Partitioning clustering

Hierarchical clustering is further subdivided into:

Agglomerative clustering

Divisive clustering

Partitioning clustering is further subdivided into:

K-Means clustering

Fuzzy C-Means clustering

PRACTICAL APPLICATIONS OF CLUSTERING

**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.

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.

Social Media: In the areas of social networking and social media, Cluster Analysis is used to identify similar communities within larger groups.

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.

Clustering VS Segmentation

**Clustering**: Analytics Technique

**Segmentation**: Business Problem

Ex: People, Products, Markets

Note: To do Segmentation we use clustering technique

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

segments should have intra-segment homogeneity and inter-segment heterogeneity

Mainly 3 types of segmentation are used for customer segmentation:

**Behavioural 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 the person’s profile and uses information such as age, gender, residence locality, income, etc.

Thus, 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 standardization.
* 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.