Cluster Analysis

In this article today we are going to introduce you to cluster analysis. Cluster analysis comes under unsupervised learning.

When we are dealing with dataset but data are not labelled, we use unsupervised techniques. Using unsupervised learning, we try to discover hidden structures of the data. We want to explore the data to find some structure in them but we do not have any reference to check. That is why, cluster analysis also comes under exploratory analysis. We can find the hidden structures in different ways. One of them is natural grouping of the data. Natural grouping of the data is an important analysis. When we are dealing with very large dataset one of the biggest challenges that we face is heterogeneity of data. So we try to divide the data into some homogeneous groups.

These groupings or labelings are done by the analysts based on observations with similar attributes. Clustering is the most popular type of unsupervised learning.

Now I will go through few interesting examples that I came across as a data scientist.

Pharmaceutical or Insurance companies or people who are going through regulations and compliance for their products need to submit a lot of reports. Pharmaceutical companies are submitting thousands of reports in prescribed formats every year for approval of new drugs. With the advance of deep learning, people are trying to create a standard template of reports so that medical writers can create a report with minimal changes. Document clustering, paragraph clustering or sentence clustering help a lot in this scenarios. It is nearly impossible to read thousands and thousands of paragraphs to find out contextually similar sentences from the documents and form a template. Cluster analysis play a key role here. Using deep learning techniques, now a days we can compute contextual similarities between pairs of two sentences. While learning clustering techniques, we will see how these similarities help in forming clusters or groups easily.

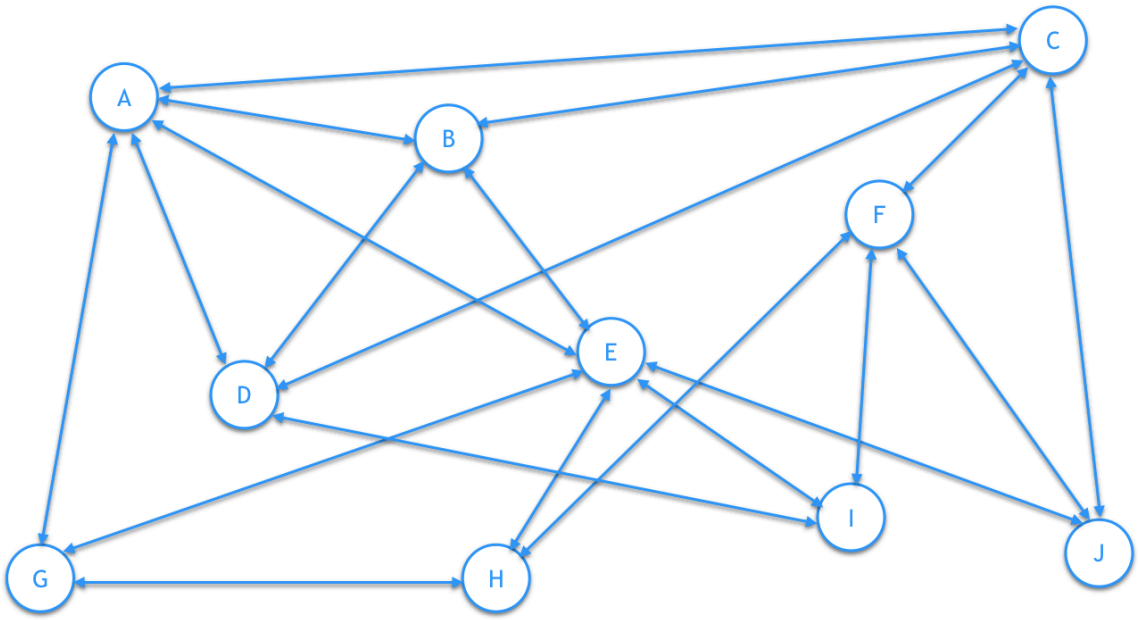
Another benefit of customer segmentation particularly in B2E business space is as follows.

ARPU or average revenue per user is the average revenue generated by a user typically on a monthly basis. This is grossly computed as total monthly revenue divided by number of subscribers. along with ARPU, companies now compute ARPU segmentwies. This helps companies to understand the effect changes in market dynamics.

Let me now introduce you to social network analysis or SNA. We will then explore how clustering on social network analysis is useful in real-life scenarios. This is something really interesting.

SNA is the analysis of relationships and flows between entities. Entity can be people, groups, organizations, countries, computers, URLs etc. The nodes in the network are entities and the links show relationships or flows between the entities. SNA provides both a visual and a mathematical analysis of the relationships.

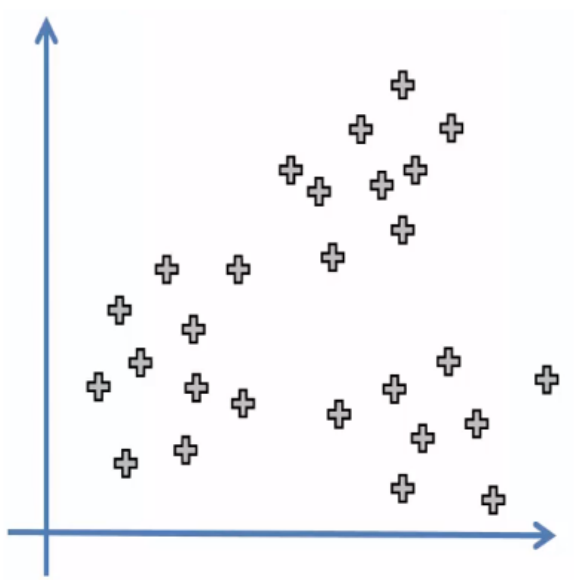
Most simple example of a Social network that we can think of is a set of telephonic conversations among us.



In the above picture we see telephonic conversations in an office in a day among people. A to J, which are the nodes of the network, denote people and the arrowed lines denote telephonic conversation between two persons.

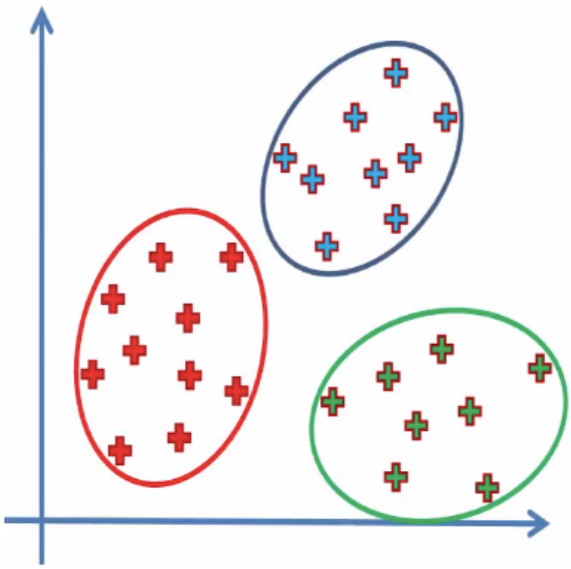
When two persons are having a telephonic conversation between them, we draw a link or an edge between two nodes that represent the two persons in the SN. The strength of relationship depends on the number of calls between two persons. One of the analysis of Social network is graph clustering. Graph clustering helps us to identify different groups, leaders in the groups etc.

Now the question is how such clusters or groups help us? One very useful application that come in my mind is fighting against terrorism. Research on usages of social network analysis to study terrorism and insurgency has increased dramatically after the 9/11 attacks against the United States. The study of relational analysis and grouping provide a variety of concepts, theories and analytical tools to better understand the behavior of militant groups. Changes in graph structure of the suspected groups in social network sometimes help to take action before any insurgency takes place.



Let us look at this example data. Here we have only data points but no prior information of grouping. But in this particular data you visually see that there are 3 natural clusters; and a clustering algorithm is supposed to come up with these 3 clusters. This is what we expect the clustering algorithm to do.

There are different aspects of clustering. First of all you need to know which clustering algorithm is suitable to your problem. There are different types of clustering algorithms available.



One of those algorithms is partition based in nature and this is the most popular one. This algorithm divides the data into a number of groups. k-means is an example of a partition based clustering algorithm. We need to set the number of clusters say 'k' at the beginning. Partitioning algorithms partition the data into k clusters. The algorithm will partition in such a way that most similar data points belong to same cluster. Later we will explore k-means in details which comes up with a local optimum based on certain criteria.

Then there are hierarchical clustering algorithms where at every step two most similar objects form a new cluster and number of clusters reduces by one. If data contain 'n' observations then it starts with n clusters and each cluster contains one observations, then repeatedly go on merging two most similar clusters until there is one cluster. This results in a nested tree.

We will talk more about k-means clustering and hierarchical clustering later.

(Dis) Similarity measure:

Throughout this article, I have repeatedly mentioned the term 'similar'. But how can we measure similarity between two

observations? Another aspect of a clustering algorithm is the metric that we use to find a distance metric or a similarity metric. There are certain distance metrics for example, one of them is Euclidean distance measure; we all know about it. The other type is Minkowski family of distance measures; The Minkowski formula for calculating distance between any two sample points **x** (x_1, x_2,…,x_n) and **y** (y_1, y_2,…y_n) is:

$$d_{x,y} = \sqrt[p]{\sum_i (|x_i - y_i|)^p}$$

Euclidean measure is a special case of it when p = 2. Cosine and Correlation are other popular measures of similarities.

Cluster validity indices:

The next aspect of clustering is how to measure the quality of the clustering algorithm. In partition based algorithms we can divide the data in 4 groups or 5 groups. Now the question is which one to choose? There are few methods that help us to find good estimate of number of clusters.