**Machine Learning Assignment 3**

**Ans 13**. Clustering methods like Hierarchical method, partitioning, density based methods, model based clustering and grid based model help in grouping the data points into clusters.

These clustering techniques help in grouping the data into similar categories and each of these categories is subdivided into subcategories to assist the exploration of queries output.

Clustering is a process that has enormous applicability. It can efficiently address diverse problems and objectives, from the simplest to the most complex.

For example, in-depth data analysis is essential for segmentation. The best way to get the most out of your clusters is to make sure the surveying is done carefully and in line with the company’s objectives.

For this, clusters need three essential steps:

1. **Identification:**Know what you are dealing with. Identifying is like putting all the pieces on the table, mapping out the situation, and sorting them using patterns.
2. **Analysis:** Analyze these patterns to make your clusters more focused and accurate.
3. **Strategy:**Create differentiated strategies for each of the clusters, with specific objectives, actions, and goals.

**Three types of clustering per application**

**Client Clustering**

Client clustering is the technique of organizing individuals according to their behavior, habits, preferences, values, social class, geographic information, and demographics, among others.

In a practical way, this technique goes beyond segmentation, because it considers more intangible variables in the group organization framework, such as mindset, opinions, and context.

It is important to know that data analysis tools are the basis of customer clustering. Getting data from various digital platforms also makes it easier to identify patterns like common interests.

Through clustering, companies optimize the quality of the messages they send to the public, such as product promotions with more acquisition power or an after-sales service in line with the previous purchase. This strengthens customer relations and consequently increases sales.

**Product Clustering**

When it comes to retail and e-commerce, products are the heart and soul of the business. A major challenge for this sector is how to be more assertive in the distribution of products.

For this to work, one of the most important techniques is the establishment of clusters.

Based on clustered data, either quantitative or qualitative (which take into account information such as: average income in the region, climate, consumption habits, age, tourist profile, among others), it is possible to predict which clothes will sell the best at each franchise. This allows logistics to send a customized shipment to each store to maximize sales.

This data is also complemented with historical information on sales in these regions; Information such as store size, average stock, and sales figures per season, among others.

In the end, clustering ensures more efficiency for retail chains and a better positioning of each unit based on its context.

**Data Clustering**

Data Clustering has more to do with operations. It’s how data is organized into clusters for archiving. This allows future queries to be answered faster, giving employees fast and easy access to data whenever they need it.

One of the many jobs of data analysts is summarizing collected information.

A very useful method is to create clusters, especially when there are a large number of observations. This can also help if the data within sets are similar to one another while at the same time being vastly different from the information in other sets,

This clustering can be applied in two ways:

1. **Hierarchical:** when the cluster’s main characteristic is an algorithm capable of providing more than one type of data partition. It generates several possible clusters, where one cluster can be merged with another.
2. **Non-Hierarchical:** non-hierarchical cluster analysis methods are characterized by the need to define an initial partition. They offer flexibility since the elements can be swapped between groups during the algorithm execution.

Clustering is very much important as**it determines the intrinsic grouping among the unlabeled data present**.

**Ans 14**. Ways to improve clustering performance are:

Tactic 1: Algorithm Tuning and Silhoutte score to identify potential clusters.

Tactic 2: Sensitivity Analysis: Calinski-Harabasz score or the Davies-Bouldin score can be used to evaluate how sensitive our tuning procedure is to the choice of the scoring function.

Tactic 3: Consensus Analysis: We need to check if a particular clustering structure is detected only by a specific algorithm (e.g., k-means). If we could replicate our findings with other clustering algorithms, we would be accumulating more substantial evidence about the fact that the clusters that we found represent some “real” characteristic of our data.

Agglomerative Clustering, confirms the presence of 6 clusters in our data when tuned with either the Silhouette score or the Calinski-Harabasz score.

When using the Davies-Boulding score, we can see a strong indication that different clusters may be present (although x clusters is the best configuration according to the scoring function).

The fact that Agglomerative Clustering also hints at the presence of clusters does not automatically imply that the clusters are the same clusters that we previously found using k-means. This is something that we need to check.

When we compare the clusters found by k-means and the clusters found by Agglomerative Clustering, we observe a one-to-one correspondence between the two. To see this, we can use a cross-tabulation.

It is useful to keep in mind that the cluster labels assigned by each algorithm are completely arbitrary: as long as we can find a one-to-one mapping between the cluster indices of one algorithm and the cluster indices of another algorithm - i.e., there are no or few off-diagonal elements after rearranging the rows or the columns of the cross-tabulation matrix - then the two algorithms are producing the same clusters. In fact, consensus between two algorithms can be found even in situations where the algorithms do not agree on the number of clusters (i.e., when the cross-tabulation matrix is not square). For instance, consider the cross-tabulation of the 6 clusters found by tuning k-means with the Silhouette score and the 8 clusters found by tuning Agglomerative Clustering with the Davies-Bouldin score.

It is a good idea to evaluate the level of agreement between different clustering algorithms on the clusters that they produce on the same dataset. After all, clustering is an unsupervised problem, and the best that we can do is to accumulate as much evidence as possible to bolster our conclusions and to try and rule out that the particular clusters that we found are a one-off result due to the choice of a specific algorithm.

It is worth pointing out that, when different clustering algorithms produce different clustering structures, we could also try to reconcile them into a final “ensemble” clustering structure. Also, as a side note, there are some situations in which different clustering algorithms are expected to produce very different clustering structures. An example is the case of clusters that have shapes which are very far from being convex or “round”. This is a situation where density-based clustering algorithms such as DBSCAN are generally expected to produce different and more meaningful results than e.g., k-means or Agglomerative Clustering.