**CLUSTERING ASSESMENT 2**

**PW SKILLS**

\*\*Q1. What is hierarchical clustering, and how is it different from other clustering techniques?\*\*

\*\*Answer\*\*: Hierarchical clustering is a clustering technique that builds a hierarchy of clusters. It doesn't require a predetermined number of clusters, unlike some other clustering techniques. It can be divided into two main types: agglomerative hierarchical clustering (bottom-up approach) and divisive hierarchical clustering (top-down approach). In agglomerative clustering, each data point starts as a separate cluster, and then pairs of clusters are merged based on similarity until all data points belong to a single cluster. Divisive clustering, on the other hand, starts with all data points in a single cluster and recursively splits them into smaller clusters until each data point is in its own cluster.

\*\*Q2. What are the two main types of hierarchical clustering algorithms? Describe each in brief.\*\*

\*\*Answer\*\*:

1. \*\*Agglomerative Hierarchical Clustering\*\*: Starts with each data point as a separate cluster and merges the closest pairs of clusters until only one cluster remains.

2. \*\*Divisive Hierarchical Clustering\*\*: Begins with all data points in a single cluster and recursively splits them into smaller clusters based on a chosen criterion until each data point is in its own cluster.

\*\*Q3. How do you determine the distance between two clusters in hierarchical clustering, and what are the common distance metrics used?\*\*

\*\*Answer\*\*: The distance between two clusters is determined using a distance metric, such as:

- Euclidean distance

- Manhattan distance

- Mahalanobis distance

- Cosine similarity

- Pearson correlation coefficient

Commonly used distance metrics depend on the type of data being clustered and the specific characteristics of the dataset.

\*\*Q4. How do you determine the optimal number of clusters in hierarchical clustering, and what are some common methods used for this purpose?\*\*

\*\*Answer\*\*: The optimal number of clusters can be determined using methods such as:

- Dendrogram visualization

- Gap statistics

- Elbow method

- Silhouette method

These methods aim to find a balance between the number of clusters and the within-cluster variance.

\*\*Q5. What are dendrograms in hierarchical clustering, and how are they useful in analyzing the results?\*\*

\*\*Answer\*\*: Dendrograms are tree-like diagrams that represent the hierarchy of clusters in hierarchical clustering. They visualize the merging (agglomerative) or splitting (divisive) of clusters at each step. Dendrograms help in understanding the relationships between clusters and deciding the optimal number of clusters by identifying the appropriate level to cut the tree.

\*\*Q6. Can hierarchical clustering be used for both numerical and categorical data? If yes, how are the distance metrics different for each type of data?\*\*

\*\*Answer\*\*: Yes, hierarchical clustering can be used for both numerical and categorical data. For numerical data, distance metrics like Euclidean distance are commonly used. For categorical data, appropriate distance metrics include:

- Jaccard distance

- Hamming distance

- Gower distance

These metrics account for the differences in the nature of categorical data compared to numerical data.

\*\*Q7. How can you use hierarchical clustering to identify outliers or anomalies in your data?\*\*

\*\*Answer\*\*: Hierarchical clustering can be used to identify outliers by examining the structure of the dendrogram. Outliers are often present as singleton clusters or as data points that are distant from any cluster. By setting a threshold distance or height in the dendrogram, outliers can be identified as points that fall beyond this threshold.

---

This assignment provides a comprehensive overview of hierarchical clustering, covering its types, algorithms, distance metrics, determination of optimal clusters, interpretation of dendrograms, handling of different data types, and identification of outliers.