SYMBIOSIS INTERNATIONAL (DEEMED UNIVERSITY)

(Established under section 3 of the UGC Act 1956)
Re - accredited by NAAC with ‘A’ Grade
Founder: Prof.Dr. S. B. Mujumdar, M.Sc.,Ph.D. (Awarded Padma Bhushan and Padma Shri by President of India)

Name: Naivedya Rai
PRN: 20070122083

Assignment 10:

Write a Python program to cluster a set of points using Hierarchical Clustering Analysis (HCA)

Theory:

Hierarchical clustering is a widely employed technique in the realms of data analysis and data mining. Its primary purpose is to categorize similar data points into clusters or groups based on their degree of resemblance or dissimilarity. This process results in the creation of a tree-like structure known as a dendrogram, where data points are gradually merged into clusters, and these clusters can, in turn, be combined into larger clusters.

The process of hierarchical clustering involves several key steps:
1. Similarity Calculation: It begins with the calculation of the similarity (or dissimilarity) between data points. This is often accomplished using distance metrics such as Euclidean distance or correlation. The chosen metric quantifies the degree of similarity between data points.

2. Linkage Method: The choice of linkage method plays a crucial role in hierarchical clustering. Methods like single linkage, complete linkage, or average linkage determine how the distance between clusters is computed. This decision significantly influences the structure of the resulting dendrogram.

One of the notable advantages of hierarchical clustering is its flexibility:
-No Predefined Cluster Number: Unlike some other clustering techniques, hierarchical clustering does not require specifying the number of clusters in advance. It naturally reveals the structure of the data by progressively forming clusters, which can be particularly advantageous when the number of clusters is not known a priori.

- Hierarchical Structure Insights: Another valuable aspect is its ability to provide insights into the hierarchical structure of the data. This means that not only does it yield information about the clusters themselves, but it also reveals how these clusters can be grouped into larger, more inclusive clusters.

Despite its strengths, hierarchical clustering has certain limitations, such as its computational cost, which can become significant when dealing with large datasets. The process of calculating distances between data points and merging clusters can be resource-intensive, making it less practical for very large datasets.

Hierarchical clustering is applied in various fields, including biology, social sciences, and marketing. Its utility lies in uncovering natural groupings or inherent structures within data, which can offer valuable insights into the relationships between data points. By revealing these groupings and hierarchical relationships, it helps researchers and analysts gain a deeper understanding of the underlying structure of their data, facilitating more informed decision-making and pattern recognition.

Code:
https://github.com/Naivedya-Rai/ML-Algo-Implement/blob/main/Lab10_HCA.ipynb