Hierarchical Ascendant Classification (HAC)

Hierarchical Ascendant Classification (HAC), often referred to as **Hierarchical Clustering**, is a clustering technique used to group similar data points into clusters. Unlike methods that create clusters all at once (like k-means clustering), HAC is a hierarchical approach that builds a tree-like structure (called a dendrogram) by gradually merging or splitting clusters. The result is a nested set of clusters that provides insight into the hierarchical relationships among the data points (or objects).

Key Characteristics of Hierarchical Ascendant Classification (HAC)

1. **Agglomerative Approach**:
    - HAC typically follows an **agglomerative** approach, which means it starts with each data point as its own individual cluster.
    - Clusters are then merged step-by-step in a "bottom-up" manner, based on their similarity, until all data points are combined into a single cluster.

2. **Dendrogram**:
    - HAC builds a **dendrogram**, a tree-like diagram that shows the merging process and represents the hierarchy of clusters.
    - By "cutting" the dendrogram at different levels, you can get different sets of clusters.

3. **Similarity and Distance Measures**:
    - HAC requires a measure of similarity or dissimilarity between data points, often expressed as **distance**.
    - Common distance measures include **Euclidean distance**, **Manhattan distance**, or **cosine similarity**.
    - The choice of distance metric affects the clusters and may depend on the nature of the data.

4. **Linkage Criteria**:
    - To decide how to merge clusters, HAC uses **linkage criteria**, which measure the distance between clusters:
        - **Single Linkage**: Uses the smallest distance between points in different clusters.
        - **Complete Linkage**: Uses the largest distance between points in different clusters.
        - **Average Linkage**: Uses the average distance between all points in the clusters.
        - **Ward's Method**: Minimizes the total within-cluster variance by merging clusters that lead to the smallest increase in variance.

How HAC Works

1. **Initialize Clusters**:
    - Begin with each data point as an individual cluster (if there are nnn data points, start with nnn clusters).

2. **Calculate Distances**:
    - Compute the pairwise distances between all clusters based on the selected distance metric.

3. **Merge Clusters**:
    - Find the two clusters that are closest to each other based on the linkage criterion and merge them into a single cluster.

4. **Update Distances**:
    - Recalculate the distances between the new cluster and all remaining clusters.

5. **Repeat**:
    - Continue merging clusters step-by-step until there is only one cluster containing all data points. Each merge forms a branch in the dendrogram, creating a hierarchy of clusters.

6. **Choose the Number of Clusters**:
    - Once the dendrogram is complete, you can "cut" it at a chosen level to obtain a desired number of clusters.
    - The optimal number of clusters can be selected by looking for large gaps between levels in the dendrogram or using criteria like the **silhouette score**.

Advantages of HAC

- **Hierarchical Insight**: HAC gives a full hierarchy of clusters, which can be helpful for understanding nested patterns within data.
- **No Need to Predefine the Number of Clusters**: Unlike kkk-means, HAC does not require specifying the number of clusters in advance.
- **Flexibility**: HAC works well with different distance metrics and linkage methods, allowing for flexibility in adapting the algorithm to various types of data.

Disadvantages of HAC

- **Computational Complexity**: HAC can be computationally expensive, especially for large datasets, as it requires calculating and updating distance matrices.
- **Sensitivity to Noise and Outliers**: Hierarchical clustering is sensitive to outliers, which may affect the quality of clustering.

Interpreting the Dendrogram

The dendrogram visually represents the hierarchical clustering process:

- **Height of Merges**: The height at which two clusters merge in the dendrogram represents the distance or dissimilarity between them. Clusters that merge at lower heights are more similar to each other.
- **Choosing a Threshold**: By selecting a threshold or "cutting" the dendrogram at a specific height, you can create a set number of clusters. For instance, a higher cut will produce fewer, broader clusters, while a lower cut will yield more, finer clusters.

See also the document entitled "8d. Hierarchical Ascendant Classification - illustrated.pdf" where you will find the step by step approach and illustrations.

References

- Murtagh, F., & Contreras, P. (2012). *Algorithms for hierarchical clustering: An overview*. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1), 86-97.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.
- Tan, P. N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining*. Pearson.
- Everitt, B., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis* (5th ed.). Wiley.
- Ward, J. H. Jr. (1963). *Hierarchical Grouping to Optimize an Objective Function*. *Journal of the American Statistical Association*, 58(301), 236-244.