Viva Questions on Hierarchical Clustering

1. **What is hierarchical clustering?**
   - Hierarchical clustering is a method of cluster analysis that seeks to build a hierarchy of clusters. It either merges smaller clusters into larger ones (agglomerative) or divides a large cluster into smaller ones (divisive).
2. **What are the two main types of hierarchical clustering?**
   - **Agglomerative Clustering:** Starts with each data point as its own cluster and merges the closest pairs of clusters iteratively.
   - **Divisive Clustering:** Starts with one, all-encompassing cluster and splits it iteratively into smaller clusters.
3. **What is a dendrogram?**
   - A dendrogram is a tree-like diagram that records the sequences of merges or splits in hierarchical clustering. It illustrates the arrangement of clusters and the distances at which clusters are merged or split.
4. **What are the common linkage methods used in hierarchical clustering?**
   - **Single Linkage (Minimum Linkage):** Measures the distance between the closest points of the two clusters.
   - **Complete Linkage (Maximum Linkage):** Measures the distance between the farthest points of the two clusters.
   - **Average Linkage (Mean Linkage):** Measures the average distance between all pairs of points in the two clusters.
   - **Ward's Method:** Minimizes the total within-cluster variance. It merges clusters that result in the smallest increase in total variance.
5. **What is the difference between single linkage and complete linkage?**
   - **Single Linkage:** Uses the shortest distance between any two points in the clusters, which can lead to "chaining" effects where clusters can form long, straggly shapes.
   - **Complete Linkage:** Uses the longest distance between any two points in the clusters, which tends to produce more compact and spherical clusters.
6. **What is Ward's method in hierarchical clustering?**
   - Ward's method aims to minimize the variance within each cluster. It merges clusters in a way that the increase in the total within-cluster variance is minimized.
7. **How do you choose the optimal number of clusters in hierarchical clustering?**
   - The optimal number of clusters can be determined by analyzing the dendrogram for the largest vertical distance that doesn't intersect any horizontal line, or by using methods like the elbow method or silhouette score on the clustered data.
8. **Explain the `linkage` function from `scipy.cluster.hierarchy`.**

   - The `linkage` function performs hierarchical clustering and returns a linkage matrix that encodes the clustering hierarchy. It takes as input a matrix of pairwise distances or an array of data points and uses a specified linkage method (e.g., single, complete, average, ward).

9. **What is the role of the `fcluster` function in hierarchical clustering?**
   - The `fcluster` function is used to form flat clusters from the hierarchical clustering result. It takes the linkage matrix and a criterion (such as the number of clusters or a distance threshold) to produce cluster labels for each data point.
10. **How is the `dendrogram` function used in hierarchical clustering?**
    - The `dendrogram` function is used to plot the hierarchical clustering result. It takes the linkage matrix and visualizes the clustering hierarchy as a tree-like structure, showing how clusters are merged or split.
11. **How do you interpret a dendrogram?**
    - In a dendrogram, the y-axis represents the distance or dissimilarity between clusters. The x-axis represents the data points or clusters. A higher point on the y-axis indicates that the clusters being merged are more dissimilar.
12. **What are the advantages of hierarchical clustering?**
    - Hierarchical clustering doesn't require the number of clusters to be specified in advance. It provides a tree-like structure of clusters which can be useful for understanding data hierarchy and relationships.
13. **What are the disadvantages of hierarchical clustering?**
    - Hierarchical clustering can be computationally intensive, especially for large datasets. It also doesn't handle noise or outliers well and can produce less interpretable results with very large datasets.
14. **How does hierarchical clustering handle noise and outliers?**
    - Hierarchical clustering doesn't handle noise and outliers explicitly. They can distort the clustering results, especially in methods like single linkage where outliers can affect cluster formation.
15. **What is the `distance` metric in hierarchical clustering?**
    - The distance metric determines how the distance between clusters is calculated. Common metrics include Euclidean distance, Manhattan distance, and cosine similarity.
16. **How does hierarchical clustering compare to k-means clustering?**
    - Hierarchical clustering creates a tree-like structure of clusters without needing the number of clusters to be specified, while k-means clustering requires specifying the number of clusters in advance and partitions data into k clusters.
17. **Can hierarchical clustering be used with categorical data?**
    - Hierarchical clustering is typically used with numerical data. For categorical data, distance measures need to be adapted, such as using Gower's distance.
18. **What is the linkage matrix in hierarchical clustering?**

- The linkage matrix is a matrix that encodes the hierarchical clustering structure. Each row of the matrix represents a merge operation and contains information on the indices of the clusters being merged and the distance between them.

19. **How do you handle large datasets in hierarchical clustering?**
    - For large datasets, hierarchical clustering can be computationally expensive. Techniques such as using a smaller sample of the data, dimensionality reduction, or approximate algorithms can help manage large datasets.
20. **What is the purpose of normalization or standardization in hierarchical clustering?**
    - Normalization or standardization ensures that all features contribute equally to the distance calculations. This prevents features with larger scales from disproportionately affecting the clustering results.
21. **How can you visualize hierarchical clustering results?**
    - Hierarchical clustering results can be visualized using dendrograms to show the merging or splitting of clusters and heatmaps to display cluster similarities.
22. **What is the impact of the choice of distance metric on hierarchical clustering?**
    - The choice of distance metric affects how clusters are formed. Different metrics can lead to different clustering results, especially in cases where clusters are not spherical or equally sized.
23. **How does the choice of linkage method affect the clustering results?**
    - Different linkage methods affect how distances between clusters are calculated, which influences the shape and compactness of the resulting clusters.
24. **What is the difference between hierarchical and partitional clustering methods?**
    - Hierarchical clustering builds a hierarchy of clusters, while partitional clustering (like k-means) divides data into a fixed number of clusters. Hierarchical methods create nested clusters, whereas partitional methods assign each data point to a specific cluster.
25. **How can you assess the quality of hierarchical clustering results?**
    - Quality can be assessed by examining the dendrogram, cluster validity indices, internal consistency, and comparing with domain knowledge or external validation metrics.
26. **Explain the role of `scipy.cluster.hierarchy.linkage` function parameters.**
    - `method`: Specifies the linkage method (single, complete, average, ward).
    - `metric`: Specifies the distance metric (e.g., euclidean, manhattan).
27. **What are the limitations of using hierarchical clustering for large datasets?**
    - Hierarchical clustering has high time complexity (O(nˆ3)) and space

complexity, making it impractical for very large datasets.

28. **Can hierarchical clustering be used for supervised learning?**
    - Hierarchical clustering is an unsupervised learning method and is not used for supervised learning tasks. However, it can be used for feature exploration and understanding the structure of data.
29. **What are the typical use cases for hierarchical clustering?**
    - Hierarchical clustering is used in gene expression analysis, taxonomy, document clustering, and any scenario where understanding data hierarchy is useful.
30. **How do you interpret cluster labels obtained from `fcluster`?**
    - Cluster labels from `fcluster` represent the assignment of each data point to a specific cluster based on the chosen criterion (number of clusters or distance threshold).
31. **What is the impact of setting different criteria in `fcluster`?**
    - Setting different criteria (like distance or number of clusters) affects how clusters are formed. For example, specifying a lower distance threshold results in more clusters.
32. **How can hierarchical clustering be combined with other clustering techniques?**
    - Hierarchical clustering can be used to determine an initial clustering structure, which can then be refined using partitional clustering methods like k-means.
33. **What are the key steps in the hierarchical clustering process?**
    - Key steps include: computing distance matrix, applying a linkage method, generating the linkage matrix, and visualizing with a dendrogram.
34. **What is the role of distance matrix in hierarchical clustering?**
    - The distance matrix is a square matrix that contains the distances between each pair of data points. It is used to determine how clusters are merged or split.
35. **How do different linkage methods handle outliers?**
    - Different linkage methods handle outliers differently. Single linkage is more sensitive to outliers, while complete linkage and Ward's method may be less affected.
36. **What is the significance of the vertical height in a dendrogram?**
    - The vertical height in a dendrogram indicates the distance at which clusters are merged. Higher heights mean that the clusters being merged are more dissimilar.
37. **How does hierarchical clustering handle non-spherical clusters?**
    - Hierarchical clustering can handle non-spherical clusters better than k-means, but the choice of distance metric and linkage method can still influence how well it captures complex shapes.
38. **What are the computational complexities of hierarchical clustering methods?**
    - Agglomerative hierarchical clustering has a time complexity of O(nˆ3) and space complexity of O(nˆ2), while divisive methods can be even

more computationally expensive.

39. **How does hierarchical clustering relate to density-based clustering methods?**
    - Hierarchical clustering creates a hierarchy based on distance, while density-based methods (like DBSCAN) focus on the density of data points. They can complement each other in identifying clusters of varying shapes.
40. **What is the role of data scaling in hierarchical clustering?**
    - Data scaling ensures that all features have equal weight in distance calculations. It prevents features with larger scales from dominating the clustering process.