5 Practical: Hierarchical Clustering and linkage measures

In Nestor, you will find the file `data_clustering.csv`, which contains 200 two-dim. feature vectors. There are no labels associated with the data points.

Implement the agglomerative hierarchical clustering algorithm as introduced and discussed in class, using simple (squared) Euclidean distance. Consider
(a) different linkage measures: *'single'*, *'average'*, *'complete'*, and *'ward'* linkage functions.
(b) three different choices for the number of clusters $K = 2, 3, 4$.

Your code should have the following structures:

- Read the file containing the data
- Apply the linkage measure and draw the dendrogram
- Define the number of clusters *K* by observing the dendrogram
- Perform the agglomerative hierarchical clustering based on the linkage function and the number of clusters
- Obtain the cluster labels for each data point
- Compute the average silhouette score as introduced in the class

You can use the build-in functions for linkage measures, the dendrogram, and the silhouette score.

Report

You should hand in a structured report comprising:

- **(1 point)** An **Introduction** section that describes your assignment.
- **(3 points)** A **Methods** section in which you explain the agglomerative hierarchical algorithm in a general manner, including different linkage measures. Code and implementation itself will also be taken into account for the grading of this section.
- **(4 points)** A **Results** section in which you provide both qualitative and quantitative results. For the qualitative results, you should include the following (for all different linkage measures and different choices of the number of clusters K):
  - A figure displaying the dendrogram and indicate the cut-off thresholds for different number of clusters using dashed horizontal lines (in total, 4 dendrograms, each with the presence of 3 dashed lines for cut-off).
  - A figure displaying the original data points and the resulting clusters (in total, 13 figures, one for the visualisation of the original data points and 12 for clustering results).

For the quantitative results, you should provide a table listing the computed silhouette score for all different linkage measures and different choices of the number of clusters (in total, 12 silhouette scores).

- **(2 points)** A **Discussion** section that includes your observations on both qualitative and quantitative results, and conclusions for the best choice of linkage measures and the number of clusters.

Bonus (suggestions)

1 point max. in total:

- Implement the silhouette score S by yourself defined as:

  $$S = \frac{1}{n} \sum_{i=1}^{n} \frac{b_i - a_i}{max(b_i, a_i)},$$

  where data point $i \in C_i$, $a_i = \frac{1}{|C_i|-1} \sum_{j \in C_i, i \neq j} d(i,j)$, $b_i = \underset{k \neq i}{min} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$, $d(i,j)$ is the (Euclidean) distance between data point i and j.

- Compute other evaluation metrics as discussed in the class, namely within cluster sum of squares (*WSS*) and between cluster sum of squares (*BSS*). Compute the sum of these two distances for different number of clusters and draw your conclusions.