CS5228 – Tutorial 3

Clustering: AGNES & Cluster Evaluation

Compared to K-Means and DBSCAN, Agglomerative Nesting (AGNES) is a hierarchical clustering method. As such, each data point may belong to different clusters depending on the hierarchy level (as AGNES yields complete clusterings, each point belongs to at least one cluster). Regarding the underlying algorithm, AGNES not only relies on the notion of distance between data points (or centroids), but also on the distance between clusters (i.e., sets of data points). This led us to the concept of *Linkage Methods* – that is, different approaches to calculate the distance between two clusters.

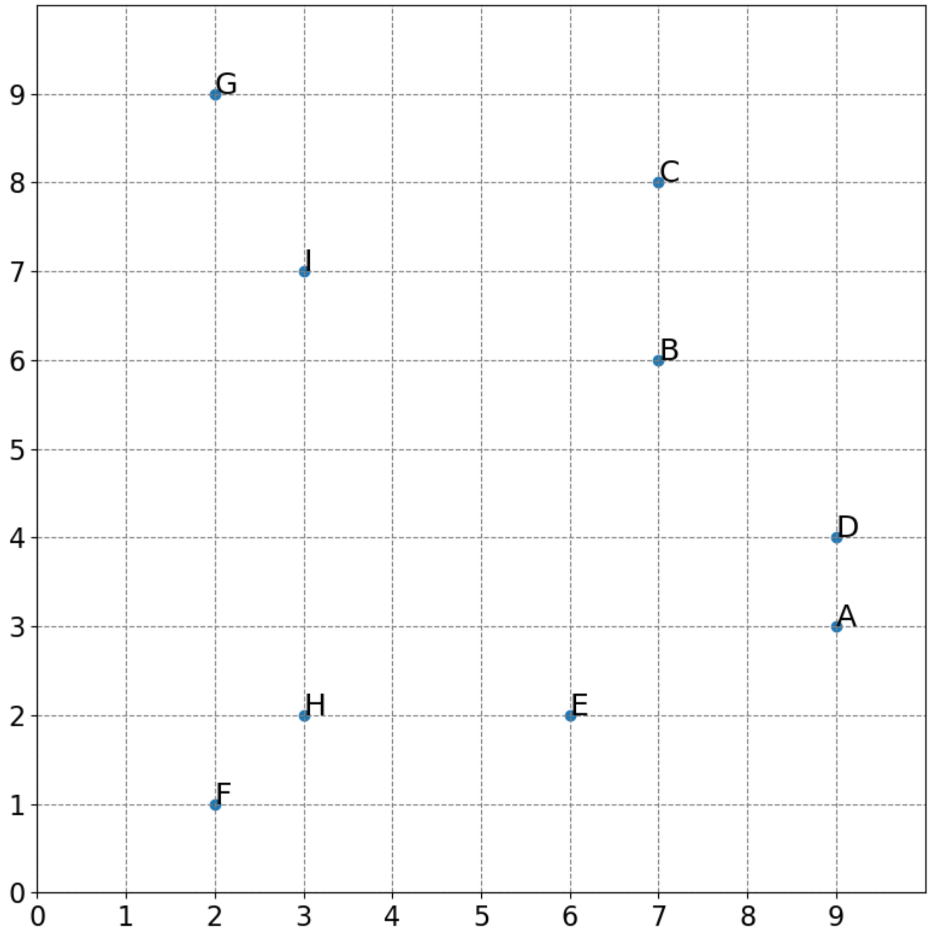

Figure 1: Toy dataset for "manually" performing AGNES.

1. **Performing AGNES "by hand".** Figure 1 shows a toy dataset with 9 data points. In the following 2 tasks, perform Hierarchical Clustering (AGNES) step by step on the dataset above. After each step write down the current set of clusters and the value of the shortest distance!

   - Denote a cluster as a sequence of points forming that cluster. For example, XYZ denotes the cluster containing the data points labeled X, Y, and Z – the order is not important
   - For each step, write down the shortest distance between the two clusters merged in each step. Note that the points are conveniently placed to make the calculation of distances pretty straightforward. If needed, round the distances to 2 decimal places (e.g., sqrt(2) = 1.41)
   - The table below shows an example of clustering 3 data points X, Y, and Z.

Start	X, Y, Z	At the start, each data point forms a cluster
1.22	Y, XZ	Cluster X and Z where closest with a distance of 1.22
2.00	XYZ	Cluster XZ and Y where closest with a distance of 2.0
End	XYZ	At the end, all data points are within a single cluster

   (a) **Perform AGNES step by step using Single Linkage!** Write down the process of forming the clusters as indicated in the example table above; you can omit the third column with the comments.

   (b) **Perform AGNES step by step using Complete Linkage!** Write down the process of forming the clusters as indicated in the example table above; you can omit the third column with the comments.

   (c) **Compare and discuss the results!** Even for this small toy dataset the resulting clustering will differ for Single Linkage and Complete Linkage. Briefly describe qualitatively, how does the choice of the linkage method affect the result! What does that mean for the potential shape of clusters on real-world data? (Hint: Check in what step the results from (a) and (b) start to differ).

   (d) **Analyze the performance!** In the lecture, we briefly mentioned that Single Linkage can be implemented more efficiently. Briefly describe qualitatively why this is the case. Having performed AGNES "by hand" in (a) and (b) should help with that.

2. **Evaluation of Clusterings** We have seen in the lecture that – apart from any ground truth in the form of labeled data – there is no perfect method to reliably evaluate the quality of a clustering. While different methods exist, they all have their limitations and it's important to be aware of those.

   (a) What are limitations of using SSE as a general metric to measure the quality of a clustering?

   (b) What does the Silhouette Score (SC) do better than SSE but which limitation still remains?

   (c) If clusterings are so difficult and unreliable to evaluate, why can we still say that clustering is such a very useful data mining method?