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.

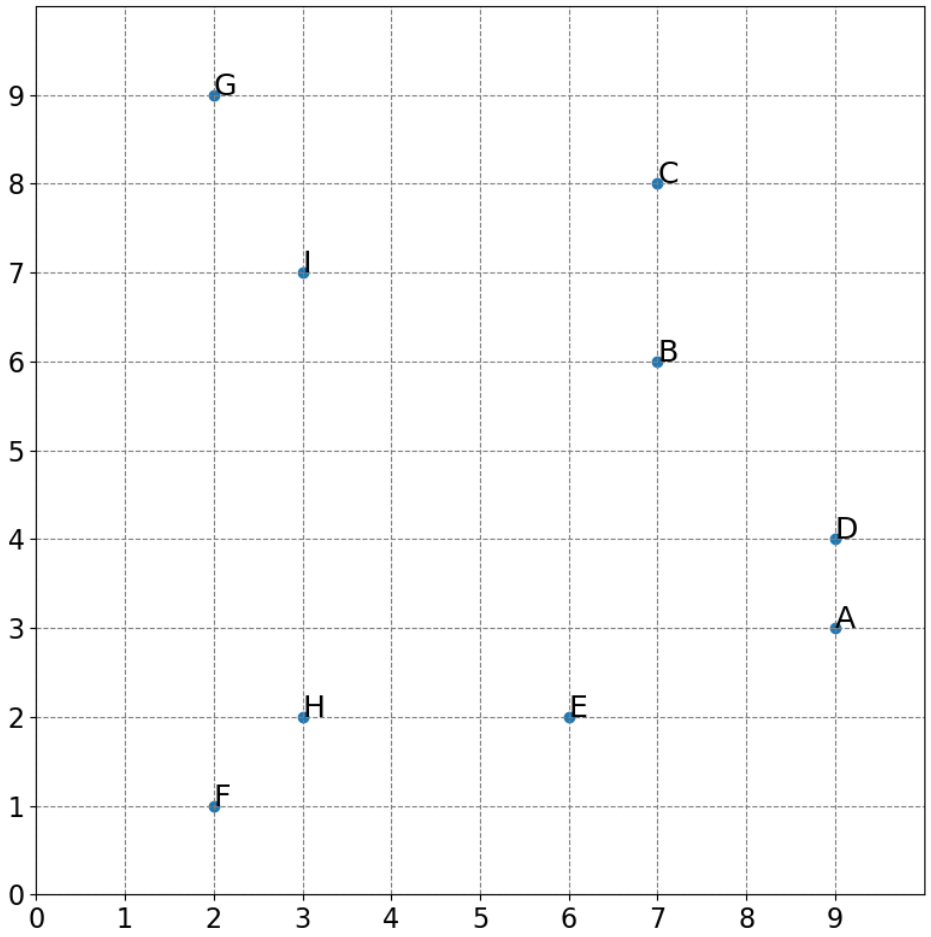

Figure 1: Toy dataset for "manually" performing AGNES.

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
   - 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.

   Solution:

Start	A, B, C, D, E, F, G, H, I
1.00	AD, B, C, E, F ,G, H, I
1.41	AD, FH, B, C, E, G, I
2.00	AD, FH, BC, E, G, I
2.24	AD, FH, BC, GI, E
2.83	ABCD, FH, GI, E
3.00	ABCD, EFH, GI
3.16	ABCDEFH, GI
4.12	ABCDEFGHI
End	ABCDEFGHI

   (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.

   Solution:

Start	A, B, C, D, E, F, G, H, I
1.00	AD, B, C, E, F ,G, H, I
1.41	AD, FH, B, C, E, G, I
2.00	AD, FH, BC, E, G, I
2.24	AD, FH, BC, GI, E
3.61	ADE, FH, BC, GI
5.83	ADE, BCGI, FH
7.62	ADEFH, BCGI
9.22	ABCDEFGHI
End	ABCDEFGHI

(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).

Solution:

- Until the 4th step resulting in (AD, FH, BC, GI, E), both Single Linkage and Complete Linkage perform the same merges; this seems intuitive as these 4 pairs of data points kind of form well-separated clusters (also: in the beginning, when most clusters still just contain a single data point, the linkage methods matters much less or not at all)
- In step 5, Single Linkage merges Clusters BC and AD since it only looks at the two closest points B and D, respectively; this would also be true if, say, Cluster would already contain more points "above" B and C as they would matter in case of Single Linkage
- In contrast, w.r.t. Complete Language, Clusters BC and AD are relatively far apart since here the two far-away points C and A matter.
- Generally speaking, Complete Linkage is more likely to yield blob-like clusters, while Single Linkage potentially may yield clusters of arbitrary shape.

(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.

Solution:

- Once we have the initial distance matrix with all pairwise distances between data points, we have already all the information to perform Single Linkage

- For Single Linkage the distance between two clusters will always be the distance between some data points (which we already calculated); this is not for the other Single Linkages.
- To convince yourself, you can create the initial 9x9 distance matrix M containing the pairwise distances between data points. You will see that all distances in the result table of (a) will be somewhere in M; the same will not hold for the result table of (b).

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?

   Solution:
   - SSE favors blob-like clusters
   - SSE is always decreasing
   - SSE does not punish large number of clusters
   - The elbow method not so straightforward to apply

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

   Solution:
   - SC is not monotonically decreasing and punishes large number of clusters
   - SC still favors blob-like clusters

   (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?

   Solution:
   - General-purpose data mining method ("only" distance/similarity measure required, works on unlabeled data).
   - Clustering methods are arguably intuitive, making the results (relatively) easy to interpret.

- Calculating, inspecting, evaluating clusters can be part of the EDA and data preprocessing (cf., binning & smoothing).
- Evaluation of clusterings in practice is often very pragmatic (e.g., parameter values given by the task; only individual clusters might be important, and not the complete clustering).
- Clustering provides a useful (and straightforward) meso-view on the data.