NUS | Computing

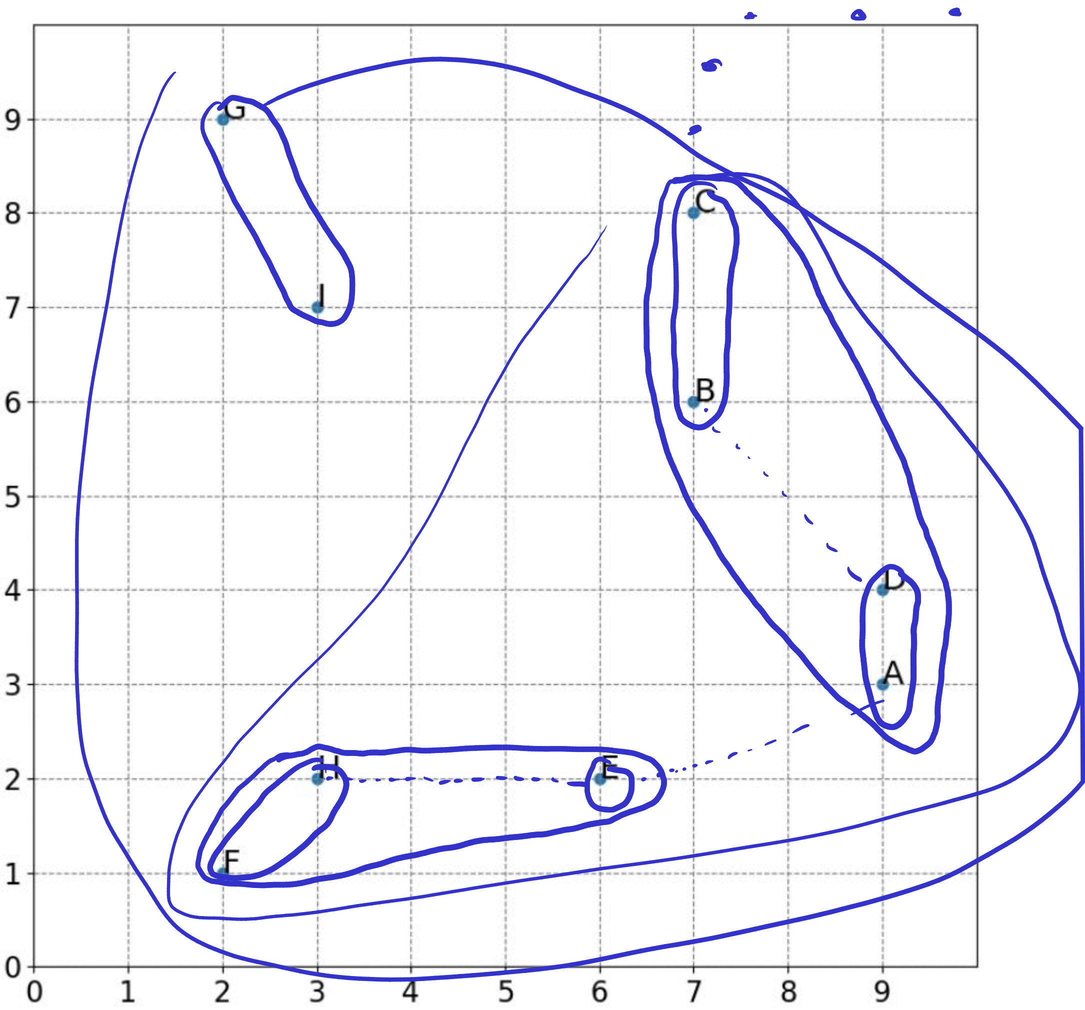
National University of Singapore

CS5228: Knowledge Discovery and Data Mining

Tutorial 3 — AGNES & Cluster Evaluation

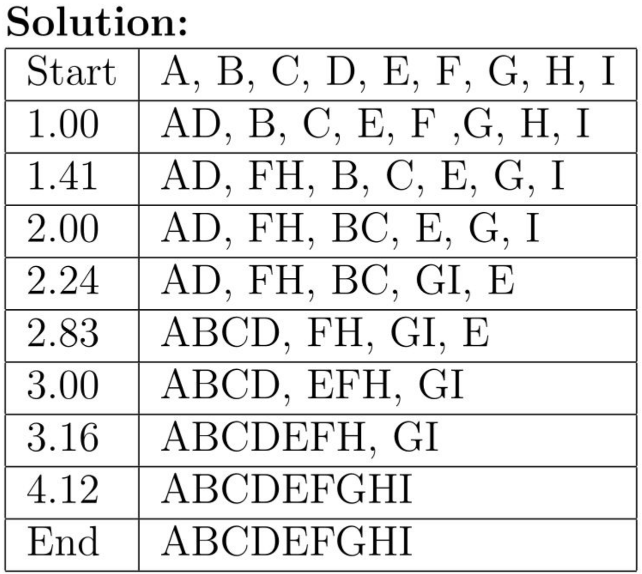
Question 1

1 a) AGNES with Single Linkage



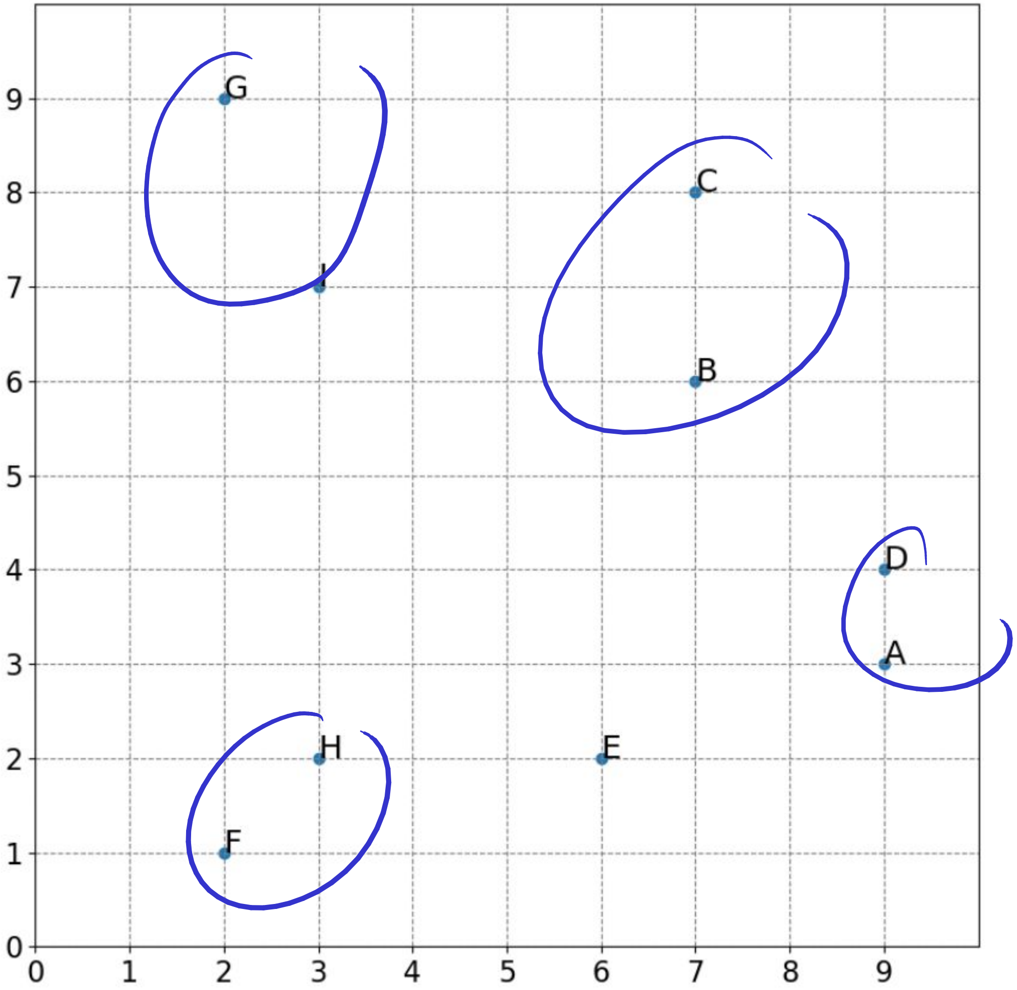
Question 1

1 a) AGNES with Single Linkage



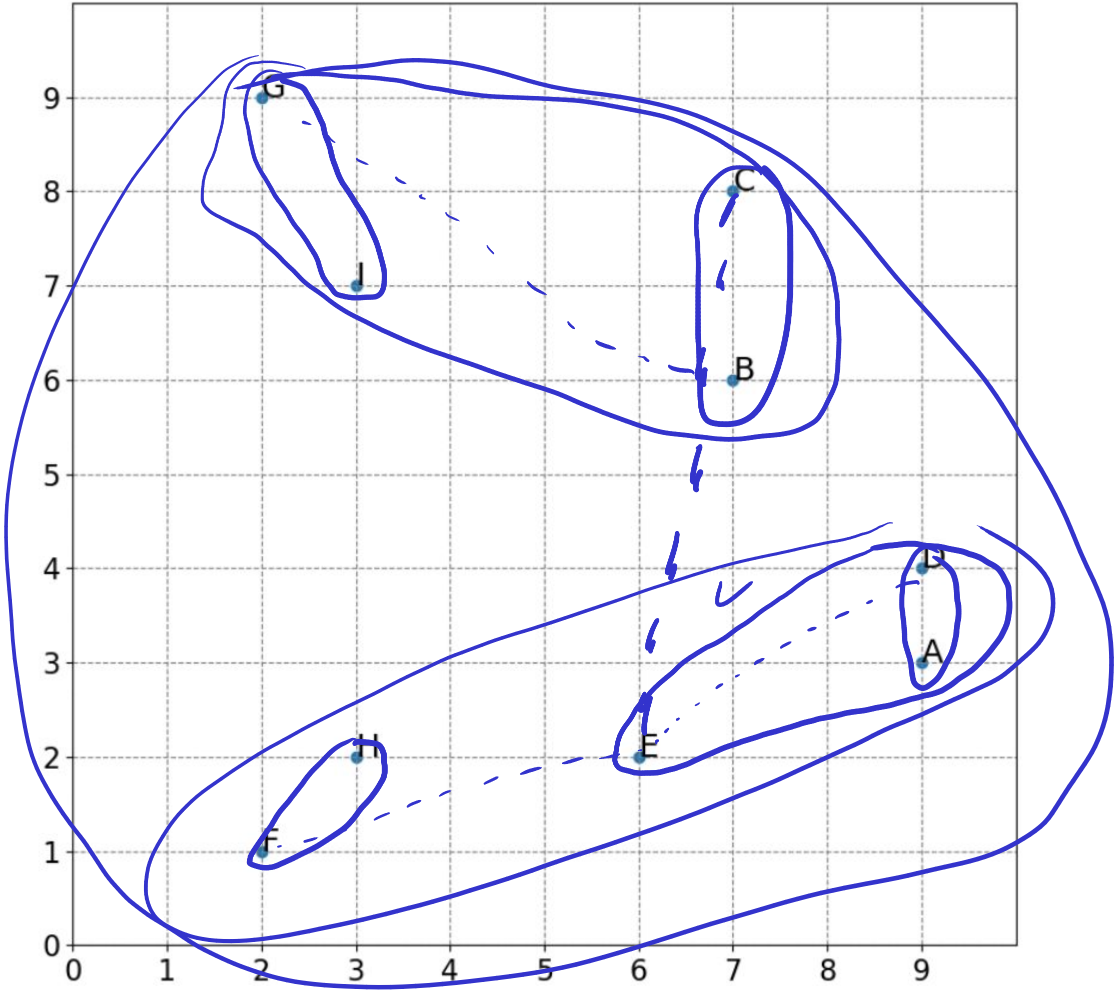
Solution:

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



Question 1

1 b) AGNES with Complete Linkage

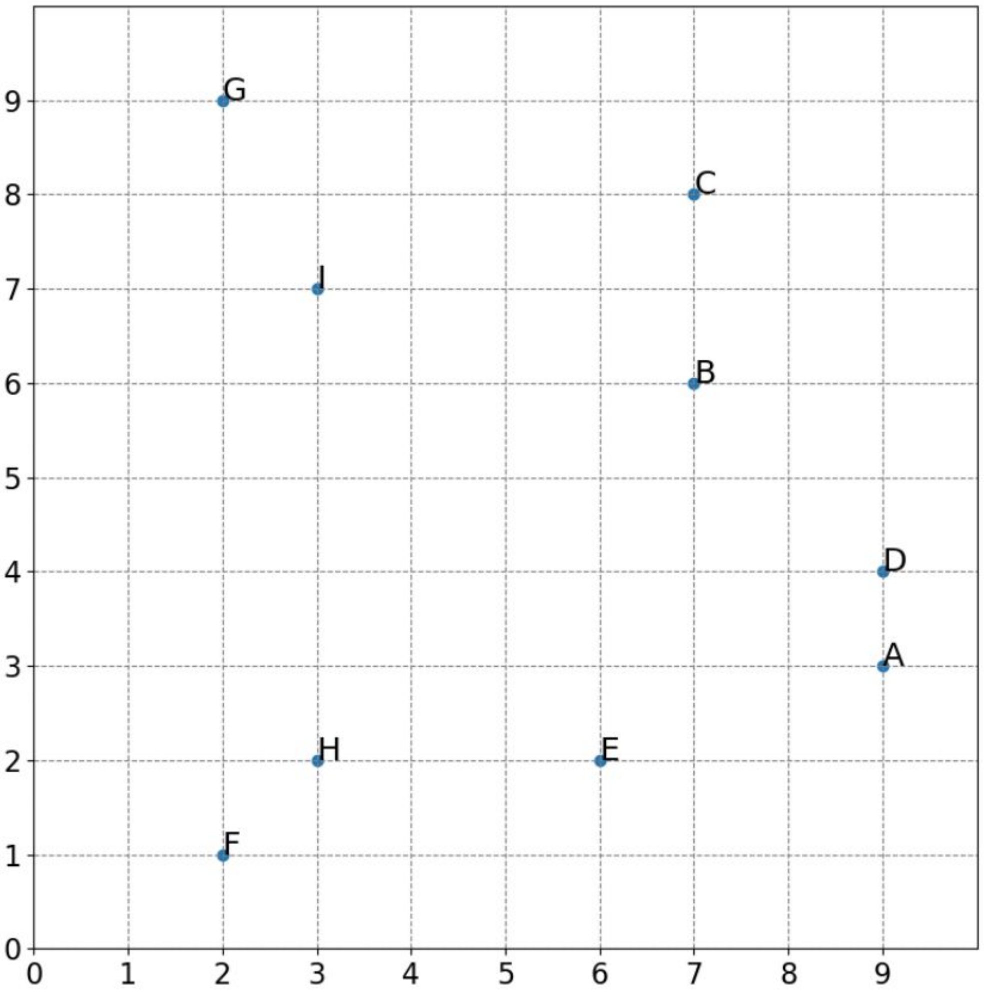


Question 1

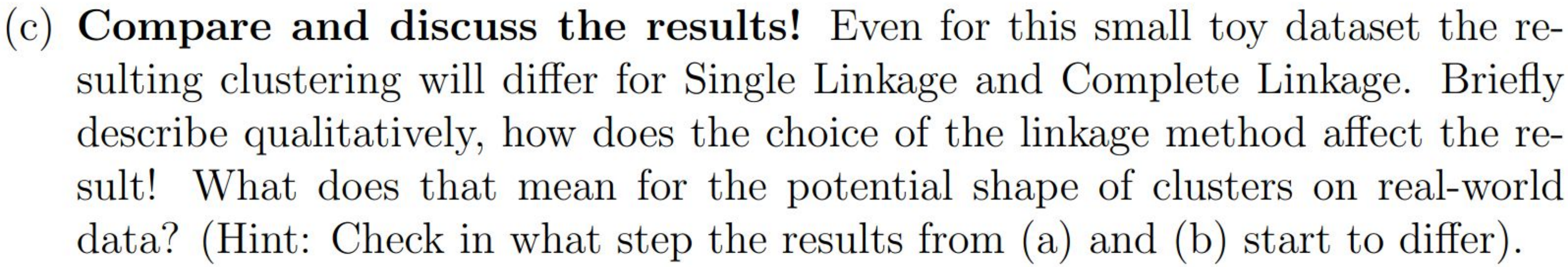
1 b) AGNES with Complete Linkage

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

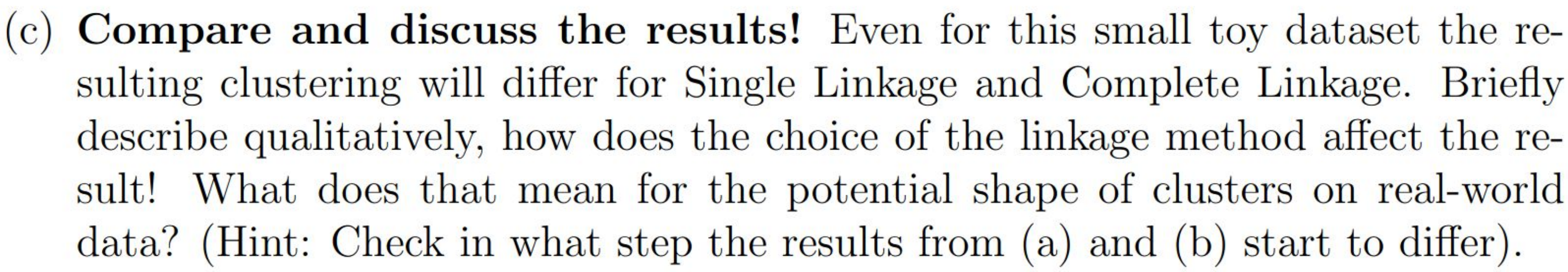


Question 1



(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).

Question 1

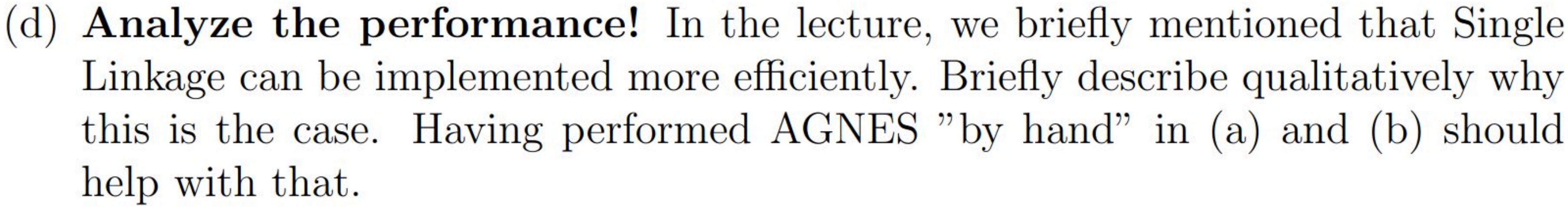


(c) **Compare and discuss the results!** Even for this small toy dataset the resulting clustering will differ for Single Linkage and Complete Linkage. Briefly describe qualitatively, how does the choice of the linkage method affect the result! What does that mean for the potential shape of clusters on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

Solution

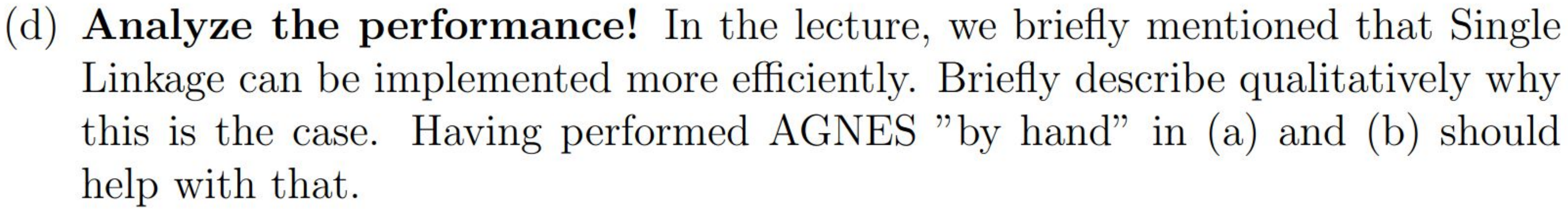
- Until the 4th step resulting in (AD, FH, BC, GI, E), both SL and CL same merges
  (this seems intuitive as these 4 pairs of data points kind of form well-separated clusters)
- In Step 5, SL Clusters BC and AD since it only looks at the two closest points B and D
  (this would also be true if, say, Cluster would already contain more points "above" B and C)
- In contrast, w.r.t. CL, Clusters BC and AD are relatively far apart
- Generally speaking, CL is more likely to yield blob-like clusters

Question 1



(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.

Question 1

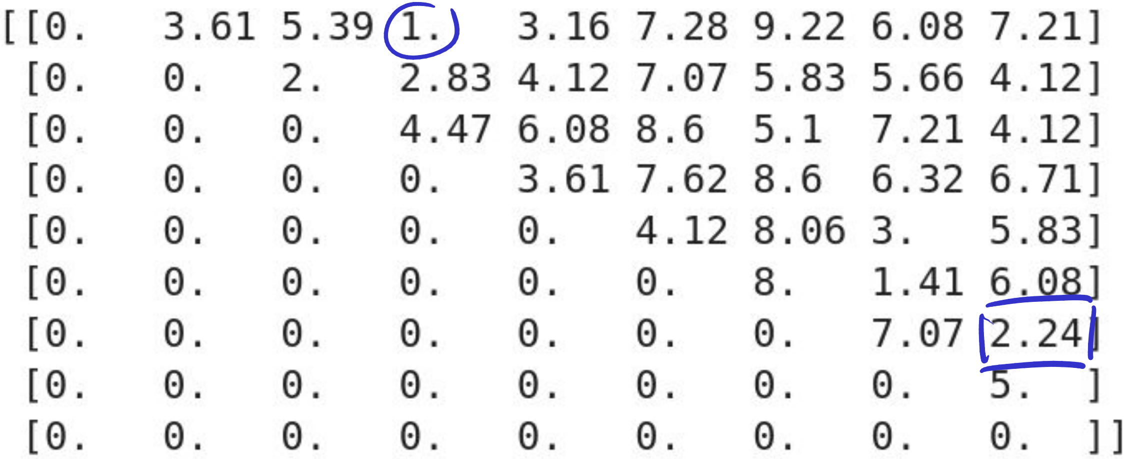


(d) **Analyze the performance!** In the lecture, we briefly mentioned that Single Linkage can be implemented more efficiently. Briefly describe qualitatively why this is the case. Having performed AGNES ”by hand” in (a) and (b) should help with that.

Solution

- For Single Linkage, we only need to ~~compare~~ Calculat all pairwise distances between only once (also true for Complete Linkage but not for Average Linkage)

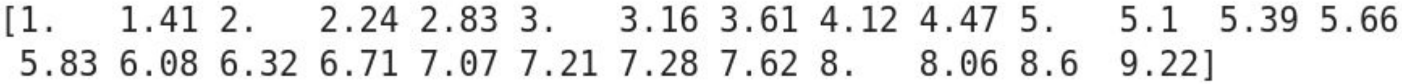
- For Single Linge we can "immediately" spot the next two cluster to by merges



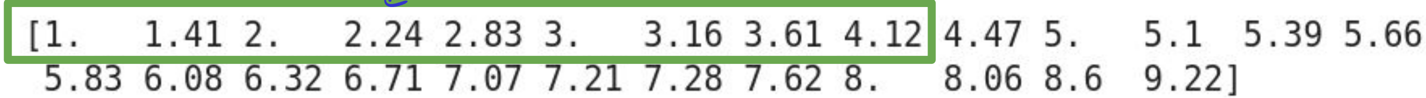
$$\begin{array}{cccccccccc}
 & A & B & C & D & E & F & G & H & I \\
A & 0. & 3.61 & 5.39 & \boxed{1.} & 3.16 & 7.28 & 9.22 & 6.08 & 7.21 \\
B & 0. & 0. & 2. & 2.83 & 4.12 & 7.07 & 5.83 & 5.66 & 4.12 \\
C & 0. & 0. & 0. & 4.47 & 6.08 & 8.6 & 5.1 & 7.21 & 4.12 \\
D & 0. & 0. & 0. & 0. & 3.61 & 7.62 & 8.6 & 6.32 & 6.71 \\
E & 0. & 0. & 0. & 0. & 0. & 4.12 & 8.06 & 3. & 5.83 \\
F & 0. & 0. & 0. & 0. & 0. & 0. & 8. & 1.41 & 6.08 \\
G & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 7.07 & \boxed{2.24} \\
H & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 5. \\
I & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0. & 0.
\end{array}$$

All pairwise distances between data points

Sort all distances (remove duplicates)

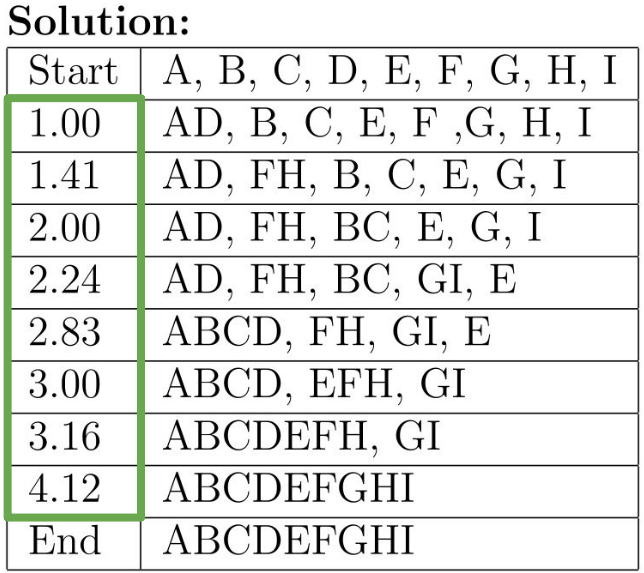


```
[1.   1.41 2.   2.24 2.83 3.   3.16 3.61 4.12 4.47 5.   5.1  5.39 5.66
 5.83 6.08 6.32 6.71 7.07 7.21 7.28 7.62 8.   8.06 8.6  9.22]
```



```
[1.   1.41 2.   2.24 2.83 3.   3.16 3.61 4.12 4.47 5.   5.1  5.39 5.66
 5.83 6.08 6.32 6.71 7.07 7.21 7.28 7.62 8.   8.06 8.6  9.22]
```

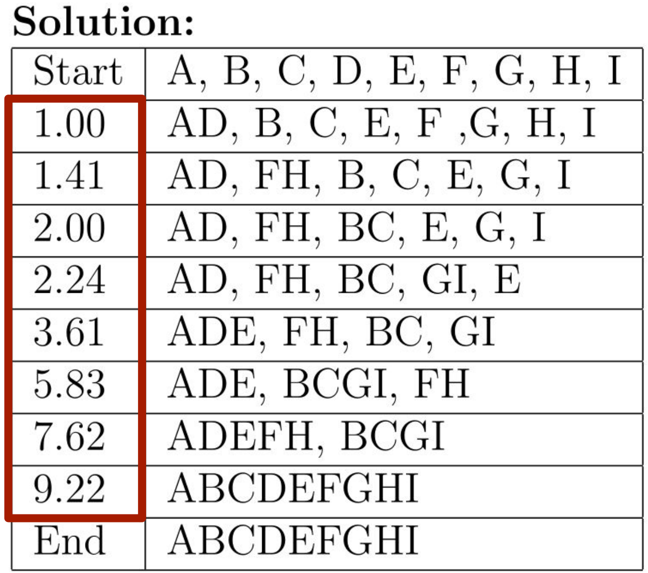
Single Linkage



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

Complete Linkage



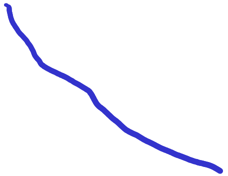
Solution:

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

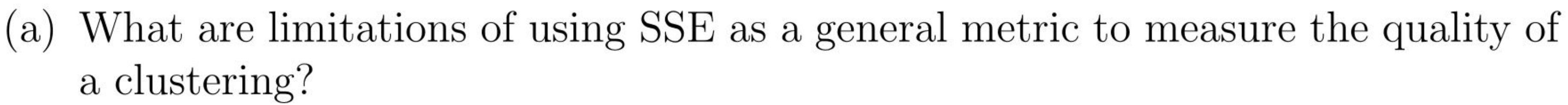
Question 2

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

- does not penalize large # cluster
- favors blob-like clusters
- elbow can be very vague



Question 2



(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

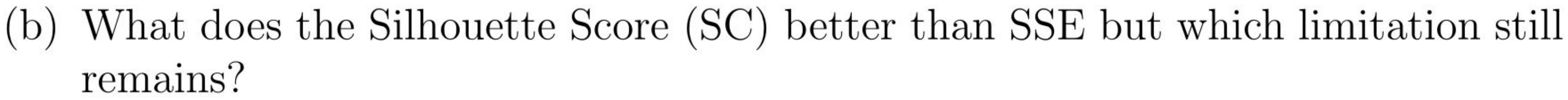
Question 2

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

+ penalizes large # clusters

− favors blobs

Question 2

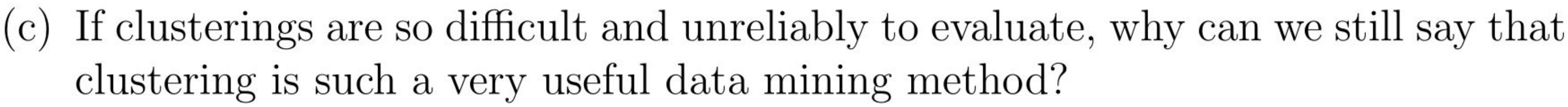


(b) What does the Silhouette Score (SC) better than SSE but which limitation still remains?

Solution

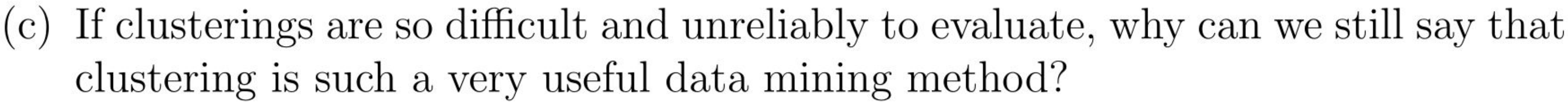
- SC is not monotonically decreasing and punishes large number of clusters
- SC still favors blob-like clusters

Question 2



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

Question 2



(c) If clusterings are so difficult and unreliably 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 often very pragmatic
  (e.g., parameter values given by the task; only individual cluster might be important, and not the complete clustering)
- Clustering provides a useful (and straightforward) meso-view on the data.