Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Thomas Seidl
Max Berrendorf, Julian Busch

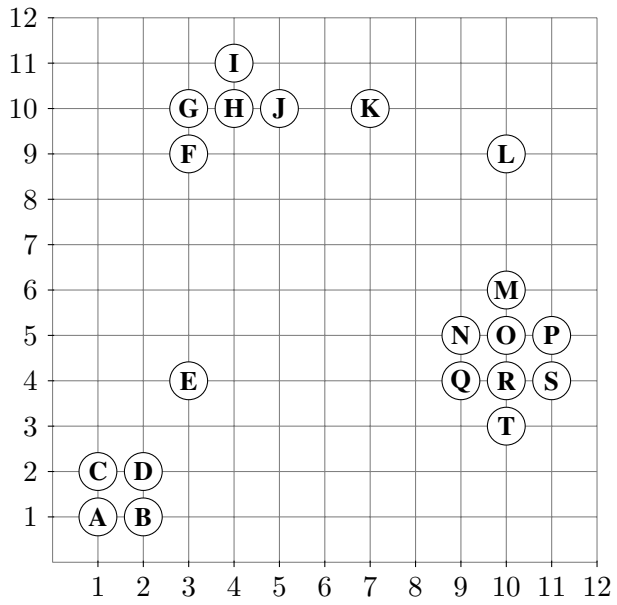
Knowledge Discovery and Data Mining I

WS 2018/19

Exercise 7: Agglomerative Clustering, OPTICS, Clustering Evaluation

Exercise 7-1 Hierarchical Clustering

Given the following data set:

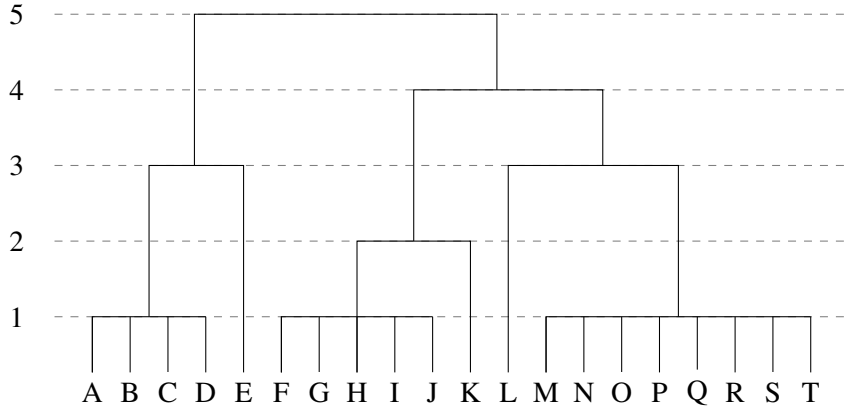


As distance function, use Manhattan Distance:

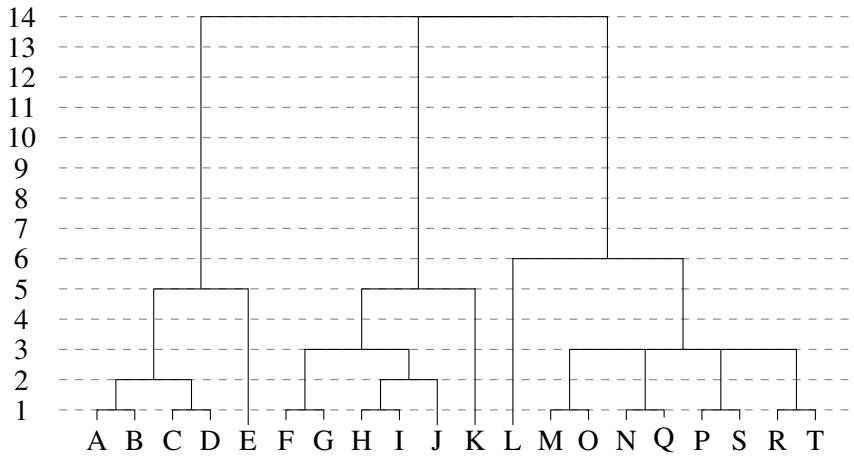
$$L_1(x, y) = |x_1 - y_1| + |x_2 - y_2|$$

Compute two dendrograms for this data set. To compute the distance of sets of objects, use

- the single-link method

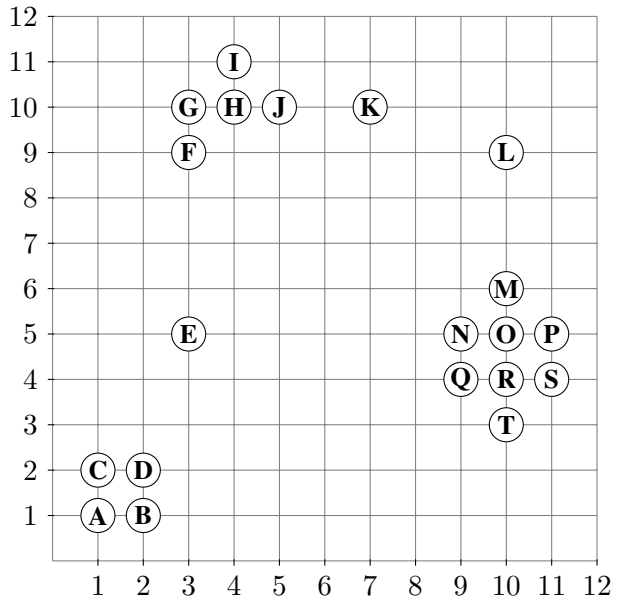


- the complete-link method



Hint: With discrete distance values, nodes may have more than two children.

Exercise 7-2 OPTICS



As distance function, use Manhattan distance $L_1(a, b) := |a_1 - b_1| + |a_2 - b_2|$.

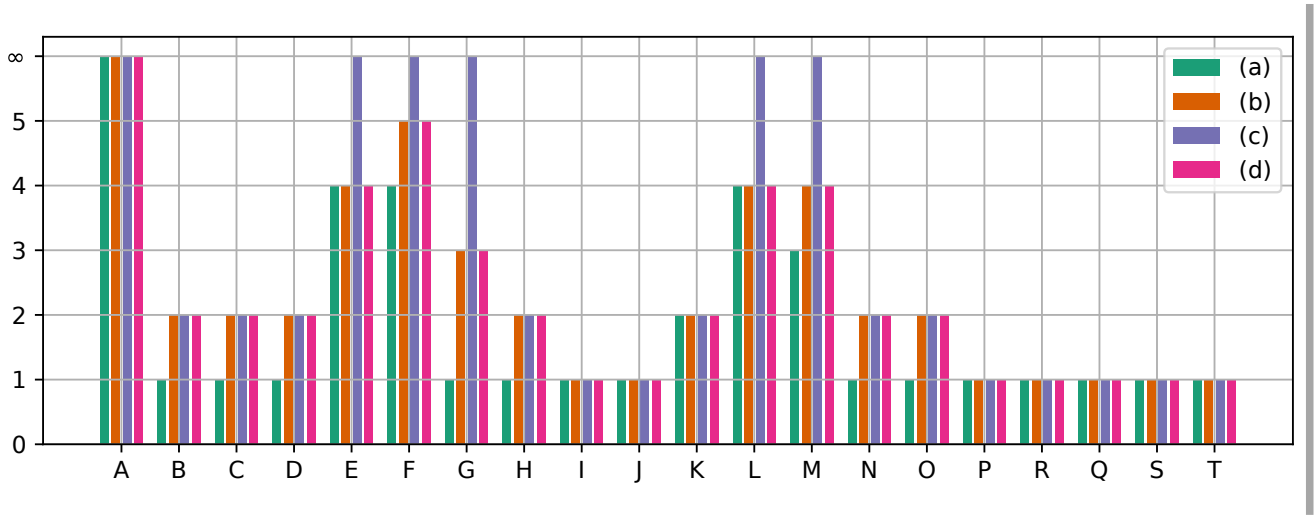
Construct an OPTICS reachability plot for each of the following parameter settings. In case of a tie always proceed with the first candidate in alphabetical order.

(a) $\varepsilon = 5$ and *minPts* $= 2$

(b) $\varepsilon = 5$ and *minPts* $= 4$

(c) $\varepsilon = 2$ and *minPts* $= 4$

(d) $\varepsilon = \infty$ and *minPts* $= 4$



Exercise 7-3 Efficient Evaluation of Clusterings

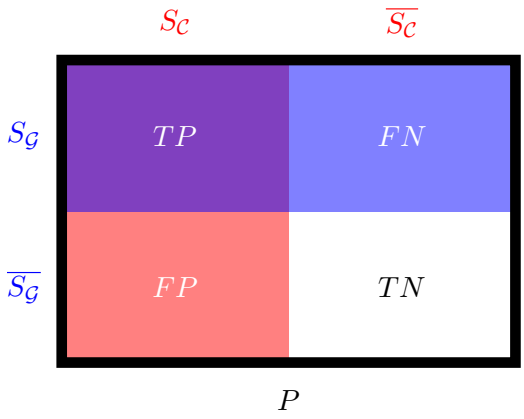
Let D be a database of size $n := |D|$, and let $\mathcal{C}, \mathcal{G}$ be two partitionings of D. Furthermore, let $k := |\mathcal{C}|$ and $l := |\mathcal{G}|$ be the number of partitions, and assume that the contingency table is provided as a $(k \times l)$ matrix, where $N_{ij} = |C_i \cap G_j|$ denotes one cell in this table.

As in the lecture slides, let $P := \{(o, p) \in D^2 \mid o \neq p\}$ denote the set of all pairs, and $S_\mathcal{C} = \{(o, p) \in P \mid \exists C_i \in \mathcal{C} : \{o, p\} \subseteq C_i\}$ be the set of pairs that are contained in a common cluster C_i in $\mathcal{C}$. In addition, $\overline{S_\mathcal{C}}$ denotes the complement of $S_\mathcal{C}$ in P, i.e. $\overline{S_\mathcal{C}} = P \setminus S_\mathcal{C}$. $S_\mathcal{G}$ and $\overline{S_\mathcal{G}}$ are used analogously.

Using these four sets, we can now define the

- *True Positives* (TP): Same labelling in $\mathcal{C}$ and same labelling in $\mathcal{G}$, i.e. $TP = |S_\mathcal{C} \cap S_\mathcal{G}|$
- *False Positives* (FP): Same labelling in $\mathcal{C}$, but different labelling in $\mathcal{G}$, i.e. $FP = |S_\mathcal{C} \cap \overline{S_\mathcal{G}}|$
- *False Negatives* (FN): Different labelling in $\mathcal{C}$, but same labelling in $\mathcal{G}$, i.e. $FN = |\overline{S_\mathcal{C}} \cap S_\mathcal{G}|$
- *True Negatives* (TN): Different labelling in $\mathcal{C}$, and different labelling in $\mathcal{G}$, i.e. $TN = |\overline{S_\mathcal{C}} \cap \overline{S_\mathcal{G}}|$

The relation of these four sets and $S_\mathcal{C}$ as well as $S_\mathcal{G}$ is also visualised in the following Venn diagram:



For each of these cardinalities, provide a method to obtain the numbers solely from the contingency table, i.e. without explicitly enumerating set of all pairs (which requires $\mathcal{O}(n^2)$ time).

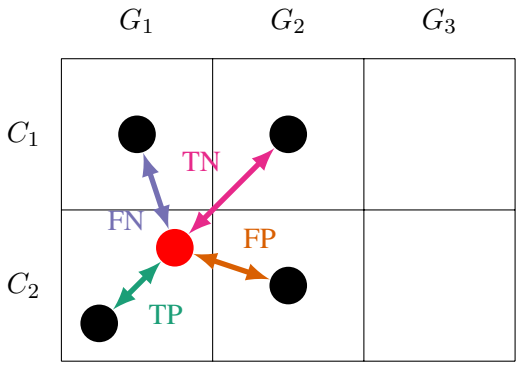
(a) $TP = |S_\mathcal{C} \cap S_\mathcal{G}|$,

(b) $FP = |S_{\mathcal{C}} \cap \overline{S_{\mathcal{G}}}|$,

(c) $FN = |\overline{S_{\mathcal{C}}} \cap S_{\mathcal{G}}|$,

(d) $TN = |\overline{S_{\mathcal{C}}} \cap \overline{S_{\mathcal{G}}}|$.

The following visualisation shall aid understanding the relation between the contingency table and the retrieval formalisation. The black and red dots represent data objects. The coloured lines indicate that this pair of data points is considered as a TP/FP/FN/TN (*Important: We only show the relation between the red dot and the black dots*). In the contingency table we only see the number of data objects that reside in a specific cell. Given those, we hence need to compute the number of pairs that fulfil the TP/FP/FN/TN constraints.



(a) The *TP* are the pairs that get assigned to the same cluster in $\mathcal{C}$ and also in $\mathcal{G}$.

$$
\begin{aligned}
TP &= |S_{\mathcal{C}} \cap S_{\mathcal{G}}| \\
&= |\{(o,p) \in P \mid \exists C_i \in \mathcal{C} : \{o,p\} \subseteq C_i\} \cap \{(o,p) \in P \mid \exists G_j \in \mathcal{G} : \{o,p\} \subseteq G_j\}| \\
&= |\{(o,p) \in P \mid (\exists C_i \in \mathcal{C} : \{o,p\} \subseteq C_i) \wedge (\exists G_j \in \mathcal{G} : \{o,p\} \subseteq G_j)\}| \\
&= |\{(o,p) \in P \mid \exists C_i \in \mathcal{C}\ \exists G_j \in \mathcal{G} : \{o,p\} \subseteq C_i \cap G_j\}| \\
&= \left| \bigcup_{i=1}^{k} \bigcup_{j=1}^{l} \{\{o,p\} \subseteq C_i \cap G_j \mid o \neq p\} \right| \\
&\overset{\clubsuit}{=} \sum_{i=1}^{k} \sum_{j=1}^{l} |\{\{o,p\} \subseteq C_i \cap G_j \mid o \neq p\}| \\
&= \sum_{i=1}^{k} \sum_{j=1}^{l} \binom{N_{ij}}{2} \\
&= \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{l} N_{ij}(N_{ij} - 1)
\end{aligned}
$$

where $\clubsuit$ uses that $C_i \cap G_j$ and $C_{i'} \cap G_{j'}$ are disjoint for $i \neq i'$ or $j \neq j'$, as $\mathcal{C}$ and $\mathcal{G}$ are partionings.

(b) The *FP* are the pairs that get assigned to the same cluster in $\mathcal{C}$ but to different ones in $\mathcal{G}$.

We have

$$|S_{\mathcal{C}} \cap S_{\mathcal{G}}| + |S_{\mathcal{C}} \cap \overline{S_{\mathcal{G}}}| \quad = \quad |S_{\mathcal{C}}|$$

as $S_{\mathcal{C}} \cap S_{\mathcal{G}}$ and $S_{\mathcal{C}} \cap \overline{S_{\mathcal{G}}}$ are a partitioning of $S_{\mathcal{C}}$, i.e. they are disjoint (no element gets counted twice), and their union equals $S_{\mathcal{C}}$ (every element is counted at least once). As, $|S_{\mathcal{C}} \cap S_{\mathcal{G}}| = TP$ is already known,

we only need $|S_{\mathcal{C}}|$:

$$\begin{aligned}
|S_{\mathcal{C}}| &= |\{(o,p) \in P \mid \exists C_i \in \mathcal{C} : \{o,p\} \subseteq C_i\}| \\
&= \left| \bigcup_{i=1}^{k} \{\{o,p\} \subseteq C_i \mid o \neq p\} \right| \\
&\overset{\clubsuit}{=} \sum_{i=1}^{k} |\{\{o,p\} \subseteq C_i \mid o \neq p\}| \\
&= \sum_{i=1}^{k} \binom{N_i^{\mathcal{C}}}{2} \\
&= \frac{1}{2} \sum_{i=1}^{k} N_i^{\mathcal{C}}(N_i^{\mathcal{C}} - 1)
\end{aligned}$$

where ♣ holds due to $\mathcal{C}$ being a partitioning, i.e., in particular, $C_i \cap C_{i'} = \emptyset$ for $i \neq i'$, and $N_i^{\mathcal{C}} = \sum_{j=1}^{l} N_{ij}$ is the row sum of the contingency table. Thus, $FP = |S_{\mathcal{C}}| - TP$.

(c) The *FN* are the pairs that get assigned to different clusters in $\mathcal{C}$ but to same in $\mathcal{G}$. We can repeat the steps from *FP* with $\mathcal{G}$ instead of $\mathcal{C}$, define $N_i^{\mathcal{G}} = \sum_{i=1}^{k} N_{ij}$ (the column sum in the contingency table), and analogously obtain

$$|S_{\mathcal{C}}| = \frac{1}{2} \sum_{i=1}^{l} N_i^{\mathcal{G}}(N_i^{\mathcal{G}} - 1), \tag{1}$$

i.e. $FN = |S_{\mathcal{G}}| - TP$.

(d) Finally, the *TN* are the pairs that get assigned to different clusters, both, in $\mathcal{C}$ and $\mathcal{G}$. Here, we can use

$$TP + FP + FN + TN = |P| = \binom{|D|}{2} = \frac{1}{2} n(n-1)$$

to obtain the value from the three others, i.e. $TN = |P| - |S_{\mathcal{C}}| - |S_{\mathcal{G}}| + TP$.

Exercise 7-4 Mutual Information

Given are two clusterings of $D = \{A, \ldots, Z\}$:

- $\mathcal{C} = \{\{A, D, E, I, K, L, M, N, T\}, \{B, F, O, Q, R, S, X, Y\}, \{C, G, H, P, V\}, \{J, U, W, Z\}\}$.
- $\mathcal{G} = \{\{A, F, M, Y\}, \{B, E, H, N, O, Q, R, Z\}, \{C, G, U, W\}, \{D, I, K, L, P, T, V\}, \{J, S, X\}\}$.

(a) Setup the contingency table, i.e. compute the sizes $|C_i \cap G_j|$ for $i = 1, \ldots, 4$, and $j = 1, \ldots, 5$.

	G_0	G_1	G_2	G_3	G_4	N_i^C
C_0	2	2	0	5	0	9
C_1	2	4	0	0	2	8
C_2	0	1	2	2	0	5
C_3	0	1	2	0	1	4
N_j^G	4	8	4	7	3	26

(b) Using the contingency table from (a), compute the entropy of $\mathcal{C}$ and $\mathcal{G}$, i.e. $H(\mathcal{C})$ and $H(\mathcal{G})$.

The entropy of $\mathcal{F} \in \{\mathcal{C}, \mathcal{G}\}$ is given by

$$H(\mathcal{F}) = -\sum_{i=1}^{4} \frac{N_i^{\mathcal{F}}}{N} \log \frac{N_i^{\mathcal{F}}}{N}$$

(we will use the logarithm to base 2 here)

	N^C	N^C/N	$\log(N^C/N)$
C_0	9	9/26	-1.531
C_1	8	8/26	-1.700
C_2	5	5/26	-2.379
C_3	4	4/26	-2.700

$H(\mathcal{C}) \approx 1.926$

	N^G	N^G/N	$\log(N^G/N)$
G_0	4	4/26	-2.700
G_1	8	8/26	-1.700
G_2	4	4/26	-2.700
G_3	7	7/26	-1.893
G_4	3	3/26	-3.115

$H(\mathcal{G}) \approx 2.223$

(c) Using the contingency table from (a), compute the mutual entropy $H(\mathcal{C} \mid \mathcal{G})$.

Mutual Entropy $H(\mathcal{C} \mid \mathcal{G})$ is computed by

$$H(\mathcal{C} \mid \mathcal{G}) \quad = \quad -\sum_{i=1}^{k} \frac{N_i^C}{N} \sum_{j=1}^{l} \frac{N_{ij}}{N_i^C} \log \frac{N_{ij}}{N_i^C}$$

First, each row is divided by the row-sum N_i^C to obtain N_{ij}/N_i

	G_0	G_1	G_2	G_3	G_4	N_i^C/N
C_0	2/9	2/9	0/9	5/9	0/9	9/26
C_1	2/8	4/8	0/8	0/8	2/8	8/26
C_2	0/5	1/5	2/5	2/5	0/5	5/26
C_3	0/4	1/4	2/4	0/4	1/4	4/26

Second, $y_{ij} = -x_{ij} \log x_{ij}$ is computed for each cell individually (where $N_{ij} > 0$), yielding:

	G_0	G_1	G_2	G_3	G_4	$\sum_j y_{ij}$	N_i^C/N
C_0	0.482	0.482		0.471		1.436	9/26
C_1	0.500	0.500			0.500	1.500	8/26
C_2		0.464	0.529	0.529		1.522	5/26
C_3		0.500	0.500		0.500	1.500	4/26

Finally, the tentative sums are weighted by N_i^C/N and added to obtain $H(\mathcal{C} \mid \mathcal{G}) \approx 1.482$

(d) Combine the results from (b) and (c) to obtain the normalised mutual information. What does this value tell about the two clusterings?

The Mutual Information (MI) is given by

$$I(\mathcal{C}, \mathcal{G}) = H(\mathcal{C}) - H(\mathcal{C} \mid \mathcal{G}) \approx 1.926 - 1.482 = 0.444$$

The Normalised Mutual Information (NMI) adds another normalisation factor

$$NMI(\mathcal{C}, \mathcal{G}) = \frac{I(\mathcal{C}, \mathcal{G})}{\sqrt{H(\mathcal{C})H(\mathcal{G})}} \approx \frac{0.444}{\sqrt{1.926 \cdot 2.223}} \approx 0.215$$

The value range of the NMI is $[0, 1]$, where a value of 1 corresponds to a perfect matching between $\mathcal{C}$ and $\mathcal{G}$. Hence, the clusterings $\mathcal{C}$ and $\mathcal{G}$ are rather dissimilar.