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 $\text{minPts} = 2$

(b) $\varepsilon = 5$ and $\text{minPts} = 4$

(c) $\varepsilon = 2$ and $\text{minPts} = 4$

(d) $\varepsilon = \infty$ and $\text{minPts} = 4$

Exercise 7-3  Efficient Evaluation of Clusterings

Let $D$ be a database of size $n := |D|$, and let $C, \mathcal{G}$ be two partitions of $D$. Furthermore, let $k := |C|$ 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 $S_C = \{(o, p) \in P \mid \exists C_i \in C : \{o, p\} \subseteq C_i\}$ be the set of pairs that are contained in a common cluster $C_i$ in $C$. In addition, $\overline{S_C}$ denotes the complement of $S_C$ in $P$, i.e. $\overline{S_C} = P \setminus S_C$. $S_G$ and $\overline{S_G}$ are used analogously.

Using these four sets, we can now define the

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

The relation of these four sets and $S_C$ as well as $S_G$ 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 $O(n^2)$ time).

(a) $TP = |S_C \cap S_G|$,  
(b) $FP = |S_C \cap \overline{S_G}|$,  
(c) $FN = |\overline{S_C} \cap S_G|$,  
(d) $TN = |\overline{S_C} \cap \overline{S_G}|$.

### Exercise 7-4 Mutual Information

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

(a) Setup the contingency table, i.e. compute the sizes $|C_i \cap G_j|$ for $i = 1, \ldots, 4$, and $j = 1, \ldots, 5$.  
(b) Using the contingency table from (a), compute the entropy of $C$ and $G$, i.e. $H(C)$ and $H(G)$.  
(c) Using the contingency table from (a), compute the mutual entropy $H(C \mid G)$.  
(d) Combine the results from (b) and (c) to obtain the normalised mutual information. What does this value tell about the two clusterings?