University of Southern Denmark
IMADA
Ricardo Campello

DM583: Data Mining

Exercise 4: Hierarchical Agglomerative Clustering (HAC)

Exercise 4-1 Single- and Complete-Linkage

a) Perform *Single-Linkage* step-by-step on the following distance matrix with pairwise distances between 5 data objects ([1], [2], …, [5]), then draw the final dendrogram:

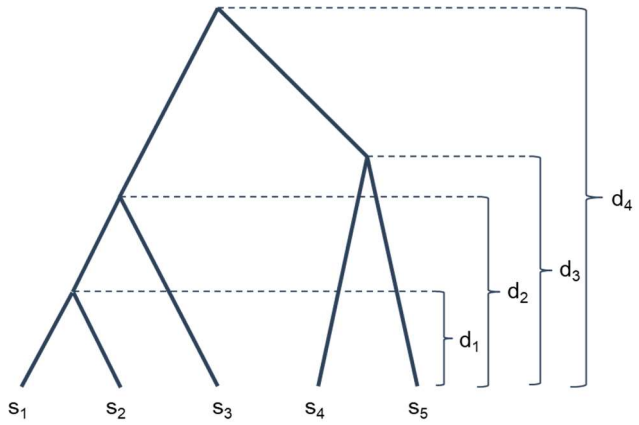
D =

	[1]	[2]	[3]	[4]	[5]
[1]	0	2	6	10	9
[2]	2	0	5	9	8
[3]	6	5	0	4	5
[4]	10	9	4	0	3
[5]	9	8	5	3	0

b) Repeat item (a) but now for *Complete-Linkage*.

Exercise 4-2 Average-Linkage (UPGMA)

a) In addition to being a general-purpose algorithm for clustering problems, irrespective of any particular domain of application, certain properties of the UPGMA algorithm make this algorithm also suitable for applications in the specific task of constructing *rooted phylogenetic trees* in biology. A rooted phylogenetic tree is essentially an ancestor-descendant tree representing the ancestry relationships between a group of species and their common ancestors, based on some measure of evolutionary distance between them. For instance, consider the following *rooted phylogenetic tree* for a group of five species:



This tree suggests that there is a common former ancestor (the root node at the top) to all 5 existing species S1 through to S5, which has split into two different descendants, one of which is an ancestor for (S1, S2, S3) whereas the other one is an ancestor for S4 and S5. The latter split into S4 and S5 through evolution, whereas

the former further branched up by first separating S3 from a common ancestor to S1 and S2, which eventually split into these two. The vertical measurements in the figure indicate the presumed evolutionary distance between groups of species and their more recent common ancestor: d1 is the evolutionary distance between (S1, S2) and their closest common ancestor, d2 is the distance between (S1, S2, S3) and their closest common ancestor, and so forth. For reasons that are beyond the scope of this exercise, if the tree and distances in the figure are an accurate representation of the evolutionary process in question, then UPGMA algorithm can be shown to be guaranteed to exactly recover the correct tree and the corresponding evolutionary distances to common ancestors assuming that pairwise distances between current (existing) species can be measured and are proportional to their evolutionary distances to their closest common ancestor. Show that this is indeed the case in the above example by manually computing the UPGMA algorithm on the following distance matrix between species, which assumes that the distance between any two species is precisely given by their evolutionary distance to their closest common ancestor, i.e. (hint: notice from the figure that d1 < d2 < d3 < d4):

	S_1	S_2	S_3	S_4	S_5
S_1	0				
S_2	d_1	0			
S_3	d_2	d_2	0		
S_4	d_4	d_4	d_4	0	
S_5	d_4	d_4	d_4	d_3	0

b) The pairwise distance between two species (as a proxy for the evolutionary distance to their common ancestor) can be estimated by measuring dissimilarity between certain molecular/biological sequences (genes or proteins). One possibility is to measure the difference between two aligned segments of amino acid sequences corresponding to a key protein family of interest, such as the following ungapped multiple alignment of the fragments of *Cytochrome C* from four different species, namely, *Rickettsia Conorii*, *Rickettsia Prowazekii*, *Bradyrhizobium Japonicum* and *Agrobacterium Tumefaciens*, as shown below (in top-down order):

NIPELMKTANADNGREIAKK
NIQELMKTANANHGREIAKK
PIEKLLQTASVEKGAAAAKK
PIAKLLASADAAKGEAVFKK

Use the evolutionary distance defined by the formula $dij = -\ln(1 - pij)$, where pij is the fraction of mismatches in the pairwise alignment of sequences i and j, to build a rooted phylogenetic tree for the given sequences by the UPGMA. Note: This problem is originally from [M. Borodovsky & S. Ekisheva "Biological Sequence Analysis", 2006].

c) If pairwise distance estimates fail to satisfy certain critical assumptions as roughly outlined in item (a), then the evolutionary tree distances computed by UPGMA (i.e., the heights of mergers along the dendrogram vertical scale) are not necessarily equal or proportional to them, as they were in item (a). Show that this is the case in the example in item (b).

Exercise 4-3 HAC in R and Dendrogram Interpretation

a) Compute and plot Ward's dendrogram for the `iris` data set using function `hclust()` from the `stats` package available in base R, with Euclidean distance, which can be readily computed using function `dist()`. Then perform a horizontal cut through the dendrogram to obtain a partition with 2 clusters using function `cutree()`. Note: Use only the numerical variables `iris[1:4]` for the clustering. The 5th variable, `iris$Species`, can only be used to (optionally) display the class labels at the bottom of the dendrogram.

b) Repeat item (a) for Single-Linkage (SL), Complete-Linkage (CL), and Average-Linkage (AL).

c) The following figure displays the dendrogram resulting from the application of a HAC algorithm (AL) to a real-world dataset consisting of gene-expression measurements (from hundreds of genes) from a particular type of tissue sampled from a few tens of different patients, which have been clustered. Interpret the dendrogram.