Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
June 16th, 2023

Problem 2: Similarities between soil pollutant molecules

We aim to investigate the similarities among various soil pollutants by analyzing the chemical formulae of their main component. One of your collaborators has devised a metric distance measure for chemical formulae based on classic distances between graph objects. The file `molecules.txt` contains the matrix of the pairwise distances [1] between 90 different molecules of soil pollutants. In order to explore the similarities between these molecules, we will employ a cluster analysis approach.

a) Which clustering methods discussed in class are suitable for this case? Provide a precise justification for your answer.

b) Perform hierarchical clustering of the molecules using *average linkage*. Report the dendrogram. Determine an appropriate number of clusters and cut the dendrogram accordingly. Report the sizes of the resulting clusters.

c) Now, let us explore the DBSCAN approach.

Using a `minPts` value of 3 and choosing a consistent value for `eps` (with an accuracy of .05), run the DBSCAN algorithm.

Justify your choice for `eps` and report any computations or plots involved in this choice. Also, provide the number of clusters discovered and their respective sizes. Is the result satisfactory?

d) Run DBSCAN again, this time with `minPts` $= 10$ and `eps` $= 0.15$. Report the number of clusters identified and their sizes.

Suggest a quantitative method for comparing the quality of the clustering results obtained from this DBSCAN run and the hierarchical clustering conducted in b). Based on this method, select the best clustering procedure.

e) Is there a way to visualize the molecules in a two-dimensional plot? Report the plot, showing through it the results of the chosen clustering procedure from the previous question. Assess whether the plot tends to underestimate or overestimate the true distances.

Upload your results here:
`https://forms.office.com/e/MPQNhPWtuq`

[1] The matrix can be read as a classic table and then converted to a R `dist()` object with the command `as.dist()`