STAT3019/M019/G019 Exercises 3

You can submit solutions on Moodle until Wednesday 31 January, 13:00, and will get personal feedback.

The workshops will be used for discussion of the Exercises.

I repeat the last question from last week's Exercises 2 as question 1 here.

1. Show that Manhattan, Simple Matching and Jaccard distance fulfill the triangle inequality. Find a counterexample that shows that the correlation dissimilarity does not fulfill the triangle equality.

   Can the triangle inequality be violated for d_G with continuous values only and d_l the absolute difference (i.e., as in the Manhattan distance) if there are missing values?

2. Consider Figure 3.1 in the course notes. Invent a dissimilarity measure for time series of the form $\mathbf{x} = (x_1, \ldots, x_T) \in I\!R^T$ so that $d(\mathbf{x}_1, \mathbf{x}_2) > d(\mathbf{x}_1, \mathbf{x}_3)$ in Figure 3.1, and that generally two time series are similar if their "up and down patterns" are similar, rather than their values for all time points. Check that it is a dissimilarity measure. Does it fulfill the triangle inequality?

   (Note that time series could be treated as variables and then the correlation dissimilarity could be used, but please try to invent something else; the correlation dissimilarity has several disadvantages for this task. Could you imagine such disadvantages?)

3. Produce a Multidimensional Scaling graph for the Veronica data with the simple matching distance and compare it with the graph from using the Jaccard distance.

   Produce Multidimensional Scaling graphs for the Old Faithful Geyser dataset using Euclidean, Manhattan and Mahalanobis distance and compare them.

4. Consider the following dataset with $n = 3$ observations and $p = 4$ variables, the first of which is categorical (for use with the simple matching distance), the second and third are binary (Jaccard distance should be used), and the fourth is on a continuous scale. "NA" denotes missing values.

$$\begin{aligned} \mathbf{x}_1 &= (\text{blue}, 1, 1, 12) \\ \mathbf{x}_2 &= (\text{red}, 0, \text{NA}, \text{NA}) \\ \mathbf{x}_3 &= (\text{red}, 1, 0, 17). \end{aligned}$$

   What is the Gower dissimilarity between all pairs of observations? Figure out how to compute this in R (function `daisy` in package `cluster`; there is also a function `gower.dist` in the `StatMatch`-package) and check against manual calculation.

5. The following is a bit of a research project, but definitely doable.

   The Gap statistic method for estimating the number of clusters is based on $\log(S_K)$. One could apply the very same approach to S_K instead (removing all the logs in the definition of the procedure).

   Do you believe that it is possible the results of this approach will differ, in general, from the one based on $\log(S_K)$? Why or why not?

   Implement in R the method based on S_K, compare it to the one based on $\log S_K$ on any dataset and check whether results differ.

   (A proper research project would run the comparison on a systematically chosen set of simulated and real datasets. The same thing can be done for other cluster validity indexes such as the Average Silhouette Width later introduced in Sec. 5, and other clustering methods. Exploring such a thing could even lead to publishable work.)