Modern Statistics and Big Data Analysis, Exercises 3

Deadline: end of Friday 20 October. Please send your solution via email to *both* `christian.hennig@unibo.i` and `gabriele.dangella2@unibo.it`.

1. (2 points) Consider the following dataset with $n = 4$ observations and $p = 5$ variables, the first of which is categorical (for use with the simple matching distance), the second, third, and fourth are binary (Jaccard distance should be used), and the fourth is on a continuous scale. "NA" denotes missing values.

$$\begin{aligned} \mathbf{x}_1 &= (\text{blue}, 1, 1, 0, 12) \\ \mathbf{x}_2 &= (\text{red}, 0, 0, \text{NA}, \text{NA}) \\ \mathbf{x}_3 &= (\text{red}, 1, 0, \text{NA}, 17) \\ \mathbf{x}_4 &= (\text{green}, 1, 0, 0, 21). \end{aligned}$$

What are the Gower (coefficient) dissimilarities between all pairs of observations?

(a) Manually compute Gower dissimilarities based on distances for all variables separately, including variables 2-4 (Jaccard distance for a single variable, see slides).

(b) Compute the Gower dissimilarites using the `daisy`-function in R and check against the manual calculation in (a) and (b).

2. (2 points)

(a) Show that Simple Matching and Mahalanobis distance are dissimilarities and also distances (i.e., fulfill the triangle inequality).

Hint: For Mahalanobis note that $\mathbf{S} = \mathbf{U}\mathbf{D}\mathbf{U}^T$, the eigendecomposition of $\mathbf{S}$, i.e., $\mathbf{U}$ is an orthonormal matrix of eigenvectors of $\mathbf{S}$, and $\mathbf{D}$ is the diagonal matrix of eigenvalues. You can then write the Mahalanobis distance between $\mathbf{x}$ and $\mathbf{y}$ as Euclidean distance between $\mathbf{D}^{-1/2}(\mathbf{U}^{-1})^T\mathbf{x}$ and $\mathbf{D}^{-1/2}(\mathbf{U}^{-1})^T\mathbf{y}$, which actually are the full principal component projections (all dimensions) of the data.

(b) Give counterexamples to show that the correlation dissimilarity (first version on slide 131) and the Gower coefficient do not fulfill the triangle inequality, i.e., in each case present three observations of which you show that they violate the triangle inequality.

Note: Just to clarify, I'm asking here for the "standard" Gower coefficient using a distance for each variable, and only L_1-, Simple Matching, or Jaccard distances. You do not have to use all of these, and in fact you can even create counterexamples using all variables of the same type with the same distance. All possible counterexamples involve missing values.

3. (2 points)

(a) A political scientist wants to cluster the respondents of a questionnaire using a distance-based method. The questionnaire has preference questions with two response options of the type "do you prefer the current level of taxes, or do you prefer higher taxes with all money from the higher taxes being invested in the health system?" Generally the option "I prefer the current situation" is coded 0 and the option that suggests something else is coded 1. Would you prefer the simple matching distance or the Jaccard distance here? Why?

(b) Geographers want to cluster areas in the Swiss alps according to danger from avalanches in order to produce a map with different colour codes for different danger levels, using a distance-based method. Their variables are, all for the year 2019: (i) the number of avalanches in the area, (ii) the average percentage of the area covered by an avalanche, (iii) number of persons injured or dead in incidents involving avalanches in the area, (iv) Swiss Francs investment in the security of the ski slopes in the area, (v) Swiss Francs budget for emergency rescue in the area. Would you prefer the Euclidean distance on raw data, the Euclidean distance on scaled data, the Manhattan distance on raw data, the Manhattan distance on scaled data, or the Mahalanobis distance for these data? Why? (Probably more than one option can convincingly be argued.)

4. (4 points) On Virtuale you can find the data set `covid2021.dat`. This data set has time series characterising the spread of Covid-19 in 179 countries. The data are from

`https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series`

(Johns Hopkins University Covid resources).

The data set has been pre-processed. Countries have been removed that were too strongly dominated by what happened on a single day, and some non-countries were either removed or included in the countries to which they belong.

The time span is 1 April 2020 to 7 October 2021. Data give for each day the number of additional cases in the previous week (to remove weekday effects) divided by the country's population (in 1,000).

The data set has 559 variables:

country Name of country,

continent The continent to which the country belongs,

latitude latitude,

longitude longitude,

X4.1.20-X10.7.21 555 daily variables giving the counts. Only these should be used for clustering, the others can be used for interpretation and visualisation.

The data set can be read by

```
covid2021 <- read.table("covid2021.dat")
cobid2021cl <- covid2021[,5:559] # This selects the variables for clustering
```

A visualisation can be produced by

```
plot(1:555,covid2021cl[1,],type="l",ylim=c(0,25),
  ylab="New cases over one week per 1000 inhabitants",
  xlab="Day (1 April 2020-7 October 2021)")
for(i in 2:179)
  points(1:555,covid2021cl[i,],type="l")
```

This will incur some heavy overplotting, and you may want to have a look at individual curves or at magnifying the lower values by setting `ylim=c(0,c)` with `c` far smaller than 25. (Later it may be of interest to look at the curves in a specific cluster.)

The task here is to cluster the countries in order to find groups of countries with similar developments. Try out one or more dissimilarity-based hierarchical clustering methods together with Euclidean and correlation dissimilarity. You may try to come up with further ideas for defining a dissimilarity for these data. Choose a number of clusters, try to understand and interpret the clusters as good as you can, using the information in the data (using visualisation as you see fit), and built yourself an opinion which of the tried out clusterings is most appropriate, and how appropriate they are in general (you may not be happy with any of them, in which case you may think about what went wrong, and how a better clustering could be achieved, even if you currently don't know how to put such ideas into practice).

Visualise the data set using multidimensional scaling (different dissimilarities will produce different MDS plots), coloring the observations according to the clusters. You may also try to produce a good heatmap of the data.