STAT3019/M019/G019 Exercises 4

You can submit solutions on Moodle until Wednesday 7 February, 13:00, and will get personal feedback.

The workshop on Thursday 8 February can be used for discussion of these exercises.

1. Produce a Multidimensional Scaling graph for the Veronica data with the simple matching distance and compared 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.

2. Choosing $K = 9$ for the unscaled Olive Oil data, compute eight K-means clusterings leaving out each single one of the variables. Compute the adjusted Rand index comparing each of these clusterings with the clustering computed on all variables.

   Do the same for the scaled Olive Oil data.

   Comment on the results.

3. For the five parties in the Bundestag dataset, construct the dendrograms with Single Linkage, Complete Linkage and Average Linkage manually (without using the computer) and compare them.

4. Prove that the Average Linkage AAHC is monotonic.

5. For the Veronica dataset, compute Single Linkage, Complete Linkage and Average Linkage clusterings for a range of values of K including $K = 8$ (e.g., $K = 2, \ldots, 20$). For each K, measure the similarity between the three clusterings by averaging the three ARI-values that you get from comparing all pairs of clusterings.

   Do you think that the value of K that maximises this is a good number of clusters for this dataset?

   Do you think that in general this is a good way to find the number of clusters?

6. Show Equation (4.1) in the course notes:

$$S(\mathcal{C}, \mathbf{m}_1, \ldots, \mathbf{m}_K) = \sum_{k=1}^{K} \frac{1}{2|C_k|} \sum_{\mathbf{x}_i, \mathbf{x}_j \in C_k} d^2_{L2}(\mathbf{x}_i, \mathbf{x}_j).$$