STAT3019/M019/G019 Exercises 1

You can submit solutions on Moodle until Wednesday 17 January, 13:00, and will get personal feedback.

The workshops will be used for discussion of the Exercises.

For some exercises you can make decisions what exactly you want to do. This is intentional. In research you often have to make your own decisions about how to proceed.

1. For one or more examples for K-means in the course notes, run K-means several (say 5 or 10) times with `nstart=1`. How many different solutions do you find?

   Doing this, is there any evidence about whether stability goes up (meaning that there are fewer different solutions) or down for larger n, and/or larger p?

   If you do the same thing with `nstart=100`, are the results always stable?

2. Run K-means for the Olive Oil data with $K = 3$ and $K = 9$. Assuming that the macro-areas are the "true" clusters for $K = 3$, use `table` to compare the macro-areas with the clustering. Do you think that this is a good clustering result in terms of matching the macro-areas? Why?

   Do the same for the regions and the $K = 9$-clustering.

   How could one measure how well the clustering matches a given grouping, i.e., how could a similarity statistic be defined that is larger if the match is better and smaller if the match is worse?

   Note that in the literature some statistics of this kind are defined, the probably most popular one being called "adjusted Rand index" (which I will introduce later in this course). Instead of trying to come up with your own idea, you may alternatively try to find out about the adjusted Rand index and compute it for the example above (the packages `mclust` and `fpc` in R can compute it).

   Note: It may be advisable to scale the data first. Techniques for comparing clusterings as discussed above could be used to decide whether using the scaled data is better here.

3. Let $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n\}$, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ for $i \in I\!N_n$, and $\mathbf{m}_1^{Km}, \ldots, \hat{\mathbf{m}}_K^{Km}$ and $c^{Km}(1), \ldots, c^{Km}(n)$ its K-means clustering for given K.

   Let $\mathcal{D}^*$ be a dataset obtained from $\mathcal{D}$ by multiplying all variables by the same constant q.

   Prove that the K-means clustering $c^{Km*}(1), \ldots, c^{Km*}(n)$ of $\mathcal{D}^*$ is the same as $c^{Km}(1), \ldots, c^{Km}(n)$. Are the corresponding centroids $\mathbf{m}_1^{Km*}, \ldots, \hat{\mathbf{m}}_K^{Km*}$ also the same as $\mathbf{m}_1^{Km}, \ldots, \hat{\mathbf{m}}_K^{Km}$? Prove it, or prove how they differ.

   (I think that it may be difficult to figure out what exactly to do here, but once you have figured it out, it should be easy to do.)