Modern Statistics and Big Data Analysis, Exercises 2

The deadline for submitting these is end of Wednesday 11 October. Please send your solution via email to *both* `christian.hennig@unibo.it` and `gabriele.dangella2@unibo.it`.

(1), 2 points

(a) Use the gap statistic method to estimate the number of clusters for the olive oil data, and for Artificial Data Set 2 (you can use $K_{max} = 10$ but you can go higher if you want). Comment on the solutions.

(b) Generate a dataset from a two-dimensional uniform distribution on the rectangle $[\min \mathbf{x}_1, \max \mathbf{x}_1] \times [\min \mathbf{x}_2, \max \mathbf{x}_2]$, where $\mathbf{x}_1, \mathbf{x}_2$ are the values of the first and second variable of Artificial Data Set 2. Compute K-means clusterings for K from 1 to 10 (obviously for $K = 1$ it's just all points in the same cluster) and show at least three scatterplots of the dataset with clusterings with different K. Compare the values of $\log S_k$ from clustering Artificial Data Set 2 in (a) with those from clustering the uniformly distributed dataset in a plot like the ones on p.94 and 100, right side, on the course slides (you don't need to plot $E(\log S_k)$ from the B uniform datasets generated by the `clusGap` function in part (a)).

(2), 4 points Often the general quality of a statistical method cannot be guaranteed by theory. Particularly in cluster analysis (but also elsewhere) theory is hard, and may cover only very special cases.

Often methods are therefore investigated by means of simulation studies, meaning that many data sets with "known truth" are artificially generated, methods are applied to them, and the quality of the outcome of the methods is then somehow measured.

The R-function `clusGap` offers various options, for example `spaceH0="original"` (using a rectangle along the main axes for the uniform distribution) and `SE.factor=1` ($q = 1$ for the factor q with which the standard error of Gap(K) is multiplied, as suggested in Tibshirani et al.'s original paper). We may be interested in whether these options could improve the results compared to the options `spaceH0="scaledPCA"` and `SE.factor=2` as used in the lecture.

As every `spaceH0` choice can be combined with every `SE.factor` choice, these options define four different ways to run the `clusGap` function.

The course notes give the code to generate data sets with the same distribution as the Artificial Data Set 1. Different data sets of the same kind can be generated by not fixing `set.seed` before generating every individual data set (it may be fixed before generating any data for making the simulation results reproducible).

Using the `gapnc`-function that automatically estimates the number of clusters for a data set with `clusGap` with given choices for `spaceH0` and `SE.factor`, generate 100 data sets according to the specifications for Artificial Data Set 1, estimate the

number of clusters by all four possible ways to run `clusGap` and compile four different vectors with all 100 estimated numbers of clusters for each version of `clusGap`. At the end look at the four distributions of estimated numbers of clusters using a suitable graphical display and comment on them. Are there better or worse results for some of the `clusGap`-versions?

Do the same thing with another model generating data. For this you need to choose a number of clusters, distributions within all clusters, true parameters for these distributions, and numbers of observations in each cluster. In a professional simulation study, these factors will be varied, to cover a range of situations, having in mind that results should give a more general indication of the quality of methods.

Here you are asked to generate data from just one further model, with 4 clusters in 4 dimensions, and numbers of observations 50, 100, 150, and 200 for the four clusters. You can use normal distributions for all clusters. These could be generated independently in the four dimensions using `rnorm`, but using the function `rmvnorm` from package `mtvnorm` you could also specify full covariance matrices. I leave to you how to choose the mean and variance (or covariance) parameters, but before running the simulation, generate a data set from your model and produce a pairs plot of it to check whether the clusters are visible (if separation between clusters is not all too strong, this is fine, because it will then not be too easy for the methods to get it right, and results can be more interesting).

(3), 2 points Regarding the gap statistic, using the notation of Sec. 2.5 of the course, consider choosing $K_{0,q}$ as optimal if it is the smallest K so that

$$\text{Gap}(K) > \text{Gap}(K^*) - qs_{K^*}, \;\; K^* = \arg\max_L \text{Gap}(L)$$

as in "Step 5". Comparing the choices $q = 1$ and $q = 2$, and assuming the same data set to be clustered, and the same B uniform data sets to be generated in "Step 2", which one of these holds: $K_{0,1} \leq K_{0,2}$, or $K_{0,1} \geq K_{0,2}$? Prove your answer (for *any* data set).

(4), 2 points Consider the following three observations from $\mathbb{R}^8$, which are visualised on p.109 of the course slides:

$$\begin{aligned} x_1 &= (1,4,5,4,2,1,1,4), \\ x_2 &= (2,3,2,2,3,3,3,3), \\ x_3 &= (7,11,11,12,9,8,8,12). \end{aligned}$$

If d is the Euclidean or L_1-distance, $d(x_1,x_2) < d(x_1,x_3)$. Invent (or find in the literature) a dissimilarity d^* on $\mathbb{R}^p$, $p > 1$, that expresses the idea that two observations are similar if they tend to be relatively larger or smaller on the same variables, so that in particular $d^*(x_1,x_2) > d^*(x_1,x_3)$.

You are asked to write down a formula for d^*, to show that it is a dissimilarity (it doesn't have to fulfill the triangle inequality but if it does and you are happy to show it, please do so), and to show by computation that indeed $d^*(x_1,x_2) > d^*(x_1,x_3)$.