Modern Statistics and Big Data Analysis, Exercises 1

The deadline for submitting these is end of Friday 29 September. Please send your solution via email to *both (!)* `christian.hennig@unibo.it` and `gabriele.dangella2@unibo.it` (Gabriele will assign the points, I will look up the solutions to see what to comment on when talking about the exercises in class).

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), 3 points Run K-means for the Olive Oil data with $K = 3$ and $K = 9$ with scaled and unscaled data. Assuming that the macro-areas are the "true" clusters for $K = 3$, use `table` to compare the macro-areas with the clustering, and compare the quality of the two clusterings with $K = 3$.

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

(2), 2 points 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 data set 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.

(3), 3 points On Moodle you can find the data set `Boston.dat`. This data set contains information collected by the U.S Census Service concerning housing in the area of Boston Mass. The data was originally published by Harrison, D. and Rubinfeld, D.L. 'Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978.

Every observation refers to a "census tract", i.e., a town or district.

The data contains the following columns:

crim per capita crime rate by town.

zn proportion of residential land zoned for lots over 25,000 sq.ft.

indus proportion of non-retail business acres per town.

chas Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).

nox nitrogen oxides concentration (parts per 10 million).

rm average number of rooms per dwelling.

age proportion of owner-occupied units built prior to 1940.

dis weighted mean of distances to five Boston employment centres.

rad index of accessibility to radial highways.

tax full-value property-tax rate per $10,000.

ptratio pupil-teacher ratio by town.

black $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town.

lstat lower status of the population (percent).

medv median value of owner-occupied homes in $1000s.

Visualise the data, produce a clustering of this data set that looks reasonable to you, and explain the reasons why you have chosen this and you think it is reasonable. You can use other clustering methods that you know other than K-means, but if you use K-means only, that's fine, too.

Note that some of the variables are very discrete, which may create problems with the clustering, so you may decide to leave one or more variables out.

Warning: Experience shows that some students put a lot of effort into a question like this, and if you have several ideas what to try out, this can easily get quite long and sophisticated. Note that you are *not* expected to do it like this. Even if you run, say, only two different clusterings such as K-means with two different numbers of clusters, and then you give a reason which one you like more, that's fine. It's surely fine if you do more, however there will not be a points premium for that, and also do not expect to get detailed feedback, as doing this for several lengthy solutions would be a lot of work. If you decide to do a lot here, it's for your own fun and exercising.

(4), 2 points "kmeans++" is the name of a method to initialise the k-means algorithm that has been proposed in the literature (for really big datasets it may be problematic to run Lloyd's or similar algorithms a lot of times from random starting points, and having just one well picked starting point will be much faster and hopefully not much worse or even better). Do some research on the internet, find out and explain how this works. It can be run by the following function `kmpp`, where `X` is a data matrix and `k` is the number of clusters. The output is of the same format as `kmeans`.

```
library(pracma)
 kmpp <- function(X, k) {
        n <- nrow(X)
        C <- numeric(k)
        C[1] <- sample(1:n, 1)

        for (i in 2:k) {
            dm <- distmat(X, X[C, ])
            pr <- apply(dm, 1, min); pr[C] <- 0
            C[i] <- sample(1:n, 1, prob = pr)
        }

        kmeans(X, X[C, ])
    }
```

Run this on one or more of the example data sets and run `kmeans` as well, both with `nstart=1` and `nstart=100`, and compare the achieved values of the objective function (`tot.withinss`-component of the output).