Master of Public Health, specialization: AI4PH

TU MET-MALE: Methods of machine learning

General instructions

This assignment consists of applying a very popular algorithm, k-means for clustering (unsupervised learning). Special attention will be paid to the interpretation and reproducibility of the results.

It is strongly recommended that you comment on your lines of code to help others understand your approach. All the implementation will be done in R language. For this, it is recommended to use RStudio to prepare your document in R markdown format.

Case study : Ketamine protocols and quality of life

Ketamine is a molecule used in the treatment of chronic pain, it acts as an anesthetic. Whatever the mechanisms by which it acts on chronic pain, there is a relationship between the dose injected and the duration of its effect. Thus, a longer duration of infusion at doses tolerated by patients seems to extend the reduction of pain.

Although recent studies confirm the therapeutic potential of ketamine in the management of chronic pain, clinical practices regarding injection protocols are far from being agreed upon.

A 12-month observational study of chronic pain patients was conducted to evaluate the effectiveness of ketamine protocols on patient quality of life.

Variables that will be used in this case-study are:

- `ag` : patient's age
- `cum_dose`: average dose of ketamine mg/kg
- `cum_days`: number of days of infusion
- `perfusion`: number of infusion hours per session
- `cost`: cost in euros per year
- `qaly`: quality of life measure

Questions

1. Install `healthdata` package from github and load the `keta` dataset (link: https://github.com/ielbadisy/healthdata/blob/master/README.md).
2. Explore the dataset by displaying its structure. Make sure that all the 6 variables mentioned above are coded in the appropriate format. Why this is important for applying k-means?
3. Use the elbow method to detect the optimal number of clusters to choose. Interpret your results. Fix the seed for reproducibility.
4. Perform the k-means clustering using the optimal k value that was found in the previous question.
5. What is the main quantity that the k-means algorithm seeks to minimize? What is its value in your results?
6. What does (between_SS / total_SS) mean? Interpret its value in your results.
7. Use the silhouette method to assess the consistency of your clustering results. You can use the `silhouette()` function from the cluster package.

8. Interpret the results based on centers' values for each clusters.
9. Now consider that the cluster membership variable is your new outcome. Use a supervised learning algorithm of your choice to build a predictive model of cluster membership. Explain the interest of this approach.
10. Why k-means is considered as an unstable algorithm?