Internship : clustering for mixed data

Laboratoire ERIC, Université Lyon 2, in collaboration with EDF

1. Context

Clustering is the task of organizing similar objects into meaningful groups. With the big data phenomenon, modern data are now high dimensional and /or heterogeneous. This provides new challenges and there is a need to develop new clustering methods adapted to such data.

In particular, we are interested in developing a clustering algorithm which is able to work with two types of data: quantitative data and functional data. Functional data are types of observation encountered when with observe a quantity over a continum, and are represented by curves.

Typical application which can be solve with such algorithm is to be able to cluster EDF customers according to both the household electricity consumption (curves) and additional information about the household (number of occupants, date of the building, …)

2. Subject

The goal of the internship is to develop a clustering algorithm for functional and quantitative data. The main missions are :

- to study the recent development in clustering methods for mixed data,
- to develop a model on the basis of an original idea proposed by the internship supervisor,
- to test this model on simulated data and onto data provided by EDF.

A publication presenting the model will be written during the internship. The intern candidate should have high skills in statistical learning, machine learning and R programming.

3. References

Y. Ben Slimen, J. Jacques and S. Allio (2020). Co-clustering for binary and functional data, Communications in Statistics - Simulation and Computation, in press.

C. Bouveyron, L. Bozzi L., J. Jacques J. and F-X. Jollois (2018). The Functional Latent Block Model for the Co-Clustering of Electricity Consumption Curves, Journal of the Royal Statistical Society, Series C, 67 [4], 897-915.

M. Selosse, J. Jacques and C. Biernacki (2020). Model-based co-clustering for mixed type data, Computational Statistics and Data Analysis, 144.

4. Internship conditions

Location : the intern will join the Data Mining & Decision team of the ERIC lab., which is composed of 11 permanent researchers in statistics and computer science.

Duration: 6 months, starting in March 2021

Salary: approx. 550€ / month

Contact: julien.jacques@univ-lyon2.fr