Projects

High-Dimensional Data Analysis and Machine Learning

Camille Mondon

1 Objectives and constraints

The objectives of this project are the following:

(i) To look for and choose an open data set from a website such as https://www.data.gouv.fr/en/datasets/. Ensure that you have a *minimum* of 80 observations, 6 numerical variables and 1 categorical variable. Clearly outline the precise source of the data and the transformations required to get your final data set in the first part of your report. Avoid relying on datasets provided by R (or Python) packages; instead generate your own data set from open data sources.

(ii) To implement the 7 statistical methods you have studied in the first part of this course (PCA, clustering, DA and CART, bootstrap, bagging and random forests) on your data set and comment the results you obtain in an appropriate and sound way.

(iii) To understand, explain and implement another specific method (depending on the project A, B, C, D, you are registered for) on your data set and comment the results you obtain in an appropriate and sound way.

You can use R (or Python) for the implementation.
Depending on the project you are registered for (A, B, C, or D), you may need to find a data set with particular characteristics. Thus, look at the new methodology you have to implement before choosing your data set. You may want to transform a numerical variable into a categorical one or a categorical variable into dummy variables.
Use simple exploratory analysis to describe your data. Provide a thorough analysis with some short explanation about the methods used to study your data. For (ii), you need to implement **at least once** each of the seven methods seen in class in your project. For (iii), the new methodology is precised below. Begin by reading the article and the R package documentation related to your project, and available on Moodle, to gain a comprehensive understanding of the methodology and how you can implement it. You are expected to give a clear presentation of this methodology during the defense and in your report. Subsequently, apply the methodology to your dataset, and provide insightful comments on the obtained results. You don't need to use all the functions of the R package.

2 Project specifics

Project A: Understand and implement on your data set the **SMOTE methodology** (package: Siriseriwan (2024), paper: Chawla et al. (2002)). Be careful that you need as the target variable a dummy variable where the number of observations in the two categories is unbalanced.

Project B: Understand and implement on your data set the **hybrid hierarchical clustering** (package: Chipman (2019), paper: Chipman and Tibshirani (2006)).

Project C: Understand and implement on your data set the **treeclust** method (package: S. Buttrey (2018), paper: S. E. Buttrey and Whitaker (2015)).

Project D: Understand and implement on your data set the **classmap** method (package: Raymaekers and Rousseeuw (2023), paper: Raymaekers, Rousseeuw, and Hubert (2022)).

3 Project report format

You need to submit a report **before 24 March 2025** satisfying the following rules:

1. It must be a **Quarto** (or Rmardown) project (a main .qmd file that can optionally call other .qmd, R scripts, etc.). See the documentation on Quarto books if necessary.
2. Your Quarto project must be **compilable** as one .pdf file which is less than 20 pages long. Be careful to set the seed so that the results you comment do not change when recompiled.
3. Please state the name of your group in the name of the file and on the cover page. Don't forget to number the pages.
4. Give a meaningful title to your project.
5. Write an introduction where you clearly state your objective, present the data you will use and give the outline of your report.
6. Write a first section where you detail the **contribution** of each of the students of the group (it cannot be the same tasks for all of you and it has to correspond to the reality).
7. Explain the difficulties you may have encountered when importing your data in R or Python, or when cleaning your data (missing values, etc.)
8. Use simple exploratory analysis to describe your data.
9. Provide a thorough analysis with explanation of the methods used to study the data (especially if the plots or the methods have not been studied during the lectures).
10. Write a conclusion to summarize your findings and provide perspectives of your work.
11. The R (or Python) chunks have to be well organized and commented.
12. Large language models can only be used *after* the code and report have been written, and only for the purpose of correcting the English at sentence level.

Don't forget to use graphics to describe your data and convey findings.

Remark. If you have difficulties uploading all the necessary sources to build your project on Moodle (for instance if the data set is too large), please contact me in order to find a solution.

4 Grading

Concerning the grading, I will take into account the interest of your project, the clarity of your objectives, the relevance of the statistical methods, the quality of your explanations and comments of the results, the quality of the R (or Python) code, the quality of the report writing.
The accent will be put on the reproducibility of your project (which means that I can compile the pdf from source while obtaining the same results) and the simplicity, homogeneity and consistence of the redaction throughout the different parts (even if written by different members of the group).

Buttrey, Sam. 2018. "treeClust: Cluster Distances Through Trees." https://cran.r-project.org/web/packages/treeClust/.

Buttrey, Samuel E., and Lyn R. Whitaker. 2015. "treeClust: An R Package for Tree-Based Clustering Dissimilarities." *The R Journal* 7 (2): 227–36. https://journal.r-project.org/archive/2015/RJ-2015-032/index.html.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. "SMOTE: Synthetic Minority Over-Sampling Technique." *Journal of Artificial Intelligence Research* 16 (June): 321–57. https://doi.org/10.1613/jair.953.

Chipman, Hugh. 2019. "Cran/hybridHclust." cran. https://github.com/cran/hybridHclust.

Chipman, Hugh, and Robert Tibshirani. 2006. “Hybrid Hierarchical Clustering with Applications to Microarray Data.” *Biostatistics* 7 (2): 286–301. https://doi.org/10.1093/biostatistics/kxj007.

Raymaekers, Jakob, and Peter Rousseeuw. 2023. “Classmap: Visualizing Classification Results.” https://cran.r-project.org/web/packages/classmap/.

Raymaekers, Jakob, Peter J. Rousseeuw, and Mia Hubert. 2022. “Class Maps for Visualizing Classification Results.” *Technometrics* 64 (2): 151–65. https://doi.org/10.1080/00401706.2021.1927849.

Siriseriwan, Wacharasak. 2024. “Smotefamily: A Collection of Oversampling Techniques for Class Imbalance Problem Based on SMOTE.” https://cran.r-project.org/web/packages/smotefamily/.