

Project - Optimal classification trees

Zacharie ALES

2024 - 2025

The objective of this project is to compute optimal classification trees for several datasets using formulation F .

1 Description of the provided Julia code

The code includes:

- 3 data sets (Iris, Wine and Seeds);
- The definition of formulation F for both axis-aligned and oblique split functions;
- A naive method to compute clusters \mathcal{C} from a dataset;
- Formulations $F_U(\mathcal{C})$, $F_S^e(\mathcal{C})$ and $F_S^h(\mathcal{C})$.

The main files and folders are:

- **File `main.jl`**

Contains one method `main()` which applies F with both axis-aligned and oblique split functions for several datasets and several tree depths ($D \in \{2, 3, 4\}$).

To execute the method, simply call `main()` in a Julia console after including the file (`include("main.jl")`).

- **File `building_tree.jl`**

Contains the definition of the functions required to solve F and F_U .

- **File `main_merge.jl`**

Similar to `main.jl` but applies F_U instead of F for partitions \mathcal{C} of different size obtained through a naive clustering method.

To execute the method, simply type `main_merge()` in a Julia console after including the file (`include("main_merge.jl")`).

- **File `main_iterative_algorithm.jl`**

Similar to `main.jl` but applies F_S^h and F_S^e instead of F for the same partitions.

To execute the method, simply type `main_iterative()` in a Julia console after including the file (`include("main_iterative_algorithm.jl")`).

- **File `merge.jl`**

Contains the functions:

- `simpleMerge()`: cluster the data using a naive algorithm described in the course;
- `exactMerge()`: cluster the data so that hypothesis H_1 is satisfied.
(`exactMerge` is not used in `main_merge()` since it does not regroup any data for the three provided datasets).

- **Folder data**
Contains the datasets.
- **Folder struct**
Contains the structures used in the code (Tree, Cluster et Distance).

2 Requested work

Apply methods `main()`, `main_merge()`, and `main_iterative()` on the 3 datasets provided and 2 additional datasets of your choice. For simplicity, only apply `main_iterative()` in the univariate case.

Remark: The time limit of each optimization model solved is initially set to 30 seconds to make your tests faster. It should be increased to at least a few minutes for the results that you will present in your report.

Address at least one of the following open questions:

1. Design and test better clustering methods which must satisfy the following conditions:
 - does not completely ignore the class of the data ;
Example : do not simply apply the k-means on all data without taking into account their class
 - does not treat each class independently
Example : do not simply apply one k-means for the data of each class

Methods that generate clusters which could contain data from several classes or that handle outliers could for example be interesting.
2. Find one or several clustering hypothesis similar to H_1 which enable to obtain optimal or near-optimal solutions;
3. Identify and use valid inequalities to improve the solution of formulation F (they can be added initially or dynamically through a callback);
Remark: Inequalities that you did not find by yourself and which do not provide an improvement will be considered insufficient.
4. Remove constraints from formulation F and add them dynamically during the solution through a callback;
5. Use opposite box corner equalities.
 - Adapt formulation F as in the practical.
 - Detect and add opposite box inequalities to the model. Evaluate their impact on the solution time, the linear relaxation and the number of node of the branch-and-bound tree.
 - To increase the number of inequalities generated, round the data features and evaluate the impact on the solution time and the performances on the training set depending on the rounding precision.
6. Find and test any ideas which can improve the performances.

3 Due date

Send me before the 31st of march:

- your files (data, code, ...);
- your report including
 - your results with comments (computation time, accuracy on the different datasets, for different tree depths, for both types of split functions).
 - the description of your work on the open question(s).

The report can be short. You do not have to present the context, Formulation F , the different types of separation or the principle of clustering data to find classification trees.