Aalto-yliopisto
Perustieteiden
korkeakoulu

Optimizing tree ensemble models

(topic presentation of BSc thesis)

Eetu Reijonen

25.08.2023

Advisor: *Nikita Belyak*

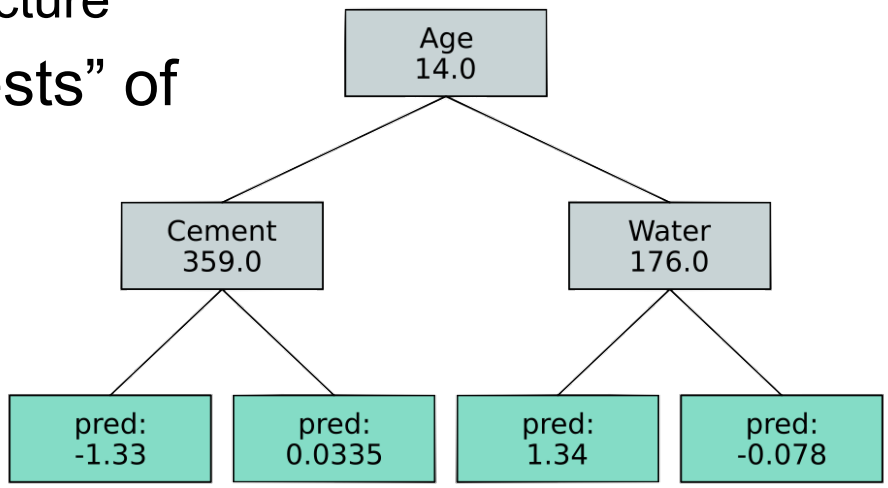
Supervisor: *Fabricio Oliveira*

Systeemianalyysin
laboratorio

Background – decision trees and tree ensembles

- A decision tree is a function mapping an input to an output "leaf"
  - Interpretable as a tree structure
- Tree ensembles are "forests" of individual decision trees
  - Popular choice in machine learning (XGBoost)

$$DT : (Age, Cement, Water) \rightarrow Concrete\ strength$$



Background – tree ensemble optimization

- How to find an input that maximizes/minimizes the tree ensemble prediction?
- Formulated as a mixed-integer optimization (MIO) problem

$$\underset{\mathbf{x},\mathbf{y}}{\text{maximize}} \quad \sum_{t=1}^{T} \sum_{\ell \in \mathbf{leaves}(t)} \lambda_t \cdot p_{t,\ell} \cdot y_{t,\ell} \tag{2a}$$

$$\text{subject to} \quad \sum_{\ell \in \mathbf{leaves}(t)} y_{t,\ell} = 1, \quad \forall\, t \in \{1,\dots,T\}, \tag{2b}$$

$$\sum_{\ell \in \mathbf{left}(s)} y_{t,\ell} \le \sum_{j \in \mathbf{C}(s)} x_{\mathbf{V}(s),j},$$

$$\forall\, t \in \{1,\dots,T\},\ s \in \mathbf{splits}(t), \tag{2c}$$

$$\sum_{\ell \in \mathbf{right}(s)} y_{t,\ell} \le 1 - \sum_{j \in \mathbf{C}(s)} x_{\mathbf{V}(s),j},$$

$$\forall\, t \in \{1,\dots,T\},\ s \in \mathbf{splits}(t), \tag{2d}$$

$$\sum_{j=1}^{K_i} x_{i,j} = 1, \quad \forall\, i \in \mathscr{C}, \tag{2e}$$

$$x_{i,j} \le x_{i,j+1}, \quad \forall\, i \in \mathcal{N},\ j \in \{1,\dots,K_i - 1\}, \tag{2f}$$

$$x_{i,j} \in \{0,1\}, \quad \forall\, i \in \{1,\dots,n\},\ j \in \{1,\dots,K_i\} \tag{2g}$$

$$y_{t,\ell} \ge 0, \quad \forall\, t \in \{1,\dots,T\},\ \ell \in \mathbf{leaves}(t). \tag{2h}$$

Mišić, V.V., 2020. Optimization of tree ensembles. *Operations Research*, *68*(5), pp.1605-1624.

Objective

- Evaluating the trade-off between prediction quality and optimization performance
  1. How tree model size affects prediction quality?
     - Depth of trees and size of the forest
  2. How tree model size affects the corresponding MIO problem optimization performance?
     - The optimization problem is NP-hard

Methods

- Programming language: Julia, tree ensemble model: EvoTrees.jl, MIO formulation: JuMP, solver: Gurobi
- Three datasets: concrete strength, drug design - OX2 and 3A4
- Computational experiments: training tree ensembles with different parameters, then optimizing the corresponding MIO problems

Limitations

- Only three datasets
  - Ideally more different types of data and numbers of variables
- Testing only with gradient boosted trees
- Only one MIO formulation used
  - Better performance could be seen with an alternative (not yet devised) formulation

Table 1: Summary of the datasets used

Dataset	No. variables	No. observations (train)	No. observations (test)
Concrete	9	772	258
OX2	5790	11151	3704
3A4	9491	37241	12338

Literature

- Mišić, V.V., 2020. Optimization of tree ensembles. *Operations Research*, *68*(5), pp.1605-1624.

Schedule

- Introduction to topic, literature and Julia 5/2023
- Code implementation 6-7/2023
- Computational experiments 7/2023
- Topic presentation 8/2023
- Thesis writing 8/2023
- Thesis presentation 11/2023