IUT de Paris - Rives de Seine
Université Paris Cité

université PARIS-SACLAY

Mathématiques Orsay

Tree-based methods (CART, Random Forests, Boosting) and interpretability

Jean-Michel Poggi
Univ. Paris Cité & LMO, Univ. Paris-Saclay, Orsay

Overview of the Course

Universidad de la República – Facultad de Ingeniería
Instituto de Matemática y Estadística "Prof. Ing. Rafael Laguardia"
Montevideo, Uruguay
February, 2025

FACULTAD DE INGENIERIA

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Overview

- *Course 1: CART trees*
- *A variant: Spatial CART*

- *Course 2: Random Forests, Variable importance and Variable Selection*
- *An application to Driver's Stress Level Classification*

- *Course 3: Boosting, from Adaboost to Gradient boosting*
- *A variant: Boosting diversity in regression ensembles*

- *Course 4: Interpretability of tree-based methods*

Course 1 - *CART trees*

- 1. Introduction
- 2. CART trees and splits
- 3. Construction: maximal tree and pruning
- 4. A typical theoretical result from the model selection viewpoint
- 5. Extensions and variants
- 6. CART in practice using the R package `rpart`

- Breiman, L., Friedman, J., Olshen, R. et Stone, C. [1984]. *Classification and Regression Trees* . Chapman & Hall, New York.
- Genuer, R., Poggi J.-M. [2020]. Random Forests with R, 98 p., Use'R!, Springer.
- Therneau, T., Atkinson, B. et Ripley, B. [2015]. `rpart`: *Recursive Partitioning and Regression Trees*

Course 1 (continuing)
Spatial CART

- 1. Spatial CART
- 2. Influence Measures for CART
- 3. Influence Measures and Stability for Graphical Models

- Bar-Hen, A., Gey, S. et Poggi, J.-M. [2021]. *Spatial CART classification trees*. Computational Statistics, 36, 2591–2613.
- Bar-Hen, A., Gey, S. et Poggi, J.-M. [2015]. *Influence measures for CART classification trees*. Journal of Classification, 32 (1), 21–45.
- A. Bar-Hen, J-M. Poggi [2016] *Influence Measures and Stability for Graphical Models*, Journal of Multivariate Analysis, Vol. 47, 145-154

Course 2 - *Random Forests, Variable importance and Variable Selection*

- 1. Introduction
- 2. Trees, Bagging and Random Forests (RF)
- 3. Extensions and variants, some theoretical results
- 4. Out-of-bag error and variable importance measure
- 5. Variable Selection using RF
- 6. RF in practice using the R packages `randomForest` and `VSURF`

- Breiman, L. [2001]. *Random forests*. Machine learning , 45 (1), 5–32.
- Genuer, R., Poggi, J.-M. et Tuleau-Malot, C. [2010]. *Variable selection using random forests*. Pattern Recognition Letters , 31 (14), 2225–2236.
- Genuer, R., Poggi, J.-M. et Tuleau-Malot, C. [2015]. `Vsurf`: *An R package for variable selection using random forests*. The R Journal , 7 (2), 19–33.
- Genuer, R., Poggi J.-M. [2020]. *Random Forests with R*, 98 p., Use'R!, Springer.

Course 2 (continuing)

An application: RF-Based Approach for Physiological Functional Variable Selection

- 1. Introduction and motivation
- 2. Physiological functional variables and wavelets
- 3. Block variables importance measure
- 4. Functional variable selection using RF
- 5. Driver's Stress Level Classification

- Gregorutti, B., Michel, B. et Saint-Pierre, P. [2015]. *Grouped variable importance with random forests and application to multiple functional data analysis*. Computational Statistics & Data Analysis, 90, 15–35.
- El Haouij N., Poggi J.-M., Ghozi R., Sevestre Ghalila S., Jaïdane M [2018], *Random Forest-Based Approach for Physiological Functional Variable Selection: Towards Driver's Stress Level Classification*. Statistical Methods & Applications Journal, 1-17

Course 3 - *Boosting, from Adaboost to Gradient boosting*

- 1. Introduction and references
- 2. Adaboost in the classification case
- 3. Additive linear models
- 4. Boosting generalizations
- 5. Gradient boosting
- 6. Boosting in practice

- Freund, Y., Schapire, R. E. [1997]. *A decision-theoretic generalization of on-line learning and an application to boosting*. Journal of computer and system sciences, 55(1) :119-139.
- Breiman, L. [1998]. *Arcing classifiers*. Annals of statistics, 26(3) :801-849.
- Friedman, J., Hastie, T., and Tibshirani, R. [2000]. *Additive logistic regression: A statistical view of boosting*. Annals of statistics, 337-374.

Course 3 (continuing) - *A variant: Boosting diversity in regression ensembles*

- 1. Introduction and motivation
- 2. Diversity
- 3. Boosting diversity algorithm
- 4. A theoretical result
- 5. Illustrative examples and hyperparameters
- 6. Real examples

- Bourel, M. Cugliari, J. Goude, Y. Poggi, J.-M. [2023]. *Boosting Diversity in Regression Ensembles*, Stat. Anal. Data Min.: ASA Data Sci. J. 1-17

Course 4

Interpretability of tree-based methods

- 1. Interpretability post-hoc: generalities and references
- 2. Interpretability in the random forests (RF) case
- 3. Importance measures: principles and usefulness
- 4. Global/Local Importance measures for RF (MDI and MDA)
- 5. Partial Dependence Plot (PDP)
- 6. Some packages

- Molnar, C. [2022]. *Interpretable Machine Learning, A guide for making black box models explainable*, 2nd edition, Lulu. com
- Geurts P. [2023]. *Random forests-based Variable importance measures,* Lecture notes, ECAS-SFdS 2023 School, 8-13 September, 2023.