Research internship: Analyzing Random Forests from a theoretical perspective

1 Introduction

Tree-based methods, and random forests (RF) in particular [3], are state-of-the-art methods to handle classification and regression tasks with tabular data. Even if RF exist for more than two decades, several of their key features remain to be understood. Such methods are based on the following elements - tree aggregation, sampling mechanisms (bootstrap), feature selection and small-sample statistics - making theoretical analyses notoriously difficult.

First axis - understanding the role of the hyperparameter `max-features` Among all RF hyperparameters, the number of directions eligible for splitting is the more difficult to tune. Modifying it can drastically change RF performances. Besides, only few results (excess risk upper bounds mostly) depend on thys hyperparameter, thus creating some uncertainty about its role on predictive performances. Recently, [4, 5] advocate that adding noisy features leads to model regularization in RF models. However their analysis was very generic. We aim at taking a step further by deriving finite-sample bounds to show that the regularization phenomenon occurs precisely for RF models. The regularization strength depends on the hyperparameter `max-features`, which could in turn provide guidance about how to choose/tune this hyperparameter.

Second axis - Variable importance Variable importances are often used to assess the role of an input variable on predicting the output. Two variable importances exist for RF: the Mean Decrease in Impurity (MDI) and the Mean Decrease in Accuracy (MDA). Recently, it has been shown that none of these measures are relevant to perform variable selections [see, e.g., 2, 6]. Other measures like Sobol-MDA [2, 1] have been proposed and their consistency has been established. In practice, one could be interested in testing whether a variable importance is null, which requires developping tailored tests. One can draw inspiration from the work of [7]. Besides, most data sets contain missing data which may modify the variable importance measure. Analyzing how these measures are impacted by different missingness types would be very useful for practitionners.

The choice between the axes will be discussed with the successful candidate. For each axis, the subject combines two aspects of a scientific work: on the one hand, a more methodological development could lead to efficient algorithms; on the other hand, a more thorough theoretical study of this issue will allow establishing nice statistical results. Both aspects are important, and can be modulated according to the candidate's affinities.

Supervisors : Claire Boyer (LMO, Université Paris-Saclay), Rafaël Pinot (LPSM, Sorbonne Université), Erwan Scornet (LPSM, Sorbonne Université)

Required skills: M1 or M2 level trainee in statistics/machine learning. Applicants should send CV, transcripts of the last two years and the name of a referee to

- claire.boyer@sorbonne-universite.fr
- rafael.pinot@sorbonne-universite.fr
- erwan.scornet@polytechnique.edu

Practical information: the internship will take place at LPSM (Sorbonne Université) in the statistical team. This is a 6-month internship that can start at the beginning of April. This internship can be followed by a PhD thesis.

References

[1] Clément Bénard, Gérard Biau, Sébastien Da Veiga, and Erwan Scornet. Shaff: Fast and consistent shapley effect estimates via random forests. In *International Conference on Artificial Intelligence and Statistics*, pages 5563–5582. PMLR, 2022.

[2] Clément Bénard, Sébastien Da Veiga, and Erwan Scornet. Mean decrease accuracy for random forests: inconsistency, and a practical solution via the sobol-mda. *Biometrika*, 109(4):881–900, 2022.

[3] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[4] Alicia Curth, Alan Jeffares, and Mihaela van der Schaar. Why do random forests work? understanding tree ensembles as self-regularizing adaptive smoothers. *arXiv preprint arXiv:2402.01502*, 2024.

[5] Lucas Mentch and Siyu Zhou. Randomization as regularization: A degrees of freedom explanation for random forest success. *Journal of Machine Learning Research*, 21(171):1–36, 2020.

[6] Erwan Scornet. Trees, forests, and impurity-based variable importance in regression. In *Annales de l'Institut Henri Poincare (B) Probabilites et statistiques*, volume 59, pages 21–52. Institut Henri Poincaré, 2023.

[7] Brian D Williamson, Peter B Gilbert, Noah R Simon, and Marco Carone. A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association*, 118(543):1645–1658, 2023.