Random Planted Forest

Munir Eberhardt Hiabu[*1], Joseph Meyer[†2], and Enno Mammen [‡2]

[1]Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5,2100 Copenhagen Ø, Denmark.
[2]Institute for Applied Mathematics, Heidelberg University, INF 205, 69120 Heidelberg, Germany.

Abstract

The goal of Planted Machine Learning is to discover and learn low dimensional structures that can get more complex along a planted path. I will introduce Random Planted Forest – an algorithm we have developed that seems to be competitive with state-of-the-art machine learning predictors with respect to accuracy while having favourable interpretability properties.

Random Planted Forest aims to estimate the unknown regression function from a functional decomposition perspective in which the functional components correspond to lower order interaction terms. More precisely, given covariates $X \in \mathbb{R}^p$, we assume that the conditional mean of a response Y can be decomposed into a sum of components with order smaller or equal r:

$$\mathbb{E}[Y|X = x] = m(x) = \sum_{S\subseteq\{1,\dots,p\},|S|\leq r} m_S(x_S).$$

Random Planted Forest can be seen as a modifiaction of the Random Forest algorithm whereby certain leaves are kept after they are split instead of deleting them. This leads to non-binary trees which we refer to as planted trees. An extension to a forest leads to our Random Planted Forest algorithm. The maximum number of covariates which can interact within a leaf, r, can be bounded. If we set $r = 1$, the resulting estimator is a sum of one-dimensional functions. In the other extreme case, if we do not set a limit ($r = p$), the resulting estimator and corresponding model place no restrictions on the form of the regression function. In a simulation study we find encouraging prediction and visualisation properties of our Random Planted Forest method.

We also develop theory for an idealized version of Random Planted Forests in cases where the interaction bound is low. We show that if it is smaller than three, the idealized version achieves asymptotically optimal convergence rates up to a logarithmic factor.

Lastly, we will discuss a so-called marginal identification of the functional decomposition leading, under suitable causal assumption, to an interpretation of the component m_S as the *average natural direct effect* of the features in S on Y. We will apply Random Planted Forest on insurance claim data and fit claim frequencies in an interpretable way while having high accuracy.

Keywords: Random Forest, Interpretable Machine Learning, Functional Decomposition.

[*]E-mail address: mh@ku.dk
[†]E-mail address: Joseph-Theo.Meyer@uni-heidelberg.de
[‡]E-mail address: mammen@math.uni-heidelberg.de

References

[1] Munir E. Hiabu, Enno Mammen, Joseph T. Meyer (2023), "Random planted forest: a directly interpretable tree ensemble." *Preprint.* https://arxiv.org/abs/2012.14563

[2] Munir E. Hiabu, Marvin N. Wright, Joseph T. Meyer (2023), "Unifying local and global model explanations by functional decomposition of low dimensional structures." *AISTATS, PMLR* vol. **206**, pp. 7040–7060.