Proceedings of Machine Learning Research vol 125:1–2, 2020 33rd Annual Conference on Learning Theory

Estimation and Inference with Trees and Forests in High Dimensions

Vasilis Syrgkanis VASY@MICROSOFT.COM
Microsoft Research

Manolis Zampetakis MZAMPET@MIT.EDU
Massachusetts Institute of Technology

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

[1] Regression Trees Breiman et al. (1984) and Random Forests Breiman (2001), are one of the most widely used estimation methods by machine learning practitioners. Despite their widespread use, their theoretical underpinnings are far from being fully understood. Recent breakthrough advances have shown that such greedily built trees are asymptotically consistent Biau (2010); Denil et al. (2014); Scornet et al. (2015) in the low dimensional regime, where the number of features is a constant, independent of the sample size. Also, the works of Mentch and Hooker (2016); Wager and Athey (2018) provide asymptotic normality results for honest versions of Random Forests.

In this work, we analyze the performance of regression trees and forests with binary features in the high-dimensional regime, where the number of features can grow exponentially with the number of samples. We show that trees and forests built greedily based on the celebrated CART criterion, provably adapt to sparsity: when only a subset R, of size r, of the features are relevant, then the mean squared error of appropriately shallow trees, or fully grown honest forests, scales exponentially only with the number of relevant features and depends only logarithmically on the overall number of features.

More precisely, we identify three regimes, each providing different dependence on the number of relevant features. When the relevant variables are "weakly" relevant (in the sense that there is not strong separation between the relevant and irrelevant variables in terms of their ability to reduce variance), then shallow trees achieve "slow rates" on the mean squared error of the order of $2^r/\sqrt{n}$, when variables are independent, and $1/n^{1/(r+2)}$, when variables are dependent. When the relevant variables are "strongly" relevant, in that there is a separation in their ability to reduce variance as compared to the irrelevant ones, by a constant $\beta_{\min}$, then we show that greedily built shallow trees and fully grown honest forests can achieve fast parametric mean squared error rates of the order of $2^r/(\beta_{\min} n)$. When variables are strongly relevant, we also show that the predictions of sub-sampled honest forests have an asymptotically normal distribution centered around their true values and whose variance scales at most as $O(2^r \log(n)/(\beta_{\min} n))$.

Thus, sub-sampled honest forests are provably a data-adaptive method for non-parametric inference, that adapts to the latent sparsity dimension of the data generating distribution, as opposed to classical non-parametric regression approaches. Our results show that, at least for the case of binary features, forest based algorithms can offer immense improvement on the statistical power of non-parametric hypothesis tests in high-dimensional regimes.

Keywords: Regression trees, random forests, high dimensions, sparsity, finite sample, mean squared error, asymptotic normality

1. Extended abstract. Full version appears in arxiv under the same title.

Acknowledgments

This work was done while M.Z. was an intern at MSR, New England. M.Z. is supported by a Google Ph.D. Fellowship award.

References

Grard Biau. Analysis of a random forests model. *Journal of Machine Learning Research*, 13, 05 2010.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone. Classification and regression trees. wadsworth int. *Group*, 37(15):237–251, 1984.

Misha Denil, David Matheson, and Nando de Freitas. Narrowing the gap: Random forests in theory and in practice. In *International Conference on Machine Learning (ICML)*, 2014. URL http://jmlr.org/proceedings/papers/v32/denil14.pdf.

Lucas Mentch and Giles Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research*, 17(26):1–41, 2016. URL http://jmlr.org/papers/v17/14-168.html.

Erwan Scornet, Grard Biau, and Jean-Philippe Vert. Consistency of random forests. *Ann. Statist.*, 43 (4):1716–1741, 08 2015. doi: 10.1214/15-AOS1321. URL https://doi.org/10.1214/15-AOS1321.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.