Preventing spurious interactions in tree-based metalearners

Roger Pros[1,2] **Jordi Vitrià**[1]

[1]Departament de Matemàtica i Informàtica, , Universitat de Barcelona, 08007 Barcelona, Spain
[2]Zenital

Abstract

In recent years, various insights have been employed to enhance causal machine learning methods by refining estimation techniques and introducing robust algorithms that account for causal structures and dependencies within the data. Building on this trend, we propose a novel method to improve the estimation of the Conditional Average Treatment Effect (CATE). A common approach in CATE estimation involves the use of metalearners, which can estimate CATE if certain identification properties are met. However, this approach employs causal knowledge only for identifying the estimand, not for the estimation process itself.
We present a new method that utilizes causal knowledge in the estimation phase by imposing variable interaction constraints during model training. These constraints are based on total or partial knowledge about the underlying data-generating process. By applying these constraints to traditional tree-based estimation algorithms, we show that models trained in this manner achieve improved performance and reduced variability in estimating CATE.

1 INTRODUCTION

Machine learning techniques are highly effective at identifying associations in i.i.d. data. However, these models often face challenges when applied to real-world scenarios where this basic assumption does not hold. It has been argued that these problems arise partly from the model's lack of causal knowledge about the data Peters et al. [2017], Goyal and Bengio [2022], Pearl [2019], Ahmed et al. [2020], Schölkopf et al. [2021]. To address this issue, there has been growing interest in Causal Machine Learning Kaddour et al. [2022]. These methods enhance traditional machine learning by incorporating causal knowledge that is typically overlooked by standard machine learning techniques.

One of the most prevalent approaches is the use of metalearners Künzel et al. [2019]. Metalearners employ various supervised learning methods to estimate the Conditional Average Treatment Effect (CATE) function. These algorithms use predictive models as base learners for CATE estimation, with each base learner bringing its own inductive biases. This context inspires our approach: proposing the use of tree-based models constrained by causal knowledge for CATE estimation.

1.1 THE NEED OF REGULARIZATION

Machine learning models are designed for predicting outcomes from observed data, but individualized causal effects cannot be directly estimated due to the fundamental problem of causal inference. Instead, causal effects must be inferred indirectly, often by computing potential outcomes. Estimating an identified distribution like $P(Y|T = 1, X) - P(Y|T = 0, X)$ can be challenging with finite data, especially if the solutions lack stability. To address this, regularization is necessary to constrain the model space and ensure more reliable solutions.

In this work, we identify and address a frequently overlooked bias in the literature: the inclusion of spurious interactions between covariates within models. Typically, we have some understanding of the data generation process for the given scenario. This knowledge limits the set of variable interactions to those consistent with the causal structure of the data thereby helping to reduce model bias.

2 METHODOLOGY

According to the **Independent Causal Mechanisms (ICM)** principle Schölkopf et al. [2021], the generative model of the variables of the causal model consists of independent modules that operate without sharing information or exerting influence on one another. These mechanisms are reflected in the factorization of the joint distribution of the data generating process which can be represented as a causal graph.

$$p(x) = p(x_1, \ldots, x_k) = \prod_{j=1}^{k} p(x_j \mid Pa(x_j)),$$

where $Pa(x_j)$ are the parents of node x_j in the causal graph. The factorization of the data generation process might not be fully known, but there might be some expert knowledge about which variables interact in specific causal mechanisms and which do not.

An interaction occurs when the effect of one variable on the outcome depends on another variable. We define a **spurious interaction** as any variable interaction that does not form part of an ICM and is therefore not included in the data-generating process. Spurious interactions are not robust to interventions, and reliance on them by models impairs the performance of causal effect estimation.

Most classical ML learners model the distribution of the target variable as $P(Y|X)$. Specifically, tree-based models use a combination of weak learners, each modeling the target variable conditioned on a random subset of the predictive features G_i:

$$P(Y|X) \sim f(P(Y|G_1), .., P(Y|G_m)).$$

By contrast, we restrict the learning of the model to a function of the distributions of $P(Y|G_{x_i})$, where G_{x_i} is the set of variables that appear in the ICM of the data generating process that corresponds to x_i:

$$P(Y|X) \sim f(P(Y|G_{x_i}), .., P(Y|G_{x_m}))$$

For a set G_{x_i}, each variable is permitted to interact only with other variables within the same set, thereby avoiding any spurious interactions.

2.1 IMPLEMENTATION

We implement the proposed method in SLearner and TLearner metalearners with gradient boosting as base learners (Chen and Guestrin [2016]). Among the hyperparameters of these methods, there is the possibility of passing feature interaction constraints as a list of sets of variables. For each set, each variable is only allowed to interact with variables in the same set. An example of implementation of the method can be found here.

3 RESULTS

Illustrative example In order to show the properties of the method, we generated a synthetic data with a graph topology that is well suited to exploit the advantages of the method. This setting is composed by three clearly independent sources of information and a complex structure. By adding the constraints we prevent the model from using any spurious correlation that might arise between an interaction of features belonging to different ICM. The constrained version of the learners obtained an average PEHE 40% with a standard deviation 7% lower than their unconstrained counterparts. In figure 1 we can see that the constrained model learns an independent distribution for each ICM in the data generating process.

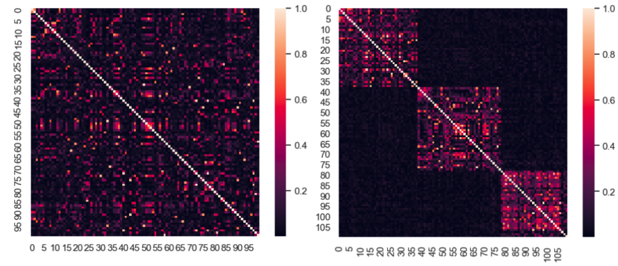

Figure 1: Heatmaps of the correlations of the individual prediction trees for each gradient boosting algorithm. On the left is the unconstrained version, and on the right is the version constrained by our method, rearranged according to the constraint group.

Benchmark results We run experiments in two different settings. The first setting is the synthetic datasets used in Nie and Wager [2021] to test model performance under different scenarios and variable relationships. In this setting, the causal graph is known. The second setting is the causal benchmark suggested in Neal et al. [2020]. The data sets in this setting are semi-synthetic, and the causal graph is unknown and discovery algorithms are used to approximate it. In the first setting, constrained models have a higher performance, with both lower error and lower error variability, than the unconstrained models in all scenarios. In the second one, despite the possible errors in the causal graph discovery, constrained models still outperform the unconstrained models.

4 CONCLUSION

We presented a novel approach to introduce model constraints into tree-based predictive models by exploiting information from the causal structure of the dataset. We then used these constrained models for CATE estimation and compared them with their unconstrained counterparts. We showed that the constrained models outperform the unconstrained ones with higher accuracy and lower variability. The results open new research opportunities related to exploiting the information encoded in the data generating process.

ACKNOWLEDGEMENTS

This research was funded by an Industrial Doctorate grant 2021 DI 41 of AGAUR (Generalitat de Catalunya) between Universitat de Barcelona and Zenital and PID2022-136436-NB-100, 2021SGR01104 grants.

References

Ossama Ahmed, Frederik Träuble, Anirudh Goyal, Alexander Neitz, Yoshua Bengio, Bernhard Schölkopf, Manuel Wüthrich, and Stefan Bauer. Causalworld: A robotic manipulation benchmark for causal structure and transfer learning. *arXiv preprint arXiv:2010.04296*, 2020.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL `http://doi.acm.org/10.1145/2939672.2939785`.

Anirudh Goyal and Yoshua Bengio. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, 478(2266):20210068, 2022.

Jean Kaddour, Aengus Lynch, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal machine learning: A survey and open problems. *arXiv preprint arXiv:2206.15475*, 2022.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

Brady Neal, Chin-Wei Huang, and Sunand Raghupathi. Realcause: Realistic causal inference benchmarking. *arXiv preprint arXiv:2011.15007*, 2020.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2): 299–319, 2021.

Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60, 2019.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.