The Journal of Open Source Software

Pycausal-Explorer: A Scikit-learn compatible causal inference toolkit

Guilherme Goto Escudero [1*¶], **Heitor de Moraes Santos** [2*], **João Vitor Tigre Almeida** [2], **and Roseli de Deus Lopes** [1]

1 Escola Politécnica - Universidade de São Paulo, Brazil **2** Independent Researcher, Brazil ¶ Corresponding author * These authors contributed equally.

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor:

Submitted: 29 January 2024
Published: unpublished

Summary

Pycausal-Explorer is an open source, scikit-learn compatible Python library which leverages causal inference and machine learning for causal reasoning and exploration. It consists of a number of algorithms built to calculate treatment effects and methods concerning causal analysis. It has the foundation to allow researchers to easily study and compare different causal inference algorithms. Pycausal-Explorer is a modular package with a common scikit-learn-like interface, reducing the gap from research to production environments.

Statement of need

Causal inference has been an ascending research topic over the last years. Some recent work conjugate machine learning and causal inference in order to estimate treatment effect from observational data, some examples being the use of meta-learners (Künzel et al., 2019) and honest causal trees (Wager & Athey, 2018).

In such a landscape, `Pycausal-Explorer` is a Python package that helps create, compare and publish methods which estimate treatment effects from observational data. The package has two main objectives: (1) leverage the study of causal inference through a common framework, and (2) close the gap between research of the field and real-world applications. The library is an open source project built in Python, providing algorithms, methods, and data sets commonly found in the causal inference literature. `Pycausal-Explorer` is compatible with scikit-learn (Pedregosa et al., 2011), which expands such a well-known library for building and monitoring causal models. It has been extensively tested and has documented guidelines so that contributions can be made to the library.

The library implements a Python class called BaseCausalModel from which all models should inherit. It is an abstract class inherited from scikit-learn's BaseEstimator. All models must implement the following methods: fit, predict, estimate_ate and estimate_ite. Whenever a model is not capable of measuring individual treatment effect, individual treatment effects will be defined as the average treatment effect, regardless of the individual's particularities. The models currently implemented include linear regression and inverse propensity score weighting (IPTW) for average treatment effect and k-nearest neighbours, S-learner, T-Learner, X-Learner, RA-Learner,
DR-Learner, DoubleML and a novel method called randomized trees ensemble embedding to calculate heterogeneous treatment effect.

Though `Pycausal-Explorer` is unique in that it enables causal model exploration in the form of a scikit-learn compatible Python tool, there are other prominent toolkits aimed at causal inference which also bring a lot of value to research and business applications.

Causal explorer (Aliferis et al., 2003) is a MATLAB package with causal inference tools meant for biomedical applications, and it is perhaps one of the first published libraries aimed at working with causality. It provides tools to model causal relations between variables as causal probabilistic networks (which are bayesian networks) and a handful of causal discovery algorithms which help choose the best causal graph assumption.

Generalized random forests (Athey et al., 2019) is an open source R package which implements causal models based on tree ensembles, supporting the so-called honest estimation of causal effects. Although it is a very popular language in academics, specially in the econometrics field, Python tends to be more used than R in business applications and is easier to integrate with APIs and external software, so there might be demand for a Python implementation of generalized random forests.

Furthermore, there are also examples of business-orientated Python libraries for causal modeling. CausalML (Chen et al., 2020) is a library made by Uber which is focused on uplift models (i.e., estimating effects of interventions in scenarios such as A/B tests). EconML (Battocchi et al., 2019) is a library made by Microsoft which leverages machine learning tools to estimate treatment effects, particularly for applications in Economics. Another library by Microsoft called DoWhy (Sharma et al., 2019) implements an end-to-end causal analysis tool, abstracting most of the data process work. This last library focuses on identification of causal effect (i.e., inspecting causal models' assumptions) and allows for integration with algorithms implemented by both CausalML and EconML.

References

Aliferis, C. F., Tsamardinos, I., Statnikov, A. R., & Brown, L. E. (2003). Causal explorer: A causal probabilistic network learning toolkit for biomedical discovery. *METMBS*, *3*, 371–376.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148–1178.

Battocchi, K., Dillon, E., Hei, M., Lewis, G., Oka, P., Oprescu, M., & Syrgkanis, V. (2019). *EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation*. https://github.com/microsoft/EconML.

Chen, H., Harinen, T., Lee, J.-Y., Yung, M., & Zhao, Z. (2020). *CausalML: Python package for causal machine learning*. https://arxiv.org/abs/2002.11631

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4156–4165.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Sharma, A., Kiciman, E., & others. (2019). *DoWhy: A Python package for causal inference*. https://github.com/microsoft/dowhy.

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