causal-machine-learning / **kdd2021-tutorial**

generated from fastai/fastpages

EconML/CausalML KDD 2021 Tutorial

causal-machine-learning.github.io/kdd2021-tutorial/

Apache-2.0 License

38 stars 6 forks

Causal Inference and Machine Learning in Practice with EconML and CausalML: Industrial Use Cases at Microsoft, TripAdvisor, Uber

Schedule

Time

- 4:00 AM - 7:00 AM August 15, 2021 SGT
- 4:00 PM - 7:00 PM August 14, 2021 EDT

- 1:00 PM - 4:00 PM August 14, 2021 PDT

Live Zoom Link

To be shared within the KDD 21 Virtual Platform during the conference.

Abstract

In recent years, both academic research and industry applications see an increased effort in using machine learning methods to measure granular causal effects and design optimal policies based on these causal estimates. Open source packages such as CausalML and EconML provide a unified interface for applied researchers and industry practitioners with a variety of machine learning methods for causal inference. The tutorial will cover the topics including conditional treatment effect estimators by meta-learners and tree-based algorithms, model validations and sensitivity analysis, optimization algorithms including policy leaner and cost optimization. In addition, the tutorial will demonstrate the production of these algorithms in industry use cases.

Target Audience and Prerequisites for the Tutorial

Anyone who is interested in causal inference and machine learning, especially economists/statisticians/data scientists who want to learn how to combine causal inference and machine learning with real industry use cases incorporated in large scaled machine learning systems at companies such as Microsoft, TripAdvisor and Uber. The tutorial assumes some basic knowledge in statistical methods, machine learning algorithms and the Python programming language.

Outline

Title	Duration	Slides	Code
Introduction to Causal Inference	20 minutes	Slides	
Case Studies Part 1 by CausalML			
Introduction to CausalML	15 minutes	Slides	
Case Study #1: Causal Impact Analysis with Observational Data: CeViChE at Uber	30 minutes	Slides	Notebook

Title	Duration	Slides	Code
Case Study #2: Targeting Optimization: Bidder at Uber	30 minutes	Slides	Notebook
Case Studies Part 2 by EconML			
Introduction to EconML	15 minutes	Slides	Notebook
Case Study #3: Customer Segmentation at TripAdvisor with Recommendation A/B Tests	30 minutes	Slides	Notebook
Case Study #4: Long-Term Return-on-Investment at Microsoft via Short-Term Proxies	30 minutes	Slides	Notebook

Presentation Abstracts

Introduction to Causal Inference

We will give an overview of basic concepts in causal inference. A quick refresher on the main tools and terminology of causal inference: correlation vs causation, average, conditional, and individual treatment effects, causal inference via randomization, Causal inference using instrumental variables, Causal inference via unconfoundedness.

Introduction to CasualML

We will provide an overview of CausalML, an open source Python package that provides a suite of uplift modeling and causal inference methods using machine learning algorithms based on recent research. We will introduce the main components of CausalML: (1) inference with causal machine learning algorithms (e.g. meta-learners, uplift trees, CEVAE, dragonnet), (2) validation/analysis methods (e.g. synthetic data generation, AUUC, sensitivity analysis, interpretability), (3) optimization methods (e.g. policy optimization, value optimization, unit selection).

Case #1: Causal Impact Analysis with Observational Data at Uber

As an introductory case study for using causal inference, we will cover the use case of understanding the causal impact from observational data in the context of cross sell at Uber. We emphasize that simple comparisons of users who make cross purchase or not will produce biased estimates and that can be demonstrated in the causal inference framework. We show the use of different causal estimation methodologies through propensity score matching and meta learners to estimate the causal impact. In addition, we will use sensitivity analysis to show the robustness of the estimates.

Case #2: Targeting Optimization: Bidder at Uber

We will introduce the audience selection method with uplift modeling in online RTB, which aims to estimate heterogeneous treatment effects for advertising. It has been studied to provide a superior return on investment by selecting the most incremental users for a specific campaign. To examine the effectiveness of uplift modeling in the context of real-time bidding, we conducted the comparative analysis of four different meta-learners on real campaign data. We adapted an explore-exploit set up for offline training and online evaluation. We will also introduce how we use Targeted Maximum Likelihood Estimation (TMLE) based Average Treatment Effect (ATE) as ground truth for evaluation.

Introduction to EconML

We will provide an overview of recent methodologies that combine machine learning with causal inference and the significant statistical power that machine learning brings to causal inference estimation methods. We will outline the structure and capabilities of the EconML package and describe some of the key causal machine learning methodologies that are implemented (e.g. double machine learning, causal forests, deepiv, doubly robust learning, dynamic double machine learning). We will also outline approaches to confidence interval construction (e.g. bootstrap, bootstrap-of-little-bags, debiased lasso), interpretability (shap values, tree interpreters) and policy learning (doubly robust policy learning).

Case #3: Customer Segmentation at TripAdvisor with Recommendation A/B Tests

We examine the scenario in which we wish to learn heterogeneous treatment effects (CATE), but observational data is biased and direct experimental data (e.g. A/B test) is plagued by imperfect compliance. In this setup, TripAdvisor would like to know whether joining a membership program compels users to spend more time engaging with the website and purchasing more products. The usual approach, a direct A/B test, is infeasible: the website cannot force users to comply and become members, hence the imperfect compliance that can bias calculations. The solution is to use an alternative A/B test that was originally designed to measure whether an easier sign-up process would promote user membership. This A/B test plays the role of an instrument that nudges users to sign up for membership. We introduce EconML's IntentToTreatDRIV estimator which can leverage this repurposed A/B test to both learn the effect of membership on user engagement and understand how these effects vary with customer features. We show how this novel methodology led to extracting key business insights and helped TripAdvisor understand and differentiate how customers engage with their platform.

Case #4: Long-Term Return-on-Investment at Microsoft via Short-Term Proxies

In this case study, we talk about using observational data to measure the long term Return-on-Investment of some types of dollar value investments Microsoft gives to the enterprise customers. There are many challenges for this setting, for instance, we don't have enough period of data to identify a long term ROI, we should control the effect coming from the future investment and we are in a high dimensional data space. We then propose a surrogate based approach assuming the long-term effect is channeled through some short-term proxies and employ a dynamic adjustment to the surrogate model in order to get rid of the effect from future investment, finally apply double machine learning (DML) techniques to estimate the ROI. We apply this methodology to answer the questions like what is the average long-run ROI on each type of the investment? What types of customers have a higher ROI to a specific investment? And how different incentives impact the different solution areas. Finally we will showcase how you could use EconML to solve similar problems by only a few lines of code.

Tutors

Presenters

- Jing Pan, Uber, CausalML
- Yifeng Wu, Uber, CausalML
- Huigang Chen, Facebook, CausalML
- Totte Harinen, Toyota Research Institute, CausalML

- Paul Lo, Uber, CausalML
- Greg Lewis, Microsoft Research, EconML
- Vasilis Syrgkanis, Microsoft Research, EconML
- Miruna Oprescu, Microsoft Research, EconML
- Maggie Hei, Microsoft Research, EconML

Contributors

- Jeong-Yoon Lee, Netflix Research, CausalML
- Zhenyu Zhao, Tencent, CausalML
- Keith Battocchi, Microsoft Research, EconML
- Eleanor Dillon, Microsoft Research, EconML

References

1. Künzel, Sören R., et al. "Metalearners for estimating heterogeneous treatment effects using machine learning." Proceedings of the national academy of sciences 116.10 (2019): 4156-4165. (paper)
2. Chernozhukov, Victor, et al. "Double/debiased/neyman machine learning of treatment effects." American Economic Review 107.5 (2017): 261-65. (paper)
3. Nie, Xinkun, and Stefan Wager. "Quasi-oracle estimation of heterogeneous treatment effects." arXiv preprint arXiv:1712.04912 (2017) (paper)
4. Tso, Fung Po, et al. "DragonNet: a robust mobile internet service system for long-distance trains." IEEE transactions on mobile computing 12.11 (2013): 2206-2218. (paper)
5. Louizos, Christos, et al. "Causal effect inference with deep latent-variable models." arXiv preprint arXiv:1705.08821 (2017) (paper)
6. Wager, Stefan, and Susan Athey. "Estimation and inference of heterogeneous treatment effects using random forests." Journal of the American Statistical Association 113.523 (2018): 1228-1242. (paper)
7. Oprescu, Miruna, et al. "EconML: A Machine Learning Library for Estimating Heterogeneous Treatment Effects." (repo)
8. Chen, Huigang, et al. "Causalml: Python package for causal machine learning." arXiv preprint arXiv:2002.11631 (2020) (repo)
9. Yao, Liuyi, et al. "A survey on causal inference." arXiv preprint arXiv:2002.02770 (2020). (paper)
10. Goldenberg, Dmitri, et al. "Personalization in Practice: Methods and Applications." Proceedings of the 14th ACM International Conference on Web Search and Data

Mining. 2021 (paper)

11. Blackwell, Matthew. "A selection bias approach to sensitivity analysis for causal effects." Political Analysis 22.2 (2014): 169-182. (paper)
12. Athey, Susan, and Stefan Wager. "Efficient policy learning." arXiv preprint arXiv:1702.02896 (2017). (paper)
13. Sharma, Amit, and Emre Kiciman. "Causal Inference and Counterfactual Reasoning." Proceedings of the 7th ACM IKDD CoDS and 25th COMAD. 2020. 369-370. (paper)
14. Li, Ang, and Judea Pearl. "Unit selection based on counterfactual logic." Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. 2019 (paper)
15. Kennedy, Edward H. "Optimal doubly robust estimation of heterogeneous causal effects." arXiv preprint arXiv:2004.14497 (2020) (paper)
16. Gruber, Susan, and Mark J. Van Der Laan. "Targeted maximum likelihood estimation: A gentle introduction." (2009) (paper)
17. D. Foster, V. Syrgkanis. Orthogonal Statistical Learning. Proceedings of the 32nd Annual Conference on Learning Theory (COLT), 2019 (paper)
18. V. Syrgkanis, V. Lei, M. Oprescu, M. Hei, K. Battocchi, G. Lewis. Machine Learning Estimation of Heterogeneous Treatment Effects with Instruments. Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS), 2019 (paper)
19. M. Oprescu, V. Syrgkanis and Z. S. Wu. Orthogonal Random Forest for Causal Inference. Proceedings of the 36th International Conference on Machine Learning (ICML), 2019 (paper)
20. Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep IV: A flexible approach for counterfactual prediction. Proceedings of the 34th International Conference on Machine Learning, ICML'17, 2017 (paper)
21. Battocchi, K., Dillon, E., Hei, M., Lewis, G., Oprescu, M., & Syrgkanis, V. (2021). Estimating the Long-Term Effects of Novel Treatments. arXiv preprint arXiv:2103.08390. (paper)
22. Lewis, G., & Syrgkanis, V. (2020). Double/Debiased Machine Learning for Dynamic Treatment Effects. arXiv preprint arXiv:2002.07285. (paper)

Releases

No releases published

Packages

No packages published

No packages published

Contributors 3

- jeongyoonlee Jeong-Yoon Lee
- vincewu51 Yifeng Wu
- paullo0106 Paul Lo

Environments 1

- github-pages Active

Languages

● Jupyter Notebook 92.9% ● HTML 3.5% ● JavaScript 1.1% ● SCSS 1.0% ● Shell 0.6%
● Python 0.3% ● Other 0.6%