Counterfactual Learning and Evaluation for Recommender Systems: Foundations, Implementations, and Recent Advances

Yuta Saito
Cornell University
Ithaca, NY, USA
ys552@cornell.edu

Thorsten Joachims
Cornell University
Ithaca, NY, USA
tj@cs.cornell.edu

ABSTRACT

Counterfactual estimators enable the use of existing log data to estimate how some new target recommendation policy would have performed, if it had been used instead of the policy that logged the data. We say that those estimators work "off-policy", since the policy that logged the data is different from the target policy. In this way, counterfactual estimators enable *Off-policy Evaluation* (OPE) akin to an unbiased offline A/B test, as well as learning new recommendation policies through *Off-policy Learning* (OPL). The goal of this tutorial is to summarize *Foundations, Implementations, and Recent Advances* of OPE/OPL. Specifically, we will introduce the fundamentals of OPE/OPL and provide theoretical and empirical comparisons of conventional methods. Then, we will cover emerging practical challenges such as how to take into account combinatorial actions, distributional shift, fairness of exposure, and two-sided market structures. We will then present *Open Bandit Pipeline*, an open-source package for OPE/OPL, and how it can be used for both research and practical purposes. We will conclude the tutorial by presenting real-world case studies and future directions.

CCS CONCEPTS

• **Computing methodologies → Batch learning**; *Learning from implicit feedback*; **Learning to rank**; **Ranking**.

KEYWORDS

counterfactuals, off-policy evaluation/learning, recommender systems, fairness of exposure

ACM Reference Format:
Yuta Saito and Thorsten Joachims. 2021. Counterfactual Learning and Evaluation for Recommender Systems: Foundations, Implementations, and Recent Advances . In *Fifthteenth ACM Conference on Recommender Systems (RecSys '21), September 27-October 1, 2021, Amsterdam, Netherlands.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3460231.3473320

1 INTRODUCTION: MOTIVATION AND TARGETED AUDIENCE

Interactive decision-making systems such as ad/recommendation/search platforms produce log data valuable for evaluating and redesigning the system. For example, the logs of a news recommendation system record which news article was presented and whether the user read it, giving the system designer a chance to redesign its recommendations to be more relevant. Exploiting log bandit data is, however, more difficult than conventional supervised machine learning, since the result is only observed for the action chosen by the system, but not for all the other actions that the system could have taken. The logs are also biased in that they over-represent the actions favored by the system. A potential solution to this problem is an A/B test that compares the performance of competing systems in an online environment. However, A/B testing systems is often difficult because deploying a new policy is time- and money-consuming and entails the risk of failure. This motivates the problem of OPE/OPL, which aims to estimate the performance of a new policy or to train it using only the log data collected by a past policy.

Because of their practical relevance, there has been a growing amount of theoretical and methodological research in OPE/OPL. However, it is not always straightforward to apply these methods to real-world applications, since there can be a number of challenges that arise in practice, such as combinatorial/continuous actions, distributional shift, and fairness of exposure requirements. This tutorial is aimed at bridging the gap between theory and practice in OPE/OPL. Specifically, we will introduce the fundamentals of OPE/OPL and compare conventional methods from both theoretical and empirical perspectives. Then, we will cover recent advances in the field to handle the emerging practical challenges. We will then present *Open Bandit Pipeline*[1] [15], an open-source package and how it helps us implement OPE/OPL for research and practical purposes. We will also present real-world case studies and future directions.

It has been five years since the related tutorial "Counterfactual Evaluation and Learning for Search, Recommendation and Ad Placement" by Thorsten Joachims and Adith Swaminathan took place at SIGIR2016 [6]. It is an excellent time to aggregate and unify the essential recent works into one coherent tutorial that is particularly valuable to the RecSys community.

The learning outcomes of this tutorial are to enable the participants (such as applied researchers, practitioners, and students):

(1) to know fundamental concepts and conventional methods of OPE/OPL
(2) to be familiar with recent advances to address practical challenges such as fairness of exposure
(3) to understand how to implement OPE/OPL in their research and applications
(4) to be aware of remaining challenges and opportunities in the area

[1]https://github.com/st-tech/zr-obp

RecSys '21, September 27-October 1, 2021, Amsterdam, Netherlands

ACM ISBN 978-1-4503-8458-2/21/09…$15.00
https://doi.org/10.1145/3460231.3473320

This tutorial is aimed at an audience with intermediate experience in machine learning, information retrieval, or recommender systems who are interested in using OPE/OPL methods in their research and applications. Participants are expected to have basic knowledge of machine learning, probability theory, and statistics. The tutorial will provide practical examples based on Python code and Jupyter Notebooks.

2 OUTLINE OF THE TUTORIAL

This tutorial consists of the following contents.

(1) **Introduction**: We will introduce conventional formulation and methods of OPE/OPL [3–5, 14, 19–21, 24]. Moreover, we will provide comprehensive comparisons of a variety of methods from both theoretical and empirical perspectives.
(2) **Recent Topic 1**: We will cover recent works on OPE/OPL methods to handle emerging practical challenges such as combinatorial actions [10, 12], continuous actions [2, 8], deficient support [13], multiple loggers [1, 7], and distributional shifts [9, 11, 16]. These challenges are closely related to real-world applications in recommender and e-commerce systems.
(3) **Recent Topic 2**: We will cover OPE/OPL with alternative and interdependent objectives (e.g., fairness, diversity, etc.) in multi-sided markets [17, 18, 22, 23, 25].
(4) **Implementations and Case-Studies**: We will introduce how to use *Open Bandit Pipeline* to implement OPE/OPL in research and applications [15]. We will also present some real-world case-studies to describe how to implement OPE/OPL in practice.
(5) **Conclusions**: We will conclude the tutorial by summarizing the previous sections and presenting remaining research challenges of the area.

All materials, including slides and code, will be available during and after the tutorial.

3 PRESENTER BIO

Yuta Saito (ys552@cornell.edu) is a Ph.D. student in the Department of Computer Science at Cornell University, advised by Prof. Thorsten Joachims. He received a B.Eng degree in Industrial Engineering and Economics from Tokyo Institute of Technology in 2021. His current research focuses on OPE of bandit algorithms and learning from human behavior data. He has been collaborating with several tech companies to implement OPE/OPL in practical situations and to conduct large-scale empirical studies. Some of his recent work has been published at top conferences, including ICML, SIGIR, SDM, ICTIR, RecSys, and WSDM.

Thorsten Joachims (tj@cs.cornell.edu) is a Professor in the Department of Computer Science and in the Department of Information Science at Cornell University, and he is an Amazon Scholar. His research interests center on the synthesis of theory and system building in machine learning, with applications in information retrieval and recommendation. His past research focused on support vector machines, learning to rank, learning with preferences, and learning from implicit feedback, text classification, and structured output prediction. Working with his students and collaborators, his papers won 9 Best Paper Awards and 4 Test-of-Time Awards. He is also an ACM Fellow, AAAI Fellow, KDD Innovations Award recipient, and member of the SIGIR Academy.

ACKNOWLEDGMENTS

This research was supported in part by NSF Awards IIS-1901168 and IIS-2008139. Yuta Saito is supported by Funai Overseas Scholarship. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Aman Agarwal, Soumya Basu, Tobias Schnabel, and Thorsten Joachims. 2017. Effective Evaluation using Logged Bandit Feedback from Multiple Loggers. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, 687–696.
[2] Mert Demirer, Vasilis Syrgkanis, Greg Lewis, and Victor Chernozhukov. 2019. Semi-Parametric Efficient Policy Learning with Continuous Actions. In *Advances in Neural Information Processing Systems*, Vol. 32.
[3] Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. 2014. Doubly Robust Policy Evaluation and Optimization. *Statist. Sci.* 29, 4 (2014), 485–511.
[4] Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. 2018. More Robust Doubly Robust Off-policy Evaluation. In *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80. PMLR, 1447–1456.
[5] Nan Jiang and Lihong Li. 2016. Doubly Robust Off-Policy Value Evaluation for Reinforcement Learning. In *Proceedings of the 33rd International Conference on Machine Learning*, Vol. 48. PMLR, 652–661.
[6] Thorsten Joachims and Adith Swaminathan. 2016. Counterfactual Evaluation and Learning for Search, Recommendation and Ad Placement. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. Association for Computing Machinery, 1199–1201.
[7] Nathan Kallus, Yuta Saito, and Masatoshi Uehara. 2021. Optimal Off-Policy Evaluation from Multiple Logging Policies. In *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139. PMLR, 5247–5256.
[8] Nathan Kallus and Angela Zhou. 2018. Policy Evaluation and Optimization with Continuous Treatments. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. PMLR, 1243–1251.
[9] Masahiro Kato, Shota Yasui, and Masatoshi Uehara. 2020. Off-Policy Evaluation and Learning for External Validity under a Covariate Shift. In *Advances in Neural Information Processing Systems*, Vol. 33. 49–61.
[10] Shuai Li, Yasin Abbasi-Yadkori, Branislav Kveton, S Muthukrishnan, Vishwa Vinay, and Zheng Wen. 2018. Offline Evaluation of Ranking Policies with Click Models. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, 1685–1694.
[11] Anqi Liu, Hao Liu, Anima Anandkumar, and Yisong Yue. 2019. Triply Robust Off-Policy Evaluation. *arXiv preprint arXiv:1911.05811* (2019).
[12] James McInerney, Brian Brost, Praveen Chandar, Rishabh Mehrotra, and Benjamin Carterette. 2020. Counterfactual Evaluation of Slate Recommendations with Sequential Reward Interactions. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, 1779–1788.
[13] Noveen Sachdeva, Yi Su, and Thorsten Joachims. 2020. Off-policy Bandits with Deficient Support. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, 965–975.
[14] Yuta Saito. 2020. Doubly robust estimator for ranking metrics with post-click conversions. In *Fourteenth ACM Conference on Recommender Systems*. Association for Computing Machinery, 92–100.
[15] Yuta Saito, Shunsuke Aihara, Megumi Matsutani, and Yusuke Narita. 2020. Open Bandit Dataset and Pipeline: Towards Realistic and Reproducible Off-Policy Evaluation. *arXiv preprint arXiv:2008.07146* (2020).
[16] Nian Si, Fan Zhang, Zhengyuan Zhou, and Jose Blanchet. 2020. Distributionally Robust Policy Evaluation and Learning in Offline Contextual Bandits. In *Proceedings of the 38th International Conference on Machine Learning*, Vol. 119. PMLR, 8884–8894.
[17] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, 2219–2228.
[18] Ashudeep Singh and Thorsten Joachims. 2019. Policy Learning for Fairness in Ranking. In *Advances in Neural Information Processing Systems*, Vol. 32.
[19] Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. 2020. Doubly Robust Off-Policy Evaluation with Shrinkage. In *Proceedings of the 37th*

International Conference on Machine Learning, Vol. 119. PMLR, 9167–9176.
[20] Yi Su, Lequn Wang, Michele Santacatterina, and Thorsten Joachims. 2019. Cab: Continuous Adaptive Blending for Policy Evaluation and Learning. In *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97. PMLR, 6005–6014.
[21] Adith Swaminathan and Thorsten Joachims. 2015. Batch Learning from Logged Bandit Feedback through Counterfactual Risk Minimization. *The Journal of Machine Learning Research* 16, 1 (2015), 1731–1755.
[22] Lequn Wang, Yiwei Bai, Wen Sun, and Thorsten Joachims. 2021. Fairness of Exposure in Stochastic Bandits. *arXiv preprint arXiv:2103.02735* (2021).
[23] Lequn Wang and Thorsten Joachims. 2020. Fairness and Diversity for Rankings in Two-Sided Markets. *arXiv preprint arXiv:2010.01470* (2020).
[24] Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudık. 2017. Optimal and adaptive off-policy evaluation in contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70. PMLR, 3589–3597.
[25] Himank Yadav, Zhengxiao Du, and Thorsten Joachims. 2019. Fair Learning-to-Rank from Implicit Feedback. *arXiv preprint arXiv:1911.08054* (2019).