A Gentle Introduction to Recommendation as Counterfactual Policy Learning

Flavian Vasile
Criteo AI Lab
f.vasile@criteo.com

David Rohde
Criteo AI Lab
d.rohde@criteo.com

Olivier Jeunen
University of Antwerp
olivier.jeunen@uantwerp.be

Amine Benhalloum
Criteo AI Lab
ma.benhalloum@criteo.com

ABSTRACT

The objective of this tutorial is to give a structured overview of the conceptual frameworks behind current state-of-the-art recommender systems, explain their underlying assumptions, the resulting methods and their shortcomings, and to introduce an exciting new class of approaches that frames the task of recommendation as a counterfactual policy learning problem. The tutorial can be divided into two modules. In module 1, participants learn about current approaches for building real-world recommender systems that comprise mainly of two frameworks, namely: recommendation as optimal auto-completion of user behaviour and recommendation as reward modelling. In module 2, we present the framework of recommendation as a counterfactual policy learning problem and go over the theoretical guarantees that address the shortcomings of the previous frameworks. We then proceed to go over the associated algorithms and test them against classical methods in RecoGym, an open-source recommendation simulation environment.

Overall, we believe the subject of the course is extremely actual and fills a gap between the consecrated recommendation frameworks and the cutting edge research and sets the stage for future advances in the field.

ACM Reference Format:
Flavian Vasile, David Rohde, Olivier Jeunen, and Amine Benhalloum. 2020. A Gentle Introduction to Recommendation as Counterfactual Policy Learning. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20), July 14–17, 2020, Genoa, Italy.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3340631.3398666

1 INTRODUCTION & MOTIVATION

Traditional implicit-feedback recommender systems aim to answer the following question: “Given the natural sequence of user-item interactions up to time t, can we predict which item the user will interact with at time $t + 1$?”. Actual real-world deployed recommender systems, however, pose a subtly different question: “Given the natural sequence of user-item interactions up to time t, can we get the user to interact with a recommended item at time $t+1$?” [7]. This includes the aspect of *interaction*, which is at the core of recommender systems research. Looking at it from a causal perspective, the recommender will perform an intervention, and we want to measure and optimise its effect. In order to build a system that optimises for this interactive aspect, we need a dataset that reflects this: logs of shown recommendations, along with users' reactions. As we can only observe the outcome of recommendations that were actually shown to the user, we call this *bandit* feedback. This has long been the focus of computational advertising with most of the practical state-of-the-art methods being around large-scale logistic regression models for click prediction, such as in [9].

For both existing frameworks, the majority of new recommendation algorithms presented in academic papers are evaluated on an offline dataset of logged user-item interactions, with results reported for some offline ranking metric. Recent work has shown repeatedly that offline evaluation results tend to diverge from online performance [5]. From a practical or industrial point of view, a need arises for offline evaluation methods that are robust, reproducible and closely related with the actual online objectives of the deployed recommender system. The reinforcement learning literature has long dealt with similar issues. Based on logged data from a certain *policy* (recommender), we want to predict what the performance would have been if another policy had been deployed. Counterfactual estimators, often based on importance sampling, are at the heart of this type of evaluation. Recent work has shown that they can be adopted for the recommendation use-case, and accurately reflect online performance [2, 3].

Recently, several simulation environments have been proposed for the recommendation setting, allowing online experiments such as A/B-tests to be simulated [4, 11]. They enable us to explore the use of bandit feedback for learning and evaluation of recommender systems, and have opened up promising new research directions. In this tutorial and its hands-on sessions, we will make us of the RecoGym simulator.[1] We introduce the idea of counterfactual evaluation, and present an overview of methods that can be used in recommender use-cases. Through the simulator, we show how these estimators succeed in providing sensible measurements for online metrics, demonstrating their merit in real-world environments. Additionally, we show where they fall short and present open research directions. Moving on from evaluation, we explore the problem of learning a new and optimal recommendation policy based on a dataset consisting of logged bandit feedback [6, 10]. Several recent papers propose solutions for this general problem setting [1, 8, 12]. There are, however, some specifics to the recommendation use-case that make it non-trivial to straightforwardly apply them. Through

UMAP '20, July 14–17, 2020, Genoa, Italy
© 2020 Copyright held by the owner/author(s).
This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20), July 14–17, 2020, Genoa, Italy*, https://doi.org/10.1145/3340631.3398666.

[1] https://github.com/criteo-research/reco-gym/

the use of RecoGym, we highlight and overcome these pitfalls, such as highly stochastic rewards, small effect sizes, and large action spaces.

2 TUTORIAL OUTLINE

In this tutorial, we aim to foster a general understanding of the modern approaches to recommendation in the UMAP research community. To this end, the goals of the tutorial are three-fold:

(1) To clarify the landscape of current approaches by by highlighting their underlying assumptions, which leads to a clear separation of the state-of-the-art in two distinct frameworks with complete different approaches, research communities and bodies of work.
(2) To go over one of biggest problems in the current recommender systems community, which is the problem of the misalignment between offline and online performance. We show the underlying causes, which are those of *Covariate Shift* and *Optimizer's Curse*, and how they relate to current approaches.
(3) To introduce the framework of Counterfactual Policy Learning [12], show how it can solve the two aforementioned issues and go over the existing state-of-the-art methods. As a note, this framework is different from online bandit methods, since in our setting, we do not control the exploration behaviour, thus having no control over the exploitation-exploration trade-off. This distinction is also known as the distinction between off-policy vs. on-policy methods.

Hands-on sessions with Jupyter notebooks running on Google Colaboratory are included in the course material.[2] The tutorial builds further on material that was previously presented at the 2019 Data Science Summer School organised by École Polytéchnique in Paris, France; as well as the 2019 ACM RecSys Summer School in Gothenburg, Sweden. We keep the material up-to-date with recent advancements so this edition is the most recent and complete version to date. The tutorial is expected to be of interest to recommender systems researchers and practitioners across the board. Aside from a general computer science and math background, no additional prerequisite knowledge is required. The outline of the tutorial is as follows:

Module I: Recommendation via Reward Modelling

(1) Classical vs. Modern Approaches: Recommendation as Behavioural Auto-complete vs. Recommendation as an Intervention Policy
(2) Recommendation Reward Modelling via Maximum Likelihood Models
(3) Shortcomings of Classical Value-based Recommendations

Module II: Recommendation as Policy Learning Approaches

(1) Policy Learning: Introduction, Concepts and Notations
(2) Issues of Value-based Models and Policy Learning
   (a) Covariate Shift and Counterfactual Risk Minimization
   (b) Optimizer's Curse and Distributional Robust Optimization

3 PRESENTERS

Flavian Vasile is part of the Criteo AI Lab where he works as the ML Recommendations Solutions Architect, with his main focus being on the development of deep learning-based recommendation systems and on introducing aspects of causal inference to recommendation. Among his recent research publications, the work on *Causal Embeddings for Recommendation* received the best paper award at RecSys 2018.

David Rohde is a Research Scientist at Criteo. His research interests are around Bayesian machine learning, offline evaluation and causal inference. He has numerous publications in applied and theoretical aspects of machine learning, from topics including variational approximations, causal inference, doubly intractable models to astronomy, analysing massive public transport datasets and evaluating recommender systems.

Olivier Jeunen is a PhD Student at the University of Antwerp, Belgium. His main line of research focuses on implicit-feedback recommender systems and applications of counterfactual and causal inference to recommendation. He regularly collaborates with industrial research labs such as Technicolor, Froomle and Criteo.

Amine Benhalloum is a Senior Machine Learning Engineer at Criteo, working on building large scale representation learning and retrieval systems for recommendation, applying Deep learning to personalize billions of daily display ads, reaching billions of users and connecting them with millions of products.

REFERENCES

[1] L. Bottou, J. Peters, J. Quiñonero-Candela, D. Charles, D. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. 2013. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research* 14, 1 (2013), 3207–3260.

[2] A. Gilotte, C. Calauzènes, T. Nedelec, A. Abraham, and S. Dollé. 2018. Offline A/B Testing for Recommender Systems. In *Proc. of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18)*. ACM, 198–206.

[3] A. Gruson, P. Chandar, C. Charbuillet, J. McInerney, S. Hansen, D. Tardieu, and B. Carterette. 2019. Offline Evaluation to Make Decisions About Playlist Recommendation Algorithms. In *Proceedings of the 12th ACM International Conference on Web Search and Data Mining* (Melbourne VIC, Australia) *(WSDM '19)*. ACM, New York, NY, USA, 420–428.

[4] E. Ie, C. Hsu, M. Mladenov, V. Jain, S. Narvekar, J. Wang, R. Wu, and C. Boutilier. 2019. RecSim: A Configurable Simulation Platform for Recommender Systems. arXiv:1909.04847 [cs.LG]

[5] O. Jeunen. 2019. Revisiting Offline Evaluation for Implicit-feedback Recommender Systems. In *Proc. of the 13th ACM Conference on Recommender Systems* (Copenhagen, Denmark) *(RecSys '19)*. ACM, 596–600.

[6] O. Jeunen, D. Mykhaylov, D. Rohde, F. Vasile, A. Gilotte, and M. Bompaire. 2019. Learning from Bandit Feedback: An Overview of the State-of-the-art. arXiv:1909.08471 [cs.IR]

[7] O. Jeunen, D. Rohde, and F. Vasile. 2019. On the Value of Bandit Feedback for Offline Recommender System Evaluation. arXiv:1907.12384 [cs.IR]

[8] T. Joachims, A. Swaminathan, and M. de Rijke. 2018. Deep Learning with Logged Bandit Feedback. In *Proc. of the 6th International Conference on Learning Representations (ICLR '18)*.

[9] H. B. McMahan, G. Holt, D. Sculley, M. Young, D. Ebner, J. Grady, L. Nie, T. Phillips, Eugene D., D. Golovin, et al. 2013. Ad click prediction: a view from the trenches. In *Proc. of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1222–1230.

[10] D. Mykhaylov, D. Rohde, F. Vasile, M. Bompaire, and O. Jeunen. 2019. Three Methods for Training on Bandit Feedback. arXiv:1904.10799 [cs.IR]

[11] D. Rohde, S. Bonner, T. Dunlop, F. Vasile, and A. Karatzoglou. 2018. RecoGym: A Reinforcement Learning Environment for the problem of Product Recommendation in Online Advertising. *ArXiv e-prints* (Aug. 2018). arXiv:1808.00720 [cs.IR]

[12] A. Swaminathan and T. Joachims. 2015. Counterfactual Risk Minimization: Learning from Logged Bandit Feedback. In *Proc. of the 32nd International Conference on International Conference on Machine Learning (ICML'15)*. JMLR.org, 814–823.

[2] https://colab.research.google.com/