Counterfactual Policy Learning for Recommendation

Olivier Jeunen[1], David Rohde[2] and Flavian Vasile[2]

[1] Adrem Data Lab, University of Antwerp, Belgium
[2] Criteo AI Lab, Paris, France

Abstract.

Conventional approaches to recommendation often do not explicitly take into account information on previously shown recommendations and their recorded responses. One reason is that, since we do not know the outcome of actions the system did not take, learning directly from such logs is not a straightforward task. Several methods for off-policy or counterfactual learning have been proposed in recent years, but their efficacy has not yet been validated in a recommender systems context. Due to the limitations of offline datasets and the lack of access of most academic researchers to online experiments, this is a non-trivial task for which simulation environments can provide a solution.

In this work, we conduct the first large-scale empirical study of counterfactual learning methods for recommendation, in a simulated environment. We consider various different policy-based methods that make use of the Inverse Propensity Score (IPS) to perform Counterfactual Risk Minimisation (CRM), as well as value-based methods based on Maximum Likelihood Estimation (MLE). We show that under certain assumptions the value- and policy-based methods have an identical parameterisation, allowing us to propose a new model that combines both the MLE and CRM objectives. This "Dual Bandit" approach achieves state-of-the-art performance in a wide range of scenarios, excelling the most in the realistic presence of large action spaces, finite training samples, and limited randomisation. We show that a logarithmic variant of the traditional IPS estimator can further improve empirical performance in *stochastic* and *sparse* settings; while drawing parallels with other conventional extensions such as variance or entropy regularisation. Our findings shed light on the practicality of counterfactual approaches in real-world recommendation environments.

The main contributions of our work are listed as follows:

1. We review the existing literature on offline learning from bandit feedback with a focus on the recommendation task, and discuss how these approaches relate to one another.
2. We present specific issues that arise within a recommender system context, such as stochastic and sparse rewards. We propose a novel logarithmic variant of the traditional IPS estimator, and show how it can improve empirical performance in such environments.
3. We show that, under certain assumptions, the value- and policy-based methods have an identical parameterisation. This allows us to propose the "Dual Bandit" objective that combines MLE and CRM without introducing additional parameters, and show that it achieves state-of-the-art performance for a wide range of settings.
4. We thoroughly investigate the impact of several factors, such as the quality of the logging policy, the number of items in the catalogue, and the size of the historical dataset used for training, on the quality of the learned models.