On the value of bandit feedback for offline recommender system evaluation

Reference:
Jeunen Olivier, Rohde David, Vasile Flavian.- On the value of bandit feedback for offline recommender system evaluation
Proceedings of the ACM RecSys Workshop on Reinforcement Learning and Robust Estimators for Recommendation Systems, September 20th, 2019, Copenhagen, Denmark - Copenhagen, ACM, 2019, 3 p.

On the Value of Bandit Feedback for Offline Recommender System Evaluation

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

David Rohde
Criteo AI Lab
Paris, France
d.rohde@criteo.com

Flavian Vasile
Criteo AI Lab
Paris, France
f.vasile@criteo.com

ABSTRACT

In academic literature, recommender systems are often evaluated on the task of next-item prediction. The procedure aims to give an answer to the 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 $t+1$?". Evaluation results obtained through said methodology are then used as a proxy to predict which system will perform better in an online setting. The online setting, however, poses 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 $t+1$?". From a causal perspective, the system performs an intervention, and we want to measure its effect. Next-item prediction is often used as a fall-back objective when information about interventions and their effects (shown recommendations and whether they received a click) is unavailable.

When this type of data *is* available, however, it can provide great value for reliably estimating online recommender system performance. Through a series of simulated experiments with the RecoGym environment, we show where traditional offline evaluation schemes fall short. Additionally, we show how so-called *bandit feedback* can be exploited for effective offline evaluation that more accurately reflects online performance.

ACM Reference format:
Olivier Jeunen, David Rohde, and Flavian Vasile. 2019. On the Value of Bandit Feedback for
Offline Recommender System Evaluation. In *Proceedings of the ACM RecSys Workshop on Reinforcement Learning and Robust Estimators for Recommendation Systems, Copenhagen, Denmark, September 20th, 2019 (REVEAL '19),* 3 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Traditionally, recommender systems are evaluated either through offline methods, online methods, or user studies [2, 17, 37]. In the offline setting, a previously collected dataset of preference expressions (be it explicit, implicit, or logged feedback) is used to assess the performance of competing recommendation methods. Online experiments deploy every competing method to a portion of real user traffic, and users' interactions with the systems are subsequently measured (often through A/B-tests, inter- or multi-leaving [5, 6, 25]). As online methods require a large number of resources and time, they are more costly than their offline counterpart [13]. Finally, the most expensive option, user studies are small-scale analyses where users' interactions with the system are studied in a more detailed manner, usually followed by qualitative questionnaires. In mixed-methods research, multiple of these variants are combined [11]. Because of the costly nature of these latter options, offline evaluation methodologies often remain a necessity when assessing algorithmic performance. However, results obtained through traditional offline evaluation schemes are often poorly correlated with true online performance [3, 12, 35]. A large portion of the academic literature surrounding recommender systems utilises offline evaluation procedures stemming from the broader field of supervised learning. In this setting, all true labels are assumed to be known and techniques like bootstrapping or k-fold cross-validation have been shown to provide accurate performance estimates [10, 24, 38]. In a recommender systems context, this line of research focuses on *organic* user behaviour: trying to find the items that naturally complement an already existing user sequence.

The recommender systems use case, however, is in some ways more closely related to reinforcement learning [39], multi-armed [33] and contextual bandits [26]. Here, the true *labels* or *rewards* are for the most part unknown. We observe rewards only for the actions (*recommendations*) that were actually performed (*shown to the user*). This is known as the *bandit feedback* setting [4, 22, 40]. Stemming from the reinforcement learning field of off-policy or counterfactual evaluation, a large body of work has recently focused on applying these techniques to provide accurate offline estimators of online recommender performance [1, 7, 13, 14, 19]. This second line of research aims to perform interventions (recommendations) that influence the user in some optimal way (leading them to click on, or purchase an item).

In this work, we present a comparison study of techniques from both fields. We investigate the value of using bandit feedback for recommender system evaluation, and show where traditional techniques (focusing solely on organic feedback) fall short. Through a range of experiments with the RecoGym environment [34], we empirically validate our findings.

2 METHODOLOGY

k-fold leave-one-out cross-validation (LOOCV) is one of the most used offline evaluation schemes in the literature. For every user sequence, an item is (or multiple items are) randomly sampled to be part of the test set. What remains of the user sequences then makes up the training set. Based on the training set, every model then generates a set of top-N recommendations for every user. Algorithms that can rank the missing sampled items highest in the set of recommendations are then assumed to be the best performers in an online environment as well. In order to provide

REVEAL '19, Copenhagen, Denmark

a robust estimate, this process is averaged over k different runs. As this technique has been used widely and recently to present new models as the state of the art [8, 9, 15, 27, 29–31, 41, 43], we adopt it as the representative for traditional offline evaluation in our experiments. Temporal evaluation procedures that take the chronological ordering of user-item interactions into account have recently gained traction as well [20, 23], but are out-of-scope for the purposes of this paper. The RecoGym environment simulates users' interests changing over time, but items remain stationary. As such, experiments showed no significant difference between a random or time-based split. We will consider hit-rate-at-1 (HR@1) as evaluation metric: the ratio of correct item predictions over all users. In this context, HR@1 is identical to Precision- and Recall@1.

Counterfactual or off-policy evaluation methods often use estimators based on importance sampling or inverse propensity scoring (IPS) [32]. By re-weighting (user, action, reward)-triplets according to how likely they are to occur under a new policy as compared to the old policy, various estimators for the performance of the new policy can be derived. The Clipped IPS (CIPS) estimator is an extension to classical IPS, exchanging variance for a pessimistic bias by putting a hard upper bound on these weights [4, 13, 18]. Assume we have a dataset $\mathcal{D}$ consisting of n logs $(x_i, a_i, p_i, \delta_i)$, where $x_i \in \mathbb{R}^d$ describes the user state, a_i is an identifier representing the action that was taken, $p_i \in (0, 1)$ denotes the probability with which that action was taken by the logging policy, and $\delta_i \in \{0, 1\}$ is the observed reward. Now, the CIPS estimator for a new policy π can be computed on samples from $\mathcal{D}$ as shown in Equation 1, where M denotes the maximally allowed sample weight.

$$\text{CIPS}(\pi, \mathcal{D}) = \frac{1}{n} \sum_{i=1}^{n} \delta_i \cdot \min\left(M, \frac{\pi(a_i|x_i)}{p_i}\right) \tag{1}$$

When the rewards δ_i are clicks, Equation 1 provides an estimate of the click-through-rate (CTR) that a new policy π will generate when deployed. Note that the logging policy π_0 needs to be stochastic, and have support over the same actions as π. The target policy π, however, can be deterministic.

3 EXPERIMENTAL RESULTS

We compare the traditional and counterfactual evaluation procedures as laid out in the previous sections with results obtained through a simulated A/B test[1]. Six recommendation approaches are compared: a random baseline, a popularity baseline recommending the item with the most organic views, a personalised popularity baseline recommending the item the specific user has organically viewed most often in the past, a latent-factor model based on a singular-value-decomposition of the user-item matrix [42], an item-based k-nearest-neighbours model [36], and a user-based k-nearest-neighbours model [16]. For illustrative purposes and brevity, we include only traditional methods in our comparison. Our findings, however, are model-agnostic and general. All models were trained on logged organic feedback obtained through the RecoGym environment with 2 000 users and 2 000 items. This leads to roughly 40 000 organic user-item interactions and 160 000 bandit-feedback samples. Note that all considered models only use organic information to generate recommendations: no bandit feedback is taken into account at learning time. We test our models on a set of 5 000 unseen users and the same set of items; with 100 000 organic and 390 000 bandit samples. The logging policy when generating test samples was stochastic and based on the personalised popularity baseline.

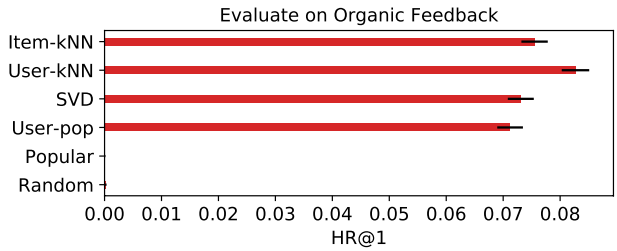

Figure 1: HR@1 as measured through 10-fold LOOCV on a logged dataset with organic feedback.

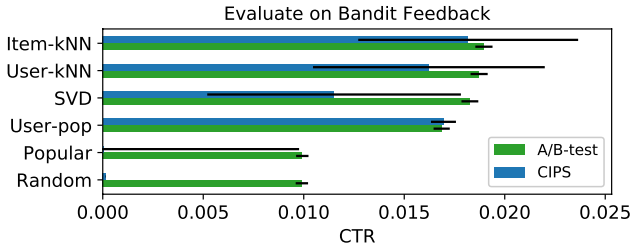

Figure 2: CTR estimate through the CIPS estimator on a logged dataset with bandit feedback ($M = 15$), and as measured through a simulated A/B-test.

Figure 1 shows the achieved HR@1 for all models, obtained through 10-fold LOOCV on the logged organic feedback. Figure 2 shows the estimated CTR for all models, obtained through the CIPS estimator on the logged bandit feedback, as well as measured results from a simulated A/B test. Key observations from these results are as follows: (1) LOOCV generates wildly different results in terms of absolute values, ratios and rankings among competing algorithms. These findings are in line with those presented in [21]. (2) The counterfactual CIPS estimator succeeds in providing sensible confidence intervals for the CTR. Although these intervals are wide, their true CTR value is almost always captured[2]. When ranking competing algorithms according to their upper confidence bound, we are able to infer the true ranking as obtained through the A/B test. (3) Due to an insufficient sample size, CIPS fails to accurately predict online performance for the Random baseline.

4 CONCLUSION

We presented an overview of the most often-used evaluation procedures for recommender systems, along with the distinction between *organic* and *bandit* feedback and how these terms relate to *supervised* and *counterfactual* evaluation techniques. Through a series of simulated experiments with RecoGym, we showed that algorithmic performance on predicting organic user behaviour is not necessarily a good proxy for the bandit task. When properly tuned, counterfactual estimators such as clipped IPS can accurately represent

[1] A notebook with all source code can be found at: https://git.io/fjyYq

[2] See [4, 28] for additional discussion regarding the -sometimes poor- coverage of traditional confidence intervals for importance sampling estimators.

model utility in an online setting. As such, when bandit feedback is available, it can be exploited for more effective evaluation.

REFERENCES

[1] A. Agarwal, S. Basu, T. Schnabel, and T. Joachims. 2017. Effective Evaluation Using Logged Bandit Feedback from Multiple Loggers. In *Proc. of the 23rd ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '17)*. ACM, 687–696.

[2] C. C. Aggarwal. 2016. *Evaluating Recommender Systems*. Springer International Publishing, 225–254.

[3] J. Beel, M. Genzmehr, S. Langer, A. Nürnberger, and B. Gipp. 2013. A Comparative Analysis of Offline and Online Evaluations and Discussion of Research Paper Recommender System Evaluation. In *Proc. of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation (RepSys '13)*. 7–14.

[4] 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.

[5] B. Brost, I. J. Cox, Y. Seldin, and C. Lioma. 2016. An Improved Multileaving Algorithm for Online Ranker Evaluation. In *Proc. of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*. ACM, 745–748.

[6] O. Chapelle, T. Joachims, F. Radlinski, and Y. Yue. 2012. Large-scale Validation and Analysis of Interleaved Search Evaluation. *ACM Trans. Inf. Syst.* 30, 1, Article 6 (March 2012), 41 pages.

[7] M. Chen, A. Beutel, P. Covington, S. Jain, F. Belletti, and E. H. Chi. 2019. Top-K Off-Policy Correction for a REINFORCE Recommender System. In *Proc. of the 12th ACM International Conference on Web Search and Data Mining (WSDM '19)*. ACM, 456–464.

[8] E. Christakopoulou and G. Karypis. 2016. Local Item-Item Models For Top-N Recommendation. In *Proc. of the 10th ACM Conference on Recommender Systems (RecSys '16)*. ACM, 67–74.

[9] E. Christakopoulou and G. Karypis. 2018. Local Latent Space Models for Top-N Recommendation. In *Proc. of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*. ACM, 1235–1243.

[10] B. Efron. 1982. *The jackknife, the bootstrap, and other resampling plans*. Vol. 38. Siam.

[11] J. Garcia-Gathright, C. Hosey, B. St. Thomas, B. Carterette, and F. Diaz. 2018. Mixed Methods for Evaluating User Satisfaction. In *Proc. of the 12th ACM Conference on Recommender Systems (RecSys '18)*. ACM, New York, NY, USA, 541–542.

[12] F. Garcin, B. Faltings, O. Donatsch, A. Alazzawi, C. Bruttin, and A. Huber. 2014. Offline and Online Evaluation of News Recommender Systems at Swissinfo.Ch. In *Proc. of the 8th ACM Conference on Recommender Systems (RecSys '14)*. 169–176.

[13] 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.

[14] 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 (WSDM '19)*. ACM, New York, NY, USA, 420–428.

[15] R. He, W. Kang, and J. McAuley. 2017. Translation-based Recommendation. In *Proc. of the 11th ACM Conference on Recommender Systems (RecSys '17)*. ACM, 161–169.

[16] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl. 1999. An Algorithmic Framework for Performing Collaborative Filtering. In *Proc. of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*. ACM, 230–237.

[17] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. 2004. Evaluating Collaborative Filtering Recommender Systems. *ACM Transactions on Information Systems* 22, 1 (Jan. 2004), 5–53.

[18] E. L. Ionides. 2008. Truncated Importance Sampling. *Journal of Computational and Graphical Statistics* 17, 2 (2008), 295–311.

[19] R. Jagerman, I. Markov, and M. de Rijke. 2019. When People Change Their Mind: Off-Policy Evaluation in Non-stationary Recommendation Environments. In *Proc. of the 12th ACM International Conference on Web Search and Data Mining (WSDM '19)*. ACM, 447–455.

[20] O. Jeunen. 2019. Revisiting Offline Evaluation for Implicit-Feedback Recommender Systems. In *Proc. of the 13th ACM Conference on Recommender Systems (RecSys '19)*. ACM, 5.

[21] O. Jeunen, K. Verstrepen, and B. Goethals. 2018. Fair Offline Evaluation Methodologies for Implicit-feedback Recommender Systems with MNAR Data. In *Proc. of the REVEAL 18 Workshop on Offline Evaluation for Recommender Systems (RecSys '18)*.

[22] 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)*.

[23] M. Jugovac, D. Jannach, and M. Karimi. 2018. Streamingrec: A Framework for Benchmarking Stream-based News Recommenders. In *Proc. of the 12th ACM Conference on Recommender Systems (RecSys '18)*. ACM, 269–273.

[24] R. Kohavi et al. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proc. of the 1995 International Joint Conference on Artificial Intelligence*, Vol. 14. 1137–1145.

[25] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne. 2009. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery* 18, 1 (01 Feb 2009), 140–181.

[26] J. Langford and T. Zhang. 2008. The Epoch-Greedy Algorithm for Multi-armed Bandits with Side Information. In *Advances in Neural Information Processing Systems 20*, J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis (Eds.). Curran Associates, Inc., 817–824.

[27] X. Li and J. She. 2017. Collaborative Variational Autoencoder for Recommender Systems. In *Proc. of the 23rd ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '17)*. ACM, 305–314.

[28] A. Maurer and M. Pontil. 2009. Empirical Bernstein Bounds and Sample Variance Penalization. *stat* 1050 (2009), 21.

[29] X. Ning and G. Karypis. 2011. SLIM: Sparse Linear Methods for Top-N Recommender Systems. In *Proc. of the 2011 IEEE 11th International Conference on Data Mining (ICDM '11)*. IEEE Computer Society, 497–506.

[30] Y. Ning, Y. Shi, L. Hong, H. Rangwala, and N. Ramakrishnan. 2017. A Gradient-based Adaptive Learning Framework for Efficient Personal Recommendation. In *Proc. of the 11th ACM Conference on Recommender Systems (RecSys '17)*. ACM, 23–31.

[31] R. Otunba, R. Rufai, and J. Lin. 2017. MPR: Multi-Objective Pairwise Ranking. In *Proc. of the 11th ACM Conference on Recommender Systems (RecSys '17)*. ACM, 170–178.

[32] A. B. Owen. 2013. *Monte Carlo theory, methods and examples*.

[33] H. Robbins. 1952. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* 58, 5 (1952), 527–535.

[34] 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:cs.IR/1808.00720

[35] M. Rossetti, F. Stella, and M. Zanker. 2016. Contrasting Offline and Online Results when Evaluating Recommendation Algorithms. In *Proc. of the 10th ACM Conference on Recommender Systems (RecSys '16)*. ACM, 31–34.

[36] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. 2001. Item-based Collaborative Filtering Recommendation Algorithms. In *Proc. of the 10th International Conference on World Wide Web (WWW '01)*. ACM, 285–295.

[37] G. Shani and A. Gunawardana. 2011. Evaluating Recommendation Systems. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor (Eds.). Springer US, 257–297.

[38] M. Stone. 1974. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)* 36, 2 (1974), 111–133.

[39] R. S. Sutton and A. G. Barto. 1998. *Introduction to reinforcement learning*. Vol. 135.

[40] 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 - Volume 37 (ICML'15)*. JMLR.org, 814–823.

[41] J. Yang, C. Chen, C. Wang, and M. Tsai. 2018. HOP-rec: High-order Proximity for Implicit Recommendation. In *Proc. of the 12th ACM Conference on Recommender Systems (RecSys '18)*. ACM, 140–144.

[42] S. Zhang, W. Wang, J. Ford, F. Makedon, and J. Pearlman. 2005. Using singular value decomposition approximation for collaborative filtering. In *Proc. of the 7th IEEE International Conference on E-Commerce Technology (CEC '05)*. 257–264.

[43] Y. Zhang, H. Lu, W. Niu, and J. Caverlee. 2018. Quality-aware Neural Complementary Item Recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys '18)*. ACM, 77–85.