Reward optimizing Recommendation using Deep Learning and Fast Maximum Inner Product Search

Imad Aouali
imadaouali9@gmail.com
Criteo

Amine Benhalloum
ma.benhalloum@criteo.com
Criteo

Martin Bompaire
m.bompaire@criteo.com
Criteo

Achraf Ait Sidi Hammou
aitsidihammou.achraf@gmail.com
Criteo

Sergey Ivanov
s.ivanov@criteo.com
Criteo

Benjamin Heymann
b.heymann@criteo.com
Criteo

David Rohde
d.rohde@criteo.com
Criteo

Otmane Sakhi
o.sakhi@criteo.com
Criteo

Flavian Vasile
f.vasile@criteo.com
Criteo

Maxime Vono
m.vono@criteo.com
Criteo

ABSTRACT

How can we build and optimize a recommender system that must rapidly fill slates (i.e. banners) of personalized recommendations? The combination of deep learning stacks with fast maximum inner product search (MIPS) algorithms have shown it is possible to deploy flexible models in production that can rapidly deliver personalized recommendations to users. Albeit promising, this methodology is unfortunately not sufficient to build a recommender system which maximizes the reward, *e.g.* the probability of click. Usually instead a proxy loss is optimized and A/B testing is used to test if the system actually improved performance. This tutorial takes participants through the necessary steps to model the reward and directly optimize the reward of recommendation engines built upon fast search algorithms to produce high-performance reward-optimizing recommender systems.

CCS CONCEPTS

• **Computing methodologies** → *Maximum likelihood modeling*; *Neural networks*; • **Information systems** → **Recommender systems**.

ACM Reference Format:
Imad Aouali, Amine Benhalloum, Martin Bompaire, Achraf Ait Sidi Hammou, Sergey Ivanov, Benjamin Heymann, David Rohde, Otmane Sakhi, Flavian Vasile, and Maxime Vono. 2022. Reward optimizing Recommendation using Deep Learning and Fast Maximum Inner Product Search. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3534678.3542622

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08.
https://doi.org/10.1145/3534678.3542622

1 TUTORIAL OUTLINE

1.1 Deep Learning Combined with MIPS - A Winning Combination

In this section we outline the capability of combining Deep Learning with Maximum Inner Product Search in a production environment [8]. The recommendation engine relies on learning both a query function and P embeddings, one per item in the catalogue, which will later be indexed by the maximum inner product search. Randomized Singular Value Decomposition [10] can be used to produce embeddings of dimension d that can be used for further training just like NLP tasks often rely on pre-trained embeddings such as Word2Vec [9] or BERT [5].

1.2 A Slate-Level Reward Model that Combines Reward and Rank

This module presents how we can leverage reward-optimizing recommendation to build an efficient and scalable slate recommender system that combines both reward information *i.e.* whether the user interacted with the banner, and rank signal *i.e.* the position of the selected item in the banner [1]. The benefits of the proposed methodology, *e.g.* recommendation performance and speed, in large-scale scenarios are illustrated by running A/B tests in a simulated environment. We compare our method with common and recently-proposed policy-learning approaches, such as inverse propensity scoring [12] and the top$-K$ heuristic proposed in [4]. We show that these baselines suffer from important caveats such as high variance, over-simplifying assumptions on the parametrised policy and poor scaling when the catalogue size becomes large. In contrast, by both combining reward and rank signals and by leveraging fast (approximate) MIPS techniques, the proposed framework shows promising recommendation results while meeting low-latency requirements.

1.3 Estimating Reward with the Horvitz-Thompson Estimator

The syllabus will be drawn from [2, 3, 7, 11, 12].

The direct way to use the Horvitz Thompson estimator embodies the assumptions of maximum inner product search but is extremely high variance:

$$E[c|\Xi, \boldsymbol{\beta}] \approx \sum_{n=0}^{N} \frac{c_n \pi_{\Xi,\beta,K}(a_1, \ldots, a_K|\Omega)}{\pi_0(a_1, \ldots, a_K|\Omega)}$$

where N is the number of data points, c_n is the reward, $\pi_{\xi,\beta}(a_1, \ldots, a_K|\Omega)$ is the policy parameterized to be maximum inner product search friendly and $\pi_0(a_1, \ldots, a_K|\Omega)$ is the propensity score of the slate.

We investigate several proposals in the slate setting that reduce the variance at the expense of introducing bias to become managable in the recommender setting. We further show that by restricting the policy we are able to optimize maximum inner product search based algorithms.

1.4 Scaling REINFORCE to large catalogs with MIPS

Given a reward estimator $\hat{R}$ (a Reward model, Horwitz-Thompson estimator, Doubly Robust estimator..), Offline Policy based methods aim at learning a parametrised policy π_θ that maximizes the average reward on the logged data $\bar{R}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{a\sim\pi_\theta(.|x_i)}[\hat{R}(a, x_i)]$. We can achieve this by leveraging the REINFORCE gradient $\nabla_\theta \bar{R}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{a\sim\pi_\theta(.|x_i)}[\hat{R}(a, x_i)\nabla_\theta \log \pi_\theta(a|x_i)]$ that enables us to optimize our objective function to obtain reward maximizing policies. In the context of large scale recommender systems, this objective can be computationally demanding as it scales linearly with the size of the catalog. In this module, we want to shed light on a newly proposed method scaling logarithmically on the catalog size by leveraging Maximum Inner Product Search algorithms, allowing faster training time without losing in the quality of the policy learned. We will cover the intuition behind the approach and provide notebooks with toy and real world examples. We will also talk about how to naturally extend the method to slate recommendation with Plackett-Luce and the problems that can be faced when using such algorithms [6].

REFERENCES

[1] Imad Aouali, Achraf Ait Sidi Hammou, Sergey Ivanov, Otmane Sakhi, David Rohde, and Flavian Vasile. 2022. Probabilistic Rank and Reward: A Scalable Model for Slate Recommendation. (2022).

[2] Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. 2011. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 19–26.

[3] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. 2013. Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising. *Journal of Machine Learning Research* 14, 11 (2013).

[4] Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H Chi. 2019. Top-k off-policy correction for a REINFORCE recommender system. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 456–464.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] Artyom Gadetsky, Kirill Struminsky, Christopher Robinson, Novi Quadrianto, and Dmitry Vetrov. 2020. Low-variance black-box gradient estimates for the plackett-luce distribution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 10126–10135.

[7] Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. 2018. Offline a/b testing for recommender systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 198–206.

[8] Olivier Koch, Amine Benhalloum, Guillaume Genthial, Denis Kuzin, and Dmitry Parfenchik. 2021. Scalable representation learning and retrieval for display advertising. *arXiv preprint arXiv:2101.00870* (2021).

[9] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 26 (2013).

[10] Tae-Hyun Oh, Yasuyuki Matsushita, Yu-Wing Tai, and In So Kweon. 2017. Fast randomized singular value thresholding for low-rank optimization. *IEEE transactions on pattern analysis and machine intelligence* 40, 2 (2017), 376–391.

[11] Otmane Sakhi, Louis Faury, and Flavian Vasile. 2020. Improving Offline Contextual Bandits with Distributional Robustness. *arXiv preprint arXiv:2011.06835* (2020).

[12] Adith Swaminathan and Thorsten Joachims. 2015. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*. PMLR, 814–823.