dversarial Personalized Ranking for Recommendation[*]

Xiangnan He
National University of Singapore
xiangnanhe@gmail.com

Zhankui He
Fudan University
zkhe15@fudan.edu.cn

Xiaoyu Du
Chengdu University of Information Technology
duxy.me@gmail.com

Tat-Seng Chua
National University of Singapore
dcscts@nus.edu.sg

BSTR CT

Item recommendation is a personalized ranking task. To this end, many recommender systems optimize models with pairwise ranking objectives, such as the Bayesian Personalized Ranking (BPR). Using matrix Factorization (MF) — the most widely used model in recommendation — as a demonstration, we show that optimizing it with BPR leads to a recommender model that is not robust. In particular, we find that the resultant model is highly vulnerable to adversarial perturbations on its model parameters, which implies the possibly large error in generalization.

To enhance the robustness of a recommender model and thus improve its generalization performance, we propose a new optimization framework, namely *dversarial Personalized Ranking* (PR). In short, our PR enhances the pairwise ranking method BPR by performing adversarial training. It can be interpreted as playing a minimax game, where the minimization of the BPR objective function meanwhile defends an adversary, which adds adversarial perturbations on model parameters to maximize the BPR objective function. To illustrate how it works, we implement PR on MF by adding adversarial perturbations on the embedding vectors of users and items. Extensive experiments on three public real-world datasets demonstrate the effectiveness of PR — by optimizing MF with PR, it outperforms BPR with a relative improvement of 11.2% on average and achieves state-of-the-art performance for item recommendation. Our implementation is available at: https://github.com/hexiangnan/adversarial_personalized_ranking.

CCS CONCEPTS

• **Information systems → Recommender systems**; **Information retrieval**; **Retrieval models and ranking**;

KEYWORDS

Personalized Ranking, Pairwise Learning, dversarial Training, Matrix Factorization, Item Recommendation

CM Reference Format:
Xiangnan He, Zhankui He, Xiaoyu Du, and Tat-Seng Chua. 2018. dversarial Personalized Ranking for Recommendation. In *SIGIR '18: 41st International CM SIGIR Conference on Research and Development in Information Retrieval, July 8-12, 2018, nn rbor, MI, US* . CM, New York, NY, US , 10 pages. https://doi.org/10.1145/3209978.3209981

1 INTRODUCTION

Recent advances on adversarial machine learning [30] show that many state-of-the-art classifiers are actually very fragile and vulnerable to *adversarial examples*, which are formed by applying small but intentional perturbations to input examples from the dataset. typical example can be found in Figure 1 of [15], which demonstrates that by adding small adversarial perturbations to an image of panda, a well-trained classifier misclassified the image as a gibbon with a high confidence, whereas the effect of perturbations can hardly be perceived by human. This points to an inherent limitation of training a model on static labeled data only. To address the limitation and improve model generalization, researchers then developed adversarial training methods that train a model to correctly classify the dynamically generated adversarial examples [15, 25].

While the inspiring progress of adversarial machine learning mainly concentrated on the computer vision domain where the adversarial examples can be intuitively understood, to date, there is no study about such adversarial phenomenon in the field of information retrieval (IR). lthough the core task in IR is ranking, we point out that many learning to rank (L2R) methods are essentially trained by optimizing a classification function, such as the pairwise L2R method Bayesian Personalized Ranking (BPR) in recommendation [28], among others [21]. This means that it is very likely that the underlying IR models also lack robustness and are vulnerable to certain kinds of "adversarial examples". In this work, we aim to fill the research gap by exploring adversarial learning methods on item recommendation, an active and fundamental research topic in IR that concerns personalized ranking.

Nevertheless, directly grafting the way of generating adversarial examples from the image domain is infeasible, since the inputs of recommender models are mostly discrete features (*i.e.,* user ID, item ID, and other categorical variables). Clearly, it is meaningless to apply noises to discrete features, which may change their semantics. To address this issue, we consider exploring the robustness of a recommender model at a deeper level — at the level of its intrinsic

[*] This work is done when during the internship of Zhankui He and Xiaoyu Du at National University of Singapore. NExT research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its IRC@SG Funding Initiative.

SIGIR'18, July 8–12, 2018, nn rbor, MI, US
© 2018 ssociation for Computing Machinery.
CM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3209981