Adversarial Machine Learning in Recommender Systems

Felice Antonio Merra
Amazon
Germany
felmerra@amazon.de*

Abstract

Recommender systems are ubiquitous. Our digital lives are influenced by their use when, for instance, we select the news to read, the product to buy, the friend to connect with, and the movie to watch. While enormous academic research efforts have been mainly focused on getting high-quality recommendations to reach maximum user satisfaction, little effort has been devoted to studying the integrity and security of these systems. Is there an underlying relationship between the characteristics of the historical user-item interactions and the efficacy of injection of false users/feedback strategies against collaborative models? Can public semantic data be used to perform attacks more potent in raising the recommendability of victim items? Can a malicious user poison or evade the image data of visual recommenders with adversarial perturbed product images? Is the family of model-based recommenders more vulnerable to multi-step gradient-based adversarial perturbations? Furthermore, is the adversarial training robustification still effective in the last scenario? Is this training defense influencing the beyond-accuracy and bias performance?

This dissertation intends to pave the way towards more robust recommender systems, beginning with understanding how to robustify a model, what is the cost of robustness in terms of reduction of recommendation accuracy, and which are the novel adversarial risks of modern recommenders. This thesis, getting inspiration from the literature on the security of collaborative models against the insertion of hand-engineered fake profiles and the recent advances of adversarial machine learning methods in other research areas like computer vision, contributes to several directions: (i) the proposal of a practical framework to interpret the impact of data characteristics on the robustness of collaborative recommenders [Deldjoo et al., 2020], (ii) the design of powerful attack strategies using publicly available semantic data [Anelli et al., 2020], (iii) the identification of severe adversarial vulnerabilities of visual-based recommender models where adversaries can break the recommendation integrity by pushing products to the highest recommendation positions with a simple and human-imperceptible perturbation of products' images [Anelli et al., 2021b], (iv) the proposal of robust adversarial perturbation methods capable of completely breaking the accuracy of matrix factorization recommenders [Anelli et al., 2021a], and (v) a formal study that examines the effects of adversarial training in reducing the recommendation quality of state-of-the-art model-based recommenders Anelli et al. [2021c].

*Work performed while at Politecnico di Bari, Italy.

Awarded by: Politecnico di Bari, Bari, Italy **on** 24 January 2022.
Supervised by: Tommaso DI Noia.
Available at: https://iris.poliba.it/retrieve/dd89f8a6-faa4-ccdd-e053-6605fe0a1b87/Adversarial_Machine_Learning_in_Recommender_Systems.pdf.

Selected Publications

Vito Walter Anelli, Yashar Deldjoo, Tommaso Di Noia, Eugenio Di Sciascio, and Felice Antonio Merra. Sasha: Semantic-aware shilling attacks on recommender systems exploiting knowledge graphs. 12123:307–323, 2020.

Vito Walter Anelli, Alejandro Bellogín, Yashar Deldjoo, Tommaso Di Noia, and Felice Antonio Merra. MSAP: multi-step adversarial perturbations on recommender systems embeddings. 2021a.

Vito Walter Anelli, Yashar Deldjoo, Tommaso Di Noia, Daniele Malitesta, and Felice Antonio Merra. A study of defensive methods to protect visual recommendation against adversarial manipulation of images. pages 1094–1103, 2021b.

Vito Walter Anelli, Yashar Deldjoo, Tommaso Di Noia, and Felice Antonio Merra. A formal analysis of recommendation quality of adversarially-trained recommenders. pages 2852–2856, 2021c.

Yashar Deldjoo, Tommaso Di Noia, Eugenio Di Sciascio, and Felice Antonio Merra. How dataset characteristics affect the robustness of collaborative recommendation models. pages 951–960, 2020.