Project proposal: Membership Inference Attacks on Sequence Recommender Systems

Huan Zhao (U00879231), Jiaqi Xu (U00787062)

Abstract

Membership Inference Attacks (MIA) are a significant threat to recommender systems, as they can reveal whether a specific user's data was part of the model's training set, posing privacy risks. Traditional MIA techniques are less effective for sequential recommendation models, where high generalization reduces output differences between training and non-training sequences. This makes it challenging to distinguish training data in sequential models, as non-training sequences can yield similar recommendations. Observing that overfitting creates distinct distributions between training and non-training data, we leverage this phenomenon to enhance MIA. Our approach balances training and non-training sequences, overfits the transformer model as target model to highlight distribution differences, and trains a binary classifier as attack model to detect training participation effectively. We validate our method using standard metrics, demonstrating its accuracy in identifying training data.

1 Introduction

Membership Inference Attacks (MIA) are a major concern in securing recommender systems, as they attempt to reveal whether a specific user's data was part of the model's training set Shokri et al., 2017. For instance, in a video streaming platform, a vulnerable recommender system could allow an attacker to determine if a user has viewed certain sensitive or stigmatized content. By identifying individuals who interacted with specific items, the attacker could infer private details about users' preferences, behaviors, or even their identities, posing a serious risk to user privacy. Carlini et al., 2022

Research such as Zhang et al., 2021 and Wang et al., 2022 has shown that model outputs often correlate more with inputs when the sequence is part of the training set, making it feasible to identify training sequences. However, as noted by Zhu et al., 2023 and Chi et al., 2024, sequential recommendation models often provide equally relevant recommendations regardless of an input's training status. This property reduces the effectiveness of traditional MIA techniques, as it minimizes the output differences between sequences inside and outside the training data.

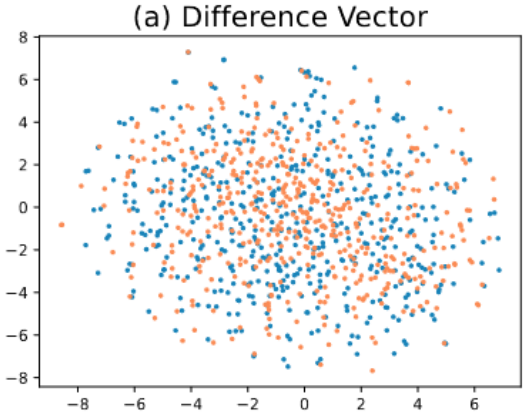

Figure 1: Visualization of Output-Input Correlation Distribution for Training and Non-Training Sequences by t-SNE algorithm.

Figure 1 illustrates this by showing similar output-input correlations for both training and non-training sequences, suggesting that high generalization by the model diminishes MIA efficacy based on output-input correlation disparities.

Identifying whether a general sequence was part of the training set is especially challenging in sequential recommender models. These models are designed to generalize effectively, making it hard to distinguish training sequences from others solely based on outputs. Additionally, non-training sequences can still yield recommendations closely resembling those of training sequences, complicating detection in sequential settings Zhu et al., 2023.

Observing that the model exhibits different distributions between training and non-training data when overfitting, our method leverages this phe-

nomenon to address the challenges. We begin by preparing balanced data, ensuring equal representation of training and non-training sequences. Next, we overfit the sequential recommender model (target model) on the training data to accentuate distinct output distributions. These labeled distributions are then used to train a Multi-Layer Perceptron (MLP) classifier (attack model), enabling effective identification of whether a general sequence was part of the training data.

Finally, we evaluate our attack method using standard metrics, including accuracy, precision, recall, and F1-score, to assess its effectiveness in detecting training participation.

2 Data and Process

3 Method Overview

In this section, we describe our method, which leverages the observation that, under overfitting conditions, the distributions of training and non-training sequences differ significantly. The accompanying figure 2 illustrates this phenomenon. To utilize this distinction for membership inference attacks (MIA), we design our approach as shown in Fig 3.

Our method begins with a dataset split into training and non-training databases. These datasets are fed into a sequential recommender model (target model), which is intentionally overfitted on the training data to accentuate the differences in output distributions. The model produces scores for each sequence, which are then categorized into two groups: member data (training data) and non-member data (non-training data). To ensure a balanced representation of both categories, we carefully preprocess the data before using it to train a Multi-Layer Perceptron (MLP) attack model. This MLP model leverages the amplified output distribution differences to effectively distinguish members from non-members, enabling precise membership inference attacks. The figure clearly outlines the pipeline, highlighting the critical role of overfitting and data balancing in our approach.

The core methods follow a pipeline approach to:

Data Processing: We use the MovieLens dataset, which includes user interaction histories (such as clicks or views) and is sufficiently large for training deep recommendation models. Each item is mapped to a continuous integer index starting from 0 for computational efficiency. To structure user interactions for sequence-based recommendations,

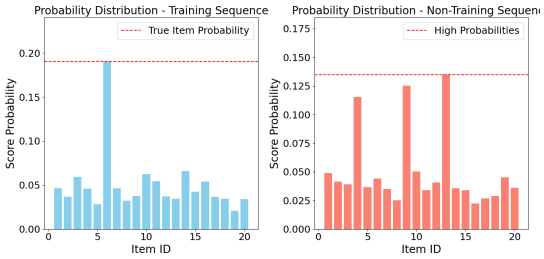

Figure 2: Visualization of Output-Input Correlation Distribution for Training and Non-Training Sequences by t-SNE algorithm.

we divide each user's history into sessions based on timestamps, where each session represents a sequence of interactions within a specific timeframe. This session-based structuring captures the temporal order of interactions, providing a realistic foundation for training sequential recommendation models.

Data Augmentation: To increase the diversity of similar sequences, we compute item similarities using an interaction matrix derived from user-item interactions. Using these similarity scores, we augment the data by replacing a random item in each session with a high-similarity item. This augmentation generates new, similar sequences, balancing the positive and negative samples in the training set and enhancing the model's robustness to similar, unseen sequences.

Sequence Processing: Each session sequence is tokenized and standardized to a fixed length. Longer sequences are truncated, while shorter ones are padded with a designated token. Additionally, start and end tokens are added to each sequence to mark its boundaries. This structured format allows the model to recognize sequence limits and interpret each session consistently during training.

Target Model Training: Using the processed and augmented sequence data, we train a transformer-based recommendation model as the target model. This model captures sequential patterns in user sessions to make item recommendations, generating outputs that reflect the characteristics of the training data. By learning these sequential dependencies, the target model becomes more responsive to patterns specific to the training sequences.

Attack Model Training: Finally, using the target model's output data, we train a binary classification neural network (MLP) as the attack model. This attack model is designed to identify subtle differences between training and non-training sequences,

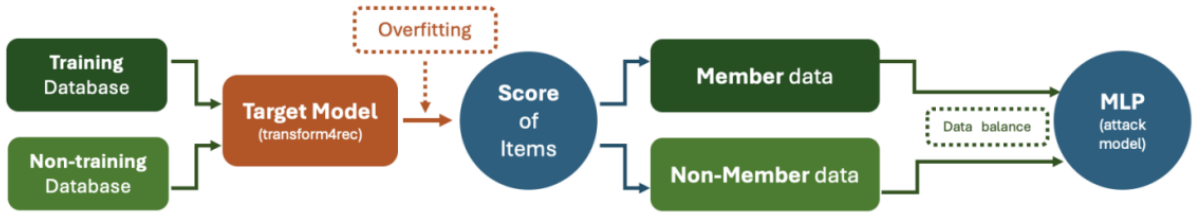

Figure 3: Pipeline of the Proposed Membership Inference Attack Method.

allowing it to detect whether a new sequence was part of the target model's training data. By distinguishing between sequences seen during training and those it hasn't encountered, the attack model effectively performs membership inference.

This five-step approach enables us to leverage the model's tendency to overfit, allowing the attack model to accurately differentiate between training and non-training sequences for MIA detection.

4 Experiment setting

In this section, we describe the dataset and running environment used in our study.

4.1 Dataset

For our experiments, we use the **YooChoose** dataset, specifically the `yoochoose-clicks.dat` file, available from the RecSys Challenge 2015 on Kaggle. This dataset contains clickstream data, recording user interactions with various items over time. Each entry includes a session ID, timestamp, item ID, and category, allowing us to reconstruct user sessions, which makes this dataset ideal for training and evaluating sequential recommendation models. We use only the use interactions table, as they form the majority of data and we are interested in next-click prediction.

Table 1: Dataset statistics

Dataset	Days	Items (K)	Sessions (M)	Interactions (M)	Sessions Length (avg.)	Gini Index
YOOCHOOSE eCommerce	182	50,549	6,756,575	26,478,390	3.83	0.89

4.2 Running Environment

We perform our experiments on a cloud-based Paperspace server, equipped with NVIDIA GPUs to accelerate training and inference. Additionally, we utilize NVIDIA's official `transformer4rec` PyTorch container for an optimized setup. Detailed information on Transformer4Rec can be found in its GitHub repository.

5 Evaluation

To evaluate the effectiveness of our proposed membership inference attack method on sequential recommenders, we conducted experiments using the widely adopted Yoochoose dataset. The evaluation metric is the area under the ROC curve (AUC), which measures the ability of the attack model to distinguish between training and non-training sequences. We compare our method against two baselines: DL-MIA, which applies a deep learning-based membership inference approach, and Biased-MIA, which incorporates bias adjustments to enhance performance. These baselines were chosen as they represent state-of-the-art methods for membership inference attacks in related domains.

5.1 Target Model: Transformer4Rec

Our target model is **Transformer4Rec**, a transformer-based sequential recommendation model developed by NVIDIA as part of the Merlin framework. Transformer4Rec utilizes the transformer architecture to capture sequential dependencies in user interactions, providing effective session-based recommendations. The model is trained on the processed dataset to learn sequential patterns and generate item recommendations.

5.2 Attack Model: Multi-Layer Perceptron (MLP)

Our attack model is a **Multi-Layer Perceptron (MLP)**, a neural network architecture suited for binary classification tasks. The MLP model is trained on the output data from the target model, allowing it to detect subtle differences between training and non-training sequences, thereby performing membership inference. The MLP architecture consists of multiple fully connected layers, effectively learning from the target model's output distributions to classify sequences based on their presence in the training data.

5.3 Baseline Model

Baselines We use Biased-MIA Zhang et al., 2021 and DL-MIA Wang et al., 2022 to compare with our model.

- Biased-MIA is the frst algorithm to implement MIA against the recommender system. They identify users through the diference between user historical behaviors and the recom mendations provided by the target model.

- DL-MIA improved Biased-MIA by debiasing learning. They design a disentangled encoder and a weight estimator to simultaneously mitigate training data and estimation biases.

5.4 Evaluation Metrics

To evaluate the performance of our attack model, we use the Accuracy (AUC). Because AUC represents the proportion of correctly classified instances (both training and non-training sequences) out of the total instances, providing an overall measure of the model's performance. We adopt the area under the ROC curve (AUC) which is threshold independent as our evaluation metric. AUC is widely used in binary classifcation problems due to its insensitivity to the label distribution of the dataset.

6 Results

The results demonstrate the superiority of our proposed method over the baselines. Specifically, our method achieves an AUC of 0.90012, significantly outperforming DL-MIA (0.50352) and Biased-MIA (0.50694) on the Yoochoose dataset. This substantial improvement highlights the effectiveness of leveraging output distribution differences in sequential recommender systems. By overfitting the target model on training data, our approach accentuates the separation between training and non-training sequences, enabling the MLP-based attack model to achieve high accuracy. These findings validate the robustness of our method in addressing the challenges of membership inference attacks in sequential recommender systems.

To further illustrate the effectiveness of our approach, we visualized the output distributions of the Yoochoose dataset using t-SNE, as shown in Figure 4. The red points represent members (training data), while the blue points denote non-members (non-training data). This visualization clearly reveals a distinct separation between the distributions of members and non-members, a phenomenon accentuated by overfitting the sequential recommender model. Our method is specifically designed to exploit this disparity by training an MLP classifier on the labeled distributions, enabling it to effectively distinguish between members and non-members. This observation underscores why our approach achieves significantly higher AUC compared to the baseline methods.

Table 2: Attack performance (AUC) over sequential recommenders

Algorithm	yoochoose
DL-MIA	0.50352
Biased-MIA	0.50694
Our	0.90012

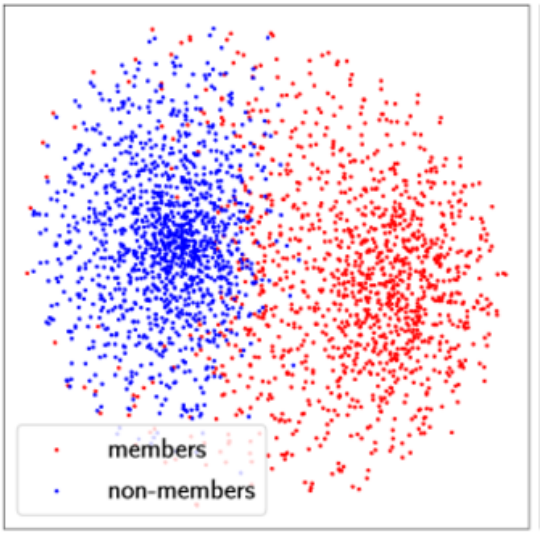

Figure 4: Visualization of yoochoose dataset results by t-SNE, where red points denote members and blue points represent non-members.

7 Discussion and Takeaways

Our proposed method introduces a novel and effective approach to membership inference attacks in sequential recommenders by leveraging the distinct output distribution differences caused by overfitting. By preparing balanced data and intentionally overfitting the target model on training data, we amplify the separation between members and non-members, as evidenced by both quantitative results and the t-SNE visualization. This clear separation enables our MLP-based attack model to achieve significantly higher accuracy than state-of-the-art baselines. The strength of our approach lies in its

systematic exploitation of overfitting-induced vulnerabilities, turning what is traditionally seen as a limitation of machine learning into a mechanism for effective attack. These findings shed light on critical privacy risks in sequential recommenders, highlighting the need for future research to develop defenses against such attacks while maintaining the utility and performance of recommender systems.

References

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914. IEEE.

Xiaoxiao Chi, Xuyun Zhang, Yan Wang, Lianyong Qi, Amin Beheshti, Xiaolong Xu, Kim-Kwang Raymond Choo, Shuo Wang, and Hongsheng Hu. 2024. Shadow-free membership inference attacks: Recommender systems are more vulnerable than you thought. *arXiv preprint arXiv:2405.07018*.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE.

Zihan Wang, Na Huang, Fei Sun, Pengjie Ren, Zhumin Chen, Hengliang Luo, Maarten de Rijke, and Zhaochun Ren. 2022. Debiasing learning for membership inference attacks against recommender systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1959–1968.

Minxing Zhang, Zhaochun Ren, Zihan Wang, Pengjie Ren, Zhunmin Chen, Pengfei Hu, and Yang Zhang. 2021. Membership inference attacks against recommender systems. pages 864–879.

Zhihao Zhu, Chenwang Wu, Rui Fan, Defu Lian, and Enhong Chen. 2023. Membership inference attacks against sequential recommender systems. In *Proceedings of the ACM Web Conference 2023*, pages 1208–1219.