Cyber Security Project Report

Privacy Risk Analysis on Machine Learning Models using Membership Inference Attacks

By

Lakshman Kruthik Manubolu
Jibin Yesudas Vargheese

Abstract

The proliferation of machine learning models in sensitive applications has underscored the criticality of understanding and mitigating privacy risks associated with such models. This study examines the adaptability of metric-based attack models on defence mechanisms, focusing on the MemGuard defence method against membership inference attacks. Building on the foundational work by Song and Mittal, which established the efficacy of metric based attacks within certain datasets, we extend the investigation to the Purchase 100 dataset, aiming to ascertain the attack's scalability. Our empirical results reveal a significant reduction in attack accuracy, challenging the prevailing assumptions about the universal effectiveness of metric-based attacks. This paper presents a systematic analysis contrasting these findings with those of previous studies and discusses the technical implications underlying the discrepancies observed in the defence mechanism's performance across various datasets.

Background

The paper, "Systematic Evaluation of Privacy Risks of Machine Learning Models,", authored by Liwei Song and Prateek Mittal,and presented at the 30th USENIX Security Symposium, 2011, evaluates the vulnerability of machine learning models to membership inference attacks. The paper's primary contribution is to assess the risk of such attacks systematically and to benchmark the effectiveness of various defence strategies.

The researchers developed new methods for testing machine learning models vulnerability to membership inference attacks. Unlike previous techniques, they used non-neural network-based attacks, particularly based on metrics like confidence, correctness and entropy, and included strategies that adjust thresholds for different classes and a new attack based on a modified version of prediction entropy.

In particular, it focuses on two defence methods: adversarial regularisation and MemGuard. These defence methods are designed to protect the privacy of the individuals whose data were used to train the models by reducing the accuracy of an attacker trying to determine if a particular data point was part of the training dataset.

Adversarial regularisation is a technique that aims to make the machine learning model's predictions on its training data indistinguishable from its predictions on new, unseen data. This is done by incorporating a term into the training objective that penalises the model when there is a significant difference between predictions on training data and non-training data. Essentially, it adds noise to the training process to obscure the traces of the individual data points.

MemGuard, on the other hand, is an approach that defends against membership inference attacks by adding carefully crafted noise to the model's outputs. This noise is intended to confuse the attacker without significantly deteriorating the utility of the model for legitimate predictions.

The paper reports that previously, with the help of these defences, the accuracy of the neural network attacks was brought down to 50% from 81% and 74%, which is close to a random guess, indicating that the defences were initially effective. However, with the help of the authors' benchmarks of their newly introduced metric based attack, the attack accuracy improved significantly, going up to 69-74% on certain datasets. This suggests that while the defences provide a certain level of privacy protection, they are not foolproof, especially when the attacker has a well-designed attack strategy.

The researchers also made use of Privacy Risk Score, which was a new way to measure how likely it is that a data sample was in the training set of a model,by accounting for a fine-grained analysis of risks where the focus shifted to individual data points instead of the whole model. The researchers show that this score can be used to make very targeted attacks and that it's related to how sensitive the model is, how well it generalises, and the nature of the data it's been trained on. Thus this metric provides a nuanced view of privacy risks, allowing for the identification of data points at high risk of exposure.

For our project, we will focus on the author’s attack’s scalability on the MemGuard defence model that was once believed to improve security significantly by reducing successful attack rates to nearly random guessing levels (50%), as their paper

demonstrated that attack accuracies could increase dramatically to 69-74% under certain conditions, indicating vulnerabilities in the defence.

Our research will expand on this by testing how well the attacks, proposed by the authors, perform on different datasets, specifically Purchase 100. We aim to analyse the scalability of the non-neural network-based attacks and assess the privacy risk score metric's effectiveness in this new context. This investigation into the scalability of the attack will contribute to our understanding of how to protect sensitive data in machine learning applications.

Research Question :

Our research question and goal are centred around testing the scalability and robustness of the attack model proposed by the researchers on the Purchase 100 dataset. Specifically, we aim to:

- Validate/Challenge the researchers' findings that non-neural network-based attacks, which utilise metrics like entropy, confidence, modified prediction entropy, and correctness, are effective in breaching privacy across different datasets.
- Assess the effectiveness of defence mechanisms, like MemGuard, when applied to datasets other than those examined in the original study.
- Investigate the performance of the metric-based attack model in both defended and undefended scenarios to understand the true resilience of machine learning models.
- Evaluate the privacy risk score metric in the context of the Purchase 100 dataset under the Memguard defence to see if the privacy risk is consistently assessed across different data samples and models.

Our project differs from the published paper in that we are extending the application and testing of their proposed attack model to an additional dataset—Purchase 100—that was not the primary focus in their MemGuard evaluations. While the researchers did use Purchase 100 to evaluate adversarial regularisation, our research will apply the MemGuard defence to this dataset to establish whether their attack model's success can be

replicated and to what extent the defence can mitigate the attacks. This will provide further evidence for or against the generalizability of the attack model and the robustness of MemGuard as a defence mechanism. By doing this, we aim to offer a broader validation of the paper's claims and contribute new insights into the efficacy of privacy-preserving techniques in machine learning.

The issue at hand is crucial because machine learning models are increasingly being utilised in a variety of applications, many of which involve sensitive personal data. If these models can easily be attacked to reveal whether certain data was used in their training, it poses a serious risk to individual privacy. In a world where data breaches are costly and damaging, ensuring the robustness of machine learning models and the scalability of the defence mechanisms against such privacy attacks is essential.

Datasets:

The authors in the paper had used two datasets namely

- **Texas 100 :** The hospital discharge data used in this dataset. Data public use files with patient's information released by the Texas Department of State Health Services. This dataset encapsulates 67,330 instances, each with 6,170 binary attributes, distilled from the 100 most commonly recorded procedures. Each data record contains the external causes of injury (e.g., suicide, drug misuse), the diagnosis (e.g., schizophrenia), the procedures the patient underwent (e.g., surgery) and some generic Information, e.g. gender, age, race and so on.[3]
- **Location 30 :** The Foursquare database is used in this dataset, which contains location "check-in" records of several thousand users. Composed of 5,010 instances, the dataset is structured with 446 binary attributes and categorised into 30 unique clusters that signify diverse geosocial categories. All of these features are related to a given region or location Type and indicate if an individual has visited. Region location, or not.[3]

While training the model with this dataset the authors have focused on the Hospital and Location details to filter out how many users are visiting a particular location or discharged from a hospital and a bit of personal information. This information is private but may not contain sensitive information like personal numbers or transaction details.

Moreover these filtering mechanisms don't focus on a single user rather they focus on the hospital or location.

In contrast to this we have used a dataset **Purchase 100** : The dataset is based on Kaggle's acquired value shopping challenge which contains the transaction records of users.The dataset comprises 197,324 instances, each described by 600 binary attributes, categorised into 100 distinct groups that reflect varying purchasing patterns. Each feature is related to the product and represents whether or not it was purchased by an individual and if yes then what is the mode of purchase and personal details of the users like phone number, email etc.. We believe that this dataset is more specific to a single user and it also contains more sensitive information than above datasets.

Methodology

The authors in the paper proposed a Metrics based Black Box Inference Attack while considering the best defence in the attack scenario called MemGuard [1]. In the paper the authors had mentioned that the attack classifier, I, is trained following the shadow training technique , which takes the model prediction F(x) with the sample label y, and outputs a score I(F(x), y) For membership inference, in the range [0,1], if the output is: If the data sample is greater than 0.5, it is considered to be a member, vice versa. They have done this using Texas 100 and Location 30 datasets. Following the methodology proposed by Jia et al. [1], MemGuard was applied to defend classifiers trained on the Location30 and Texas100 datasets, employing a neural network architecture with four hidden layers. The numbers of neurons for these layers are 1024, 512, 256, and 128, respectively, and all utilise the rectified linear unit (ReLU) as the activation function. This configuration is crucial for understanding the environment in which MemGuard's effectiveness is evaluated and serves as a benchmark for comparison against the Purchase 100 dataset outcomes.

By applying the attack model to the Purchase 100 dataset, our project seeks to challenge the original paper’s findings regarding the efficacy of metric based attack and other non-neural network-based attacks. This is especially significant as the original paper did not fully explore MemGuard’s defence against the particular attacks in the context of

many other dataset. Our work stands to validate the generalizability of the researchers' attack model and the robustness of the defence mechanisms proposed.

We have used the same defence mechanism and shadow training technique but with Purchase 100 dataset. First we need to prepare the Purchase 100 dataset. After downloading the original dataset, a Python script has been implemented that automates the preparation of a dataset related to purchases. It will check if there is a directory for the dataset, extract it from an available tar archive to be processed and saved as an NPZ file containing input features and labels where appropriate. In each step, the code provides a status message and is capable of specifying batch sizes. Overall, for additional analyses or machine learning tasks, the process of preparing data sets is streamlined. We use that script to create an array of user details by focusing on the user-id that is mentioned in the original dataset. We have used user-id as our filtering feature because any other feature might be ambiguous and multiple users can have the same names. So user-id is a unique number in which each user once completed a transaction has given an id and that has been used as a feature.

After the data is preprocessed, two NPZ files called shuffle_index.npz and data_complete.npz were obtained for training and testing the shadow model. The shadow model was already created by the authors so we took help of that but with the Purchase100 dataset. Here data_complete.npz file will be having total dataset as an array with user_id as index and shuffle_index.npz file is the randomly picked 30% of the whole data for testing the shadow model.

Once the data is prepped and ready the attack which is present is the MIA_Evaluate.py file will be implemented using the prompt

Python MIA_Evaluate.py --dataset [texas or location or purchase] --defended [0 or 1].

Here we can select the dataset and whether the MemGuard functionality should be implemented or not. This prompt will give the accuracy of the attack for the respective dataset in terms of the performance metrics (Correctness, Entropy, Confidence, Modified

Entropy) and Privacy Risk Score which tells us how many data point are at privacy risk and how many are not.

Findings and Reflection

In this section, we focus on the results of our experiment regarding the adaptability of metric-based attack models, particularly those evaluated in the seminal work by Liwei Song and Prateek Mittal, which scrutinized the efficacy of MemGuard. Our replication of their benchmarks on the Location30 and Texas100 datasets corroborated their observations, revealing vulnerabilities in the defended models against sophisticated membership inference attacks.

Metric	**Defended Model**	**Undefended Model**
Correctness	68.7%	68.7%
Confidence	69.1%	76.3%
Entropy	52.1%	61.6%
Modified Entropy	68.8%	78.1%

Table 1. Our Findings on Location 30 Dataset which corroborated the Paper's Findings

Metric	**Defended Model**	**Undefended Model**
Correctness	74.2%	74.2%
Confidence	74.1%	79.0%
Entropy	54.6%	66.6%
Modified Entropy	74.0%	79.4%

Table 2. Our Findings on Texas 100 Dataset which corroborated the Paper's Findings

Our empirical analysis then extended to the Purchase 100 dataset, a considerably larger and possibly more complex dataset than those originally evaluated. This dataset's distinct properties provided a novel context for assessing the scalability of the attack models. Interestingly, our experimentation surfaced an intriguing anomaly: the attack accuracies for both defended and undefended models, especially when utilizing entropy as the attack

vector, dropped significantly—down to approximately 53% for the defended model and 66% for the undefended model. This is a marked decrease compared to the Location30 and Texas100 datasets, where attack accuracies were substantially higher. Moreover, in the defended model, attack accuracies for both confidence and modified entropy metrics were recorded below the 50% threshold, at 48.7% and 48.9% respectively—figures that beckon an analysis of the possible intricacies involved.

Metric	**Defended Model**	**Undefended Model**
Correctness	50%	50%
Confidence	48.7%	47.9%
Entropy	53.6%	66.4%
Modified Entropy	48.9%	47.4%

Table 3. Our Findings on Purchase 100 Dataset

This divergence prompts a technical discussion on the potential reasons why metric-based attacks may fall short in their scalability and adaptability:

Complexity and Size of the Dataset: Purchase 100's larger and more complex dataset may dilute the distinctiveness of the model's output on members versus non-members, hindering the attack's efficacy. This complexity may not be captured effectively by the metrics used in the attack models.

Feature Space and Model Overfitting: Given how the attacks depend on the threshold, if the attack models were implicitly tuned to the specific feature spaces and distribution patterns of the original datasets, they might fail to capture the nuances of a new dataset with different attributes. This suggests an overfitting of the attack models to the original datasets, limiting their transferability.

Defensive Mechanisms and Noise: The defended models introduce noise that may impact the confidence scores or entropy values used by the attack models. The efficacy of this noise introduction by MemGuard seems to be more pronounced in Purchase 100,

which could be attributed to the way noise interacts with the feature space of this particular dataset.

Metric Robustness: Metrics like confidence and entropy are commonly leveraged in membership inference attacks under the assumption that they are robust indicators of membership likelihood. However, our findings suggest that their robustness may be overstated and not as reliable in the face of varying dataset characteristics.

Class Imbalance and Distribution Shifts: Any class imbalance or distribution shift in Purchase 100 compared to Location30 and Texas100 could affect the attack models' accuracy. Metric-based models may not be equipped to adjust their thresholds or strategies to account for such disparities.

Dataset Specificity:
Finally, the Purchase 100 dataset could inherently possess a degree of specificity that renders metric-based attack models less effective. This could include unique inter-class relationships, feature correlations, or a more even spread of the underlying data that resists membership disclosure.

Furthermore, defence mechanisms like MemGuard may introduce noise that disproportionately affects metrics like confidence and modified entropy, leading to reduced attack success. It is also possible that attack models, if overly tailored to the features of the original datasets, could face limitations when confronted with new datasets like Purchase 100 that exhibit different characteristics.

The disparity in attack accuracy between defended and undefended models for certain metrics raises additional questions. It hints at the possibility that the noise introduced to confound attacks could be creating unintended patterns that attackers might exploit. This could also suggest an imbalance in the training of attack models, which might be more adept at countering specific types of attacks while being susceptible to others.

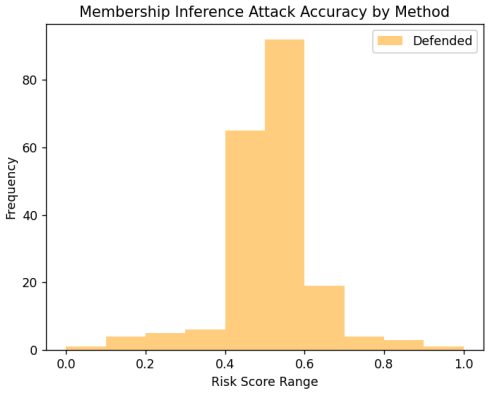

Fig.1-Purchase Defended Privacy risk score

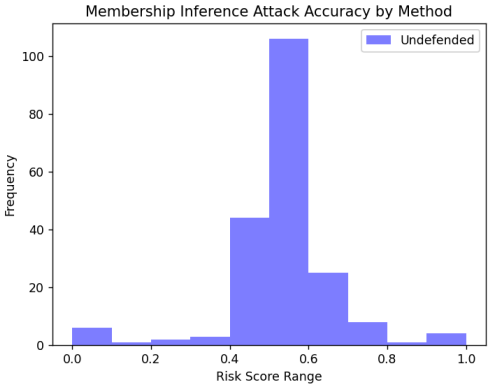

Fig.2-Purchase Undefended Privacy risk score

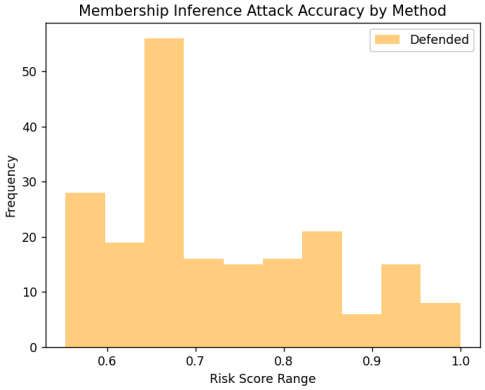

Fig.3-Location Defended Privacy risk score

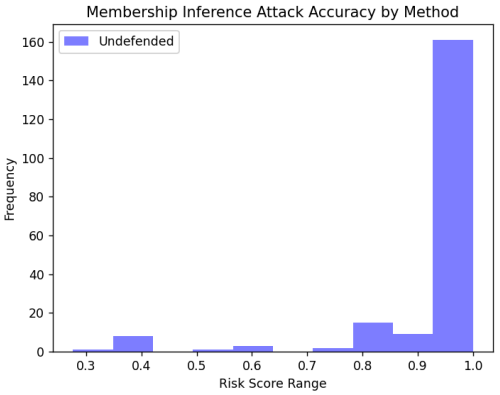

Fig.4-Location Undefended Privacy risk score

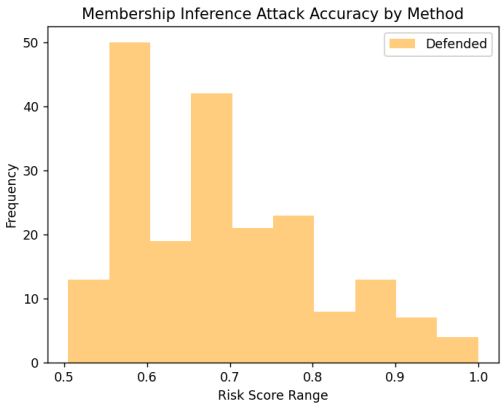

Fig.5-Texas Defended Privacy risk score

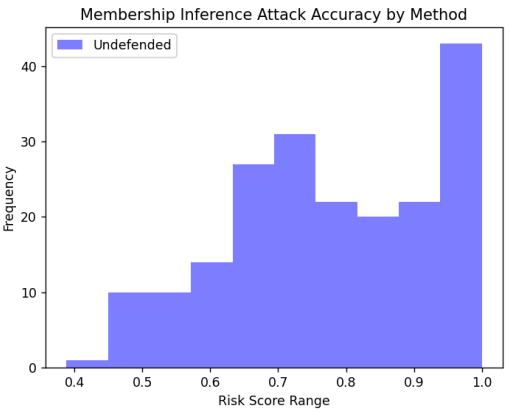

Fig.6-Texas Undefended Privacy risk score

Privacy Risk Scores Analysis: In the undefended scenarios, our results indicate that the Purchase 100 dataset exhibits a narrower risk score distribution with a substantial frequency in the higher risk score range, as evidenced by the histogram presented in Fig

2. This contrasts with a broader distribution observed in Fig 6 and Fig 4, suggesting a heightened privacy risk inherent to the Purchase 100 dataset in the absence of defenses. Upon deploying MemGuard, Fig 1 reflects a shift towards a lower risk score range. This shift is more pronounced than that observed in the Fig 5 and Fig 3, which implicates a relative efficacy of MemGuard in modulating privacy risk in the Purchase dataset.

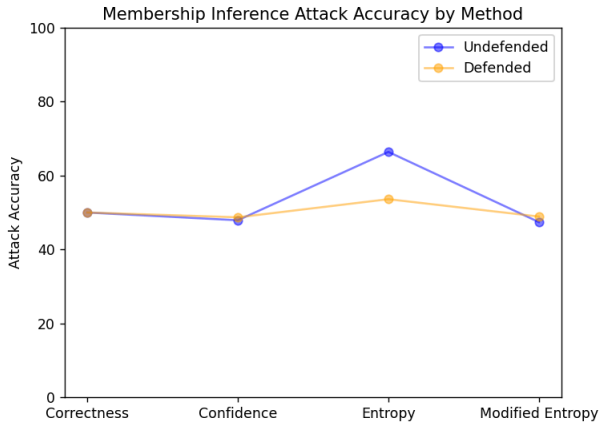

Fig.7 -Attack Accuracy Graph on Purchase 100

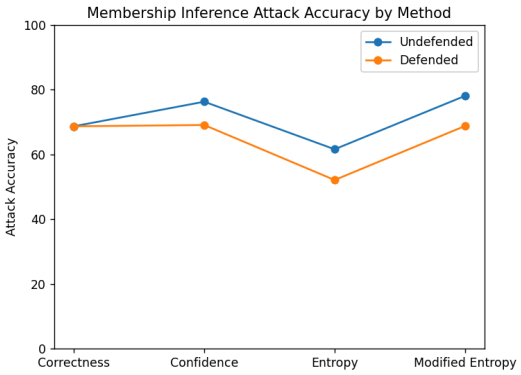

Fig.8 -Attack Accuracy Graph on Location30

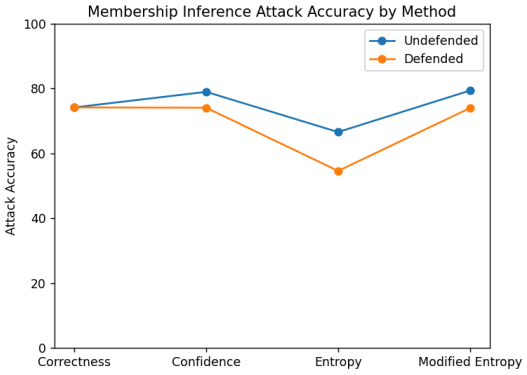

Fig.9 - Attack Accuracy graph on Texas100

Membership Inference Attack Accuracy: Our analyses present a noteworthy distinction in the effectiveness of membership inference attacks when considering entropy as a metric. In the case of the Purchase 100 dataset, the entropy-based attack accuracy decreases from 66.4% in the undefended model to 53.6% in the defended model (Fig 7). This diminution of attack accuracy is less pronounced in Location and Texas datasets, as depicted in Fig 8 and Fig 9, wherein the defended models' entropy-based attack accuracies show less variation from their undefended counterparts.

In light of these findings, we assert that while metric-based attack models present a significant threat to model privacy, their performance is evidently contingent on dataset-specific characteristics. The noted drop in attack efficacy on the Purchase 100 dataset suggests that the scalability of these attack models is not absolute and that their adaptability to datasets with differing intrinsic properties may be limited. Our research, therefore, contributes a critical perspective to the ongoing discourse on the privacy risks posed by machine learning models, highlighting the necessity for a dynamic approach to developing both attack and defence mechanisms.

Conclusion

The analysis of the Membership Inference Attack against various dataset shows that as the dataset becomes larger and more specific to the single user that is more private details of the user the MIA attack accuracy against any defence and even without any defence degrades and provide very less accuracy. In our analysis the attack with Texas and Location datasets show around 70% to 75% accuracy when the defence is in place and 75% to 80% when it is undefended. When it comes to the Purchase dataset it shows around 50% in both defended and undefended cases.

Our findings underscore the importance of context and dataset characteristics in assessing the vulnerability of machine learning models to privacy attacks. The specificity of the Purchase 100 dataset, characterised by unique user transactions and sensitive personal details, suggests that not all datasets are equally susceptible to MIAs, and thus, defence strategies need to be tailored to the specific context of the dataset. This complexity necessitates a dynamic and adaptable approach to privacy protection in machine learning models, as static defence mechanisms may not suffice against evolving threats.

Ultimately, our research expands on existing literature by providing empirical evidence of the variegated performance of attack models across datasets, thereby reinforcing the call for more nuanced and flexible privacy-preserving methodologies in machine learning.

The code and methodology supporting our research are accessible on the project's GitHub repository, as cited in the references [2], facilitating future studies and collaborations in this critical field of cybersecurity.

References

1. Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. Memguard: Defending against black-box membership inference attacks via adversarial examples. In ACM Conference on Computer and Communications Security, 2019.
2. https://github.com/Kruthikmanubolu/MIA Code for the project can be found here.
3. The paper that we followed : Song, Liwei and Prateek Mittal. "Systematic Evaluation of Privacy Risks of Machine Learning Models." *USENIX Security Symposium* (2020).