Membership Inference Attacks and Defences: A Survey

Noopa Jagadeesh
noopa.jagadeesh@ontariotechu.net
Ontario Tech University
Oshawa, Ontario

ABSTRACT

Current rise of Artificial intelligence (AI) is credited to machine learning and Deep learning. Machine Learning has emerged to be a strong and efficient framework that can be applied to a wide range of complex learning problems. As a result, machine learning is being extensively used in most of the current everyday applications. The use of machine learning on sensitive and private information, such as financial transaction data, conversations with friends,purchase histories, and health-related data, has expanded in the past years and so has the research on vulnerabilities within those machine learning systems. In this paper, we attempt to provide a detailed discussion on one type of attack—the membership inference attacks with various threat models and also elaborate the efficiency and challenges of recent countermeasures against them. In this paper, we attempt to comprehensively summarize the work that design membership inference attacks, analyze the existence of such attacks and also elaborate the efficiency and challenges of recent countermeasures against them.

KEYWORDS

Machine Learning, black-box attack, white-box attack, machine learning as a service

ACM Reference Format:
Noopa Jagadeesh. 2019. Membership Inference Attacks and Defences: A Survey. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Artificial intelligence makes it possible for machines to perform human-like tasks with high accuracy by learning from experience and adjusting to new inputs. The applications of artificial intelligence technologies in fields like healthcare, education, autonomous vehicles, e-commerce, finance have been rapidly developed recently. Much of AI is powered by breakthroughs in machine learning and deep learning. Machine learning feeds a computer data and uses statistical techniques to help it "learn" how to get progressively better at a task, without having been specifically programmed for that task, and thereby eliminating the need for millions of lines of code. Deep learning is a subset of machine learning that runs inputs through a biologically-inspired neural network architecture. The neural networks consist of number of hidden layers through which the data is processed, thereby allowing the machine to go "deep" in its learning, making connections and weighting input for the best results.

Machine learning(ML) is actively being used today, to improve business decisions, increase productivity, detect disease, forecast weather, image recognition, sentiment analysis, news classification, video surveillance, speech recognition, online fraud detection, medical services, recommendation for products and services, online customer supports, information retrieval. The success of ML has driven leading Internet companies to deploy machine learning as a service (MLaaS). Amazon's Amazon ML, Microsoft's Azure ML, IBM's Watson and Google Cloud ML are some of the leading providers of MLaaS services. MLaaS is an array of services that provide machine learning tools as part of cloud computing services. Under such services, a user uploads their own dataset to a server and the server returns a trained ML model to the user, typically in the form of a black-box API. MLaaS helps clients take the advantages from machine learning without worrying about time, cost, computational power and building an inhouse machine learning team. Infrastructural concerns such as data pre-processing, exploratory data analysis, model training, model evaluation, and model predictions on new input can be mitigated through MLaaS.

Despite being popular with all the shiny new capabilities that ML promises, ML models are susceptible to various security and privacy attacks, available for attackers to exploit. Some of these vulnerabilities can be leveraged by attackers to mess with model's learning and to get the model output an incorrect prediction. Other attacks, however, allow the attacker to extract sensitive and personal information from your model, such as the underlying data, hyperparameters or the model itself. The attacks on ML models can be organized into membership inference attack, adversarial attacks, model inversion, unintended memorization, model extraction or model stealing and hyperparameter stealing.

In this paper, we review recent studies on artificial intelligence membership inference attack and defense technologies. Membership inference attack happens when the attacker has a data point at hand and wants to check if that datapoint belong to the original training dataset used to train an ML model. Imagine, for example, if someone wanted to find out if your name was in some sensitive medical list or you participated in a survey? Shokri et al.[15] presented the first membership inference attack against machine learning models. In the case where samples are linked to a person, such as medical or financial data, inferring whether samples come from the training dataset of the ML model constitutes a privacy threat. For instance, membership inference attack can be done to check if a child participates to a study that aims to design the effectiveness of serious game designed for the improvement of communication skills and social behavior, social conversation, imaginative skills, sensory

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

integration and learning accounts in ASD children [22]. Successful membership inference attacks can lead to severe consequences.

In this paper, we review recent findings on membership inference attack and defense technologies and present a detailed understanding of the attack models and methodologies. In Section 2, we provide a taxonomy of the related terms and keywords. In Section 3, we introduce the causes and characteristics of membership inference attack, as well as the adversarial capabilities and goals. In Section 4, we conclude the existing defense methods against membership inference attack. In Section 5, we identify other types of attacks. We conclude in Section 6.

2 BACKGROUND

Machine learning is an application of artificial intelligence (AI) that focuses on the development of computer programs that can access data, automatically learn and improve from experience to become more accurate in predicting outcomes without being explicitly programmed. The majority of practical machine learning uses supervised learning. *Supervised learning* is where you have a labeled dataset and the aim of the algorithm is to construct a model that can learn relationships and dependencies between the target prediction output and the input attributes. The goal is to approximate the mapping function so well that your model generalizes well on new input data. The model is said to *overfit* if its prediction error is less on the training dataset while very large on the test/validation dataset. Many techniques such as increasing the size of the training dataset, reducing the number of attributes, regularization, early stopping have been proposed to prevent overfitting in ML models.

A number of internet companies have launched cloud-based ML services. These platforms provide APIs for uploading the training data sets, have the service provider chose and run training algorithms on the data, and make the resulting models available for prediction queries. The details of the models and the training algorithms used to build the model are hidden from the data owners. A model is *white-box* if a user may download a representation suitable for local use. A model is *black-box* if accessible only via a prediction query interface.

The attackers in the membership inference system has access to different levels of information. In *black-box attack* settings the adversaries have no knowledge about target model and its network parameters. The attacker has only access to the output for a given input. We refer to the *white-box attack* setting as the case where the attacker have total knowledge about the target model , including the training algorithm, data distribution , and model parameters .

3 MEMBERSHIP INFERENCE ATTACK STRATEGIES

Membership inference, is a type of information leakage where the attacker's goal is to determining whether a given data record was part of the model's training dataset or not. Previously, membership inference has been successfully conducted in many other domains, such as biomedical data [1] and mobility data [11]. The work of Backes et al. [1] sheds light on privacy risks stemming from microRNA expression data, showing that it is possible to detect membership in microRNA-based studies' datasets by relying on their published mean statistics with their experimental results clearly showing that membership is much easier to detect in disease-specific datasets than in random ones. Pyrgelis et al. [11] focus on membership inference attacks, whereby an adversary attempts to determine whether or not location data of a target user is part of the aggregates.

Shokri et al. [15] present the first membership inference attack against machine learning models in a black-box setting were the adversary's access to the model is limited to queries that return the model's output on a given input. To answer the membership inference question, they trained an attack model whose purpose is to distinguish the target model's behavior on the training inputs from its behavior on the inputs that it did not encounter during training thereby turning membership inference problem into a classification problem. The general idea behind this attack is to use several ML models (one for each prediction class), called as *attack models*, to make membership inference over the target model's output. Given that the target model is a black-box API, Shokri et al. propose to construct multiple *shadow models* to mimic the target model's behavior and derive the data necessary, to train attack models.The developed black-box membership inference techniques against ML models perform best when the target model is overfitted to the training data.

Shokri et al. [15] uses two strong assumptions which largely reduce the scope of membership inference attacks against ML models. First, the attacker needs to establish multiple shadow models with each one sharing the same structure as the target model. Second, the dataset used to train shadow models comes from the same distribution as the target model's training data. Salem et al. [13] gradually relaxed these assumptions in order to show that far more broadly applicable attack scenarios are possible and shows that one shadow model and one attack model are sufficient to achieve an effective attack compared to the proposal of multiple shadow models and attack models by Shokri et al. [15].

Truex et al. [19] extend and generalize this work to black-box and federated-learning settings. By exploring a variety of machine learning model types and their correlations with respect to the three phases of the attack generation process, Truex et al. [19] present five interesting characteristics of membership inference attacks: they are data-driven attacks, attack models are transferable, target model type is a strong indicator of model vulnerability, attack data generation techniques need not explicitly mirror the target model, and membership inference attacks can persist as insider attacks in federated systems.

Rahman et al. [12] use membership inference to evaluate the tradeoff between test accuracy and membership privacy in differentially private ML models. One of the promising approaches for privacy-preserving deep learning is to employ differential privacy during model training which aims to prevent the leakage of sensitive information about the training data via the trained model. The experimental results indicate that differentially private deep models may provide privacy protection against strong adversaries only by sacrificing model utility by a considerable margin. As a result, such models may be vulnerable to modern and sophisticated privacy attacks such as the membership inference attack, when providing a competitive utility level. This, advocate for an empirical value that best trade-offs the utility and privacy for the current application at hand.

Recent work on membership inference has shown that federated learning exposes participants to significant privacy risk [19], [10]. Black-box attacks might not be effective against deep neural networks that generalize well. Nasr et al. [10] designed and evaluated the first white-box membership inference attacks against neural network models in the stand-alone and federated settings by exploiting the privacy vulnerabilities of the stochastic gradient descent algorithm and demonstrated that even well-generalized models are significantly susceptible to white-box membership inference attacks.

Hayes et al. [4] gave the first study of membership inference against generative models based on both white-box and black-box access to the target model. This attack leverage Generative Adversarial Networks (GANs), which combine a discriminative and a generative model, to detect overfitting and recognize inputs that were part of training datasets, using the discriminator's capacity to learn statistical differences in distributions. Hilprecht et al. [5] proposed two membership inference attacks for generative models: the *Monte Carlo attack* and the *Reconstruction attack*. While the first is applicable to all generative models the latter is specialized for Variational Autoencoders (VAE). It was observed in this work that VAEs are more vulnerable to the membership inference attacks which in turn suggests that VAEs are more prone to overfitting than GANs if the same amount of training data is available.

Many recent works have also studied membership inference against machine learning from different angles [8] , [21], [7], [16]. Long et al. [8] show that well-generalized models can leak membership information, but the adversary must first identify a handful of vulnerable records in the training dataset. Long et al. [8] demonstrates that overfitting contributes to the information leaks but is not the fundamental cause of the problem. Yeom et al. [21] explores the relationships between privacy, overfitting, and influence in machine learning models. Their results confirm that models become more vulnerable to both membership inference and attribute inference attacks as they overfit more but also highlights that overfitting is not the only factor that can lead to privacy risk. Long et al. [7] propose *Differential Training Privacy (DTP)*, an empirical metric to estimate the privacy risk of publishing a classifier when methods such as differential privacy cannot be applied. DTP estimates the privacy risk of a training record by measuring its influence on the predictions of machine learning models with a large DTP indicating that the record's influence is strong enough to indicate its presence in the training dataset. To help enforce data-protection regulations and detect unauthorized uses of personal data, Song et al. [16] developed a model auditing technique that helps users check if their data was used to train a machine learning model.

4 DEFENSE STRATEGY

[15] identifies overfitting as an important (but not the only) reason why machine learning models leak information about their training datasets. Regularization techniques such as dropout can help defeat overfitting and also strengthen privacy in neural networks. The effectiveness of membership inference attacks in [13] is also mainly due to the overfitting nature of ML models and therefor the defense techniques proposed are designed to increase ML models' generalizability, i.e., prevent them from being overfitted. Their first technique is dropout which is designed for neural network-based classifiers and the second technique is model stacking which is suitable for all ML models, independent of the classifier used to build them.

[8] propose adding noise to the training set or to the model to achieve differential privacy to suppress the information leak. It also identifies that there is a fundamental contention between selecting useful training instances, which bring in additional information, and suppressing their unique influence to protect their privacy. An important step that could be taken here is to automatically identify outliers and drop those not contributing much to the utility of the model.

[18] presents a membership privacy analysis and evaluation system, called MPLens. [14] proposed a new distributed training technique, based on selective stochastic gradient descent. This system lets participants train independently on their own datasets and selectively share small subsets of their models' key parameters during training. This offers an attractive point in the utility/privacy tradeoff space were participants preserve the privacy of their respective data while still benefitting from other participants' models and thus boosting their learning accuracy beyond what is achievable solely on their own inputs.

5 RELATED WORK

Adversarial attacks: Besides membership inference, there exist multiple other types of attacks against ML models. Another major family of attacks against machine learning is adversarial samples [9]. Adversarial samples are generated by perturbing correctly classified inputs to cause Deep neural network to misbehave (e.g., misclassification). In this setting, an attacker adds a controlled amount of noise to a data point for the purpose of fooling a trained ML model to wrongly classify the data point.

Model inversion: Model inversion uses the output of a model applied to a hidden input to infer certain features of its input. For example, in the specific case of pharmacogenetics analyzed in [3], the model captures the correlation between the patient's genotype and the dosage of a certain medicine. This attack learns aggregate statistics of the training data, potentially revealing private information.Another example is, consider a face recognition model: given an image of a face, it returns the probability the input image is of some specific person. Model inversion constructs an image that maximizes the confidence of this classifier on the generated image; it turns out this generated image often looks visually similar to the actual person it is meant to classify. No individual training instances are leaked in this attack, only an aggregate statistic of the training data (e.g., what the average picture of a person looks like).

Model Extraction: Tramèr et al. [17] demonstrated Model Extraction attack on online ML service providers such as BigML and Amazon Machine Learning. The authors presented simple attacks to extract target machine learning models for popular classification problems such as logistic regression, neural networks and decision trees. Attacks presented are strict black box attacks, but could build models locally that are functionally close to target. Also called as model stealing it attempts to extract the parameters from a remote

model, so that the adversary can have their own copy.Later work extended model extraction/model stealing attacks to hyperparameter stealing attacks [20].

Memorization in ML models: Carlini et al. [2] show that a black-box adversary can extract specific numbers that occur in the training data of a generative model, given some prior knowledge about the format (e.g., a credit card number). The authors describes a testing methodology for quantitatively assessing the risk that rare or unique training-data sequences are unintentionally memorized by generative sequence models.

Privacy-Preserving Machine Learning: Along with the attacks described above, there has been a large amount of effort spent on training private machine learning algorithms. The centerpiece of these defenses is often differential privacy. Both [12] and [6] investigate differential privacy as a mitigation technique for membership inference attacks. Both indicate that existing differential privacy techniques do not display viable accuracy and privacy protection trade-offs.

6 CONCLUSION

Despite their high accuracy and performance, machine learning algorithms have been found vulnerable to different types of attacks. The threat becomes more severe and dominant when the applications operate on private sensitive data. Since Shokri et al.[15] proposed that machine learning algorithms are vulnerable to membership inference attacks, researchers have conducted a large number of studies on membership inference attacks and defense methods and produced good results.

REFERENCES

[1] M. Backes, P. Berrang, M. Humbert, and P. Manoharan. 2016. Membership Privacy in MicroRNA-based Studies. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, Article 7, 319–330 pages.

[2] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. In *28th USENIX Security Symposium*. 267–284.

[3] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. 2014. Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing. In *23rd USENIX Security Symposium*.

[4] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. 2019. LOGAN: Membership inference attacks against generative models. In *Privacy Enhancing Technologies (PoPETs)*.

[5] Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. 2019. Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models. In *Privacy Enhancing Technologies*. 232—-249.

[6] Bargav Jayaraman and David Evans. 2019. Evaluating differentially private machine learning in practice. In *28th USENIX Security Symposium (USENIX Security 19)*.

[7] Yunhui Long, Vincent Bindschaedler, and Carl A. Gunter. [n.d.]. Towards Measuring Membership Privacy. ([n. d.]).

[8] Yunhui Long, Vincent Bindschaedler, Lei Wang, Diyue Bu, Xiaofeng Wang, Haixu Tang, Carl A. Gunter, and Kai Chen. [n.d.]. Understanding Membership Inferences on Well- Generalized Learning Models. ([n. d.]).

[9] Shiqing Ma, Yingqi Liu, Guanhong Tao, Wen-Chuan Lee, and Xiangyu Zhang. 2019. NIC: Detecting Adversarial Samples with Neural Network Invariant Checking. In *Network and Distributed Systems Security (NDSS) Symposium 2019*. https://dx.doi.org/10.14722/ndss.2019.23415

[10] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2019. Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, San Francisco, CA.

[11] A. Pyrgelis, C. Troncoso, and E. D. Cristofaro. 2018. Knock Knock, Who's There? Membership Inference on Aggregate Location Data. In *Proceedings of the 2018 Network and Distributed System Security Symposium (NDSS)*. Internet Society.

[12] Md Atiqur Rahman, Tanzila Rahman, Robert Laganiere, Noman Mohammed, and Yang Wang. [n.d.]. Membership Inference Attack against Differentially Private Deep Learning Model. *TRANSACTIONS ON DATA PRIVACY* 11 ([n. d.]).

[13] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. 2019. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. *Network and Distributed Systems Security (NDSS) Symposium* (Feb. 2019). https://dx.doi.org/10.14722/ndss.2019.23119

[14] Reza Shokri and Vitaly Shmatikov. 2015. Privacy-Preserving Deep Learning. In *22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM, 1310–1321.

[15] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. 2017. Membership Inference Attacks Against Machine Learning Models. In *Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 3–18.

[16] Congzheng Song and Vitaly Shmatikov. 2019. Auditing Data Provenance in Text-Generation Models. In *25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*. ACM, 196–206.

[17] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. 2016. Stealing Machine Learning Models via Prediction APIs. In *USENIX Security Symposium*. 601—-618.

[18] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Wenqi Wei, and Lei Yu. 2019. Effects of Differential Privacy and Data Skewness on Membership Inference Vulnerability. (Nov. 2019). https://arxiv.org/pdf/1911.09777.pdf

[19] Stacey Truex, Ling Liu, Mehmet Emre Gursoyand Lei Yu, and Wenqi Wei. [n.d.]. Towards demystifying membership inference attacks. ([n. d.]).

[20] Binghui Wang and Neil Zhenqiang Gong. 2018. Stealing hyperparameters in machine learning. In *39th IEEE Symposium on Security and Privacy*. IEEE.

[21] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. In *2018 IEEE Computer Security Foundations Symposium (CSF)*. IEEE.

[22] Hanan Makki Zakari, Minhua Ma, and David Simmons. [n.d.]. A Review of Serious Games for Children with Autism Spectrum Disorders (ASD). *International Conference on Serious Games Development and Applications* ([n. d.]).