Introduction to Membership Inference Attacks

Lukas Gehrke
lg58weky@studserv.uni-leipzig.de
Leipzig University

March 31, 2021

Table of Contents

Abstract. With more and more companies offering Machine Learning as a Service (MLaaS) a novel threat of data breaches has emerged: Membership Inference Attacks aim at identifying the fact that given data instances were among the training data of a machine learning model available to an adversary. This knowledge might be of fatal consequences, if the membership exposes sensitive information about people, such as a disease or financial dept. This paper gives a general introduction about membership inference attacks. After discussing enabling factors and underlying theory, core studies as well as mitigation strategies are surveyed, followed by a discussion.

Keywords: Membership Inference Attacks · Machine Learning Security · Machine Learning as a Service

1 Introduction

1.1 Machine Learning

With computers, automation is one of the most simple yet impactful aims that can be achieved. However, teaching computers to solve complex tasks automatically is hard, as it would require huge amounts of instructions and decision rules for the computer. In order to avoid that, computers are programmed to "learn" the rules on their own using Machine Learning models. A model's learning process is based on experience [1], which again comes from **data**, called *training data*. If with experience, that is to say appropriate data, a model gets better at its task, it is said to have successfully "learned".

1.2 Machine Learning as a Service

Thanks to vast scientific efforts, Machine learning is applicable for many use cases nowadays. These range from simple classification of objects to complicated recognition of signals, such as visual or audio detection. Consequently, numerous possibilities for companies have emerged to improve their products, services or internal processes. However, infrastructure for machine learning is expensive, as huge amounts of data have to be stored and transferred and many calculations have to be conducted. To address these problems, the market of **Machine Learning as a Service** (MLaaS) has emerged [2]. The idea behind MLaaS is similar to Software as a Service: A customer has access to remote hardware and software running on this hardware, so that the customer has no need to own them. Examples of companies offering MLaaS are Amazon, Google, IBM or Microsoft. These companies already need the infrastructure and knowledge to process huge amounts of data for their own services. With MLaaS, they gain additional profit by renting hardware that is not fully occupied.

1.3 Membership Inference

With ready-to-use machine learning models being used and often being accessible online, a new kind of privacy risk has evolved: The *Membership Inference Attack*.

Membership Inference asks the simple question, whether a given data record was used to train a machine learning model in question [3]. This leads to critical privacy breaches, if membership in the training dataset allows to imply certain facts. For example, when considering a model that uses cellular structure to recommend ideal medicine dosage for a disease, membership of a person's record in the training data reveals that the person is suffering from this very disease.

This report first explains the theoretical foundation of membership inference attacks. Afterwards, already conducted studies about membership inference attacks are surveyed. Finally, mitigation strategies are discussed, followed by a discussion.

The explanation of theoretical foundations of a membership inference attack are mainly based on the publication by Shokri et al. [3], who first defined the attack concept. They introduced an architecture based on an attack machine learning model and shadow machine learning models which imitate the behavior of the target model and create training data for the attack model.

2 Membership Inference Attacks

The idea behind membership inference attacks is to make use of the fact that machine learning models often exploit different behaviors on previously unseen data in comparison to the data they were trained on[1] [3]. The main reason for this phenomenon is *Overfitting*. Overfitting describes the tendency of machine learning models or statistical classification models to perform significantly better on their training data in comparison to data they have not been trained on [4]. That is to say, the model fails to generalize to data records apart from its training data. Often, this effect is caused by training data with low degree of diversity. The model then does not to learn enough information about the underlying population of the data, as its learning sample is not representative. Model selection also influences the success of membership inference: Some models *remember* more details about their training data, leading to more leakage of information.

2.1 Formal Definition

Prerequesits In this report, **Membership Inference** is defined as the (theoretical) *question*, whether a given data record was part of the training data of a given, ready-to-use machine learning model. A **Membership Inference Attack** describes the (theoretical) *act* of performing membership inference against a machine learning model the adversary does not know the training data for and which - if successful - leads to potentially harmful privacy breaches.

[1] In the following sections, these two sets of data will be referred to as *testing* and *training data*.

Machine Learning Models As introduced by Shokri et al. [3], the basic membership inference attack setup consists of three machine learning models:

1. The **target model** is the model under attack. The attacker has access to the target model but does not know its exact training dataset.
2. The **attack model** performs membership inference by deciding, whether the attack model's output using a given record as input, was part of the training dataset of the attack model.
3. The **shadow models** imitate the target model as good as possible. They are used to teach the different reactions of the target model to seen and unseen data to the attack model.

Attack Definition Let $f_{target}()$ be the target model and $f_{attack}()$ be the attack model. D_{train}^{target} is the training dataset of $f_{target}()$, consisting of data records $(\mathbf{x}^i, y^i)$, with $|D_{train}^{target}| = n$, $i \in (1, \ldots, n)$. Here, $\mathbf{x}^i$ is a vector describing an entity and y^i refers to the class of the entity with respect to the task of $f_{target}()$. Assume the number of possible classes for the target model is c_{target}. The output of the target model is a vector of values of size c_{target} with each value being in $[0, 1]$ and indicating the probability that the entity in question belongs to the respective class. The vector is denoted by $\mathbf{y}_{train}^i$.

The attack model f_{attack} uses labeled data entries and the respective outputs generated by the target model for these entries as input x_{attack}. Its purpose is to recognize whether a data entry was part of the training dataset based on these inputs. This task leads to a binary output scheme: "*in*" for data records of the training dataset and "*out*" for non-members. The attack procedure is as follows: For a labeled data record $(\mathbf{x}, y)$ the output of $f_{target}()$ is generated: $f_{target}(\mathbf{x}) = \mathbf{y}$. Afterwards, the tuple $(y, \mathbf{y})$ is queried to the attack model. Ideally, the attack model then recognizes patterns in distribution of $\mathbf{y}$ around the actual y [3]. The attack model then calculates the pobability $P(((\mathbf{x}, y) \in D_{train}^{target})$ that the record $(\mathbf{x}, y)$ was part of the training dataset of the target function or belongs to the "*in*" class. Figure 1 gives an overview over the attack procedure. Note that it would also be possible to use only $\mathbf{y}$ or to use $((\mathbf{x}), \mathbf{y})$ as input for the target model. In the latter case the model could identify patterns in the relationship between $\mathbf{x}$ and $\mathbf{y}$.

Attack Model Training One fundamental part of the aforementioned process remains open: The training of an attack model. Therefore, Shokri et al. [3] propose a technique called "*shadow training*". Multiple shadow models $f_{shadow_i}()$ are generated with the main aim being to make their behavior as similar to the target model's behavior as possible. The training datasets $D_{train}^{shadow_i}$ are all known and may partially overlap. Training data for the shadow models may originate from existing sources. Otherwise data records have to be generated as "*synthetic*" training data. The idea of using synthetic training data is based on empirical evidence following the idea that similar machine learning models provided by the same service and trained on similar data behave similarly [3].

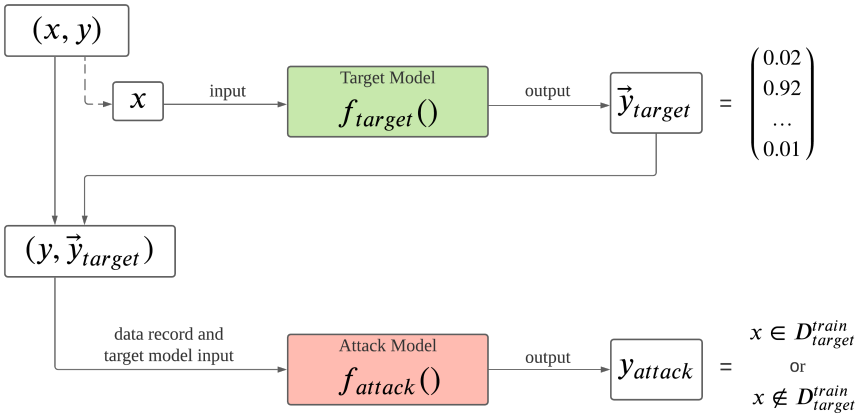

Fig. 1. The black-box Membership Inference Attack scenario. A data record $(\mathbf{x}, y)$ is first queried to the target model. Afterwards, the output and of the target model and the true class of the data record are used as input for the attack model. The attack model decides, whether the data record was in the training dataset of the target model or not.

Moreover, Shokri et al. claim that a higher amount of shadow models provides more training experience for the attack model and thus enhances the attack accuracy. To generate training data for the shadow models, Shokri et al. [3] propose different methods:

- **Model-based Synthesis** uses the target model for the generation of synthetic training data. It is based on the idea that if the target model classifies a randomly generated data record with high confidence, that is to say with a value close to 1 for the predicted class and with values close to zero for all other classes, the data record is similar to a real entry from the target models training data.
- **Statistics-based Synthesis** bases generation of synthetic data on statistical knowledge about the underlying population of the training dataset of the target model.
- **Noisy real data**: If real data records of the training dataset of the target model are accessible, these can be used to generate new records by changing a selected number of random properties of the records at hand.

The Model-based synthesis as introduced by Shokri et al. [3] uses a two-phase process: First, records that are classified by the target model with high confidence are searched. This procedure is iterative and involves randomly initializing data records. Afterwards, features are changed until the prediction confidence by the target model for a fixed class surpasses the confidence for all other classes and a threshold. Additionally, already accepted synthetic records are used as a base for new ones. In each iteration, a new record is created by randomly changing

features of the last accepted record. The second phase consists of sampling new records for the synthetic datasets if the acceptance criteria from phase one are met.

To train the attack model, each shadow model $f_{shadow_i}()$ has to classify all records of its respective training dataset $D^{train}_{shadow_i}$ and test dataset $D^{test}_{shadow_i}$. Given a record $(\mathbf{x}, y)$ from a training set, the output $f_{shadow_i}(\mathbf{x}) = \mathbf{y}$ will be extended with the label "*in*". Thus, the tuple (y, y, in) will be added to the attack model training dataset D^{train}_{attack}. Note that y, the true label, is added to enable the attack model to learn the distribution of $\mathbf{y}$ around y as mentioned earlier. For all entries of $D^{test}_{shadow_i}$, tuples of the form (y, y, out) are added to D^{train}_{attack}. Afterwards, D^{train}_{attack} is split into c_{target} partitions, one for each value of y, the true class label. Then, according to Shokri et al. [3], for every class y_i a separate attack model f_{attack_i} should be trained. Figure 1 illustrates the attack model training process. Ultimately, the resulting task is a binary classification task. This is a standard task in machine learning, allowing the attack model to be build with any machine learning framework. Within this binary classification task, the attack model learns how to keep apart records that belong to the "*in*" or "*out*" group and where *both classified with high confidence by the target model*, as Shokri et al. [3] highlight. A task much subtler than only distinguishing between records classified with high or low confidence, which is not what the attack model learns [3].

2.2 Scientific Studies

In this chapter, studies about membership inference attacks are presented. The chapter begins with an overview of most notable scientific publications. Afterwards, key studies are reviewed in detail. As many publications include a chapter about mitigation strategies, these are also mentioned in the overview. Mitigation strategies are discussed later on.

Overview of Studies Several studies about membership inference attacks have been conducted. Shokri et al. [3] use six different datasets and 4 different target models to explore the possible threats of membership inference attacks. They find out that machine learning models indeed leak information about their training data and show the success of their shadow model training technique. They also discuss multiple mitigation techniques such as regularization or restriction of the prediction vector. Truex et al. [5] directly tie on this work by developing a framework for membership inference attacks. They survey attacks against different types of machine learning models and explore the success factors of membership inference attacks. A study of mitigation strategies such as model hardening and differential privacy is also conducted, revealing need for more studies on this matter. Hayes et al. [6] focus on attacks against generative machine learning models, a version of neural networks that is very popular in many applications nowadays. Generative adversarial networks use a combination of a discriminative and a generative model and are mostly used to generate images

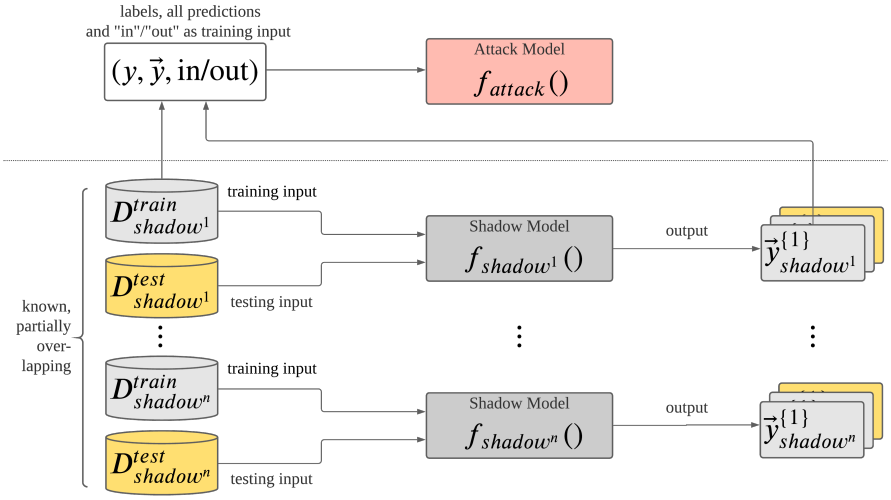

Fig. 2. The attack model training process involves n shadow model with partially overlapping training and testing datasets.For every shadow models, all prediction results are obtained and additionally, each result is flagged with "*in*" if the respective record was in the training dataset of the shadow model and "*out*" otherwise. Resulting tuples are extended with a records true class y and then used as training data for the attack model.

or videos [7]. Hayes et al. show that membership inference attacks work better against discriminative models than against generative models. Moreover, they experiment with regularization techniques as mitigation strategies and report mixed success as well as need for more research on that topic. Salem et al. [8] study applicability of membership inference attacks by gradually reducing the assumptions on the attack model. They experiment with a lower amount of shadow models and attack training data generation through so-called data transferring attacks. They also explore means of reduction of overfitting to oppose membership inference attacks. Jia et al. [9] develop a defense framework against membership inference attacks. The framework enables adding noise to output vectors of potential target models in order to confuse attack models.

Original Studies Shokri et al. [3] use six datasets to explore the possibilities of membership inference attacks against different target models. Table 1 gives an overview of the datasets. As target models, they use three MLaaS models and one model that is implemented locally. The models are listed in table 2.

Shokri et al. [3] use precision and recall as well as accuracy as measures of attack success. Precision is defined as the fraction of records the attack model identifies as members of the training dataset that indeed are members. Recall is

Dataset	Task	Features	Entities
UCI Adult	Binary classification	14 features (binary and numerical)	48,842
Purchases	2 to 100 class classification	600 binary features	197,324
CIFAR	10 to 100 class classification	32 x 32 color images	60,000
MNIST	10 class classification	32 x 32 images of handwritten digits	70,000
Locations	128 class classification	446 binary features	11,592
Hospital Stays	100 class classification	6,170 binary features	67,330

Table 1. Overview of the datasets initially used to study membership inference attacks [3]

Model	Parameters
Google Prediction API	No configuration parameters available
Amazon ML version I	maximum number of passes over the training data, L2 regularization amount: $(10, 1e-6)$
Amazon ML version II	maximum number of passes over the training data, L2 regularization amount: $(100, 1e-4)$
Neural networks	Torch7, Convolutional neural network with two convolutions and max pooling layers, fully connected layer of size 128, SoftMax layer

Table 2. Overview of the machine learning models used in the initial publication about membership inference attacks [3]

defined as the fraction of the training dataset that is correctly identified by the attack.

First, the accuracy of the attack is measured. The measurement is based on training and testing datasets of the same size, thus the baseline accuracy is $0,5$. On the neural-network, the authors report a high degree of overfitting when measuring test dataset accuracy with the CIFAR dataset. The train-test accuracy gap on the CIFAR dataset is 0.5 when using Amazon ML version II. Regarding the attack precision, on the CIFAR dataset with ten classes (CIFAR-10), the median precision is 0.72 for a training set size of $10,000$. For the CIFAR-100 dataset, the precision is almost 1, while recall is almost 1 for both scenarios. Against the Google model, the precision is even higher in the same setup. On the Hospital Stays and Locations dataset, Shokri et al. [3] evaluate an attack against the Google model, reaching precision of more than 0.7 for half of the classes and a precision between 0.6 and 0.8 respectively.

Furthermore, the *effect of the different shadow model training techniques* was studied. For the noisy training data, precision is reported as dropping with increasing noise. However, the authors note that with noisy data the attack precision is still above baseline. Therefore they deduce that membership inference attacks even work with vague assumptions about the training data. To evaluate the approach of model-based synthesis on the Purchase dataset, $30,000$ fully synthetic records and $187,300$ marginal-based synthetic records are gen-

erated. The marginal-based approach uses marginal distributions of individual features to generate synthetic data. For the fully synthetic records, on average 156 queries to the target model are needed in order to generate a single entry. Against the Google model with black-box access, the attack precision and recall are 0.935 and 0.994 for real-data, 0.896 and 0.526 for model-based synthetic data as well as 0.795 and 0.991 for the marginal-based synthetic data. The authors mention that the attack model performs very poorly on some classes when being trained on the model-based synthetic data. This effect might be due to the under-representation of some classes in the target-models training dataset, leading to bad synthetic records. Overall, Shokri et al. [3] report that membership inference attacks perform very well even without any knowledge about the data in the training dataset of a black-box-access target model. A prerequisite is the ability of an adversary to access the target model and obtain predictions with high confidence from it.

Additional studies by Shokri et al. focus on the "*effect of the number of classes and training data per class*" and the consequences of *overfitting* for attack success. The authors claim that the more classes, the better a membership inference attack works. A study using the Purchases dataset and the Google model supports this claim. While with 20 classes average precision is around 0.6, with 50 classes average precision rises to almost 0.9. The authors try to explain this effect by analyzing the behavior of the target model. With more classes, the model has to use more separating information from the records to be able to distinguish inputs from the different classes accurately. Thus, the model leaks more information.

Finally, the effect of overfitting is considered with a calculation of the train-test accuracy gap. Overfitting is directly related to a high train-test accuracy gap, because a high gap indicates that the model in question fails to generalize from its training data to the test data and is thus overfitted. Shokri et al. identify a positive correlation between the size of the accuracy gap of a model and the accuracy of the membership inference attack against the same model. However, they also find counter-examples, where a model with a smaller accuracy gap is more vulnerable to membership inference attacks. The authors conclude that overfitting and diversity of training data straightly enable success of membership inference. They thereby identify privacy leakage as another negative effect that overfitting causes for machine learning models next to performance loss.

Further Studies Following the initial paper about membership inference attacks, Truex et al. [5] conduct experiments with the same datasets, Adults, MNIST, CIFAR and Purchases. Additionally, they carefully define the attack setup and process. They highlight that the training dataset of the target model heavily influences the success of an attack: The higher the number of classes and the higher the in-class standard deviation, the higher the accuracy of a membership inference attack. They explain the effect of the number of classes with the underlying space $\mathbb{R}$ of the dataset, that they assume to be more and more divided for higher numbers of classes. This would result in smaller areas

for single classes and the areas would more closely frame the single data records from the associated class. Consequently, the borders between class areas become much and much smaller with all records being closer to the borders. This results in higher influence of single records on the decision borders. Thus, Truex et al. argue, an adversary would identify a data record with higher probability as the model under attack would more likely be "*impact*(ed)" by the instances for its decisions. Regarding the in-class standard deviation, Truex et al. claim that higher standard deviation inside a class favors membership inference attacks. They argue that if records within the same group have huge differences, then the consideration of every single instance will influence the decision structure of the target model. The authors underpin their claims with a comparison of the aforementioned datasets using the respective in-class standard deviation, number of classes and attack accuracy. In general they claim that the higher the complexity of the classification problem, the higher the accuracy of an attack and that success of membership inference attacks is driven by the underlying data of the target model. They add that some factors may be dominant over others, as they find Purchases-100 to result in higher attack accuracy than Purchases-50 while the in-class standard deviation is nearly equal. Additionally, they study the selection of model type for all three models involved in an attack: The target model, the attack model and the model for attack model data generation ("shadow model"). The experiments involved measurement of attack accuracy and standard deviation in attack accuracy with changing models. They reveal that attack success depends on the target model while the other models are interchangeable. Therefore, the authors described membership inference attacks as "*transferable*" [5]. As a consequence, an adversary might create a successful attack model without the knowledge about the best model type for membership inference. Moreover, Truex et al. study the influence of the knowledge of an attacker about the target data, similarly to the experiments by Shokri et al. with noisy data. The receive similar findings: Even with 10% data loss, features replaced by random values ("noise"), attack accuracy remains high. When inspecting the shadow model training data, they find a similar noise resilience. They furthermore suggest that shadow model training datasets should be sufficiently large, as they find some significant increases in performance when carrying out experiments on the Purchases-20 and 50 datasets with varying shadow model training dataset sizes. To sum up, Truex et al. report new insights about membership inference attacks such as that the attacks are mainly influenced or enabled by the target model and the underlying data and that attack and attack data generation models are interchangeable.

Hayes et al. [6] focus on membership inference attacks against generative machine learning models. They again use CIFAR-10 as well as Labeled Faces in the Wild (LFW), a dataset consisting of images with faces and Diabetic Retinography (DR), consisting of images of retinas. The authors perform white-box and black-box attacks against three different generative adversarial networks (GAN) with one fixed GAN as attack model. They also used a black-box setting with limited auxiliary knowledge: Here, an adversary has knowledge of about

$20-30\%$ of the training or testing set of the target model. The black-box attack without and with auxiliary knowledge on the LFW dataset yields an accuracy of 0.4 and 0.63 respectively. With the CIFAR-10 dataset, accuracies 0.37 and 0.58 respectively are reached, while 0.22 and 0.81 are the results on DR. Note that random guessing yielded 0.1 accuracy on LFW and CIFAR-10 and 0.2 on DR. Additionally, Hayes et al. [6] underpin the finding of Truex et al. [5] that different attack models yield similar attack performance. Moreover, the authors claim that attack accuracy is directly correlated to with the quality of generated training data for the attack model. They report that the target model, if used for data generation, generates better data than an attack data generation model and that in general better training data is created after more training epochs of the models. All in all, Hayes et al. show that membership inference attacks against GANs are possible and discover techniques to improve attack performance.

Salem et al. [8] simplify membership inference attacks by removing multiple requirements towards the adversary. First, they perform membership inference using only one shadow model in contrast to Shokri et al. [3] using the same datasets. Notably, their attack with only one shadow model yields almost equal performance regarding precision and recall on all datasets. They also study the need of knowledge about the structure of the target model. By changing the parameters of the target neural network model and performing an attack using the Purchase-100 dataset, they receive similar results despite changed parameters. Therefore, the authors claim that knowledge about the hyperparameters of a target model is not required to perform the attack. To test the influence of knowledge about the classification algorithm of the target model, Salem et al. created a shadow model that combines different algorithms. Thereby, the behaviors of all algorithm can be taught to an attack model with one shadow model. They combine logistic regression, multilayer perceptron and random forests and evaluate the combined model on the Purchase-100 dataset against each of these algorithms alone as target model. Moreover, they use a single shadow model against each target model, always with the same algorithm as the target, for comparison. The results indicate that the combined model performs similarly to the respective single models, revealing only a little skew for random forests as target algorithm. A second study by Salem et al. introduces a novel technique for synthetic attack data generation. Here, the shadow model generates training data for the attack model independently from the distribution of the target models training data. The underlying idea is to only teach the attack model general knowledge about membership and non-membership in a training dataset of a target machine learning model. In their experiment, the authors randomly pick on of the datasets for attack data generation and one as target training dataset and overall receive precision of more than 0.7 and similar rec all, although there are some outliers. The authors claim that this procedure also simplifies generation of attack training data, as huge amounts of queries against the target model become unnecessary. The third approach by Salem et al. focusses on minimizing all requirements to the attack model by completely withdrawing shadow models and using unsupervised binary classification [8]. The authors claim that mem-

bership inference with solid accuracy is even possible in this setting. Overall, Salem et al. present some effective simplifications for membership inference attacks such as the reduction of the amount of shadow models or the usage of training data for attack data generation, that is completely unrelated to the target dataset.

2.3 Mitigation Strategies

In order to prevent membership inference attacks, several studies share their insights about the best mitigation strategies. Shokri et al. [3] discuss four techniques: They (1) leave out classes with low probability from the prediction vector, (2) round the values in the prediction vector, (2) increase the output vector's entropy and (3) apply regularization. Regularization aims at reducing model complexity, so that the fit of the model to its training data, especially outliers, is not to tight [10]. The authors find their membership inference attacks to withstand most of their mitigation attempts. For example, leaving out classes of the prediction vector, even reducing the number of output classes to 1, the class with the highest probability does not decrease attack accuracy significantly. Shokri et al. report regularization as the only useful technique but at the same time warn to not over-use it to prevent the potential target model from losing its prediction performance. Truex et al. [5] present "*model hardening*" and "*API hardening*" to prevent membership inference attacks. In case of model hardening, they suggest (1) choosing a machine learning model or algorithm that is resilient against membership inference, (2) controlling their model while training in order to prevent overfitting, e.g. by using certain model parameters, (3) applying regularization and (4) apply anonymization techniques. With API hardening, Truex et al. describe techniques that affect the working model in use. Here, they recommend techniques similar to Shokri et al. [3]. Moreover, they recommend usage of differential privacy mechanisms. Hayes et al. [6] study usage of "*Weight Normalization*" and "*Dropout*". They report Dropout to be way more effective than Weight Normalization but also claim that Dropout significantly prolongs the training phase. Similarly to Truex et al. they also study differential privacy mechanisms. Salem et al. [8] introduce "*model stacking*" as mitigation strategy in addition to Dropout, which is only applicable to neural networks. This technique uses ensemble learning to split training of the components of a potential target model with different parts of the training dataset. Thereby, Salem et al. want to reduce to degree of overfitting of the model. Their experiments conducted with model stacking revealed reductions of attack performance on different datasets of up to 30% and more. To sum up, different privacy preserving techniques are researched, ranging from simple reduction of model output complexity over implementation of new model training procedures up to privacy systems such as differential privacy.

3 Conclusion

3.1 Summary

Machine Learning offers many possibilities to create new services or simplify existing ones but also has high requirements regarding hardware, software and know-how. With Machine Learning as a Service, the possibilities of machine learning have become accessible for a broad range of institutions. However, large amounts of data underlie these systems and are the basis for successful learning. Often, these data are privacy-sensitive. MLaaS and the potential privacy-sensitivity of underlying data have caused a new threat for privacy: Membership Inference Attacks - the identification of the fact that a given data record was used to train a machine learning model at hand.

This report gives a general introduction about the developments that made membership inference attacks possible, the theoretical process of performing an attack with an attack model and shadow models as well as an overview of key studies about membership inference attacks and related defense strategies.

3.2 Discussion

Membership inference attacks offer an interesting field of studies. In some studies, single questions remain open. Shokri et al. [3] claim that the more shadow models are used for attack data generation, the more accurate the attack model will be. However, they provide no empirical evidence for their claim. Moreover, Salem et al. [8] show that a single shadow model offers similar performance in comparison to multiple shadow models. Therefore, the aforementioned un-tested claim in the study by Shokri et al. is simply wrong. The mistake could have been avoided by doing an additional experiment on the number of shadow models.

The discussion of the effect of the number of classes as explained by Truex et al. [5] seems to link to the "curse of dimensionality" defined by Bellman [11]. The curse of dimensionality states that with higher amount of dimensions, a latent space gets larger, increasingly empty and the instances in it gather close together. Following their argumentation, higher numbers of dimensionality of the training dataset under attack should cause higher vulnerability for membership inference attacks, as the instances would get closer together and influence the decision processes of the target model [5]. Truex et al. state that in class standard deviation directly affects membership inference attack success but do not give real empirical evidence for that claim in their publication. In the corresponding table, they even have an example that opposes their hypothesis with CIFAR-10 having lower in-class standard deviation than Purchases-10 but higher attack accuracy. Both examples have 10 classes. In general, Truex et al. [5] seem to replicate or at most refine lot of the work of Shokri et al. [3] such as a giving formal definition of membership inference attacks and their order of events with shadow models which they call "*(...) the first generalized framework for the development of a membership inference attack model*". This is a critical claim, as Shokri et al.

[3] introduced a very similar architecture before, though never called it "*framework*". Additionally, Truex et al. [5] do not reference Shokri et al. [3] for first finding out that the number of classes is positively correlated with the accuracy of a membership inference attack. Instead, they avoid to directly describe the phenomenon in their publication (chapter 4.2) although it is observable in the table in chapter 4.2.

3.3 Outlook

With new machine learning models and refinements of already available models being published all the time, membership inference attacks remain an important topic. Every new model is a potential target model and a potential attack model. Therefore, studies about frameworks for quick checks of vulnerability of new models would be an interesting topic for further research. In general, the study of mitigation techniques remains important, as the aforementioned studies show, that many mitigation ideas fail to prevent membership inference attacks. Additionally, topics like the influence of the dimensionality of the data under attack on the attack performance could be addressed in future work.

References

[1] Thomas M. Mitchell. *Machine Learning*. 1st ed. USA: McGraw-Hill, Inc., 1997. ISBN: 0070428077.

[2] Mauro Ribeiro, Katarina Grolinger, and Miriam AM Capretz. "Mlaas: Machine learning as a service". In: *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. IEEE. 2015, pp. 896–902.

[3] Reza Shokri et al. "Membership inference attacks against machine learning models". In: *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2017, pp. 3–18.

[4] Andreas Weigend. "On overfitting and the effective number of hidden units". In: *Proceedings of the 1993 connectionist models summer school*. Vol. 1. 1994, pp. 335–342.

[5] Stacey Truex et al. "Towards demystifying membership inference attacks". In: *arXiv preprint arXiv:1807.09173* (2018).

[6] Jamie Hayes et al. "Logan: Membership inference attacks against generative models". In: *arXiv preprint arXiv:1705.07663* (2017).

[7] Ian J Goodfellow et al. "Generative adversarial networks". In: *arXiv preprint arXiv:1406.2661* (2014).

[8] Ahmed Salem et al. "Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models". In: *arXiv preprint arXiv:1806.01246* (2018).

[9] Jinyuan Jia et al. "Memguard: Defending against black-box membership inference attacks via adversarial examples". In: *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. 2019, pp. 259–274.

[10] Andrew Y Ng. "Feature selection, L 1 vs. L 2 regularization, and rotational invariance". In: *Proceedings of the twenty-first international conference on Machine learning*. 2004, p. 78.

[11] Richard Bellman. *Dynamic Programming*. Dover Publications, 1957. ISBN: 9780486428093.