Summary of the Paper "Membership Inference Attacks Against Machine Learning Models"

anonymous authors

1 Introduction

This paper (written by R. Shokri, M. Stronati, C. Song and V. Shmatikov) investigates the leak of membership information through attacks on prediction outputs of machine learning models. On the basis of data records and a black-box access, the authors try to examine whether a certain record is part of the original training dataset or not.

Machine learning is widely used by companies for example to improve marketing or to conduct individual advertising. Providers like Amazon or Google offer machine-learning-as-a-service which is mostly available as a black-box API with hidden processes and algorithms. Therefore costumers can use their own training data and produce predictive data as a outcome (as seen in Figure 1). With this scenario being a difficult setting, the authors approach the question of membership inference. To proof the existing threat, especially in terms of sensitive data, the authors develop a so-called shadow training. It consists of a shadow model, with knowledge of the training data set and an attack model which is trained to classify labeled inputs and outputs. Given this, the authors turn the membership inference problem into a classification problem.

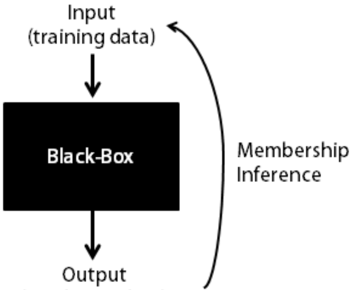

Figure 1: Machine-learning-as-a-service with the problem of membership inference

The authors rely on previous research concerning the leakage of training records but are the first to try several methods in a black-box setting. The research question is also reviewed in terms of privacy.

2 Machine Learning Background

This paper is focused on machine learning which is supervised, using training data records as input and labels as an output. With this they try to understand the relationship between the input and output by creating a model classify the results. In addition there are companies which provide machine learning as a service.

Those are mainly cloud services APIs where customers can use a platform to train a model but are not able to see the algorithms which are used for example Amamzon Machinen Learning. Also the comapnies don't reveal what they actually do with the uploaded data from the customer and are using those data to make profit with other companies. This paper is going to use machine learning as a service as a black-box.

3 Privacy in Machine Learning

When it comes up to privacy in context of machine learning the author addresses two topics. For once the inference about members of the population and the inference about members of the training dataset.

The author says that there are privacy rules in statistical disclosures which says that the input data of the members should not display more after being applied than when they are not applied but there is a gap. When a model is based on statistical facts it is possible for competitors to infer the unintentionally input values from the output values. This applies to the members of an entire population.

4 Problem Statement

The author investigates two scenarios where a classification model gets attacked. For one the attacker pretends to be the machine learning platform which teaches the model how to gain a connection between the content of a data record and a label and second the attacker has query access to the model with the help of an black-box and second

In the first case the attacker knows the structure of the model and the meta-parameters for the second case not. The goal of the membership inference to figure out if a data record is part of a the target models training set.

5 Membership Inference

There are inconsistencies in the behaviors of the machine learning models with the trained data and the actually used data which an attacker has to consider while creating an attack model. The created attack model is trained to find those inconsistencies and is able to differ between members and non-members. Before using the attack model the attacker is training shadow models. Each output class has its own shadow model which makes sure to give precise results and to understand the behavior. Shadow models use datasets which has the same structure and exude similar to the target model dataset which was used to train the target model. There are three methods the author describes in this papers to generate a dataset. The first one is the mode-based synthesis where the attacker has no informationen. He is creating synthetic training data by using the target model. This method has two steps search, where an algorithm is searching for data which are classified as high confidence and sample, where those synthetic data are stored. In the first step the attacker has to determine the class c which is the class where the attacker wants to generate synthetic data from. After that the attacker generates a record x and sample the values and classify them. In the second part those proposed data records probabilities are compared to the other data records probabilities to make sure that the labels and the prognosis are right.

The second method is statistics-based synthesis where the attacker got some information. The mentioned that the were sampling values and that this attack was very effectiv.

The third method is the noisy real data where the attacker has some information but the target model's training data are from another population than the training data for the shadow model.

After Training the shadow models there are two outputs label "in" and "out" which now has to be added to the training set of the attack model.

6 Evaluation

To proof their thesis, the authors chose several realistic scenarios for membership inference attacks on machine learning models. Not only did they test two popular cloud-based services from Google and Amazon, they also setup a local neural network based on widely-used frameworks. Their testing datasets also consisted of publicly available information. CIFAR-10 and CIFAR-100 are benchmark datasets for object recognition in images. MNIST uses images of normalized handwritten digits to identify text. Kaggle contains real data from online-shops. In addition, the authors also chose to use location-based data from a Bangkok social network, patient data from a Texas hospital and American census data.
The characteristics of the machine learning as a service providers are, that they provide very little to none options to interfere with the learning process of the model nor do they grant access to the model itself. The only way to access these models is via the API they provide. Whereas the local neural network gives room for experiments to see how changes in the underlying parameters affect the effectiveness of the membership inference attack.

For every case a target model, several shadow models and an attack model where created. Each target and shadow model had to be trained with specific data and in a second step it was tested with different data. These training and test sets contained randomly selected data from the before mentioned sources and are the same size. Also, the training and test sets of the target model and the corresponding shadow models share no data, are therefore disjoint to create a harder setup.

To measure the effectiveness of an attack, each target was confronted with an evenly shuffled dataset from the training and test data, which contained equal numbers of members and non-members in the dataset. This represents a baseline accuracy of 0.5 and could also be achieved by blind guessing. To compare the impact of the different factors, several testing environments for the underlying data where created. In this way, the authors could not only measure the information leakage from different cloud-based platforms or local neural networks, but they also found the affect the data size itself has and whether or not it was classified into little or many classes.

One of the findings is, that even when the attacker has inaccurate assumption about the real data's distribution, synthetic data could lead to robust precision in the attack reaching up to 89% for the purchasing data in the Google cloud. Also, the more output classes a target model has, the more information of the underlying training data has to be remembered. This leads to the result that more classification let the target model leak more information and therefore making the attack more efficient. One problem in the training process of a model is overfitting. When a model is overfitted, it resembles the training data too well, so that it has negatively impact on new data. One of the findings is, that overfitting is also a cause which makes a model vulnerable to an attack.

Overall the authors could create efficient attacks in all environments and constellations. This means, that a membership inference attack can be accomplished with very little knowledge of the target model or its training data.

7 Mitigation Strategies

The success of the membership inference is defined and related to the diversity of the training data and the generalizability of the target model. The authors introduce several mitigation strategies. It seems that the accuracy is generally better with fewer classes. To prevent membership inference the authors suggest to restrict the prediction vector of the top k classes as less information is leaked the smaller k is. Additional to that d should be kept small and make a coarsen precision of the prediction vector. Furthermore the use of regularizations and the increase of the entropy of the prediction vector is recommended.

To evaluate the suggested mitigation strategies the authors implemented all of them in fully controlled models. The attack was still based on the black-box setting. The results of the mitigation testing show, that even restricting the prediction vector to a single label does not fully prevent membership inference. The attack still seems to be robust against the suggested strategies. Nevertheless regularization approaches are useful and can even increase the prediction accuracy of the target model.

8 Conclusion

The results of the attacks are interesting for both machine learning and privacy research. The attack like the one created by the authors in form of a shadow training could be used to review the machine learning services and improve them. It can be created effectively by using synthetic or noisy data. Providers like Google or Amazon should be aware of the possible risks their software carries and should inform potential users and costumers.

9 Sources

R. Shokri, M. Stronati, C. Song, V. Shmatikov. *Membership Inference Attakcs Against Machine Learning Models.* Security and Privacy (SP), 2017 IEEE Symposium. May 2017