Privacy and Big Data: Reviews 1 [H00Y3a]

Mariya Hendriksen (MAI, r0690989)

November 25, 2018

Contents

1 Introduction

For the second assignment, we are to present a critical review of one of the two papers and give a summary and ask 3-4 question for the other paper.

2 Critical Review

In this part of the assignment, we give a critical review of the paper 'Membership Inference Attacks Against Machine Learning Models' written by Reza Shokri, Marco Stronati, Congzheng Song and Vitaly Shmatikov [1]. Below, summary, evaluation and synthesis of the paper are presented.

2.1 Summary

Machine learning is getting increasingly popular over the last years, the models are being trained on different types of data. Since some time ago, tech giants like Google and Amazon have even started to offer "machine learning as a service": a customer can upload a dataset and a task on the platform and receive a black box model in return. In connection to it, Shokri et al. address the problem of **membership inference** in machine learning, i.e., a possibility to

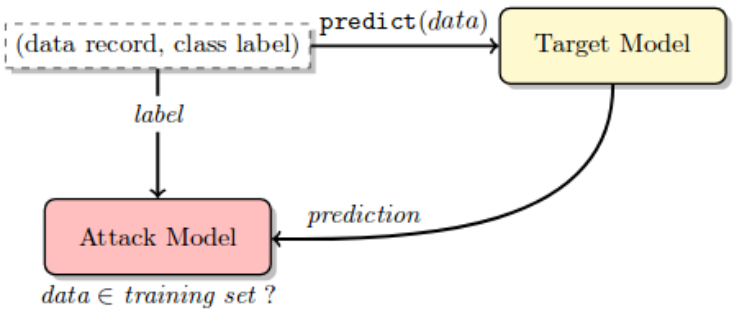

Figure 1: End-to-end membership inference attack process [1]

determine whether a given data point was part of the training set of the model. Membership inference allows an adversary to infer private information about the users whose data was used for the model training. Therefore, the aim of the paper is to quantify the data leakage which results from conducting the membership inference attack.

The main contribution of the authors is designing an attack model for membership inference. This model is essentially a binary classifier which goal is to determine if the datapoint belonged to the training set of the target model or not. The attack model is a set of models where each model is a binary classifier for every output class of the target model.

Shokri et al. choose to perform the attack on a black box model as they believe it to be the most difficult case from the adversary's perspective since the target model parameters are unknown. The possibility to conduct the attack lies on the assumption that target model behaves differently when encountering a datapoint which was used during training.

To train the attack model, the authors suggest **shadow training**, a method they invented specifically for the purpose. The idea is to build a bunch of shadow models which are to reproduce the behaviour of the target model. In case of using machine learning API, the shadow models are trained on the same platform as the target model. The training of these models is supervised, i.e., it is completely clear whether a fed datapoint belongs to a dataset or not.

The shadow models can be trained on synthetic or noisy data. The generated data should have the same distribution as the training data of the target model. To achieve this, Shokri et al. suggest three novel approaches on how to generate training data for the shadow models:

- model-based synthesis: this method is based on the idea that the target model classifies with a high confidence the data points which are closer to the distribution of its training dataset
- statistics-based synthesis: the attacker leverages the statistical information about the population from which the training set was drawn

- real noisy data: assume access to a possibly noisy version of the target dataset, the attacker is allowed to query the target model only once

To obtain a training set for the attack model, the shadow models are queried. Each model is fed datapoints of two types: 1) data points used during training (the shadow model output is labelled as 'in'), 2) data points not used during training (the shadow model output is labelled as 'out'). The attack model is trained on the labelled inputs and outputs of the shadow models.

The code for the attack experiment is available on GitHub[2].

To evaluate the inference membership attack, the authors used CIFAR dataset for image recognition, purchases dataset from a Kaggle competition, location dataset from Foursquare, Texas hospital stays, MNIST and census income dataset. Shokri et al. also choose three black box models as their targets: two models constructed with machine learning API (namely, Amazon ML[3] and Google Prediction API[1]), the last model built locally. The authors train the attack model and evaluate their accuracy, the lowest test accuracy being 46.8%, the highest – 67%.

The experiments with the noisy data allow to us conclude that even when the assumptions about the target distribution are rather distorted, the attack model performance is still robust. Furthermore, the results achieved with the models trained on the synthetic data brings authors to the conclusion that the there is no need to access the actual target distribution for carrying out a successful attack. What is important for the attack model is to generate data which is classified by a target model with high confidence.

In general, models with a bigger number of classes and overfitting are more likely to leak data. At the same time, a number of examples per class and regularization help to decrease privacy breach.

There are some mitigation strategies to prevent security breach. One of the examples is restricting the prediction vector to the top n classes. In the most extreme case, the model is to output one class without its probability. The other strategies include lowering the precision of the vector of predictions, an increase of entropy of the output vector and application of regularization. However, in all the cases the attack model is still capable to produce the information leakage even though to a lesser extent.

Overall, the success of the attack models depends on the capability of the target model to generalize and on a diversity of its training dataset. Since those are also the characteristics the ideal machine learning model should possess, the attack model can be used as one of the metrics for evaluation of machine learning models.

2.2 Evaluation

Shokri et al. tackle the problem of membership inference in machine learning. This problem is becoming more significant as the popularity of the models increase. The authors offer an attack technique to detect data leakage of the

[1] shut down on April 2018[4], replaced by Cloud Machine Learning Engine[5]

target model and run a series of experiments to prove its applicability. The results of their experiments prove that the attack model can perform considerably well even though it is trained on noisy or synthesized data. The attack technique can serve as a metric for measuring data leakage of a target model. Besides, Shokri et al. investigated the mitigation strategies for decreasing the information leakage and test their attack on them. Their experiments allow suggesting that mitigation strategies can dimish information leakage, but cannot completely avoid it.

Unlike the prior work in the field of attack on machine learning models, the authors target black box models which parameters are unknown. The shadow attack technique is also different from the model inversion presented in related work in the way that the latter does not inevitably imply the information leakage. Moreover, model extraction attacks are different in a way that they are to identify the parameters of the trained model. Even though it can be considered as a first step of the shadow training attack, the ultimate goal of the latter is to infer members of the private training dataset, not the model parameters itself. The previous work done in the field of the privacy-preserving machine learning focuses mainly on how to train models without direct access to the data or on the differential privacy. The results presented in the paper allow suggesting that those techniques will not prevent privacy breach as the model is still capable to leak information and the data needed to train shadow models can be synthesized.

Overall, the authors propose a robust membership inference technique the models for which can be trained on noisy target data or on fully synthesized data.

The authors supported their claims by launching attacks on black box models built with ML API as well as on the locally built model. They trained their models on six different datasets. The achieved accuracy was ranging from 46.8% to 67%.

2.3 Synthesis

The crux of the research problem is the membership inference in machine learning. This is directly connected to the other biggest problem raised by the research – namely, the security breach of the machine learning models. Even though Shokri et al. offered several techniques to minimize the information leakage, the models are generally vulnerable for it.

The authors did an impressive work by proposing a new type of attack model and proving its viability by running a set of experiments on it. However, it is not completely clear how did the authors pick the number of shadow models to train. They claim that an increased number of shadow models improve the accuracy of the attack but make it more expensive. Therefore, it would be interesting to further study this relationship and identify an optimal number of shadow models for an attack.

Besides, Shokri et al. claim that the most secure models are those which do not overfit and which dataset is diverse. It would be interesting to analyze

those accurately trained models and see if it is possible to improve the attack technique so that more data can be leaked.

3 Summary & Questions

In this part of the assignment, the main points of the paper 'Accountability for the use of algorithms in a big data environment' by Anton Vedder and Laurens Naudts[6] is given and the questions related to the paper are presented.

3.1 Summary

The presence of algorithms in our life is constantly increasing, hence their impact should be assessed and they should be evaluated. One of the recent examples of regulations of the kind is the General Data Protection Regulation (GDPR) implemented on 25 May 2018. The authors consider implicit and explicit accountability mechanisms described in GDPR and evaluate them in a Big Data context.

Vedder and Naudts present the analysis of the complexities of algorithms from technical and contextual perspectives. They also discuss the accountability techniques present in GDPR, in particular, self-assessing mechanisms, the right not to be subject to automated decision-making and transparency principle. Besides, Vedder and Laurens Naudts analyze all the techniques in a Big Data context and elaborate on cases where accountability techniques cannot be applied to the data processing.

The authors conclude that the accountability of algorithms in Big Data is very important in terms of data protection. Even though GDPR provides some implicit and explicit accountability techniques, they are not fully sufficient in a Big Data context. Hence, the more suitable techniques should be developed and data protection practices should be established.

3.2 Questions

1. What do the authors imply by the technical and contextual complexity of an algorithm in a Big Data context?

2. What are the major implicit accountability mechanisms present in GDPR?

3. Who is the data controller and what are his main responsibilities according to GDPR?

4. Can accountability mechanisms presented in GDPR be applied to the algorithms in a Big Data context?

References

[1] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Security and Privacy (SP), 2017 IEEE Symposium on*. IEEE, 2017, pp. 3–18.

[2] "Code for Membership Inference Attack against Machine Learning Models (in S&P 2017)," https://github.com/csong27/membership-inference.

[3] "Machine Learning on AWS," https://aws.amazon.com/machine-learning/.

[4] "Cloud Prediction API is deprecated," https://cloud.google.com/prediction/.

[5] "Cloud Machine Learning Engine," https://cloud.google.com/ml-engine/.

[6] A. Vedder and L. Naudts, "Accountability for the use of algorithms in a big data environment," *International Review of Law, Computers & Technology*, vol. 31, no. 2, pp. 206–224, 2017.