Institut National des Sciences Appliquées

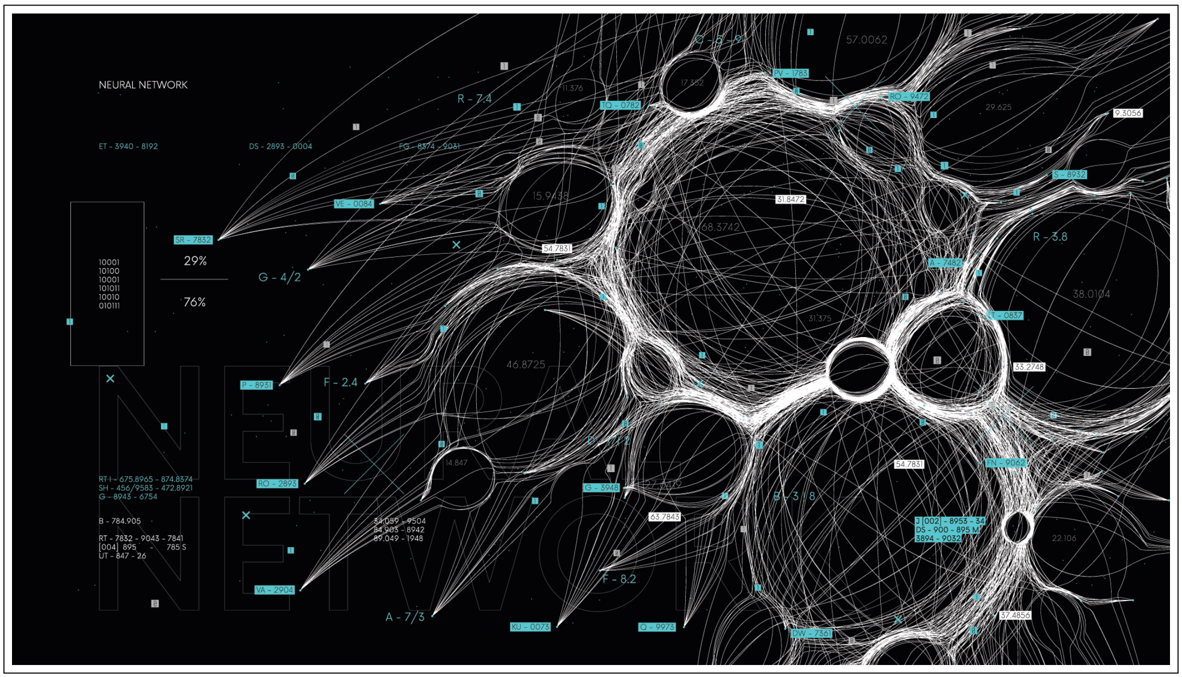
Université de Rennes

On the use of Shadow Models to perform Membership Inference Attacks (MIA)

An abstract

Author:
Thomas Aubin

English Professor :
Olga Moine
Project Manager :
Cédric Eichler



December 11, 2024

Abstract

Machine Learning could be correctly defined as the Science of Artificial Intelligence (in the sense of algorithms capable of imitating some behaviour until now considered as strictly human). This activity-growing topic is in fact composed of numerous and complex scientific domains. Among them can be found of course statistics, but also other mathematics subdomains like probabilities, linear algebra, numerical analysis and optimization, as well as Computer Science skills, to cite Data Science, high-level programming and complexity theory. Numerous Machine Learning algorithms have been designed and implemented for the last two decades, and they all have but one thing in common : they use tremendous amounts of data to be operational. Data for Machine Learning Models can basically be seen as fuel for cars : nothing happens until you have provided enough of it to your machine. In regard to the surprisingly fast ascent of AI-usage in civil and industrial activities, usage of personal data has also risen up as a global concern, since many models are accused of using data they're not allowed to, or on the contrary, to leak the data used in the training phases to their users.

As a consequence of this problem, a whole new research field in Machine Learning is born somewhere around 2015, called Adversarial Machine Learning. It basically consists in attacking Machine Learning Algorithms to understand their behavior, and especially which datasets they are training on.

The papers presented herein aim to propose a starting point in the fundamental question : how can we reproduce several avatars of a Machine Learning model to conclude on how it produces its results, and therefore determine if it has actually used a specific part of a dataset or not. Those avatars are called shadow models, and this attacking process is called a Membership Inference Attack[1].

The main difficulty of the study is that the model is considered as a "black box", which means we can't evaluate its training techniques, hyperparameters and overall behaviour. Despite this issue, the results of the study are quite impressive : thanks to specific tools to be found on the Internet, anyone can upload a dataset and build a public ML algorithm with little to no skills in Computer Science. This leads the researchers to the conclusion that the ML platforms can leak quite a lot of information about their dataset. To cite some relevant numbers : some of Google's and Amazon's ML services input data has been extracted by the SM, with an accuracy of 94% and 74% respectively. Another concerning result is that health-care datasets belong to the most private and therefore the most dangerous type of data and can be easily accessed through a MIA.

Of course the study is presenting some more technical results, like "a new shadow learning technique that works with minimal knowledge about the target model and its dataset". Here is the rough principle of this new technique : the attacker creates k shadow models, each of them being trained on a different dataset, whose format and distribution must be similar to the target's training dataset. The training datasets can be generated using one method among the following three.

1. Model-based synthesis : the target model itself can be used to generated synthetic data, which in turn can take place of a new training dataset.
2. Statistics-based synthesis : this second method is similar to the previous one, except that the data is then slightly modified to correspond to statistical expectations of the attacker
3. Noisy real data : if the attacker has access to another dataset considered as a correct input for the model, it can be used as well

For the study, six datasets are divided into smaller subset, then analysed :

1. CIFAR : one of the most popular datasets used to train image recognation algorithms
2. Purchases : this one can be found on a famous challenges platform named Kaggle, which is an equivalent of root-me for Machine Learning
3. Locations : a publicly available set of location check-ins in Bangkok collected in 2012 and 2013 on Foursquare, a social network
4. Texas hospitals stay : released by the Texas Department of State Health Service, previously mentioned as leading to one of the most striking conclusions of the report
5. MNIST : a set of number centered in images
6. UCI Adult : some records of socio-economical attributes of a population, whose goal is to verify the conditions under which a person can perceive a $50K annual income

[1] for the rest of the abstract, we'll use the following abreviations : ML, SM and MIA

To determine the success of an attack, two metrics are computed : precision, corresponding to the fraction of records inferred as members that are actually part of the dataset, and recall, which is the fraction of the training dataset's members correctly inferred by the attacker. The models attacked are not described hereby, but we can keep in mind that the attacks go successful on different forms of Machine Learning algorithms.

We shall not go into the details of the curves presented, and just recall that attack precision remains high (despite decreasing) when the noise increases. This means that the attacks are resilient in regard of distribution of the target's training data. Therefore, they can be performed on a large number of models, even the ones for which we have minimum knowledge about their training sets. A crucial parameter for the success of the attack (but more generally for any ML model) is its overfitting. This parameter basically describes if the model is more accurate in fitting known data (hindsight) but less accurate in predicting new data. Applied to our case, it is identified as the main cause of leaking data, the structure and type of the model being the second factors.

The ending of the paper presents several preventive strategies that can be deployed to face MIA, like the regularization (a statistical process that basically consists in turning answers into simpler one) as the main efficient technique against overfitting, and decreasing the number of classes of the prediction vector.

Bibliography

[1] *Adversarial Machine Learning.* Page Wikipedia de l'Adversarial Machine Learning. Nov. 2024. URL: https://en.wikipedia.org/wiki/Adversarial_machine_learning#Adversarial_attacks_and_training_in_linear_models.

[2] Tristan Allard and Mathias Bernard. "Snakes Strikes Back". In: (Oct. 2024).

[3] Tatev Aslanyan. *Machine Learning in 2024 – Beginner's Course.* Feb. 2024. URL: https://www.youtube.com/watch?v=bmmQA8A-yUA&t=1769s.

[4] Author. *Membership inference attacks from first principles.* How published. Some note. Month Year. URL: https://www.youtube.com/watch?v=1CNxfhMlk-A.

[5] Chloé-Agathe Azencott. *Introduction au Machine Learning.* (2nd). InfoSup. Dunod, Feb. 2022.

[6] *Classification naïve bayésienne.* Page Wikipedia du modèle bayésien naïf. Aug. 2024. URL: https://fr.wikipedia.org/wiki/Classification_na%C3%AFve_bay%C3%A9sienne.

[7] *Gaussian Naive Bayes Explained With Scikit-Learn.* Tutoriel pour construire un classifieur bayésien naïf. Nov. 2023. URL: https://builtin.com/artificial-intelligence/gaussian-naive-bayes.

[8] *Generative adversarial network.* Page Wikipedia du modèle GAN. Nov. 2024. URL: https://en.wikipedia.org/wiki/Generative_adversarial_network.

[9] Matt Harrison. *Machine Learning - Les Fondamentaux. Exploiter des données structurées en Python.* O'Reilly, 2020.

[10] Benjamin JOURDAIN. *Probabilités et statistiques pour l'ingénieur.* Jan. 2018.

[11] Learndataa. *81 Scikit-learn 78 Supervised Learning 56 Naive Bayes classifiers.* Youtube. 2021. URL: https://www.youtube.com/watch?v=9Tmnr4L5ZOQ.

[12] Zinan Lin et al. "Using GANs for Sharing Networked Time Series Data : Challenges, Initial Promise, and Open Questions". In: (Jan. 2021). Présentation du modèle DoppelGANger. URL: https://arxiv.org/abs/1909.13403.

[13] *Machine Learning.* Page Wikipedia du Machine Learning. Nov. 2024. URL: https://en.wikipedia.org/wiki/Machine_learning.

[14] MarkovML. *Comparing and Evaluating Datasets: A Simplified Guide.* Nov. 24, 2024. URL: https://www.markovml.com/blog/compare-datasets.

[15] Boris Meinardus. *How I'd learn ML in 2024 (if I could start over).* Youtube. 2024. URL: https://www.youtube.com/watch?v=gUmagAluXpk.

[16] *Méthode des k plus proches voisins.* Page Wikipedia anglophone de la recherche des plus proches voisins. Mar. 2024. URL: https://fr.wikipedia.org/wiki/M%C3%A9thode_des_k_plus_proches_voisins.

[17] *Nearest neighbor search.* Page Wikipedia anglophone de la recherche des plus proches voisins. Aug. 2024. URL: https://en.wikipedia.org/wiki/Nearest_neighbor_search.

[18] *Overfitting.* Page Wikipedia de l'Overfitting. Nov. 2024. URL: https://en.wikipedia.org/wiki/Overfitting#Machine_learning.

[19] *Régression logistique.* Page Wikipedia de la régression logistique. Nov. 2024. URL: https://fr.wikipedia.org/wiki/R%C3%A9gression_logistique.

[20] Reza Shokri. *Membership Inference Attacks against Machine Learning Models.* Vidéo de vulgarisation du papier du même nom. May 2017. URL: https://www.youtube.com/watch?v=rDm1n2gceJY&t=53s.

[21] Reza Shokri et al. "Membership Inference Attacks Against Machine Learning Models". In: ().

[22] *Training, validation, and test data sets.* Page Wikipedia rappelant la différence entre les datasets d'entraînement et de test. Nov. 2024. URL: https://en.wikipedia.org/wiki/Training,_validation,_and_test_data_sets.