ML Privacy Meter: Aiding Regulatory Compliance by Quantifying the Privacy Risks of Machine Learning*

Sasi Kumar Murakonda, Reza Shokri

Data Privacy and Trustworthy ML Research Lab
National University of Singapore
{murakond,reza}@comp.nus.edu.sg

ABSTRACT

When building machine learning models using sensitive data, organizations should ensure that the data processed in such systems is adequately protected. For projects involving machine learning on personal data, Article 35 of the GDPR mandates it to perform a Data Protection Impact Assessment (DPIA). In addition to the threats of illegitimate access to data through security breaches, machine learning models pose an additional privacy risk to the data by indirectly revealing about it through the model predictions and parameters. Guidances released by the Information Commissioner's Office (UK) and the National Institute of Standards and Technology (US) emphasize on the threats to data from models and recommend organizations to account for and estimate these risks to comply with data protection regulations. Hence, there is an immediate need for a tool that can quantify the privacy risks to data from models.

In this paper, we focus on this indirect leakage about training data from machine learning models. We present ML Privacy Meter, a tool that can quantify the privacy risk to data from models through state of the art membership inference attack techniques. We discuss how this tool can help practitioners in compliance with data protection regulations, when deploying machine learning models.

1. DATA PRIVACY RISKS OF MACHINE LEARNING MODELS

Organizations are collecting massive amounts of personal information for building applications that are powered by machine learning. This data, which is used to train the models, typically contain sensitive information about individuals. Machine learning models encode information about the datasets on which they are trained. The encoded information is supposed to reflect the general patterns underlying the population data. However, it is commonly observed that these models memorize specific information about some members of their training data [5] or be tricked to do so [10].

Models with high generalization gap as well as the models with high capacity (such as deep neural networks) are more susceptible to memorizing data points from their training set. This is reflected in the predictions of the model, which exhibits a different behavior on training data versus test data, and in the model's parameters which store statistically correlated information about specific data points in their training set [9, 7]. This vulnerability of machine learning models was shown using membership inference attacks, where an attacker detects the presence of a particular record in the training dataset of a model, just by observing the model. Machine learning models were shown to be susceptible to these attacks in both the black-box [9] and white-box settings [7].

In the black-box setting, we can only observe predictions of the model. This setting models the scenario of machine learning as a service offered on cloud platforms by companies such as Amazon,[1] Microsoft,[2] and Google.[3]. It can be used to measure the privacy risks against legitimate users of a model who seek predictions on their queries. In the white-box setting, we can also observe the parameters of the model. This reflects the scenario where a model is outsourced to a potentially untrusted server or to the cloud, or is shared with an aggregator in the federated learning setting [6, 8]. The privacy risks of machine learning models can be evaluated as the accuracy of such inference attacks against their training data.

2. DATA PROTECTION REGULATIONS

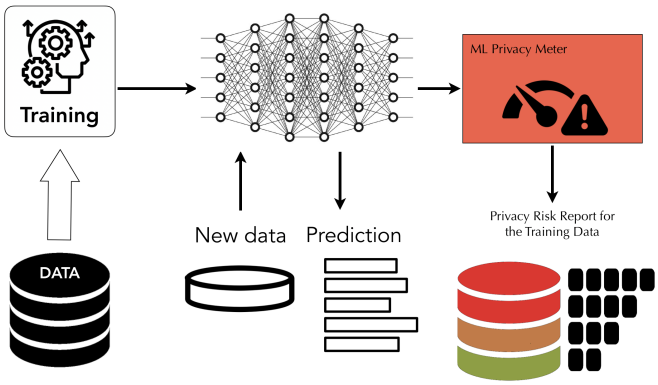
For a safe and secure use of machine learning models, it is important to have a quantitative assessment of the privacy risks of these models, and to make sure that they do not reveal sensitive information about their training data. Data protection regulations, such as GDPR, and AI governance frameworks require personal data to be protected when used in AI systems, and that the users have control over their data and awareness about how it is being used.

For projects involving innovative technologies such as machine learning, it is mandatory from Article 35 of the GDPR to perform a Data Protection Impact Assessment (DPIA).[4] The key steps in DPIA are to identify the potential threats to data and assess how they might affect individuals. In general, risk assessment in DPIA statements focuses on the risk of security breaches and illegitimate access to the data. Machine learning models pose additional privacy risk to the training data by indirectly revealing about it through the model's predictions and parameters. Hence, special attention needs to be paid for data protection rules in AI regulation frameworks. Guidances released by both the European Commission and the White House call for protection of personal data during all the phases of deploying AI systems

*Repository for the code and tutorials is available at `https://github.com/privacytrustlab/ml_privacy_meter`

[1] `https://aws.amazon.com/machine-learning`
[2] `https://studio.azureml.net`
[3] `https://cloud.google.com/prediction`
[4] `https://gdpr-info.eu/art-35-gdpr/`

Figure 1: ML Privacy Meter is a python library that enables quantifying the privacy risks of machine learning models to members in the training dataset. The tool provides privacy risk scores which help in identifying the data records that are under high risk of being revealed through the model parameters or predictions.



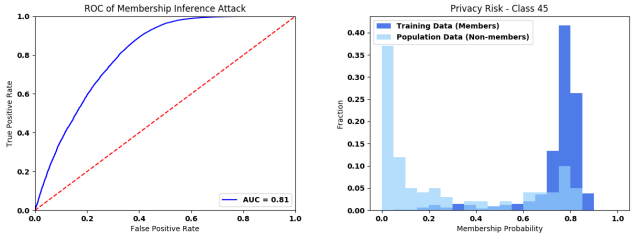
and build systems that are resistant to attacks [2, 3]. Recent reports published by the Information Commissioner's Office (ICO) for auditing AI [1] and the National Institute of Standards and Technology (NIST) for securing applications of Artificial Intelligence [4] highlight the privacy risk to data from machine learning models. And they specifically mention membership inference as a confidentiality violation and potential threat to the training data from models. It is recommended in the auditing framework by ICO for organizations to identify these threats and take measures to minimize the risk [1]. As the ICO's investigation teams will be using this framework to assess the compliance with data protection laws, organizations must account for and estimate the privacy risks to data through models.

3. ML PRIVACY METER

A tool that can automatically assess the privacy risks of machine learning models to their training data can aid practitioners in compliance with data protection regulations. But how do we measure the risk of indirect information leakage about training data from complex ML models? We present ML Privacy Meter that can quantify the privacy risks to training data and is based on well-established algorithms to measure privacy risks of machine learning models through membership inference attacks [7, 9]. The tool provides privacy risk scores that help in identifying the data records that are under high risk of being revealed through the model parameters or predictions. The tool can generate extensive privacy reports about the aggregate and individual risk for data records in the training set at multiple levels of access to the model. It can estimate the amount of information that can be revealed through the predictions of a model (referred to as Black-box access) and through both the predictions and parameters of a model (referred to as White-box access). Hence, when providing query access to the model or revealing the entire model, the tool can be used to assess the potential threats to training data.

ML Privacy Meter works by implementing membership inference attacks against machine learning models. It simulates attackers with different levels of access and knowledge about the model. It considers attackers who can exploit only

Figure 2: ML Privacy Meter quantifies the privacy risk to training data from machine learning models. The risk is measured through success of membership inference attacks quantified by an ROC curve representing the trade-off between true positive and false positive rates. It also allows for comparison of privacy risk across records from different classes.



the predictions of the model, the loss values, and the parameters of the model. For each of the simulated attacks, the tool reports risk scores for all the data records. These scores represent the attacker's belief that the record was part of the training dataset. The larger the gap between the distribution of these scores for records that are in the training set versus records that are not in the training set, the larger is the leakage from the model would be.

Success of the attacker can be quantified by an ROC curve representing the trade-off between False Positive Rate and True Positive Rate of the attacker. True positive represents correctly identifying a member as present in the data and False positive refers to identifying a non-member as member. An attack is successful if it can achieve larger values of True Positive rate at small values of False Positive rate. A trivial attack such as random guess can achieve equal True Positive and False Positive Rates. ML Privacy Meter automatically plots the trade-offs that are achieved by our simulated attackers. The area under those curves quantifies the aggregate privacy risk to the data posed by the model. The higher the area under curve, larger the risk. These numbers not only quantify the success of membership inference attacks, but they can also be seen as a measure of information leakage from the model.

When deploying machine learning models, this quantification of risk can be useful while performing a Data Protection Impact Assessment. The aim of doing a DPIA is to analyze, identify and minimize the potential threats to data. ML privacy meter can guide practitioners in all the three steps. It can help in estimating the privacy risk to data and to identify the potential causes of this risk. It can also be useful in selecting and deploying appropriate risk mitigation measures.

The tool produces detailed privacy reports for the training data. It allows comparing the risk across records from different classes in the data. We can also compare the risk posed by providing black box access to the model with the risk due to white box access. As the tool can immediately measure the privacy risks for training data, practitioners can take simple actions such as finetuning their regularization techniques, sub-sampling, re-sampling their data, etc., to reduce the privacy risk. Or they can even choose to learn with a privacy protection, such as differential privacy, in place.

Differential Privacy is a cryptographic notion of privacy, wherein the outputs of a computation should be indistin-

guishable when any single record in the data is modified. The level of indistinguishability is controlled by a privacy parameter ϵ. Open source tools such as OpenDP [5] and TensorFlow Privacy [6] are available for training models with differential privacy guarantees. Selecting an appropriate value for ϵ is highly non-trivial when using these tools. Models learned with smaller value of ϵ provide better privacy guarantees but are also less accurate. ϵ represents a worst case upper bound on the privacy risk and the practical risk might be much lower. ML Privacy Meter can help in the selection of privacy parameters (ϵ) for differential privacy by quantifying the risk posed at each value of epsilon. Compared to just relying on the guarantees provided by epsilon, using this method helps in deploying models with higher accuracy. By letting practitioners choose models with better utility, ML Privacy Meter can enable the use of privacy risk minimization techniques.

4. SUMMARY

By leaking information through predictions and parameters, machine learning models pose an additional privacy risk to data in AI systems. To comply with data protection regulations, we need to assess these risks and take possible mitigation measures. ML Privacy Meter quantifies the privacy risk of machine learning models to their training data. It can guide practitioners in regulatory compliance by helping them analyze, identify, and minimize the threats to data. By permitting for deploying models with better accuracy, through practical estimates of utility-privacy trade-offs, we expect the tool to boost adaptation of privacy enhancing techniques in machine learning.

5. REFERENCES

[1] Guidance on the ai auditing framework draft guidance for consultation. information commissioner's office (2020). https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf.

[2] On artificial intelligence - a european approach to excellence and trust. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

[3] A taxonomy and terminology of adversarial machine learning. https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf.

[4] A taxonomy and terminology of adversarial machine learning. https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8269-draft.pdf.

[5] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium*, pages 267–284, 2019.

[6] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282, 2017.

[7] M. Nasr, R. Shokri, and A. Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *IEEE Symposium on Security and Privacy (SP)*, pages 1022–1036, 2019.

[8] R. Shokri and V. Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321, 2015.

[9] 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*, pages 3–18. IEEE, 2017.

[10] C. Song, T. Ristenpart, and V. Shmatikov. Machine learning models that remember too much. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 587–601, 2017.

[5] https://github.com/opendifferentialprivacy/

[6] https://github.com/tensorflow/privacy