
IOWA STATE UNIVERSITY
Department of Computer Science

Machine Learning with Membership Privacy using Adversarial Regularization

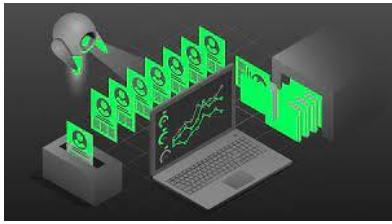
Milad Nasr, Reza Shokri, Amir Houmansadr

Yu-Pin Liang (Jasmine)

IOWA STATE UNIVERSITY

INTRODUCTION – Risk in Machine Leaning

- Machine learning models leak information about the datasets on which they are trained. (machine learning service from Google, Amazon, Microsoft…)



- Wide range of sensitive data, such as online and offline profiles of users, location traces, personal photos, speech samples, medical and clinical records, and financial portfolios, is used as input for training machine learning models.

SENSITIVE DATA

- The leakage through complex machine learning models maybe less obvious, compared to, for example, linear statistics

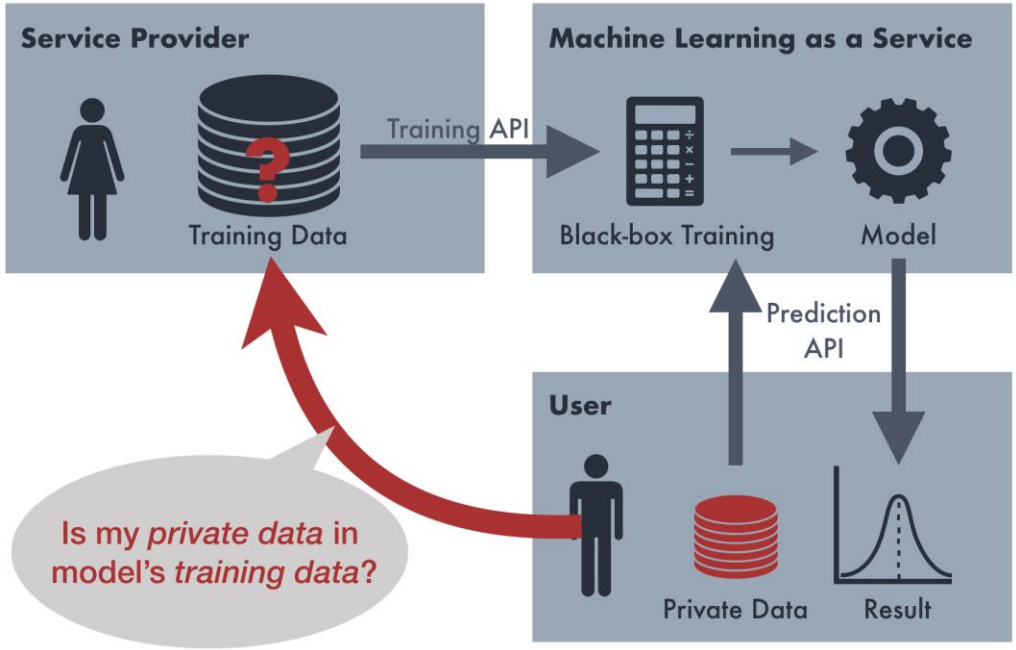
INTRODUCTION – Membership Inference Attack

- An adversary algorithm to trace individual members of a model's training dataset

  -> aims to distinguish between data points that were part of the **model's training set** and any other data points from the same distribution.

- This is known as the tracing (membership inference) attack.
- Membership inference attacks pose a threat to the privacy of individual data points in a dataset used for machine learning, with even black-box access to a model, can perform a membership inference attack against the model to determine whether or not a target data record is a member of its training set.

INTRODUCTION – Membership Inference Attack



Model - membership inference attack

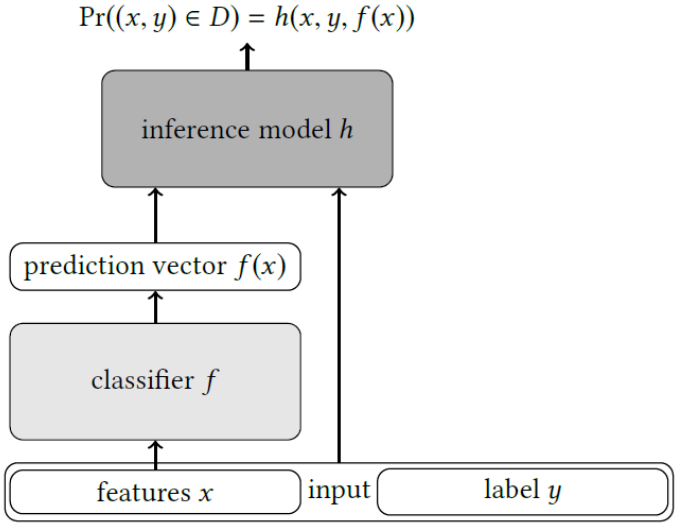


Figure 1: The relation between different elements of the black-box classification model f and the inference model h.

Purpose

In this paper, authors focus on protecting machine learning models against this exact threat: **black-box membership inference attacks**.

Existing Defense Mechanism

Two major groups of existing defense mechanisms

- Simple mitigation techniques, such as limiting the model's predictions to top-k classes, therefore reducing the precision of predictions, or regularizing the model

  -> impose a negligible loss to model but cannot guarantee any rigorous notion of privacy.

- Differential privacy mechanisms

  ->do guarantee (membership) privacy up to privacy parameter ϵ. Impose a significant classification accuracy loss for protecting large models on high dimensional data for small values of ϵ.

Contribution

Authors design rigorous privacy mechanism for protecting a given training dataset, against a particular adversarial objective. We want to train machine learning models that guarantee membership privacy

Model - idea

- Author model this optimization as a min-max privacy game between the defense mechanism and the inference attack. - > optimization problem

- Train the model in an adversarial process ->To protect data privacy, add the gain of the inference attack as a regularizer for the classifier.

- Negligible loss in classification accuracy for a significant gain in membership privacy

Model – How it works

- The classification model maps features of a data record to classes, and computes the probability that it belongs to any class. The primary objective of this model is to minimize prediction error.

- The inference model maps a target data record, and the output of the classifier on it, to its membership probability. The objective of the inference model is to maximize its membership inference accuracy.

Model

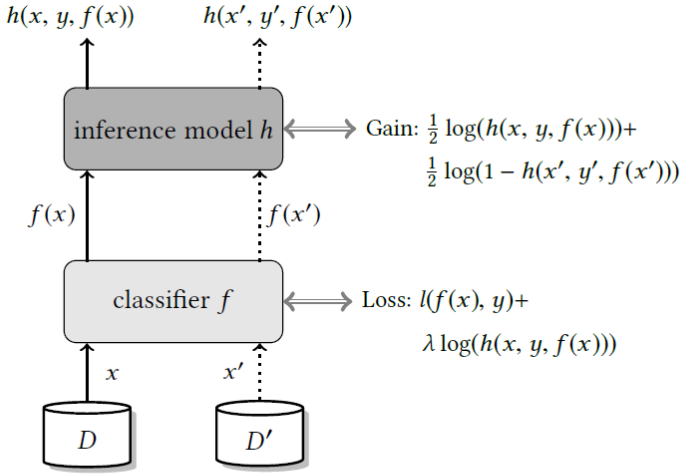


Figure 2: Classification loss and inference gain, on the training dataset D and reference dataset D', in our adversarial training. The classification loss is computed over D, but, the inference gain is computed on both sets. To simplicity the illustration, the mini-batch size is set to 1.

Result- with and without Defense

	Without defense			With defense		
Dataset	Training accuracy	Testing accuracy	Attack accuracy	Training accuracy	Testing accuracy	Attack accuracy
Purchase100	100%	80.1%	67.6%	92.2%	76.5%	51.6%
Texas100	81.6%	51.9%	63%	55%	47.5%	51.0%
CIFAR100- Alexnet	99%	44.7%	53.2%	66.3%	43.6%	50.7%
CIFAR100- DenseNET	100%	70.6%	54.5%	80.3%	67.6%	51.0%

Table 4: Comparison of membership privacy and training/test accuracy of a classification model (without defense), and a privacy-preserving model (with defense) on four different models/datasets. Compare the two cases with respect to the trade-off between testing accuracy and attack accuracy. See Table 3 for the experimental setup.

Result- different L2–regularization

L2-regularization factor	Training accuracy	Testing accuracy	Attack accuracy
0 (no regularization)	100%	80.1%	67.6%
0.001	86%	81.3%	60%
0.005	74%	70.2%	56%
0.01	34%	32.1%	50.6%

Table 6: The results of using a *L2*–regularization as a mitigation technique for membership inference attack. The model is trained on the Purchase100 dataset. Compare these results with those in Table 4 which shows what we can achieve using the strategic min-max optimization.

Result- Reference set size

Reference set size	Testing accuracy	Attack accuracy
1,000	80.0%	59.2%
5,000	77.4%	52.8%
10,000	76.8%	52.4%
20,000	76.5%	51.6%
30,000	76.4%	50.6%

Table 7: The effect of the size of the reference set D' on the defense mechanism for the Purchase100 dataset. Note that (as also shown in Table 3) the size of the training set is 20,000.

Conclusion

- A new privacy mechanism for mitigating the information leakage of the predictions of machine learning models about the membership of the data records in their training sets.

  ->jointly maximize privacy and prediction accuracy

- The solution will be a model whose predictions on its training data are indistinguishable from its predictions on any data sample from the same underlying distribution.