LTU Attacker for Membership Inference

Joseph Pedersen [1] Rafael Muñoz-Gómez [2] Jiangnan Huang [2] Haozhe Sun [2] Wei-Wei Tu [3,4] Isabelle Guyon [2,4]

[1]Rensselaer Polytechnic Institute, Troy, NY, USA [2]LISN/CNRS/INRIA Université Paris-Saclay, France [3]4Paradigm, China [4]ChaLearn, USA

Introduction

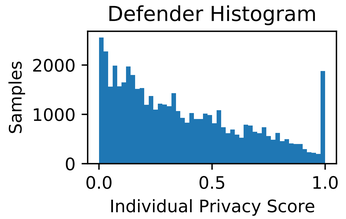
The data used to train machine learning models often needs to be kept private, but the model trained on the data may be susceptible to privacy attacks.

Methodology

Each round, give an **LTU Attacker** the trained Defender model, the Defender's model trainer with hyper-parameters, and all data except two membership labels.

The membership classification error from N independent LTU rounds is a **global privacy score**.

An **individual privacy score** for any sample $d \in \mathcal{D}_D$ computed using that sample for all N rounds, and only drawing $r \sim \mathcal{D}_R$ at random. Often some samples are 100% exposed.

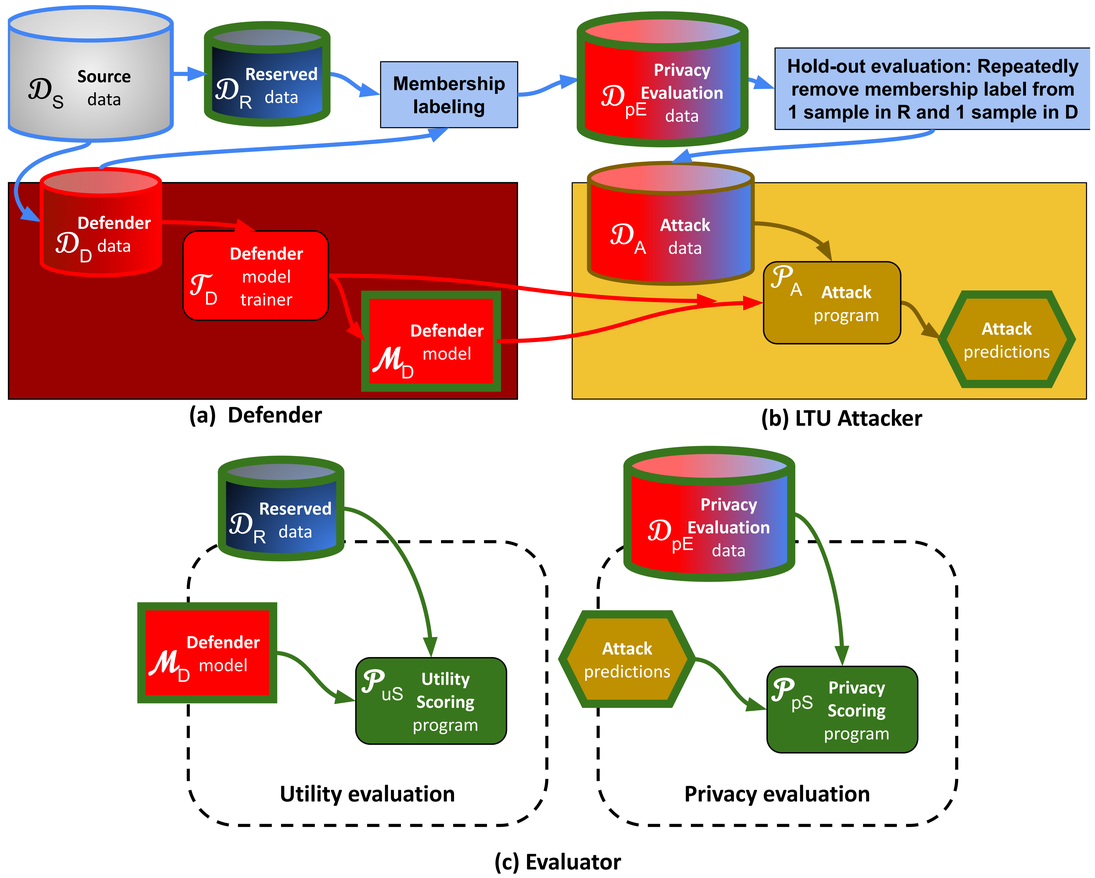


Privacy Scores for CIFAR100, using AlexNet for the Defender model trainer and the ML Privacy Meter of Shokri et al. for the Attack Program

Theoretical Results

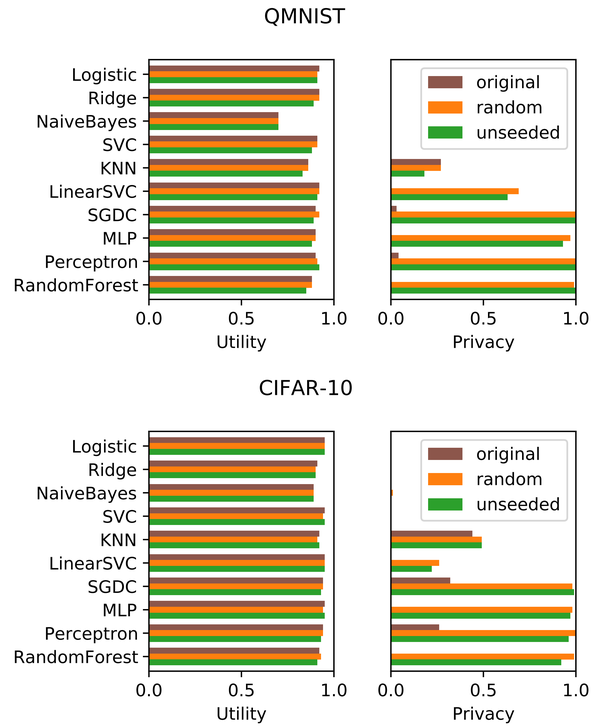
- Two theorems with lower bounds on LTU Attacker membership classification accuracy for **overfit** models.
- A third theorem showing that the LTU Attacker can achieve 100% accuracy if the Defender's model trainer is **deterministic**, invariant to the order of the training data, and injective.

Leave-Two-Unlabeled (LTU) Attacker framework for assessing privacy loss



Methodology Flow Chart. (a) Defender: uses Defender data and Defender model trainer to train Defender model. **(b) LTU Attacker:** each round, the Evaluator gives the LTU Attacker all of the data including their origin, Defender or Reserved, except for the membership labels from one Defender sample and one Reserved sample, which the LTU Attacker must predict. **(c) Evaluator:** Computes two scores: LTU Attacker prediction error over N rounds (Privacy metric), and Defender model classification performance (Utility metric).

Experimental Results



Utility and Privacy of different scikit-learn models (missing bar indicates 0)

Conclusion

Many model trainers are insufficiently random or overfit particular training samples, making models that are susceptible to privacy attacks. This framework can help identify vulnerability, especially for the most exposed samples.