2017 IEEE Symposium on Security and Privacy

Membership Inference Attacks Against Machine Learning Models

Reza Shokri
Cornell Tech
shokri@cornell.edu

Marco Stronati*
INRIA
marco@stronati.org

Congzheng Song
Cornell
cs2296@cornell.edu

Vitaly Shmatikov
Cornell Tech
shmat@cs.cornell.edu

***Abstract*—We quantitatively investigate how machine learning models leak information about the individual data records on which they were trained. We focus on the basic membership inference attack: given a data record and black-box access to a model, determine if the record was in the model's training dataset. To perform membership inference against a target model, we make adversarial use of machine learning and train our own inference model to recognize differences in the target model's predictions on the inputs that it trained on versus the inputs that it did not train on.**

We empirically evaluate our inference techniques on classi-

part of the model's training dataset or not. We investigate this question in the most difficult setting, where the adversary's access to the model is limited to **black-box** queries that return the model's output on a given input. In summary, we quantify membership information leakage through the prediction outputs of machine learning models.

To answer the membership inference question, we turn machine learning against itself and train an *attack model* whose purpose is to distinguish the target model's behavior