Review of Membership Inference Attacks Against Machine Learning Models

Mohammadmahdi Abdollahpour
mabdollahpour@aut.ac.ir

CONTENTS

1 HOW TO ATTACK A MODEL

This paper[1] introduces concept of Membership Inference Attacks Against Machine Learning Models. Membership inference is determining whether a given data record was part of the models training dataset or not. This can be dangerous because sometimes training data could be sensitive data. This problem is investigated in black box scenario in which adversary can only supply input to model and get the output.

1.1 *Training attack model*

An adversary tries to predict whether a record is included in training set by training attack models, one for each class of target output. Assume we have record x and target model output $y = f_{target}(x)$ and we want to see if it is in target training set or not. Attack model takes in(y, Y) where Y is true label of x and $y = f_{target}(x)$ as inputs, and predicts in or out. Question is how do we train this model? In order to train attack models we train multiple shadow models and shadow models are intended to behave similar to target model but for a shadow model we know the ground truth, it means we know if a record was included in training set. Shadow models must be trained similar to target model so same algorithm or same ML service must be used. For each shadow model we provide train data and test data and we put (y, Y, in) for training records (D^{shadow}_{train}) and (y, Y, out) for test records (D^{shadow}_{test}) in attack model training set where Y is true label of $x \in D^{shadow}_{train} or D^{shadow}_{test}$ and y is output of shadow model, $y = f_{shadow}(x)$. We train attack model to perform this binary classification.

we need to provide data to train shadow models. This data can be noisy real data or can be generated by picking records that are classified by high confidence or it can be from statistical synthesis like having knowledge about marginal distributions.

[1] Membership Inference Attacks Against Machine Learning Models

1.2 *Model-based synthesis*

Paper proposes a hill-climbing algorithm to find inputs that are classified as class c by high confidence for each output class. briefly, it starts by taking a random record and at each iteration proposing a new record by changing k (k is initialized to k_{max} and divided by 2 each time until it gets to k_{min}) features of record until probability for class c gets larger than previous accepted record, then if this probability is larger than a threshold and it is the maximum probability in y, it is accepted with probability of y_c that is probability of this record belonging to class c. If after rej_{max} iteration we do not find a record with higher confidence than previous accepted record, search is terminated and started again.

By the way this method may not be good for images because of difficulties exploring the space.

2 RESULTS HIGHLIGHTS

Success of membership inference is directly related to the (1) generalizability of the target model and (2) diversity of its training data. If the model overfits and does not generalize well to inputs beyond its training data, or if the training data is not representative, the model leaks information about its training inputs.[1]

2.1 *Effect of the number of classes and training data per class*

Results shows that more classes contributes to information leakage. briefly, models with more output classes need to remember more about their training data, thus they leak more information.

In general, the more data in the training dataset is associated with a given class, the lower the attack precision for that class.[1]

2.2 *Effect of overfitting*

Assuming train-test accuracy gap as a measure of overfitting, bigger train-test accuracy gaps in a model indicate more information leakage. overfitting is not the only cause of information leakage but it contributes to that. So, the leakage of sensitive information about the training data is introduced as another form of overfitting by the paper.

Different machine learning models, due to their different structures, remember different amounts of information about their training datasets.This leads to different amounts of information leakage even if the models are overfitted to the same degree.[1]

3 HOW TO DEFEND AGAINST ATTACKS

Regularization techniques such as dropout can help defeat overfitting. differentially private models are secure against this type effects. ML as service platforms need to explicitly warn customers about this risk and provide more visibility into the model and the methods that can be used to reduce this leakage.

3.1 *Mitigation strategies and evaluations*

- Restrict the prediction vector to top k classes in order to leak less information. It turns out that restricting it to only the most likely label does not foil the attack because of attack can still exploit the mislabeling behavior of the target model because members and non-members of the training dataset are mislabeled differently (assigned to different wrong classes) [1]

- Coarsen precision of the prediction vector.
- Increase entropy of the prediction vector: for logit vector z output probabilities are $e^{z_i/t} / \sum_j e^{z_i/t}$ to increase entropy and leak less information.
- Use regularization.

Overall, the attack is robust against these mitigation strategies.

4 CRITICAL THOUGHTS AND TECHNICAL SUGGESTIONS

4.1 *Actor Critic Design for training*

In training process, model should be trying to satisfy two properties, increasing train accuracy (resulting overfitting if done too much) and preserving privacy (in contradiction with previous one). In other words we must maximize accuracy while minimizing information leakage. (similar to GAN networks)

we must design a training process to satisfy both properties to fit the data very well ans also preserve privacy.

REFERENCES

[1] Reza Shokri, Marco Stronati, Congzheng Song, Vitaly Shmatikov *Membership Inference Attacks against Machine Learning Model*. IEEE Symposium on Security and Privacy (SP), Oakland , 2017.