MEMBERSHIP INFERENCE ATTACK AGAINST MACHINE LEARNING MODELS

ECE 535 Project
Master's in Data Science,
University of Victoria, Canada
Mentored by,
Dr. Stephen W. Neville

Submitted by,
Shaima Patel
V009499940
Spring 2020
Email-id: shaimapatel24@gmail.com

***Abstract*—The researchers have demonstrated how machine learning leaks information about the data they were trained on. They have discussed Membership Inference Attack using data record and black box access. To perform the experiment, they have trained their own inference model and have used realistic datasets like Amazon, Google etc. A hospital discharge dataset has also been used to show how privacy is at stake. Finally, they investigate the factors that influence leakage and give some mitigation techniques.**

Keywords — machine learning, membership inference attack, security, testing and training data, black box, shadow models

I. INTRODUCTION

This paper focuses on Membership Inference attack that is possible on different platforms. What is Membership Inference Attack? To infer membership of individual training instance of model to which adversary has black box access through Machine Learning as service API. In simple words, this attack infers membership of data even without having knowledge of training datasets. Basically, ML collects data to give feedback. Example, Mobile App Maker analyze user activity and in app purchases where user most likely to respond. The researchers of the paper have tried to solve this problem of data leakage by using black box arrangement which will return output for only the selective given input.

The researchers have trained attack model to illustrate the Membership Inference Attack. For this purpose, they have distributed the data into training and test datasets. For choosing training data they have turned ML against itself and trained attack model. The purpose of doing this is to distinguish target model behavior on training input from the behavior that it did not encounter during training. This is also called Classification Problem. They have used shadow training technique for training the attack model. Now there are 3 ways in which models are trained:

- Create multiple shadow models where training dataset is already known and so it follows target model.
- Train attack model by using input output obtained from shadow model.
- Generate training data for shadow model by using black box access given to target model(real) or by using the statistics about population from which target and training data is drawn(synthetic).

The first method has no prior knowledge of training dataset while the 2nd and 3rd allow attacker to query target model (only once before inferring whether record is trained data or not).

This paper tries to achieve the goal by inferring membership of the given black box

API, which is unknown to model and opposed to statistics, discuss the root cause of such attacks and compares various ways to stop it.

The question is why should the inference attack be mitigated? Well, if we have data about a patient's medical or financial things, inferring whether samples belong to a person becomes a privacy threat. Such information can be leaked to the attacker and is said to be leakage of confidential information that can be used to target he person. This kind of attack on sensitive machine learning data can also be used in biasing[1].

II. ASSUMPTIONS

This paper tries to explain the concept of Overfitting. What is Overfitting? Machine Learning models often behave different on data they were trained versus the data they see for the first time. Now it is assumed that using the concept of overfitting the attacker will create model in such a way that he knows all the background knowledge of training algorithm and the model. If this is the case and if the output of the attack is also the output member of data originally then it gives rise to Membership Precision Attack. We can calculate Recall with respect to the attacker as:
Recall = Fraction of training records that attacker correctly infers as members.

The other assumption paper tries to make is with respect to the concept studied in class called Maximum Likelihood. The classification model that we build to classify training and testing dataset outputs probability (prediction vector) for input provided. It is assumed that the class with highest probability is assigned to the data record.[2]

Using the concept of Supervised Learning, paper tries to display that when attacker knows or even has a rough idea regarding the input and output of the targeted model than he can use black box to get output from the given input.

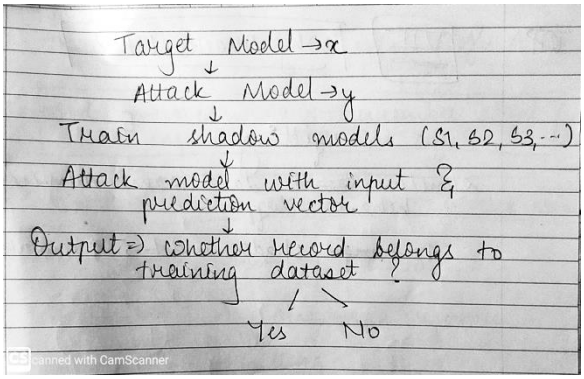
The paper also assumes that when attack is able to determine the class of input records it succeeds. This is also termed as accuracy of the attack. As mentioned above it uses the concept of Recall and also Precision to determine this where Precision is:
Precision = Fraction of records inferred as member of training datasets that are indeed members.
It also states that larger the gap between training and testing accuracy, more are the chances of Overfitting.

During the training of shadow models, researchers have stated that more the number of shadow models trained higher the attack accuracy and more the cost. And, when dealing with the accuracy of attack of a model if the number of classes are more the training data will remember more and thus will leak more information. These assumptions are made with respect to attack accuracy and the researchers also end up saying that more data in training dataset more would be the attack precision.

Researchers assume that the training data which would be selected randomly and the testing data for target model and shadow model will not overlap whereas inter-shadow model datasets may overlap amongst themselves.
$D_{trainshadow} \cap D_{traintarget} = null$
This is the worst case for attackers.

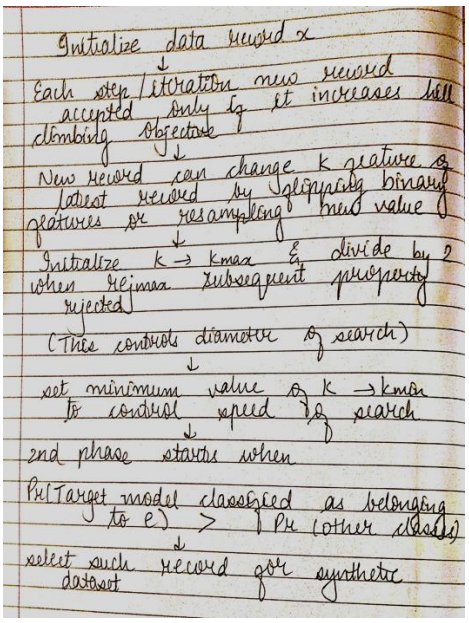
III. EXPERIMENT SETUP

The process of attack is shown below:



The figure tries to explain how this research is trying to achieve its goal of detecting Membership Inference attack and then mitigate it. The process starts with taking a target model and an attack model. We will train shadow models which are basically like target models. Now when this shadow models are ready; we will attack the model with input and prediction vector [2]. The

output will be whether the record belongs to the training set or not. The shadow models can be trained in 3 ways:
Model based synthesis: Attacker may have training data but do not contain statistics about distribution. This method will generate synthetic training data from target model in such a way that search is performed using hill climbing algorithm to find data records that are classified with high confidence.



Statistics based synthesis: Attacker has some statics information about target model's training data.
Noise real data: Attacker has access of data similar to target model's training data but is considered noisy.

More the number of shadow models → More accurate attack

IV. EVALUATION

This paper uses 6 different types of dataset:

- CIFAR
- Purchases
- Locations
- Texas hospital stays
- MNIST
- UCI Adult

The general evaluation is that synthetic data has less attack accuracy then real data as it is different for all the classes. As mentioned above that if the number of classes increase, they need to remember more information which then leads to information leakage. Now considering real data it performs well even for noisy real data. When we replace 10% features in shadow data with random value this will also be able to match the original attack and hence it is robust to noisy data. Also, the attack precision will decrease if the noise increases. To overcome the overfitting the paper also provides few mitigation techniques [2]:

- We can limit the number of different classes along with prediction vector (up to k classes)
- We can also increase the entropy of prediction vector and decrease the precision to few decimal digits.
- Finally, the researchers try to improve prediction power of model and reduce the leakage of data using the concept of regularization.

V. THOUGHTS

Now coming to the assumptions paper made;

- The main idea of this paper revolved around Overfitting and it successfully proved that this is one of the reasons but not the only reason behind data leakage. The researchers have made an assumption that more the overfitting in the model → more will be the leakage. It has been then proved with Amazon data record where one dataset(100,1e-4) is more overfitted than the other (100, 1e-6) and so the it (100,1e-4) leaks more information. But when we compare Amazon and Google, Google is less overfitted than Amazon yet leaks more information. Thus, this is how they provide satisfactory evidence to say that Overfitting is not the only factor that makes a model vulnerable to membership inference attack, model's structure and type also contribute.
- Now in the later part of the paper, researchers state that more general the target model → better representation of

training data → less data leakage. This idea opposes the first assumption.

- "Important Observation: Here the data privacy and the model's accuracy have similar objectives and are not in dispute as usual. A common cause of the degradation of both the quantities can be observed, the cause being Overfitting. The data privacy objective: Model should not leak information about the it's training data. The machine learning objective: The model must be generalized for data out-side the training data. Overfitting causes a crack in both these objectives hence is the common cause." [2]
- As mentioned in IV. Evaluation that paper using synthetic data the paper does not work good compared to the real ones. Having mentioned this there is no approach or strategy given to improve it. The paper also makes an assumption that shadow models are just like target model as they are trained using them. They have proved this assumption by training the Google, Amazon etc. records to make shadow models which are just like the original data record. In the research paper the mention of Machine learning as service, the service here are data records like Amazon. The models which are different can be used as attack model. Coming to the attack model, paper mentions that there is one attack model for each class and such arrangement will increase accuracy is proved with a valid reason, but no experimental results are given in support [2].

  Now as the paper uses the concept of Maximum Likelihood in a way that it selects the data for shadow models which have high confidence, the ones with less are ignored. I think an attack should also be displayed using the data classified having low confidence.

VI. DISCUSSION

The paper uses 6 different types of datasets and so it can hold for any given dataset. Moreover, the reason this paper was written was to mitigate data leakage as many data might have private information and hence this proves that data leak can occur for any data, this paper holds for all datasets.

This paper states that different types of ML models like logistic regression, linear regression etc. can be generated locally and can be used to compare its data leak, success rate, attack accuracy, overfitting. This can tell us if the paper holds for other ML algorithm/ intelligent adversaries. So, when testing on other adversaries this paper will hold but the result might change. E.g. For neural network the amount of overfitting may be less but for logistic regression the overfitting would be more thus having more data leaks [2].

The experiment done here for privacy protection mainly focuses on how to learn without direct access to the main data record. The algorithm would remain the same for non-privacy preserving case. This also holds for model trained by encrypted data. The researches have also mentioned that they have applied this experiment setup to linear and logistic regression, support vector machines, risk minimization, deep learning etc. over a discrete population using random samples and it was quite successful. Thus, this paper also holds on real world scenarios regardless of privacy and only certain conditions like High confidence distribution, selection of training and testing data etc. should remain constant.

VII. CONCLUSION

Data leakage is a very crucial problem and Membership Inference Attack can do it with the given black box setup. The researchers have very well tried to portray the importance of mitigating it and have discussed few ways to mitigate it. They proved that this problem holds regardless of dataset used and can be done using and ML algorithm.

They have also mentioned that their attack experiment setup can be used as one of the selection metrics when choosing the type of model to train or using ML as a service.

Their main innovation is shadow models which trains attack model to recognize target models member. They also have tried to show how shadow models can be generated in different ways and how each of them performs. Thus, in a gist this paper really strives to achieve privacy implications from their results.

REFERENCES

[1] Demystifying the Membership Inference Attack. URL: https://medium.com/disaitek/demystifying-the-membership-inference-attack-e33e510a0c39

[2] Membership Inference Attack Against Machine Learning Models. URL: http://home.iitk.ac.in/~anmolp/Review_MIA_paper.pdf