Introduction:

Growing integration of machine learning (ML) into important applications ranging from banking to healthcare raises privacy and security issues. Although their great performance is appealing, ML models can learn from sensitive datasets and so expose themselves to risks that take advantage of the learnt representations. For that reason, the concern for data privacy and security has also become essential. Among the emerging threats, Membership Inference Attacks (MIA)- was first introduced introduced by Reza Shokri and his colleagues in their 2017 research paper (Membership Inference Attacks Against Machine Learning Models), a way that allows adversaries to determine whether a specific data point was part of a machine learning model’s training dataset. MIA poses serious privacy risks especially when the data is sensitive, such as medical records or financial transactions.Training several shadow models to replicate the behaviour of a target model first proved the viability of MIA in black-box environments in Shokri et al. (2017).

Types of Membership Inference Attacks:

White-box attacks: A white-box attack happens when the adversary have enough knowledge about trained model, including model architecture, weights and parameters, gradients and loss values.

Black-box attacks: A black-box attacker does not have the model’s inside knowledge means they have no prior knowledge of architecture, weights and parameters, gradients.Attacker determines if a data point was used in the model by observing the confidence of the output.

Gray-box attacks:

This type of attackers have partial knowledge about model architecture and also may have a little access to training data.

Shadow model attacks:

The attacker trains shadow models that mimic the target model's behavior, then queries them with training and non-training data to observe differences. This information is used to train an attack model that can infer whether inputs were used during training.

Metric-based attacks:

Unlike shadow model attacks, which involve training new models to mimic the target model, metric-based attacks are simpler. They look directly at the output of the target model, observes the output confident and use that information to guess whether a specific data point was part of the training set..

Likelihood ratio attacks:

Noisy label/confidence score attacks:

Literature Review:

Introduction:

This chapter reviews several existing solutions following the research of using different membership inference attacks on different sectors with defence mechanisms.It has been observed that in the previous research many mechanisms and features were employed, particularly several Membership Inference Attacks algorithms. This section has examples of several analyses, each having its own set of defense mechanisms and remedies.

ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models

The study discusses loosened adversarial models that operate under more realistic limitations- finally needing neither model nor data access.By challenging conventional wisdom and assessing the viability and efficiency of fresh adversarial models, the ML-Leaks article broadens the horizon of MIAs. It shows that even with low assumptions, effective inference results from which almost any implemented ML model is maybe vulnerable. This disproves past ideas that just certain, overfitted, or poorly used models are vulnerable.

This Study showed three varieties of adversaries, each needing less past information than the next, under relaxed adversarial assumptions. These comprise a completely model and data-independent adversary, an adversary utilising only one shadow model, and an adversary using data from another distribution.

Showing the wide relevance of MIAs, they undertake extensive experiments across eight different datasets including image (CIFAR, MNIST), text (20 Newsgroups), and structured data (Purchase, Adult).

The authors offer two defenses,dropout and model stacking—which reduce MIA success while maintaining model value.

Adversarial Models and Methods

ML-Leaks is strong in its methodical assessment of attackers under progressively limited limits. Every opponent stands for a particular degree of realism about attacking capacity.

Adversary 1:

This adversary shows that one shadow model suffices to produce almost equal performance as utilising many, hence improving on the baseline by Shokri et al.(2017). The adversary's performance in terms of precision and recall remains strong even with looser architectural knowledge, hence proving the redundancy of several shadow models in many situations.

Adversary 2:

This approach eliminates the need for training data coming from the same distribution which is more realistic scenario than the first adversary type. The attacker nevertheless does successful inference by teaching a shadow model on a totally unrelated dataset. For example, a shadow model trained on the 20 Newsgroups text dataset may efficiently target a CNN built on CIFAR-100 picture data. Dubbed the data transferring attack, this amazing finding shows that MIAs rely more on learning overfitting behaviour than on accurate data or task similarity.

Adversary 3:

The third opponent calls neither data nor a shadow model. Rather, it only searches the target model and bases membership on statistics including maximal posterior or entropy of the output. Surprisingly effective, the attack earns high AUC ratings in several scenarios. To estimate the non-member distribution, the authors suggest a percentile-based threshold selecting based on random inputs.

Results of Experimentation

By means of evaluation on datasets like MNIST, CIFAR-10, CIFAR-100, Purchase-100, and others, this study demonstrates that all three adversaries can succeed in membership inference under various assumptions. Their findings repeatedly reveal: the main weakness is overfitting. Training and test accuracy directly correspond with MIA success rates.

Model-agnostic attacks (Adversary 3) even unsupervised are successful, particularly in cases when the model is quite confident on training data.

Protection Strategies

Salem et al. suggest two sensible defences to offset these attacks:

Dropout:

Dropout is a common regularising method used in training to randomly remove connections in the network therefore preventing overfitting. While in some circumstances it maintains or even increases model accuracy, it greatly lowers MIA performance.

Model Staking:

Though the Dropout mechanism is effective, this mechanism works for deep neural networks.For other machine learning classifiers this study introduces model staking mechanism- is an ensemble-based defence is training several models on disparate data subsets utilising their outputs to feed a meta-model. This produces non-overlapping data exposure, which makes overfitting pattern discovery more difficult for attackers.

Moreover, their defence ideas provide sensible trade-offs between utility and privacy. The research combines privacy preservation with good machine learning techniques by connecting overfitting reduction with MIA mitigating – an idea that can be generally used.

Salem et al.'s ML-Leaks publication marks a historic progress in the field of membership inference attacks' research. The authors clearly show that membership inference is not only possible but shockingly generalisable by building and analysing a range of adversaries ranging from strong, model-aware attackers to totally data-agnostic ones.

Their results highlight the critical need of privacy-preserving training methods and model deployment strategies particularly as ML gets more included into consumer and business environments. Dropout and model stacking are among the suggested defences that provide easily available answers balancing strong usefulness with privacy resilience.