**ARTICLE : A Survey of Privacy Attacks in Machine Learning**

**AUTHORS : MARIA RIGAKI, SEBASTIAN GARCIA**

To understand and defend against attacks in machine learning from a privacy perspective, it is useful to have a general model of the environment, the different actors, and the assets to protect. From a threat model perspective, the assets that are sensitive and are potentially under attack are the training dataset D, the model itself, its parameters 𝜃, its hyper-parameters, and its architecture. The actors identified in this threat model are:

* **(1) The data owners**, whose data may be sensitive.
* **(2) The model owners**, which may or may not own the data and may or may not want to share
* information about their models.
* **(3) The model consumers**, that use the services that the model owner exposes, usually via

some sort of programming or user interface.

* **(4) The adversaries**, that may also have access to the model’s interfaces as a normal consumer

does. If the model owner allows, they may have access to the model itself.

**Figure 1** depicts the assets and the identified actors under the threat model, as well as the

information flow and possible actions. This threat model is a logical model and it does not preclude

the possibility that some of these assets may be collocated or spread in multiple locations.

Diagram

Description automatically generated

Figure 1 : Threat Model of privacy and confidentiality attacks against machine learning systems.

**ARTICLE : A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View**

**AUTHORS : QIANG LIU, PAN LI, WENTAO ZHAO, WEI CAI, SHUI YU AND VICTOR C. M. LEUNG**

**Categories of threats to IA:**

In the perspective of the influence on classifiers, security threats towards machine learning can be classified into two categories: (a) ***Causative attack***. It means that adversaries have the capability of changing the distribution of training data, which induces parameter changes of learning models when retraining, resulting in a significant decrease of the performance of classifiers in subsequent classification tasks. (b) ***Exploratory attack***. Such attack does not seek to modify already trained classifiers. Instead, it aims to cause misclassification with respect to adversarial samples or to uncover

sensitive information from training data and learning models.

In the perspective of the security violation, threats towards machine learning can be categorized into three groups: (a) ***Integrity attack***. It tries to achieve an increase of the false negatives of existing classifiers when classifying harmful samples. (b) ***Availability attack***. Such attack, on the contrary, will cause an increase of the false positives of classifiers with respect to benign samples. (c) ***Privacy violation attack***. It means that adversaries can obtain sensitive and confidential information from training data and learning models.

In the perspective of the attack specificity, security threats towards machine learning have two types as follows: (a) ***Targeted attack***. It is highly directed to reduce the performance of classifiers on one particular sample or one specific group of samples. (b) ***Indiscriminate attack***. Such attack causes the classifier to fail in an indiscriminate fashion on a broad range of samples.

**Poisoning attack :**

The poisoning attack is a type of causative attack, which disrupts the availability and the integrity of machine learning models via injecting adversarial samples to the training data set. Typically, such adversarial samples are designated by adversaries to have similar features with malicious samples but incorrect labels, inducing the change of training data distribution. Therefore, adversaries can reduce the performance of classification or regression models in terms of accuracy.

These systems are generally required to periodically update their decision models to adapt to varying application contexts. Taking the adaptive FRS (*Facial Recognition System*) as an example, an attacker utilizes the periodic update characteristic and injects masqueraded samples into the training data used for retraining decision models, resulting in the change of normal data classification centroid.

**Method for generating adversarial sample :**

Recently, it is worth noticing that a more effective method for generating adversarial samples is to adopt generative adversarial network (**GAN**), which consists of a generative model and a discriminative one. Specifically, the generative model is trained to generate candidate adversarial samples. Then, the discriminative model is used to select the optimal samples with a specific loss function. Comparative results between GAN and **direct gradient methods** on MNIST and CIFAR-10 data sets validated that **GAN was able to rapidly generate high-quality adversarial samples with a larger loss value**.

**Evasion attack :**

Evasion attack was proposed to compromise machine learning in information security : spam detection, PDF malware detection… The main idea of this attack is that an attacker generates some adversarial samples that are able to evade detection such that the overall security of target systems is significantly reduced. There are several studies on attacking and defense echniques with respect to evasion attacks. For example, the authors in [55] proposed **to generate the optimal adversarial samples to evade detection via gradient algorithms**. Recent studies also demonstrated that evasion attacks were feasible for use to attack against FRS and malware detection in real world, resulting in severe security threats towards target systems.

**Impersonate attack :**

Similar to the above attack, the impersonate attack prefers to **imitate data samples from victims**, particularly in application scenarios of image recognition, malware detection, intrusion detection based on machine learning. Specifically, an attacker aims to generate specific adversarial samples such that existing machine learningbased systems wrongly classify the original samples with different labels from the impersonated ones. By doing so, the attacker can gain the victims' authority in practical access control systems. **Such attack is particularly effective in attacking DNN algorithms** because DNN usually extracts a small feature set to facilitate the object identification. Hence, an attacker can easily launch impersonate attacks by modifying some key features. Moreover, there are many impersonate attacks to imitate images.

**Model inversion :**

Given the access to a model, inversion attack aims to recover private training given some target label. Two types of inversion attack : **Blackbox** or **Whitebox**. In Blackbox, the attacker can only query the model whereas in Whitebox, the attacker has access to the model parameters. In recent years, many platforms release the parameters of their model. Therefore, there is an urgent need to evaluate the privacy risk in the Whitebox setting.

**Membership inference / Extraction:**

A type of attack called “**membership inference**” makes it possible to detect the data used to train a machine learning model. In many cases, the attackers can stage membership inference attacks without having access to the machine learning model’s parameters and just by observing its output. Membership inference can cause security and privacy concerns in cases where the target model has been trained on sensitive information. The **MIA** (Membership Inference Attack) is the process of determining whether a sample comes from the training dataset of a trained ML model or not.

**WIKI : Adversarial ML**

An **adversarial example** refers to specially crafted input which is design to look "normal" to humans but causes misclassification to a machine learning model. Often, a form of specially designed "noise" is used to elicit the misclassifications. One can design adversarial examples for Whitebox attacks or Blackbox **evasion attacks**.

Whitebox attacks :

* Fast Gradient Sign Method (FGSM)
* Carlini & Wagner (C&W)