SoK: Security and Privacy in Machine Learning
Paper Review
Shivam Bhat
bhat41@purdue.edu
10/23/2022

Summary

Machine Learning has today permeated most aspects of our lives. Applied ML has found its application in various domains like Autonomous Systems, Data Analytics and Security. With the rapid advancements and cheap availability of computing resources, machine learning models are today being deployed at a rapid pace to facilitate software based inference and decision making. However, often a lot of focus is put on the accuracy of the model and not on their security aspects. This has resulted in these software systems being exposed to newer vulnerabilities. This paper attempts to systematize and consolidate the knowledge around Machine Learning Security and Privacy.

While expanding on the topic, the author proposes a unified threat model which departs from existing contemporary research by considering the entire pipeline instead of just the machine learning algorithms in isolation. This is followed up with a thorough analysis of the attack surfaces, adversarial goals and possible defense strategy against such attacks. The author also does a good job of providing a gentle introduction to machine learning concepts and nomenclatures, whilst also mapping classical security aspects of confidentiality, integrity, and availability (CIA) model to them. While the paper does seminal work in this broader area of trustworthy ML, there were few assumptions and concepts that could have been further expounded upon, particularly how the two mentioned attack types - black and white box are achieved in real life scenarios and not just their characteristics and environments. Further there are some broader assumptions being made about parameters and hypothesis of machine learning models which often differ depending on the chosen algorithm. The cost functions, learning strategy ,data compatibility, tuning parameters etc are different for each algorithm and hence generalizing might not be correct. The paper, nonetheless, is very much relevant and also the first to propose such an elaborate framework to analyze and reason about various threat models for applied machine learning in security. The detailed process, my comments and suggestions are listed in the next section.

Comments

The rapid advancement in machine learning, availability of cheap hardware and computing resources coupled with the rise of ML as a service on cloud has today made machine learning ubiquitous in the digital world. Machine learning is being applied across industries in autonomous vehicles, aviation, health care, data analytics etc. However this has meant an exposure to new vulnerabilities for such software systems trying to leverage machine learning. Since a lot of ML research is more focused on improving the efficacy and accuracy of such algorithms, the security and privacy of machine learning systems is today still an active yet relatively nascet area. In an effort to further research in this area , this paper consolidates the knowledge around security and privacy of ml systems and eventually helps develop countermeasures. To help develop threat models, the paper views the systems from the prism of the CIA model(confidentiality, integrity, and availability). In this context, confidentiality attacks are defined as attacks on model or data which reveal the model parameters or data structure. Such attacks affect the data privacy aspects. Similarly attacks on integrity are defined as those where an adversary is able to elicit a specific output or behavior. Lastly, in an availability attack, an adversary tries to prevent authorized or legitimate users from accessing feature or model outputs

The paper does a good job of providing a brief introduction to various classes of machine learning techniques, namely Supervised Learning, Unsupervised Learning and Reinforcement Learning. This is followed by a discussion on the training and inference stages of model learning where the reader is also exposed to concepts like loss function, parametric tuning, generalization, feature space etc. In the following discussion on threat models and attack surfaces we learn how the attack surfaces for the system can be defined with respect to the data processing pipeline. An adversary agent can try to attack by manipulating data collection, cause model corruption or modify outputs. In the case of an online environment , attacks can be quite impairing, as the adversary can slowly but eventually alter the model using modified inputs supplied at runtime. Such attacks have been previously seen on anomaly detectors used for spam or network intrusion detection

In the discussion on Adversarial capabilities we learn about the range of capabilities possessed by the adversary while attacking in the inference or training phase. The author classified the attacks into two types - black and white box attacks. A white box attack is an attack in which the adversary has some information about the model or its training methods (in inference phase) whereas in case of a black box attack the adversary has no knowledge about the model and hence depends on setting or past inputs to discover vulnerabilities in the

model. We also learn how in training phrase attacks the adversary can manipulate the learning algorithm easily by colluding with untrusted ML training components. Such attacks are broadly referred to as logic corruption and are difficult to protect against.

While discussing Adversarial goals, which are the last but essential components of a threat model, the paper focuses on four properties - confidentiality, integrity, availability and privacy. It's observed that system integrity and availability are closely related in goal and method, as are confidentiality and privacy. Integrity and availability attacks are defined with respect to model outputs where the goal is to incite the model into behaving as desired by the adversary for example in case of an integrity attack on a face recognition system, false positives are induced to discredit the authentication process's integrity. In case of availability , the goal of an attacker is to make the model inconsistent.

We also learn about poisoning attacks which modify the training dataset by inserting, editing, or removing data points in order to modify the model's decision boundary. Such modification also changes the distribution of training and inference data leading to a phenomenon called "distribution drifts". If the adversary is only able to change the data labels, its attack surface gets restricted and hence it needs to find the most damaging labels to perturb in the data. In the case of SVM classifiers it was shown that an adversary only needed to flip 40% of the training labels. In case of a 'Input Manipulation' threat model, the adversary can corrupt the training features of the model in addition to the labels. For models with an online learning phase, the attack surface is further exacerbated since there is a direct poisoning of the learning inputs. For example, in the case of clustering algorithms, Kloft et al. showed in their paper how data poisoning shifted the decision boundary of the centroid model. Opposite to this, in case of offline learning, Biggio et al. in their paper introduced an attack which identified poisoning points by gradient ascent which when added back to the training set resulted in degraded classification accuracy.

In the closing sections of the paper, the authors discuss open problems related to robustness to distribution drifts, fairness and learning privacy-preserving models. We learn how most of the present methods rely on poisoned samples being out of input distribution for their defense strategies. In one such approach taken by Rubinstein et. al, a constrained PCA-based detection model is used to limit the influence of outliers to training distribution. In another approach, Biggio et. al limit the vulnerability of SVMs by adding a loss regularization term. For a model to be privacy-protecting , it should not reveal any additional information about subjects during the model's training phase. This is beautifully encapsulated by a framework known as "differential privacy". To ensure differential privacy , it's

necessary to randomize parts of ML systems' pipeline. This is also shown by Chaudhuri et. al in their paper titled "Differentially private empirical risk minimization".

Recommendations and takeaways

Certain generalizations made about ML algorithms might not be true in all cases. While on a broader level all algorithms try to generalize a model for unseen data but on a lower level most machine learning algorithms differ in terms of learning methodology, cost function, hypothesis, hyper parameters, data ingestion, accuracy etc.

Further, although the author has extensively discussed the black and white box attack model's working and characteristics, very little was discussed about what leads to or allows creation of such attacks. Discussion about the origins of such attacks and what leads to creation to white box over black box and vice versa could have been helpful.

Lastly the table in figure 4 was difficult to comprehend due to usage of reference numbers instead of contextual information. This required the user to scroll till the end of the page whenever they wanted to read the table and hence made this less accessible.