SoK: Security and Privacy in Machine Learning

Sowmya Jayaram Iyer
jayarami@purdue.edu
PUID: 0033742039

SUMMARY:

The paper presents a large, rather unnecessarily detailed, information on threat model for a ML algorithm and its environment and a survey on attacks and how to defend them. The three main contributions mentioned were (1) a unified threat model which is not limited to ML algorithm but also its data pipeline, as a basis for introducing the existing security and privacy of systems that use or is built upon ML. (2) Taxonomizing attacks when model is not trained using adversarial data as well as when it is, and also, defenses in existing literature especially elaborating on differentially private learning. (3) Highlighting properties with scope of improvement for robust machine learning. Though detailed, this paper was far from being thorough. The misuse of words such as "this security model", "taxonomize attacks" misleads the reader at various points in the paper. Though the attacks and defenses presented in this paper focusses mainly on classification tasks in ML, the threat model was generalized and elaborated well with all attack surfaces well-defined. With more evidences for topics mentioned in the comments, and by presenting a more variegated types of attacks and corresponding defense methods, this paper might have a good standing. That being said, this paper still provides little to no contribution towards existing literature and the structuring of the content needs attention.

DETAILED COMMENTS:

The abstract of the paper, particularly the second part, gives a decent description of the paper's contents. However, it fails to highlight the significance of this paper. The author mentions about constructing a PAC theory for better understanding of how sensitivity of modern ML algorithms to the data they analyze which in turn helps understand the science behind security and privacy in them. Though the latter was elaborated one, the prior was not properly introduced as PAC and the link was hence not properly established.

This paper has a concrete goal of introducing readers to various attacks in ML, their nature, their target and their outcomes, and presenting current findings about defenses against them. The author in Page 399 describes his work as "This security model serves as a roadmap for surveying knowledge about attacks and defenses of ML systems.". It is important to note that no such model is discussed in the paper and simply

current works were surveyed. The content of the latter part of Introduction section (Section 1) were repetitive. A more concise description of the paper's work would establish a more interesting and engaging introduction. In Section 2 the author gives a brief introduction to various parts of ML providing a nice revision on concepts in ML and hence gives a more in-depth understanding of various attack scenarios discussed in the later sections.

In section 3, the author begins with a claim that "we taxonomize the definition and scope of threat models in ML systems and map the space of security models". However, taxonomizing in research holds an entirely different meaning from the author's work which is merely introducing the concepts in the field of security and elaborating on them. In section 3.1, the author says that this work, unlike previous works, takes even the data pre-processing step in ML into consideration. However, that attacks mentioned mainly focused on training and inference attacks and sufficient attention to attacks reliant on data-preprocessing was not given to back such a grand claim. Otherwise, in section 3, the author gives a good introduction to black box attacks and white box attacks and what meaning they hold in the domain of ML. Further, the difference between attacks in training and inference phase was well defined using simple yet sufficient examples. The author also differentiates well between Confidentiality and Privacy in ML and also how integrity and Availability of Ml are compromised using essentially the same attack methods with neat examples to substantiate them.

In section 4, the author begins explaining training in adversarial settings with a rather specific attack – poisoning. And also, the dependency of attacks on decision boundaries is elaborated only with respect to SVMs. A more generalized approach of compromise of Integrity in ML would have made the attacks taken into account more compelling. A very important note to be made is that, the author only elaborates on poisoning attack for inferring model parameters. This paper was compiled in 2018 and by 2017, two other attacks of significant importance in ML domain were popular called *Trojaning*, where the existing behavior of the model stays even after the attacker has changed its behaviors in some circumstances, and *Backdooring*, a difficult to trace backdoor is injected to a model along with additional behaviors which stay even after a model has been retrained. These were hardly mentioned in the survey presented.

In section 5, the author moves towards Inferring in Adversarial settings and explains it separately for white-box adversary and black-box adversary scenarios. The white-box adversarial references provide a good understanding of size of perturbations, their dependency on the feature space of the model and their effect on model performance. The take-aways mentioned by the author also clearly summarizes the content. In section 5.2, the author shows how an attacker learns the least cost method to create an adversarial example in a black-box scenario. However, how the attacker can directly manipulate model inputs is quite disorganized and vague. The property of

transferability of adversarial examples is said to hold true even when models are trained on different datasets. However, immediately in the next part author comments that the in the reference given "their attack does not generalize well to other application domains or models". Such a contradictory statement is not elaborated upon. Additionally, the transferability property was supported by a statement, "For instance, the attack on a logistic regression hosted by Amazon has a success rate of 96%.", with no evidence or explanation given. In take-away 5.4, author says, two "models solving the same ML task with comparable performance are likely to have similar decision boundaries". A simple counter example would be a decision tree trained on 1000 samples vs a decision tree trained on 10 samples for the same task.

Membership attacks on ML is a highly researched topic in Security and this was not elaborated enough. In training data extraction, author introduces *Model inversion*. He fails to state the similarity or difference between this and the *Reverse engineering* mentioned in Section 3 for they were given the same definition. In section 6.1, the author mentions violation of security fundamentals when defender holds out a set of data as secret. What are these fundamentals? Why wasn't this considered important enough to elaborate upon? The working and failure of Gradient masking method against transferability property is well explained. Though, the failure of defense against larger perturbations using adversarial training when an attacker uses a different heuristic is not elaborated upon. Adversarial training, being one of the major breakthroughs against attacks, deserves more evidence when being voted down.

Finally, the author does establish the dependance on algorithm's sensitivity to training inputs of differential privacy. For models with loose bounds, the citation of randomization and its effects in giving a guaranteed differential privacy shows light on how models can improve security from a Privacy of data perspective.

RECOMMENDATIONS:

This paper does a good job in explaining the scope for improvement in security of ML under various scenarios with respect to various attack surfaces and capabilities of attackers. However, the paper is too narrowed down to specific kinds of models or attacks and hence does not give a generalized road map to security in ML. This paper's attempt to give a complete run down on ML, its working, threat model, attack scenarios and defense formulated till date is undermined by numerous false claims, narrowed application to classification algorithms under very few specific attacks and missing explanation or examples. At its present form, with improvements suggested in the previous section, the paper makes an acceptable case for publication.