A Review of Adversarial Machine Learning (4th ACM Workshop on AISec, Oct. 2011,43-58)

Rohit Das
ID No.: 11910230
M. Tech. (CSE)
Reviewers: Dr. Sk Subidh Ali, and Dr. Subhajit Sidhanta

August 30, 2019

I. Summary

The manuscript by Huang et. al. focuses primarily on the security fault lines in modern Machine Learning, how they can be utilized to break into secure learning systems, and manipulate data and users. It also sequentially outlines how such situations can be avoided by keeping some points in mind while designing a new or improving an existing learning algorithm or model, forming the crux of Adversarial Machine Learning.

II. Outline

Listed below are the milestones covered by the paper:

- A good overview of Adversarial Machine Learning and its impact in the field of modern Artificial Intelligence.
- Taxonomy of influence, security violations and specificity of attack by a potential adversary on a learning system.
- Modeling secure learning systems as a game between an attacker and a defender, and how both can influence the learning algorithm, learning and evaluation of data, as well as the data itself. It also subtly creates the idea that a learning model can be made more robust by artificially creating an adversary and training the model with it.
- An elaborate section on Causative attacks to learning systems, well-documented with graphs and data. It also contains case studies on two algorithms: SpamBayes and Principal Component Analysis (PCA)-based anomalous network traffic detector. The authors have highlighted the strengths of an attacker given certain degrees of freedom, and how they can be countered.
- A section on Exploratory attacks, discussing various evasion techniques, theoretical and practical, and how the most optimum evasion techniques are not the ones generally implemented in real time.
- Privacy violations, and advantages and roles of randomization of functions and parameters in learning models.

III. Best Elements

Below mentioned are the best elements of the paper:

- All conclusions, facts and experiments have been well-supported with relevant data and graphs. The graphs clearly highlight the differences between using a robust and a non-secure learning model.
- The paper mainly focuses on causative attacks, the most malicious of attacks, and extensive research has been done into its types, variants and countermeasures.
- The two algorithms used as case studies are very much used in real life situations, and hence makes the paper very relevant in today's new age of Artificial Intelligence, thereby easily underlining the need for the topic researched on.
- Privacy, being a trending concept throughout history and a fundamental human right, is also a part of this paper, and makes it a must-read for researchers in this field.

IV. Limitations

- A section on image and pattern recognition, and how adversaries can trick such models, could have been made a part of the research, as it is a budding novel area in Machine Learning, and specifically, Deep Learning.
- The paper is a bit too textual. Adding more images would have helped better illustrate the various attacks and countermeasures presented here.

V. Opinion

The paper gives a good introduction and an elaborate explanation on Adversarial Machine Learning, and how to counter it. It aptly highlights the necessity and relevance of securing learning models, especially when they will be planned to be used in mission-critical situations. However, the organization of the paper could be improved by including more references in the analyses.