Fostering Trustworthiness in Machine Learning Algorithms

Mengdi Huai

Department of Computer Science, Iowa State University
mdhuai@iastate.edu

Abstract

Recent years have seen a surge in research that develops and applies machine learning algorithms to create intelligent learning systems. However, traditional machine learning algorithms have primarily focused on optimizing accuracy and efficiency, and they often fail to consider how to foster trustworthiness in their design. As a result, machine learning models usually face a trust crisis in real-world applications. Driven by these urgent concerns about trustworthiness, in this talk, I will introduce my research efforts towards the goal of making machine learning trustworthy. Specifically, I will delve into the following key research topics: security vulnerabilities and robustness, model explanations, and privacy-preserving mechanisms.

Introduction

Machine learning is integral to machine intelligence, encompassing techniques that allow computers to identify patterns in data through adaptable modeling methods and variable selection. These techniques play a crucial role in the development of artificial intelligence (AI) applications. In recent years, there has been a significant increase in research that develops and applies machine learning algorithms to create intelligent learning systems, such as self-driving cars, recommendation systems, and medical diagnostic systems.

However, traditional machine learning algorithms primarily emphasize optimizing accuracy and efficiency, and they often fail to consider how to foster trustworthiness in their design. Trustworthiness reflects the degree of a user's confidence that the deployed machine learning system will function as expected under various challenges, such as human errors, system faults, malicious attacks, or environmental disturbances. Key elements of trustworthiness include model transparency, robustness against malicious attacks, privacy preservation, and fairness. Without fully studying the trustworthiness of real-world intelligent learning systems, we will face a variety of severe social and environmental consequences. For instance, in autonomous driving, attackers could manipulate the output of autonomous vehicles' perception systems via slightly changing the driving environment, which can potentially lead to a range of catastrophic outcomes, ranging from a life-threatening accident to major disruptions in transportation services.

Motivated by the above, in this talk, I will cover the following research topics: First, I will investigate the security vulnerabilities and robustness of machine learning algorithms to analyze adversaries' actions in adversarial environments and thereby derive robust learning strategies. In particular, I will delve into the study of malicious attacks based on machine unlearning, a new learning paradigm that aims to erase (or unlearn) the impact of some training data from a well trained model and generate an unlearned model without needing to retrain it from scratch (Zhao et al. 2023; Qian et al. 2023; Bourtoule et al. 2021). For such attacks, the adversaries are some malicious training data providers, and their goal is to induce malicious behavior in the unlearned model by submitting deceptive unlearning requests. Note that existing works that study the vulnerabilities of machine learning models to adversarial and data poisoning attacks only focus on the testing and training stages, and they fail to address the failure modes of models caused by malicious unlearning attacks during the unlearning stage. Subsequently, I will discuss model interpretation methods that can provide insights into machine learning models' working mechanisms by interpreting what they have learned and hence help increase the trust in model decisions. Finally, I will explore privacy-preserving mechanisms designed to share sensitive private information in a secure manner.

References

Bourtoule, L.; Chandrasekaran, V.; Choquette-Choo, C. A.; Jia, H.; Travers, A.; Zhang, B.; Lie, D.; and Papernot, N. 2021. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, 141–159. IEEE.

Qian, W.; Zhao, C.; Le, W.; Ma, M.; and Huai, M. 2023. Towards understanding and enhancing robustness of deep learning models against malicious unlearning attacks. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1932–1942.

Zhao, C.; Qian, W.; Ying, Z.; and Huai, M. 2023. Static and Sequential Malicious Attacks in the Context of Selective Forgetting. In *Thirty-seventh Conference on Neural Information Processing Systems*.