A framework for robustifying linear regression models against adversarial manipulations

P. G. Arce, C. Lopez Amado, R. Naveiro Flores and D. Rios Insua

Abstract

In addition to the positive developments attributable to machine learning, significant misuses have also been reported. Many of these issues arise from attempts by adversaries to fool machine learning algorithms to attain a benefit, giving rise to the relatively recent field of adversarial machine learning. Conventional machine learning models rely on the assumption of independent and identically distributed data during both training and operations. However, in a wide range of applications, it is important to consider that an adversary may modify model inputs, thereby altering the incumbent distributions and potentially compromising the integrity of the system. Consequently, adversarial machine learning aims to provide algorithms that are more robust against adversarial manipulations, thus enhancing the security and reliability of machine learning systems. This field primarily focuses on three key issues: studying attacks to machine learning algorithms to understand their vulnerabilities, designing effective defenses against such attacks to better protect the algorithms and the systems they serve, and providing frameworks that determine the best pipelines to mitigate potential attacks while maintaining optimal performance.

This paper introduces a pipeline aimed at enhancing the robustness of linear regression models against adversarial attacks. The pipeline incorporates strategies to protect models throughout both their training and operational phases, drawing from prior research in classification to tailor approaches to the linear regression setting. During training, the pipeline implements robust training methodologies, imposing a regularity condition on the learned distribution to fortify the model against adversarial manipulation during inference. Similarly, protection during operations involves adjustments to the inference process to counteract potential attacks. Both forms of protection are designed in a Bayesian manner. Additionally, the pipeline integrates techniques for detecting attacks and monitoring changes

in attack patterns. A change in attack patterns showcases the need to adapt the pipeline, either by retraining (in a robust manner) or adjusting inference. Several examples illustrate the role of the pipeline and compare the results against known defenses.

Keywords: Adversarial Risk Analysis; Machine Learning; Regression; Bayesian.