

Hugging Face

The science of defenses for machine learning is somewhat less well developed. Here we consider several defensive goals. We introduce a unifying threat model to allow structured reasoning about the security and privacy of systems that incorporate machine learning. We explore defenses that provide fairness (preventing biased outputs) and accountability (explanations of why particular outputs were generated, also known as transparency). We taxonomize attacks and defenses as informed elements of PAC learning theory. And we introduce a no free lunch theorem for adversarial machine learning, which characterizes the tradeoff between accuracy and robustness to adversarial efforts, especially when learning from limited data.

ChatGPT

This passage discusses the challenges and goals in defending machine learning systems. It outlines three main defensive objectives: robustness to distribution drifts during training and inference, formal privacy guarantees through differential privacy, and the achievement of fairness and accountability in model outputs. The text introduces a comprehensive threat model, departing from previous efforts by considering the entire data pipeline rather than isolating machine learning algorithms. It also categorizes attacks and defenses using PAC learning theory, addressing challenges in adversarial settings and presenting desirable properties to enhance security and privacy in machine learning. Additionally, the passage introduces a "no free lunch theorem" in Section VII, highlighting the tradeoff between accuracy and robustness to adversarial efforts when learning from limited data.