Abstract

With the pervasive deployment of machine learning algorithms in mission-critical AI systems, it is imperative to ensure that these algorithms behave predictably in the wild. Current machine learning algorithms rely on a tacit assumption that training and test conditions are similar, an assumption that is often violated due to changes in user preferences, blacking out of sensors, etc. Worse, these failures are often silent and difficult to diagnose.

We propose to develop a new generation of machine learning algorithms that come with strong static and dynamic guarantees necessary for safe deployment in open-domain settings. Our proposal focuses on three key thrusts: *robustness* to context change, *inferring the underlying process* from partial supervision, and *failure detection* at execution time. First, rather than learning models that predict accurately on a target distribution, we will use minimax optimization to learn models that are suitable for *any* target distribution within a "safe" family. Second, while existing learning algorithms can fit the input-output behavior from one domain, they often fail to learn the *underlying reason* for making certain predictions; we address this with moment-based algorithms for learning latent-variable models, with a novel focus on structural properties and global guarantees. Finally, we propose using *dynamic testing* to detect when the assumptions underlying either of these methods fail, and trigger a reasonable fallback. With these three points, we aim to lay down a framework for machine learning algorithms that work reliably and fail gracefully.