The Falsificationist View of Machine Learning

Popperian falsificationism provides a sobering view about scientific progress–an insight generally neglected by engineers. Applying modern scientific advancements requires making decisions in a highly complex environment, but optimizing performance often sacrifices robustness.

Nassim Nicolas Taleb argues that the 21st century challenges humanity with Black Swans–highly improbable events with considerable losses**.** Such events are the reality of solutions in medicine, autonomous driving, and finance–often powered by artificial intelligence. Although researchers made notable progress in protecting neural networks against adversarial examples and in quantifying uncertainty, the authors argue that the field could benefit from the principles of Popper's philosophy.

The belief of obtaining reliable, task-specific models with a limited amount of data and the ever-increasing state-of-the-art performance obscure the fragility of the quest for the perfect decisions in a noisy setting: the need for a decision disregards whether the best solution is superior. Given the noise and the highly unexpected, models can notoriously fail.

The Popperian flavor of mathematical methods is not unknown: statistical hypothesis testing provides conclusions based on falsifying hypotheses. This paper examines modern machine learning methods in the falsificationist context, arguing that constraining the hypothesis space would improve decision quality: the illusion of an unambiguous decision is less likely but more informative.

First, we discuss the learning paradigms of supervised, self-supervised, unsupervised, and reinforcement learning. We conclude that the Popperian approach is applicable for all but unsupervised learning–as the latter paradigm lacks hypotheses. We emphasize that the supposed controversy of supervision—as Popper denies the access to the ground truth—in (self-)supervised learning is only due to different terminologies. Although the desired output is present, the scientific inquiry is about the mapping itself, which is unavailable.

Second, we contrast classification and regression methods–pointing out that falsificationism naturally fits only the former. Nonetheless, this still elucidates many methods from reinforcement learning to self-supervised learning or generative adversarial networks. The unique role of ensembles–namely, enabling regression tasks' entrance into the Popperian framework–is also within the scope of our analysis.

By providing an epistemological context for machine learning algorithms, we hope to inspire a discussion that helps to ensure the robust deployment of artificial intelligence.