Abstract

We study a sequential prediction problem in which an adversary is allowed to inject arbitrarily many adversarial instances in a stream of i.i.d. instances, but at each round, the learner may also *abstain* from making a prediction without incurring any penalty if the instance was indeed corrupted. For this problem, Goel et al. (2023) showed that, if the learner knows the distribution μ of clean samples in advance, learning can be achieved for all VC classes without restrictions on adversary corruptions. Knowledge of μ, however, is a strong assumption for deployed learning systems: the clean distribution is typically unknown and may be represented only through samples. To learn without knowledge of the distribution μ, we first provide a weaker learner (agent) for the simpler problem where a very small set of clean samples are available. Then, we design a boosting strategy that adaptively combines suggestions from those agents, so that the final algorithm AbstainBoost eventually achieves sublinear error for general VC classes in *distribution-free* abstention learning for oblivious adversaries. Our algorithm also enjoys similar guarantees for adaptive adversaries, for structured function classes including linear classifiers. These results are complemented with corresponding lower bounds, which reveal an interesting polynomial trade-off between misclassification error and the number of erroneous abstentions.

Link to paper: https://arxiv.org/abs/2602.17918