STEVENS
INSTITUTE OF TECHNOLOGY
1870

Ph.D. DISSERTATION DEFENSE

Candidate:	Shiwei Zeng
Degree:	Doctor of Philosophy
School/Department:	Charles V. Schaefer, Jr. School of Engineering and Science, Department of Computer Science
Date:	Wednesday, April 24th, 2024
Time/Location:	12:00 p.m. / https://stevens.zoom.us/my/jieshen
Title:	Data-Efficiency and Robustness in Machine Learning
Chairperson:	Dr. Jie Shen, Department of Computer Science, Stevens Institute of Technology
Committee Members:	Dr. Wendy Hui Wang, Department of Computer Science, Stevens Institute of Technology Dr. Philippos Mordohai, Department of Computer Science, Stevens Institute of Technology Dr. Shucheng Yu, Department of Electrical and Computer Engineering, Stevens Institute of Technology Dr. Yudong Chen, Department of Computer Science, University of Wisconsin-Madison Dr. Yu-Xiang Wang, Department of Computer Science, University of California, Santa Barbara

ABSTRACT

Machine learning has been a powerful tool in the modern world. In the past decades, due to the explosion of unverified data sources, and the increasing interaction between human and computer, it is of concern whether machine learning algorithms are robust to data corruption or even adversarial attacks. On the other hand, in many real-world scenarios, it is hardly the case that we can gather enough data for training a good model, making data-efficiency another important aspect of algorithmic designs.

In this talk, I will address these concerns by presenting how to design machine learning algorithms that are provably robust to noise and have efficiency guarantees in terms of the computation, amount of labels, queries, and total number of samples. Several challenging learning settings are considered. Concretely, under the crowdsourced learning setting, where the unlabeled instances are given to the learner and the learner can choose to make queries from a pool of crowd workers, I will present algorithms that can generalize from the noisy crowd even when the majority is incorrect. Moreover, the algorithms are query and label-efficient, i.e. they make a constant number of queries on each unlabeled instance and in total only a logarithmic number of labels. Under the list-decodable learning setting, I will discuss the fundamental problem of mean estimation and present an attribute-efficient algorithm that can recover a list of hypotheses at least one of which is close enough to the ground truth. For the problem of learning polynomial threshold functions under the nasty noise, I will present an attribute-efficient algorithm that outputs a function that enjoys PAC guarantees with dimension-independent noise rate. When the underlying models have sparse structures, an attribute-efficient algorithm enjoys a sample complexity that depends polynomially on the sparsity parameter and poly-logarithmically on the ambient dimension.