Fair and Optimal Prediction via Post-Processing

Han Zhao

Department of Computer Science, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA
hanzhao@illinois.edu

Abstract

In this talk I will discuss our recent work on characterizing the inherent tradeoff between fairness and accuracy in both classification and regression problems. I will also present a post-processing algorithm that derives optimal fair predictors from Bayes score functions.

Introduction

With the development of machine learning algorithms and the increasing computational resources available, artificial intelligence has achieved great success in many application domains, including computer vision, natural language processing, healthcare, etc. However, the success of machine learning has also raised concerns about the *fairness* of the learned models. For instance, the learned models can perpetuate existing bias and discrimination in the training data. This issue has become a major obstacle to the deployment of machine learning systems in high-stakes domains, e.g., criminal judgment, medical testing, online advertising, hiring process, etc. To address these issues, it is crucial to understand the potential tradeoff between fairness and accuracy and develop fair machine learning algorithms (Zhao et al. 2019; Hu et al. 2023; Chi et al. 2022, 2021; Wang, Li, and Zhao 2022; Zhao et al. 2022) that are optimal in terms of both efficiency and accuracy.

To mitigate the bias exhibited by machine learning models, fairness criteria can be integrated into the training process to ensure fair treatment across all demographics, but it often comes at the expense of model performance. Understanding such tradeoffs, therefore, underlies the design of optimal and fair algorithms. In this talk, I will discuss our recent work on characterizing the inherent tradeoff between fairness and accuracy in both classification and regression problems (Zhao and Gordon 2022; Xian, Yin, and Zhao 2023; Zhao 2022), where we show that the cost of fairness could be characterized by the optimal value of a Wasserstein-barycenter problem. Then I will show that the complexity of learning the optimal fair predictor is the same as learning the Bayes predictor, and present a post-processing algorithm based on the solution to the Wasserstein-barycenter problem that derives optimal fair predictors from Bayes score functions.

Acknowledgments

The work of Han Zhao was partially supported by the Defense Advanced Research Projects Agency (DARPA) under Cooperative Agreement Number: HR00112320012, a research grant from the Amazon-Illinois Center on AI for Interactive Conversational Experiences (AICE), and the IBM-Illinois Discovery Accelerator Institute (IIDAI).

References

Chi, J.; Shen, J.; Dai, X.; Zhang, W.; Tian, Y.; and Zhao, H. 2022. Towards Return Parity in Markov Decision Processes. In *International Conference on Artificial Intelligence and Statistics*, 1161–1178. PMLR.

Chi, J.; Tian, Y.; Gordon, G. J.; and Zhao, H. 2021. Understanding and mitigating accuracy disparity in regression. In *International conference on machine learning*. PMLR.

Hu, Y.; Wu, F.; Zhang, H.; and Zhao, H. 2023. Understanding the Impact of Adversarial Robustness on Accuracy Disparity. In *International Conference on Machine Learning*, 13679–13709. PMLR.

Wang, H.; Li, B.; and Zhao, H. 2022. Understanding gradual domain adaptation: Improved analysis, optimal path and beyond. In *International Conference on Machine Learning*, 22784–22801. PMLR.

Xian, R.; Yin, L.; and Zhao, H. 2023. Fair and Optimal Classification via Post-Processing. In *International Conference on Machine Learning*.

Zhao, H. 2022. Costs and Benefits of Fair Regression. *Transactions on Machine Learning Research*.

Zhao, H.; Coston, A.; Adel, T.; and Gordon, G. J. 2019. Conditional Learning of Fair Representations. In *International Conference on Learning Representations*.

Zhao, H.; Dan, C.; Aragam, B.; Jaakkola, T. S.; Gordon, G. J.; and Ravikumar, P. 2022. Fundamental limits and tradeoffs in invariant representation learning. *The Journal of Machine Learning Research*, 23(1): 15356–15404.

Zhao, H.; and Gordon, G. J. 2022. Inherent tradeoffs in learning fair representations. *The Journal of Machine Learning Research*, 23(1): 2527–2552.