Xuezhou Zhang

University of Wisconsin-Madison
Department of Computer Sciences
1210 West Dayton Street
Madison, WI, USA 60208-3119

Phone: (310) 717-9440
Email: zhangxz1123@cs.wisc.edu
Homepage: http://pages.cs.wisc.edu/∼zhangxz1123

Research Interests

My research has covered several topics around machine learning, including adversarial machine learning, interpretable machine learning and reinforcement learning.

Education

Ph.D. Computer Sciences, University of Wisconsin-Madison, expected 2021.

Advisor: Xiaojin (Jerry) Zhu

M.S. Computer Sciences, University of Wisconsin-Madison, 2018.

B.S. Applied Mathematics, *Summa Cum Laude*, University of California, Los Angeles, 2016.

Research Advisor: Andrea L. Bertozzi

Major GPA: 4.00/4.00

Work Experience

Alibaba Group, Redmond, WA Summer 2019

Research Intern, hosted by Shaojian He.

Designed and implemented an RL-based recommendation system for the Alibaba Taobao e-commerce platform, with the goal of optimizing long-term user retention.

Microsoft Research, Redmond, WA Summer 2018

Research Intern, hosted by Rich Caruana.

Studied the interpretability of multi-class generalized additive models (GAMs). Paper published in KDD 2019, Research Track, Oral Presentation. [Paper]

Helped develop a scalable Interpretable Machine Learning package *InterpretML*, that includes the state-of-the-art methods for training interpretable models as well as explaining black-box models. Open-sourced under MIT license. [Github]

Publication

Xuezhou Zhang, Yuzhe Ma, Adish Singla. Task-agnostic Exploration in Reinforcement Learning. *arXiv Preprint.* 2020.

Xuezhou Zhang, Shubham Kumar Bharti, Yuzhe Ma, Adish Singla, Xiaojin Zhu. The Teaching Dimension of Q-learning. *arXiv Preprint.* 2020.

Xuezhou Zhang*, Yun-Shiuan Chuang*, Yuzhe Ma, Mark Ho, Joe Austerweil, Xiaojin Zhu. Using Machine Teaching to Investigate Human Assumptions when Teaching Reinforcement Learners. *arXiv Preprint.* 2020.

Rishabh Agarwal, Nicholas Frosst, Xuezhou Zhang, Rich Caruana, Geoffrey Hinton. Neural Additive Models:Interpretable Machine Learning with Neural Networks. *arXiv Preprint.* 2020.

Huajie Shao, Jun Wang, Jinnian Zhang, Xuezhou Zhang, Aston Zhang, Ze Tian, Shaojian He, Xin Li, Tarek Abdelzaher. APEX: PID-Based Controllable and Diverse Text Generation. *arXiv Preprint.* 2020.

Xuezhou Zhang, Yuzhe Ma, Adish Singla, and Xiaojin Zhu. Adaptive reward-poisoning attacks against reinforcement learning. *In The 37th International Conference on Machine Learning (ICML).* 2020.

Xuezhou Zhang, Xiaojin Zhu, Laurent Lessard. Online Data Poisoning Attack. *In Learning for Dynamics and Control (L4DC).* 2020. **Oral Presentation.**

Yuzhe Ma, Xuezhou Zhang, Wen Sun, and Xiaojin Zhu. Policy poisoning in batch reinforcement learning and control. In *The Thirty-Third Conference on Neural Information Processing Systems(NeurIPS)*, 2019.

Xuezhou Zhang, Sarah Tan, Paul Koch, Urszula Chajewska, and Rich Caruana. Axiomatic Interpretability for Multiclass Additive Models. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, Research Track, 2019. **Oral Presentation.**

Xuezhou Zhang*, Laurent Lessard*, and Xiaojin Zhu. An Optimal Control Approach to Sequential Machine Teaching. In *The 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Xuezhou Zhang, Xiaojin Zhu, and Stephen J. Wright. Training Set Debugging using Trusted Items. In *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, 2018. **Oral presentation.**

Yuzhe Ma, Philippe Rigollet, Xuezhou Zhang, Robert D Nowak, and Xiaojin Zhu. Teacher Improves Learning by Selecting a Training Subset. In *The 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

Ayon Sen, Scott Alfeld, Xuezhou Zhang, Ara Vartanian, Yuzhe Ma, and Xiaojin Zhu. Training set camouflage. In *Conference on Decision and Game Theory for Security (GameSec)*, 2018. **Oral presentation.**

Xuezhou Zhang, Hrag Gorune Ohannessian, Ayon Sen, Scott Alfeld, and Xiaojin Zhu. Optimal teaching for online perceptrons. In *NIPS 2016 workshop on Constructive Machine Learning*, 2016.

Honors and Awards

Summer Research Fellowship, UW Madison, 2017

Rose Gilbert Honors Program Scholarship, University of California, Los Angeles, 2015.

Meritorious Winner (top 9%), COMAP Mathematical Contest in Modeling, 2015.