Thesis Summary

Fulton Wang

Overview I am broadly interested in 1. developing machine learning methods that provide human understandable insight about the world or their own inner workings, and 2. benefiting society through my work. I pursue these two interests in my thesis. Within the first interest, I have worked on creating interpretable models - models whose form or predictions can be easily understood. Within the second interest, I have worked on creating methods that avoid bias, broadly construed, and also working on real-life problems where my contributions can provide social good. Below, I detail my thesis work within those two interests.

Interpretable Machine Learning I am interested in creating interpretable models because humans are more likely to trust and use models whose form or predictions they understand and can draw insight from. One interpretable model I created is the falling rule list (Wang and Rudin 2015b), which is a decision list classifier where the predicted risk probabilities monotonically decrease from the top to bottom of the list. This imposed structure orders rules from highest to lowest risk, allowing the user to check if the ordering agrees with their existing knowledge, which is likely also mentally represented as some ordering over rules. This work was a winner of the 2015 ASA Statistical Learning and Data Mining section student paper competition, as well as a finalist for the 2015 Informs Data Mining Section Best Student Paper Award. As a follow-up work, I created the causal falling rule list (Wang and Rudin 2015a), where the decision list now groups subjects so that the average treatment effect for each group decreases monotonically down the list. I created a generative model in which the average treatment effects are parameters, and formulated the choice over possible causal falling rule lists as a Bayesian model selection problem. The model is interpretably prescriptive, ordering subject groups by treatment priority.

Another interpretable model I created was a Bayesian model for the trajectory of one's sexual function score following a prostatectomy (Wang et al. 2017). Urologists have a strong expectation of what such a trajectory looks like. Thus to gain their trust in the model, we constrained the model so that its predictions would meet those expectations. Furthermore, the model represents uncertainty interpretably, imposing a strong prior that the distribution over scores (which lie in a closed interval) be unimodal; a doctor would be distrustful of a bimodal posterior, which they would interpret as making 2 predictions for the same patient. A poster explaining this work won a best poster award at the 2014 ASA Statistical Learning and Data Mining conference.

Handling Bias in Machine Learning I am concerned about negative consequences that may arise when machine learning methods or datasets are biased, broadly defined. To this end, I have worked on handling covariate shift, in which the covariate distribution of the training data differs from that of the test population for which predictions are needed. Ignoring this shift when training a model may result in subpopulations underpresented in the training data receiving inaccurate predictions. A popular approach to handling covariate shift minimizes a weighted training loss over predictors, emphasizing training samples in regions of high test density. My unpublished work (Wang and Rudin 2017) adds dimension reduction into this procedure to reduce its variance, which is high when the training and test populations are divergent, even among non-predictive covariates. In particular, I assume that covariates will be projected to a subspace before the predictor is learned using weighted loss minimization. Using techniques from gradient descent for hyperparameter optimization, I then jointly minimize the weighted loss over subspaces and predictors acting on it. The variance is reduced because weights are estimated in lower dimensions and because I add a regularizer discouraging subspaces within which the two distributions are overly divergent.

In ongoing collaborative work, I am helping to develop a classifier that does not produce predictions that are systematically too high or low for any subgroup, defined as subpopulations satisfying a conjunction of rules. The past work of my collaborators (Zhang and Neill 2016) has presented a computationally efficient way to detect whether the predictions of a given classifier are overly biased in either direction within any of the exponentially many possible subgroups. We are now trying use this bias detector constructively, to create a classifier the detector would not flag as being biased.

Data Science for Social Good Aside from methodological research, I am interested solving societal problems with data, and also defining the quantitative problem to be solved

given qualitative directions from social scientists. In a collaboration with the Cambridge Police department, I applied a computationally efficient subset scan method to identify concentrations of crimes in location, time, and feature space, crimes likely committed by the same individual. This project required clear communication of quantitative findings with police officers, which I enjoyed. This, as well as my experience with health data (including readmissions data, urology surveys, electronic medical records) has reinforced my desire to develop methods that can be applied to problems of societal importance.

Future Work Going forward, I plan to pursue my interest in methods that provide human understandable insight. This includes continuing my work on creating models that directly explain their decisions and when they are confident, perhaps for new settings such as structured prediction, extreme label classification, and reinforcement learning, as well as generating post-hoc interpretations of (black-box) models, including explanations that do not reveal sensitive or proprietary information, and those bridging the gap between local and global explanations. Towards my interest in benefiting society, i plan to continue work on methods that avoid bias, and understand and overcome the ethical and computational barriers that might hinder their deployment. Finally, I hope work on these methods will all be motivated by applications for social good, inspired by interactions with those would both deploy and benefit from them.

References

Wang, F., and Rudin, C. 2015a. Causal falling rule lists. *arXiv preprint arXiv:1510.05189*. Appeared in 2017 Workshop on Fairness, Accountability, and Transparency in ML (FATML).

Wang, F., and Rudin, C. 2015b. Falling rule lists. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, 1013–1022.

Wang, F., and Rudin, C. 2017. Dimension reduction for robust covariate shift correction. *ariv preprint arXiv:1711.10938*.

Wang, F.; McCormick, T. H.; Rudin, C.; and Gore, J. 2017. Modeling recovery curves with application to prostatectomy. *Biostatistics*. Forthcoming.

Zhang, Z., and Neill, D. B. 2016. Identifying significant predictive bias in classifiers. *arXiv preprint arXiv:1611.08292*.