Alexis Kwan
PA 470

Machine Learning for Fairness

Machine learning is often used to improve prediction accuracy, but rarely is it used to evaluate the potential side effects of its application. The purpose of research from Zheng et al. is to explore using machine learning to evaluate those effects, based on a probabilistic measure of the potential inequities. The metric the author described compares opportunities across protected attributes with all other factors held constant. Using this metric they compare two major algorithms for their disparities, deep neural networks (DNN) and discrete choice models (DCN). Their express purpose of implementing this metric is their realization that “models inheriting the inherent biases can risk perpetuating the existing inequality in the data structure, and the biases in modeling can further exacerbate it” (Zheng et al., 2021). In this case they are looking specifically at modeling the disparities with transit opportunities.

Their experimental procedure consists of essentially two parts. They initially identify prediction disparities using logistic regression and DNN. Then they attempt to mitigate the biases by modifying an approach from Beutel et al. which uses correlation regularization (added to the loss function used in the training the models) to rebalance the biases. The data they use for the experiment include both synthetic data and real world data. The synthetic data are constructed such that there is covariance between the protected variable (e.g. some representation of race in the data) and another explanatory variable. The real data sets they used are 2017 National Household Travel Survey data and Chicago travel survey.

Overall the methods they used appeared to be fairly effective with only a small loss in over accuracy. From a more macroscopic policy perspective there is one flaw that could immediately result in the definition of the “equality of opportunity” metric they used. When there is covariance between a protected attribute and an explanatory variable, usually that variable

then becomes a proxy for bias. What they have done in the paper appears to address this at least partially. However if there are multiple nested covariances, the bias becomes more obfuscated. For example, some areas may require more transportation resources because people have migrated there in greater numbers due to disinvestment in their original neighborhoods, which could then further increase disparities. Even the authors admit that their algorithms seem to produce a tradeoff between protected variables. This warrants further research into how the biases themselves are represented in the data and in the relationships between variables.

References

Transportation equity. (2021). CMAP. https://www.cmap.illinois.gov/2050/mobility/transportation-equity

Zheng, Y., Wang, S., & Zhao, J. (2021). Equality of opportunity in travel behavior prediction with deep neural networks and discrete choice models. *Transportation Research Part C: Emerging Technologies*, *132*, 103410. https://doi.org/10.1016/j.trc.2021.103410