Multiaccurate Proxies for Downstream Fairness

Emily Dina (edina@wharton.upenn.edu)
with Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, Aaron Roth, Saeed Sharifi-Malvajerdi
University of Pennsylvania

### ALGORITHMIC FAIRNESS IN THE NEWS

![Machine Bias](image)

**Predictive policing algorithms are racist. They need to be dismantled.**

### ALGORITHMIC FAIRNESS IN THE LITERATURE

![Data Mining](image)

**Fairness Through Awareness [5]**

### RESEARCH QUESTION

- Algorithmic fairness aims to understand and prevent bias in machine learning models.
- Often one wants to train a model that is fair with respect to a sensitive feature that has been redacted from training data?
- Could be for legal or policy reasons:
  - In the United States it is against the law to use race as an input to consumer lending models.
  - Many large consumer-facing organizations choose not to ask their customers for such information.

How do we make a model fair with respect to race if we don’t have data about race?

### FRAMEWORK

- Data domain $\Omega$ divided into $K$ groups:
  - $\Omega = \{\text{non-sensitive features}\} \times \{\text{sensitive feature}\}$
  - $\Omega = \{\text{non-sensitive features}\} \times \{\text{sensitive feature}\}$
- Proxy model class $\hat{g} : \mathbb{X} \rightarrow \mathbb{R}^k$
- Proxy $\hat{z} \in \Omega$; vector of $K$ real numbers (\hat{z}_1, \ldots, \hat{z}_K)
- Downstream model class $H : \mathbb{X} \rightarrow \mathbb{Y}$

Proxy Learner aims to find proxy $\hat{z}$ such that if a Downstream Learner trains a model $h$ that is fair with respect to $\hat{z}$, $h$ is also fair with respect to $z$.

### EXPERIMENTS: OVERVIEW

Simulating a downstream learner, we train a model to be fair with respect to four representations of the sensitive feature and evaluate its performance:

- True Labels: $Z$
- Baseline Proxy: Logistic regression of $Z$ on $X$
- $\hat{z}$-Proxy: Solution to Program (1) without squared error objective
- MSE Proxy: Solution to Program (1) with squared error objective

Conducted experiments on American Community Survey (ACS) datasets and tasks from [2].

### EXPERIMENTS: ACS DATA

**Figure: Proxy results on the ACSIncome dataset with race as sensitive feature**

**Figure: Proxy results on the ACSIncome dataset with age as sensitive feature**

**Figure: Proxy results on the ACSIncome dataset with sex as sensitive feature**

### SELECTED REFERENCES

1. Scheidegger CE, Parno A, Shokri R. "Multiaccurate Classification for Non-Discriminative Model Training." In: Proceedings of the 33rd International Conference on Machine Learning, 2016.
2. Zhang C, Zhao J. "Learning Fair Representations with Fairness-Aware Priors." In: Proceedings of the 34th Conference on Neural Information Processing Systems, 2020.
3. Dwork C, Hardt M, Pitassi T, Reingold E, Zemel R. "Fairness Through Awareness." In: Proceedings of the 50th Annual ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, 2011.
4. Gehrke J, Ramakrishnan R, Srikant R. "Data Mining Algorithms in R with Applications in Biostatistics." In: Springer, 2018.

### SELECTED REFERENCES

- We have shown that it is possible to efficiently train proxies that can stand in for missing sensitive features to effectively train downstream classifiers subject to a variety of demographic fairness constraints.
- Our theoretical and empirical results demonstrate that proxies trained using our methods can stand in as near perfect substitutes for sensitive features in downstream training tasks.
- Results crucially depend on the assumption that the data that the Proxy Learner uses to train its proxy is distributed identically to the data that the Downstream Learner uses.
- In real applications, either of these assumptions can fail (or can become false due to distribution shift, even if they are true at the time that the proxy is trained).

### CONCLUSION