Multiaccurate Proxies for Downstream Fairness

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

**Algorithmic Fairness in the News**

- **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 due to 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 model a fair model with respect to race if we don’t have data about race?

**Algorithmic Fairness in the Literature**

- Inherent Trade-Offs
- In real applications, either of these assumptions can fail (or can become false due to
- 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 the Proxy Learner uses to train its proxy is distributed identically to the data that 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 moment the proxy is trained).

**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 due to 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 model a fair model with respect to race if we don’t have data about race?

**Framework**

- Data domain $\Omega$ divided into $K$ groups:
  $$\Omega = \{x \in \mathbb{R}^d : y, z \in \{0, 1\}\} = \{x \in \mathbb{R}^d : y, z \in \{0, 1\}\}$$

- Proxy model class $\mathcal{F} : \mathbb{X} \to \mathbb{R}^K$.
- Proxy $\tilde{y} \in \mathcal{F}$: vector of $K$ real numbers ($\tilde{z}_1, \ldots, \tilde{z}_K$).
- Downstream model class $\mathcal{H} : \mathbb{X} \to \mathbb{R}^K$.

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

**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 $X$
  - $\gamma$-Proxy: Solution to Program (1) with squared error objective
  - MSE Proxy: Solution to Program (1) with squared 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

**Conclusion**

- 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 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 the proxy is trained).

**Selected References**

1. **[1]** Aleksandar Nikolov, Carsten Sinno Janson, and Paul Grigorescu. "Adversarial training for fairness against individual attributes."