Multiaccurate Proxies for Downstream Fairness

Emily Diana

ediana@wharton.upenn.edu
THANK YOU TO MY COLLABORATORS

Wesley Gill  Michael Kearns  Krishnaram Kenthapadi

Aaron Roth  Saeed Sharifi-Malvajerdi
Algorithmic fairness aims to understand and prevent bias in machine learning models.
Challenges: How do we decide which definitions to use? How do we decide what constitutes harm? When and how do we intervene? How do we balance trade-offs?

1. Pedreshi, Ruggieri, and Turini. Discrimination-Aware Data Mining. KDD ’08
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.

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

---

5 In the United States it is against the law to use race as an input to consumer lending models
6 Many large consumer-facing organizations choose not to ask their customers for such information.
PLAN

Key Insight

Framework

Experiments

Main Result

Takeaways
Data domain $\Omega = \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ divided into $K$ groups

Proxy model class $\mathcal{G} : \mathcal{X} \to \mathbb{R}^K$

Proxy $\hat{z} \in \mathcal{G}$ is a vector of $K$ real numbers $(\hat{z}_1, ..., \hat{z}_K)$

Downstream model class $\mathcal{H} : \mathcal{X} \to \mathcal{Y}$

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 $z$. 
FRAMEWORK

Proxy Learner

$\hat{z}(x) \in \mathcal{G}$

Downstream Learner

$h(x) \in \mathcal{H}$

Fair Machine Learning Model

$\hat{z}(x,y) \in \mathcal{G}$
We can write fairness constraints, usually defined with respect to binary valued group membership using a real valued proxy:

\[
\Pr[h(x) \neq y | z_k = 1] = \frac{\Pr[z_k = 1, h(x) \neq y]}{\Pr[z_k = 1]}
= \frac{\mathbb{E}[1[z_k = 1]1[h(x) \neq y]]}{\mathbb{E}[1[z_k = 1]]}
= \frac{\mathbb{E}[z_k 1[h(x) \neq y]]}{\mathbb{E}[z_k]}
\]
KEY INSIGHT: REPLACE $Z$ WITH $\hat{Z}$

If the following holds:

$$\frac{\mathbb{E}[z_k 1 \{h(x) \neq y\}]}{\mathbb{E}[z_k]} = \frac{\mathbb{E}[\hat{z}_k(x) 1 \{h(x) \neq y\}]}{\mathbb{E}[\hat{z}_k(x)\]}$$

Then if a model is fair with respect to $\hat{z}$

$$\frac{\mathbb{E}[\hat{z}_{k_i}(x) 1 \{h(x) \neq y\}]}{\mathbb{E}[\hat{z}_{k_i}(x)]} = \frac{\mathbb{E}[\hat{z}_{k_j}(x) 1 \{h(x) \neq y\}]}{\mathbb{E}[\hat{z}_{k_j}(x)]}$$

it also satisfies fairness constraints with respect to the true attribute $z$. 
We say $\hat{z}$ is an $\alpha$-proxy for $z$ if for all classifiers $h \in \mathcal{H}$, and all groups $k \in [K]$,

$$\left| \frac{\mathbb{E}_{(x,z)} [z_k 1 \{ h(x) \neq y \}]}{\mathbb{E}_{(x,z)} [z_k]} - \frac{\mathbb{E}_{(x,z)} [\hat{z}_k(x) 1 \{ h(x) \neq y \}]}{\mathbb{E}_{(x,z)} [\hat{z}_k(x)]} \right| \leq \alpha$$
KEY INSIGHT: MULTIACCURACY

Then to learn a proxy, we can solve the linear program:

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{n} \sum_{i=1}^{n} (z_i - \hat{z}(x_i))^2 \\
\text{subject to} & \quad \sum_{i=1}^{n} z_i = \sum_{i=1}^{n} \hat{z}(x_i), \\
& \quad \sum_{i=1}^{n} z_i \mathbb{1}[h(x_i) \neq y_i] = \sum_{i=1}^{n} \hat{z}(x_i) \mathbb{1}[h(x_i) \neq y_i], \quad \forall h \in \mathcal{H}
\end{align*}
\]

(1)

These constraints are multiaccuracy constraints – we want \( \hat{z} \) to be an unbiased estimator for \( z \) on the set of points where \( h \) errs.
STRONG DUALITY AND LOW-REGRET DYNAMICS

Linear Program

Dual Problem (min max Lagrangian)

Solution to Game

Lagrangian

\[ L(\hat{z}, \lambda) = \sum_{i=1}^{n} \left[ \left( z_i - \hat{z}(x_i) \right)^2 \right] \]

\[ + \lambda_0 \left( \frac{\sum_{i=1}^{n} \hat{z}(x_i)}{\sum_{i=1}^{n} z_i} - 1 \right) \]

\[ + \sum_{h \in \mathcal{H}} \lambda_h \sum_{i=1}^{n} (z_i - \hat{z}(x_i)) \mathbb{1} [h(x_i) \neq y_i] \]

Under appropriate conditions\(^7\), Solving Program (1) is equivalent to solving:

\[ \min_{\hat{z} \in \mathcal{G}} \max_{\lambda} L(\hat{z}, \lambda) = \max_{\lambda} \min_{\hat{z} \in \mathcal{G}} L(\hat{z}, \lambda) \]

\(^7\)Primal variable space is convex and compact, dual variable space is convex, and Lagrangian is convex-concave in primal and dual variables respectively.
ALGORITHM OVERVIEW: NO-REGRET DYNAMICS

Can cast problem as zero-sum game between Learner and Auditor

▶ Proxy Learner uses Online Projected Gradient Descent to select $\hat{z}$ minimizing $L(\hat{z}, \lambda)$

▶ Auditor best responds, appealing to an oracle over downstream model class $\mathcal{H}$ to select $\lambda$ maximizing $L(\hat{z}, \lambda)$

Freund and Schapire show that if a sequence of actions for the two players jointly has low regret, then the uniform distribution over each player’s actions forms an approximate equilibrium.

---

8 Here we consider the simpler case in which $\hat{z}$ is a linear function in its parameter space, so both $\hat{z}$ and its negation are convex. More details on the non-convex case are provided in the paper.
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 $X$
- **$H$-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\textsuperscript{9}

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Samples</th>
<th>Dim</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACSEmployment</td>
<td>196104</td>
<td>12</td>
<td>Employment</td>
</tr>
<tr>
<td>ACSIncome</td>
<td>101270</td>
<td>4</td>
<td>Income &gt; $50K</td>
</tr>
<tr>
<td>ACSIncomePovertyRatio</td>
<td>196104</td>
<td>15</td>
<td>Income-Poverty Ratio &lt; 250%</td>
</tr>
<tr>
<td>ACSMobility</td>
<td>39828</td>
<td>17</td>
<td>Same address one year ago</td>
</tr>
<tr>
<td>ACSPublicCoverage</td>
<td>71379</td>
<td>15</td>
<td>Health Insurance</td>
</tr>
<tr>
<td>ACSTravelTime</td>
<td>89145</td>
<td>8</td>
<td>Commute &gt; 20 minutes</td>
</tr>
</tbody>
</table>

EXPERIMENTS: ACSIncome Race

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

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

▶ 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.

▶ Results crucially depend on assumption that the data that the Proxy Learner uses to train its proxy is distributed identically to the data that the Downstream Learner uses.
THANK YOU!

QUESTIONS?
THANK YOU TO MY COLLABORATORS

Wesley Gill        Michael Kearns        Krishnaram Kenthapadi

Aaron Roth        Saeed Sharifi-Malvajerdi