A Convex Framework for Fair Regression


Motivation

- Machine learning (ML) increasingly used to make critical decisions, e.g. hiring and sentencing
- Problem: there are many examples of ML that is discriminatory or unfair
- There is a large body of work on fair classification; we instead focus on fair regression

Fairness Definitions

- Adapts idea that similar individuals (similar ground-truth label) should be treated similarly (similar predicted label) [Dwork et. al.] by introducing sample fairness penalties
- Individual Fairness penalty:
  \[ f_1(w, S) = \frac{1}{n_1 n_2} \sum_{(x_i, y_i) \in S_1} \sum_{(x_j, y_j) \in S_2} d(y_i, y_j) (w \cdot x_i - w \cdot x_j)^2 \]
  - Each pair of similar examples classified dissimilarly adds loss – no “cancellation”, most stringent fairness requirement
- Group Fairness penalty:
  \[ f_2(w, S) = \left[ \frac{1}{n_1 n_2} \sum_{(x_i, y_i) \in S_1} \sum_{(x_j, y_j) \in S_2} d(y_i, y_j) (w \cdot x_i - w \cdot x_j)^2 \right]^{\frac{1}{2}} \]
  - Pairs of similar examples classified dissimilarly can be cancelled out by pairs classified dissimilarly in the opposite direction, least stringent fairness requirement
- Hybrid Fairness: cancellation only among cross-pairs within “buckets” – interpolates between individual and group fairness
- Fairness loss minimized by constant predictors, but this incurs bad accuracy loss
- How to trade off accuracy and fairness losses?

The Optimization Problem

- Overall loss function to minimize is
  \[ \min_w E_{(x, y) \sim P} [(w \cdot x - y)^2] + \lambda f(w) + \alpha(\lambda) \| w \|_2 \]
  - Accuracy loss + fairness loss + \ell_2 regularizer
  - Benefit: convex optimization problem \implies tractable

Summary of Datasets

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Type</th>
<th>( n )</th>
<th>( d )</th>
<th>Minority Protected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult</td>
<td>logit</td>
<td>32561</td>
<td>14</td>
<td>10771 gender</td>
</tr>
<tr>
<td>Comm. &amp; Crime</td>
<td>linear</td>
<td>1994</td>
<td>128</td>
<td>227 race</td>
</tr>
<tr>
<td>COMPAS</td>
<td>logit</td>
<td>3373</td>
<td>19</td>
<td>1455 race</td>
</tr>
<tr>
<td>Default</td>
<td>logit</td>
<td>30000</td>
<td>24</td>
<td>11888 gender</td>
</tr>
</tbody>
</table>

Pareto Curves

Price of Fairness

\[ \text{PoF}(\alpha) = \min_w \text{err}(w) \text{ subject to } f(w) \leq \alpha f(w^*) \]

- The increment in error for any given fairness level of \( \alpha \) compared to the best unfair predictor

Quantitative Measure of Trade-off

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Type</th>
<th>( n )</th>
<th>( d )</th>
<th>Minority Protected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult</td>
<td>logit</td>
<td>32561</td>
<td>14</td>
<td>10771 gender</td>
</tr>
<tr>
<td>Comm. &amp; Crime</td>
<td>linear</td>
<td>1994</td>
<td>128</td>
<td>227 race</td>
</tr>
<tr>
<td>COMPAS</td>
<td>logit</td>
<td>3373</td>
<td>19</td>
<td>1455 race</td>
</tr>
<tr>
<td>Default</td>
<td>logit</td>
<td>30000</td>
<td>24</td>
<td>11888 gender</td>
</tr>
</tbody>
</table>

Takeaways

- Notion of fairness that’s tractable to optimize
- The detailed trade-offs between fairness and accuracy and different notions of fairness appear to be quite data-dependent and lack universals
- Possibly consistent with emerging theoretical literature demonstrating the lack of a unified, comprehensive fairness definition