Certifying Robustness to Programmable Data Bias in Decision Trees

Anna P. Meyer, Aws Albarghouthi, and Loris D’Antoni
dataset $\rightarrow$ learning algorithm $\rightarrow$ model
Is the model fair? accurate? trustworthy?
Is the dataset biased?  
complete?  
representative?
Is the dataset biased?
  complete?
  representative?

Probably not.

What is the impact on the model’s predictions?
Goal: certify robustness to training-data bias
Types of data bias

Incorrect labels

e.g., historical biases like women marked as “not hired” for a job even though they were qualified

Missing data

e.g., neglected to collect data from a minority neighborhood

Fake data

e.g., fake answers submitted through crowdsourcing
for all $D'$ that disagree with $D$ on $\leq n$ labels, show that $f_{D'}(x) = f_D(x)$
bias robustness of $x$

for all $D'$ that disagree with $D$ on $\leq n$ labels

show that $f_{D'}(x) = f_D(x)$

Dataset $D$
etc.
etc.
for all $D'$ that disagree with $D$ on $\leq n$ labels,

show that $f_{D'}(x) = f_D(x)$

Key challenge

Combinatorial explosion in the number of datasets

$|D| = 1000$

$n = 10$

$\sim 10^{23}$ datasets!
large set of training datasets → abstract learning algorithm → large set of trained models
abstract
decision-tree learning algorithm
large set of training datasets
large set of decision trees
Dataset $D$
$\phi := \text{value } \leq 3$
Dataset $D$

$\phi := \text{value} \leq 3$

Number $\checkmark = 4$
Number $\times = 1$
\[ \phi := \text{value} \leq 3 \]

Dataset \( D \)

\[
\text{Gini Impurity} = \sqrt{\frac{4}{5}} \cdot (1 - \sqrt{\frac{4}{5}}) + \frac{1}{5} \cdot (1 - \frac{1}{5}) = 0.32
\]
Abstraction of Dataset $D$

$\phi := \text{value} \leq 3$
Abstraction of Dataset $D$

$\phi := \text{value} \leq 3$

Number $\checkmark = 4$
Number $\times = 1$

Number $\checkmark = [3, 5]$
Number $\times = [0, 2]$

$\exists 0, 1, 2, 3, 5, 6, 7, 8, 9$
Gini Impurity = \( \checkmark \cdot (1-\checkmark) + \times \cdot (1- \times) \)
= \( \frac{3,5}{5}(1 - \frac{3,5}{5}) + \frac{0,2}{5}(1 - \frac{0,2}{5}) \)
= \([0, 0.8]\)
Abstract decision-tree-learner pipeline

1. Build an abstract decision tree

2. Find the prediction of $\mathbf{x}$ under each of the trees constructed with the best predicates

3. See whether all predictions agree

   If so, $\mathbf{x}$ is certifiably robust!

   If not, inconclusive.
Experimental results
Certification rate

Given n% bias, what percentage of test data points are certifiably robust?

<table>
<thead>
<tr>
<th>Bias type</th>
<th>Dataset</th>
<th>Bias amount as a percentage of training set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.05</td>
</tr>
<tr>
<td>MISS (missing data)</td>
<td>Drug Consumption</td>
<td>94.5</td>
</tr>
<tr>
<td></td>
<td>COMPAS</td>
<td>89.0</td>
</tr>
<tr>
<td></td>
<td>Adult Income (AI)</td>
<td>96.0</td>
</tr>
<tr>
<td></td>
<td>COMPAS targeted</td>
<td>89.0</td>
</tr>
<tr>
<td></td>
<td>AI targeted</td>
<td>98.8</td>
</tr>
</tbody>
</table>
## Certification rate

Given $n\%$ bias, what percentage of test data points are certifiably robust?

<table>
<thead>
<tr>
<th>Bias type</th>
<th>Dataset</th>
<th>Bias amount as a percentage of training set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.05</td>
</tr>
<tr>
<td>MISS (missing data)</td>
<td>Drug Consumption</td>
<td>94.5</td>
</tr>
<tr>
<td></td>
<td>COMPAS</td>
<td>89.0</td>
</tr>
<tr>
<td></td>
<td>Adult Income (AI)</td>
<td>96.0</td>
</tr>
<tr>
<td></td>
<td>COMPAS targeted</td>
<td>89.0</td>
</tr>
<tr>
<td></td>
<td>AI targeted</td>
<td>98.8</td>
</tr>
</tbody>
</table>

### Bias-set size color scheme

- $< 10^{10}$
- $< 10^{50}$
- $< 10^{100}$
- $< 10^{500}$
- $> 10^{500}$
- infinite
Certification discrepancy between demographic groups

COMPAS dataset (but discrepancies exist for Adult Income, too)
Future work

- Extensions to other ML algorithms
- Counter-examples to robustness