**PROBLEM STATEMENT**

Let $X = \text{non-protected attributes}$, $A = \text{protected attributes}$, $\hat{Y} = \text{predicted class label}$

Given

- binary classifier $\mathcal{M} : (X, A) \rightarrow \{0, 1\}$ and
- probability distribution $X \sim \mathcal{D}$,

verify whether $\mathcal{M}$ achieves independence and separation fairness metrics with respect to the distribution $\mathcal{D}$

**Fairness Metrics**

**Independence:** A classifier satisfies $(1 - \epsilon)$-disparate impact (DI) if, for $\epsilon \in [0, 1]$,

$$\min_{a \in A} \Pr[\hat{Y} = 1|a, \mathcal{M}] \geq (1 - \epsilon) \max_{b \in A} \Pr[\hat{Y} = 1|b, \mathcal{M}].$$

**Separation:** A classifier satisfies $\epsilon$-statistical parity if, for $\epsilon \in [0, 1]$,

$$\max_{a, b \in A} |\Pr[\hat{Y} = 1|a, \mathcal{M}] - \Pr[\hat{Y} = 1|b, \mathcal{M}]| \leq \epsilon.$$

**CONTRIBUTION**

A formal and scalable fairness verification framework, named Justicia, based on Stochastic SAT

- Two fairness definitions: independence and separation
- Handle compound protected groups such as White-male, Black-female etc.

**Python library:** pip install justicia

**KEY OBSERVATION**

Computing the positive predictive value (PPV) of the classifier

$$\Pr[\hat{Y} = 1|A = a]$$

is the building block of verifying different fairness metrics

**STOCHASTIC SAT (SSAT)**

Compute probability of satisfaction of a CNF formula $\Phi$ given quantification over its variables

$$\Phi = Q_1X_1, \ldots, Q_MX_M; \phi_{\text{prefix}}$$

where $Q_i \in \{\exists, \forall, b^0\}$ is either an existential ($\exists$), an universal ($\forall$),or a randomised ($b^0$) quantifier with $p_i = \Pr[X_i = \text{TRUE}]$

**Semantics.** Recursively eliminate the outermost quantifier of $X$

1. $\Pr[\text{TRUE}] = 1$, $\Pr[\text{FALSE}] = 0$.
2. $\Pr[\Phi] = \max_{X} \{\Pr[\Phi|X], \Pr[\Phi|\neg X]\}$ if $X$ is $\exists$ quantified
3. $\Pr[\Phi] = \min_{X} \{\Pr[\Phi|X], \Pr[\Phi|\neg X]\}$ if $X$ is $\forall$ quantified
4. $\Pr[\Phi] = \rho \Pr[\Phi|X] + (1 - \rho) \Pr[\Phi|\neg X]$ if $X$ is $b^0$ quantified

**Example.** $\Phi = b^{0.25} X_1, \exists X_2, \exists X_3; (X_1 \lor \neg X_2) \land (\neg X_1 \lor X_2 \lor X_3) \land (\neg X_1)$ such that $\Pr[\Phi] = 0.75$

**ENCODING CORRELATION**

Use conditional probability $\Pr[F|\text{age} \geq 40]$ instead of $\Pr[F]$

$$\Phi_{\text{age} \geq 40} := b^{0.01} F, b^{0.09} I, b^{0.08} J; \exists A; (\neg F \lor I) \land (F \lor J) \land A$$

Disparate impact $= \frac{0.18}{0.72}$; Statistical parity $= |0.18 - 0.72| = 0.54$

**APPRAOCH 2 : LEARNING**

Learning the most favored group

$$\Phi := \exists A, b^{0.41} F, b^{0.93} I, b^{0.09} J; (\neg F \lor I) \land (F \lor J)$$

Learning the least favored group

$$\Phi := \forall A, b^{0.41} F, b^{0.93} I, b^{0.09} J; (\neg F \lor I) \land (F \lor J)$$

**EXPERIMENTAL RESULTS**

**Accuracy:** Justicia shows less than 1%-error

<table>
<thead>
<tr>
<th>Metric</th>
<th>FairSquare</th>
<th>VeriFair</th>
<th>AIF360</th>
<th>Exact</th>
<th>Justicia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disparate impact</td>
<td>0.99</td>
<td>0.99</td>
<td>0.25</td>
<td>0.26</td>
<td>0.25</td>
</tr>
<tr>
<td>Statistical parity</td>
<td>—</td>
<td>—</td>
<td>0.54</td>
<td>0.53</td>
<td>0.54</td>
</tr>
</tbody>
</table>

**Scalability:** Justicia shows 1 to 3 orders of magnitude speed-up

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Ricci</th>
<th>Titanic</th>
<th>COMPAS</th>
<th>VERI FAIR</th>
<th>Justicia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classifier</td>
<td>DT</td>
<td>LR</td>
<td>DT</td>
<td>LR</td>
<td>DT</td>
</tr>
<tr>
<td>FairSquare</td>
<td>4.8</td>
<td>—</td>
<td>16.0</td>
<td>—</td>
<td>36.9</td>
</tr>
<tr>
<td>VeriFair</td>
<td>5.3</td>
<td>2.2</td>
<td>1.2</td>
<td>0.8</td>
<td>15.9</td>
</tr>
<tr>
<td>COMPAS</td>
<td>11.3</td>
<td>9.3</td>
<td>15.9</td>
<td>11.3</td>
<td>295.6</td>
</tr>
<tr>
<td>Justicia</td>
<td>0.1</td>
<td>0.2</td>
<td>0.1</td>
<td>0.2</td>
<td>1.0</td>
</tr>
</tbody>
</table>

**Verification of compound protected groups and robustness**

<table>
<thead>
<tr>
<th>Metric</th>
<th>FairSquare</th>
<th>VeriFair</th>
<th>AIF360</th>
<th>Exact</th>
<th>Justitia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disparate impact</td>
<td>0.99</td>
<td>0.99</td>
<td>0.25</td>
<td>0.26</td>
<td>0.25</td>
</tr>
<tr>
<td>Statistical parity</td>
<td></td>
<td></td>
<td>0.54</td>
<td>0.53</td>
<td>0.54</td>
</tr>
</tbody>
</table>