Justicia: A Stochastic SAT Approach to Formally Verify Fairness

Bishwamittra Ghosh[1], Debabrota Basu[2,3], Kuldeep S. Meel[1]

[1]National University of Singapore, Singapore, [2]Chalmers University of Technology, Goteborg, Sweden, [3]Scool, Inria Lille-Nord Europe, France

Problem Statement

Let X = non-protected attributes, A = protected attributes, $\hat{Y}$ = 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 $\epsilon \in [0, 1]$,

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

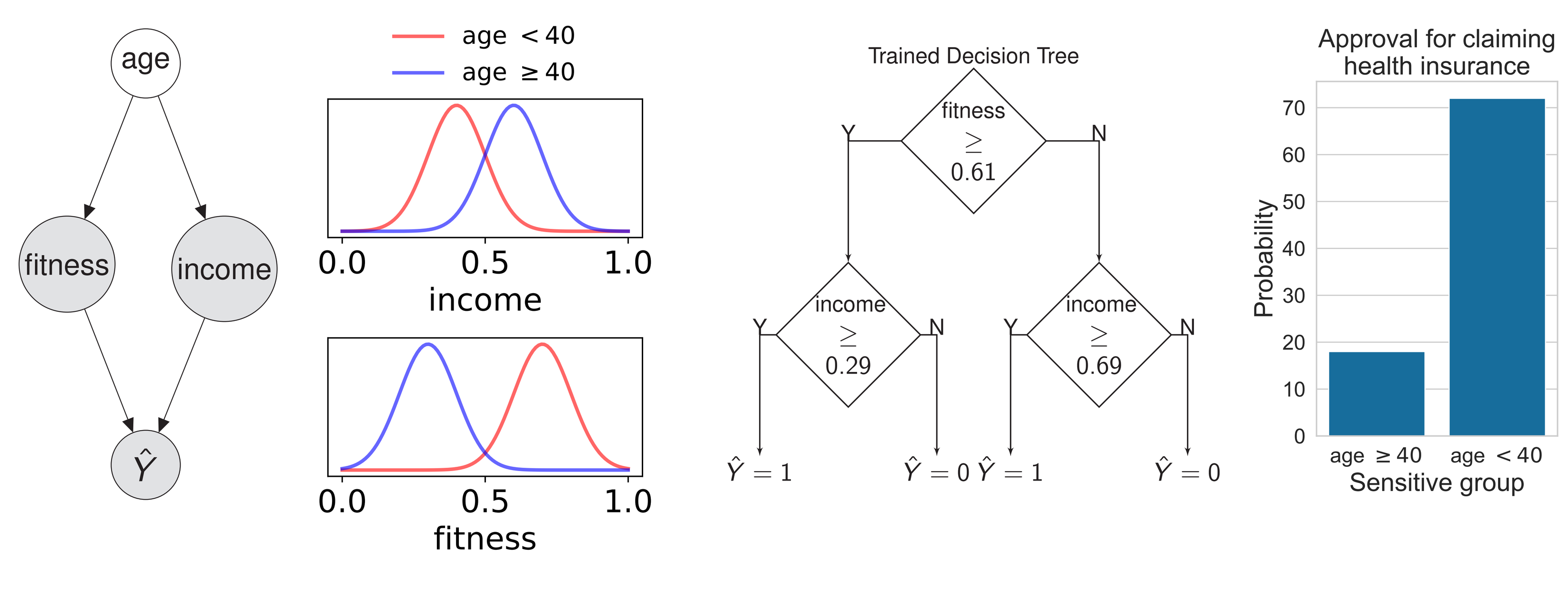
Separation: A classifier satisfies ϵ-statistical parity if, for $\epsilon \in [0, 1]$,

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

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

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

is the building block of verifying different fairness metrics

Stochastic SAT (SSAT)

Compute probability of satisfaction of a CNF formula ϕ given quantification over its variables

$$\Phi = \underbrace{Q_1 X_1, \dots, Q_m X_m}_{\text{prefix}};\ \underbrace{\phi}_{\text{CNF}}$$

where $Q_i \in \{\exists, \forall, \rotatebox[origin=c]{180}{R}^{p_i}\}$ is either an existential ($\exists$), an universal($\forall$),or a randomized($\rotatebox[origin=c]{180}{R}^{p_i}$) 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. $\Pr[\Phi] = p \Pr[\Phi|_X] + (1-p) \Pr[\Phi|_{\neg X}]$ if X is $\rotatebox[origin=c]{180}{R}^p$ quantified

Example. $\Phi = \rotatebox[origin=c]{180}{R}^{0.25} X_1, \exists X_2, \exists X_3;\ (X_1 \vee \neg X_2) \wedge (\neg X_1 \vee X_2 \vee X_3) \wedge (\neg X_1)$ such that $\Pr[\Phi] = 0.75$

Approach 1 : Enumeration

Given a CNF formula $\phi_{\hat{Y}}$ representing the classifier, $\Pr[\hat{Y} = 1 | A = \mathbf{a}]$ can be computed by solving

$$\Phi_{\mathbf{a}} := \underbrace{\rotatebox[origin=c]{180}{R}^{p_1} X_1, \dots, \rotatebox[origin=c]{180}{R}^{p_m} X_m}_{\text{non-protected attributes}}, \underbrace{\exists A_1, \dots, \exists A_n}_{\text{protected attributes}};\ \phi_{\hat{Y}} \wedge (A = \mathbf{a})$$

Example. Let $A \triangleq \text{age} \geq 40$, $F \triangleq \text{fitness} \geq 0.61$, $I \triangleq \text{income} \geq 0.29$, $J \triangleq \text{income} \geq 0.69$

$$\Phi_{\text{age} \geq 40} := \rotatebox[origin=c]{180}{R}^{0.41} F, \rotatebox[origin=c]{180}{R}^{0.93} I, \rotatebox[origin=c]{180}{R}^{0.09} J, \exists A;\ \underbrace{(\neg F \vee I) \wedge (F \vee J)}_{\text{classifier}} \wedge \underbrace{A}_{\text{group}}$$

Disparate impact $= \frac{\Phi_{\text{age} \geq 40}}{\Phi_{\text{age} < 40}} = \frac{0.43}{0.43} = 1$

Statistical parity $= |\Phi_{\text{age} \geq 40} - \Phi_{\text{age} < 40}| = 0$

Encoding Correlation

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

$$\Phi_{\text{age} \geq 40} := \rotatebox[origin=c]{180}{R}^{0.01} F, \rotatebox[origin=c]{180}{R}^{0.99} I, \rotatebox[origin=c]{180}{R}^{0.18} J, \exists A;\ (\neg F \vee I) \wedge (F \vee J) \wedge A$$

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

Approach 2 : Learning

Learning the most favored group

$$\Phi := \exists A, \rotatebox[origin=c]{180}{R}^{0.41} F, \rotatebox[origin=c]{180}{R}^{0.93} I, \rotatebox[origin=c]{180}{R}^{0.09} J;\ (\neg F \vee I) \wedge (F \vee J)$$

Learning the least favored group

$$\Phi := \forall A, \rotatebox[origin=c]{180}{R}^{0.41} F, \rotatebox[origin=c]{180}{R}^{0.93} I, \rotatebox[origin=c]{180}{R}^{0.09} J;\ (\neg F \vee I) \wedge (F \vee J)$$

Experimental Results

Accuracy: Justicia shows less than 1%-error

Metric	FairSquare	VeriFair	AIF360	Exact	Justicia
Disparate impact	0.99	0.99	0.25	0.26	0.25
Statistical parity	—	—	0.54	0.53	0.54

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

Dataset	Ricci		Titanic		COMPAS		Adult	
Classifier	DT	LR	DT	LR	DT	LR	DT	LR
FairSquare	4.8	—	16.0	—	36.9	—	—	—
VeriFair	5.3	2.2	1.2	0.8	15.9	11.3	295.6	61.1
Justicia	0.1	0.2	0.1	0.9	0.1	0.2	0.2	1.0

Verification of compound protected groups and robustness