Fairness Quantification in Machine Learning: The Power of Stochastic Satisfiability

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

[1]National University of Singapore, Singapore, [2]Équipe Scool, Univ. Lille, Inria UMR 9189-CRIStAL, CNRS, Centrale Lille, France

Probabilistic Fairness Quantification

Notation: sensitive features $\mathbf{A}$, non-sensitive features $\mathbf{X}$
Given

- a binary classifier $\mathcal{M} : (\mathbf{X}, \mathbf{A}) \to \widehat{Y} \in \{0, 1\}$
- the probability distribution of features $(\mathbf{X}, \mathbf{A}) \sim \mathcal{D}$

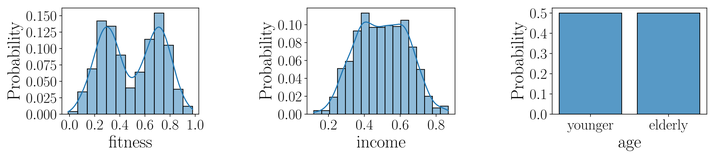
Problem: quantify the fairness of $\mathcal{M}$ with respect to $\mathcal{D}$

Example of Group Fairness Metric

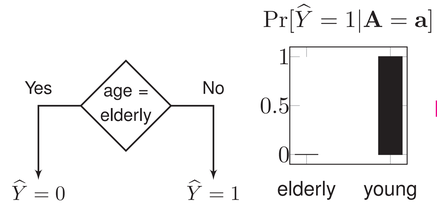
$$\text{Disparate impact, DI} = \frac{\min_{\mathbf{a}} \Pr[\widehat{Y} = 1 | \mathbf{A} = \mathbf{a}]}{\max_{\mathbf{a}} \Pr[\widehat{Y} = 1 | \mathbf{A} = \mathbf{a}]}$$

Illustration of (Un)Fairness

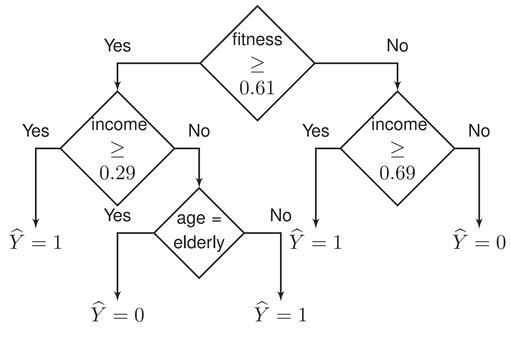
- sensitive features $\mathbf{A}$ = [age], age ∈ {elderly, young}
- non-sensitive features $\mathbf{X}$ = [fitness, income]



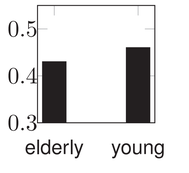
$\Pr[\widehat{Y} = 1 | \mathbf{A} = \mathbf{a}]$



DI $= \frac{0}{1} = 0$



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



DI $= \frac{0.43}{0.46} = 0.93$

References

[GBM-22] Algorithmic Fairness Verification with Graphical Models, AAAI-2022

[GBM-21] Justicia: A Stochastic SAT Approach to Formally Verify Fairness, AAAI-2021

Code

Contribution

Justicia, a Stochastic Satisfiability (SSAT) based framework for fairness quantification with scalable and accurate performance

Key Idea: Encoding conditional probability as SSAT

Stochastic Satisfiability (SSAT)

Compute the probability of satisfaction of a CNF formula ϕ with quantified Boolean variables

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

Quantifier $Q_i \in \{\exists, \forall, \text{ᗺ}^{p_i}\}$ is an existential (∃), an universal(∀), or a random(ᗺ^{p_i}) quantifier with $p_i = \Pr[X_i = 1]$

Semantics

1. $\Pr[\text{ᗺ}^p X \Phi] = p \Pr[\Phi|_X] + (1-p) \Pr[\Phi|_{\neg X}]$ (**expectation**)
2. $\Pr[\exists X \Phi] = \max_X \{\Pr[\Phi|_X], \Pr[\Phi|_{\neg X}]\}$ (**maximization**)
3. $\Pr[\forall X \Phi] = \min_X \{\Pr[\Phi|_X], \Pr[\Phi|_{\neg X}]\}$ (**minimization**)
4. $\Pr[1] = 1, \Pr[0] = 0$

Existential-Random SSAT

$\Phi_{\text{ER}} = \exists X_2, \exists X_3, \text{ᗺ}^{0.75} X_1, \ (\neg X_1 \vee X_2) \wedge (X_1 \vee \neg X_2 \vee \neg X_3) \wedge X_3$
Solution: $\Pr[\Phi_{\text{ER}}] = 0.75, \sigma* = \{X_2 = 1, X_3 = 1\}$

Universal-Random SSAT

$\Phi_{\text{UR}} = \forall X_2, \forall X_3, \text{ᗺ}^{0.75} X_1, \ (\neg X_1 \vee X_2) \wedge (X_1 \vee \neg X_2 \vee \neg X_3) \wedge X_3$
Solution: $\Pr[\Phi_{\text{ER}}] = 0, \sigma* = \{X_2 = 0, X_3 = 0\}$

Fairness Quantification Using SSAT

For the maximum conditional probability $\max_{\mathbf{a}} \Pr[\widehat{Y} = 1 | \mathbf{A} = \mathbf{a}]$, solve

$$\Phi_{\text{ER}} := \underbrace{\exists A_1, \ldots, \exists A_n}_{\text{sensitive features}}, \underbrace{\text{ᗺ}^{p_1} X_1, \ldots, \text{ᗺ}^{p_m} X_m}_{\text{non-sensitive features}}, \underbrace{\phi_{\widehat{Y}}}_{\text{Classifier}}$$

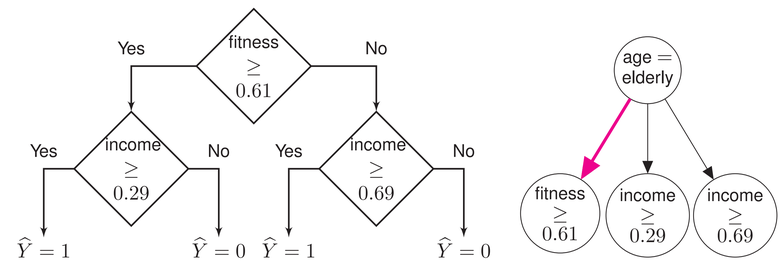
For the mimimum conditional probability $\min_{\mathbf{a}} \Pr[\widehat{Y} = 1 | \mathbf{A} = \mathbf{a}]$, solve

$$\Phi_{\text{UR}} := \forall A_1, \ldots, \forall A_n, \text{ᗺ}^{p_1} X_1, \ldots, \text{ᗺ}^{p_m} X_m, \phi_{\widehat{Y}}$$

$$\text{Disparate impact } = \frac{\Pr[\Phi_{\text{UR}}]}{\Pr[\Phi_{\text{ER}}]}$$

Feature Correlation as Bayesian Network

Encode conditional probability $p = \Pr[X = 1 | \text{parent}(X) = \mathbf{x}]$ from a Bayesian Network using additional variables and clauses



Variables

- $X_1 :=$ fitness ≥ 0.61
- $A_1 :=$ age = elderly
- $Z_1 :=$ fitness $\geq 0.61 \mid$ age = elderly

Clauses

- $A_1 \wedge Z_1 \to X_1$
- $A_1 \wedge \neg Z_1 \to \neg X_1$

Experimental Results

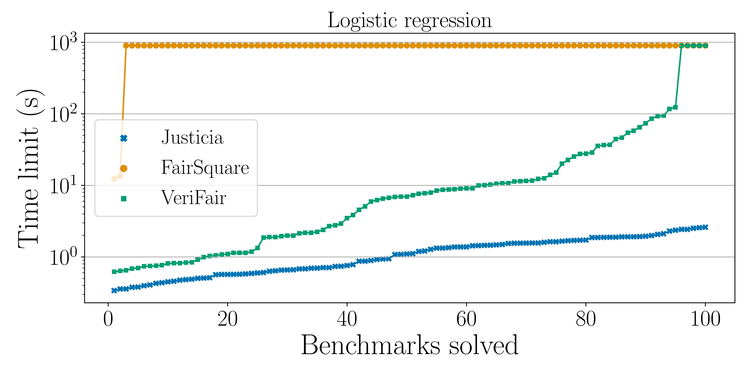


Figure 1: *Higher scalability*: Justicia achieves 1 to 2 orders improvement

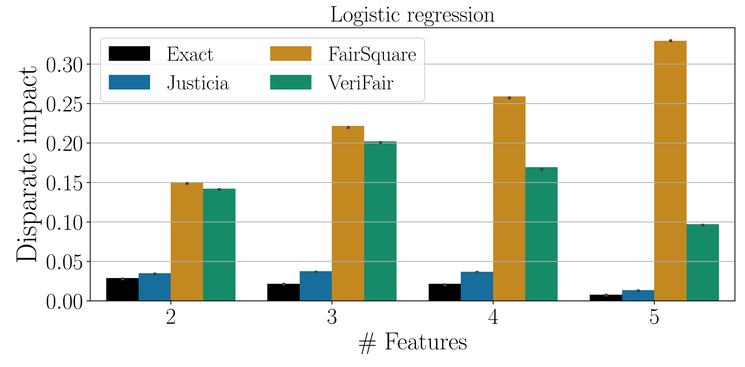


Figure 2: *Improved accuracy* due to encoding feature correlations