Fairness Quantification in Machine Learning: The Power of Stochastic Satisfiability

Bishwamittra Ghosh
Ph.D. candidate
National University of Singapore (NUS)

- Non-sensitive features, $\mathbf{X} = [\text{fitness, income}]$
- Sensitive features, $\mathbf{A} = [\text{age}]$
  - Two sensitive groups: age $\in$ {elderly, young}

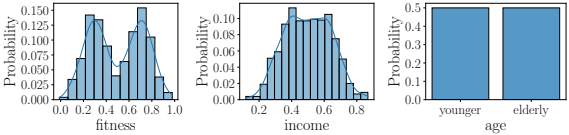


Figure: Distribution of features

- Non-sensitive features, $\mathbf{X} = [\text{fitness, income}]$
- Sensitive features, $\mathbf{A} = [\text{age}]$
  - Two sensitive groups: age $\in$ {elderly, young}

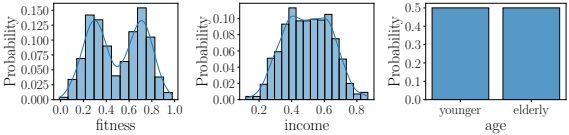


Figure: Distribution of features

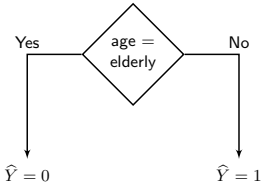
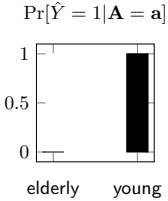


Figure: A decision tree



- Non-sensitive features, $\mathbf{X} = [\text{fitness, income}]$
- Sensitive features, $\mathbf{A} = [\text{age}]$
  - Two sensitive groups: age $\in$ {elderly, young}

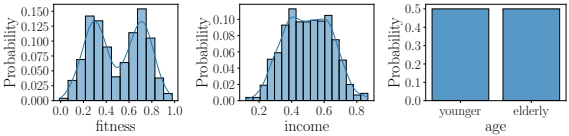


Figure: Distribution of features

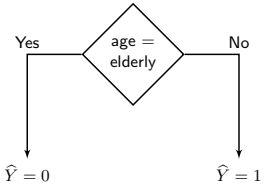
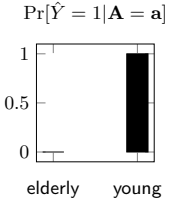


Figure: A decision tree



$$\text{Disparate impact} = \frac{\min_{\mathbf{a}} \Pr[\widehat{Y} = 1|\mathbf{A} = \mathbf{a}]}{\max_{\mathbf{a}} \Pr[\widehat{Y} = 1|\mathbf{A} = \mathbf{a}]} = \frac{0}{1} = 0$$

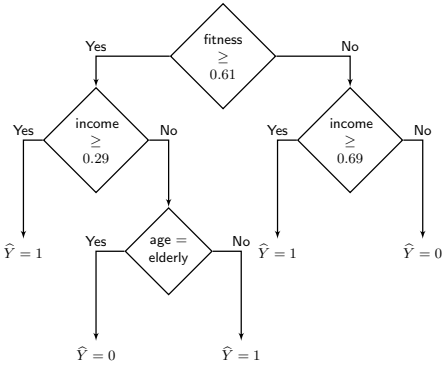


Figure: Another decision tree

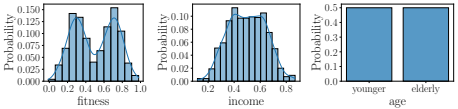
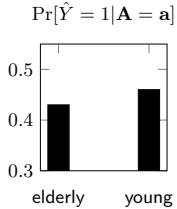


Figure: Distribution of features



$$\text{Disparate impact} = \frac{\min_{\mathbf{a}} \Pr[\widehat{Y} = 1|\mathbf{A} = \mathbf{a}]}{\max_{\mathbf{a}} \Pr[\widehat{Y} = 1|\mathbf{A} = \mathbf{a}]} = \frac{0.43}{0.46} = 0.93$$

Probabilistic Fairness Quantification

Given

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

Problem: quantify the fairness of the classifier $f(\mathcal{M}, \mathcal{D})$

Probabilistic Fairness Quantification

Given

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

Problem: quantify the fairness of the classifier $f(\mathcal{M}, \mathcal{D})$

Group fairness metrics

- Disparate impact
- Statistical parity
- Equalized odds
- Predictive parity

Disparate impact [FFM+15]

$$f_{\mathsf{DI}}(\mathcal{M}, \mathcal{D}) = \frac{\min_{\mathbf{a}} \Pr[\widehat{Y} = 1 | \mathbf{A} = \mathbf{a}]}{\max_{\mathbf{a}} \Pr[\widehat{Y} = 1 | \mathbf{A} = \mathbf{a}]}$$

Probabilistic Fairness Quantification

Given

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

Problem: quantify the fairness of the classifier $f(\mathcal{M}, \mathcal{D})$

Group fairness metrics

- Disparate impact
- Statistical parity
- Equalized odds
- Predictive parity

Disparate impact [FFM+15]

$$f_{\mathsf{DI}}(\mathcal{M}, \mathcal{D}) = \frac{\min_{\mathbf{a}} \Pr[\widehat{Y} = 1 | \mathbf{A} = \mathbf{a}]}{\max_{\mathbf{a}} \Pr[\widehat{Y} = 1 | \mathbf{A} = \mathbf{a}]}$$

Scope of our work

- Classifiers: decision trees, linear classifiers, SVM with linear kernels
- Input distribution: (i) product distribution and (ii) Bayesian Network

Existing approaches

- FairSquare: computing weighted volume of SMT encoding of conditional probability [ADDN17]
  - monotonic convergence to exact values
- VeriFair: a probabilistic fairness quantifier based on sampling [BZSL19]
  - correctness via adaptive concentration inequalities

Limitations of state of the art

- Limited scalability due to integrating SMT formulas or via sampling
- Limited accuracy due to handling specific distribution
- Limited to a single Boolean sensitive feature
  - e.g., either race or gender

Our Contribution [GBM22, GBM21]

Justicia: a probabilistic fairness quantifier based on Stochastic Satisfiability (SSAT)

- Encoding (maximum or minimum) conditional probability as an SSAT formula

Compound sensitive groups

- Sensitive features = [race, gender]
- Compound sensitive groups = {[White, male], [Black, male], [White, female], [Black, female]}

Multiple fairness metrics

- Disparate impact
- Statistical parity
- Equalized odds

Application: fairness auditing

- fairness improvement algorithm [CWV+17, KC12, ZWS+13, ZLM18, HPS16, KKZ12, SBC20]
- fairness attack algorithm [SBC20]

Satisfiability (SAT)

Given

- a Boolean formula ϕ in CNF (Conjunctive Normal Form) defined over Boolean variables $\mathbf{X}$

Problem: find a satisfying assignment of $\mathbf{X}$ that evaluates ϕ to true

Example

$$\phi = (\neg X_1 \vee X_2) \wedge (X_1 \vee \neg X_2 \vee \neg X_3) \wedge X_3$$

- Satisfying assignment: $X_1 = \text{true}$, $X_2 = \text{true}$, $X_3 = \text{true}$

Stochastic SAT (SSAT)

An SSAT formula Φ has a prefix and a CNF formula ϕ

$$\Phi = \underbrace{q_1 X_1, \ldots, q_n X_n}_{\text{prefix}}, \phi$$

SSAT computes the **probability of satisfaction** $\Pr[\Phi]$

Quantifier $q_i \in \{\text{R}^{p_i}, \exists, \forall\}$

- Random variables
  - random quantifier R^{p_i} with $p_i = \Pr[X_i = \text{ true}]$
- Optimization variables
  - existential quantifier $\exists$ to maximize $\Pr[\Phi]$
  - universal quantifier $\forall$ to minimize $\Pr[\Phi]$

Semantics of SSAT

Let X_i be the left-most variable in the prefix of Φ

1. If X_i is randomized quantified ($\mathbb{R}^{p_i}$)

$$\Pr[\mathbb{R}^p X_i \Phi] = p_i \Pr[\Phi|_{X_i}] + (1 - p_i) \Pr[\Phi|_{\neg X_i}]$$

2. If X_i is existentially quantified ($\exists$)

$$\Pr[\exists X_i \Phi] = \max_{X_i} \{\Pr[\Phi|_{X_i}], \Pr[\Phi|_{\neg X_i}]\}$$

3. If X_i is universally quantified ($\forall$)

$$\Pr[\forall X_i \Phi] = \min_{X_i} \{\Pr[\Phi|_{X_i}], \Pr[\Phi|_{\neg X_i}]\}$$

4. $\Pr[\text{true}] = 1$, $\Pr[\text{false}] = 0$

Examples of SSAT Formulas

- **Existential-random SSAT formula**

$$\Phi_{\mathsf{ER}} = \exists X_2, \exists X_3, \text{ꓤ}^{0.75} X_1,\ (\neg X_1 \vee X_2) \wedge (X_1 \vee \neg X_2 \vee \neg X_3) \wedge X_3$$

  - $\Pr[\Phi_{\mathsf{ER}}] = 0.75$
  - Optimal assignment of $\exists$ variables: $X_2 = \mathsf{true}$, $X_3 = \mathsf{true}$

Examples of SSAT Formulas

- **Existential-random SSAT formula**

$$\Phi_{\mathsf{ER}} = \exists X_2, \exists X_3, \rotatebox[origin=c]{180}{R}^{0.75} X_1,\ (\neg X_1 \vee X_2) \wedge (X_1 \vee \neg X_2 \vee \neg X_3) \wedge X_3$$

  - $\Pr[\Phi_{\mathsf{ER}}] = 0.75$
  - Optimal assignment of $\exists$ variables: $X_2 =$ true, $X_3 =$ true

$$(\neg X_1 \vee X_2) \wedge \underbrace{(X_1 \vee \neg X_2 \vee \neg X_3)}_{\text{expected probability of satisfaction } = 0.75} \wedge X_3$$

Examples of SSAT Formulas

- **Universal-random SSAT formula**

$$\Phi_{\mathsf{UR}} = \forall X_2, \forall X_3, \mathbb{R}^{0.75} X_1,\ (\neg X_1 \vee X_2) \wedge (X_1 \vee \neg X_2 \vee \neg X_3) \wedge X_3$$

  - $\Pr[\Phi_{\mathsf{UR}}] = 0$
  - Optimal assignment of $\forall$ variables: $X_2 = \text{false},\ X_3 = \text{false}$

Examples of SSAT Formulas

- **Universal-random SSAT formula**

$$\Phi_{\mathsf{UR}} = \forall X_2, \forall X_3, \mathfrak{R}^{0.75} X_1,\ (\neg X_1 \vee X_2) \wedge (X_1 \vee \neg X_2 \vee \neg X_3) \wedge X_3$$

  - $\Pr[\Phi_{\mathsf{UR}}] = 0$
  - Optimal assignment of $\forall$ variables: $X_2 = \text{false},\ X_3 = \text{false}$

$$(\neg X_1 \vee X_2) \wedge (X_1 \vee \neg X_2 \vee \neg X_3) \wedge \underbrace{X_3}_{\text{unsatisfied}}$$

Examples of SSAT Formulas

- **Universal-random SSAT formula**

$$\Phi_{\mathsf{UR}} = \forall X_2, \forall X_3, \mathfrak{R}^{0.75} X_1,\ (\neg X_1 \vee X_2) \wedge (X_1 \vee \neg X_2 \vee \neg X_3) \wedge X_3$$

  - $\Pr[\Phi_{\mathsf{UR}}] = 0$
  - Optimal assignment of $\forall$ variables: $X_2 = \text{false}$, $X_3 = \text{false}$

$$(\neg X_1 \vee X_2) \wedge (X_1 \vee \neg X_2 \vee \neg X_3) \wedge \underbrace{X_3}_{\text{unsatisfied}}$$

The decision problem of existential-random and universal-random SSAT formulas is $\mathrm{NP}^{\mathrm{PP}}$ [LMP01]

Fairness Quantification Based on SSAT Encoding

Preprocessing

- features $\mathbf{X} \cup \mathbf{A}$ are Boolean and independent variables
- predicted class-label $\widehat{Y}$ is encoded as a CNF formula $\phi_{\widehat{Y}}$ defined on $\mathbf{X} \cup \mathbf{A}$

$$\widehat{Y} = 1 \Leftrightarrow \phi_{\widehat{Y}} \text{ is satisfied}$$

Fairness Quantification Based on SSAT Encoding

Preprocessing

- features $\mathbf{X} \cup \mathbf{A}$ are Boolean and independent variables
- predicted class-label $\widehat{Y}$ is encoded as a CNF formula $\phi_{\widehat{Y}}$ defined on $\mathbf{X} \cup \mathbf{A}$

$$\widehat{Y} = 1 \Leftrightarrow \phi_{\widehat{Y}} \text{ is satisfied}$$

The computation of

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

is equivalent to solving

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

Fairness Quantification Based on SSAT Encoding

Preprocessing

- features $\mathbf{X} \cup \mathbf{A}$ are Boolean and independent variables
- predicted class-label $\widehat{Y}$ is encoded as a CNF formula $\phi_{\widehat{Y}}$ defined on $\mathbf{X} \cup \mathbf{A}$

$$\widehat{Y} = 1 \Leftrightarrow \phi_{\widehat{Y}} \text{ is satisfied}$$

The computation of

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

is equivalent to solving

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

For computing $\min_{\mathbf{a}} \Pr[\widehat{Y} = 1 | \mathbf{A} = \mathbf{a}]$, solve

$$\Phi_{\mathsf{UR}} := \underbrace{\forall A_1, \ldots, \forall A_n}_{\text{sensitive features}}, \underbrace{\mathsf{R}^{p_1} X_1, \ldots, \mathsf{R}^{p_m} X_m}_{\text{non-sensitive features}}, \phi_{\widehat{Y}}$$

Illustration of Fairness Quantification via SSAT Encoding

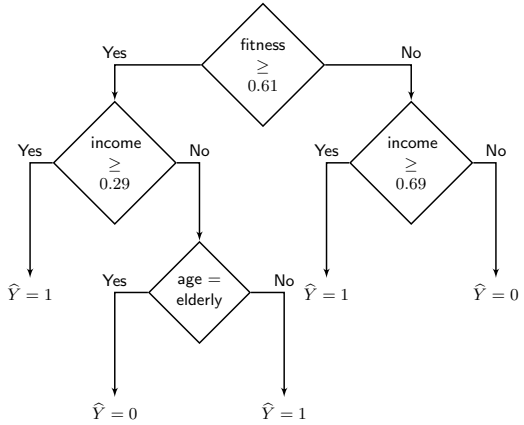
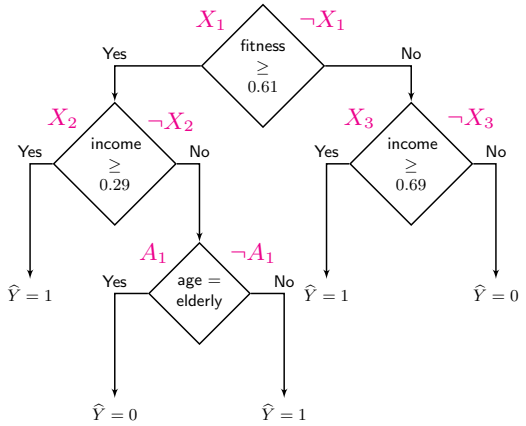


Illustration of Fairness Quantification via SSAT Encoding



CNF representation: $(\neg X_1 \vee X_2 \vee \neg A_1) \wedge (X_1 \vee X_3)$

Illustration of Fairness Quantification via SSAT Encoding

- CNF classifier $(\neg X_1 \vee X_2 \vee \neg A_1) \wedge (X_1 \vee X_3)$
- Independent probabilities
  - $\Pr[X_1 = \mathsf{true}] = 0.41$
  - $\Pr[X_2 = \mathsf{true}] = 0.93$
  - $\Pr[X_3 = \mathsf{true}] = 0.09$

Illustration of Fairness Quantification via SSAT Encoding

- CNF classifier $(\neg X_1 \vee X_2 \vee \neg A_1) \wedge (X_1 \vee X_3)$
- Independent probabilities
  - $\Pr[X_1 = \text{true}] = 0.41$
  - $\Pr[X_2 = \text{true}] = 0.93$
  - $\Pr[X_3 = \text{true}] = 0.09$
- To compute $\max_{a_1} \Pr[\widehat{Y} = 1 | A_1 = a_1]$, solve

$$\Phi_{\mathsf{ER}} = \exists A_1, \text{ꓤ}^{0.41} X_1, \text{ꓤ}^{0.93} X_2, \text{ꓤ}^{0.09} X_3, \ (\neg X_1 \vee X_2 \vee \neg A_1) \wedge (X_1 \vee X_3)$$

Illustration of Fairness Quantification via SSAT Encoding

- CNF classifier $(\neg X_1 \vee X_2 \vee \neg A_1) \wedge (X_1 \vee X_3)$
- Independent probabilities
  - $\Pr[X_1 = \text{true}] = 0.41$
  - $\Pr[X_2 = \text{true}] = 0.93$
  - $\Pr[X_3 = \text{true}] = 0.09$
- To compute $\max_{a_1} \Pr[\widehat{Y} = 1 | A_1 = a_1]$, solve

$$\Phi_{\mathsf{ER}} = \exists A_1, \text{ꓤ}^{0.41} X_1, \text{ꓤ}^{0.93} X_2, \text{ꓤ}^{0.09} X_3, \ (\neg X_1 \vee X_2 \vee \neg A_1) \wedge (X_1 \vee X_3)$$

- $\Pr[\Phi_{\mathsf{ER}}] = 0.46$

Illustration of Fairness Quantification via SSAT Encoding

- CNF classifier $(\neg X_1 \vee X_2 \vee \neg A_1) \wedge (X_1 \vee X_3)$
- Independent probabilities
  - $\Pr[X_1 = \text{true}] = 0.41$
  - $\Pr[X_2 = \text{true}] = 0.93$
  - $\Pr[X_3 = \text{true}] = 0.09$
- To compute $\max_{a_1} \Pr[\widehat{Y} = 1 | A_1 = a_1]$, solve

$$\Phi_{\mathsf{ER}} = \exists A_1, \text{ᴙ}^{0.41} X_1, \text{ᴙ}^{0.93} X_2, \text{ᴙ}^{0.09} X_3,\ (\neg X_1 \vee X_2 \vee \neg A_1) \wedge (X_1 \vee X_3)$$

- $\Pr[\Phi_{\mathsf{ER}}] = 0.46$
- To compute $\min_{a_1} \Pr[\widehat{Y} = 1 | A_1 = a_1]$, solve

$$\Phi_{\mathsf{UR}} = \forall A_1, \text{ᴙ}^{0.41} X_1, \text{ᴙ}^{0.93} X_2, \text{ᴙ}^{0.09} X_3,\ (\neg X_1 \vee X_2 \vee \neg A_1) \wedge (X_1 \vee X_3)$$

- $\Pr[\Phi_{\mathsf{UR}}] = 0.43$

Illustration of Fairness Quantification via SSAT Encoding

- CNF classifier $(\neg X_1 \vee X_2 \vee \neg A_1) \wedge (X_1 \vee X_3)$
- Independent probabilities
  - $\Pr[X_1 = \text{true}] = 0.41$
  - $\Pr[X_2 = \text{true}] = 0.93$
  - $\Pr[X_3 = \text{true}] = 0.09$
- To compute $\max_{a_1} \Pr[\widehat{Y} = 1 | A_1 = a_1]$, solve

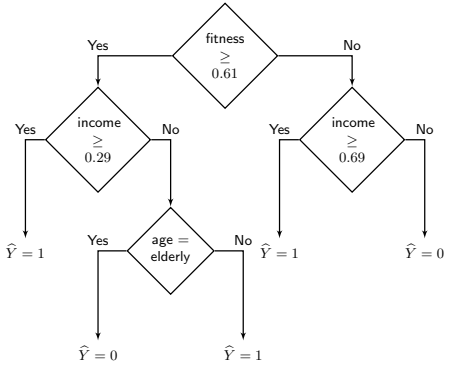
$$\Phi_{\mathsf{ER}} = \exists A_1, \text{ᴙ}^{0.41} X_1, \text{ᴙ}^{0.93} X_2, \text{ᴙ}^{0.09} X_3,\ (\neg X_1 \vee X_2 \vee \neg A_1) \wedge (X_1 \vee X_3)$$

- $\Pr[\Phi_{\mathsf{ER}}] = 0.46$
- To compute $\min_{a_1} \Pr[\widehat{Y} = 1 | A_1 = a_1]$, solve

$$\Phi_{\mathsf{UR}} = \forall A_1, \text{ᴙ}^{0.41} X_1, \text{ᴙ}^{0.93} X_2, \text{ᴙ}^{0.09} X_3,\ (\neg X_1 \vee X_2 \vee \neg A_1) \wedge (X_1 \vee X_3)$$

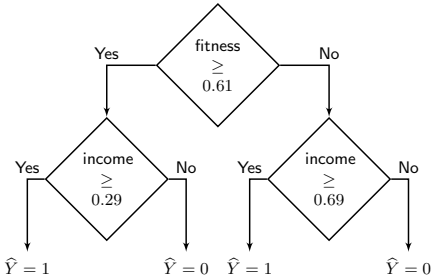
- $\Pr[\Phi_{\mathsf{UR}}] = 0.43$
- Disparate impact is $\frac{0.43}{0.46} = 0.93$

Illustration of Fairness Quantification via SSAT Encoding



CNF: $(\neg X_1 \vee X_2 \vee \neg A_1) \wedge (X_1 \vee X_3)$

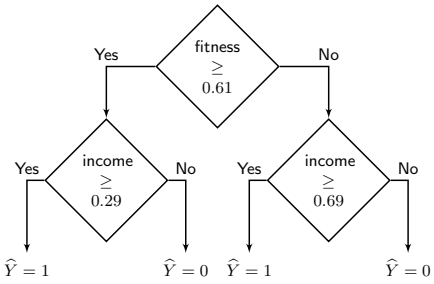
Disparate impact $= \frac{0.43}{0.46} = 0.93$

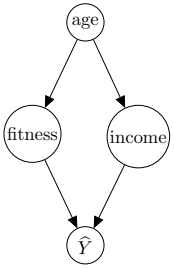


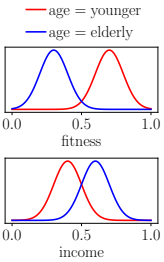
CNF: $(\neg X_1 \vee X_2) \wedge (X_1 \vee X_3)$

Disparate impact $= \frac{0.43}{0.43} = 1$

Example: Feature Correlation in Fairness Quantification







Disparate impact $= \frac{0.18}{0.72} = 0.25$

Bayesian Network as Input Distribution

When X_i is randomly quantified $\mathfrak{R}^p$

- Independent probability:

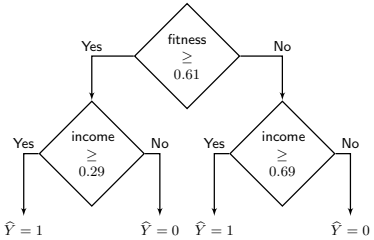
$$p = \Pr[X_i = \text{ true}]$$

- Conditional probability:

$$p_{\text{cond}} = \Pr[X_i = \text{ true } | \text{parent}(X_i) = \mathbf{x}]$$

  - $\text{parent}(X_i) \subseteq \mathbf{X} \cup \mathbf{A}$

Illustration of Encoding Conditional Probabilities [CD08]



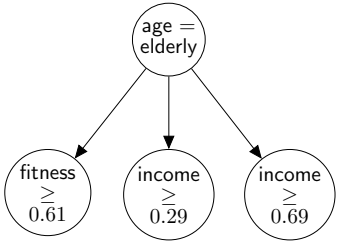
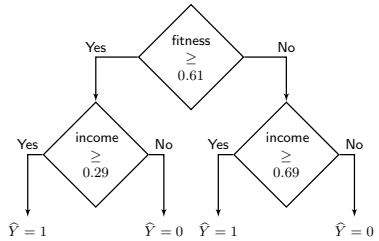
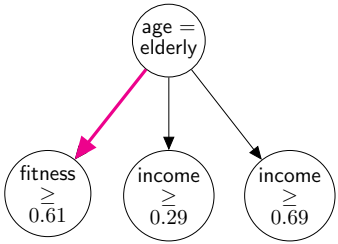


Illustration of Encoding Conditional Probabilities [CD08]





Variables:

$X_1 :=$ fitness ≥ 0.61

$A_1 :=$ age $=$ elderly

$Z_1 :=$ fitness $\geq 0.61 \mid$ age $=$ elderly

$\ldots$

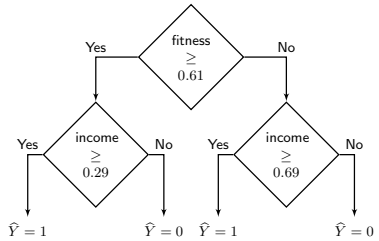
Clauses to be added:

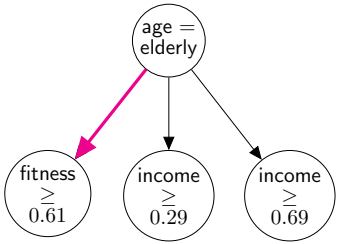
$A_1 \wedge Z_1 \rightarrow X_1$

$A_1 \wedge \neg Z_1 \rightarrow \neg X_1$

$\ldots$

Illustration of Encoding Conditional Probabilities [CD08]





Variables:

$X_1 :=$ fitness ≥ 0.61

$A_1 :=$ age $=$ elderly

$Z_1 :=$ fitness $\geq 0.61 \mid$ age $=$ elderly

$\ldots$

Clauses to be added:

$A_1 \wedge Z_1 \rightarrow X_1$

$A_1 \wedge \neg Z_1 \rightarrow \neg X_1$

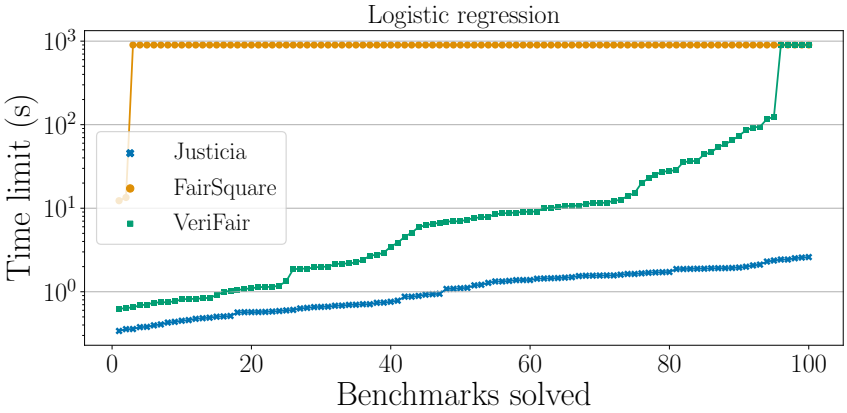
$\ldots$

For a Bayesian Network with n vertices and c independent parameters

- additional variables $= n + c$
- additional clauses $= 2(n + c)$

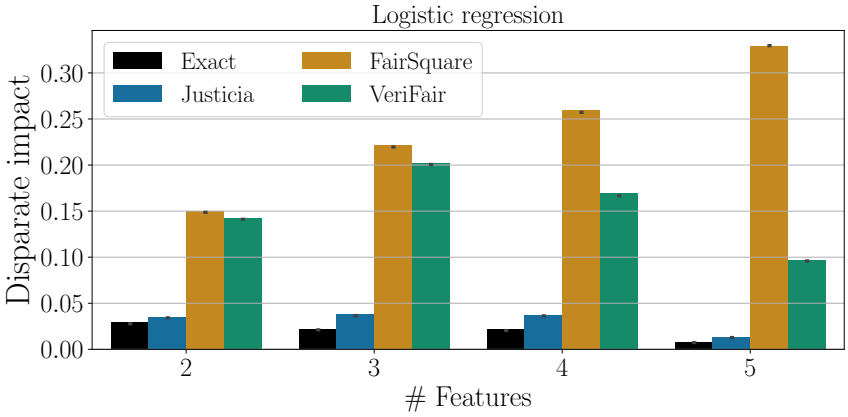
Experimental Results: Scalability on Linear Classifiers

Tractable fairness quantification based on dynamic programming [GBM22]



- The number of features varies between 5 to 26

Accuracy of Quantifying Disparate Impact



- Exact value of disparate impact is calculated for Gaussian non-sensitive features

Conclusion

- Probabilistic fairness quantifier estimates classifiers' unfairness beyond dataset
- Prior approaches demonstrate limitation in scalability or accuracy or both
- Combination of formal methods and machine learning addresses the limitations
  - Justicia demonstrates higher **scalability** and **accuracy** than the state of the art

Open-source tools

Justicia

Conclusion

- Probabilistic fairness quantifier estimates classifiers' unfairness beyond dataset
- Prior approaches demonstrate limitation in scalability or accuracy or both
- Combination of formal methods and machine learning addresses the limitations
  - Justicia demonstrates higher **scalability** and **accuracy** than the state of the art

Open-source tools

Justicia

Thank you all!

Contact: bghosh@u.nus.edu

References I

[ADDN17] Aws Albarghouthi, Loris D'Antoni, Samuel Drews, and Aditya V Nori.
FairSquare: probabilistic verification of program fairness.
Proceedings of the ACM on Programming Languages, 1(OOPSLA):1–30, 2017.

[ALSA+17] Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin.
Certifiably optimal rule lists for categorical data.
In *Proceedings of the 23rd ACM SIGKDD Conference of Knowledge, Discovery, and Data Mining (KDD)*, volume 2, pages 229–246. Springer, 2017.

[BDH+18] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang.
Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, Oct 2018.

[BHN17] Solon Barocas, Moritz Hardt, and Arvind Narayanan.
Fairness in machine learning.
NIPS Tutorial, 1, 2017.

[BHO09] Christian Bessiere, Emmanuel Hebrard, and Barry O'Sullivan.
Minimising decision tree size as combinatorial optimisation.
In *International Conference on Principles and Practice of Constraint Programming*, pages 173–187. Springer, 2009.

[BSFF20] Tom Begley, Tobias Schwedes, Christopher Frye, and Ilya Feige.
Explainability for fair machine learning.
arXiv preprint arXiv:2010.07389, 2020.

[BZSL19] Osbert Bastani, Xin Zhang, and Armando Solar-Lezama.
Probabilistic verification of fairness properties via concentration.
Proceedings of the ACM on Programming Languages, 3(OOPSLA):1–27, 2019.

[CD08] Mark Chavira and Adnan Darwiche.
On probabilistic inference by weighted model counting.
Artificial Intelligence, 172(6-7):772–799, 2008.

[CN89] P. Clark and T. Niblett.
The CN2 induction algorithm.
Mar. 1989.

References II

[Coh95] William W Cohen.
Fast effective rule induction.
In *Machine learning proceedings 1995*, pages 115–123. Elsevier, 1995.

[CS99] W. W. Cohen and Y. Singer.
A simple, fast, and effective rule learner.
In *Proc. of AAAI*, Orlando, FL, Jul. 1999.

[CWV+17] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney.
Optimized pre-processing for discrimination prevention.
In *Advances in Neural Information Processing Systems*, pages 3992–4001, 2017.

[DG21] Sanjeeb Dash and Joao Goncalves.
LPRules: Rule induction in knowledge graphs using linear programming.
2021.

[EKAK17] Bradley J Erickson, Panagiotis Korfiatis, Zeynettin Akkus, and Timothy L Kline.
Machine learning for medical imaging.
volume 37, pages 505–515. Radiological Society of North America, 2017.

[FFM+15] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian.
Certifying and removing disparate impact.
In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[GBM21] Bishwamittra Ghosh, Debabrota Basu, and Kuldeep S. Meel.
Justicia: A stochastic SAT approach to formally verify fairness.
In *Proceedings of AAAI*, 2 2021.

[GBM22] Bishwamittra Ghosh, Debabrota Basu, and Kuldeep S. Meel.
Algorithmic fairness verification with graphical models.
In *Proceedings of AAAI*, 2 2022.

[GM19] Bishwamittra Ghosh and Kuldeep S. Meel.
IMLI: An incremental framework for MaxSAT-based learning of interpretable classification rules.
In *Proceedings AIES*, 2019.

References III

[GMM20] Bishwamittra Ghosh, Dmitry Malioutov, and Kuldeep S. Meel.
Classification rules in relaxed logical form.
In *Proceedings of ECAI*, 2020.

[GMM22] Bishwamittra Ghosh, Dmitry Malioutov, and Kuldeep S. Meel.
Efficient learning of interpretable classification rules.
In *Proceedings of JAIR*, 2022.

[HPS16] Moritz Hardt, Eric Price, and Nati Srebro.
Equality of opportunity in supervised learning.
In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[IIMS20] Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva.
On explaining decision trees.
2020.

[IMSNS21] Alexey Ignatiev, Joao Marques-Silva, Nina Narodytska, and Peter J Stuckey.
Reasoning-based learning of interpretable ML models.
In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

[KC12] Faisal Kamiran and Toon Calders.
Data preprocessing techniques for classification without discrimination.
Knowledge and Information Systems, 33(1):1–33, 2012.

[KKZ12] Faisal Kamiran, Asim Karim, and Xiangliang Zhang.
Decision theory for discrimination-aware classification.
In *2012 IEEE 12th International Conference on Data Mining*, pages 924–929. IEEE, 2012.

[KMRB20] Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren.
Secure, privacy-preserving and federated machine learning in medical imaging.
volume 2, pages 305–311. Nature Publishing Group, 2020.

[KOAV18] Ram Shankar Siva Kumar, David R O'Brien, Kendra Albert, and Salome Vilojen.
Law and adversarial machine learning.
2018.

References IV

[Kon01] Igor Kononenko.
Machine learning for medical diagnosis: history, state of the art and perspective.
volume 23, pages 89–109. Elsevier, 2001.

[LBL16] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec.
Interpretable decision sets: A joint framework for description and prediction.
In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1675–1684, 2016.

[LKCL17] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec.
Interpretable & explorable approximations of black box models.
2017.

[LMP01] Michael L Littman, Stephen M Majercik, and Toniann Pitassi.
Stochastic boolean satisfiability.
Journal of Automated Reasoning, 27(3):251–296, 2001.

[LRMM15] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, and David Madigan.
Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model.
volume 9, pages 1350–1371. Institute of Mathematical Statistics, 2015.

[Luc18] Rosemary Luckin.
Machine Learning and Human Intelligence: The future of education for the 21st century.
ERIC, 2018.

[Lun20] Scott M Lundberg.
Explaining quantitative measures of fairness.
In *Fair & Responsible AI Workshop@ CHI2020*, 2020.

[MV13] Dmitry Malioutov and Kush Varshney.
Exact rule learning via boolean compressed sensing.
In *International Conference on Machine Learning*, pages 765–773. PMLR, 2013.

[NIP+18] Nina Narodytska, Alexey Ignatiev, Filipe Pereira, Joao Marques-Silva, and IS RAS.
Learning optimal decision trees with sat.
In *IJCAI*, pages 1362–1368, 2018.

References V

[PRP19] Inon Peled, Filipe Rodrigues, and Francisco Câmara Pereira.
Model-based machine learning for transportation.
In *Mobility patterns, big data and transport analytics*, pages 145–171. Elsevier, 2019.

[Qui93] J Ross Quinlan.
C4. 5: Programming for machine learning.
1993.

[Rud19] Cynthia Rudin.
Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.
volume 1, pages 206–215. Nature Publishing Group, 2019.

[SBC20] David Solans, Battista Biggio, and Carlos Castillo.
Poisoning attacks on algorithmic fairness.
arXiv preprint arXiv:2004.07401, 2020.

[Sur14] Harry Surden.
Machine learning and law.
volume 89, page 87. HeinOnline, 2014.

[SWVM16] Guolong Su, Dennis Wei, Kush R Varshney, and Dmitry M Malioutov.
Learning sparse two-level boolean rules.
In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2016.

[VR18] Sahil Verma and Julia Rubin.
Fairness definitions explained.
In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE, 2018.

[WR15] Fulton Wang and Cynthia Rudin.
Falling rule lists.
In *Artificial Intelligence and Statistics*, pages 1013–1022. PMLR, 2015.

[ZKKK19] Fotios Zantalis, Grigorios Koulouras, Sotiris Karabetsos, and Dionisis Kandris.
A review of machine learning and IoT in smart transportation.
volume 11, page 94. Multidisciplinary Digital Publishing Institute, 2019.

References VI

[ZLM18] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell.
Mitigating unwanted biases with adversarial learning.
In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[ZWS+13] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork.
Learning fair representations.
In *International Conference on Machine Learning*, pages 325–333, 2013.