Fairness: What is the Right Thing to Do? A Comparative Study of Fairness-Preserving Machine Learning Algorithms

Guanqun Yang

University of California, Los Angeles
Department of Electrical and Computer Engineering

CS 260 Course Project, Fall 2018

Outline

Background and Motivation

- Algorithmic Decision Making (ADM) system is widely used in daily life
  - GRE e-Rater
  - Credit scoring
  - Job applicant selection
  - Many others...
- But they do not necessarily give fair predictions

Background and Motivation

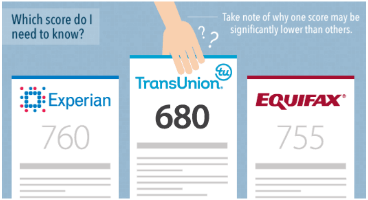
- Algorithmic Decision Making (ADM) system is widely used in daily life
  - GRE e-Rater
  - Credit scoring
  - Job applicant selection
  - Many others...
- But they do not necessarily give fair predictions

UCLA

Background and Motivation

- Algorithmic Decision Making (ADM) system is widely used in daily life
  - GRE e-Rater
  - Credit scoring
  - Job applicant selection
  - Many others...
- But they do not necessarily give fair predictions

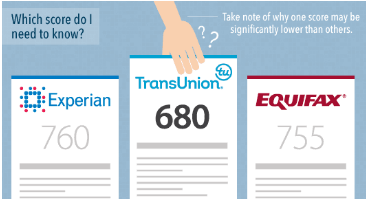
Background and Motivation



- Algorithmic Decision Making (ADM) system is widely used in daily life
  - GRE e-Rater
  - Credit scoring
  - Job applicant selection
  - Many others...
- But they do not necessarily give fair predictions

Background and Motivation

ETS e-rater

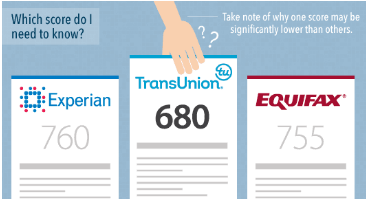


- Algorithmic Decision Making (ADM) system is widely used in daily life
  - GRE e-Rater
  - Credit scoring
  - Job applicant selection
  - Many others...
- But they do not necessarily give fair predictions

UCLA

Background and Motivation

ETS e-rater

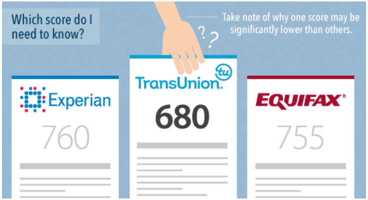


- Algorithmic Decision Making (ADM) system is widely used in daily life
    - GRE e-Rater
    - Credit scoring
    - Job applicant selection
    - Many others...
- But they do not necessarily give fair predictions

UCLA

Background and Motivation

ETS e-rater

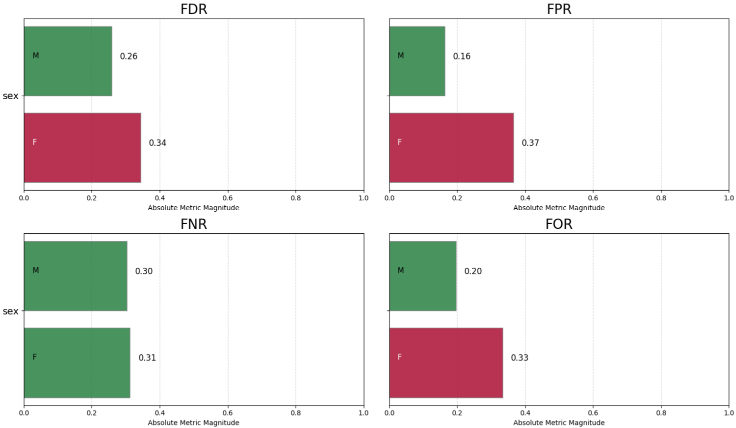


- Algorithmic Decision Making (ADM) system is widely used in daily life
  - GRE e-Rater
  - Credit scoring
  - Job applicant selection
  - Many others...
- But they do not necessarily give fair predictions

UCLA

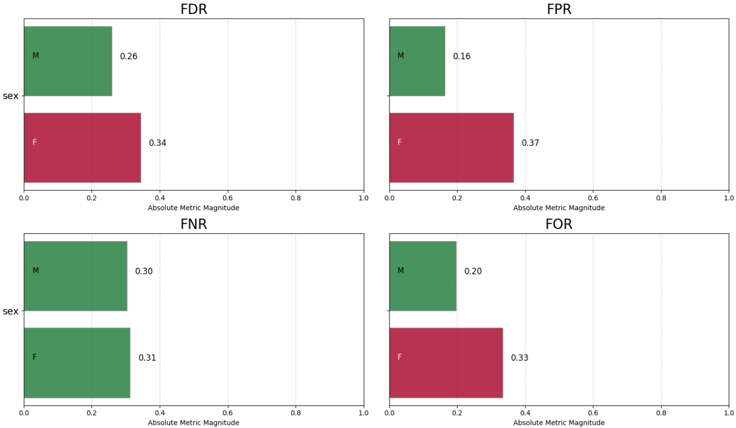
Two Numerical Examples - Student Performance Dataset

- Underrepresented groups are biased by machine learning algorithms
- Female



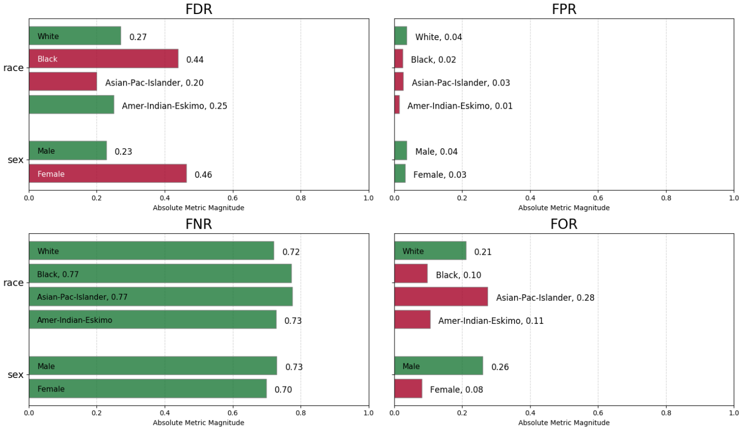
Two Numerical Examples - Student Performance Dataset

- Underrepresented groups are biased by machine learning algorithms
- Female



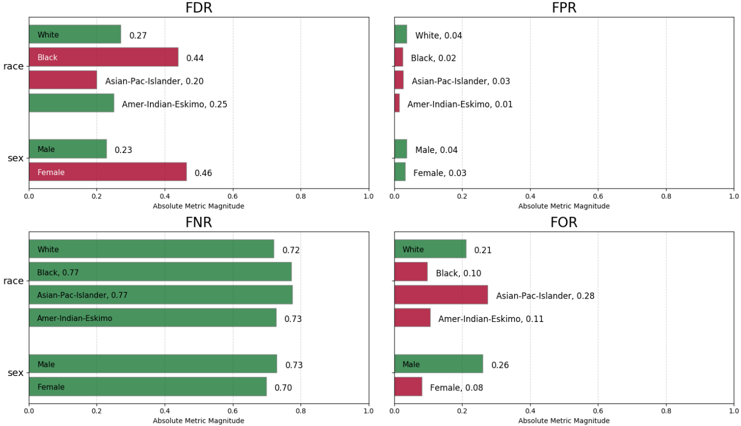
Two Numerical Examples - Adult-Income Dataset

- Underrepresented groups are biased by machine learning algorithms
- African-American, Asian-Pacific-Islander, Amer-Indian-Eskimo, Female



Two Numerical Examples - Adult-Income Dataset

- Underrepresented groups are biased by machine learning algorithms
- African-American, Asian-Pacific-Islander, Amer-Indian-Eskimo, Female



Description

Definition

A machine learning algorithm is said to be fair when predicted outcomes operating on data is *non-discriminatory* for people based on their protected status such as race, sex, etc.

- How to characterize the *fairness* (non-discrimination) of prediction?
- What *methods* are available to enforce non-discriminatory prediction?

UCLA

Description

Definition

A machine learning algorithm is said to be fair when predicted outcomes operating on data is *non-discriminatory* for people based on their protected status such as race, sex, etc.

- How to characterize the *fairness* (non-discrimination) of prediction?
- What *methods* are available to enforce non-discriminatory prediction?

UCLA

Description

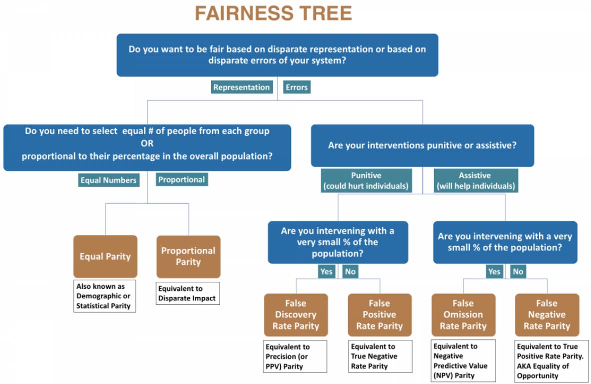
Definition

A machine learning algorithm is said to be fair when predicted outcomes operating on data is *non-discriminatory* for people based on their protected status such as race, sex, etc.

- How to characterize the *fairness* (non-discrimination) of prediction?
- What *methods* are available to enforce non-discriminatory prediction?

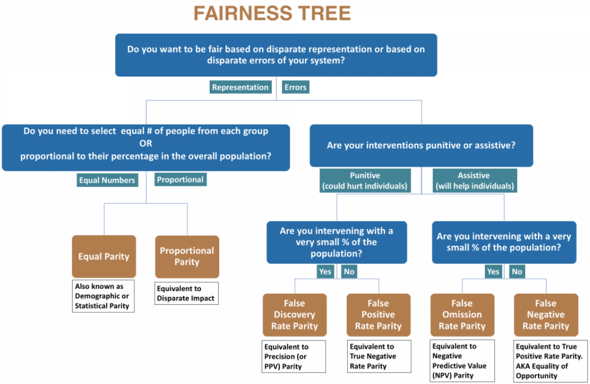
Fair Metrics Overview

- Two types of principle
  - "We Are Equal" (WAE): all groups are similar abilities with respect to the task
  - "What You See is What You Get" (WYSIWYG): observations reflect ability with respect to the task.
- Turn to fairness tree



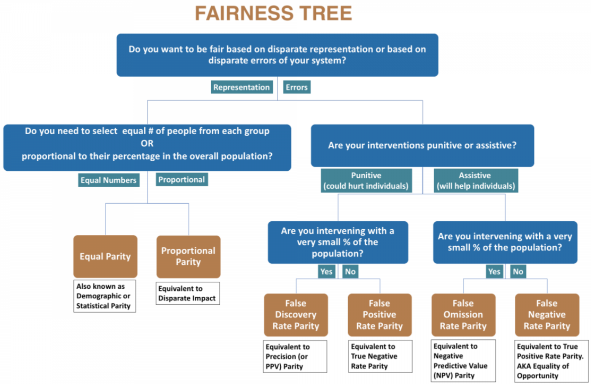
Fair Metrics Overview

- Two types of principle
  - "We Are Equal" (WAE): all groups are similar abilities with respect to the task
  - "What You See is What You Get" (WYSIWYG): observations reflect ability with respect to the task.
- Turn to fairness tree



Fair Metrics Overview

- Two types of principle
  - "We Are Equal" (WAE): all groups are similar abilities with respect to the task
  - "What You See is What You Get" (WYSIWYG): observations reflect ability with respect to the task.
- Turn to fairness tree



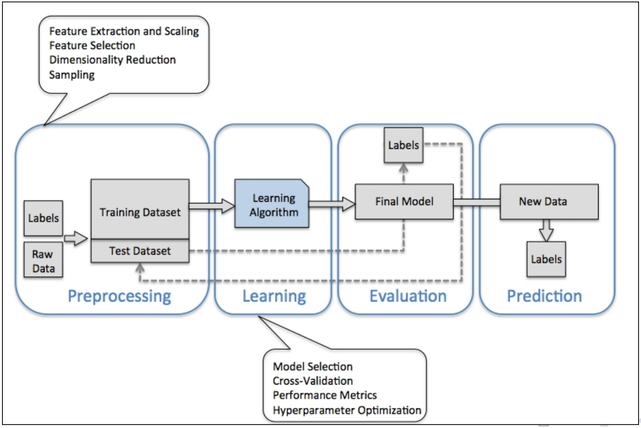
UCLA

Fairness-Preserving Algorithms Overview

Proprocessing Methods Adjust feature space

In-Processing Methods Adjust machine learning algorithms with fairness constraints

Postprocessing Methods Adjust prediction result

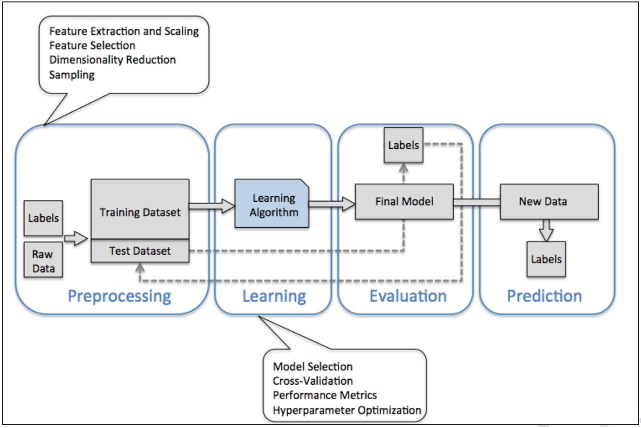


Fairness-Preserving Algorithms Overview

Proprocessing Methods Adjust feature space

In-Processing Methods Adjust machine learning algorithms with fairness constraints

Postprocessing Methods Adjust prediction result

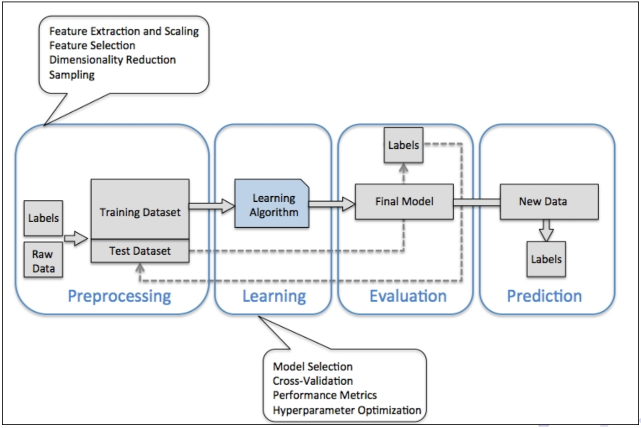


Fairness-Preserving Algorithms Overview

Proprocessing Methods Adjust feature space

In-Processing Methods Adjust machine learning algorithms with fairness constraints

Postprocessing Methods Adjust prediction result



Preprocessing Methods - Uniform Sampling

- $\forall A \in \{a_1, a_2, \cdots, a_n\}$, $y \in \{c_1, c_2, \cdots, c_m\}$, compute weight associated with each group (a_i, c_j)

$$W(a_i, c_j) = \frac{|\{x \in \mathcal{X} : A = a_i\}||\{x \in \mathcal{X} : y = c_j\}|}{|\mathcal{D}||\{x \in \mathcal{X} : A = a_i, y = c_j\}|}$$
$$= \frac{\Pr[A = a_i]\Pr[y = c_j]}{\Pr[A = a_i, y = c_j]}$$

- Uniform sampling $\mathcal{D}$ with weights $W(a_i, c_j)$

Preprocessing Methods - Uniform Sampling

- $\forall A \in \{a_1, a_2, \cdots, a_n\}$, $y \in \{c_1, c_2, \cdots, c_m\}$, compute weight associated with each group (a_i, c_j)

$$W(a_i, c_j) = \frac{|\{x \in \mathcal{X} : A = a_i\}||\{x \in \mathcal{X} : y = c_j\}|}{|\mathcal{D}||\{x \in \mathcal{X} : A = a_i, y = c_j\}|}$$
$$= \frac{\Pr[A = a_i]\Pr[y = c_j]}{\Pr[A = a_i, y = c_j]}$$

- Uniform sampling $\mathcal{D}$ with weights $W(a_i, c_j)$

In-Processing Methods - Fairness through Regularization

- Add additional fairness regularizer $R(\mathcal{D}, \theta)$ and minimize

$$-\mathcal{L}(\mathcal{D}; \theta) + \eta \mathrm{R}(\mathcal{D}, \theta) + \frac{\lambda}{2} \|\theta\|_2^2$$

- Inspired from KL divergence

$$R(\mathcal{D}, \theta) = \sum_{(\mathbf{x}_i, a_i) \in \mathcal{D}} \sum_{y \in \{0,1\}} \Pr[y | \mathbf{x}_i, a_i; \Theta] \ln \frac{\hat{\Pr}[y | a_i]}{\hat{\Pr}[y]}$$

- Minimize the difference of distribution $\Pr[y|a_i]$, $\Pr[y]$

UCLA

In-Processing Methods - Fairness through Regularization

- Add additional fairness regularizer $R(\mathcal{D}, \theta)$ and minimize

$$-\mathcal{L}(\mathcal{D}; \theta) + \eta \mathrm{R}(\mathcal{D}, \theta) + \frac{\lambda}{2}\|\theta\|_2^2$$

- Inspired from KL divergence

$$R(\mathcal{D}, \theta) = \sum_{(\mathbf{x}_i, a_i) \in \mathcal{D}} \sum_{y \in \{0,1\}} \Pr[y|\mathbf{x}_i, a_i; \Theta] \ln \frac{\hat{\Pr}[y|a_i]}{\hat{\Pr}[y]}$$

- Minimize the difference of distribution $\Pr[y|a_i]$, $\Pr[y]$

In-Processing Methods - Fairness through Regularization

- Add additional fairness regularizer $R(\mathcal{D}, \theta)$ and minimize

$$-\mathcal{L}(\mathcal{D};\theta) + \eta \mathrm{R}(\mathcal{D}, \theta) + \frac{\lambda}{2}\|\theta\|_2^2$$

- Inspired from KL divergence

$$R(\mathcal{D}, \theta) = \sum_{(\mathbf{x}_i, a_i) \in \mathcal{D}} \sum_{y \in \{0,1\}} \Pr[y|\mathbf{x}_i, a_i; \Theta] \ln \frac{\hat{\Pr}[y|a_i]}{\hat{\Pr}[y]}$$

- Minimize the difference of distribution $\Pr[y|a_i]$, $\Pr[y]$

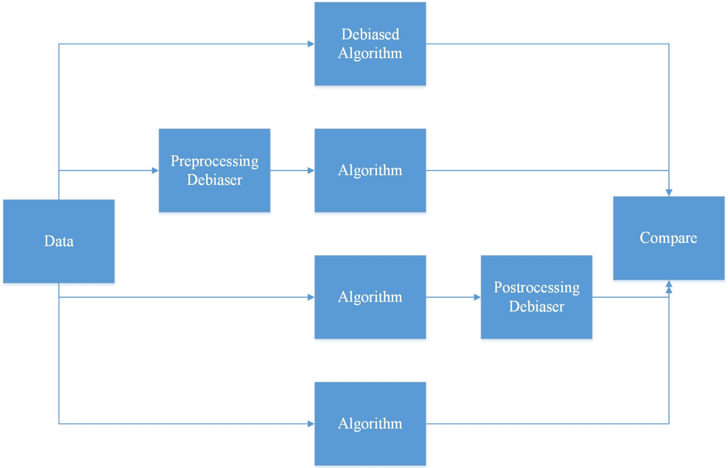
UCLA

Postprocessing Methods - Rejection Option Classification

- Adjust uncertain prediction based on group membership
- Critical region
  $\forall \mathbf{x} \in \{\mathbf{x} \in \mathcal{X} : \max\{\Pr[c^+|\mathbf{x}], 1 - \Pr[c^+|\mathbf{x}]\} < \theta\},\ 0.5 < \theta < 1$
  - If $\mathbf{x} \notin \mathcal{X}^p$, then $c_i = c^+$
  - If $\mathbf{x} \in \mathcal{X}^p$, then $c_i = c^-$
- Standard decision rule
  $\forall \mathbf{x} \in \{\mathbf{x} \in \mathcal{X} : \max\{\Pr[c^+|\mathbf{x}], 1 - \Pr[c^+|\mathbf{x}]\} \geq \theta\},\ 0.5 < \theta < 1$
  - $c_i = \arg\max_{\{c^+, c^-\}}\{\Pr[c^+|\mathbf{x}], \Pr[c^-|\mathbf{x}]\}$

UCLA

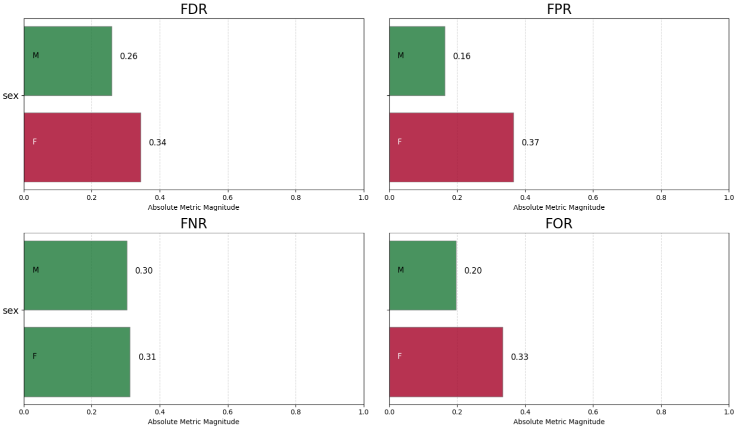
Experiment Setting



Dataset	Protected Attribute	Target
Student Performance Dataset	Sex	Grade $\geq$ 60%?
Adult Income Dataset	Race, Sex	Salary $\geq$ 50K?

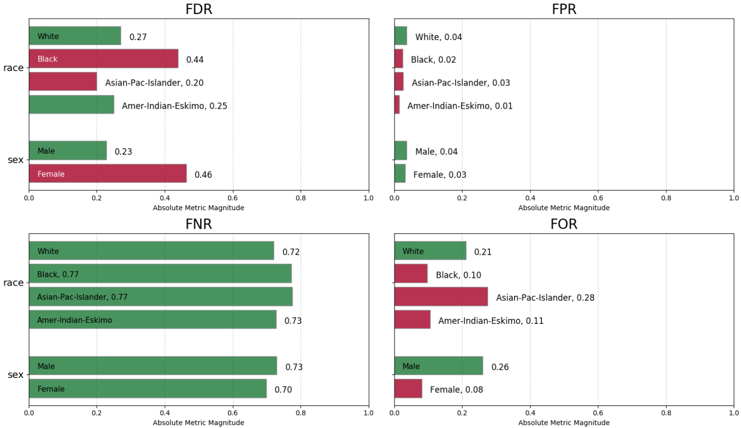
Prediction Bias Revisited

- Underrepresented groups are bias by machine learning algorithm
- African-American, Native-American, Asian, Female...

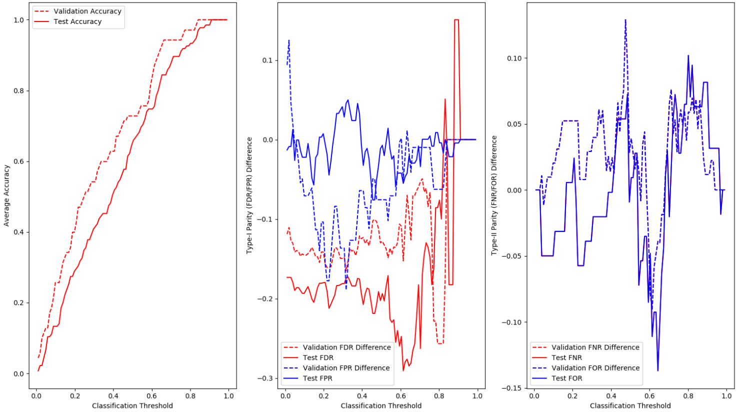


Prediction Bias Revisited

- Underrepresented groups are bias by machine learning algorithm
- African-American, Native-American, Asian, Female...

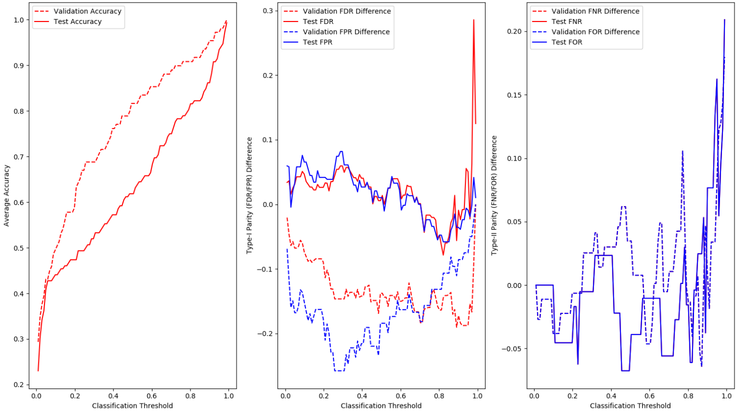


Student Performance Dataset - Preprocessing



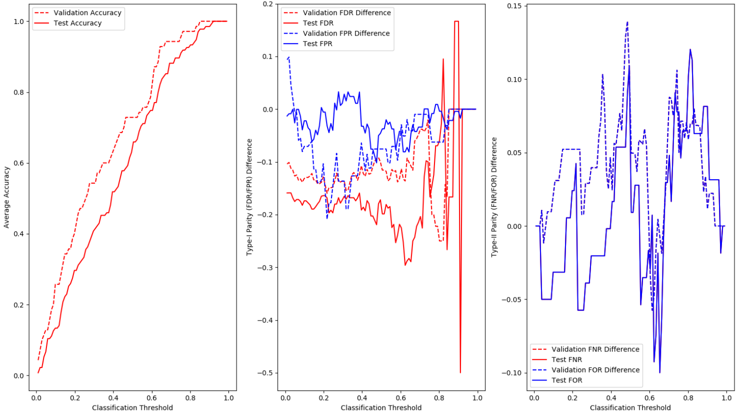
UCLA

Student Performance Dataset - In-Processing

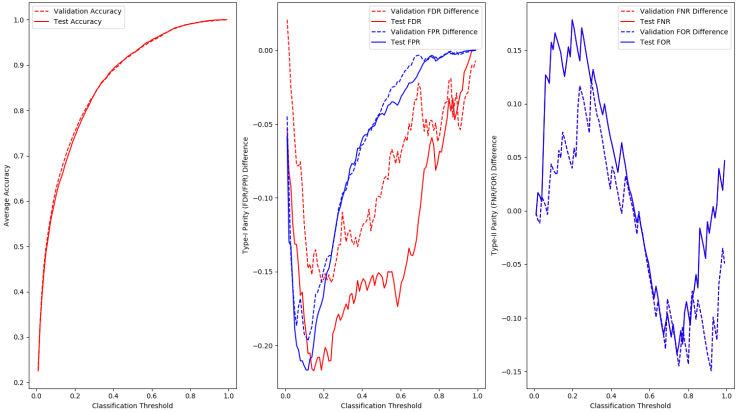


UCLA

Student Performance Dataset - Postprocessing

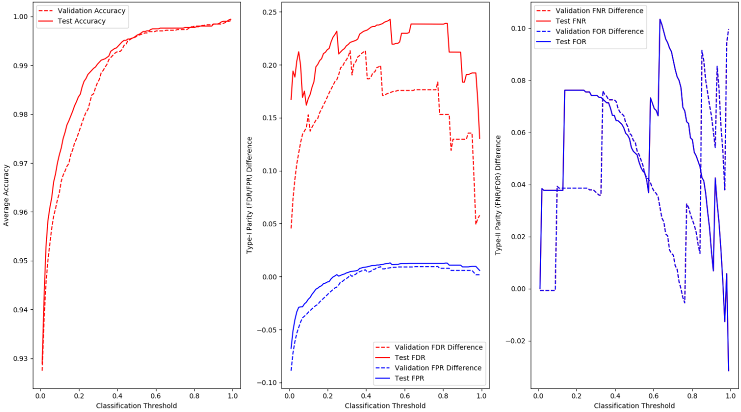


Adult-Income Dataset - Preprocessing

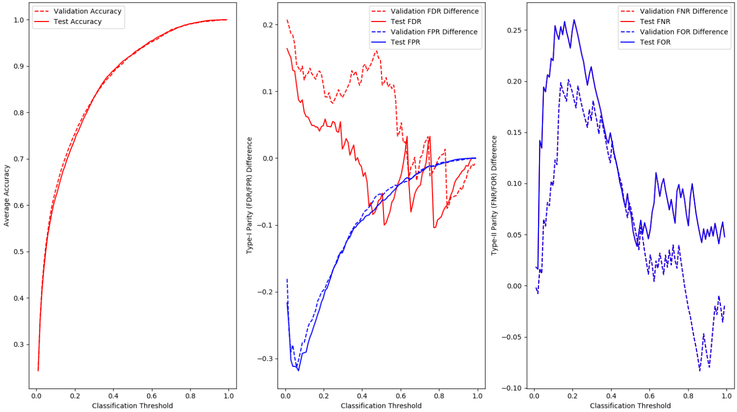


UCLA

Adult-Income Dataset - In-Processing



Adult-Income Dataset - Postprocessing



Summary and Future Work

Summary

- Review of different metrics and fairness-preserving algorithms
- Comparison of intervention methods in different phases of machine learning application

Future Work

- Fairness in deep learning and reinforcement learning
- Insight and interpretation from causal reasoning

Summary and Future Work

Summary

- Review of different metrics and fairness-preserving algorithms
- Comparison of intervention methods in different phases of machine learning application

Future Work

- Fairness in deep learning and reinforcement learning
- Insight and interpretation from causal reasoning

UCLA

Thank you for your listening!

Question and Answer