Learning Fair Representations

Published 2013 by:

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork

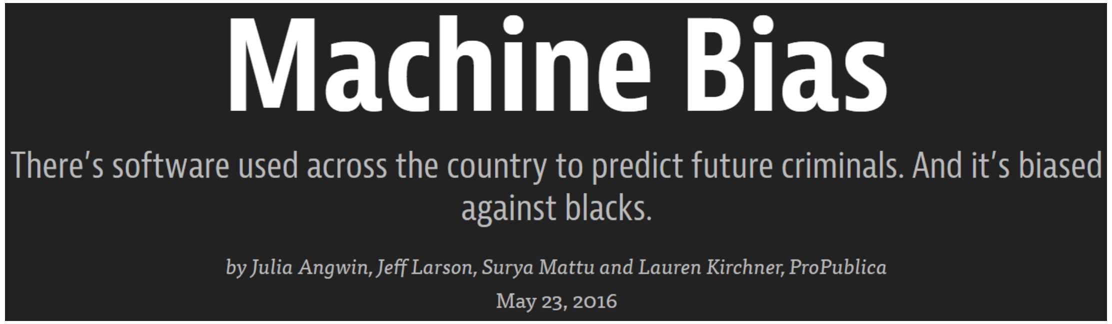
University of Toronto, Microsoft

Presented by: Robin Chan

Content

1. Motivation
2. Theoretical Background and Previous Work
3. Learning Fair Representations (LFR) Model
4. Experiments
5. Overview + Critic

Motivation



Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

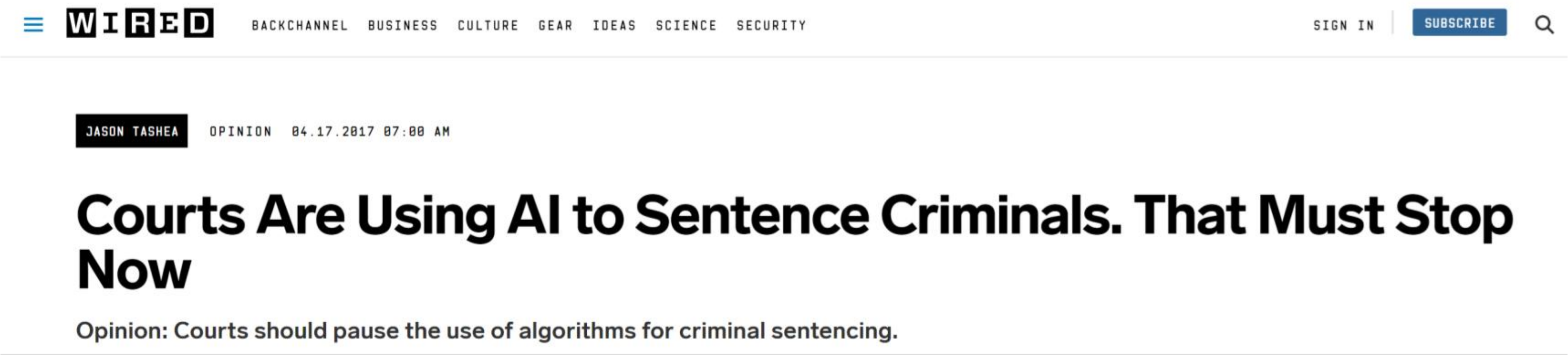
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Fairness in Machine Learning Systems

Further Examples:

- Mortgage discrimination
- Screening candidates to hire



JASON TASHEA OPINION 04.17.2017 07:00 AM

Courts Are Using AI to Sentence Criminals. That Must Stop Now

Opinion: Courts should pause the use of algorithms for criminal sentencing.

Examples of legally recognized protected groups: Race, Color, Sex, Religion, National origin, Citizenship, Age, Familial status, Disability status, …

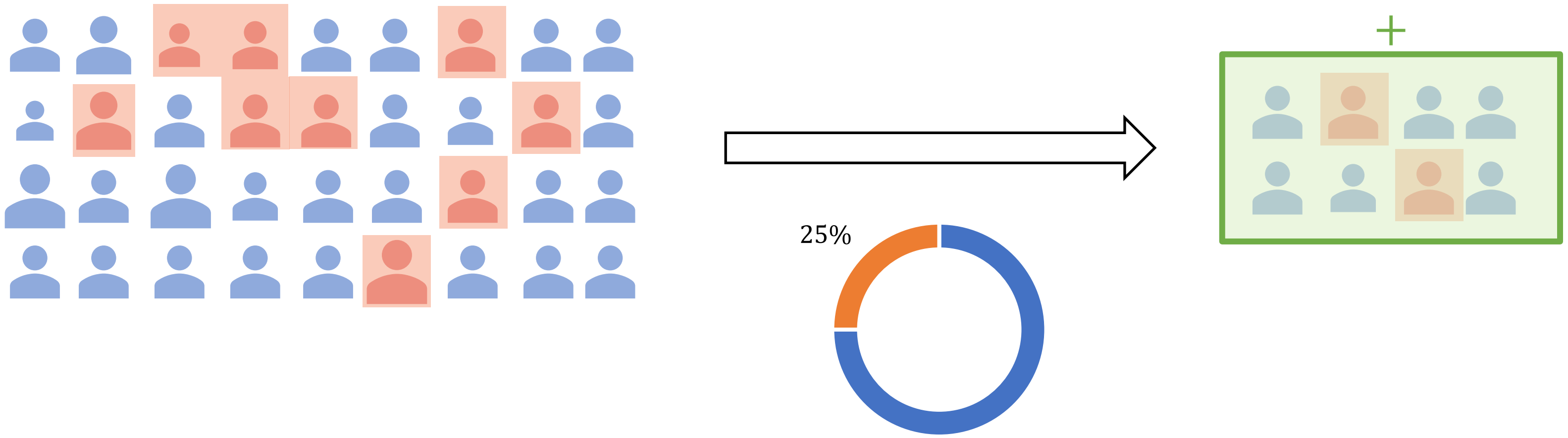
What is the actual problem?
How can we combat this?

Theoretical Background

Group Fairness and Individual Fairness

Group Fairness (= Statistical Parity)

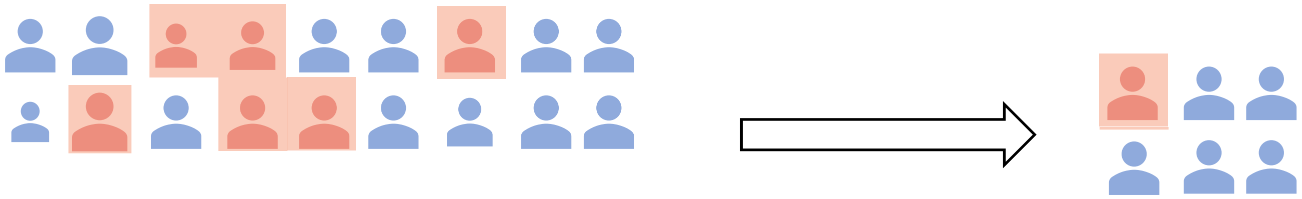
- Proportion of members of a protected group receiving positive/negative classification are identical to the proportion of the protected group in the population.



Theoretical Background

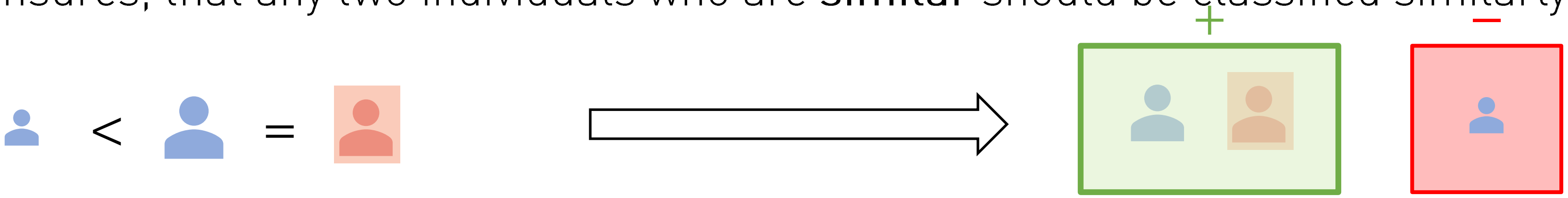
Group Fairness and **Individual Fairness**

What does group fairness miss?



Individual Fairness

- Ensures, that any two individuals who are **similar** should be classified similarly



Theoretical Background

Prior Work

Fairness Through Awareness – Dwork et al. (2011)

- Introduces the concept of a **hypothetical** measure of similarity between individuals with respect to the classification task at hand.

- Method: Define probabilistic mapping from individuals to an intermediate representation, which achieves the above goals.

Theoretical Background

Prior Work - Dwork et al. (2011) - Shortcomings

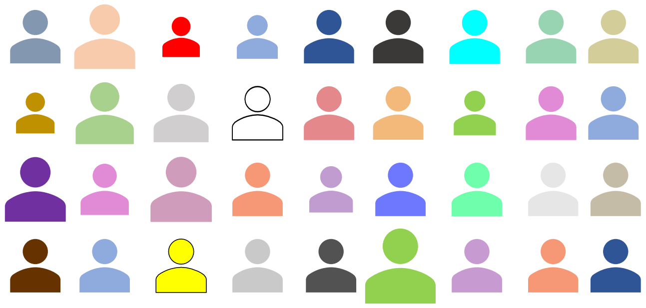
1) A fair similarity measure between individuals is assumed to be given. Finding a fair similarity measure is challenging!

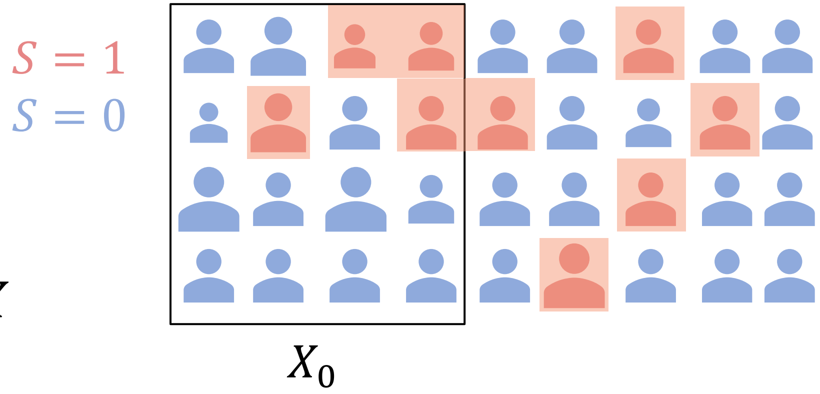
2) The mapping to intermediate representations are only defined for the given set of individuals → Lacks generalization for unseen data.

The LFR Model

LFR Model

Notation



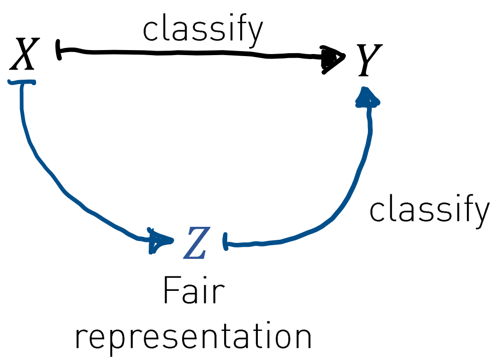


- Dataset X of individuals $x \in X \in \mathbb{R}^D$
  - Qualitative or numerical
- S: is x a member of the **protected group?**
  - Subset of individuals in the protected group: $X^+ \subset X$
  - Subset of individuals not in the protected group: $X^- \subset X$
- Training set $X_0 \subset X$
  - Subset of individuals in the protected group: $X_0^+ \subset X$
  - Subset of individuals not in the protected group: $X_0^- \subset X$
- Y is the binary random variable (classification for each individual)

LFR Model

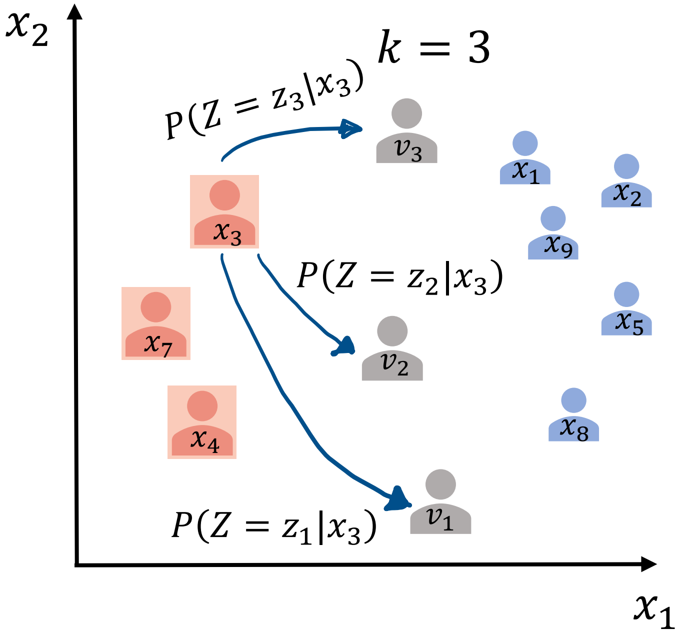
The Basic Idea

- Two-step model via a **fair**, intermediate representation
- Probabilistic mapping: $X \rightarrow Z$ to a set of prototypes.
  - Mapping should **hide the membership** of an individual in the protected group
  - Mapping should **retain** as much **information** about the individual as possible
- Mapping from prototypes to classification decision: $Z \rightarrow Y$



The LFR Model

Clustering: Prototypes



- $Z = [z_1, \dots, z_K]$: set of prototypes. "Centroid" vector $\boldsymbol{v}_k \in \mathbb{R}^D$ for each z_k

$$P(Z = k|\boldsymbol{x}_n) = \mathrm{softmax}\big(-d(\boldsymbol{x}_n, \boldsymbol{v}_k, \alpha)\big)$$

$$= \frac{\exp(-d(\boldsymbol{x}_n, \boldsymbol{v}_k, \alpha))}{\sum_{j=1}^{K} \exp\big(-d(\boldsymbol{x}_n, \boldsymbol{v}_k, \alpha)\big)}$$

LFR Model

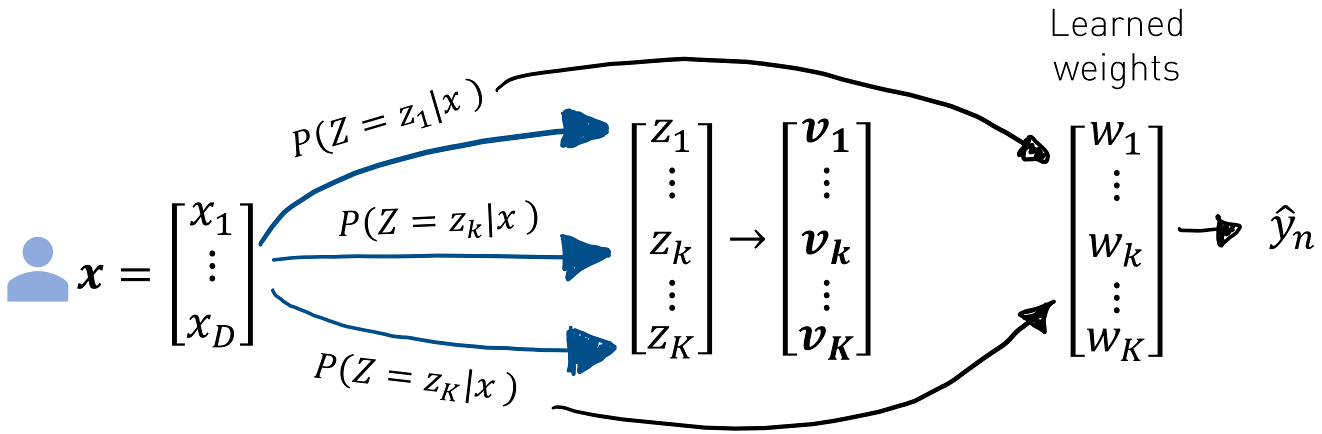
The Model Specifics

- Probabilistic mapping: $X \rightarrow Z$ (between individuals and set of prototypes)

$$P(Z = k|\boldsymbol{x_n}) = \mathrm{softmax}\big(-d(\boldsymbol{x_n}, \boldsymbol{v_k}, \alpha)\big) = \frac{\exp(-d(\boldsymbol{x_n}, \boldsymbol{v_k}, \alpha))}{\sum_{j=1}^{K} \exp\big(-d(\boldsymbol{x_n}, \boldsymbol{v_k}, \alpha)\big)}$$

- Mapping from prototypes to classification decision: $Z \rightarrow Y$

$$\hat{y}_n = \sum_{k=1}^{K} P(Z = k|x_n) \cdot w_k$$



LFR Model

Objectives

1) We want an accurate prediction $\hat{y}_n$

2) Intermediate representation should be accurate: $\hat{\boldsymbol{x}}_n = \sum_{k=1}^{K} P(Z = k|\boldsymbol{x}_n) \cdot \boldsymbol{v}_k$

3) Obfuscate membership in protected group: $P(Z = k|\boldsymbol{x} \in X^+) = P(Z = k \mid \boldsymbol{x} \in X^-)$

Estimated on the training data:

$$\mathbb{E}_{\boldsymbol{x} \in X^+} P(Z = k|\, \boldsymbol{x}) = \mathbb{E}_{\boldsymbol{x} \in X^-} P(Z = k|\, \boldsymbol{x}) \leftrightarrow M_k^+ = M_k^-, \qquad \forall k$$

Optimize: $\alpha_i, \{\boldsymbol{v}_k\}_{k=1}^{K}, \boldsymbol{w}$

Objective Function:

$$L = A_z \cdot \sum_{k=1}^{K} \left| M_k^+ - M_k^- \right| + A_x \cdot \sum_{n=1}^{N} (\boldsymbol{x}_n - \hat{\boldsymbol{x}}_n)^2 + A_y \cdot \sum_{n=1}^{N} -y_n \log \hat{y}_n - (1 - y_n) \log(1 - \hat{y}_n)$$

Experiments

Experiments
Evaluation

How do we quantify the quality of a <u>fair model</u>?

- Accuracy: $1 - \frac{1}{N}\sum_{n=1}^{N}|y_n - \hat{y}_n|$
- Discrimination: $\left|\frac{\sum_{n:s_n=1}\hat{y}_n}{\sum_{n:s_n=1}1} - \frac{\sum_{n:s_n=0}\hat{y}_n}{\sum_{n:s_n=0}1}\right|$
- Consistency: comparison to $kNN(\boldsymbol{x})$: $1 - \frac{1}{Nk}\sum_n\left|\hat{y}_n - \sum_{j\in kNN(\boldsymbol{x}_n)}\hat{y}_j\right|$

Model Selection?

- Min discrimination
- Max. Delta (between accuracy and discrimination)

Experiments
Compared Models and Datasets

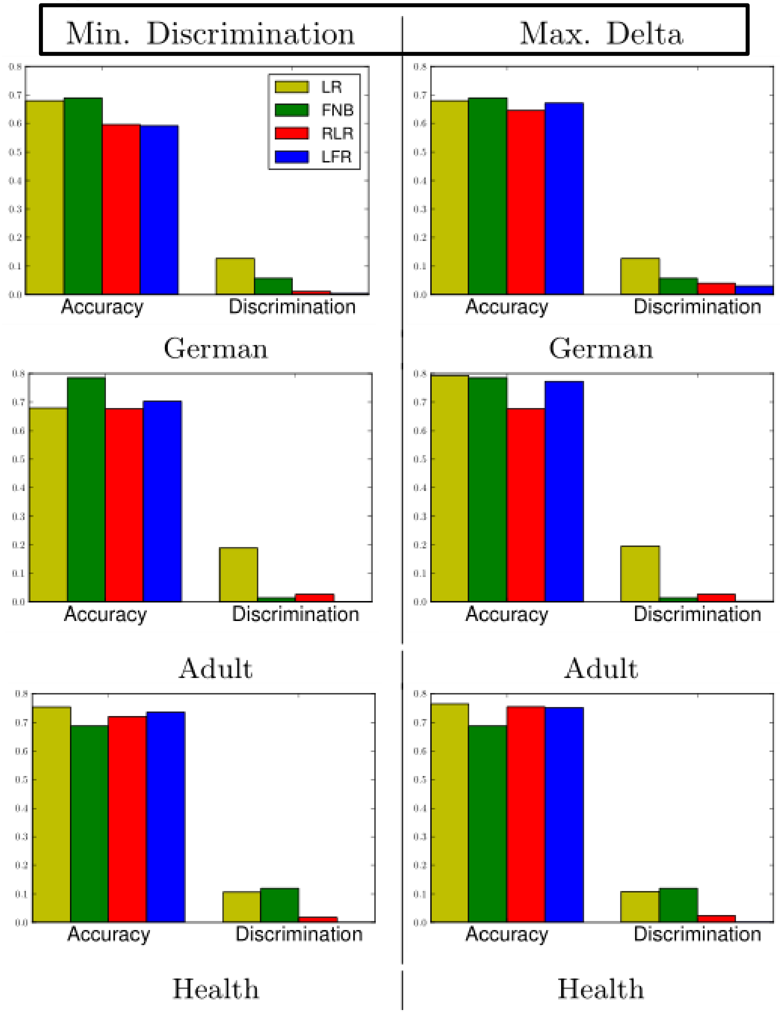
- German credit dataset (1000 samples)
  - Each individual is represented by **20 attributes**
  - Classify bank account holders into a “Good” or “Bad”
  - Considered attribute = **Age**
- Adult income dataset (45’222 samples)
  - Each individual is represented by **14 attributes**
  - Classify whether the income is larger than 50’000 dollars
  - Considered attribute = **Gender**
- Health Heritage Dataset (147’473 samples)
  - Each individual is represented by **139 attributes**
  - Classify whether a person will be in the hospital in a particular year
  - Considered attribute = Age

Models:
- LR = Logistic Regression
- FNB = Fair Naïve Bayes
- RLR = Regularized Logistic Regression
- LFR = Learned Fair Representation

Experiments

Results + Discussion: Accuracy and Discrimination

Model selection criteria



Models (Legend):

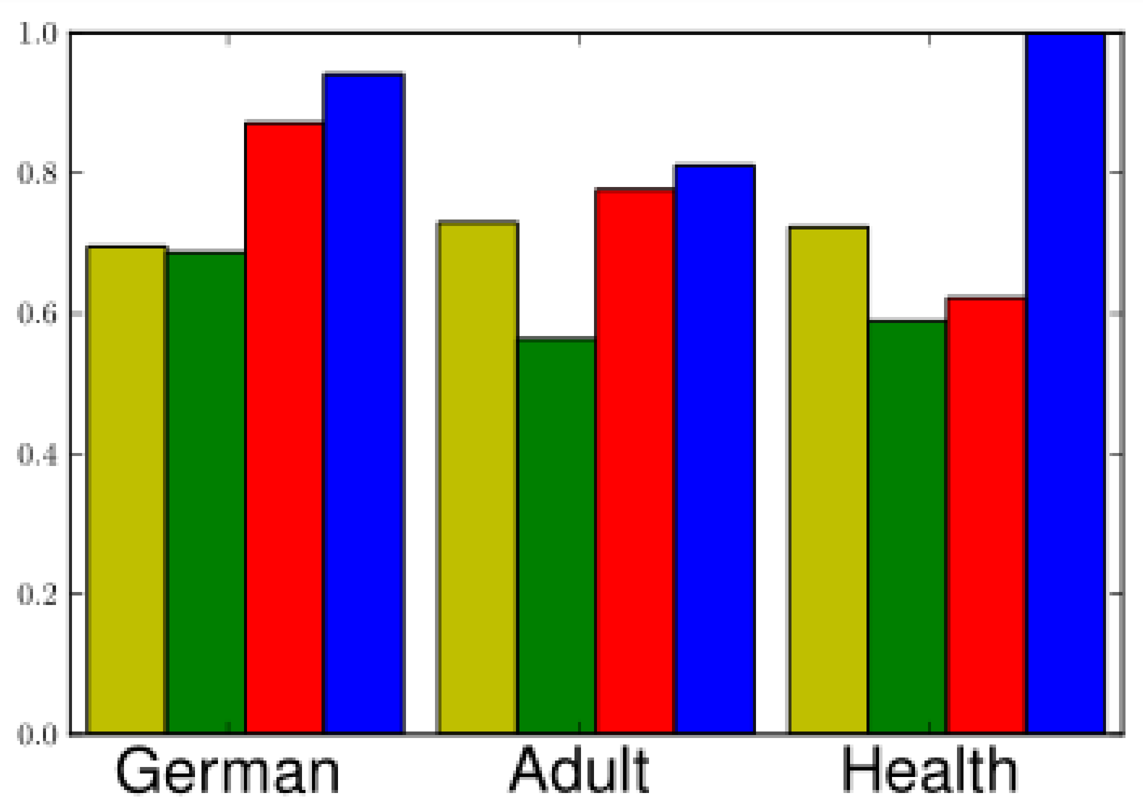
- LR = Logistic Regression
- FNB = Fair Naïve Bayes
- RLR = Regularized Logistic Regression
- LFR = Learned Fair Representation

Experiments

Results + Discussion 2: Consistency (Measure of individual fairness)

Models (Legend):

- LR = Logistic Regression
- FNB = Fair Naïve Bayes
- RLR = Regularized Logistic Regression
- LFR = Learned Fair Representation

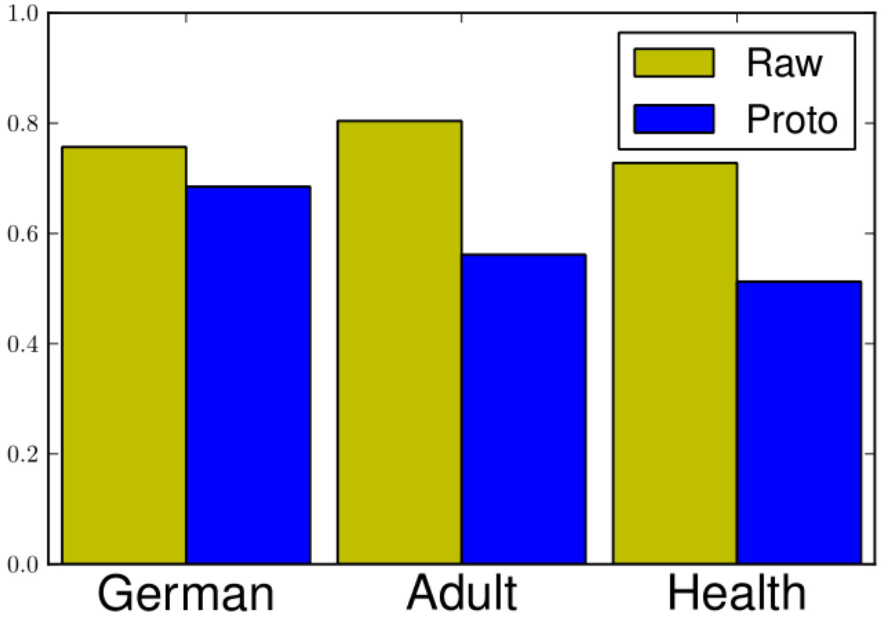


Experiments

Results + Discussion 3: How well is information about $\boldsymbol{S}$ obfuscated?

Create a predictor which learns $\boldsymbol{S}$ from Z: $\hat{s}_n = \sum_{k=1}^{K} P(Z = k|\boldsymbol{x}_n)\, u_k$

What are desirable values? → lower bound: 0.5



Big Picture Contributions

Technical Contributions

- Framework achieves both group and individual fairness
- Learning framework: learn the weights of the distance function, as well as a fair intermediate representation with good properties.
- Mapping $X \rightarrow Z$ can be generalized to samples not in the training set!

Critic of the Paper

+ very simple, intuitively understandable framework
+ novel formulation of fairness as optimization

- Achieving individual fairness is either way very hard to achieve. Is individual fairness truly achieved here?
- LFR is currently only considered for binary classification.

End.