MIT-IBM Watson AI Lab

Industry Showcase

AI Fairness Workshop

May 9, 2022
AI Fairness

Mikhail Yurochkin and Onkar Bhardwaj
Roadmap

- AI is prone to biases
- Definitions of algorithmic fairness
- Practical fairness methods
  - Identifying fairness violations
  - Training fair models
  - Post-processing for fairness
Fairness: A case study

Example: Sentiment analysis – classify words as positive or negative

Positive: admirable, adorable, joy, lucky, talented, … ✔

Negative: aggressive, distrust, nasty, radical, … ❌

Deep Learning + Word Embeddings -> 95% test accuracy.

Success?
Deployment Concerns

What is a sentiment of a name?

Common European-American names: 
Adam, Ryan, Paul, ... , Courtney, Meredith, Megan, ...

Common African-American names: 
Alonzo, Leroy, Tyree, ... , Shereen, Sharise, Tawanda, ...

Names are from “Semantics derived automatically from language corpora contain human-like biases” (Caliskan et al., 2017)
Fairness Violations

Protected Attribute
- Race: Caucasian
- Race: African-American

Baseline

Sentiment score

-20
-10
0
10
20
Does Group Fairness help?
Fairness is Violated for Individuals

Protected Attribute
- Race: Caucasian
- Race: African-American

Baseline
Group Fairness

Ryan
Tyree

sentiment score

Parameter not selected

Individual Fairness

Protected Attribute
- Race: Caucasian
- Race: African-American

Baseline
Group Fairness
Yurochkin et al., ICLR 2020
Accuracy is Preserved

Protected Attribute
- Race: Caucasian
- Race: African-American

sentiment score

Baseline

Group Fairness

Yurochkin et al., ICLR 2020

95%

94%
Roadmap

AI is prone to biases

Definitions of algorithmic fairness

Practical fairness methods
- Identifying fairness violations
- Training fair models
- Post-processing for fairness
Group Fairness

Algorithm performs similarly on groups of individuals

\[ Y \rightarrow \text{true label} \]
\[ A \rightarrow \text{protected attribute} \]
\[ \hat{Y} \rightarrow \text{prediction} \]

Demographic Parity: \[ \hat{Y} \perp\!
\perp A \]

Equalized Odds: \[ \hat{Y} \perp\!
\perp A \mid Y \]
Evaluating Group Fairness

Demographic Parity: \( \hat{Y} \perp\!\!\!\!\!\!\!\!\perp A \)

Compare average outcome for men and women

Test data: \((x_1, a_1), \ldots, (x_N, a_N)\);
model to audit \( h : \mathcal{X} \rightarrow \mathcal{Y} \)

Output DP = \[
\left| \frac{\sum_i \mathcal{I}(a_i=\text{male}, h(x_i)=1)}{\sum_i \mathcal{I}(a_i=\text{male})} - \frac{\sum_i \mathcal{I}(a_i=\text{female}, h(x_i)=1)}{\sum_i \mathcal{I}(a_i=\text{female})} \right|
\]
Evaluating Group Fairness

Equalized Odds: \( \hat{Y} \perp A \mid Y \)

Compare class accuracies for men and women

Test data: \((x_1, y_1, a_1), \ldots, (x_N, y_n, a_N)\); model to audit \( h : \mathcal{X} \to \mathcal{Y} \)

Measure \( \text{EO}_0 \) = \[
\frac{\sum_i I(a_i=\text{male}, y_i=0, h(x_i)=0)}{\sum_i I(a_i=\text{male}, y_i=0)} - \frac{\sum_i I(a_i=\text{female}, y_i=0, h(x_i)=0)}{\sum_i I(a_i=\text{female}, y_i=0)}
\]

Measure \( \text{EO}_1 \) = \[
\frac{\sum_i I(a_i=\text{male}, y_i=1, h(x_i)=1)}{\sum_i I(a_i=\text{male}, y_i=1)} - \frac{\sum_i I(a_i=\text{female}, y_i=1, h(x_i)=1)}{\sum_i I(a_i=\text{female}, y_i=1)}
\]

Output \( \text{EO} = \frac{1}{2}(\text{EO}_0 + \text{EO}_1) \)
What Group Fairness definition did we check?
Individual Fairness
(Dwork et al. 2012)

Algorithm treats similar individuals similarly

\[ d_Y(h(x_1), h(x_2)) \preceq d_X(x_1, x_2) \text{ for all } x_1, x_2 \in \mathcal{X} \]

- ML model is a map \( h : \mathcal{X} \to \mathcal{Y} \)
- \( d_Y \) measures similarity between outputs
- Fair metric \( d_X \) measures similarity between inputs
Evaluating Individual Fairness

Prediction Consistency

*Compare predictions on similar inputs*

Occupation prediction from a person’s biography:

*He graduated from law school with honors*  ➔  Attorney

*She graduated from law school with honors*  ➔  ???Paralegal???

\[
\text{Output PC} = \frac{\sum_i \mathbb{I}(h(x_i[\text{he}])=h(x_i[\text{she}]))}{N}
\]
Questions?
Is “blindness” a solution?

The @AppleCard is such an sexist program. My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time. Yet Apple’s black box algorithm thinks I deserve 20x the credit limit she does. No appeals work.
I noticed that DistilBERT loves movies filmed in India, but not in Iraq, so I plotted the result for each country: the resulting map is scary. #aibias
from transformers import pipeline

classifier = pipeline("sentiment-analysis")
classifier(["The movie was filmed in India.",
            "The movie was filmed in Iraq."])
Roadmap

AI is prone to biases
Definitions of algorithmic fairness

Practical fairness methods
  • Identifying fairness violations
  • Training fair models
  • Post-processing for fairness
Identifying Fairness violations

Group Fairness: measure DP (EO) on audit data

Test data: \((x_1, a_1), \ldots, (x_N, a_N)\);
model to audit \(h : \mathcal{X} \rightarrow \mathcal{Y}\)

Output DP = \[ \left| \frac{\sum_i \mathcal{I}(a_i=\text{male}, h(x_i)=1)}{\sum_i \mathcal{I}(a_i=\text{male})} - \frac{\sum_i \mathcal{I}(a_i=\text{female}, h(x_i)=1)}{\sum_i \mathcal{I}(a_i=\text{female})} \right| \]

Four-Fifths Rule, US Equal Employment Opportunity Commission:
“selection rate for any race, sex, or ethnic group [must be at least] four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate”
Identifying Fairness violations

Individual Fairness: Prediction Consistency

Occupation prediction from a person’s biography:

*He graduated from law school with honors* → Attorney

*She graduated from law school with honors* → ???Paralegal???

Output PC = \[ \frac{\sum_i \mathbb{I}(h(x_i[\text{he}])=h(x_i[\text{she}]就来看看))}{N} \]
Individual Fairness in Social Science


• The investigators responded to job ads in Boston and Chicago newspapers with fictitious resumes.
• They randomly assigned African-American or white sounding names to the resumes.
• The investigators concluded there is discrimination against African-Americans: the resumes assigned white names received 50% more callbacks for interviews.
Demonstration
Distributional Individual Fairness (DIF)

Find individual fairness violations algorithmically

\[ \text{DIF}(h) \triangleq \left\{ \sup_{T: \mathcal{X} \to \mathcal{X}} \mathbb{E}_{P_X} \left[ d_Y(h(x), h(T(x))) \right] \quad \text{subject to} \quad \mathbb{E}_{P_X} \left[ d_X(x, T(x)) \right] \leq \epsilon. \right\} \]

- **Auditor** \( T \) is a map that finds fairness violations
- \( d_Y \) measures similarity between outputs
- **Fair metric** \( d_X \) measures similarity between inputs

Yurochkin & Sun, ICLR 2021
Auditing for IF violations

Test data: \((x_1, y_1), \ldots, (x_N, y_N)\);
DIF map \(T(x)\) for the model \(h\) that we are auditing;
some loss function \(\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+\)

Hypothesis (\(h\) is individually fair)
\(H_0\) : loss ratio on similar individuals is at most \(\delta\)

Compute loss ratios \(R = \left\{ \frac{\ell(h(T(x_i),y_i))}{\ell(h(x_i),y_i)} \right\}_{i=1}^N\)

Reject \(H_0\) with confidence \((1 - \alpha)\) if \(\text{Mean}(R) - \frac{z_{1-\alpha}}{\sqrt{N}} \text{Var}(R) > \delta\)
Demonstration
Learning fair metrics from data

- Samples with protected attributes
- Groups of comparable samples
- Pairs of comparable and incomparable samples

\[ d_{\mathcal{X}}(x_1, x_2) = (x_1 - x_2)^\top \Sigma (x_1 - x_2) \]

Mukherjee et al., ICML 2020
Learning fair metrics from data

Samples with protected attributes:
gender/race information in the Adult dataset

Learn “sensitive” directions with Logistic Regression, i.e. $V = \{v_{\text{gender}}, v_{\text{race}}\}$.

Ignore them in the fair metric: $\Sigma = I - P_{\text{span}(V)}$.

$$d_X(x_1, x_2) = (x_1 - x_2)^\top \Sigma (x_1 - x_2)$$

Yurochkin et al., ICLR 2020
Group of comparable samples:
word embeddings of popular baby names

Find directions of major variation with PCA, i.e. \( V = \{v_1, \ldots, v_K\} \).

Ignore them in the fair metric: \( \Sigma = I - P_{\text{span}(V)} \).

\[
d_X(x_1, x_2) = (x_1 - x_2)^\top \Sigma (x_1 - x_2)
\]
Questions?
Roadmap

AI is prone to biases

Definitions of algorithmic fairness

Practical fairness methods

• Identifying fairness violations
• Training fair models
• Post-processing for fairness
A variant of adversarial training: Train model accurate on the available data and data similar in the fair metric

• Observe data
• Audit model with DIF: Find similar data where algorithm performs differently
• Update model parameters to minimize prediction error and DIF
• Repeat
Sensitive Set Invariance (SenSeI)

$$\min_{h \in \mathcal{H}} L(h) + \rho \text{DIF}(h)$$

$$L(h) \triangleq \mathbb{E}[\ell(y, h(x))]$$

- $\mathcal{H}$: model class (e.g. neural nets with a certain architecture)
- $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is a loss function
- $\rho$: regularization parameter
Relation to Adversarial Robustness

Adversarial training: Train model accurate on the available data and visually similar data. Different “fair” metric.

Image is from “A Brief Introduction to Adversarial Examples” (Màdry & Schmidt, 2018)
Demonstration
Training Group Fair models

Optimization with (data-dependent) constraints: Train model accurate on the available data subject to group fairness constraints

\[
\min_{h \in \mathcal{H}} L(h)
\]
subject to \( DP < \delta \), where

\[
DP = \left| \frac{\sum_i \mathbb{I}(a_i = \text{male}, h(x_i) = 1)}{\sum_i \mathbb{I}(a_i = \text{male})} - \frac{\sum_i \mathbb{I}(a_i = \text{female}, h(x_i) = 1)}{\sum_i \mathbb{I}(a_i = \text{female})} \right|
\]
What is Your type of Fairness?

**Group Fairness:**
- Carefully choose GF notion appropriate for the application
- Many open-source solutions (AIF360, Fairlearn, TFCO)
- Check individual fairness!

**Individual Fairness:**
- Carefully choose data for learning the fair metric
- inFairness package is soon to be open-source
- Check group fairness!
Questions?
Roadmap

- AI is prone to biases
- Definitions of algorithmic fairness

Practical fairness methods
- Identifying fairness violations
- Training fair models
- Post-processing for fairness
Post-processing for Individual Fairness

Measuring IF on a graph:

\[
\sum_{i,j} W_{ij} (f_i - f_j)^2 = 2f^\top \mathbb{L} f
\]

Closed-form solution:

\[
f = (I + \lambda \mathbb{L})^{-1} \hat{y}
\]

Original predictions

Graph Laplacian

Penalize individual fairness violations

Stay close to original predictions

Original predictions

Post-processed fair predictions

Petersen et al., NeurIPS 2021
I noticed that DistilBERT loves movies filmed in India, but not in Iraq, so I plotted the result for each country: the resulting map is scary. #aibias
from transformers import pipeline

classifier = pipeline("sentiment-analysis")
classifier([
    "The movie was filmed in India.",
    "The movie was filmed in Iraq."])

[
    {'label': 'POSITIVE', 'score': 0.9783285856246948},
    {'label': 'NEGATIVE', 'score': 0.9872057437896729}]
Demonstration
Post-processing for Group Fairness

Optimized Score Transformation for Consistent Fair Classification

*Wei et al., 2021*

FairScoreTransformer (FST): Available in AIF360
Algorithmic Fairness pipeline

1. Choose IF fair metric / GF notion
2. Audit trained ML model for fairness violations
3. Post-process trained model to improve fairness
4. Train Fair model
Questions?
Yuekai  11:44 PM

LOL I’m getting depressed

we write all these papers and all we keep hearing about is the f**ckups
We ask Your input!

Let us know your thoughts in a follow up survey.
Group Fairness References


• D. Wei, K. Ramamurthy, F. Calmon. Optimized Score Transformation for Consistent Fair Classification. JMLR 2021.

• R. Bellamy et al. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias.


• AIF360: https://github.com/Trusted-AI/AIF360

• Fairlearn: https://github.com/fairlearn/fairlearn

• TFCO: https://github.com/google-research/tensorflow_constrained_optimization
Individual Fairness References


Blog-posts and Media

AI fairness

In today’s data-driven world, machine learning (ML) systems are increasingly used to make high-stakes decisions in domains like criminal justice, education, lending, and medicine. For example, a judge may use an algorithm to assess a defendant’s chance of re-offending before deciding to detain or release the defendant. Although replacing humans with ML systems appears to eliminate human biases in the decision-making process, they can perpetuate or even exacerbate biases in the training data. Such biases are especially objectionable when it adversely affects underprivileged groups of users.

The most obvious remedy is to remove the biases in the training data, but carefully curating the datasets that modern ML systems are trained on is impractical. This leads to the challenge of developing ML systems that remain “fair” despite biases in the training data.

But what is fair?

There are two major families of definitions of fairness: (1) group fairness; (2) individual fairness. Group fairness requires certain constraints to be satisfied at the population level, e.g. proportion of hired job applicants should be similar across different demographic groups. Individual fairness (also known as Lipschitz fairness) states that hiring decisions for any pair of similar applicants (e.g. equally qualified applicants with different names) should be the same.

New research helps make AI fairer in decision-making

Our team developed the first practical procedures and tools for achieving Individual Fairness in machine learning (ML) and artificial intelligence (AI) systems.
Onkar Bhardwaj  Mayank Agarwal  Aldo Pareja

inFairness team
Collaborators

University of Michigan: Yuekai Sun, Amanda Bower, Songkai Xue, Debarghya Mukherjee, Moulinath Banerjee, Alexander Vargo, Fan Zhang, Subha Maity, Hamid Eftekharri

IBM Research: Mark Weber, Ben Hoover, Mayank Agarwal, Aldo Pareja, Onkar Bhardwaj, Uri Kartoun, Bum Chul Kwon, Kenney Ng, Zahra Ashktorab

University of Konstanz: Felix Petersen

Wells Fargo: Sherif Botros, Vanio Markov
Thank You!

Paper links, videos, news, and code are on my website
moonfolk.github.io