Counterfactual Fairness

Matt J. Kusner*,1,2, Joshua R. Loftus*,1,3, Chris Russell*,1,4, Ricardo Silva1,5

*Equal Contribution 1The Alan Turing Institute 2University of Warwick 3New York University 4University of Surrey 5University College London

Causal Inference

[Pearl et al., 2016]

Given an individual:

\[ A, G, L, Y \]

1. Change race attribute

\[ \hat{A} = a' \]

2. Compute unobserved variables in causal model

\[ \hat{A} = a'' \]

3. Recompute observed variables in causal model

\[ \hat{A} = a''' \]

\[ \hat{Y} = \hat{f}(\hat{U}, \ldots) \]

\[ \hat{Y} \sim \hat{f}(\hat{U}, \ldots) \]

Our Solution

A fair classifier gives the same prediction had the person had a different race/sex.

Prior Fairness Definitions Are Insufficient

Fairness through Unawareness

Equality of Opportunity

Unfair Influences

Teachers may believe students have behavior issues

Class placement may be different

Lack of minority rate teachers affects minority student outcomes

We present a metric to check if any algorithm which is fair

And an algorithm to learn fair classifiers

Results: US law schools

causal models

\[ G \sim \text{Normal}(b_G + [A, A, |w_G, a_0]) \]

\[ L \sim \text{Poisson}(\exp(b_L + [A, A, |w_L]) \]

\[ Y \sim \text{Normal}(U, A, A, |w_Y, 1) \]

\[ U \sim \text{Normal}(0,1) \]

\[ G = b_G + [A, A, |w_G + U_G \]

\[ L = b_L + [A, A, |w_L + U_L \]

\[ Y = b_Y + [A, A, |w_Y + U_Y \]

\[ U_G \sim \text{P}(U_G) \]

\[ U_L \sim \text{P}(U_L) \]

\[ U_Y \sim \text{P}(U_Y) \]

Compares the same individual with a different version of low/school

Results: NYC stop-and-frisk

Arrest rate (data)

Counterfactuals

Unaware

C-Fair (Non-Det.)

C-Fair (Det.)

RMSE 0.873 0.894 0.929 0.918

Related Work

Compares different individuals with the same observed features

\[ \hat{P}(\hat{Y} = y | do(A = a), X = x) = \hat{P}(\hat{Y} = y | do(A = a'), X = x) \]

\[ \hat{P}(\hat{Y} = y | do(A = a'), X = x) \]

Feature Importance

\[ \text{35% of black students} \]

\[ \text{50% of white students} \]

[Boyd, 2017]

Law School Admissions

Data

Goal

Full

Unaware

C-Fair

C-Fair

male white

1.0

1.0

0.8

0.6

male black

0.0

1.0

0.4

0.2

female white

0.0

0.5

0.5

0.3

female black

0.5

0.0

0.4

0.2

sensitive attributes

features \( \mathcal{X} \)

label \( Y \)

GPA

race

college GPA

SAT score

L SAT score

law knowledge

1st year low grade

3. Recompute observed variables in causal model

\[ \hat{A} = a''' \]

\[ \hat{Y} = \hat{f}(\hat{U}, \ldots) \]

\[ \hat{Y} \sim \hat{f}(\hat{U}, \ldots) \]