Fair Causal Reasoning

May 22, 2020
Dr. Wei Wei, Prof. James Landay

CS 335: Fair, Accountable, and Transparent (FAccT) Deep Learning
Stanford University
Recap

- Counterfactual Explanations

Grath et al, 2018
Recap

- Counterfactual Fairness

\[ P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a) \]
Recap

- Counterfactual Face Attribution

\[ P(\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A\leftarrow a'}(U) = y \mid X = x, A = a) \]
Recap

\[ S_f(d) = \mathbb{E}_{z \sim p(z)}[f(G(z + d)) - f(G(z))] \]
Outline

● Fair Causal Reasoning
● Counterfactual Fairness
  ○ Formal Methods
  ○ Law School
  ○ Crime Rates in NYC
● Equalized Counterfactual Odds
● Multiple Causal Worlds
Fair Causal Reasoning

Causal Graph
- Observed Data
- Latent Data
- Relations

Causal Fairness Criteria
- Counterfactual Fairness

Latent
● Observed

Observed Data

Observational Fairness Criteria
- Fairness Through Unawareness
- Demographic Parity
- Equalized Odds/Opp
Causal Graph

\[ z \sim N(\mu, \Sigma) \]
Why Do We Need Causal Fairness?

- Recover Latent Variables

Kusner et al, 2018
Why Do We Need Causal Fairness?

- Recover Latent Variables

Kusner et al, 2018
Why Do We Need Causal Fairness?

- Dealing with Inherent bias

Inherent Biases

Sampling Biases
Inherent bias

- Race groups live in certain regions due to socioeconomic status
- Latent Socioeconomic factors
  - More police resources in regions with low economic status
  - Results in more arrests

Kusner et al., 2018
Inherent bias

- Observational Fairness Criteria Won't work
  - Dataset (observed variables) contains inherent selection biases
  - Concentration of police resources resulted in high arrests
  - Attributing regions (and eventual race) unfairly to arrests in the dataset

\[
P(\hat{Y} = 1|A = 1) = P(\hat{Y} = 1|A = 0)
\]
Inherent bias

- Causal Fairness
  - Intervene variables in a causal graph
  - Generating samples with races that live in neighborhoods that have high police resources

\[
P(\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A\leftarrow a'}(U) = y \mid X = x, A = a)
\]
Intervention on Causal Graphs

Causal Graph with A, Z, Y

Intervene on A

Intervene on Y

Loftus et al, 2018
Outline

● Fair Causal Reasoning
● Counterfactual Fairness
  ○ Formal Methods
  ○ Law School
  ○ Crime Rates in NYC
● Equalized Counterfactual Odds
● Multiple Causal Worlds
Counterfactual Fairness Revisited

\[ P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a) \]

Real Examples

Intervention on A\leftarrow a

Counterfactual Examples

Intervention on A\leftarrow a'
Counterfactual Fairness

- Level 1
  - Build predictors using only the observable non-descendants of A

Fairness Through Unawareness
Counterfactual Fairness

- Level 2
  - Build Predictors using the parents of the observable variables
Counterfactual Fairness

- Level 3
  - Build Predictors by adding independent error terms
Outline

- Fair Causal Reasoning
- Counterfactual Fairness
  - Formal Methods
  - Law School
  - Crime Rates in NYC
- Equalized Counterfactual Odds
- Multiple Causal Worlds
Law School Success Dataset

- Conducted by Law School Admission Council in US
  - 21,790 law students
  - Entrance exam scores (LSAT)
  - Grade-point average (GPA) collected prior to law school
  - Prediction $Y =$ first year average grade (FYA)
  - Protected features = \{Gender, Race\}
Level 2 Counterfactual Fairness

- Build Predictors using the parents of the observable variables

\[
K \sim \mathcal{N}(0, 1)
\]

\[
\begin{align*}
FYA & \sim \mathcal{N}(w^K_F K + w^R_FR + w^S_FS, 1) \\
GPA & \sim \mathcal{N}(b_G + w^K_G K + w^R_GR + w^S_GS, \sigma_G) \\
LSAT & \sim \text{Poisson} \left( \exp(b_L + w^K_L K + w^R_LR + w^S_LS) \right)
\end{align*}
\]

Gaussian Dist.  
Parameters

Kusner et al., 2018
Level 3 Counterfactual Fairness

- Build Predictors by adding independent error terms

\[
\begin{align*}
\text{GPA} &= b_G + w^R_G R + w^S_G S + \epsilon_G, \quad \epsilon_G \sim p(\epsilon_G) \\
\text{LSAT} &= b_L + w^R_L R + w^S_L S + \epsilon_L, \quad \epsilon_L \sim p(\epsilon_L) \\
\text{FYA} &= b_F + w^R_F R + w^S_F S + \epsilon_F, \quad \epsilon_F \sim p(\epsilon_F)
\end{align*}
\]

Kusner et al, 2018
Baselines

full - using all features
unaware - fairness through unawareness

Kusner et al, 2018
## Results

<table>
<thead>
<tr>
<th></th>
<th>Full</th>
<th>Unaware</th>
<th>Fair $K$</th>
<th>Fair Add</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>0.873</td>
<td>0.894</td>
<td>0.929</td>
<td>0.918</td>
</tr>
</tbody>
</table>

Kusner et al, 2018
Outline

● Fair Causal Reasoning
● Counterfactual Fairness
  ○ Formal Methods
  ○ Law School
  ○ Crime Rates in NYC
● Equalized Counterfactual Odds
● Multiple Causal Worlds
Causal Graph

- Assess the fairness of the NYC arrest dataset
  - 38,609 records
  - White individuals (4492)
  - Black Hispanic individuals (2414)

Kusner et al, 2018
Assessment Results

<table>
<thead>
<tr>
<th>Race (data)</th>
<th>Arrested (data)</th>
<th>Arrests decreases from 5659 to 3722</th>
<th>Arrests increases from 5659 to 6439</th>
</tr>
</thead>
<tbody>
<tr>
<td>White (4492)</td>
<td>White (12.1%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black Hispanic (2414)</td>
<td>Black Hispanic (19.8%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Arrests decreases from 5659 to 3722

Arrest if White (counterfactual)

Kusner et al, 2018
Outline

● Fair Causal Reasoning
● Counterfactual Fairness
  ○ Formal Methods
  ○ Law School
  ○ Crime Rates in NYC
● Equalized Counterfactual Odds
● Multiple Causal Worlds
Equalized Counterfactual Odds

Equality of Odds

$$P(\hat{Y} = 1|A = 0, Y) = P(\hat{Y} = 1|A = 1, Y)$$

Counterfactual Fairness

$$P(\hat{Y}_{A\leftarrow a}(U) = y | X = x, A = a) = P(\hat{Y}_{A\leftarrow a'}(U) = y | X = x, A = a)$$

Equalized Counterfactual Odds

$$p(\hat{Y}_{A\leftarrow a}(U) | X = x, Y_{A\leftarrow a} = y, A = a) = p(\hat{Y}_{A\leftarrow a'}(U) | X = x, Y_{A\leftarrow a'} = y, A = a)$$
Healthcare Equality

- Protected Features $A = \{\text{Gender}\}$
- Features $X$, vector representation of coded diagnoses, procedures, medication orders, lab results, and clinical notes
- Prediction $Y$, a binary indicator of the occurrence of a clinically relevant outcome

\[
\begin{align*}
    u &\sim p(U) = \text{Normal}(0, I) \\
    a &\sim p(A) = \text{Categorical}(A \mid \pi) \\
    x, y &\sim p(X, Y \mid U, A) = p(X \mid U, A)p(Y \mid U, A)
\end{align*}
\]
Training Objective

● $\sigma$ - sigmoid function
● $h$ - predictor
● $J$ - cross entropy loss

$$
\mathcal{L} = J(h_\theta(x, a), y) + \lambda_{\text{CF}} \sum_{a_k \in A} \mathbb{1}[a \neq a_k] J(h_\theta(x_{A \leftarrow a_k}, a_k), y_{A \leftarrow a_k}) +
\lambda_{\text{CLP}} \sum_{a_k \in A} \mathbb{1}[a \neq a_k] \mathbb{1}[y = y_{A \leftarrow a_k}] \left( \sigma^{-1}(h_\theta(x_{A \leftarrow a_k}, a_k)) - \sigma^{-1}(h_\theta(x, a)) \right)^2
$$

logits
## Dataset Overview

<table>
<thead>
<tr>
<th>Group</th>
<th>Count</th>
<th>Length of Stay $\geq$ 7 Days</th>
<th>Inpatient Mortality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asian</td>
<td>17,465</td>
<td>0.187</td>
<td>0.025</td>
</tr>
<tr>
<td>Black</td>
<td>5,202</td>
<td>0.239</td>
<td>0.020</td>
</tr>
<tr>
<td>Hispanic</td>
<td>21,978</td>
<td>0.196</td>
<td>0.019</td>
</tr>
<tr>
<td>Other</td>
<td>11,004</td>
<td>0.200</td>
<td>0.022</td>
</tr>
<tr>
<td>Unknown</td>
<td>3,593</td>
<td>0.201</td>
<td>0.072</td>
</tr>
<tr>
<td>White</td>
<td>70,391</td>
<td>0.204</td>
<td>0.021</td>
</tr>
<tr>
<td>Female</td>
<td>72,556</td>
<td>0.167</td>
<td>0.018</td>
</tr>
<tr>
<td>Male</td>
<td>57,076</td>
<td>0.245</td>
<td>0.029</td>
</tr>
<tr>
<td>[18, 30)</td>
<td>15,291</td>
<td>0.180</td>
<td>0.007</td>
</tr>
<tr>
<td>[30, 45)</td>
<td>27,155</td>
<td>0.140</td>
<td>0.007</td>
</tr>
<tr>
<td>[45, 65)</td>
<td>43,529</td>
<td>0.222</td>
<td>0.025</td>
</tr>
<tr>
<td>[65, 89)</td>
<td>43,658</td>
<td>0.226</td>
<td>0.036</td>
</tr>
<tr>
<td>All</td>
<td>129,633</td>
<td>0.201</td>
<td>0.023</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Group</th>
<th>Metric</th>
<th>N/A</th>
<th>0.0</th>
<th>0.01</th>
<th>0.1</th>
<th>1.0</th>
<th>10.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asian</td>
<td>AUC-PRC</td>
<td>0.605</td>
<td>0.563</td>
<td>0.555</td>
<td>0.561</td>
<td>0.56</td>
<td>0.562</td>
</tr>
<tr>
<td></td>
<td>AUC-ROC</td>
<td>0.86</td>
<td>0.853</td>
<td>0.853</td>
<td>0.854</td>
<td>0.849</td>
<td>0.851</td>
</tr>
<tr>
<td></td>
<td>Brier</td>
<td>0.106</td>
<td>0.11</td>
<td>0.109</td>
<td>0.109</td>
<td>0.11</td>
<td>0.112</td>
</tr>
<tr>
<td>Black</td>
<td>AUC-PRC</td>
<td>0.579</td>
<td>0.548</td>
<td>0.55</td>
<td>0.545</td>
<td>0.563</td>
<td>0.573</td>
</tr>
<tr>
<td></td>
<td>AUC-ROC</td>
<td>0.838</td>
<td>0.825</td>
<td>0.82</td>
<td>0.825</td>
<td>0.823</td>
<td>0.823</td>
</tr>
<tr>
<td></td>
<td>Brier</td>
<td>0.124</td>
<td>0.135</td>
<td>0.129</td>
<td>0.128</td>
<td>0.127</td>
<td>0.129</td>
</tr>
<tr>
<td>Hispanic</td>
<td>AUC-PRC</td>
<td>0.592</td>
<td>0.558</td>
<td>0.565</td>
<td>0.57</td>
<td>0.564</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>AUC-ROC</td>
<td>0.862</td>
<td>0.855</td>
<td>0.856</td>
<td>0.861</td>
<td>0.853</td>
<td>0.854</td>
</tr>
<tr>
<td></td>
<td>Brier</td>
<td>0.113</td>
<td>0.117</td>
<td>0.115</td>
<td>0.114</td>
<td>0.117</td>
<td>0.118</td>
</tr>
<tr>
<td>Other</td>
<td>AUC-PRC</td>
<td>0.549</td>
<td>0.557</td>
<td>0.557</td>
<td>0.563</td>
<td>0.553</td>
<td>0.561</td>
</tr>
<tr>
<td></td>
<td>AUC-ROC</td>
<td>0.824</td>
<td>0.827</td>
<td>0.819</td>
<td>0.824</td>
<td>0.819</td>
<td>0.827</td>
</tr>
<tr>
<td></td>
<td>Brier</td>
<td>0.122</td>
<td>0.124</td>
<td>0.121</td>
<td>0.121</td>
<td>0.122</td>
<td>0.124</td>
</tr>
<tr>
<td>Unknown</td>
<td>AUC-PRC</td>
<td>0.675</td>
<td>0.616</td>
<td>0.616</td>
<td>0.606</td>
<td>0.614</td>
<td>0.633</td>
</tr>
<tr>
<td></td>
<td>AUC-ROC</td>
<td>0.9</td>
<td>0.891</td>
<td>0.888</td>
<td>0.893</td>
<td>0.891</td>
<td>0.887</td>
</tr>
<tr>
<td></td>
<td>Brier</td>
<td>0.104</td>
<td>0.106</td>
<td>0.103</td>
<td>0.103</td>
<td>0.105</td>
<td>0.111</td>
</tr>
<tr>
<td>White</td>
<td>AUC-PRC</td>
<td>0.575</td>
<td>0.568</td>
<td>0.564</td>
<td>0.559</td>
<td>0.562</td>
<td>0.563</td>
</tr>
<tr>
<td></td>
<td>AUC-ROC</td>
<td>0.847</td>
<td>0.84</td>
<td>0.839</td>
<td>0.838</td>
<td>0.838</td>
<td>0.837</td>
</tr>
<tr>
<td></td>
<td>Brier</td>
<td>0.118</td>
<td>0.12</td>
<td>0.118</td>
<td>0.12</td>
<td>0.12</td>
<td>0.121</td>
</tr>
</tbody>
</table>
Results

- Difference in the counterfactual versus factual predicted probability
Outline

● Fair Causal Reasoning
● Counterfactual Fairness
  ○ Formal Methods
  ○ Law School
  ○ Crime Rates in NYC
● Equalized Counterfactual Odds
● Multiple Causal Worlds
Multiple Causal Graphs

- Whether a student can graduate on time

Russell et al, 2017
Alternative Definitions of Counterfactual Fairness

**Exact Formulation**

\[
P(\hat{Y}_{A\leftarrow a} (U) = y \mid X = x, A = a) = P(\hat{Y}_{A\leftarrow a'} (U) = y \mid X = x, A = a)
\]

**\(\epsilon\) - Approximate Formulation**

\[
\left| f(x_{A\leftarrow a}, a) - f(x_{A\leftarrow a'}, a') \right| \leq \epsilon
\]

**(\(\delta, \epsilon\)) - Approximate Formulation**

\[
\mathbb{P}(\left| f(x_{A\leftarrow a}, a) - f(x_{A\leftarrow a'}, a') \right| \leq \epsilon \mid \mathcal{X} = x, A = a) \geq 1 - \delta
\]
Multi-world Counterfactual Fairness

\[
\min_{f} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i, a_i), y_i) + \lambda \sum_{j=1}^{m} \frac{1}{n} \sum_{i=1}^{n} \sum_{a' \neq a_i} \mu_j(f, x_i, a_i, a') \\
\text{loss of the data} \quad \text{world } j \quad \text{counterfactual examples}
\]

\[
\mu_j(f, x_i, a_i, a') := \frac{1}{S} \sum_{s=1}^{S} \max\{0, |f(x_i^{s}, A_j \leftarrow a_i, a_i) - f(x_i^{s}, A_j \leftarrow a', a')| - \epsilon\}
\]

Monte-carlo Samples \quad \epsilon - \text{Approximate Counterfactual Fairness}
Law Graduate School

L3 Method

\[ G = b_G + w_G^A A + \epsilon_G \]
\[ L = b_L + w_L^A A + \epsilon_L \]
\[ Y = b_Y + w_Y^A A + \epsilon_Y \]
\[ \epsilon_G, \epsilon_L, \epsilon_Y \sim \mathcal{N}(0, 1) \]

L2 Method

\[ G \sim \mathcal{N}(b_G + w_G^A A + w_G^U U, \sigma_G) \]
\[ L \sim \text{Poisson}(\exp(b_L + w_L^A A + w_L^U U)) \]
\[ Y \sim \mathcal{N}(w_Y^A A + w_Y^U U, 1) \]
\[ U \sim \mathcal{N}(0, 1) \]

Russell et al, 2017
Results

\[ \left| f(x_{A \leftarrow a}, a) - f(x_{A \leftarrow a'}, a') \right| \leq \epsilon \]

Russell et al., 2017
COMPAS

\[ T \sim \text{Bernoulli}(\phi(b_T + w_C^{UD} U_D + w_C^E E + w_C^A A)) \]
\[ C \sim \mathcal{N}(b_C + w_C^{UD} U_D + w_C^E E + w_C^A A + w_C^T T + w_C^P P + w_C^{JF} J_F + w_C^{JM} J_M, \sigma_C) \]
\[ P \sim \text{Poisson}(\exp(b_P + w_P^{UD} U_D + w_P^E E + w_P^A A)) \]
\[ J_F \sim \text{Poisson}(\exp(b_{J_F} + w_{J_F}^{UD} U_D + w_{J_F}^E E + w_{J_F}^A A)) \]
\[ J_M \sim \text{Poisson}(\exp(b_{J_M} + w_{J_M}^{UD} U_D + w_{J_M}^E E + w_{J_M}^A A)) \]
\[ [U_J, U_D] \sim \mathcal{N}(0, \Sigma) \]

Russell et al., 2017
Results

Russell et al, 2017
Reading Assignments

- Chiappa, Silvia. Path-specific counterfactual fairness, AAAI 2019
- Balcan, Maria-Florina F., Travis Dick, Ritesh Noothigattu, and Ariel D. Procaccia. Envy-free classification, NeurIPS 2019
- Zhang, Junzhe, and Elias Bareinboim. Equality of opportunity in classification: A causal approach, NeurIPS 2018