CS 229br Lecture 6: Causality, Fairness, Privacy
Boaz Barak

Yamini Bansal
Official TF

Javin Pombra
Official TF

Dimitris Kalimeris
Unofficial TF

Gal Kaplun
Unofficial TF

Preetum Nakkiran
Unofficial TF
Outline

• Part I: Causality
• Part II: Fairness
Causality

PATTERNS, PREDICTIONS, AND ACTIONS
A story about machine learning

Moritz Hardt and Benjamin Recht
Causality

Correlation ≠ Causation

But what is causation?

“Smoking causes cancer”

“Obesity causes heart disease”
Causality theory

Understand the conditions under which *correlation* = *causation*

Setup:

Observables: $A, B, C, D, ...$

Interventions: “do $A \leftarrow a$”

Correlation: $\Pr[ B = b \mid A = a ]$

Causation: $\Pr[ B = b \mid \text{do } A \leftarrow a ]$
Correlation: \( \Pr[B = b \mid A = a] \)
Causation: \( \Pr[B = b \mid \text{do } A \leftarrow a] \)

Scenario 1: \( X \leftarrow B(1/2) \)

\[
W \leftarrow \begin{cases} 
0, & X = 1 \\
B(1/2), & X = 0 
\end{cases} \\
H \leftarrow \begin{cases} 
0, & X = 1 \\
B(1/2), & X = 0 
\end{cases}
\]

<table>
<thead>
<tr>
<th>( X )</th>
<th>( W )</th>
<th>( H )</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1/2</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/8</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1/8</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1/8</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1/8</td>
</tr>
</tbody>
</table>

Scenario 2: \( W \leftarrow B(1/4) \)

\[
X \leftarrow \begin{cases} 
0, & W = 1 \\
B(1/3), & W = 0 
\end{cases} \\
H \leftarrow \begin{cases} 
0, & X = 1 \\
B(1/2), & X = 0 
\end{cases}
\]

<table>
<thead>
<tr>
<th>( X )</th>
<th>( W )</th>
<th>( H )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( \Pr[W = 1 \mid X = 0] )</th>
<th>Scenario 1</th>
<th>Scenario 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/2</td>
<td>1/2</td>
<td></td>
</tr>
<tr>
<td>( \Pr[W = 1 \mid \text{do } X \leftarrow 0] )</td>
<td>1/2</td>
<td>1/4</td>
</tr>
</tbody>
</table>
Correlation: \( \Pr[B = b \mid A = a] \)
Causation: \( \Pr[B = b \mid \text{do } A \leftarrow a] \)

**Scenario 1:** \( X \leftarrow B(1/2) \)
- \( W \leftarrow \begin{cases} 0, & X = 1 \\ B(1/2), & X = 0 \end{cases} \)
- \( H \leftarrow \begin{cases} 0, & X = 1 \\ B(1/2), & X = 0 \end{cases} \)

**Scenario 2:**
- \( W \leftarrow B(1/4) \)
- \( X \leftarrow \begin{cases} 0, & W = 1 \\ B(1/3), & W = 0 \end{cases} \)
- \( H \leftarrow \begin{cases} 0, & X = 1 \\ B(1/2), & X = 0 \end{cases} \)

Cannot distinguish Scenario 1 and 2 from observations alone!
Estimating causal probabilities

Assume: Know causal graph

Goal: Compute $\Pr[A = a \mid \text{do } B \leftarrow b]$

$\Pr[H = 1 \mid W = 0] = 1/6$

$\Pr[H = 1 \mid \text{do } W \leftarrow 0] = 1/4$

Controlling for $X$:

$\Pr[H = 1 \mid \text{do } W \leftarrow 0] = \Pr[H = 1 \mid W = 0, X = 0] \Pr[X = 0]$

Apriori unknown

$\Pr[H = 1 \mid W = 0, X = 0] \Pr[X = 0] + \Pr[H = 1 \mid W = 0, X = 1] \Pr[X = 1]$
Adjustment formula

\[
\Pr[ Y = y \mid \text{do } X \leftarrow x ] = \sum \Pr[ Y = y \mid X = x, Z = z ] \cdot \Pr[ Z = z ]
\]

Apriori unknown

Known* from observations
Control for wrong things

\[ \text{Pr}[X = 1|Y = 1] = \text{Pr}[X = 1 \mid \text{do } Y \leftarrow 1] = p \]

Controlling for \( Z \):

\[
\text{Pr}[X = 1|Y = 1, Z = 1] \cdot \text{Pr}[Z = 1] + \text{Pr}[X = 1|Y = 1, Z = 0] \cdot \text{Pr}[Z = 0] \approx p^2
\]

\[
\approx \frac{p^2}{2p} = \frac{p}{2} \approx 2p = 0
\]
\[
Pr[Y = y | do X \leftarrow x ] \quad \text{vs} \quad Pr[ Y = y | X = x ]
\]

\[
Pr[Y = y | do X \leftarrow x ] \quad \text{vs} \quad \sum Pr[ Y = y | X = x, Z] Pr[Z]
\]
Casual Models

"Frequentist":
\[ \text{Pr}[A \mid \text{do } B] \] is frequency of times that \( A \) occurs if we do \( B \)

"Bayesian":
\[ \text{Pr}[A \mid \text{do } B] \] is probability \( A \) would have happened in "counter-factual" world where we did \( B \)

Exogenous randomness

\[ U_1 \]

\[ X_1 = f_1(U_1) \]

\[ U_2 \rightarrow X_2 = f_2(X_1; U_2) \]

\[ U_3 \rightarrow X_3 = f_3(X_1; U_3) \]

Time
Def: $X, Y$ are confounded if

Thm: If $X, Y$ not confounded then $\Pr[Y = y \mid \text{do } X \leftarrow x] = \Pr[Y = y \mid X = x]$

Proof:
Conditioning
Conditioning

\[ Z = z \]
Conditioning

\[ Z = z \]

Introduce spurious correlations here

Kill these paths
Average Treatment Effect

\( T \in \{0,1\} \) – Treatment variable

Goal: Estimate \( \mathbb{E}[Y_1] - \mathbb{E}[Y_0] \)

Def: \( T, Y \) “ignorable” controlling for \( Z \) if:

\[ T \perp (Y_0, Y_1) \mid Z \]

i.e: choice of \( T = 0,1 \) independent of \( Y \mid \text{do } T \leftarrow t \)
Average Treatment Effect

\( T \in \{0,1\} \) – Treatment variable

Goal: Estimate \( \mathbb{E}[Y_1] - \mathbb{E}[Y_0] \)

Def: \( T, Y \) “ignorable” controlling for \( Z \) if:

\( T \perp (Y_0, Y_1) \mid Z \)

i.e: choice of \( T = 0,1 \) independent of \( Y | \)do \( T \leftarrow t \)

Claim: If \( T, Y \) ignorable controlling for \( Z \) then

\[
\Pr[Y = y \mid \text{do } T \leftarrow t] = \sum \Pr[Y = y \mid T = t, Z = z] \Pr[Z = z]
\]

Pf:

\[
\sum \Pr[Y = y \mid T = 0, Z = z] \Pr[Z = z] = \sum \Pr[Y_0 = y \mid Z = z] \Pr[Z = z]
\]
Propensity scores:

Let $e(z) = \mathbb{E}[T|Z = z]$.

CLAIM: If $Z$ admissible, $\mathbb{E}[Y \mid \text{do } T \leftarrow 1] = \mathbb{E}\left[\frac{Y \cdot T}{e(Z)}\right]$.

Pf: $\Pr[Y = y \mid \text{do } T \leftarrow 1] = \sum_z \Pr[Y = y \mid T = 1, z] \Pr[z]$

$$= \sum_z Pr[z] \frac{Pr[Y=y, T=1 | z]}{Pr[T=1 | z]} = \mathbb{E}_z \left[\frac{Pr[Y=y, T=1 | z]}{e(Z)}\right] = \mathbb{E}_z \left[\frac{Pr[YT=y | z]}{e(Z)}\right]$$

$\mathbb{E} [Y \mid \text{do } T \leftarrow 1] = \sum_y \Pr[Y = y \mid \text{do } T \leftarrow 1] \cdot y$

$$= \sum_y \mathbb{E}_z \left[\frac{Pr[YT = y | z]}{e(Z)} y\right] = \mathbb{E}_z \left[\frac{Y \cdot T}{e(Z)}\right]$$
Double ML

Let $e(z) = \mathbb{E}[T|Z = z]$.

Assume $Y = \psi(Z) + \tau \cdot T + \text{Noise}$.

$	au$ = treatment effect

Observe $(Z, T, Y)$, learn model $f(z) \approx \mathbb{E}[Y|Z = z]$.

\[
f(z) \approx \psi(Z) + \tau \cdot e(z)
\]

\[
\Rightarrow Y - f(z) \approx \tau \cdot (T - e(z))
\]

Can estimate from data.
**Instrumental variables**

$W$ is unobserved: can’t control for

Assume $Y = \tau \cdot T + f(W)$  \hspace{1cm} Cov$(Z, f(W)) = 0$

$\tau = \text{treatment effect}$

$\Rightarrow \quad \tau = \frac{\text{Cov}(Z,Y)}{\text{Cov}(Z,T)}$
Counterfactuals

Let $u$ realization of $U_1 \ldots U_n$

$Y_{X \leftarrow x}(u)$ = output of $Y$ if $U = u$ and $X = x$
Fairness

Fairness and machine learning

Limitations and Opportunities

Solon Barocas, Moritz Hardt, Arvind Narayanan

NIPS 2017 Tutorial on Fairness in Machine Learning

Solon Barocas, Moritz Hardt

Note: Focus on fairness in classification, not representation
Google Algorithm Detects Lung Cancer Better Than Human Doctors

By Stephanie Mlot 05.21.2019 8:15 AM

By Gary N

Can an Algorithm Write a Better News Story Than a Human Reporter?

Are Self-Driving Cars on the Road to OVERTAKING TRADITIONAL VEHICLES?

By Nikolas Perrault

Can an Algorithm Hire Better Than a Human?

By Claire Cain Miller @clairecm JUNE 25, 2015

Hiring and recruiting might seem like some of the least likely jobs to be automated. The whole process seems to need human skills that computers just can't perform as well as humans.
Risk of Recidivism

<table>
<thead>
<tr>
<th></th>
<th>WHITE</th>
<th>AFRICAN AMERICAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labeled Higher Risk, But Didn’t Re-Offend</td>
<td>23.5%</td>
<td>44.9%</td>
</tr>
<tr>
<td>Labeled Lower Risk, Yet Did Re-Offend</td>
<td>47.7%</td>
<td>28.0%</td>
</tr>
</tbody>
</table>

Angwin, Larson, Mattu, Kirchner 2016
Gender detection

99.7% correct

65.3% correct

Buolamwini, Gebru, 2018
“White names receive 50 percent more callbacks for interviews. Callbacks are also more responsive to resume quality for White names than for African-American ones.”
Automated underwriting in mortgage lending: Good news for the underserved?

Susan Wharton Gates, Vanessa Gail Perry & Peter M. Zorn

Published online: 31 Mar 2010

Figure 6. Effect of Introducing More Accurate Underwriting Models

- Risk cutoff with a more accurate model
- More accurate underwriting model
- Risk cutoff with a less accurate model
- Less accurate underwriting model
To predict and serve?

Predictive policing systems are used increasingly by law enforcement to try to prevent crime before it occurs. But what happens when these systems are trained using biased data? **Kristian Lum** and **William Isaac** consider the evidence – and the social consequences.
Arrests

Drug usage
Positive feedback loop

Predicted crime

FIGURE 2. (a) Number of days with targeted policing for drug crimes in areas flagged by PredPol analysis of Oakland police data. (b) Targeted policing for drug crimes, by race. (c) Estimated drug use by race.
Making it formal
Unfairness definitions

Components:

- Protected class*
- Unfairness measurement

Disparate treatment

Disparate impact

Race (Civil Rights Act of 1964); Color (Civil Rights Act of 1964); Sex (Equal Pay Act of 1963; Civil Rights Act of 1964); Religion (Civil Rights Act of 1964); National origin (Civil Rights Act of 1964); Citizenship (Immigration Reform and Control Act); Age (Age Discrimination in Employment Act of 1967); Pregnancy (Pregnancy Discrimination Act); Familial status (Civil Rights Act of 1968); Disability status (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990); Veteran status (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act); Genetic information (Genetic Information Nondiscrimination Act)
https://research.google.com/bigpicture/attacking-discrimination-in-ml/
Total profit = -79200

Correct 50% loans granted to paying applicants and denied to defaulters
Incorrect 50% loans denied to paying applicants and granted to defaulters

True Positive Rate 100% percentage of paying applications getting loans
Positive Rate 100% percentage of all applications getting loans
Profit: -39600

Blue Population

Orange Population

Profit: +39600
Maximize profit

Total profit = 32400

True Positive Rate: 60% percentage of paying applications getting loans
Positive Rate: 34% percentage of all applications getting loans
Profit: 12100

True Positive Rate: 78% percentage of paying applications getting loans
Positive Rate: 41% percentage of all applications getting loans
Profit: 20300
Ignore group

Calibrated from lender POV

Unfair from applicant POV
Demographic parity

Accuracy advantage split between lender and applicant

Same total loans
Equal opportunity

Fair from applicant POV

No demographic parity
Real world example: FICO scores

Hardt, Price, Srebro 2016
False Positives, False Negatives, and False Analyses: A Rejoinder to “Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And It’s Biased Against Blacks.”

<table>
<thead>
<tr>
<th></th>
<th>WHITE</th>
<th>AFRICAN AMERICAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labeled Higher Risk, But Didn’t Re-Offend</td>
<td>23.5%</td>
<td>44.9%</td>
</tr>
<tr>
<td>Labeled Lower Risk, Yet Did Re-Offend</td>
<td>47.7%</td>
<td>28.0%</td>
</tr>
</tbody>
</table>

Angwin, Larson, Mattu, Kirchner 2016
### Data

<table>
<thead>
<tr>
<th></th>
<th>Low Risk</th>
<th>High Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>1000</td>
<td>800</td>
</tr>
<tr>
<td></td>
<td>550</td>
<td>1400</td>
</tr>
<tr>
<td>White</td>
<td>1150</td>
<td>350</td>
</tr>
<tr>
<td></td>
<td>450</td>
<td>500</td>
</tr>
</tbody>
</table>

#### Defendant POV

**Did not recidivate**

- **Pr[HR | No rec.]**: \( \frac{800}{1800} \approx 44\% \)

#### Predictor POV

**Recidivate**

- **Pr[No Rec. | HR]**: \( \frac{800}{2200} \approx 36\% \)

- **Pr[HR | No rec.]**: \( \frac{350}{1450} \approx 24\% \)

- **Pr[No Rec. | HR]**: \( \frac{350}{850} \approx 41\% \)
Fairness and causality

Berkeley graduate admissions, 1973

- 44% of male applicants admitted
- 35% of female applicants admitted

Department level:

Female acceptance rate *higher*
”Fair” casual model:

Content of boxes matter (e.g. Griggs v. Duke Power Co., 1971)
Bottom line

Can’t come up with **universal observational** fairness criteria

Fairness is based on **assumptions** on:

- Representation of data
- Relation to **unmeasured** inputs and outcomes
- Causal relation of inputs, predictions, outcomes
Left: Construct spaces are idealized versions of features and decisions and may be unobservable.

Right: Observed spaces are the typical inputs (features) and outputs (decisions) of machine learning procedures.

Example Constructs
- Intelligence
- Grit
- Success in High School

Example Observations
- IQ Score
- SAT Score
- High School GPA

Features
- CFS

Mechanisms
- Construct
- Decisions
- Success in College
- Potential after College

Decisions
- CDS

Observed Process
- OFS

Mechanisms
- College GPA
- Years to Graduate
- Post-College Salary

Observed Process
- ODS