Introduction: Big Causal Inference

Learning Causal Models from Passive Data Sources

Phil Chodrow

August 9th, 2017

Outline

Prediction

Data set $\mathcal{D} = \{\mathsf{X}, \mathsf{Y}\}$:

1. $\mathsf{X} = \{X_i\}_{i=1}^n$: features
2. $\mathsf{Y} = \{Y_i\}_{i=1}^n$: target variable

Prediction

Find $f : X \mapsto Y$ to minimize the **expected loss** $\mathbb{E}[\mathcal{L}(f(X), Y)]$, where X and Y are drawn from the same distribution as $\mathcal{D}$.

Prediction

Data set $\mathcal{D} = \{\mathsf{X}, \mathsf{Y}\}$:

1. $\mathsf{X} = \{X_i\}_{i=1}^n$: features
2. $\mathsf{Y} = \{Y_i\}_{i=1}^n$: target variable

Prediction

Find $f : X \mapsto Y$ to minimize the **expected loss** $\mathbb{E}[\mathcal{L}(f(X), Y)]$, where X and Y are drawn from the same distribution as $\mathcal{D}$.

Example

X and Y jointly Gaussian, $\mathcal{L}(\hat{y}, y) = \|\hat{y} - y\|^2 \implies$ linear regression.

Classical Inference

Inference

Approximate the conditional probability distribution $p_{Y|X}(Y|X)$ from which the $\mathcal{D}$ was sampled. In parametric inference, we select $p_{Y|X}$ from a parameterized family $\{p_{Y|X;\Theta}\}$

Classical Inference

Inference

Approximate the conditional probability distribution $p_{Y|X}(Y|X)$ from which the $\mathcal{D}$ was sampled. In parametric inference, we select $p_{Y|X}$ from a parameterized family $\{p_{Y|X;\Theta}\}$

Example

Θ might be the coefficients of linear regression, the estimated difference between two means, or something more complex.

Good inference is sufficient but not necessary for good prediction.

Limitations of Classical Inference

The questions that motivate most studies in the health, social and behavioral sciences are not associational but ***causal*** *in nature. For example, what is the efficacy of a given drug in a given population? Whether data can prove an employer guilty of hiring discrimination? What fraction of past crimes could have been avoided by a given policy? … These are causal questions because they require some knowledge of the data-generating process; they cannot be computed from the data alone, nor from the distributions that govern the data. – [Pearl, 2009]*

Associations and Causes

1. "Smoking is *correlated* with lung cancer."

Associations and Causes

1. "Smoking is *correlated* with lung cancer."
2. "There is a genetic predisposition to both smoking and lung cancer." (R.A. Fisher)

Associations and Causes

1. "Smoking is *correlated* with lung cancer."
2. "There is a genetic predisposition to both smoking and lung cancer." (R.A. Fisher)
3. "Smoking is *caused by* lung cancer." (R.A. Fisher)

Associations and Causes

1. "Smoking is *correlated* with lung cancer."
2. "There is a genetic predisposition to both smoking and lung cancer." (R.A. Fisher)
3. "Smoking is *caused by* lung cancer." (R.A. Fisher)
4. "Smoking *causes* lung cancer."

How do we use statistics to argue that X causes Y? We need two ingredients:

1. Classical, *associational* inference.
2. Additional *causal* assumptions about our study phenomena.

Classical probability theory allows us to manipulate the former; “causal calculus” allows us to manipulate the latter.

Outline

Structural Causal Models

Example [Pearl, 2009]:

$$
\begin{aligned}
Z &= f_Z(U_X) \\
X &= f_X(z, U_X) \\
Y &= f_Y(x, U_Y)
\end{aligned}
$$

X, Y, Z are variables endogenous to the model. U_X, U_Y, and U_Z are independent exogeneous variables.

"Z may effect X. Holding X and U_Y fixed, Z does not effect Y."

Given the model, a probability distribution on U_X, U_Y induces a unique probability distribution on $\mathbb{P}(X, Y, Z)$.

Do Operator

Full model:

$$
\begin{aligned}
Z &= f_Z(U_X) \\
X &= f_X(Z, U_X) \\
y &= f_Y(X, U_Y)
\end{aligned}
$$

Do Operator

Do $X = x_0$:

$$
\begin{aligned}
z &= f_Z(u_X) \\
X &= \mathsf{x_0} \\
y &= f_Y(\mathsf{x_0}, u_Y)
\end{aligned}
$$

Do Operator

Do $X = x_0$:

$$\begin{aligned} z &= f_Z(u_X) \\ X &= x_0 \\ y &= f_Y(x_0, u_Y) \end{aligned}$$

Obtain *intervention distribution* $\mathbb{P}(Y, Z | do(x_0))$.

Can now compute causal impacts, e.g. effect size:

$$\Delta[Y, x_0, x_1] \triangleq \mathbb{E}[Y|do(x_1)] - \mathbb{E}[Y|do(x_0)] \ .$$

Calculating $\mathbb{P}(Y|do(x))$

Identification

Can we calculate properties of the *controlled distribution* $\mathbb{P}(Y = y|do(x))$ from data sampled from the *uncontrolled distribution* $\mathbb{P}(X, Y, Z)$?

Calculating $\mathbb{P}(Y|do(x))$

Identification

Can we calculate properties of the *controlled distribution* $\mathbb{P}(Y = y|do(x))$ from data sampled from the *uncontrolled distribution* $\mathbb{P}(X, Y, Z)$?

Examples: in our model,

1. $\mathbb{P}(Y|do(x_0) = \mathbb{P}(Y|X = x_0)$.
2. $\mathbb{E}[Y|do(x_0) = \mathbb{E}[Y|X = x_0]$

Calculating $\mathbb{P}(Y|do(x))$

Identification

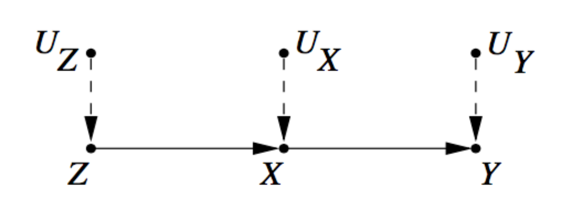
Can we calculate properties of the *controlled distribution* $\mathbb{P}(Y = y|do(x))$ from data sampled from the *uncontrolled distribution* $\mathbb{P}(X, Y, Z)$?

Examples: in our model,

1. $\mathbb{P}(Y|do(x_0) = \mathbb{P}(Y|X = x_0)$.
2. $\mathbb{E}[Y|do(x_0) = \mathbb{E}[Y|X = x_0]$

These are theorems about our model, *not* definitions.

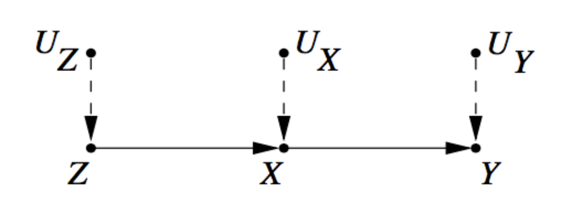
From Models to DAGs



$$Z = f_Z(U_X)$$
$$X = f_X(Z, U_X)$$
$$Y = f_Y(X, U_Y)$$

- $A \rightarrow B \equiv$ "*A might* effect *B*."
- Causal assumptions are encoded by the absence of arrows

d-Separation



d-Separation

A set $\mathcal{S}$ of nodes **blocks** a path p through a causal DAG if either:

1. p contains an arrow-emitting node in $\mathcal{S}$, or;
2. p contains a collision node $\rightarrow v \leftarrow$ such that $v \notin \mathcal{S}$ and no descendant of v is in $\mathcal{S}$.

A set $\mathcal{S}$ ***d*-separates** sets $\mathcal{R}$ and $\mathcal{T}$ if it blocks all paths between them.

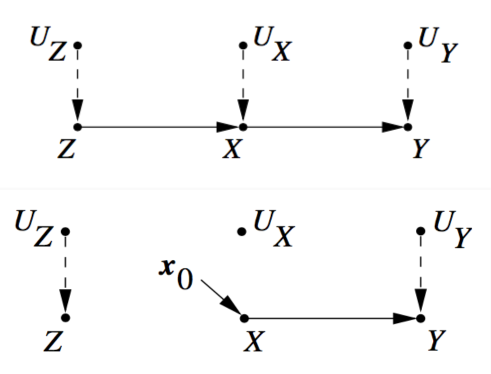
Separation and Independence

Theorem

If $\mathcal{S}$ d-separates $\mathcal{R}$ and $\mathcal{T}$, then $\mathcal{R} \perp \mathcal{T} | \mathcal{S}$.

Most criteria determining whether a particular causal inference is possible are expressed via blocking and d-separation of sets of variables.

d-Separation Example



1. Path $U_Z \to Z \to X \to Y$ is blocked by X.
2. X d-separates $\{U_Z\}$ from $\{Y\}$.
3. **Consequence:** $Y \perp U_Z | X$.

Outline

Common Issues with Massive, Passive Data

Selection Bias

You want to study all Boston commuters, but you only sample those who possess cell phones.

Transportability

You want to say something about likely energy usage patterns in New York, but you only have data for Boston.

Selection Bias

Let X be a treatment, Y an outcome.

$$S = \begin{cases} 1 & \text{Individual in data} \\ 0 & \text{Otherwise} \end{cases} \tag{1}$$

If S is dependent on X, we have selection bias.

We can estimate $\mathbb{P}(y, x | S = 1)$ from our biased experiment.

Can we get the unbiased $\mathbb{P}(y | do(x))$?

Overcoming Selection Bias

Theorem

If we have an adjustment *variable Z measured in both our biased study and the unbiased study which is* ***selection backdoor admissible***,[1] *then*

$$\mathbb{P}(y|do(x)) = \sum_z \underbrace{\mathbb{P}(y|x, z, S=1)}_{\text{from biased study}} \overbrace{\mathbb{P}(z)}^{\text{population-level measurement}} \tag{2}$$

[1] [Bareinboim et al., 2014]

Overcoming Selection Bias

Example:

- X is change in monthly transportation cost.
- Y is access to a ride-sharing program.
- Z is demographics – e.g. age, race, gender, residence.

1. Surveys of MIT students and faculty about their transportation costs before and after Lyft and Über came to Boston.

Overcoming Selection Bias

Example:

- X is change in monthly transportation cost.
- Y is access to a ride-sharing program.
- Z is demographics – e.g. age, race, gender, residence.

1. Surveys of MIT students and faculty about their transportation costs before and after Lyft and Über came to Boston.
2. **Biased demographics!**

Overcoming Selection Bias

Example:

- X is change in monthly transportation cost.
- Y is access to a ride-sharing program.
- Z is demographics – e.g. age, race, gender, residence.

1. Surveys of MIT students and faculty about their transportation costs before and after Lyft and Über came to Boston.
2. **Biased demographics!**
3. If Z is selection backdoor-admissible for our bias, then we can use a reweighting formula:

$$\mathbb{P}(y|do(x)) = \sum_z \underbrace{\mathbb{P}(y|x,z,S=1)}_{\text{from biased study}} \overbrace{\mathbb{P}(z)}^{\text{Census data}}$$

Transportability

We did our survey for MIT. Can we estimate $\mathbb{P}^*(y|do(x))$ at Harvard?

If our model says that the only relevant differences are ones we can measure, then yes.

Transportability

We did our survey for MIT. Can we estimate $\mathbb{P}^*(y|do(x))$ at Harvard?

If our model says that the only relevant differences are ones we can measure, then yes.

Example: "Harvard is just like MIT, but snootier."

Transportability

We did our survey for MIT. Can we estimate $\mathbb{P}^*(y|do(x))$ at Harvard?

If our model says that the only relevant differences are ones we can measure, then yes.

Example: "Harvard is just like MIT, but snootier."

$$\mathbb{P}^*(y|do(x)) = \sum_{\text{snootiness}} \underbrace{\mathbb{P}(y|do(x), \text{snootiness})}_{\text{Measure at MIT}} \overbrace{\mathbb{P}^*(\text{snootiness})}^{\text{Measure at Harvard}}$$

General Criteria are More Complex

Selection bias [Bareinboim et al., 2014]

Definition 4 (Selection-backdoor criterion). *Let a set* $\mathbf{Z}$ *of variables be partitioned into* $\mathbf{Z}^+ \cup \mathbf{Z}^-$ *such that* $\mathbf{Z}^+$ *contains all non-descendants of* X *and* $\mathbf{Z}^-$ *the descendants of* X. $\mathbf{Z}$ *is said to satisfy the selection backdoor criterion (s-backdoor, for short) relative to an ordered pairs of variables* (X, Y) *and an ordered pair of sets* $(\mathbf{M}, \mathbf{T})$ *in a graph* G_s *if* $\mathbf{Z}^+$ *and* $\mathbf{Z}^-$ *satisfy the following conditions:*

(i) $\mathbf{Z}^+$ *blocks all back door paths from* X *to* Y*;*

(ii) X *and* $\mathbf{Z}^+$ *block all paths between* $\mathbf{Z}^-$ *and* Y*, namely,* $(\mathbf{Z}^- \perp\!\!\!\perp Y | X, \mathbf{Z}^+)$*;*

(iii) X *and* $\mathbf{Z}$ *block all paths between* S *and* Y*, namely,* $(Y \perp\!\!\!\perp S | X, \mathbf{Z})$*;*

(iv) $\mathbf{Z} \cup \{X, Y\} \subseteq \mathbf{M}$*, and* $\mathbf{Z} \subseteq \mathbf{T}$.

Transportability [Bareinboim and Pearl, 2014]

Theorem 1 ([13]). *Let* $\mathcal{D} = \{D^{(1)}, \ldots, D^{(n)}\}$ *be a collection of selection diagrams relative to source domains* $\Pi = \{\pi_1, \ldots, \pi_n\}$*, and target domain* π^**, respectively, and* $\mathbf{S_i}$ *represents the collection of* S*-variables in the selection diagram* $D^{(i)}$*. Let* $\{\langle P^i, I_z^i \rangle\}$ *and* $\langle P^*, I_z^* \rangle$ *be respectively the pairs of observational and interventional distributions in the sources* Π *and target* π^**. The effect* $R = P^*(\mathbf{y}|do(\mathbf{x}))$ *is* mz*-transportable from* Π *to* π^* *in* $\mathcal{D}$ *if the expression* $P(\mathbf{y}|do(\mathbf{x}), \mathbf{S_1}, \ldots, \mathbf{S_n})$ *is reducible, using the rules of the do-calculus, to an expression in which (1) do-operators that apply to subsets of* I_z^i *have no* $\mathbf{S_i}$*-variables or (2) do-operators apply only to subsets of* I_z^**.*

Outline

Requirements for Causal Inference

Causal inference requires:

1. A theory of which phenomena can causally effect which other phenomena (structural causal model).
2. Data on the phenomena contained in your causal model.
3. A probabilistic model (parametric or nonparametric) consistent with your causal model.

Why or why not?

Pro: Causal inference supports much stronger scientific claims than predictive or classical associational inference.

Con: Requirements are much higher: data, math.

- Need a **strong theory** of how the variables in your study do and don't.
- **The math isn't magic** – if you don't have the right data in sufficient quantity, you can't do successful inference.

Key Question

Is it important for your result that you make a statistically validated, causal claim about the phenomena your study phenomenon?

Big data doesn't make theory obsolete. High quality causal inference depends on both.

Learn More

Two Nice Reviews :

- Gentler and more thorough: [Pearl, 2009].
- Data fusion [Bareinboim and Pearl, 2016]

PNAS colloquiuim : [Shiffrin, 2016].

Foundational Text : [Pearl, 2000].

Selection Bias : [Bareinboim et al., 2014].

Transportability : [Bareinboim and Pearl, 2013, Bareinboim and Pearl, 2014].

Nice Videos : `https://modu.ssri.duke.edu/module/introduction-causal-inference`

Nice notes on d-separation : `https://www.andrew.cmu.edu/user/scheines/tutor/d-sep.html`

References i

Bareinboim, E. and Pearl, J. (2013).
Meta-Transportability of Causal Effects: A Formal Approach.
Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS), 31:135–143.

Bareinboim, E. and Pearl, J. (2014).
Transportability from Multiple Environments with Limited Experiments: Completeness Results.
Advances in Neural Information Processing Systems, 27(November):280–288.

Bareinboim, E. and Pearl, J. (2016).
Causal inference and the data-fusion problem.
Proceedings of the National Academy of Sciences, 113(27):7345–7352.

References ii

- Bareinboim, E., Tian, J., and Pearl, J. (2014).
  Recovering from Selection Bias in Causal and Statistical Inference.
  Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI 2014), (Pearl):2410–2416.
- Pearl, J. (2000).
  Causality : models, reasoning, and inference.
  Cambridge University Press.
- Pearl, J. (2009).
  Causal inference in statistics: An overview.
  Statistics Surveys, 3(0):96–146.

Shiffrin, R. M. (2016).
Drawing causal inference from Big Data.
Proceedings of the National Academy of Sciences, 113(27):7308–7309.