STAT 8101: Applied Causality (Spring'22) Week 1: Jan 24, 2022 - Jan 30, 2022

Weekly Reading Nodes: Causal Inference In Statistics: A Primer

Han Lin hl3199@columbia.edu MS in Computer Science

1 Brief Self Introduction

I'm a second semester master student in the computer science program. My past research focused on the theories of structured random features for kernel approximation, and their applications to build efficient Transformers and GNNs. (Welcome for collaboration if you are also interested in this area!)

Causal inference seems an emerging area for exploration, especially its intersection with reinforcement learning. I'm planning to spend this whole semester delving into causal reinforcement learning. Without much background in causal inference, I know how much I need to catch up :). But I'm willing to take such challenge and have a try!

Progress this week:
Finished reading chapter 2,3 and 4 in the book "Causal Inference In Statistics: A Primer".

2 Chapter 2

2.1 Chains, Forks, Colliders

Chains: $X \to Y \to Z$: Notice that Z and X are independent, conditional on Y.

Definition: Conditional on Y: filter the data into groups based on the value of Y.

Rule 1: Two variables, X and Y, are conditionally independent give Z, if there is only one unidirectional path between X and Y and Z is any set of variables that intercepts that path.

Fork: $Y \leftarrow X \to Z$: Notice that Z and X are independent, conditional on Y. X is called common cause.

Rule 2: If a variable X is a common cause of variables Y and Z, and there is only one path between Y and Z, then Y and Z are independent conditional on X.

Colliders: $X \to Z \leftarrow Y$: X and Y are independent, but dependent conditional on Z.

Rule 3: If a variable Z is the collision node between two variables X and Y, and there is only one path between X and Y, then X and Y are unconditionally independent but are dependent conditional on Z and any descendants of Z.

2.2 d-separation

Definition: d-separated: the variables tey represent are definitely independent.
Definition: d-connected: they are possibly, or most likely dependent.
Definition: d-separation: A path p is blocked by a set of nodes Z if and only if:

(1) p contains a chain/fork, and the middle node is in Z.

(2) p contains a collider, and the middle node is NOT in Z, and NO descendant of the middle node is in Z.

If Z blocks every path between two nodes X and Y, then X and Y are d-separated, conditional on Z, and thus are independent conditional on Z.

2.3 Model Testing and Causal Search

We can test for conditional independence using a data set.

Example: guess model as $Z_1 \to X \to W$: if r_1 is not zero in regression $w = r_X X + r_1 Z_1$, then W depends on Z_1 given X (conditional correlation implies conditional dependence), so the model is wrong.

Shortcoming of regression: (1) if any parameter cannot be estimated, then joint distribution cannot be estimated, model cannot be tested. (2) tests models globally. (3) # variables involved is large.

3 Chapter 3. The Effects of Interventions

3.1 Interventions

Definition: intervention: When we intervene on a variable in a model, we fix its value. We change the system, and the values of other variables often change as a result.
Definition: condition: When we condition on a variable, we change nothing, merely narrow our focus to the subset of cases in which the variable takes the value we are interested in.

Definition: $P(Y = y|X = x)$**:** reflects the population distribution of Y among individuals whose X value is x.
Definition: $P(Y = y|do(X = x))$**:** represents the population distribution of Y if **everyone** in the population has their X value **fixed at x.**

3.2 Adjustment

Task: estimate causal effect difference/average causal effect: $P(Y = 1|do(X = 1)) - P(Y = 1|do(X = 0))$.

$X \to Y \leftarrow Z$
$P(Y = y|do(X = x)) = P_m(Y = y|X = x)$

Equations of invariance:
(1) $P_m(Y = y|Z = z, X = x) = P(Y = y|Z = z, X = x)$, (2) $P_m(Z = z) = P(Z = z)$

Adjustment Formula: (adjusting for Z / controlling for Z)

$$P(Y = y|do(X = x)) = P_m(Y = y|X = x) \quad \text{by definition} \tag{1}$$
$$= \sum_z P_m(Y = y|X = x, Z = z)P_m(Z = z|X = x) \quad \text{conditional on Z} \tag{2}$$
$$= \sum_z P_m(Y = y|X = x, Z = z)P_m(Z = z) \quad \text{X and Z independent in m} \tag{3}$$
$$= \sum_z P(Y = y|X = x, Z = z)P(Z = z) \quad \text{use the two equations of invariance} \tag{4}$$

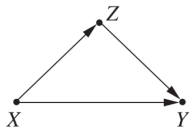
Rule 1: The Causal Effect Rule:

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, PA = z)P(PA = z) \tag{5}$$
$$= \sum_z \frac{P(X = x, Y = y, PA = z)}{P(X = x|PA = z)} \quad \text{multiply } P(X = x|PA = z) \tag{6}$$

Z ranges over all the combinations of values that the variables PA can take.

Remark: X's parents might contain unobserved variables that prevent us from calculating the conditional probabilities in the adjustment formula. We can adjust for other variables in the model to substitute for the unmeasured elements of PA(X).

Multiple Interventions and the Truncated Product Rule



Pre-intervention distribution: $P(x, y, z) = P(z)P(x|z)P(y|x, z)$

Post-intervention distribution: $P(z, y|do(x)) = P_m(z)P_m(y|x, z) = P(z)P(y|x, z)$,
sum over z: $P(y|do(x)) = \sum_z P(z)P(y|x, z)$

combine above two: $P(z, y|do(x)) = \frac{P(x,y,z)}{P(x|z)}$
Remark: conditional prob $P(x|z)$ is all we need to know in order to predict the effect of an intervention $do(x)$ from non-experimental data governed by the distribution $P(x, y, z)$.

Truncated Product Formula / g-formula:
given an intervention set X, $P(x_1, x_2, ..., x_n|do(x)[x \in X]) = \prod_i P(x_i|pa_i), \forall i$ with x_i not in X

3.3 Backdoor Criterion

Motivation: some variables have unmeasured parents that may be inaccessible for measurement.

Definition: Given an ordered pair of variables (X, Y) in a DAG, a set of variables Z satisfies the backdoor criterion relative to (X, Y) if:
(1) no node in Z is a descendant of X
(2) Z blocks every path between X and Y that contains an arrow into X.

If Z satisfies backdoor criterion:

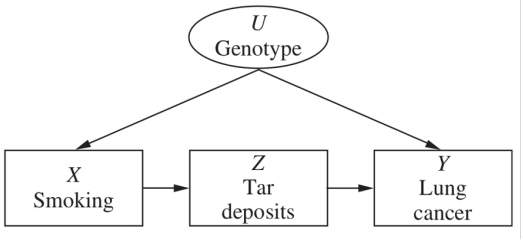
$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z) \tag{7}$$

If an empty set satisfies backdoor criterion, then no adjustment is needed: $P(y|do(x)) = P(y|x)$

Example: $X \to W \leftarrow Z \leftrightarrow T \to Y$:
Q: Compute the causal effect of X on Y for/conditioning on a specific value w of W:
A: by adjusting for set T: (section 3.5 for more detail)

$$P(Y = y|do(X = x), W = w) = \sum_t P(Y = y|X = x, W = w, T = t)P(T = t|W = w) \tag{8}$$

3.4 Frontdoor Criterion



The effect of X on Z is identifiable: (no backdoor path from X to Z)

$$P(Z = z|do(X = x)) = P(Z = z|X = x) \tag{9}$$

The effect of Z on Y is identifiable: (the backdoor path from Z to Y can be blocked by conditioning on X)

$$P(Y = y|do(Z = z)) = \sum_x P(Y = y|Z = z, X = x)P(X = x) \quad \text{Error in the book here!} \tag{10}$$

chain together: Frontdoor Formula

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|do(Z = z))P(Z = z|do(X = x)) \tag{11}$$

$$= \sum_z \sum_{x'} P(Y = y|Z = z, X = x')P(X = x')P(Z = z|X = x) \tag{12}$$

Definition: Frontdoor: Given an ordered pair of variables (X, Y) ~~in a DAG~~, a set of variables Z satisfies the frontdoor criterion relative to (X, Y) if:
(1) Z intercepts all directed paths from X to Y
(2) There is no unblocked path from X to Z
(3) All backdoor paths from Z to Y are blocked by X

Theorem 3.4.1. Front-door Adjustment If Z satisfies the front-door criterion relative to (X, Y) and if $P(x, z) > 0$, then the causal effect of X on Y is identifiable and is given by the formula:

$$P(Y = y|do(X = x)) = \sum_z P(Z = z|X = x) \sum_{x'} P(Y = y|Z = z, X = x')P(X = x') \tag{13}$$

Remark: do-calculus uncovers ALL causal effects that can be identified from a given graph.

3.5 Conditional Interventions and Covariate-Specific Effects

Motivation: Interventions may involve dynamic policies in which a variable X is made to respond in a specified way to some set Z of other variables.
(1) through functional relationship $x = g(z)$
(2) through stochastic relationship $P^*(x|z)$

Example: $do(X = g(Z))$, $g(Z) = 1$ if $Z > z$, and $g(Z = 0$ otherwise.

Goal: estimate $P(Y = y|do(X = g(Z)))$

Definition: "z-specific effect" of X: $P(Y = y|do(X = x), Z = z)$ measures the distribution of Y in a subset of the population for which Z achieves the value z **after** the intervention.

Identification: The z-specific effect can be identified by a procedure similar to the backdoor adjustment.

Remind Example: $X \to W \leftarrow Z \leftrightarrow T \to Y$:
Q: Compute the causal effect of X on Y for/conditioning on a specific value w of W:
A: by adjusting for set T: **z-specific effect**

$$P(Y = y|do(X = x), W = w) = \sum_t P(Y = y|X = x, W = w, T = t)P(T = t|W = w) \tag{14}$$

Rule 2: The z-specific effect $P(Y = y|do(X = x), Z = z)$ is identified whenever we can measure a set S of variables that $S \cup Z$ satisfies the backdoor criterion. The z-specific effect is given by the following adjustment formula:

$$P(Y = y|do(X = x), Z = z) = \sum_{\mathbf{s}} (Y = y|X = x, S = s, Z = z)P(S = s) \tag{15}$$

Then Goal could be estimated as:

$$P(Y = y|do(X = g(Z))) = \sum_z P(Y = y|do(X = g(Z)), Z = z)P(Z = z|do(X = g(Z))) \tag{16}$$

$$= \sum_z P(Y = y|do(X = g(Z)), Z = z)P(Z = z) \text{ (since Z occurs before X)} \tag{17}$$

$$= \sum_z P(Y = y|do(X = x), Z = z)|_{x=g(z)}P(Z = z) \text{ (rewrite: substitute } g(z) \text{ for x)} \tag{18}$$

Remark: Tells us that the causal effect of a conditional policy $do(X = g(Z))$ can be evaluated directly from the expression of $P(Y = y|do(X = x), Z = z)$ simply by substituting $g(z)$ for x and taking the expectation over Z (using the observed distribution $P(Z = z)$).

3.6 Inverse Probability Weighting

Motivation: Adjusting for Z might be problematic in practice, need to look at each value or combination of Z separately. Number of data samples falling within each $Z = z$ cell may be too small to provide reliable estimates of the conditional probabilities.

Definition: propensity score: $g(x, z) = P(X = x|Z = z), \forall x, z$. Can be obtained by fitting the parameters of a flexible function $g(x, z)$ to the data (minimize MSE).

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z) \text{ (by definition)} \tag{19}$$

$$= \frac{\sum_z P(Y = y|X = x, Z = z)P(X = x|Z = z)P(Z = z)}{P(X = x|Z = z)} \text{ (multiply by } P(X = x|Z = z)) \tag{20}$$

$$= \frac{\sum_z P(Y = y, X = x, Z = z)}{P(X = x|Z = z)} \tag{21}$$

$$\tag{22}$$

Intuition: Re-weight each sample by a factor $\frac{1}{P(X=x|Z=z)}$ as if they were generated from P_m rather than P, and proceed to estimate $P(Y = y|do(x))$ accordingly.

Remark: IPW is only valid when the set Z entering the factor $\frac{1}{P(X=x|Z=z)}$ satisfies the backdoor criterion. (or will introduce more bias than the one obtained through naive conditioning)

Remark: Precision (shorter confidence intervals) is another aspect.

3.7 Mediation

Motivation: Useful to know how much of variable X's effect on Y is **direct** and how much is mediated.

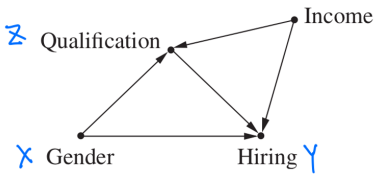
Traditional Way: by conditioning on the mediating variable.
Conceptral Way: holding the mediating variable steady without conditioning on it: intervene on it.

Definition: Controlled direct effect (CDE) on Y of changing the value of X from x to x':

$$CDE = P(Y = y|do(X = x), do(Z = z)) - P(Y = y|do(X = x'), do(Z = z)) \tag{23}$$

CDE of X on Y, mediated by Z, is identifiable if:
(1) There exists a set S_1 of variables that blocks all backdoor paths from Z to Y (S_1: I in below example, which is the set we need to sum over to remove the "do" in do(Z=z)).
(2) There exists a set S_2 of variables that blocks all backdoor paths from X to Y, after deleting all arrows entering Z. (S_2: mediating set, which is Z in below example)

Example:



$$\begin{aligned}
CDE &= P(Y = y|do(X = x), do(Z = z)) - P(Y = y|do(X = x'), do(Z = z)) && (24)\\
&= P(Y = y|X = x, do(Z = z)) - P(Y = y|X = x', do(Z = z)) && (25)\\
&\text{(since no backdoor from X to Y after intervention on Z)} && (26)\\
&= \sum_i [P(Y = y|X = x, Z = z, I = i) - P(Y = y|X = x', Z = z, I = i)]P(I = i) && (27)\\
&\text{(two backdoor paths exist from Z to Y, one through X and one through I.} && (28)\\
&\quad \text{The first is blocked since X is conditioned on. The second can be blocked by adjusting for I.)} && (29)
\end{aligned}$$

Remark:
In non-linear systems: total effect - direct effiect $\neq$ indirect effect.
Could be solved by counterfactuals.

3.8 Causal Inference in Linear Systems

Assumption: the relationships between variables are linear, all error terms have Gaussian distribution.

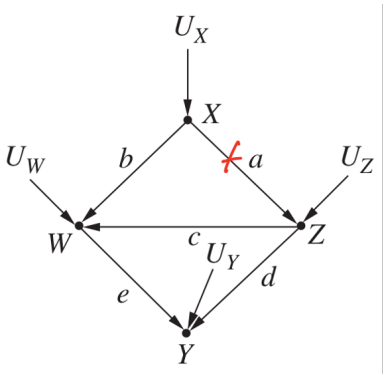
 # parameters: a set of N normally distributed variables $X_1, X_2, ..., X_N$: $2N + N(N-1)/2$ parameters, first term for N mean and N variance parameters, the second for # correlations.

Every conditional expectation $E[Y|X_1, X_2, ..., X_N]$ is given by a linear combination of the conditioning variables: (understand: could use E instead of the general P because in linear system)

$$E[Y|X_1 = x_1, X_2 = x_2, ..., X_N = x_N] = r_0 + r_1 x_1 + r_2 x_2 + ... + r_n x_n \tag{30}$$

Definition: path coefficient/structural coefficient: coefficient in the causal model. Error denoted as $\epsilon_1, \epsilon_2, ...$ Represents a direct effect, regardless of how the error terms are distributed (even if the error terms are correlated).
Definition: regression coefficient: $r_1, r_2, ..., r_N$, statistical, descriptive. Makes no assumptions about causation. $y = r_1 x + r_2 z$ make the best linear approximation of $E[y|x, z]$. Error denoted as $U_1, U_2, ...$

Causal interpretation of structural coefficients:



Direct effect of Z on Y:

$$DE = E[Y|do(Z = z + 1), do(W = w)] - E[Y|do(Z = z), do(W = w)] \tag{31}$$
$$= (d(z + 1) + ew) - (dz + ew) = d \text{ (same value as path coefficient)} \tag{32}$$

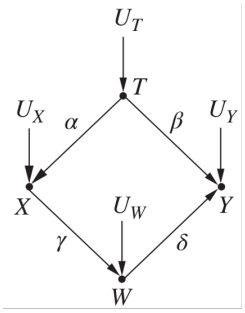
Remark: Every structural coefficient represents a direct effect, regardless of how the error terms are distributed (even if the error terms are correlated).

Total Effect of Z on Y: τ
(1) first intervene on Z, removing all arrows going into Z (including U_Z)
(2) express Y in terms of Z in the remaining model
The sum of the products of the coefficients of the edges on every non-backdoor path from Z to Y.

$$Y = dZ + eW + U_Y \tag{33}$$
$$= dZ + e(bX + cZ) + U_Y + eU_W \tag{34}$$
$$= (d + ec)Z + ebX + U_Y + eU_W \tag{35}$$
$$= \tau Z + U \ (\tau = d + ec, \text{ U contains only terms that do not depend on Z in the modified model}) \tag{36}$$

Estimating total & direct effect from non-experimental data:

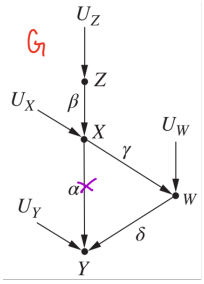
1. Steps to Estimate the Total Effect of X on Y: r_X

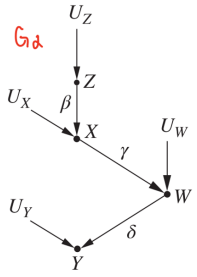


(1) find a set of covariates Z (T in example above) that satisfies the backdoor criterion from X to Y.
(2) regress Y on X and Z. $y = r_X X + r_T T + \epsilon$
(3) the coefficient of X (r_X) represents the true causal effect of X on Y.

Remark: regressing on Z adds those variables into the equation, blocking all backdoor paths from X and Y, thus preventing the coefficient of X from absorbing the spurious information those paths contain.
Remark: Direct effect of X on Y is zero in this example.

2. Steps to Estimate the Direct Effect of X on Y:





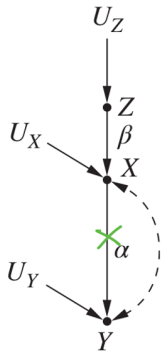
need to block not only backdoor paths, but also indirect paths going from X to Y
(1) remove the edge from X to Y, denote the resulting graph G_α
(2) if G_α exists a set of variables Z (W in above example) that d-seperates X and Y, then we regress Y on X and Z ($Y = r_X X + r_W W + \epsilon$)

3. Instrumental variable:

useful when there is no set of variables d-separates X and Y in G_α.
direct effects can be identified from total effects with instrumental variables

Definition: instrumental variable: (1) d-separated from Y in G_α, (2) d-connected to X

Example:
X and Y have an unobserved common cause represented by the dashed double-arrowed arc.
Since it hasn't been measured, we can't condition on it.



Z is an instrument wrt the effect of X on Y.

regression: $y = r_1 Z + \epsilon, x = r_2 Z + \epsilon$

$r_2 = \beta$: No backdoor between X and Z
$r_1 = \beta\alpha$: total effect of Z on Y
$\alpha = r_1/r_2$

Mediation in Linear Systems:

Estimating direct effect: same as estimating path coefficient between the two variables.
Estimating indirect effect: $IE = \tau - DE$, where τ is total effect estimated with above procedure.

Mediation in non-linear Systems:

Estimating direct effect:

$$CDE = P(Y = y|do(X = x), do(Z = z)) - P(Y = y|do(X = x'), do(Z = z)) \tag{37}$$

$$DE = E[Y = y|do(X = x), do(Z = z)] - E[Y = y|do(X = x'), do(Z = z)] \tag{38}$$

where $Z = z$ is a specific stratum of all other parents of Y (besides X).

Estimating indirect effect: cannot be given by do-expressions (IE $\neq$ DE - τ)

Chapter 4 still In Progress

4 Chapter 4. Counterfactuals and Their Applications

4.1 Counterfactuals

$$E[Y_{X=1}|X=0, Y=Y_0=1] \tag{39}$$

4.2 Defining and Computing Counterfactuals

4.2.1 Structural Interpretation of Counterfactuals

Three Steps in Computing Counterfactuals:

(1) Abduction: use evidence $E=e$ to determine the value of U

explains the past (U) from current evidence e

(2) Action: modify model M, by removing the structural equations for the variables in X and replacing them with $X=x$, to obtain the modified model M_X

bends the course of history to comply with the hypothetical antecedent $X=x$

(3) Prediction: use M_x and U to compute Y, which is the consequence of the counterfactual

predict the future (Y) based on our new understanding of the past and our new condition $X=x$

4.3 Non-deterministic Counterfactuals

4.4 Practical Uses of Counterfactuals

4.5 Mathematical Tool Kits for Attribution and Mediation