Introduction to Causal Inference in Machine Learning

1 First application: causal effect of NICU on premature baby delivery [paper]

We hope to find evidence about whether it is truly useful to invest in strengthening perinatal regionalization system, i.e., build more high-level NICU (neonatal-ICU with advanced life support equipments) to improve survival rate for premature babies. Binary variable D is defined as: $D = 1$ indicates a baby being delivered in high-level NICU, and $D = 0$ indicates a baby being delivered in low-level NICU. Binary variable Y is the outcome: $Y = 1$ indicates a survived baby, and $Y = 0$ indicates a passed-away baby. The goal is to estimate

$$E\big(Y(1) - Y(0)\big|\text{confounders}\big)$$

note that the authors are not using $Y(D)$, and we will explain what $Y(1)$ and $Y(0)$ are later. If the conditional expectation is positive, then we would suggest to the policy maker to invest more in building high-level NICUs. Because babies are precious to families expecting them, making sure to improve their surviving rate is crucial.

Continue with variable identification. Besides treatment D and outcome Y, we need to address the idea of having high-level NICU close to neighborhoods. The authors propose a binary random variable Z, where $Z = 1$ means there is a high-level NICU within 10 minutes of drving distance, and $Z = 0$ means the traveling time is longer than 10 minutes. Now

$$Y = Y(Z)$$

is the potential outcome we focus on. Next, there are some covariates X, which includes infant's gestational weeks, the month of pregnancy that prenatal care started, and mother's education. Finally, there is an unmeasured confounder U, which characterizes whether a mother is willing to go to high-level NICU: $U = `c'$ means a mother is compliant to the distance factor $Z$, $U = `a'$ means a mother will always go to high-level NICU no matter what, $U = `n'$ means a

mother will not go to high-level NICU no matter what, and $U =$ 'd' means a mother chooses the opposite to Z.

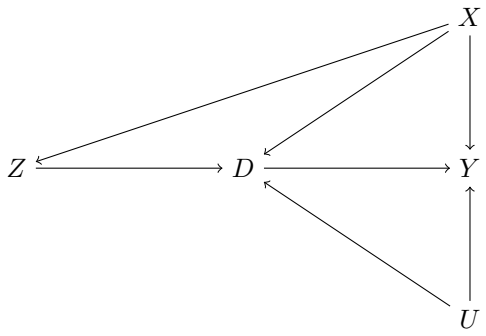
With the notations above, the goal is to estimate

$$E\left(Y(1) - Y(0) \,\middle|\, X = x, U = c\right)$$

Now, let's go through main assumptions.

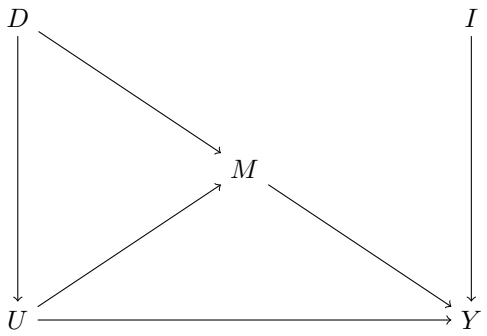
- (SUVTA) This allows us to write $D_i(z) = D_i(z_i), Y_i(z) = Y_i(z_i)$ meaning i-th individual's potential outcome is not affected by others' status. Since each mother is independent, this is plausible.
- Z affects D: i.e., $E(D(1) - D(0)) \neq 0$
- (Exchangeability) $(Y(0), Y(1), D(0), D(1)) \perp Z \,|\, X$. This is because X in some sense fully defines Z, i.e., when a mother decides where to live, X is the only deciding factor because we assume mothers are not in general expecting premature babies. And this roughly means Z will not suddenly become 1 if $Y(0)$ goes low, etc.
- $D(1) \geq D(0)$: This is plausible because if $D(1) = 0$, i.e., don't go to high-level NICU even it's close, then very likely $D(0) = 0$ given a longer distance.

Finally, the authors present a DAG as follows,



2 Second Application: Recommendation System [paper]

This paper incorporates some do-calculus techniques to alleviate the population bias issue in classic recommendation systems. We start off with a brief discussion on what is population bias and why it happens in classic algorithms. Population bias refers to an item from a popular category receiving high score even when the user may not like it that much. One of the main reasons is that classic algorithms use click-frequency as a factor in computing user-item interaction. And items from popular category will receive more clicks no matter whether the user like them or not. In term of causal inference, the user's click distribution over different categories acts like a confounder as explained below.

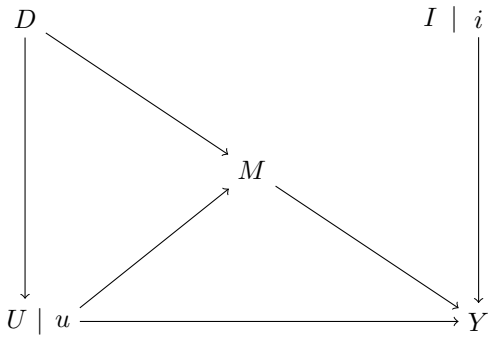


We first provide definitions of the notations. In general, we use H dimensional vector encoding. $U = (u_1, ..., u_K)$ with $u_i \in \mathbb{R}^H$ denotes the K user related features. $x_u = (a_{u,1}, ...a_{u,K}) \in \mathbb{R}^K$ is the coefficient vector of a particular user u. So, to encode a user u, we use $\sum_1^K a_{u,i}u_i \in \mathbb{R}^H$. Similarly, we let $V = (v_1, ..., v_N)$ with $v_i \in \mathbb{R}^H$ encode the N item categories. $z_{\text{item}} = (b_{\text{item},1}, ..., b_{\text{item},N}) \in \mathbb{R}^N$ is the coefficient vector for an item. Like before, we can use $\sum_1^N b_{\text{item},i}v_i \in \mathbb{R}^H$ to encode for a particular item (symbol I in the above DAG). Next, D is for user's click history frequencies. Suppose there are N item groups, $d_u = (p_u(1), ..., p_u(N)) \in \mathbb{R}^N$ is this user's click frequency. For example, if there are 3 item categories, $d_u = (0.5, 0.4, 0.1)$ means the user doesn't really click into the third item group. Then, M denotes the interaction between user and his/her click history. Formally, we write $M_u = f(d_u, u) \in \mathbb{R}^H$. A reasonable choice for f can be defined as,

$$M_u = \sum_{i=1}^{N}\sum_{j=1}^{K} d_u(i)v_i \odot x_{u,j}u_j$$

where $\odot$ stands for component-wise product.

Classic recommender system algorithms tend to implicitly have the arrow from D to U, meaning the user's click history will affect how the algorithm would characterize the user. The goal of a recommender system is to estimate $E(Y \mid U = u, I = i)$. But when we fix U, I, node D still keeps a path from U to Y: $U \to D \to M \to Y$. And we observe node D indeed satisfies the backdoor criteria (we present a SWIG),



Since a user's click history constantly changes, we wouldn't fix D. Before we present the backdoor adjustment, we first go through a quick computation from the original paper to see how exactly D affects Y (another way to see why D is a confounder),

$$
\begin{aligned}
E(Y \mid U = u, I = i) &= \frac{\int_{\mathcal{Y}} y \cdot (\sum_{d \in \mathcal{D}} \sum_{m \in \mathcal{M}} P(d)P(u|d)P(m|d,u)P(i)P(y|u,i,m))dy}{P(u)P(i)} \\
&= \int_{\mathcal{Y}} y \cdot \left(\sum_{d \in \mathcal{D}} \sum_{m \in \mathcal{M}} P(d|u)P(m|d,u)P(y|u,i,m) \right) dy \\
&= \int_{\mathcal{Y}} y \cdot \left(\sum_{d \in \mathcal{D}} P(d|u)P(y|u,i,M(d,u)) \right) dy \\
&= \int_{\mathcal{Y}} y \cdot P(y|u,i,M(d_u,u))dy
\end{aligned}
$$

The first equality follows from Bayes's rule with the assumption that I is independent from other random variables. The third equality follows from our definition that M is deterministic when given U and I. The last equality follows from the fact that given a user, his/her history click frequency is fixed. The last equality shows that the history click distribution affects Y via the mediator M. The good news is that D can be measured (online platform can collect users'

purchase history information). So, in order to cut the arrow from D to U, we can use the *backdoor adjustment*.

Following the theorem from lecture notes (here we put $X_i = (U, I)$, $X_j = Y$, and $X_A = D$), we have

$$E(Y(u,i)) = E(E(Y \mid U = u, I = i, D)) = E(g(D)) = \sum_{d \in \mathcal{D}} g(d)P(d)$$

where $g(D) = E(Y \mid U = u, I = i, D)$ is some unknown function, which may be approximated/fitted via machine learning methods.

In summary, the only tool we see here is the backdoor adjustment. So the assumptions needed are two-folds: firstly, we assume our SWIG is correct (mainly there are no other arrows connecting I and other nodes); secondly, we would need the assumptions for performing backdoor adjustment: consistency of Y and d-separation (which follows from SWIG).