STAT 8101: Applied Causality (Spring'22) Week 2: Jan 31, 2022 - Feb 7, 2022

CH21. Estimating Causal Effects from Observations

Han Lin hl3199@columbia.edu MS in Computer Science

This chapter talks about how to estimate $P(Y=y|do(X=x))$ from data.

If S satisfies back-door criteria, and $\hat{P}(Y=y|X=x,S=s)$, $\hat{P}(S=s)$ are consistent estimators, then $\sum_s \hat{P}(S=s)\hat{P}(Y=y|X=x,S=s)$ will be a consistent estimator of $P(Y=y|do(X=x))$. We could also estimate the average causal effect $E[Y|do(X=x)] = \sum_s P(S=s)E[Y|X=x,S=s]$, where the inner conditional expectation is a point-prediction of Y from X and S. However, $P(S)$ and $P(Y|X,S)$ are two potentially high-dimensional distributions. Rather than enumerating all possible values of s, we could use the law of large numbers: if we have IID $s_1, s_2, ..., s_n \sim S$, then $P(Y=y|do(X=x)) \approx \frac{1}{n}\sum_{i=1}^{n} P(Y=y|X=x,S=s_i)$.

If our interest is the average treatment effect, with $ATE = \sum_s P(S=s)(E[Y|X=1,S=s] - E[Y|X=0,S=s])$, we could also use law of large numbers to estimate it as $\frac{1}{n}\sum_{i=1}^{n} \mu(1,s_i) - \mu(0,s_i)$. To estimate $\mu(1,s_i), \mu(0,s_i)$, we could group our data by X and S. Such matching gives us a consistent estimate of ATE without using any explicit regression. Remark: (1) matching is one way of estimating identified ATE, but cannot solve identification problems. (2) it is just nearest neighbor regression, so it also suffers from the curse of dimensionality.

To alleviate problem (2), we could use the smallest set R than S that satisfies backdoor criteria. If $R = f(S)$, then $X \perp\!\!\!\perp S|R$ (R is a sufficient statistic for predicting X from S). If X is binary, then the propensity score $f(S) = P(X=1|S=s)$ could make an arbitrarily large set of control variable S builed down to a single number between 0 and 1. Since the functional form of $f(S)$ is unknown, we usually use logistic regression for modeling. By combing the propensity score and matching, it will also be easier to do matching on R than the potential large set S.

The next section, which talks about instrumental variables estimates, shares some similarity wrt the front door criteria to me, which also uses another set I that could capture the flow of effect of X on Y, and the causal effect could be estimated with wald estimator of β: $\beta = \frac{Cov(I,Y)}{Cov(I,X)}$. The intuition is reasonable to me: with certain control set S, the only way I could affect Y is through their effect on X. As mentioned in this chapter, we need to put care to ensure that there are no other unblocked paths connecting I to Y without passing through X.

The subsection 21.4 is is not quite easy to understand for me, especially the different preference of practitioners between instrumental variables and matching. But it mentioned an important point that we should know the DAG in order to do all the tricks in this chapter.