Adjustment

(a) prior to intervention

(b) after $do(X = x)$

Figure 1: Data generating process

The idea of causal inference is to use data generated by one process (prior to an intervention) to infer characteristics of another process (the process after intervention). Causal graphical model how an intervention changes the process. For each node with incoming links there is a (deterministic) function that determines its value. We also implicitly assume that these functions and the distributions of all variables without incoming links does not change as a consequence of the intervention. The specific case of causal estimation by adjusting for known confounders is shown in figure 1.

To estimate the causal effect of X on Y, we need to condition on Z so as to block the backdoor path $X \leftarrow Z \rightarrow Y$.

$$P(Y|do(X = x)) = P(Y), \text{ in graph 1b} = \sum_{z \in \mathcal{Z}} P(Y|X = x, Z = z)P(Z), \text{ in graph 1a}$$

With a binary treatment X we can also write:

$$\begin{aligned} ITE &= Y|do(X = 1), Z \leftarrow \text{ individual treatment effect, defined only if there's no unobserved noise } \epsilon_y \\ PTE(Z) &= \mathbb{E}_{\epsilon_y}\left[Y|do(X = 1), Z\right] - \mathbb{E}_{\epsilon_y}\left[Y|do(X = 0), Z\right] \leftarrow \text{ personalized treatment effect} \\ ATE &= \mathbb{E}_{Z,\epsilon_y}\left[Y|do(X = 1)\right] - \mathbb{E}_{Z,\epsilon_y}\left[Y|do(X = 0)\right] = \mathbb{E}_Z\left[ITE(Z)\right] \leftarrow \text{ average treatment effect} \end{aligned}$$

This assmumes Z is fully observable, ie there are no unobservable variables, U that cause both X and Y.

Counterfactuals

Let $Y = f_y(X, Z, \epsilon_y)$. Suppose we have a sample $\{x_i = 1, z_i, y_i\}$ and corresponding unobserved $\{\epsilon_{x,i}, \epsilon_{y,i}\}$ from the model in figure 1a. The counterfactual Y_1, the value of y that would have been observed had $x = 1$, is y_i (since x did equal 1). Y_0 is the value of y that would have been observed had $x = 0$. Even if we know f_y, we cannot in general determine Y_0 because we did not observe ϵ_y.

So now we have one thing that we know with certainty and another about which we could say something about in expectation. Does defining treatment effects in terms of counterfactuals in this way have lower noise - than in the other way because we are taking expectations only with respect to one of the parts?

Covariate Shift

Seperate from the problem of adjustment, lets define covariate shift.

- We have training data $d_{train} = \{(x_1, y_1), ..., (x_n, y_n)\}$ sampled from $P(X, Y) = P(X)P(Y|X)$.
- The test data $d_{test} = \{(x'_1, y'_1), ..., (x'_{n'}, y'_{n'})\}$ is assumed to be sampled from $Q(X)P(Y|X)$, where $Q(X) \neq P(X)$. As usual for test data, we observe only the inputs $\{x'_1, ..., x'_{n'}\}$
- We wish to use the training data to learn a hypothosis/function $h : \mathcal{X} \to \mathcal{Y}$ that minimises $\sum_{i=1}^{n'} \mathcal{L}(h(x'_i, y'_i))$, where $\mathcal{L}$ is some loss.

The covariate shift assumption is that $P(Y|X)$ has not changed.

Although we have assumed the underlying mapping, $P(Y|X)$, we are trying to learn has not changed between the test and the train data we still need to adapt standard prediction algorithms as otherwise (through a preference for 'simple' functions) we may select a function that fits well where $P(X)$ is high but performs badly where $Q(X)$ is high. There are some situations that require no adaptation. For example, if the true function is linear and we fit an (unregularized) linear model, then model that is in expectation optimal on the training data is also optimal for the test data. Another interesting example is if the model is a Gaussian process .

Covariate shift in the adjustment problem

In the adjustment problem, we have observational training data, $d_{train} = \{(x_1, z_1, y_1), ...(x_n, z_n, y_n)\} \sim P(Z, X)P(Y|Z, X)$. If we intervene and set $X = x$ then the data will be sampled from $P(Z)\delta(X - x)P(Y|Z, X)$. Note that the covariate shift assumption is satisfied in this scenario, $P(Y|Z, X)$ does not change between the train and test settings. What has changed is the joint distribtion $P(X, Z)$.