Lecture Notes 21
36-705

1 Causal Inference

Much of statistics and machine learning focuses on questions of association. Are X and Y correlated? Is X predictive of Y, and so on.

In many applications however, our questions are inherently causal: is a medication effective against a disease? Do masks prevent the spread of Covid? Was someone fired because of their age? Does making an ad larger on a website make people buy more?

These are not questions of association. Aspirin is strongly associated with headaches but we don't think that aspirin causes headaches. We often experience turbulence aafter the seat belt sign comes on in a plane. The association is strong. But turning on the seaat belt sign does not cause turbulence. This is what we mean by the phrase: "correlation does not imply causation."

2 The Potential Outcomes Framework

There are two essentially equivalent languages for causation: the first is called potential outcomes or counterfactuals. The second is structural equation models or directed acyclic graphs. We'll start with the first one.

Suppose we have two random variables (A, Y) where A is an exposure or treatment and Y is an outcome. For now, assume that A is binary such as "take aspirin $(A = 1)$" and "don't take aspirin $(A = 0)$." A typical dataset looks like this:

A	1	1	1	1	0	0	0	0
Y	97	76	83	93	100	89	13	67

Now introduce more random variables called *potential outcomes* (or *counterfactuals*). Let $Y(0)$ be the outcome that would have been observed if $A = 0$ and let $Y(1)$ be the outcome that would have been observed if $A = 1$. Causal questions involve comparisons of these two potential outcomes. Note that

$$Y = \begin{cases} Y(0) & \text{if } A = 0 \\ Y(1) & \text{if } A = 1. \end{cases}$$

We can write this as

$$Y = Y(A)$$

or

$$Y = (1 - A)Y(0) + AY(1).$$

So now we have four random variaables $(Y, A, Y(0), Y(1))$ where Y is related to $Y(0)$ and $Y(1)$ by the above consistency relations. Our data set now looks like this:

A	1	1	1	1	0	0	0	0
Y	97	76	83	93	100	89	13	67
Y(0)	?	?	?	?	100	89	13	67
Y(1)	97	76	83	93	?	?	?	?

Much of the data are missing because we don't observe $Y(0)$ when $A = 1$ and we don't observe $Y(1)$ when $A = 0$.

More generally, if $A \in \mathbb{R}$ then the set of counterfactuals is $(Y(a) : \ a \in \mathbb{R})$. In this case there are infinitely many counterfactuals. The observed Y is

$$Y = Y(A).$$

You can think of $Y(a)$ as a curve and we get to observe $Y(a)$ evaluated at A.

While all of this might seem rather obvious, thinking formally about treatment and control, and the potential outcomes is extremely important to causal inference. A point of particular emphasis is that if you are asking a causal question, ideally you need to be able to meaningfully say what the "treatment" is and what the potential outcomes are.

Here are a few examples of statements:

1. "Aspirin cures headaches." In order to cast this is the potential outcomes framework we could imagine that for a person with a headache (a unit) we could either give the person aspirin (treatment) or a placebo (control), and observe the corresponding potential outcome.

2. "She has long hair because she is a girl." This sounds like a causal statement so we should be able to describe the experiment. Is a unit a girl/boy? What exactly is a treatment? Can we meaningfully say what the potential outcomes are?

For some causal questions we can naturally define an associated "experiment". Murky causal questions are ubiquitous, and are in some sense interesting and challenging.

3 Causal Estimands

There are many possible parameters of interest. For example, $\mathbb{E}[Y(a)]$ which is the outcome if everyone had $A = a$. Here is some other notation that is sometimes used:

$$\mathbb{E}[Y(a)] = \mathbb{E}[Y|\text{set } A = a] = \mathbb{E}[Y|\text{do } A = a].$$

In general, $\mathbb{E}[Y(a)] \neq \mathbb{E}[Y|A=a]$!

When A is binary, it is often of interest to estimate the *average treatment effect (ATE)*

$$\psi = \mathbb{E}[Y(1)] - \mathbb{E}[Y(1)].$$

Think of this as the mean of Y if everyone took treatment minus the mean of Y if nobody took treatment. In prediction and machine learning one instead focuses on quantities like

$$\alpha = \mathbb{E}[Y|A=1] - \mathbb{E}[Y|A=0]$$

which is not, in general, the same as ψ. The latter is some measure of association.

How are we going to estimate ψ?

4 Randomized Experiments

Suppose that A was randomly assigned. (Think of the vaccine trials for covid.) In that case, A is independent of $(Y(0), Y(1)$ which we write as

$$A \perp\!\!\!\perp (Y(0), Y(1))$$

then we have

$$\alpha = \mathbb{E}[Y|A=1] - \mathbb{E}[Y|A=0] = \mathbb{E}[Y(1)|A=1] - \mathbb{E}[Y(0)|A=0] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \psi.$$

Randomization ensures that association IS causaation. And we can estimate α easily. Suppose, for example, that we assigned treatment by flipping a coin. Let

$$\widehat{\alpha} = \frac{1}{n_1} \sum_{i:A_i=1} Y_i - \frac{1}{n_0} \sum_{i:A_i=0} \equiv \overline{Y}_1 - \overline{Y}_0$$

where $n_1 = \sum_i I(A_i = 1)$ and $n_0 = \sum_i I(A_i = 0)$. It is easy to see that $\sqrt{n}(\overline{Y}_1 - \overline{Y}_0) \rightsquigarrow N(\theta, \tau^2)$ where $\tau^2 = 2\sigma_1^2 + 2\sigma_2^2$ and $\sigma_j^2 = \text{Var}[Y|A=j]$. Inference is easy. This is why those companies are spending millions of dollars doing randomized trials.

5 Hypothesis testing: Fisher's Exact p-values

Fisher was one of the first to understand the power of a randomized trial. In agricultural experiments, he advocated randomized experiments in order to draw rigorous causal conclusions. A natural subsequent problem is: given an estimate of the causal effect, assess its significance (or construct confidence intervals for it).

Fisher gave a way to construct valid p-values under what is called the *sharp null*, i.e. the null hypothesis that for every unit i the potential outcomes are the same under the treatment and control, i.e. the treatment has no effect. The method is reminiscent of the permutation method we used for two-sample testing.

Suppose we test $H_0 : \theta = 0$ by rejecting when $|\widehat{\alpha}|$ is large. Under the null hypothesis, we can determine both potential outcomes $Y_i(0)$ and $Y_i(1)$ for all the units.

We can now use the permutation method. Say there are n subjects and m were treated. Permute the values of A_i and let T' denote the m units who receive treatment: then our estimate would be:

$$\widehat{\psi}_{T'} = \frac{1}{m}\sum_{i\in T'} Y_i(1) - \frac{1}{n-m}\sum_{i\notin T'} Y_i(0),$$

where we can use the sharp null hypothesis to "fill in" the potential outcomes we do not observe. We can repeat this many times (say B) and compute the p-value:

$$\text{p-value} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(|\widehat{\psi}_{T_b}| \geq |\widehat{\psi}|).$$

It is easy to verify that this is a valid p-value.

6 Confounding

For many policy questions, we cannot actually do a randomized trial. For instance, if I wanted to know if smoking caused lung cancer, there are ethical issues with trying to run a randomized trial. In this case, we have to use *observational* i.e. we have information about many people who are smokers and not, and whether they have lung cancer or not. It is clear that we can measure the correlation between smoking and lung cancer: the main question is when, if ever, can we claim a causal relationship?

Here is a motivating example: Suppose that our population has two kinds of people, those who are always healthy ($Y_i(1) = Y_i(0) = 1$) irrespective of whether they take the treatment or not, and those who are always unhealthy ($Y_i(1) = Y_i(0) = 0$) irrespective of whether they take the treatment or not. Then $Y_i(1) - Y_i(0) = 0$ for all i so there is no causal effect. Suppose further that mostly healthy people take the treatment, while the unhealthy ones do not take the treatment. The causal effect is $\psi = 0$, but the estimator above would yield, $\widehat{\psi} \approx 1$, and we might incorrectly conclude that the treatment is beneficial. The data would look like this:

A	1	1	1	1	0	0	0	0
Y	1	1	1	1	0	0	0	0
Y(0)	1	1	1	1	0	0	0	0
Y(1)	1	1	1	1	0	0	0	0

Suppose however, that we knew who the healthy people were and who the unhealthy people were (we could gather such information by asking people questions about their lifestyle and other things). Then we could try to compare healthy people who took the treatment with healthy people who did not and similarly compare unhealthy people who took the treatment with unhealthy people who did not (and then try to combine these two estimates in some way). In this case, when we compared two healthy people who took the treatment and who did not we would see the treatment had no effect, and similarly for the unhealthy ones. We would correctly conclude that the treatment has no effect.

The key assumption that makes causal inference from observational data possible is the assumption of *no unmeasured confounding* or *selection on observables* or *ignorability*. Formally, we suppose that we have access to covariates X (think demographic information) such that,

$$A \perp\!\!\!\perp (Y(1), Y(0))|X.$$

This is an assumption. Roughly the assumption is plausible in settings where we believe we can measure all of the covariates that explain the decision to take the treatment. We also need the assumption that $\mathbb{P}(A = 1|X = x)$ is bounded away from 0 and 1, so that every individual has some non-zero chance of being either treated or in the control group.

One way to think about this assumption, is that conditional on X we have a randomized trial: the treatment is independent of the potential outcomes. So if we condition on the confounders X we no longer have any selection bias.

In what follows we will assume we have random variables $(X, A, Y, Y(0), Y(1))$ where

$$Y = AY(1) + (1 - A)Y(0) = Y(A).$$

7 Identification under no unmeasured confounding

We want to estimate:

$$\psi = \mathbb{E}[Y(1) - Y(0)]$$

assuming that

$$A \perp\!\!\!\perp (Y(1), Y(0))|X.$$

Now

$$\begin{aligned}\mathbb{E}[Y(1)] &= \int \mathbb{E}[Y(1)|X = x]p(x)dx = \int \mathbb{E}[Y(1)|X = x, A = 1]p(x)dx \\ &= \int \mathbb{E}[Y|X = x, A = 1]p(x)dx = \int \mu_1(x)p(x)dx\end{aligned}$$

where $\mu_a(x) = \mathbb{E}[Y|X = a, A = a]$. Note that thus is NOT equal to $\mathbb{E}[Y|A = 1] = \int \mu_a(x)p(x|1)dx$. Similarly, $\mathbb{E}[Y(0)] = \int \mu_0(x)p(x)dx$. So

$$\psi = \mathbb{E}[Y(1) - Y(0)] = \int [\mu_1(x) - \mu_0(x)]p(x)dx.$$

This is a function of the observed data (X, A, Y) so we can estimate it.

In the case that A is continuous, the same argument shows that

$$\mathbb{E}[Y(a)] = \int \mu_a(x)p(x).$$

8 Estimation under no unmeasured confounding

The most direct way to estimate ψ is to estimate:

$$\mu_0(x) = \mathbb{E}[Y|X = x, W = 0]$$
$$\mu_1(x) = \mathbb{E}[Y|X = x, W = 1].$$

These are two functions of the covariates X, one of them is the average outcome of the treatment group as a function of the covariates, and the other is the average outcome of the control group as a function of the covariates.

Estimating a conditional expectation is a problem is probably the most common problem in statistics – it is known as *regression*. We will delve into this formally in the next few lectures but for now let us suppose that someone hands us estimators $\widehat{\mu}_0$ and $\widehat{\mu}_1$ of these two functions.

Then we can compute the plug-in estimator:

$$\widehat{\psi} = \widehat{\mathbb{E}}_X \left[\widehat{\mu}_1(X) - \widehat{\mu}_0(X)\right] = \frac{1}{n}\sum_{i=1}^{n} \left[\widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)\right]$$

which is just the average of the difference between two regression functions. One approximately correct way to think about this is that we are using regression to impute the missing potential outcomes for each individual.

There are other ways to try to estimate ψ. The other popular estimator is called the inverse propensity score estimator. The *propensity score* is

$$\pi(x) = \mathbb{P}(A = 1|X = x),$$

which represents the probability that a unit with covariates x receives treatment. Note that,

$$\mathbb{E}[A|X=x] = \pi(x)$$
$$\mathbb{E}[1-A|X=x] = 1-\pi(x).$$

Let $p(y|x,a)$ denote the density of Y given x and a and recall that $\pi(x) = \mathbb{P}(A=1|X=x)$. So, when $a=1$,

$$p(x,a,y) = p(x)p(a|x)p(y|x,a) = p(x)\pi(x)p(y|x,1)$$

and when $a=0$,

$$p(x,a,y) = p(x)p(a|x)p(y|x,0) = p(x)(1-\pi(x))p(y|x,0).$$

So, for $a=1$,

$$\begin{aligned}
\mathbb{E}[Y(1)] &= \int \mathbb{E}[Y|X=x, A=1]p(x)dx = \int\int yp(y|x,1)p(x)dxdy \\
&= \int\int \frac{y}{\pi(x)} p(y|x,1)\pi(x)p(x)dxdy \\
&= \int\int \frac{y}{\pi(x)} p(x,a=1,y)dxdy \\
&= \int\int \frac{ay}{\pi(x)} p(x,a=1,y)dxdy \\
&= \sum_{a=0}^{1} \int\int \frac{ay}{\pi(x)} p(x,a,y)dxdy \\
&= \mathbb{E}\left[\frac{AY}{\pi(X)}\right].
\end{aligned}$$

Similarly,

$$\mathbb{E}[Y(1)] = \mathbb{E}\left[\frac{(1-A)Y}{1-\pi(X)}\right].$$

Therefore,

$$\psi = \mathbb{E}\left[\frac{YA}{\pi(X)}\right] - \mathbb{E}\left[\frac{Y(1-A)}{(1-\pi(X))}\right].$$

This suggests the estimator

$$\widehat{\psi} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_iA_i}{\pi(X_i)} - \frac{Y_i(1-A_i)}{1-\pi(X_i)}\right].$$

This is called the Horvitz-Thompson estimator or the inverse probability weighted (IPW) estimator. This requires that $\pi(x)$ be known as it would be in a randomized experiment. Otherwise we have to insert an estimate of $\pi(x)$. This is again a problem of regression except the outcome is binary.

9 Advanced topics

This is just the tip of the iceberg. If you take a course in Causal Inference you will see many other interesting things such as:

1. No unmeasured confounding is just one assumption that leads to identification of a causal effect. More broadly, in economics, political science and other fields people look for what are called natural experiments, i.e. roughly some subset of the population for which the assignment to treatment/control is nearly random.

2. Even in a randomized trial you might have something called non-compliance, i.e. some people don't do what they are told. In this case, you need to adjust your estimates. This is a canonical example of something called an instrumental variable problem.

3. There are many things beyond the average treatment effect that you might want to estimate. They all have different assumptions under which they are identified (i.e. can be written in terms of observable quantities) and there are different strategies to estimate them.

4. There is a very nice/simple way to combine the regression-based and propensity-score based estimators from above to construct what are called *doubly robust* estimators. These have the property that they are consistent if you can estimate either the regression function or the propensity score well (i.e. you do not need to estimate both well).

5. The plug-in estimator $\widehat{\psi} = n^{-1} \sum_i \int [\widehat{\mu}(X_i, 1) - \widehat{\mu}(X_i, 0)]$ is not optimal. Finding optimal estimators of functionals is part of semiparametric theory.

6. There are many different languages for talking about causality and causal inference. We used potential outcomes. Many people use structural equation models and directed graphs. These lead to the same formulas for causal effects. We might revisit this later.