INTRODUCTION TO CAUSAL INFERENCE

Author
Vinitra Muralikrishnan

October 7, 2024

Contents

1 Overview

Causal inference involves trying to establish causal factors or the causal covariates that influence the outcome or Y variable of a dataset. Machine learning as a domain, especially supervised learning tries to approach problems from a predictive modeling perspective. Generate new images? Build a model that can predict pixels given a prior image. This will generate a new image, similar to the prior, with some variations. Given a dataset consisting of multivariate random variables, we try to predict something using these variables. Machine learning tries to establish association or relations between covariates and the target prediction variable.

So what does causal inference do? It establishes the direct cause-and-effect link between the covariates and the target outcome variable. How is that different from association? Let's consider the example of 2 schools, 1 of which is treated with a program wherein students get tablets to conduct their studies whereas the other doesn't. In order to measure the efficacy of the program, data is collected on both schools about their tuition, facilities, amenities and also whether or not student possesses a tablet. The outcome variable here is a metric that measures their performance. Now if we use this data for supervised learning, it would involve establishing correlations between the covariates and the performance of the student. It answers the question that given specific values X, the outcome could like by Y. But does it necessarily mean that X directly causes Y? This can be understood intuitively. Suppose we change the value of X to something, that has not been seen in the dataset. Would the model give a reliable performance metric Y? In the real world as well, we can see in many cases were machine learning models perform notoriously bad at predicting values for data that is not known. While they are able to establishing complex correlations, they are not able to establish direct causal factors. We can predict about the unknown if we know exactly the causes of the known and the effect of each of the causal factor. With this information, we can tell what happens if some of the causal factors are changed, then how it would affect the outcome. ML predictive models cannot do that. This is where causal inference comes in. It tries to delve deeper into problem solving, by querying about the exact effect of covariates, by asking *what if*? in order to understand, the exact cause-and-effect relationships.

Causal inference tries to establish the causation using certain metrics that measure the effect. It can be done by say toggling on and off features of the data and see the effect. This toggling action is called treatment. In the above tablet example, this treatment is giving the tablet. Now consider Y_1 to be the outcome of the students who were given table and Y_0 to be the outcome variable of the students who were not given tablets. In order to measure the effect of the treatment (i.e tablet given or not), on the students who got tablet and students who didn't get it, we would need their outcomes if they hadn't received the treatment they got. These are counterfactuals. This means, for $T = 1$, the counter factual, would be outcome if the students who received the tablet, didn't receive it which is given by $Y_0|T = 1$. Similarly, the counterfactual outcome for students who didn't receive the tablet, would be the outcome if they had received it i.e $Y_1|T = 0$. Now in reality, these two variables are unobservable. They are not in our data and that itself poses the underlying challenge of causal inference. Without the information, about the counter factuals, we cannot establish with certainty the causal effect link between covariates and the outcome variable. We will denote the counter factuals as Y_0' and Y_1'.

2 Mathematical proof of association ! = causation

Let's look at the mathematic proof that association ! = causation.

In order to establish the causal effect link and understand how covariates affect the outcome, we need to understand the average effect of the treatment on the outcome variable. This requires access to the counter factuals. Imagine for a second, we do have access to such counter factuals through some god-given power. We get the average effect of treatment by the metric average treatment effect (ATE) given by $E[Y_0 - Y_1]$. In this case Y_0, Y_1, can contain observed and counter factual values. Another metric used

is the average treatment effect on the treated (ATT). Both these metrics are explained in the example below:

T	Y_0	Y_1	$Y_1 - Y_0$
0	500	450	-50
0	600	600	0
1	800	600	-200
1	700	750	50

$ATE = \frac{-50+0-200+50}{4} = -50$

$ATT = \frac{-200+50}{2} = -75$

Intuitively we know that association is not equal to causation. This is because when presented with outcomes, we are thinking, about other factors that could have caused the outcome, shadowing other covariates and influenced the outcome. Suppose the outcome, is shown to increase in the above example, we are not sure if this is directly because of the tablet introduction or if it is because schools who can provide tablets have better amenities, teaching facilities as well. This is bias and this is what differentiates association from causation. The bias can cause some of the other hidden or non-obvious factors to shadow the other seemingly plausible causal factors and influence the outcome that may not be apparent. Only if we remove this bias i.e eliminate the effect of all other covariates, can we measure the effect of the target covariate i.e treatment of tablet on the outcome i.e student performance. This in above example means, both schools would have to be comparable or match in every other perspective such as amenities, facilities, in order to truly study and understand the effect of the tablet treatment. Now let's look at mathematically.

Now association is measured by $E[Y|T = 1] - E[Y|T = 0]$. This is same as the ATE $E[Y_1 - Y_0]$ in the above example.
We can rewrite the above equation as follows:

$$E[Y_1|T = 1] - E[Y_0|T = 0]$$

Adding and subtracting $E[Y_0|T = 1]$ we get;

$$E[Y_1|T = 1] - E[Y_0|T = 1] + E[Y_0|T = 1] - E[Y_0|T = 0]$$
$$E[Y_1 - Y_0|T = 1] + \{E[Y_0|T = 1] - E[Y_0|T = 0]\}$$
$$ATT + BIAS$$

Hence bias is given by how the treated and the controlled group differ before treatment, in case neither of them has received the treatment.
When presented with data, we think that there lies some bias which means $E[Y_0|T = 1] > E[Y_0|T = 0]$. This means, that the outcome i.e performance of student would be greater than the performance of the students who were not given tablet, regardless of the tablet administration due to other factors. The goal of causal inference is reducing the bias so the cause and effect links can be established.
Now let's say $E[Y_0|T = 1] = E[Y_0|T = 0]$ i.e there is no bias and the treated and controlled groups are comparable. This implies that the outcome of the students who were given tablets, matches the performance of the students without the tablet, if the tablet was taken away from the former. This means association is equal to causation. This also means that $E[Y_0|T = 1]$ and $E[Y_0|T = 0]$ are interchangeable. Now consider the average treatement effect;

$$ATE = E[Y_1 - Y_0]$$
$$E[Y_1|T = 1] - E[Y_0|T = 0]$$
$$E[Y_1|T = 1] - E[Y_0|T = 1]$$
$$E[Y_1 - Y_0|T = 1] = ATT$$

You can find more information in Chapter 1 of Causality Handbook.