1 Causality Overview

Many questions in a variety of fields and industries, from health-care, to economics, and even increasingly in tech are inherently causal in nature. For example, what ad will make a user click on this link? What is the efficacy of this treatment for this disease? What are the effects of these incentives on these users’ behavior? These questions are all asking something about the underlying causal structure, and their answers cannot be based solely on data, since the data-generating process itself is key to answering these questions. In most of the machine learning problems we have seen so far, we have not made the distinction between correlation and causation, primarily because the causal structure is obvious or assumed. As we move towards performing learning on data coming from humans or strategic agents, or learning across time, the idea of causality becomes more important.

A simple example of causal inference - the process of drawing a conclusion about a causal connection based on the conditions of the occurrence of an effect - is that of a user taking aspirin for recurring headaches. We would like to understand the effect of aspirin for dealing with headaches, and the next time the patient gets a headache we make them take an aspirin. We observe that after taking the aspirin, the patient reported that their headache went away. Can we conclude that the aspirin led to the headache going away? The link we are seeing, is a correlation, but it by no means gives us the causal link.

What if, to take the aspirin, the patient drank a glass of water. Maybe the aspirin had no effect on the headache, and the patient was merely dehydrated. How can we test for true effect of aspirin, while controlling for the fact that the patient drinks water? In essence we want to estimate the counterfactual: If the patient had not taken the aspirin, would the headache have gone away?
This simple example illustrates two of issues at the heart of causal inference. The first problem, is the fundamental misery of causal inference: we cannot observe the counterfactual. Put in simpler terms: we cannot observe the effect of not having done something in the past.

The aspirin example also illustrates the idea of a confounding variable, namely the drinking of the water. Since it is not implausible that the water could heal the headache, we may be mistakenly attributing the benefits of drinking water to aspirin if we are not careful.

In this lecture, we will introduce three of the main paradigms for causality:

1. Granger Causality
2. Structured Causal Models
3. Rubin’s causal framework

Though we define each type, we will focus mainly on Rubin’s causal framework and the setting up, and interpretation of randomized control trials to be able to infer causal effects. Though all three types of causality have slightly different definitions and quirks, we begin by introducing the idea of causality, and the main motivation for causal inference. We remark that the act of actually learning causal structures, is an active area of research in many fields, and sample papers and resources for interested students can be found on the course website. We will take the structure in most of our problems as given, and focus on understanding causal effects.

2 Granger Causality

Figure 1: In the plot, we see two time-series, $X$ and $Y$, where it can be seen empirically that with time lag of 5 time steps, process $X$ is highly correlated with process $Y$. Thus, $X$ Granger causes $Y$.

We begin by introducing Granger causality. Granger causality is maybe the most intuitive to understand, and can be described by the simple idea of correlation with a time delay. It is computationally simple to check. Granger causality is built around two simple ideas:

1. The cause must happen before its effect
2. The cause has unique information about the future values of its effect
A time series $X$ is said to Granger-cause $Y$ if it can be shown, (usually through a series of t-tests and F-tests on lagged values of $X$), that those $X$ values provide statistically significant information about future values of $Y$. In more mathematical terms, let $X_t, Y_t$ be two stochastic processes indexed by time $t \geq 0$. Let $I_t$ be all of the information in the universe up to and including time $t$. Then if:

$$P(Y_{t+1} | I_t) \neq P(Y_{t+1} | I_t / X_t)$$

or equivalently:

$$Var(Y_{t+1} | I_t) < Var(Y_{t+1} | I_t / X_t)$$

Where $I_t / X_t$ denotes all of the information at time $t$ not including the process $X_t$. Then $X_t$ contains some information about $Y(t)$, and thus $X_t$ is said to Granger-cause $Y(t)$.

Therefore the salient feature of Granger causality, when compared to other notions of causality is the dependence on time. What are the flaws of Granger Causality? If $X$ Granger causes $Y$, does $X$ really cause $Y$?

3 Structured Causal Models

We now discuss Structured Causal Models (SCMs). SCMs are at the heart of Judea Pearl’s contributions to the literature of causality, and causal inference. The main idea behind SCMs is representing causal structures as graphical models. Graphical models, as their name suggest are graphical ways for representing a probability distribution that describe the dependence between random variables as edges in a graph. SCMs are graphical models that can be represented as directed, acyclic graphs. The graphical model shown in Figure 3 shows that the joint probability distribution $P(Z) = Pr(X_1, X_2, X_3, X_4, Y)$ can be factored as:

$$Pr(X_1, X_2, X_3, X_4, Y) = Pr(X_1)Pr(X_2)Pr(X_3|X_1, X_2)Pr(X_4|X_2)Pr(Y|X_3, X_4)$$

Where $Z = (X_1, X_2, X_3, X_4, Y)$. In this form, the causal dependencies are clear.

![Directed Acyclic Graph](image)

**Figure 2:** In this figure, we can see the result of imputing the random variable $X_3$ to $i$.

Given a causal structure in this form, Pearl popularized the idea of imputation to investigate causal effects. Imputation on a graphical model, is described by the $do$ operator. $do(X = i)$ 'sets' the value of the random variable $X$ to $i$, and then cuts off the links to the parents of $X$ in the causal structure. Thus, the result of $do(X_3 = i)$ in Figure 3, is a joint probability distribution:

$$P_2(Z) = Pr(X_1, X_2, X_3, X_4, Y) = \begin{cases} 
Pr(X_1)Pr(X_2)Pr(X_3|X_1, X_2)Pr(Y|X_3 = i, X_4) ; X_3 = i \\
0 ; X_3 \neq i 
\end{cases}$$
Now, by comparing the two probability distributions, we can investigate the causal effect of $X_3$ on $Y$. For example, we can directly compare $\mathbb{E}_{P_1}[Y]$ and $\mathbb{E}_{P_2}[Y]$. Going back to the aspirin example, imputing water, would mean we make the patient drink water irrespective of whether or not they take aspirin. If we see that the headaches do not improve, we can conclude that the water has no effect on the patient.

**Example:** We now show an illustrative example for why imputation is fundamentally different than conditioning. Suppose we want to know if giving people a smart meter will decrease their energy consumption. Let’s denote income by $X_1 \sim \text{Uniform}(0, 1)$, owning a smart meter by $X_2 \sim \text{Bernoulli}(X_1)$, and monthly energy consumption by $Y \sim \text{Normal}(1 + 2X_1 - \frac{X_2^2}{2}, \sigma^2)$. Thus, the higher your income, the more likely you are to have a smart meter, and the more income you have, the higher your energy consumption.

![Figure 3: SCM for Example](image)

In the naive case, we would simply calculate $\mathbb{E}[Y|X_2 = 1] - \mathbb{E}[Y|X_2 = 0]$, and claim that this is the effect of giving people a smart meter. This would be disregarding the fact, that conditioning on smart meters, the subject has higher income, and thus higher energy usage. Thus we might conclude that having a smart meter increases your energy consumption. This is exactly what we find:

$$\mathbb{E}[Y|X_2 = 1] = \frac{11}{6}$$
$$\mathbb{E}[Y|X_2 = 0] = \frac{5}{3}$$

Therefore, conditioning on smart meters we see a $1/6$ increase in energy consumption.

Knowing about causal inference, let us see the difference when we impute $X_2$. Now, having a smart meter tells us nothing about the subjects income, so we correctly see that smart meters decrease energy consumption:

$$\mathbb{E}[Y|\text{do}(X_2 = 1)] = \frac{3}{2}$$
$$\mathbb{E}[Y|\text{do}(X_2 = 0)] = 2$$

Therefore, when imputing on smart meters we see a $1/2$ decrease in energy consumption.

## 4 Rubin’s Causality

Rubin’s causal framework, also known as the potential outcomes framework, is perhaps the most direct of the three frameworks for estimating causal effects. The following is a broad introduction to the potential outcomes framework, for a much more thorough treatment see the book *Causal Inference of Statistics, Social, and Biomedical Sciences*, by Guido Imbens, and Donald Rubin.

The framework directly addresses the fundamental misery of causal inference, and hinges on a notion of causality dealing with actions or treatments on individuals or units (e.g. people taking or not taking aspirin).
Subject’s Causal Effect ($Y_T - Y_c$)

<table>
<thead>
<tr>
<th>Subject</th>
<th>$Y_c$</th>
<th>$Y_T$</th>
<th>Subject’s Causal Effect ($Y_T - Y_c$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>12</td>
<td>+2</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>16</td>
<td>+2</td>
</tr>
</tbody>
</table>

Table 1: In an ideal setting our data would look like this. There would be no missing information.

<table>
<thead>
<tr>
<th>Subject</th>
<th>$Y_c$</th>
<th>$Y_T$</th>
<th>Subject’s Causal Effect ($Y_T - Y_c$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>2</td>
<td>?</td>
<td>16</td>
<td>?</td>
</tr>
</tbody>
</table>

Table 2: Real data in the potential outcomes framework looks like this. How can we estimate the causal effects?

The goal, in this framework is to understand the effects of the treatment on the subject, and it is best understood through an example. Let $Y^i_c$ and $Y^i_T$ be random variables describing the outcomes of applying the control and treatment actions to individual $i$ respectively. $Y^i_c$ and $Y^i_T$ are known as potential outcomes, and the fundamental problem is that we cannot observe both $Y^i_c$ and $Y^i_T$. What this means is that, in an ideal setting, our data would resemble Table 1, but in practice it most likely resembles Table 2.

The approach that Rubin popularized to solve the problem of estimating the potential outcomes relies on the fact that:

1. The unit level causal effects cannot be observed. By extension, this means that population level causal effects also cannot be observed.

Thus, we must give the different treatments to different individuals in a population and if the population satisfies some homogeneity assumption, we can estimate $Y_c$ and $Y_T$ and get an estimate of the causal effect. Thus, the potential outcomes framework rests on two main insights, and a key assumption. The two insights are:

1. To learn about causal effects, we must have replication. That is, for some units or subjects, we must observe $Y^i_c$ and for others we must observe $Y^i_T$.

2. The assignment mechanism for the treatment (whether a unit gets the control or treatment action) must be known to us, and is critical in estimating the causal effects.

These insights rely heavily on the stable unit treatment value assumption (SUTVA), that the potential outcome observation for one unit should be unaffected by the assignment of treatments to other units in the study. In effect, this assumption allows us to treat all subjects as independent.

Can you come up with examples where this assumption does not hold?

Though the assignment mechanism can be quite intricate, since it can be used to maximize the statistical power of the experiment or take into account factors affecting the involvement of subjects in the study, we will now focus on the simple case of uniform random assignment of the treatment.

**Example: Randomized Control Trials:** Randomized control trials (RCTs), mainly used in economics and (less formally) in industrial settings, are ways of estimating the causal effect of a given treatment in the potential outcomes framework. As their name suggest, the assignment rule for the subjects to either the treatment or control groups is randomized, meaning that every unit or subject in the trial has a non-zero chance of receiving the treatment. Under the simplest assignment rule, this simplifies the task of estimating the causal effect.

Assume that you have a population of $N$ subjects. For each subject, we randomly assign them to the control (0) or treatment group (1) with probability $1/2$. Denote subject $i$’s group with the random variable $X_i \in \{0, 1\}$.

After applying the treatment we observe the potential outcomes for both groups, and calculate $Y^i_T = \mathbb{E}[Y_i | X_i = 1], Y^i_c = \mathbb{E}[Y_i | X_i = 0]$. The treatment effect for an individual is then:
\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|}
\hline
Time & Control & Treatment & Difference \\
\hline
Pre & $X_c$ & $X_T$ & $X_T - X_c$ \\
Post & $Y_c$ & $Y_T$ & $Y_T - Y_c$ \\
\hline
\end{tabular}
\caption{RCT table}
\end{table}

\[E[Y_i|X_i = 1] - E[Y_i|X_i = 0]\]

Clearly, we cannot calculate this. But if the each subject \(i\) can be seen as independent samples from the same distribution \(P_{Y|X}\), then we can estimate \(Y\), where \(Y_i \sim P_{Y|X}\):

\[E[Y|X_i = 1] - E[Y|X_i = 0] = \frac{1}{N_T} \sum_{i=1}^{N_T} Y_T^i - \frac{1}{N_c} \sum_{i=1}^{N_c} Y_c^i\]

Where \(N_c\) and \(N_T\) are the number of subjects in the control and treatment groups respectively.

Note that we have implicitly used the SUTVA, and a homogeneity assumption on the population of \(N\) subjects.

Given this quick introduction, what are the drawbacks to RCTs and Rubin’s causal model?

\section{Conclusion}

In these notes we introduced 3 causal models:

\begin{enumerate}
\item Granger causality - which deals with explanatory power with a direct dependence on time
\item Structured causal models - which represents causal dependencies using graphical models and is focused on imputation to measure causal effects
\item Rubin’s causal framework - which is focused on counterfactual estimation, through experimental setup.
\end{enumerate}

This was meant as a broad overview of some ideas in causal estimation, but it should hopefully get you thinking about causality when dealing with machine learning on data generated from humans.