1

Causality: Hey, We Have Machine Learning, So Why Even Bother?

Our journey starts here. In this chapter, we’ll ask a couple of questions about causality. What is it? Is causal inference different from statistical inference? If so – how? Do we need causality at all if machine learning seems good enough? If you have been following the fast-changing machine learning landscape over the last five to ten years, you might have noticed many examples of – as we like to call it in the machine learning community – unreasonable effectiveness of neural networks (and other algorithms) in computer vision, natural language processing and other areas.

You might ask yourself – if all this stuff works so well, why would we bother and look into something else?

We’ll start this chapter with a brief discussion on the history of causality. Next, we’ll consider a couple of motivations for using causal rather than purely statistical approach to modeling and we’ll introduce the concept of confounding. Finally, we’ll see the examples of how causal approach can help solve challenges in marketing and medicine. By the end of this chapter, you will have a good idea of why and when causal inference can be useful. You’ll be able to explain what confounding is and why it’s important.

In this chapter, we will cover:

* A brief history of causality
* Motivations to use causal approach to modeling
* How not to lose money… and human lives?

A brief history of causality

Historically, many people were thinking and writing about the nature of causality, providing us with valuable or at least interesting insights. Aristotle – one of the most prolific philosophers of ancient Greece – claimed that understanding causal structure of a process is a necessary ingredient of knowledge about this process. Moreover, he argued that being able to answer why questions is the essence of scientific explanation (Falcon, 2006; 2022). Aristotle distinguishes four types of causes (material, formal, efficient and final), an idea that might capture certain interesting aspects of reality as much as it might be counter-intuitive to a contemporary scientist or researcher. David Hume, a famous 18th century Scottish philosopher, proposed a more unified framework for cause-effect relationships. Hume starts with an observation that we never experience cause-effect relationships in the world. The only thing we experience is that some events are conjoined:

“We only find, that the one does actually, in fact, follow the other. The impulse of one billiard-ball is attended with motion in the second. This is the whole that appears to the outward senses. The mind feels no sentiment or inward impression from this succession of objects: consequently, there is not, in any single, particular instance of cause and effect, any thing which can suggest the idea of power or necessary connexion” (original spelling; Hume & Millican, 2007; originally published in 1739).

One interpretation of Hume’s theory of causality (here simplified for clarity) is the following:

* We only observe how movement or appearance of object A precedes movement or appearance of object B
* If we experience such a succession sufficient number of times, we’ll develop a feeling of expectation
* This feeling of expectation is the essence of our concept of causality (it’s not about the world, it’s about a feeling we develop).

Hume’s theory of causality

The description of Hume’s theory of causality that we gave above should be taken with a grain of salt. First, Hume presented another definition of causality in his later work “An Enquiry Concerning the Human Understanding” (1758). Second, not all scholars would necessarily precisely agree with our interpretation (e.g. Archie (2005)). All this does not change the fact that the interpretation we presented captures an important way of thinking about causality and we’ll use it as a reference point across the book.

This theory is very interesting from at least two points of view.

First, elements of this theory have high resemblance to a very powerful idea in Psychology called conditioning. Conditioning is a form of learning. There are multiple types of conditioning, but they all rely on a common foundation – namely, association (hence the name for this type of learning – associative learning). In any type of conditioning, we take some event or object (usually called stimulus) and we associate it with some behavior or reaction. Associative learning works across species. You can find it in humans, apes, dogs or cats, but also much simpler organisms like snails (Alexander, Audesirk & Audesirk, 1985).

Conditioning

The way we presented conditioning is largely simplified. If you’re interested in more details, you might want to see a great TED-Ed video by Peggy Andover (https://www.youtube.com/watch?v=H6LEcM0E0i) or search for phrases like classical conditioning vs operant conditioning and names like Ivan Pavlov and Burrhus Skinner, respectively.

Second, most classic machine learning algorithms also work on a basis of association. When we’re training a neural network in a supervised fashion, we’re trying to find a function that maps inputs to the outputs. To do it efficiently, we need to figure out which elements of the input are useful for predicting the output. And... in most cases association is just good enough for this purpose.

Why causality? Ask babies!

Is there anything missing form David Hume’s theory of causation? Although many other philosophers tried to answer this question, we’ll focus on one particularly interesting answer that comes from… human babies.

Interacting with the world

Alison Gopnik is an American child psychologist who studies how babies develop their world models. She works with computer scientists, helping them understand how human babies build common-sense concepts about the external world. Children – to even greater extent than adults – make use of associative learning. Nonetheless, they are also insatiable experimenters. Have you ever seen a parent trying to convince their child to stop throwing around a toy? Some parents tend to interpret this type of behavior as “rude”, “destructive” or “aggressive”, but babies usually have a very different set of motivations. They are running systematic experiments that allow them to understand the laws of physics and rules of social interactions (Gopnik, 2009). Infants as young as 11 months prefer to perform experiments with objects that display unpredictable properties (e.g. can pass through a wall) than with objects that behave predictably (Stahl & Feigenson, 2015). This preference allows them to build more efficient models of the world.

What we can learn from babies is that we’re not limited to observe the world. We can also interact with it. In the context of causal inference these interactions are called interventions and we’ll learn more about them in Chapter 2, Judea Pearl and the Ladder of Causation. Interventions are at the core of the Holy Grail of the scientific method: randomized controlled trial or RCT for short.

Confounding – relationships that are not real

The fact that we can run experiments enhances our palette of possibilities beyond what Hume thought about. This is very powerful! Although experiments cannot solve all of the philosophical problems related to gaining new knowledge (more on this in Chapter 12, Three Families of Causal Discovery Methods), they certainly can solve some of them. A very important aspect of a truly randomized experiment is that it allows us to avoid confounding. Why is it important?

Confounding variable influences two or more other variables and produces a problematic spurious association between them. Such an association is visible from purely statistical point of view like any other association, but does not make sense from the causal point of view. Let’s see an example.

Imagine you work at a research institute and you’re trying to understand causes of people drowning. Your organization provides you with a huge database of socioeconomic variables. You decide to run a regression model over a large set of these variables to predict the number of drownings per day in your area of interest. When you check the results, it turns out that the biggest coefficient you obtained is for daily regional ice cream sales. Interesting! Ice cream usually contains large amounts of sugar, so maybe sugar affects people’s attention or physical performance while they are in the water? This could make a sensible explanation!

That’s true, but let’s try to get some perspective. How about other variables? Did we add enough predictors to the model? What if we added too many predictors?

Adding too many predictors

Adding too many predictors to the model might be harmful from both statistical and causal points of view. Statistical and causal criteria can lead to incongruent feature selection strategies. We will learn more on this topic in Chapter 3, Regression, Observations and Interventions

Do you think that adding one more variable to the model could completely change the outcome? It turns out that this is possible. Let me introduce you to daily average temperature – our confounder. Higher daily temperature makes people more likely to buy ice cream and more likely to go swimming. When there are more people swimming, there are also more accidents. Let’s try to visualize this relationship:

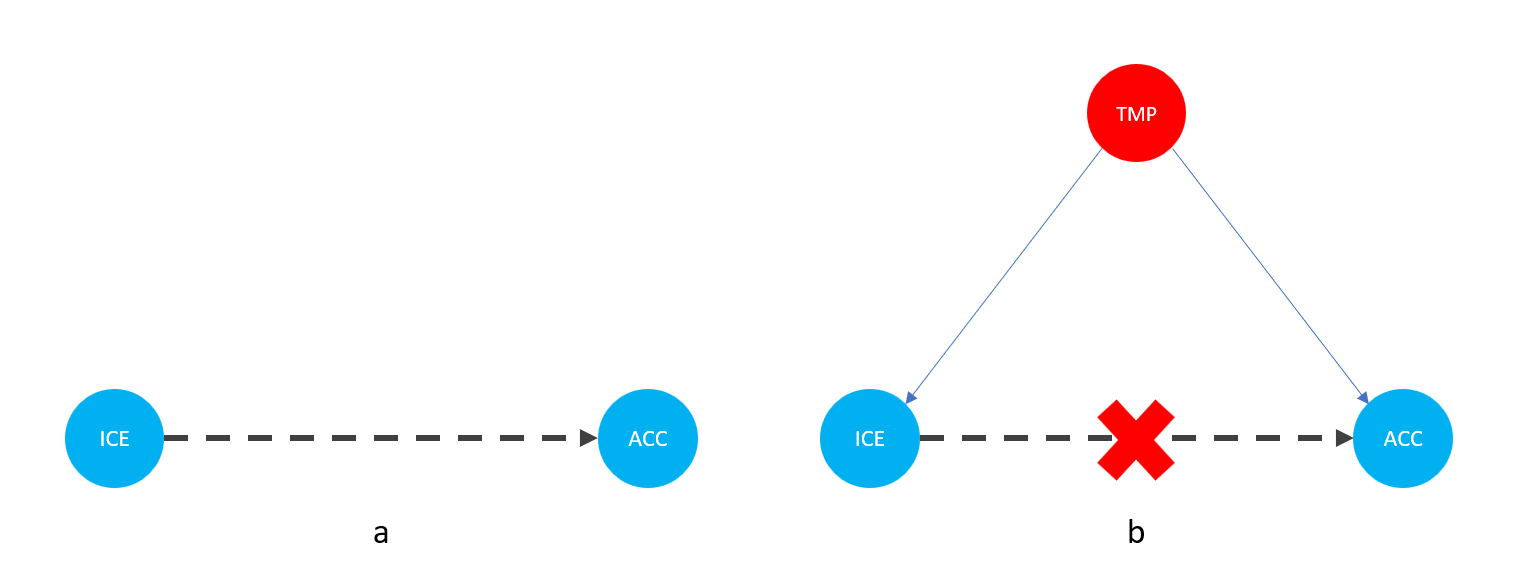


Figure 1.1. Graphical representation of models with two (a) and three variables (b). Dashed lines represent association, solid lines represent causation. ICE = ice cream sales, ACC = number of accidents, TMP = temperature.

In Figure 1.1, we can see that adding average daily temperature to the model removes the relationship between regional ice cream sales and daily drownings. Depending on your background this might or might not be surprising to you. We’ll learn more about the mechanism behind this effect in Chapter 3, Regression, Observations and Interventions.

Before we move further, we need to state one important thing explicitly: confounding is a strictly causal concept. What does it mean? It means that we’re not able to say anything about confounding using purely statistical language. To see this clearly let’s look at Figure 1.2.

Chart, scatter chart

Description automatically generated

Figure 1.2 – Pairwise scatterplots of relations between a, b and c. The code to recreate the preceding plot can be found in the Chapter\_01.ipynb notebook (<https://github.com/PacktPublishing/Causal-Inference-and-Discovery-in-Python/blob/main/Chapter_01.ipynb>).

In Figure 1.2, blue points signify causal relationship while red points signify spurious relationship and variables a, b and c are related in the following way:

b causes a and c

a and c are causally independent

As you can see, non-spurious (blue) and spurious (red) relationships look pretty similar to each other and their correlation coefficients would be similarly large. In practice, most of the time they just cannot be distinguished based on solely statistical criteria and we need causal knowledge to distinguish between them.

Asymmetries and causal discovery

If fact, in some cases we can use asymmetries . Under certain circumstances, they can be leveraged to recover causal structure form observational data. We’ll learn more about this in Part 3, Causal Discovery.

You might be asking now – ok, we said that there are some spurious relationships in our data, we added another variable to the model and it changed the model’s outcome. Nonetheless, I was still able to make useful predictions without this variable. Why would I care if the relationship is spurious or non-spurious? Why would I care if the relationship is causal or not?

How not to lose money… and human lives?

We learned that experiments can help us avoid confounding. Unfortunately, they are not always available. Sometimes experiments can be too costly to perform, unethical or virtually impossible (e.g. running an experiment regarding migration of a large group of people). In this section, we’ll look at a couple of scenarios, where we’re limited to observational data, but we still want to draw causal conclusions. These examples will provide us with a solid foundation for the next chapters.

Marketer’s dilemma

Imagine you are a tech-savvy marketer and you want to effectively allocate your direct marketing budget. How would you approach this task? When allocating the budget for a direct marketing campaign, we’d like to understand what return we can expect if we spend certain amount of money on a given person. In other words, we’re interested in estimating the effect of a certain action on some customer outcome (Gutierrez, Gérardy, 2017). Could we use supervised learning to solve this problem? To answer this question, let’s take a closer look at what exactly we want to predict.

We’re interested in understanding how a given person would react to our content. Let’s encode it in a formula:

Formula\_01\_001

In the preceding formula:

* is the treatment effect for person

* is the outcome for person  when they received the treatment  (in our example: they received marketing content from, us)

* is the outcome for the same person  given they did not receive the treatment

Formula\_01\_002

Formula\_01\_003

Formula\_01\_004

Formula\_01\_005

Formula\_01\_006

Formula\_01\_007

Formula\_01\_008

Formula\_01\_009

What the formula says is that we want to take the person ’s outcome , when this person does not receive treatment and subtract it from the same person’s outcome when they receive treatment .

Formula\_01\_010

Formula\_01\_011

Formula\_01\_012

Formula\_01\_013

An interesting thing here is that to solve this equation, we need to know what is person ’s response under the treatment and under no treatment. In reality, we can never observe the same person under two mutually exclusive conditions at the same time. To solve the equation in formula above we need counterfactuals.

Formula\_01\_014

Counterfactuals are the estimates of how the world would look like if we changed a value of one or more variables, holding everything else constant. Because counterfactuals cannot be observed, the true causal effect is unknown. This is one of the reasons why classic machine learning cannot solve this problem for us. A family of causal techniques usually applied to problems like this is called uplift modeling and we’ll learn more about it in Chapter 9, Causal Inference and Machine Learning – Part I.

Formula\_01\_015

Let’s play doctor!

Let’s take another example. Imagine you’re a doctor. One of your patients, Jennifer, has a rare disease D. Additionally, she was diagnosed with high risk of developing blood clot. You study the information on two most popular drugs for D. Both drugs have virtually identical effectiveness on D, but you’re not sure which drug will be safer for Jennifer, given her diagnosis. You look into the research data presented in Table 1.1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Drug | A | | B | |
| Blood clot | Yes | No | Yes | No |
| Total | 27 | 95 | 23 | 99 |
| Percentage | 22% | 78% | 19% | 81% |

Table 1.1 – Data for drug A and drug B..

The numbers in Table 1.1 represent the number of patients diagnosed with disease D who were administered drug A or drug B. Row 2 (blood clot) gives us information if blood clot was find in patients or not. Note that percentage scores are rounded. Based on this data, which drug would you choose? The answer seems pretty obvious. 81% of patients who received drug B did not develop blood clots. The same was true for only 78% of patients who received drug A. The risk of developing blood clot is around 3% lower for patients receiving drug B comparing to patients receiving drug A.

This looks like a fair answer, but you feel skeptical. You know that blood clot can be very risky and you want to dig deeper. You find more fine-grained data that takes patient’s gender into account. Let’s look at Table 1.2:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Drug | A | | B | |
| Blood clot | Yes | No | Yes | No |
| Female | 24 | 56 | 17 | 25 |
| Male | 3 | 39 | 6 | 74 |
| Total | 27 | 95 | 23 | 99 |
| Percentage | 22% | 78% | 18% | 82% |
| Percentage (F) | 30% | 70% | 40% | 60% |
| Percentage (M) | 7% | 93% | 7.5% | 92.5% |

Table 1.2 – Data for drug A and drug B with gender-specific results added. F = female, M = male. Color coding added for ease of interpretation, with better results marked in green and worse results marked in orange.

Something strange has happened here. We have the same numbers as before and drug B is still preferable for all patients, but it seems that drug A works better for females and for males! Have we just found a medical Schrödinger's cat (https://en.wikipedia.org/wiki/Schr%C3%B6dinger%27s\_cat) that flips the effect of a drug when patient’s gender is observed? If you think that we might have messed up the numbers – don’t believe me, just check the data for yourself. The data can be found in data/ch\_01\_drug\_data.csv (https://github.com/PacktPublishing/Causal-Inference-and-Discovery-in-Python/blob/main/data/ch\_01\_drug\_data.csv).

What we’ve just experienced is called Simpson’s paradox (also known as Yule-Simpson effect). Simpson’s paradox appears when data partitioning (that we can achieve by controlling for additional variable(s) in regression setting) significantly changes the outcome of the analysis. In the real world, there are usually many ways to partition your data. You might ask: ok, so how do I know which partitioning is the correct one? We could try to answer this question from a pure machine learning point of view: perform cross-validated feature selection and pick the variables that contribute significantly to the outcome. This solution is good enough in some settings. For instance, it will work well when we only care about making predictions (rather than decisions) and we know that our production data will be independent and identically distributed (https://en.wikipedia.org/wiki/Independent\_and\_identically\_distributed\_random\_variables). If we want more than this, we’ll need some sort of a (causal) world model.

Wrapping it up

“Let the data speak” is a catchy and powerful slogan, but as we’ve seen earlier data itself is not always enough. It’s worth remembering that in many cases “data cannot speak for themselves” (Hernán, Robins, 2020) and we need more information than just observations.

In this chapter, we learned that we’re not limited to observations as David Hume thought. We can also experiment – just like babies. Unfortunately, experiments are not always available. When this is the case, we can try to use observational data to draw causal conclusion, but data itself is usually not enough for this purpose. We need a causal model to achieve this. In the next chapter, we’ll introduce The Ladder of Causation – a neat metaphor for understanding three levels of causation proposed by Judea Pearl.

References

* Alexander, J. E., Audesirk, T. E., & Audesirk, G. J. (1985). Classical Conditioning in the Pond Snail Lymnaea stagnalis. The American Biology Teacher, 47(5), 295–298. <https://doi.org/10.2307/4448054>
* Archie, L. (2005). Hume's Considered View on Causality. [Preprint] Retrieved from: [http://philsci-archive.pitt.edu/id/eprint/2247](http://philsci-archive.pitt.edu/id/eprint/2247%20) (accessed 2022-04-23).
* Falcon, A. "Aristotle on Causality", The Stanford Encyclopedia of Philosophy (Spring 2022 Edition), Edward N. Zalta (ed.), URL = (<https://plato.stanford.edu/archives/spr2022/entries/aristotle-causality/>). Retrieved 23 April 2022
* Gopnik, A. (2009). The philosophical baby: What children's minds tell us about truth, love, and the meaning of life. New York: Farrar, Straus and Giroux.
* Gutierrez, P., & Gérardy, J. (2017). Causal Inference and Uplift Modelling: A Review of the Literature. Proceedings of The 3rd International Conference on Predictive Applications and APIs in Proceedings of Machine Learning Research, 67, 1-13.
* Hernán M. A., & Robins J. M. (2020). Causal Inference: What If. Boca Raton: Chapman & Hall/CRC.
* Hume, D., & Millican, P. F. (2007). An enquiry concerning human understanding. Oxford: Oxford University Press.
* Kahneman, D. (2011). Thinking, Fast and Slow. Farrar, Straus and Giroux.
* Lorkowski, C. M. <https://iep.utm.edu/hume-causation/> Retrieved 23 April 2022
* Stahl, A. E., & Feigenson, L. (2015). Cognitive development. Observing the unexpected enhances infants' learning and exploration. Science, 348(6230), 91–94. <https://doi.org/10.1126/science.aaa3799>