Chapters 8 to 10

Vladimir Feinberg

December 15, 2018

Notes on Chapters 8 to 10 of Pearl and Mackenzie's *Book of Why* [1].

8 Counterfactuals

Counterfactual reasoning is essential for:

1. Assigning blame (e.g., court room case)
2. Affecting future policy (e.g., climate change) – with an inductive assumption, we can propose policy reducing greenhouse gas emissions by showing that reducing certain types of air pollution would have averted natural disasters.

8.1 Types of Counterfactual Analysis

There are two causal classifications. A cause can be none, one, or both.

1. **Necessary**: A necessarily causes B if without A, B would not have happened
2. **Sufficient**: A sufficiently causes B if A happening results in B happening

Example. Necessary not sufficient cause. See below piano example in Section 8.2.

Example. Sufficient not necessary cause. Consider the firing squad with two soldiers from Chapter 1. If both soldiers fire, then *in that world* neither is necessary (since the other would have killed the prisoner) but both are sufficient.

Why do we need to bother with sufficiency? Consider using a match to burn down a house. Is the arson the match's fault or the oxygen's? Both are necessary, but only one is useful, and that's the one that has stronger sufficient causal strength.

8.2 Directness and Causation

Directness and causation are orthogonal. For instance, Joe can block the fire exit with a large object, causing someone to die. Even though we have a mediator (large object blocking a pathway), we still have sufficient indirect causation.

8.3 A New Language for Counterfactual Quantification

Both necessity and sufficiency of a cause have numerical extensions quantifying to what degree A causes B. Gradation of the degree to which A causes B is essential because A may only increase the probability of B, rather than deterministically inducing it, or other confounders may be involved.

It's too weak to have binary notions of necessary and sufficient:

> Alice attempts to shoot Bob, but misses. Bob, in an attempt to flee, runs under a building where someone is moving in with a piano. The piano falls by accident, killing Bob.

Was the gunshot sufficient to kill Bob? Not really, the gunshot missed and the probability of the piano falling on Bob is low. It would be unreasonable for Alice to plan to kill Bob by scaring him with a gunshot into running under a piano. We wouldn't charge Alice with manslaughter, just attempted manslaughter.

Was the gunshot necessary to kill Bob? This shows that the "necessary" definition is underspecified. In general, Bob can die from natural causes, so a gunshot is not strictly necessary. However, we probably want to hold Bob's general health constant when evaluating whether the gunshot was necessary to kill Bob. In that case, the gunshot was necessary to kill Bob, since Bob only fled due to the gunshot.

This tells us we need a richer language to express what we mean by "strength of a cause" as well as sufficiency and necessity because causality is dependent on what world we're analyzing the causal effect in.

8.4 Expressing Causality Formally

For simplicy, we're querying the causal effect of the indicator X being activated on the indicator Y.

The probability of necessity, PN, is:

$$\text{PN} = \mathbb{P}\left\{Y_{X=0} = 0 \,|\, X = 1, Y = 1\right\} .$$

The probability of sufficiency, PS, is:

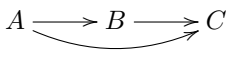
$$\text{PS} = \mathbb{P}\left\{Y_{X=1} = 1 \,|\, X = 0, Y = 0\right\} .$$

What does $\mathbb{P}\left\{Y_{X=a} = b|E\right\}$ mean? Why is it not $\mathbb{P}\left\{Y = b \,|\, \text{do}(X = a), E\right\}$? We intentionally consider a new, "imagined" variable $Y_{X=a}$ that behaves like Y would when $X = a$, but we hold all other so-called exogenous factors, E, constant. This lets us control for exogenous factors that may contradict $X = a$. This construction avoids previous difficulties probabilistic definitions of causality had [2].

This construction is, more formally, a new causal submodel, and avoids the apparent contradiction that the first do-based interpretation has [2]. The submodel considers a copy of the original model's subgraph, but fixes $X = a$ (overriding any setting E gives for X).

9 Mediation

There exist causal paths that are not one edge long:



where A is taking extra melatonin, B is initiating your body's circadian night cycle, and C is falling asleep. The act of taking melatonin itself might be relaxing to people (via the placebo effect), which could help them fall asleep.

In these kinds of situations, the effect of the mediator B is in question, relative to A, in terms of the result, C. This brings us to the general question we might ask about mediators:

1. How much of the root cause is the direct effect or indirect effects? This affects future treatments, and informs, e.g., how much melatonin you should take.

2. Can a mediator with an opposite effect explain some paradoxes away?[1]

The second question has a salient example, with A being gender, B being the department applied to, and C being the admission indicator. This explains the Simpson's paradox in Berkeley admissions, where per-department females are favored, yet school-wide males are. It's fully explained by an inverse mediator effect.

In both cases, however, we have the difficult task of identifying *to what degree* effects of mediation impact a final result.

9.1 Measures of Mediation

The book proposes **controlled** and **natural** measures of mediation. For simplicity, we consider binary indicator variables when estimating the effect of X on Y with a mediator M [3].[2]

For starters, we have the total effect of the treatment:

$$\text{TE} = \mathbb{E}\left[Y \,|\, \text{do}(X = 1)\right] - \mathbb{E}\left[Y \,|\, \text{do}(X = 0)\right]$$

Then we have the effect due directly to the treatment

$$\begin{aligned}\text{CDE}(m) &= \mathbb{P}\left\{Y = 1 \,|\, \text{do}(X = 1), \text{do}(M = m)\right\} - \mathbb{P}\left\{Y = 1 \,|\, \text{do}(X = 0), \text{do}(M = m)\right\} \\ \text{NDE} &= \mathbb{P}\left\{Y_{M|X=0} = 1 \,\middle|\, \text{do}(X = 1)\right\} - \mathbb{P}\left\{Y_{M|X=0} = 0 \,\middle|\, \text{do}(X = 0)\right\}\end{aligned}$$

The CDE controls M and asks how X increases the probability of Y. The NDE considers how much the treatment affects the outcome, had the mediator $M|X = 0$ behaved as it would without the treatment.

$$\text{NIE} = \mathbb{P}\left\{Y_{M|X=1} = 1 \,\middle|\, \text{do}(X = 0)\right\} - \mathbb{P}\left\{Y_{M|X=0} = 0 \,\middle|\, \text{do}(X = 0)\right\}$$

The net indirect effect measures what the impact of the treatment X is through the mediator M. Note by its definition, we can't have a controlled indirect effect (we're controlling the thing that we would be having an effect through).

[1] Yet another question, related to mediation, which may explain $A \to C$ effects *through* proposed intermediary B effects, is whether we can identify such effects, and what are the accuracies of models missing such intermediaries? When are we done finding mediators?

[2] This chapter helps assess proposed mediators versus their lack of mediators, but there's a more metaphiscally dubious question at hand here where we wonder "where are we done?" Why is the model $X \to Y$ vs $X \to A \to Y$? Why can't it be $X \to A \to B \to Y$ or $X \to A \to B \to C \to Y$? When do you stop being fine-grained enough?

One nice property is that TE = NIE + NDE.

10 Conclusion

Pearl concludes with three notes on modern AI and why we need to apply more of his research there.

10.1 Current Approaches Importantly Miss Causality

TODO(vlad17): review discussion in ch10 about this, deep learning, RL.

10.2 Causality Addresses Distribution Shift

Handling distribution shift is essential to strengthening AI because it improves the ability to generalize by enabling the agent to learn how to deal with new environments.

Transportability Fig. 10.1, 10.2. TODO(vlad17): find the associated paper and write this up – how do we robustly estimate $P(W|X)$?

footnote: vs. hindsight experience replay, how is that not this.

10.3 AI Safety

TODO(vlad17): flesh out high-level points based on interpretability

A Personal Note

Note this section is dedicated to my general afterthoughts, not the original work.

TODO(vlad17): true automation through proposition of causal structure.

References

[1] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.

[2] Judea Pearl. Probabilities of causation: three counterfactual interpretations and their identification. *Synthese*, 121(1-2):93–149, 1999.

[3] Judea Pearl. Interpretation and identification of causal mediation. *Psychological methods*, 19(4):459, 2014.