Chapters 8 to 10

Vladimir Feinberg

December 2, 2018

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

The probability of necessity, PN, is:

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

The probability of sufficiency, PS, is:

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

What does $\mathbb{P}\left\{Y_{X=a} = b|E\right\}$ mean? Why is it not $\mathbb{P}\left\{Y = b \,|\, \mathrm{do}(X = a), E\right\}$? $Y_{X=a}$ is the variable Y after we update our belief about the nature of Y and its relationship to X given that we observe E had occured (Pearl calls this "hindsight"). The original paper gives a full elaboration: in particular, the do version calculates the probability that $Y = b$ removing all arrows into X in the original causal model. $Y_{X=a} = b$ considers the causal *submodel* (literally the subgraph of the original causal model with a fixed to X), and uses the induced conditional probability tables from the subgraph to compute the behavior of Y, but with respect to the missing nodes taking on values from the original causal model (letting us ask questions that seemingly contradict themselves like $\mathbb{P}\left\{Y_{X=1} = 1, X = 0\right\}$). [1].

TODO(vlad17): expand/distill above paragraph, add a diagram constructed base on the above paper.

TODO(vlad17): add a bibtex entry for the paper above and add bibtex to the compilation process.

9 Mediation

There exist causal paths that are not one edge long (TODO(vlad17): example, with picture; graphxy?).

Wherever this happens a natural question how much of the root cause is the direct effect or indirect effects? This affects future treatments.

(TODO(vlad17): draw a picture; use graphxy?). Image description: The mediator of "Department you apply to" explains Simpson' paradox because gender affects the department applied to and both of those affect the admissions outcome. In this case, the direct and indirect effects have a reverse orientation.

TODO(vlad17): aligned confounders: smoking gene and smoking (cite corresponding paper mentioned by Pearl, mention effect sizes.

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.[1]

TODO(vlad17): draw the chain

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

The CDE controls M and asks how X increases the probability of Y. The NDE considers how M changes naturally due to changes in X, and lets those affect the probability of $Y_{M=m} = 0$, but only keeping M constant in a counterfactual.

TODO(vlad17): look for an example distinguishing the two in the book

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

TODO(vlad17): NDE? Look for an example distinguishing the two in the book

Alternatively consider (and ref for generality the summary paper)

9.2 Total Effect

CDE + CIE does not equal total effect, but NDE - NIE does! TODO look at an example from the book or paper.

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.

[1] 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?

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