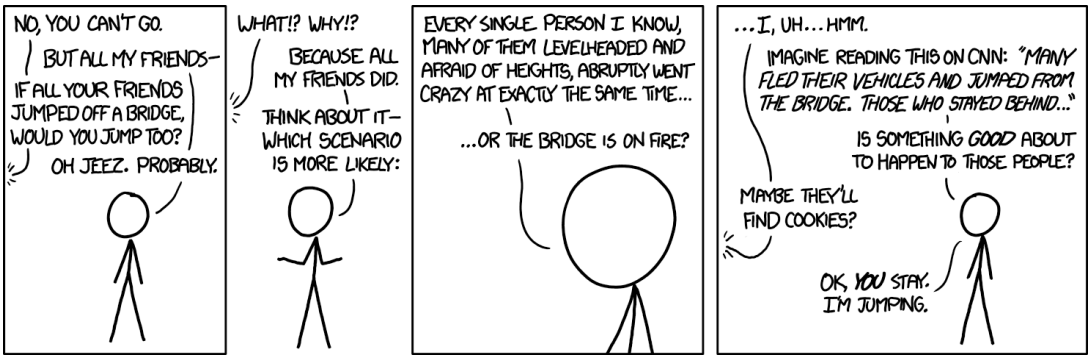
Counterfactuals for explainable machine learning

James Liley, Louis Aslett: {james.liley,louis.aslett}@durham.ac.uk



If your friends all jumped off a bridge, would you?

The question which Cueball (the visible character) and the unseen friend are debating suggests an interesting dilemma. If we take as read that our friends jumped off a bridge (let's call this the event '$J = 1$'), do we mean:

1. that we are in a possible world in which the event $J = 1$ occurs naturally; the bridge is on fire, or something similar?

2. that we are in exactly the world we are in now, but for the fact that our friends jumped off?

This question is at the core of a distinction between *conditionals* and *counterfactuals*, and a lot of argument and debate comes from not distinguishing which we mean.

It is quite difficult to phrase the difference using the standard language of probability theory. Calling the event that you jump off the bridge '$Y = 1$', we would usually write the probability expressed in the first interpretation as just the standard conditional, but the second a little differently:

$$1: \ P(Y = 1|J = 1); \qquad 2: \ P_{J\leftarrow 1}(Y)$$

How do we differentiate these formally? We are interested here in *causes*. Did your friends jumping off the bridge *cause* you to do so? Or did something else cause both?

A rich field of contemporary statistics concerns description and evaluation of such probabilities. This project will focus on differentiating these scenarios in order to glean meaningful information from real-world observations.

Causal graphs and causal mechanisms

A standard starting point in such analysis is a *causal graph*. Loosely, this concerns a set of vertices representing random variables joined acyclically by arrows such that an arrow joins A to B if A has a causal influence on B.

For instance, suppose we consider the medical data of a given person, which we model using the random variables:

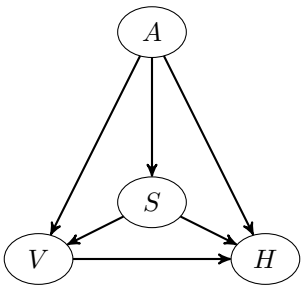
A : Age

S : Smoking status; 1 for currently smoking; 0 for not

V : Presence of vascular disease; 1 for disease, 0 for not

H : An indicator variable: 1 if the individual has a heart attack in the coming year, 0 if not.

We may consider the following causal graph:



which indicates that age causally affects smoking risk (age causes individuals to take up smoking), vascular disease risk, and heart attack risk; and that smoking affects heart attack risk both directly and through affecting risk of vascular disease.

A causal graph essentially defines a probability distribution over the variables it represents, but imbues them with an extra structure: it tells us what happens when we *perturb* the graph in various ways.

Causality and counterfactuals in machine learning

We might be interested in a probability like:

$$P(H = 1|A = 76, S = 1, V = 0)$$

that is, the probability that an individual who is 76 and smokes but does not have vascular disease has a heart attack in the coming year.

However, if we are that individual, that is not a great deal of help. We might want to know how our risk would change if we stopped smoking. But here we hit a trap: this is not just:

$$P(H = 1|A = 76, S = 0, V = 0)$$

since this describes the set of individuals who are 76, do not have vascular disease, and *already* do not smoke: who are probably healthier in other ways. What we actually want is:

$$P_{S\leftarrow 0}(H = 1|A = 76, S = 1, V = 0)$$

that is (analogously to the bridge) what happens if we *start in exactly the world we are now*, but *force* the value of S to be 0. To evaluate this probability we need a causal graph.

For essentially this reason, counterfactuals are a useful way to interpret predictive models (and other complex functions) in that they say *what might happen if we were to change something.*

Potential projects

In this project you will firstly give a clear mathematical exposition of the causal objects with which you will be dealing. After that, the project is more open. Several options include:

- Implementing and interpreting counterfactual explanations in a real dataset.
- A survey of recent research, and an analysis of how these ideas can be applied
- Examining the idea of 'counterfactual fairness': evaluating whether a machine learning model is 'fair' in a sense using these methods
- Looking at the method of 'Mendelian randomisation' in genetic datasets, and various mathematical ideas around this

Mode of operation and evidence of learning

The project will firstly involve coming to understand causal models through reading a combination of theoretical and applied papers. After choosing an idea, students will then develop that idea using simulation and potentially real-world examples. There will be an expectation that students give a strong exposition of causal models and counterfactual reasoning, and understand the ideas to the point of being able to implement them in code.

Recommended prerequisites and corequisites

This project would be suited to students in statistics or probability with an interest in understanding causal modelling. A *keenness to learn about these ideas* is more important than already being good at them!

There will be plenty of scope for programming, and a *strong* need for well-worked examples. Either Python or R would be fine for this.

Pre-requisities

Essential:	Probability II
Recommended:	Data Science and Statistical Modelling II, Statistical Inference II, Statistical Modelling III

Co-requisities

Essential:	None
Recommended:	High dimensional statistics IV

Recommended resources

1. Judea Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, 2003. pp. 467470. ISBN 0-521-59271-2. MR 1992316. A standard reference (and the original) for causal models
2. Christof Molnar. *Interpretable Machine Learning.* Online here. An excellent practical overview of many methods in explainable machine learning.
3. See general discussion on Mendelian randomisation and counterfactual fairness.

Final word

Please feel free to e-mail us with any questions.