Learning how to act: making good decisions with machine learning

Finnian Lattimore

April 28, 2017

In vain the grave, with retrospective Eye,
Would from the apparent what conclude
the why, Infer the Motive from the Deed,
and show That what we chanced, was
what we meant, to do.

Alexander Pope

Chapter 1

Random Notes

- With enough variables it becomes almost a given that we will be able to classify observations by what period they were collected in (eg detecting covariate shift). This means we can conclude the model will fail (when in most cases it will not). Is a form of regularization/reparameterization with an penalty on covariate shift the solution? Focus detection of covariate shift to those variables that actually influence the prediction.
- If variables are (conditionally) independent they are unlikely to be directly causally linked.

Introduction (3500 words)

Motivation

Many of the most important questions in science and in our personal lives are about the outcomes of doing something. Will asking people to pay upfront at the doctors reduce long term health expenditure? If we developed a drug to suppress particular genes, could we cure MS and would delaying teen-aged pregnancies improve the outcome for their kids.

These are hard questions because they require more than identifying a pattern in data. Correlation is not causation. Causal inference has proven so difficult using standard methods (such as instrumental variables and propensity scoring) that there is barely any consensus on even enduring questions like the returns to education or the long-term consequences of early life events – like teenage pregnancy, where the variables involved are susceptible to human intuition and understanding.

We now live in a world of data. Hours of our lives are spent online, where every click can be recorded, tiny computers and sensors are cheap enough to incorporate into everything and the US Institute of Health is considering if all infants should be genetically sequenced at birth. Such data gives us a window into many aspects of our lives at an unprecedented scale and detail but it is messy, complicated and often generated as a by product of some other purpose. It does not come from the controlled world of a randomised experiment.

Traditional techniques that assume linearity or substantial prior knowledge of causal structure are poorly suited to data sets that may be high dimensional, have non-linear relationships between variables, and where we have limited theory specifying how the variables are related.

Statistical machine learning has been very successful, but generalizing remains very challenging.

How much data do humans get?

Big data and machine learning are a huge deal

Data age Exciting times, larger data sets. Hype, opportunities for Machine learnig. Algorithms beign incorporated everywhere.

The end of theory (it is not) correlation is not enough no matter how many petabytes sit behind it. There are two seperate issues with a correlation only one of which is addressed by larger datasets. One is noise - it may have happened by chance. The other is much more fundamental.

Large datasets have allowed us to relax assumptions We can fit more flexible models, rely less on strong parametric assumptions about the form of the curve.

Theory is what allows you to extrapolate outside the observed domain (Hal Varian)

A cautionary tale Much hard work remains to be done before we can call and end to theory or the for the automation of science.

There is nothing special about humans. We do have some limited capacity to experiment - but can (or believe we can) draw inferenes about the likely role of variables we have never directy manipulated ourselves. Current ML algorithms are limited. If we ignore this, we will pay a price.

What is causality and why do we care? (2000 words)

To explain or to predict [**?**] The two cultures [**?**]

Defining causality

- widely debated in science and philosophy (FIND SOME REFERENCES)
- what is explaination?
- any model that aims to predict the outcome of an action or intervention in a system
- I do not see the distinction between explaination and (causal) prediction. Explaination is all about the ability to compress and to generalize. The more a model can do this, the more we view as providing an understanding of the why.
- mediation?

Identifying when we have a causal problem

Examples of typical machine learning probems. Are they causal?

- Speech recognition (for systems like Siri or Google)
- Machine translation
- Image classification
- Forecasting the weather
- Playing Go
- Identiying spam emails
- Automated essay marking
- Predicting the risk of death in patients with pnumonia.
- Predicting who will re-offend on release from prison

- Predicting which customers will cease to be your customers
- Demand prediction for inventory control
- Predicting who will click on an ad
- Financial trading
- Recommending movies
- Online search
- Self driving cars
- Pricing insurance

For image recognition, we do not particularly care about building a strong model for exactly how the thing that was photographed translates to the image we see. In fact this is because we already have one. We can be confident that the process will not change. If we develop a discriminative model that is hightly accurate at classifying cats from dogs, we do not care a lot about its internal workings (provided we have strong grounds to believe that the situations in which we will be using our model will match those under which it was trained). Covariate shift clearly comes in here. Because there are areas where mechanisms are understood it is relatively easy to argue that covariate shift is not occuring and that results will be transferable. The mechanism is known but the function may be complex.

What aspects of a problem determine if causal inference is required?

(When is pure prediction useful?)

- To decide between actions we only need to rank them (not estimate their actual effect).
- The predicted outcome in the absence of an intervention provides a single point. We can use this to find which problems are most serious if left alone - and priortise those for modelling changes.
- Any decision we take does not significanly impact the system from which the data was drawn to make it (for repeat decision making)
- Does acting on the result of the prediction change the predictive distribution p(y—x)? Ie change people's behaviour.
- Ethics - ... People's viewpoint on if its ok...

Approaches to causality (1000 words)

Two broad approaches

- Build a model to map the natural behaviour of the system to what will happen for some action
- Take the action and see what happens

The first is causal inference

The second is reinforcement learning

Both generalize from randomized experiment Reinforcment learning to sequential decisions, causal inference to non-experimental conditions

Both approaches involve assumptions the latter that we can group context and actions.

Limitations of causal inference

Limitations of experimets What are the issues with standard randomized experiments?

Causal models (3000 words)

Causal inference aims to infer the outcome of an intervention in a system from data obtained by observing (but not intervening in) the system. To do this we need terminology to describe actions and how we anticipate the system should respond to them. Three key approaches have emerged; counterfactuals, structural equation models and causal bayesian networks. In this chapter we describe these approaches, examine what problems they allow us to solve and the assumptions they rely on and discuss their differences.

Counterfactuals

Counterfactuals (or potential outcomes) are a way of describing distributions under different actions that were developed from the starting point of generalizing randomized experiments.

There are philosophical objections (references) to counterfactuals because of the way the describe alternate universes that were never realized and are not empirically testable (example).

For interventional queries, of the form; what is the is the probaiblity distribution for the variable Y if we intervene to set X = x, and the system is otherwise unchanged? Counterfactuals are a short hand. They say what is the distribution of Y had X = x (regardless of the value X actually took).

Causal graphical models

Although seemingly simplistic, the notion of hard interventions is surprisingly powerful.

A complaint leveled against this view point of causality is that the 'surgery' is too precise and that, in the real world, any intervention will effect many variables (eg Cartwright 2007). However,

Provides an explicit mechinism to map knowelede from one setting to another.

A fully observed causal bayesian network allows asymtoptic point estimates of the causal effect of any intervention (assuming positive density).

Structural Equation models

Comparing and unifying the models

Representation equivelent for interventional queries This means it is straightforward to take the best elements of work done in any of the frameworks. For example, draw a graphical network to determine if a problem is identifiable and which variables we need to adjust for to obtain an unbiased causal estimate. Then use propensity scores or ... to estimate that effect.

Alternatively, make parametric assumptions, to make the problem into a structural equation model.

SWIGs [?] [?] Causal inference without counterfactuals

Determinisim inherent in counterfactuals and SEMs.

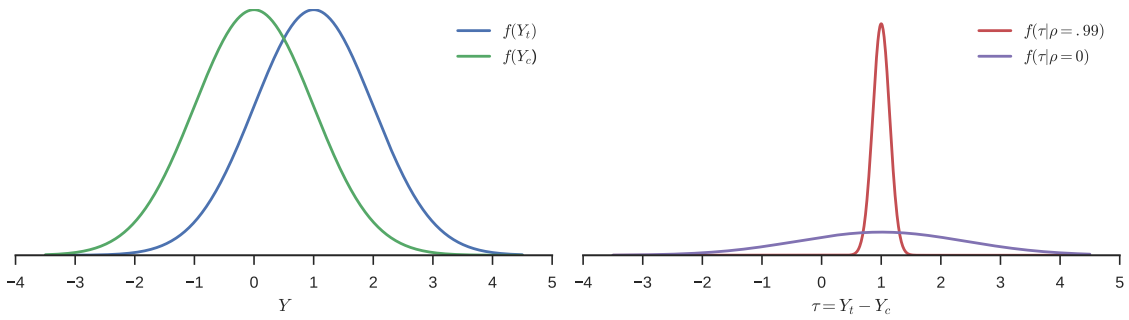
Key issue is that we can't observe the joint distribution over Y_t and Y_c even with randomized experiment.

Suppose the counterfactuals Y_t and Y_c are jointly normal (with equal variance),

$$P(Y_t, Y_c) \sim N()$$

then their difference is also normal,

$$P(\tau) = N(\mu_t - \mu_c, 2\sigma^2(1-\rho))$$



The variance of τ is not identifiable, even from experimental data. But it seems on the face of it relevent to our decision making. If the example above, if $\rho = 1$ then almost everyone benifits slightly from the treatment. However if $\rho = 0$, there is a wide range, with some people benifiting a lot and others suffering significant harm.

Can these issues be resolved by considering personalized causal effects as random variables and avoiding counterfactuals?

Counterfactuals make a natural way to talk about causal effects in natural language. They could be seen as a short hand for projecting the world forward, holding everything else constant. An (impresise) way of expressing that we wish to consider the effect of only a single change.

What assumptions are required for the kind of counterfactual analysis like X would have been higher had B ...?

What does a causal model give us? Resolving Simpson's paradox

A translator from graphical independence to counter factual statements

Defining causal effects

Summarising the difference between two distributions. There is no one answer. [?]

Two key questions (5000 words)

We can rougly categorize the problems studied within causal inference into two groups, causal effect estimation and causal discovery. In causal effect estimation we assume (at least implicitly) that key aspects of the causal graph are known. The goal is then to estimate the causal effect of some action or range of actions. WHERE DOES MEDIATION FIT IN? THIS IS ALREADY HUGE, and is central to 1000 of papers published each year. Causal discovery aims to leverage much broader assumptions to learn the stucture of causal graph from data. THIS IS THE AUTOMATION OF SCIENCE.

Causal Inference

You are willing to assume the causal graph

extremely widely appliled Implicitly accounts 10,000 of studies in phycology, medicine, business, etc.

Identifiability

Definition:asymptopic point estimate is possible, without parametric assumptions

Under what conditions in is the problem solveable

The Do Calculus

Estimation

[**?**] Review of non-parametric estimation

How well do we actually do with finite datasets?

When is regression causal?

Non identifiable queries

Parametric assumptions

eg linearity

Bounds

Instrumental variables are an example

Causal Discovery

It is possible to infer some aspects of causal structure with very general assumptions. The set of conditional independences in a non-experimental data set indicates some causal structures are more likely than others. In addition, there are subtle asymmetries in the relationship between the joint distribution of cause and effect and the distributions of cause given effect and effect given cause. These clues are the key to causal discovery algorithms, which attempt to learn causal structure from non-experimental data.

You want to learn the graph

Equates to the aim of automating scientific discovery

Increadibly hard

Methods can also be divided into contraint based and search and score

Discovery with Conditional independence

If variables are (conditionally) independent they are unlikely to be directly causally linked.

Discovery with Functional Models

Granger causality

The interventionalist viewpoint (5000 words)

Instead of trying to infer the outcome of an intervention in a system from passive observations we could just try the intervention and see what happens.

Randomization is not a requirement. -¿ control over treatment assignment is.

Experiments, bandits, reinforcement learning

internal validity Are differences in treatment and control groups down to intervention or the result of bias?

external validity Are the results of the study applicable to the broader population of interest.

Fundamental role of randomization

Feature selection Is it even more important here? How do algorithms degrade as irrelevent features are added? In the supervised vs the RL setting? What would be a fair comparison?

Image of randomization breaking any confounding links

Randomization does not ensure target and control group are exactly alike. The more features or variables you include, the more likely that there will be a significant difference across at least one variable. But the within group variance also increases, the net effect is that it becomes harder to draw a conclusion but not more biased.

Practical limitations of experimentation Failure to generalize. Transportability, imperfect compliance. Too many contexts.

What is the role of randomization? How do bandits algorithms work despite being only partially randomized? What else can you do to improve randomized studies (variance reduction, lower regret).

Multi armed bandits

Motivation Why we need to extend from simple randomized experiment.

Regret

Relationship to Markov decision processes

The exploration/exploitation trade-off

Contexual bandits

Dynamic Systems

- An explicit model of actions in a partially known system (eg HMM)
- Feynman-Kac Lemma; Solving a PDE can be converted to a stochastic process

Causal Bandits: Unifying the approaches (5000 words)

Learning from log data

Causal Inference & Machine Learning (5000 words)

A more detailed discussion on where causal inference sits within machine learning and what it can offer.

Different approaches

The two cultures

Translating terminology

Economist -¿ ML

Challenges for the Machine learning approach

No cross validation

The challenge of model selection

Does predictive accuracy indicate a good causal model?

Less large, real world datasets

The dearth of experimental data

Data for testing causal models

Simulators. Open competitions. Converting other data sets. Existing data sets.

Are models that predict better more likely to be the correct model? For polynomials, if you know the data was generated by a higher order polynomial, you may still be better of predicting it with a lower order one.

Relation to other areas of ML

Covariate shift

Generalizability

Invariants

more stable Variables causally directly causally related to the outcome (either causes or effects) should be more stable predictors over time. The assumption is there are less places for change to come in.

change input distribution If a feature is a cause of an outcome then changing the input distribution over that feature won't break the model. If its an effect it could.

feature selection The direct causes (and effects) of a variable of interest make up a sufficient set for prediction (is this true)? This may be a reason for using structure learning type algorithms even if you are simply doing prediction.

Interpretabilty

Interpretable models as proxies for causal models Let the human do the work. If we know the training and test data will be sampled from different distributions, knowing what the features that the model is looking at are, allows people use their background understanding of the world to evaluate whether or not those features are likely to be transferable to the test domain.

Causal models are more interpretable? A desire for interpretability indicates that something has been left out of the loss function.

One form of interpretability gives people insight into what the features are that the model is relying on.

Specifically, people can

- rule out many possible features as highly unlikely to be relevent to a problem

People have access to a lot of detailed prior knowledge.

http://www.news.com.au/finance/money/costs/insurance-companies-secrets-spilt/news-story/f6ef17ae73e3a56
insurance decisions voodoo because of lack of transparancy and absense of obvious causal link.
Claim: People are more comfortable with desicion making on the basis of factors they believe to be causally relevet.

Transfer Learning

find a feature representation in which $P(Y|X)$ is the same in many different domains (or stable over time). Causal models predict the outcome of actions. We could directly take these actions and learn $P(Y|a, X)$ for every (a, X) but, in reality, no two situations (or actions) are exactly alike. So we have to make representations such that things are stable.

This is tightly related to generalizability. If we take a person undergoing a medical test, we might describe the situation by the year and location, the person's age, gender, heart rate, medical condition and test results. We don't include , the color of the doctors shirt, the size of the room, ...

For example, in the advertising setting, we want to know how our on expenditure on paid search ads is linked to sales. However, this relationship may be very unstable over time because the ad slots are sold at auction. The amount we have to pay to obtain a given position for keyword depends crucially on the amount our competitors are bidding for that keyword. However, the relationship between displaying the ad at a particular position and the probability that someone clicks it an then makes a purchase may be much more consistent.

Fair Machine Learning

Exchangeability of people Is there any connection to the concept of imagining you could be born in any position and designing ethics from there?

- what determines what attributes we deam worthy of protection (eg race, gender, sexuality)
- what do we protect (and what exceptions do we give)?
- Do we think it is bad to select on the basis of protected attributes not directly related to desired outcome?
- Do we think it is bad to select in a way that has a disparate impact on a minority/disadvantaged group.

Causal inference with Bayesian Graphical models

The likelyhood principle

Look at some examples from the stan book.

Both frequentist and bayesian methods can be used to approach problems of infering the outcome of an intervention. However, Bayesian modelling allows you to very naturally encode aditional assumptions and causal prior knowledge. This seems especially critical for problems that are not non-parametrically identifiable.

In the same way that regression may be either causal or non causal, depending on what interpretation and assumptions the researcher places on the equations relating the variables, Bayesian models may or may not be interpretable causally.

Is setting certain priors equivelent to conditioning? How must the model be set up to ensure unbiased causal estimates (in what way are the assumptions different to if you are using bayesian estimation techniques for prediction?)

Conclusions (1000 words)

Open questions

Cycles - a huge issue. Not covered by Pearl,Rubin etc.

Places to look, statistical control theory, etc. any interesting papers along these lines?