Causality, Ethics and Identifiability

Finnian Lattimore

December 10, 2016

1 What, Why, When Where, How? (Finale Doschi-Velez)

Interpretability - What Why When Where How (Finale Doschi-Velez)

- What: To give or provide meaning to; explain [to humans] (dictionary.com)
- Related: transparent, accountable, trustworthy, fair, actionable, reliable

Why,When,Where needed?

- Science (why) - understand 'nature of the thing'
- debugging
- safety - decisions are sound, what if inputs are wrong
- miss-matched objectives - eg side-effects of medication not included in objective
- legal/ethics
- unknown, unknowns - something we can't properly quantify in model plus humans involved in decision making.

How (to evaluate)?

- need scientific way of measuring interpret-ability
- indirect measurement in terms of goal - eg interpret-ability in context of application success measured in terms of some specific application outcome - like increased use of model, better patient welfare, etc.
- measure interpret-ability in terms of proxies like sparsity, 'size of model', 'additive', non-negativity.
- Test how well explanation worked by testing how well people can predict the output of a model for a given input (simulation)
- Test how well explanation worked by testing how well people can answer counter-factual queries of the form: - What changes to input, weight, cost would change [prediction for X]

Where are the unknown unknowns?

- in the structure of the data (eg science)
- in reward function (eg missmatched objective)
- inputs (ie safety checks)
- internal features (ie debugging)

What kind of interpret-ability do we want?

- Global:What factors important for decision making in general

- Local: What factors lead to a particular decision.

1.1 Feature Importance

Instance based approach - compute conditional expected value of classifier output given each pixel individually. (Empirical approach).

Unification of model interpret-ability SHAP (read this paper for sure).

2 Finn Random ideas

There is a hierarchy of ethical/philosophical difficulty relating to disparate impact.

1. The disparate impact is due to biased data collection (solution fix the data)
2. The disparate impact is due to historical/short term causative factors (eg race $\rightarrow$ disadvantage)
3. The disparate impact is due to fundamental/long term causative factors (eg women $\rightarrow$ pregnancy)

Are simple models (like logistic regression) really interpretable? Not in the presence of unobserved confounding where we have an objective miss-match.

A thing to do

1. enumerate desirable properties of 'fair' algorithms
2. quantify trade-offs between these properties
3. connect these properties for underlying reasons for desiring fair algorithms

What determines which attributes we consider protected? eg why don't we protect things like intelligence?

- Not a choice
- Generally not 'fundamentally' relevant
- Easily measurable/observable

Justifications for interpretability

- better integration with human decision making
- improved generalization
- 'debuging'
- ability to incorporate additional constraints without explicitly defining loss function
- better adherence

Fundamentally causal relationships - those which we expect to be invariant (the system cannot change to remove the relationship).

The closer a variable is to being a direct cause - the more stable that relationship should be over time (as other things can vary) - also the better the relationship should generalize across situations (eg different countries).

The idealised scenario, under which the target variable and optimisation metric truly capture the desired outcome and the training data is sampled iid from the population of interest can almost never be expected to hold.

Claim: Many of the scenarios motivating interpretable models really require casual models. This is an important distinction to make. Although historically, causal modelling is generally based on interpretable models such as linear or logistic regression, there is no fundamental requirement that this be the case. There has been substantial work on estimating causal effects with complex non linear models, eg .

Some of the major risks highlighted with the use of big data and machine learning algorithms for critical decision making are the result of apply non causal algorithms to fundamentally causal problems.

Causality mentioned as a motivating factor for interpretable models **?**

Interpretable model not consistently defined because they are motivated by a number of different considerations & applications. We should first ask why we want an interpretable model then construct a definition that is compatible with our goals and finally develop a solution.

? consider including a variable indicating membership of a protected class in a model as formal discrimination on the basis of disparate treatment, even in the absence of a discriminatory effect.

data mining may require us to re-evaluate why and whether we care about not discriminating **?**

role of randomisation and experiment

Protected class membership is a proxy for variables which are "genuinely relevant in making rational and well- informed decisions" **?**

Other potential issues with use of ML/Big Data

- Lack of a 2nd chance. If everyone has the same picture of you, through a federation of many data sources then failure in one domain may lead to rejection and failure in many others. (at an individual level)
- Overly homogeneous decision making (at a group level). If everyone is using the same data and algorithms, then we could reinforce current stereotypes and remove opportunities for gradual change. For example, suppose theoretically

Problems motivating interpretable models

- Models are deployed where their use alters the environment, invalidating future predictions
- Trust that a model will perform well with respect to real (unmeasurable/hard to quantify) objectives
- Adversarial settings.
- Participatory - critical decisions otherwise made by a small technical elite.
- Understand, validate, edit trust **?**
- "trust, which is fundamental if one plans to take action based on a prediction" **?**
- trust so that a user will use the model.
- real world data different (to that used to train/validate)
- evaluation function used doesn't reflect true goal.

? Problems with ML: propagation of feedback loops, failing to notice data leaks, over-estimating model accuracy.

? define a causal model as one which can predict the outcome of an intervention.

Strong arguments for building an interpretable model, at least in an early phase to diagnose potential problems. Humans can evaluate for which features P(Y—X) is unlikely to change. People can (at least sometimes) identify features that will not lead to good generalization.

It is important to distinguish 'trust' required for no other reason than to make the user use the model versus 'trust' associated with properties of the model that make it more accurate (ie via being more robust to changes in environment).

In the former case, one valid solution would be to train the users to accept machine decisions or hire less questioning people. However, if the fundamental issue is a miss-match between the real world problem and the quantity being optimised by the machine learning algorithm, this would be disastrous. More subtlety, solutions that lead users to place a greater trust in the model that do not alleviate any fundamental mis-match are similarly bad.

Note: related to. Shown that the pattern of likes on facebook can significantly predict sensitive aspects.

Properties of causal models

- may be more easily contrasted or combined with human reasoning.
- come with theory on transferability
- are explicitly designed to predict the outcome of an intervention
- can remove conflict between transparency and continued effectiveness. In some situations, making a model transparent is directly in conflict with its predictive accuracy.

Causality is frequently mentioned in conjunction with interpretebility , however the relationship between them remains unclear.

Despite recent work in the area , there is still not a unified motivation for or definition of interpretability.

The extent to which interpretability is plausible or genuinely desirable is still under debate in the ML community .

There is growing concern in the wider community over the use of automated or machine decision systems and a strong desire for transparency and intpretability. A particularly pertinant example is the introduction of legislation in the European union that grants a right to XXX. This is due to come into effect in 2018 and has major ramifications for the application of machine learning. There are components of this relating to explaination and discrimination.

Demonstrate the need to increase the clarity around issues such as algorithmic discrimination and interpretability.

? raise the interesting point that interpretability can only be important if we care about more than simply high predictive accuracy. In other words a desire for interpretability implies a miss-match between the real world goal and the optimisation problem presented to a machine learning algorithm.

If a model is non-causal making it transparent can lead to changes in people's behaviour that reduce the predictive accuracy of the model. This creates a trade-off between transparency and the utility of the model that remains even without imposing any requirements for model simplicity. If the model is causal, then any changes in the choices people make will be reflected in the predictions, removing this conflict. For example, if students know exactly what factors lead to a high score from an automated essay marking system they are strongly motivated to change their writing to reflect this. If these factors are direct causes of the desired outcome (clear and compelling writing) this is a good outcome. However, if they are merely associated with good writing then we will end up with impressive sounding gobbledygook. As a more serious example, consider an algorithm used to help make parole decisions that looks at prisoners participation in various programs to assesses the risk of recidivism. Suppose the algorithm determines that book clubs are associated with lower risk, whilst sex offender programs are associated with higher risk. If the model is made transparent to prisoners, then we expect to see the participation in sex offender programs plummet and a strong demand for book clubs. Again if the effects are causal, perhaps the book club fundamentally changes peoples outlook on life and the sex offender programs were poorly designed, this is a positive outcome. However it might be that participation in sex offender programs actually lowers the risk of re-offending in those who take them but that the group who currently participate are simply higher risk than the general prison population. Similarly, the majority of the apparent benefit of the book club could be a function of who opts to take part.

Note: the causal attribute needs to be sufficiently measurable. Proxies may still be subject to manipulation.