Transparancy, Causality & Fairness

Finnian Lattimore
Australian National University
finnlattimore@gmail.com

Introduction

The rise of machine learning and big data is accompanied by increasing concerns over transparency and fairness [1]. A recent EU directive requires automated decision systems used to profile individuals must provide *"meaningful information about the logic involved."* and *"shall not be based on special categories of personal data"* [3]

Fairness & Discrimination

Consider a binary outcome of interest Y, protected variable X, other covariates Z and a model f that outputs the probability of a positive Y for each individual.

1. *disparate treatment* Two people with otherwise identical attributes should be treated the same. $f(x, z) = f(x', z) \; \forall \; x, x', z$.
2. *disparate impact* The distribution of the outcome should be the same for all values of the protected variable. $\sum_z f(x, z)P(z|x) = C \; \forall \; x$

Disparate treatment can be trivially avoided by excluding the protected variable from the model. However, this is

5. user *trust* leading to increased adherence to the model (irrespective of any actual advantages).

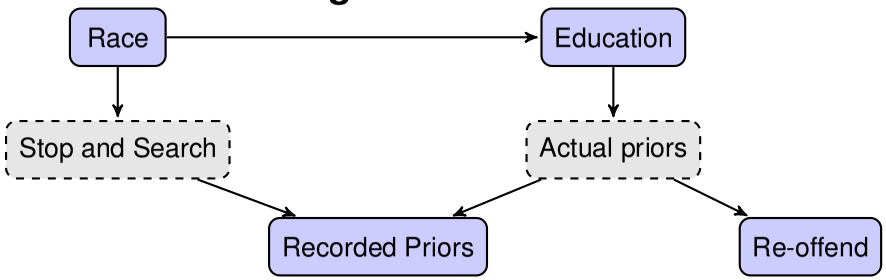
Causal Models

We define a causal model as a model which can be used to predict the outcome of an intervention or change to a system.

1. Causal models explicitly predict the outcome of an intervention. If taking actions based on the outcome of the model significantly changes the system, a causal model is required.
2. Interpretable models can mitigate some, but not all problems associated with not building a causal model eg [2] (figure 2).
3. 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: Consider making the details of an automated essay marking system public.
4. If membership of a protected class, X, is not causally

deeply unsatisfying given the presence of proxy variables and can increase bias (figure 1)
Avoiding *disparate impact* may be be expensive in terms of predictive accuracy and/or require *disparate treatment*.

Figure 1: *Proxies*



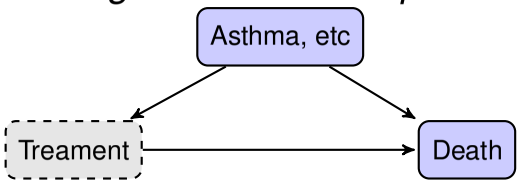
Transparancy/Interpretability

A desire for interpretability implies a miss-match between the real world goal and the optimisation problem presented to a machine learning algorithm [4].

1. situations where the true objective is hard to measure or quantify. (model can be assessed in multiple different ways for hard to define characteristics such as *fairness*)
2. integration with human decision making
3. improved generalizability (users can detect and eliminate features that are irrelevant/artefacts of training data)
4. increased *trust* due to the ability to validate the model for problems such as data leaks & feedback, where acting on model predictions changes the system.

1. If membership of a protected class, X, is not causally related to the outcome of interest, Y, then there exits a set of variables, Z, such that $X \perp\!\!\!\perp Y|Z$.

Figure 2: *Predicting risk of death in pneumonia patients*



References

[1] Solon Barocas and Andrew Selbst. Big Data's Disparate Impact. *California Law Review*, 104, 2016.

[2] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible Models for HealthCare. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, pages 1721–1730, 2015.

[3] Bryce Goodman and Seth Flaxman. EU regulations on algorithmic decision-making and a "right to explanation". In *ICML Workshop on Human Interpretability in Machine Learning*, 2016.

[4] Zachary C Lipton. The Mythos of Model Interpretability. In *ICML Workshop on Human Interpretability in Machine Learning*, pages 96–100, 2016.