Interpret-ability, Causality and Transparency

Finnian Lattimore
Australian National University and Data61/NICTA
finn.lattimore@gmail.com

December 9, 2016

Abstract

As machine learning is incorporated into decision making systems that have fundamental impacts on people's lives, such as in employment, criminal justice, health and financial services, there are increasing concerns over transparency and fairness. The European Union's General Data Protection Regulation, due to come into effect in 2018, requires that people can obtain "meaningful information about the logic involved" in an automated decision process. It also stipulates that such decisions should not be based on special categories of personal data (related to ethnicity, political and religions affiliation or sexuality) unless there "suitable measures" to safeguard individual interests.

A key concern is discrimination against disadvantaged groups. Discrimination is frequently defined in terms of either 'disparate treatment'- treating otherwise similar individuals differently on the basis of a protected attribute such as race or gender, or 'disparate impact' - a substantial difference in outcome between groups. We consider how the notions of disparate treatment and impact may be formalised and the implications of how this is done for machine learning. We examine the overlap between the motivations for interpretable and causal models, especially with respect to assessing fairness. We look at how causal models mitigate some of the trade-offs between transparency and predictive accuracy and we examine to what extent the causal relationship between an attribute, any protected attributes and the outcome of interest is relevant to assessing the impact on fairness of its inclusion in a machine learning model.

1 Introduction