Calibration for the (computationally-identifiable) masses

Omer Reingold, Stanford University

Abstract

As algorithms increasingly inform and influence decisions made about individuals, it becomes increasingly important to address concerns that these algorithms might be discriminatory. The output of an algorithm can be discriminatory for many reasons, most notably: (1) the data used to train the algorithm might be biased (in various ways) to favor certain populations over others; (2) the analysis of this training data might inadvertently or maliciously introduce biases that are not borne out in the data. This work focuses on the latter concern.

We develop and study multicalbration: a new measure of algorithmic fairness that aims to mitigate concerns about discrimination that is introduced in the process of learning a predictor from data. Multicalibration guarantees accurate (calibrated) predictions for every subpopulation that can be identified within a specified class of computations. We think of the class as being quite rich, in particular it can contain many and overlapping subgroups of a protected group.

We show that in many settings this strong notion of protection from discrimination is both attainable and aligned with the goal of obtaining accurate predictions. Along the way, we present new algorithms for learning a multicalibrated predictor, study the computational complexity of this task, and draw new connections to computational learning models such as agnostic learning.

Joint work with Ursula Hebert-Johnson, Michael P. Kim and Guy Rothblum