Introduction to Algorithmic Fairness

Marcello Di Bello

Different perspectives, disciplines

Technical v. theoretical, conceptual, philosophical v. lived experiences, stories

Different disciplines: computer science, statistics, philosophy, law, economics, sociology, history, anthropology

Algorithms in general

Definition: a series of precisely defined steps that perform a task

Question: How precisely specified?

General structure: Inputs –> algorithm –> output

Example: sorting numbers in ascending order

Question: How would such an algorithm look like?

Tradeoffs: speed v. memory

Often written by a human, e.g. a programmer using Python

Machine Learning algorithms

Meta-algorithms whose input are historical data and whose output is another algorithm. ML algorithms are *self-programming*

Applications: face recognition, translation, prediction

Example:

input: historical (training) data about high school GPA and SAT score and college graduation are fed to the meta-algorithm

model search: meta-algorithm searches all possible models, say possibles lines (straight or curved) through the data.

Lines are good for 2-dimensional data (e.g. SAT and GPA) and a binary outcome (graduate/not graduate).

To guard against *overfitting*, split between training and validation data.

optimization: meta-algorithm selects the model (output algorithm) that perform best, often the model that minimizes errors.

output: model (predictive algorithm) for predicting college graduation give high school SAT and GPA

task: make a prediction given high school SAT and GPA about college graduation for new individuals not in the historical data

Error Minimization v. Fairness (According to Computer Scientists)

> . . . it may be that the model that minimizes the overall error in predicting collegiate success, when used to make admission decisions, happens to falsely reject qualified black applicants more often than qualified white applicants. Why? Because the designer didn't anticipate it. She didn't tell the algorithm to try to equalize the false rejection rates between the two groups, so it didn't. (p. 10).

> Writing down precise definitions that capture the essence of critical and very human ideas without becoming overly complex is something of an art form, and it is inevitable that in many settings, simplifications—sometimes painful ones—are necessary . . . [This] tension is not an artifact . . . it reflects the inherent difficulty of being precise about concepts that previously have been left vague, such as "fairness." We believe that the only way to make algorithms better behaved is to begin by specifying what our goals for them might be in the first place (p. 12-13).[1]

Allegheny Family Screening Tool (AFST)

Outcome variables: proxies for child maltreatment: (a) call and referral to Child and Youth Services (CYS), and (b) child placement in foster care.

Predictive variables: stepwise probit regression, regression tested 287 variables and eliminated 156, leaving out 131 predictors

Validation data: Receiver Operating Characteristics (ROC) is 76%

> In 2016, therwe

[1] Kearns and Roth (2020), *The Ethical Algorithm*, Oxford University Press