Using STAN to Explore Fairness in University Admission Policies

Michael Mathioudakis
University of Helsinki
Helsinki, Finland
michael.mathioudakis@helsinki.fi

Carlos Castillo
Universitat Pompeu Fabra
Barcelona, Catalunya, Spain
chato@acm.org

1 THE PROBLEM

Algorithmic fairness is an emerging topic in computer science that seeks to detect and mitigate discrimination introduced by algorithms [2]. One focus of such research is data-driven decision-making algorithms used to allocate a benefit. In the anti-discrimination literature, typical benefits include access to jobs, education, training, and government benefits [4].

Our research studies the algorithm used for admission of university students in Chile. As in other countries, this admission is generally based on a weighted sum of grades in high school and scores in a standardized test, i.e., it is a linear model.

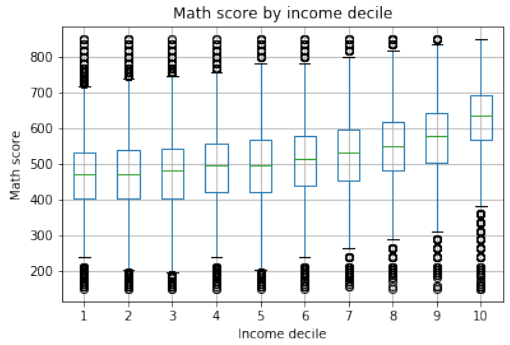

Figure 1: Math scores in the admission test by income decile.

An important societal problem, identified by previous work, is that scholastic achievement and in particular the results in these standardized tests are strongly affected by income [1]. Figure 1 shows this effect is quite dramatic in our data.

2 OUR APPROACH

Our approach follows a recent method dubbed *counterfactual fairness* [3] and implemented in STAN. In this method, one aims to infer the value of a latent variable associated with each student, which authors of [3] call *knowledge* but we prefer to call *aptitude*, given the context of our problem. This variable captures student qualities that might determine scholastic achievement – including, for instance, intelligence and persistence. The model distinguishes between *aptitude* and other factors that might *unfairly* determine achievement – including, for instance, income and gender.

One difference between our analysis and previous works is that we use two datasets that cannot be joined directly. The first corresponds to *national-level data* about students taking the admission test; the second corresponds to *university-level data* corresponding to students admitted to a specific university program. For the national-level data, we have access to gender, high school type (public or private), income decile, grades in high school, and results in standardized tests. For the university-level data, we have access to the same variables, except for income, and we also have access to student grades during the first year.

Our ultimate goal is to determine an admission policy, e.g., a linear function of grades and scores, that is a good predictor of university grades, as a proxy for scholastic achievement. The idea is that it is in the best interest of the university and of the students, that the students that are admitted are those who are likely to obtain good grades at the university. A more immediate goal is to determine whether the inferred student *aptitude* values are a good such predictor. If that were the case, we would have an admission policy that is not only accurate, but also fair – in the sense that it would discount the role of factors (e.g., gender, income) that *unfairly* affect scholastic achievement.

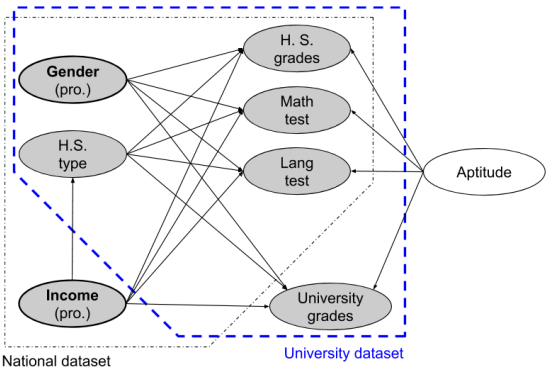

Figure 2: Dependencies in our model. Gender and income are protected or sensitive attributes.

Our model is depicted in Figure 2. Our preliminary results are promising but also show an important challenge. While we are able to model aptitude in the national dataset using assumptions similar to the ones in previous work [3], using the inferred *aptitude* values for students in the *university-level dataset* as predictor for scholastic achievement proves more challenging. The purpose of this poster presentation is to start a discussion with STAN experts and practitioners about this project and to get feedback on our modeling choices and methods. Our STAN code will be public.

STAN'18, August 2018, Helsinki, Finland

REFERENCES

[1] Rodrigo Cornejo Chávez. 2006. El experimento educativo chileno 20 años después: una mirada crítica a los logros y falencias del sistema escolar. *REICE. Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación* 4, 1 (2006).

[2] Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2125–2126.

[3] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*. 4066–4076.

[4] John E Roemer. 2009. *Equality of opportunity*. Harvard University Press.