Algorithmic Fairness – Fairness Criteria

Marcello Di Bello - ASU - Fall 2021 - Week #5

Our goal is to understand the different formal criteria of algorithmic fairness that have been proposed in the computer science literature, their pros and cons. Over the last five years, there has been a proliferation of formal criteria (metrics, measures, definitions) of algorithmic fairness. Unfortunately, the terminology in the literature is not uniform. We should keep track of the concepts rather than the terms.

The set up

An algorithm makes predictions (also called decisions) about an individual based on a collection of attributes had by the individual. The attributes are therefore the predictors. The attributes are distinguished in protected attributes (race, gender, etc.) and unprotected attributes (income, education, prior crimes, etc.).

The predicted outcome is an event of interest, such as criminal activity, defaulting on a loan, etc. The algorithm, however, does not usually make a black-and-white prediction. Rather, based on the attributes an individual possesses or lacks, the algorithm assigns to each individual a risk score that expresses the probability that the outcome of interest will occur. The prediction (decision) depends on whether or not the score assigned to the individual meets a certain threshold. Say, if an applicant is assigned a risk score for loan default of 6 or higher, the algorithm will recommend that no loan be granted to the applicant.

The optimal threshold can be set by aiming at maximizing expected utility.

Algorithms are trained on historical data. When we assess their performance—in terms of fairness and accuracy—we do so by comparing the prediction and the true outcome. We make this comparison using the historical data which contain the true outcome. However, at the time of making the decision, the outcome to be predicted is not known. That is why we are making a prediction in the first place.

Popular criteria

To illustrate some of the most popular criteria of algorithmic fairness, we will use the loan example for concreteness. The outcome is binary and is defined as "the applicant defaults" or "the applicant does not default". The two groups being compared are generically called Circles and Squares.

- *Anti-classification*: protected attributes should not be part of the

predictors used by the algorithm to make decisions about loans.

- *Equal treatment*: any two individuals who possess the same relevant attributes should be treated the same by the algorithm. Relevantly similar individuals should be treated the same.[1]
- *Same threshold*: the decision threshold to recommend that an individual be given a loan or not is the same across Circle applicants and Square applicants.
- *Causal paths*: the algorithm does not rely on data in which race is implicated in morally objectionable causal paths. We will discuss this conception more in detail later in the course.
- *Demographic parity* (also called *statistical parity*[2]): a proportion of Circle applicants are granted a loan and a proportion of Square applicants are granted a loan. *The two proportions are the same*.
- *Equality in false positives* (also called *equal opportunity*[3]): a proportion of Circle applicants *who do not (would not) default* are not granted a loan and a proportion of Square applicants *who do not (would not) default* are not granted a loan. *The two proportions are the same*.
- *Calibration*: a proportion of Circle applicants *who are labeled at a certain level of risk of default (e.g. a risk of 8)* do default and a proportion of Square applicants *who are labeled at a certain level of risk of default (e.g. a risk of 8)* do default. *The two proportions are the same*.

[1] This definition is implicit is the piece Martinez and Kirchner (2021), The Secret Bias Hidden in Mortgage-Approval Algorithms, The Markup. See also Dwork et al. (2012). Fairness Through Awareness. In Proceedings of the 3rd innovations in theoretical computer science conference, 214–226.

[2] Kearns and Roth (2020), The Ethical Algorithm, Chapter 2, p. 69

This is the diagnostic perspective.

[3] Hardt et al (2016), Equality of Opportunity in Supervised Learning. In Advances In Neural Information Processing Systems, 3315–3323. A more general notion is *equality of false positives and false negatives* (also *called equalized odds*): Definition left as an exercise.

This is the predictive perspective. See also predictive parity: definition left as an exercise.

Arguments against...

.... anti-classification

When gender or other protected traits add predictive value, excluding these attributes will in general lead to unjustified disparate impacts; when protected traits do not add predictive power, they can be safely removed from the algorithm. But we note one curiosity in the latter case. If protected attributes are not predictive, one could in theory build an accurate risk model using only examples from one particular group (e.g., white men). Given enough examples of white men, the model would learn the relationship between features and risk, which by our assumption would generalize to the entire population. This phenomenon highlights a tension in many informal discussions of fairness, with scholars advocating both for representative training data and for the exclusion of protected attributes. In reality, representative data are often most important precisely when protected attributes add information, in which case their use is arguably justified (p. 10)[4]

[4] Corbett-Davies and Goel, The Measure and Mismeasure of Fairness

. . . . equal false positives

> Suppose . . . prosecutors start enforcing low-level drug crimes that disproportionately involve black individuals . . . suppose that the newly arrested individuals have low risk of violent recidivism, and thus are released pending trial. This stylized policy change alters the risk distribution of black defendants, adding mass to the left-hand tail. As a result, the false positive rate for blacks would decrease. To see this, recall that the numerator of the false positive rate (the number of detained defendants who do not reoffend) remains unchanged while the denominator (the number of defendants who do not reoffend) increases. Without considering the distribution of risk—and in particular, the process that gave rise to that distribution—false positive rates can be a misleading measure of fairness (p. 15).

Exercise: Break down this argument. Draw the risk distribution before and after. Satisfy yourself that the reasoning is sound. Do you agree with the criticism?

The general point, of which the argument above is an illustration:

> To the extent that error metrics differ across groups, that tells us more about the shapes of the risk distributions than about the quality of decisions . . . it is hard to determine whether differences in error rates are due to discrimination or to differences in the risk distributions (p. 12).

. . . . calibration

> . . . imagine a bank that wants to discriminate against black applicants. Further suppose that: (1) within zip code, white and black applicants have similar default rates; and (2) black applicants live in zip codes with relatively high default rates. Then the bank can surreptitiously discriminate against blacks by basing risk estimates only on an applicant's zip code, ignoring all other relevant information. Such scores would be calibrated (white and black applicants with the same score would default equally often), and the bank could use these scores to justify denying loans to nearly all black applicants. The bank, however, would be sacrificing profit by refusing loans to creditworthy black applicants . . . [this] lending strategy is indeed closely related to . . . redlining (p. 16).

Exercise: Break down this argument. Is the point here that within ZIP code there could great variability in default rates? Do you agree with the criticism?

> This redlining example can be generalized . . . we can aggregate low-risk and high-risk individuals and re-assign them risk scores equal to the group average. . . . Depending on where the mean lies relative to the decision threshold, this process can be used to change the number of individuals receiving positive or negative classifications (p. 16).

Trade-offs: Accuracy v Fairness

Accuracy and fairness (understood as comparative error) do not always go together. Consider for simplicity a college admission process based on just LSAT scores. Admission officers are confronted with the question, where should the LSAT score threshold for admission

be set? A certain threshold may be more accurate (fewer errors overall) but may lead to more unfairness (more errors to some group and fewer to another). Another threshold may be more fair but less accurate. Which threshold should be used?

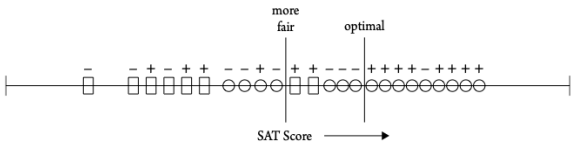

Figure 1: The optimal thresholdf is not the same as the fair threhsold. Source: Kearns and Roth, 2020.

Accuracy and fairness can be both improved by relying group-based thresholds. This would violate same threshold fairness, however.

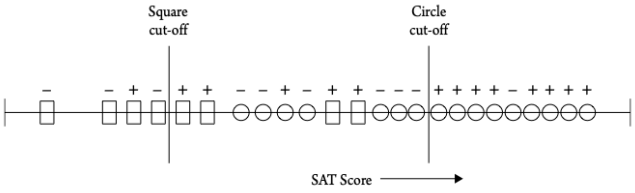

Figure 2: Two thresholds. Source: Kearns and Roth, 2020.

The performance of a family of algorithms (predictive models) could be compared by means of a Pareto frontier plotting fairness against accuracy.

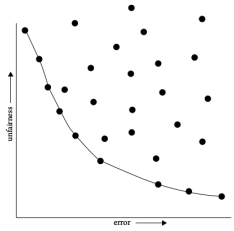

Figure 3: Pareto Frontier. Source: Kearns and Roth, 2020.

Fairness for which groups?

In aiming to be fair toward groups, which groups are we considering? The standard answer is that we should be concerned with protected groups. But why stop there? Wouldn't the economically disadvantaged be entitled to be treated fairly even though they are not a protected group? How about the intersections of different groups, maybe some protected and others unprotected? Should fairness be relative to subgroups, as well?

Examining fairness at the group or subgroup level can also give rise to odd results. It is in theory possible to be fair toward Men compared to Women and also Circles compared to Squares, and yet be unfair across the intersections of these groups.

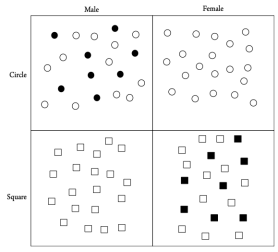

Figure 4: Fairness gerrymandering. Source: Kearns and Roth, 2020.