Fairness definitions

Sarah Gillespie

12/16/2021

Most common fairness definitions

So first, what is fair?

The paper, On Formalizing Fairness in Prediction with Machine Learning, discusses a number of definitions on how to decide if an algorithm is "fair," specifically by breaking the concept down into five ways that an outcome can fail fairness. These are just the most popular of many different ways to look at a "fair" algorithm. It cannot be emphasized enough that fairness is not a simple accuracy metric that is easy to optimize a model to achieve. It is crucial to look at your model through different fairness lenses to see if your model has negative externalities that fail different fairness measures, despite succeeding on a specific fairness measure. This week's classes will focus on trying to optimize algorithms to meet the largest possible number of fairness definitions and how to document that an algorithm is or is not being "fair."

Counterfactual fairness

Concept	Definition	Example
Counterfactual fairness	Decision is fair towards an individual if the decision is the same in both the actual world and counterfactual worlds where the individual belongs to a different demographic group.	Amazon's resume algorithm rejecting the resumes of anyone with the word "woman" on her resume.

Group fairness

Concept	Definition	Example
Group fairness	The outcome prediction for individuals should be the same outcome across different demographic groups with almost equal probability. Collectivist egalitarianism.	Northepoint sentencing algorithm failed group fairness between different racial groups. The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white

Individual fairness

Concept	Definition	Example
Individual fairness	A predictor is fair if it produces similar outputs for similar individuals. Associated with individual egalitarianism.	A victim of the Northpointe algorithm was Brisha Borden was given a disproportionately harsh sentencing suggestion compared to another data point, Vernon Prater.

Equality of opportunity

Concept	Definition	Example
Equality of opportunity	In a scenario where there is only one beneficial outcome, the true positive rate should be the same for all groups.	Not accounting for systemic discrimination that may affect one group's data used in the algorithm in an adverse way

Preference-based fairness

Concept	Definition	Example
Preference-based fairness	An algorithm satisfies an individual's preferred treatment and has the highest rate of matching an individual's real preference to the algorithm's prediction for the individual's preference.	An algorithm that has a high error percentage for preference matching, such as an unsuccessful roommate algorithm at a large university.

Common confusion points

Common confusion point: group vs. individual fairness

One common confusion point when comparing fairness definitions is the difference between group fairness and individual fairness. One might wonder how they are different, since treating groups differently means being unfair to all the individuals in those groups.

Consider two different groups of students: students at Smith College studying computer science and students at Mount Holyoke College studying computer science. All students are applying for jobs at the same career fairs and the same LinkedIn posts and the same company websites.

Group fairness in action

If there was group fairness, then the Smith College computer science students would get the same number of interview invitations as Mount Holyoke College students. If Smith computer science students received more or less interview invitations than the Mount Holyoke computer science students then this would fail group fairness.

Individual fairness in action

Individual fairness is focused on individual points in the data set and checking to make sure identical data points receive the same treatment. So, if two Smith computer science students had the same academic background, grades, and experiences and applied to the same jobs, then the interview process would be individually fair if the two Smith students received the same number of interview invitations. If the students did not receive the same number of interview invitations, then the process would not be individually fair. A company might offer more interview opportunities to people in groups underrepresented in their employee population to achieve a group fairness goal. This decreases individual fairness in the applicant population but has a potential to increase group fairness in the company's employee population.

Common confusion point: treatment vs. impact

Treatment is the specific type of action that an algorithm suggests for each data point, while impact is the effect on that data point's well-being. Since each person tends to be more unique and has their own set of skills and resources, providing the exact same treatment to people is unlikely to lead to the exact same outcome

Treatment vs. impact

Treatment vs. impact in action

An educational algorithm provides the same treatment that suggests two first grade students read the same chapter in a book. If one student has an older sibling around to help look up challenging words and explain deeper literary elements like foreshadowing and social context, that student will gain a larger benefit and a different impact than a student who reads the book without her support system around helping her to understand the chapter. Despite having the same treatment, these two students had a very different impact.

What's the best fairness definition?

Find out together in class!

Can we meet all the fairness definitions for every algorithm?

Find out together in class!

A hint

A quote from an Expo Talk Panel at NeurIPS 2020 titled “From scikit-learn and fairness, tools and challenges” by Adrin Jalali best summed up the current algorithm fairness terrain when he said “If you take an algorithm, it’s not too hard for me to find a fairness definition for which my algorithm does not perform badly. So we need to be really careful about that.”