Algorithmic Fairness – Biased data

Marcello Di Bello - ASU - Fall 2021

Terminology refresher (from last time): False positive. False negative. Confusion matrix. Sensitivity. Specificity. Positive Predictive Value (PPV). Negative Predictive Value (NPV). False positive classification error rate. False negative classification error rate. False positive prediction error rate. False negative classification error rate.

Our goal is to make some progress in understanding the expression "algorithmic bias", focusing on biases on the input side, namely the historical data on which predictive algorithms are trained.

Let's start with some observations about different forms of bias. Like "being" in Aristotle's *Metaphysics*, bias is said in many ways. Two well-known forms of bias are cognitive bias and social bias.

Cognitive and social bias

A *cognitive bias* is a deviation from the standards of rationality. Examples of cognitive biases are logical fallacies, confirmation bias, availability heuristics, framing effects.[1]

A *social bias* is a deviation from the (moral?) norm of treating everybody equally (or treating everybody with the concern and respect they deserve given the circumstances). It is an individual or group attitude, consciously accessible or not, that results in beliefs or actions that are discriminatory against members of certain social groups.

So, in its most general sense, we could define bias as any *deviation from an accepted standard or norm*.

[1] See, for example, Kahneman et al (eds) (1982), Judgment under Uncertainty: Heuristics and Biases, Cambridge UP and more recently Gilovich et al (eds) (2002), Heuristics and Biases: The Psychology of Intuitive Judgement, Cambridge UP.

This is a working defintion which could be too general and thus unhelpful.

Information bias and stereotype bias

Our goal here is to understand algorithmic bias (more on this soon). Perhaps, defining bias as deviation from a norm or standard is not the right definition for describing algorithmic bias. Here are two other forms of bias. First, bias as prior information:

> In AI and machine learning, *bias* refers to prior information … Yet bias can be problematic when prior information is derived from precedents known to be harmful … we will call harmful biases 'prejudice'….prejudice is a special case of bias identifiable only by its negative consequences (p. 2/14)[2]

[2] Caliskan et al (2017), Semantics Derived Automatically from Language Corpora Contain Human-like Biases

Second, bias as a reasoning, decision or action pattern that instantiates social-kind inductions:

> … bias … [consists] …. of mental entities that take propositional mental states as inputs and return propositional mental states as outputs in a way that instantiates social-kind inductions. ….
>
> (i) Jan is elderly.
> (ii) Elderly people are bad with computers.
> (iii) Jan is bad with computers. [3]

[3] Johnson (2020), The Structure of Bias, Mind, 129(516): 1193–1236. This definition is also connected to the notion of stereotypes or group generalizations.

These two quotations introduce other forms of bias, which we could call *information bias* and *stereotype bias*. These forms of bias may intersect with cognitive and social bias. For example, stereotype bias might be what social bias ultimately consists in.

Algorithmic bias

When we speak of algorithmic bias, we are not using the word "bias" as in "cognitive bias" or "social bias". Algorithms follow the instructions given to them and apply them uniformly. Arguably, they cannot be cognitively or socially biased. Or can they?

Is algorithmic bias a form or byproduct of information or stereotype bias?

Algorithms can be biased if they inherit—and then reproduce and even amplify—biases embedded in the historical data used to train them or biases embedded in society and reflected in the data. We often talk about biased data, for example, racially biased, gender biased or age biased data.

What do these expressions mean? What standards would data deviate from?

Examples

Let's consider a few examples of claims that have been made about biases in society, data or algorithms:

For each example, identify the norm from which the deviation occurs and the kind of data or society bias at issue.

a. Disparities in the severity of charges against black defendants compared to white defendants after controlling for several factors.[4]

[4] See Starr and Rehavi (2014), Racial Disparity in Federal Criminal Sentences, J of Pol Ec, 122(6):1320-1354.

b. A SAT question that asks students to complete the verbal analogy *Runner:Marathon=x:Regatta*. The solution is *x=oarsman*. The question is biased against students unfamiliar with regattas, because of country of origin, language spoken at home, socioeconomic status.

Think of the SAT exam as an algorithm to score academic preparation.

c. The analogy *Man:Programmer=Woman: Homemaker* completed by Google algorithm word2vec. It reproduces gender stereotypes.

Possible solution: distinguish between gendered words (such as queen and king) and non-gendered words (such as programmer and homemaker). See Kearns and Roth (2020), The Ethical Algorithm, p. 63. Is this a good solution?

d. Amazon algorithm for screening job applicants. It penalized applicants for engineering positions whose resumes contain the word "woman".

e. COMPAS algorithm for guiding decisions about bail, preventative detention or sentencing. Its false positive classification error rate for black people is twice as high as that for white people.[5]

[5] ProPublica (2016), Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks.

f. For the same risk score assigned by a health algorithm, black patients are sicker than white patients. Risk scores are based on health costs per patient as a proxy for level of sickness.[6]

[6] Obermeyer et al (2019), Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations, Science, 366(6464): 447-453.

g. Other example?

Some distinctions

data // algorithm

Under what circumstances is the algorithm biased and under what circumstances are the data biased? Or is this distinction misleading?

Consider a perfectly accurate classification algorithm. Could this algorithm still be regarded as biased?

supervised // unsupervised learning

Machine learning algorithms can be distinguished between supervised learning algorithms (e.g. regression, SVM, random forest, Bayes classifier) and unsupervised ones (e.g. clustering such as Google algorithm word2vec).

Does the notion of bias differs in the context of supervised and supervised learning?

related words

We use the word bias along with: disparity; discrimination; prejudice; stereotype; injustice; unfairness; etc. Obviously, we should not confuse them, even though there seems to be points of overlap.

What else?

Biased data

Let's focus on the data on which algorithms are trained. Data can be biased for several difference reasons. In the most general sense, data are biased if they unequally portray different socially relevant groups (say racial or gender groups). This definition is general (which is a good thing), but also over-inclusive (which is a bad thing).

Here are some examples:

Which are instances of bias in the data and why?

- The outcome of interest is represented by a proxy in the historical data. This proxy tracks the actual outcome unevenly depending on the social group (see, for example, item f. above)
- Data reflect biases in society (see, for example, items c. and d.)
- Uneven representation in the training data of people belonging to different racial, gender or other socially relevant groups.[7]
- Data for different social groups might have different sample sizes, which leads to differences in the variability of the parameters estimates. The sample size for minority groups tends to be smaller than the sample size for majority groups.

  > Our models about minorities generally tend to be worse than those about the general population ... this is assuming the classifier learned on the general population does not transfer to the minority faithfully.[8]

What are other examples of biases in society reflected in the data? Are biases in society different from social bias as defined earlier?

[7] For an example about the medical data, see Perez (2019), Invisible Women, p. 167.

[8] Hardt (2014), How Big Data is Unfair. For example, compare confidence intervals for the effectiveness of COVID-19 vaccine for white people and minorities. See Polack et al (2020), Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine, N Engl J Med, 383:2603-2615.

- The correlations between attributes (predictors) and the outcome of interest is not uniform across groups. The same set of predictors could perform differently for different social groups thus requiring group-specific predictive models. Here is an example:

Figure 1: Fitting a Support Vector Machine (SVM) liner model to the ProPublica COMPAS data.

Risks in predictive algorithm

Predictive algorithms output a risk score, an assessment of how probable it is that an individual would, say, commit a crime or default on a loan. What is the meaning of this risk? Does it make sense to assign risk to an individual?

> ... current risk assessment instruments are unable to distinguish one person's risk of violence from another's ... a predictive algorithm might confidently estimate a person's risk of arrest as somewhere between a range of five and fifteen percent. Studies have demonstrated that predictive models can only make reliable predictions about a person's risk of violence within very large ranges of likelihood, such as twenty to sixty percent. As a result, virtually everyone's range of likelihood overlaps. When everyone is similar, it becomes impossible to differentiate people with low and high risks of violence.[9]

[9] Barabas et al, Technical Flaws of Pretrial Risk Assessments Raise Grave Concerns. See Hart and Cooke (2013), Another Look at the (Im-)precision of Individual Risk Estimates Made Using Actuarial Risk Assessment Instruments, Behavioral Science and Law, 31(1):81-102.