Strategies and criteria for bias avoidance in machine learning

Examples of implicit bias abound in machine learning and statistics:

- Beauty filters make features more European (whiter skin, narrower noses). Sources: madamenoire, TechCrunch.
- An Google automated captioning system labeled Black people as "gorillas". Source: Forbes
- Word embeddings produce analogies such as "man is to software engineer as woman is to homemaker". Source
- Crime prediction tools give higher false positive rates on Black people. Source: ProPublica. Needless to say, this is a controversial topic. I haven't read it all and thought it through; that may be a future post. The ProPublica analysis is here and there is a rebuttal from the original authors here.
- Polygenic scores (an application of machine learning to genetics and medicine) are expected to exacerbate health disparities. Source: Martin et al arXiv preprint.

If this makes you upset, check out the work being done by the folks in the algorithmic justice league or at Google or at Microsoft. It makes me upset, but it also make me curious...

Is it possible to do better? If yes, then how?

We're not going to solve sexism or racism by blogging about statistics, but if this is a tiny little corner of the universe that we have some control over, then let's have a look at how to do things justly. There is legitimate interest from a lot of folks in this topic.

This current post summarizes and evaluates strategies for avoiding bias. The scope is narrow: I consider simple supervised machine learning problems with a single outcome of interest, where predictions directly affect a reward or penalty someone is offered. Examples might include predicting law school performance from pre-admission surveys (reward: admission), predicting crime rates by location (penalty: heavy policing), or predicting banner ad clicks based on user data (reward: access to high-paying jobs). This post was prompted by the wellness rewards calculation discussed in this other post???.

Some examples match the statistical format -- supervised ML with a single outcome -- but don't really fit what I have in mind because the resulting decision does not offer anyone a reward or penalty. Discrimination arises differently in such situations, and different fairness criteria are needed. An example is clinical use of polygenic risk scores, where it makes sense to use an *equal accuracy* criterion that I argue against below.

Some notation

The following conventions are used in the remainder of the post.

- Y denotes the outcome to be predicted. For example, this might be car accident risk if you work at an insurance company or it might be some aspect of criminal record if you're working on predictive policing.
- X denotes the set of features to be used in making predictions. For example, this might include location of residence or income. X may be statistically related to A: for example, in America, neighborhoods are highly segregated by race.
- A denotes a protected characteristic such as race or gender identity. For example, $A = 0$ could encode someone who is transgender and $A = 1$ someone who is cisgender.
- $\hat{Y}$ denotes a prediction of Y. It is a function of X and A.
- U denotes unobserved variables that may be correlated and/or causally linked with X, A, and Y.
- Conditional probability distributions will be written $Pr(Z|W = w)$. The intuitive meaning of this is that you pick pairs Z, W at random and discard them unless $W = w$. For example, to sample from $Pr(horsepower|carcolor = red)$, you pick cars until you get a red one, then measure its horsepower. For another example, you sample from $Pr(interview|CandidateName = StereotypicallyBlack)$ by flipping through resumés until you get one with a stereotypically Black name, and then you figure out whether your company called them for an interview.
- Counterfactual probability distributions will be written with an extra subscript: $Pr(Z_{W \leftarrow u}|W = w)$. The intuitive meaning of this is that you pick pairs Z, W at random and discard them unless $W = w$, but then there's an additional step where you intervene and set W to u. For example, to sample from $Pr(horsepower_{carcolor \leftarrow black}|carcolor = red)$, you pick cars until you get a red one, you paint it black, and then you measure its horsepower. For another example, to sample from

$$Pr(interview_{CandidateName \leftarrow StereotypicallyWhite}|CandidateName = StereotypicallyBlack)$$

, you pick a resumé as before, change the name, and figure out whether your company called them for an interview. (Similar studies have been done to rigorously assess racial preferences in hiring.)

What does it mean to be fair?

Most debates in this arena begin and end not with data analysis, but with selection of fairness criteria. (They sometimes brush it under the rug or mix it up with data analysis or , which may keep readers engaged, but it becomes very difficult to.) In some debates in this arena, people appear to be arguing about "what the data show", but they're actually not when they

These are drawn from the following paper, but in some cases where only a binary outcome was considered, I have formulated a version of the criterion for quantitative outcomes.

> Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2017). Fairness in criminal justice risk

> assessments: the state of the art. arXiv preprint arXiv:1703.09207.

(Un?)fairness through unawareness

A naive method might be to exclude protected factors A from the set of features used for training. Unfortunately, this does not work, and I have never heard anybody consider it seriously as a method for getting fair results. The issue is that protected factors often explain part of the link between non-protected features X and the outcome Y. For instance, grocery purchases carry information about race and diagnosis of diabetes. Race and diabetes are informative about health expenses. A predictive method using grocery bills for health insurance pricing is likely to exploit race and pre-existing conditions whether you mean it to or not.

Rating: F

Setting various things equal within groups

For to achieve a criterion called *equal calibration*, predicted rates are required to match actual rates within each stratum. For all possible values of a,

$$Pr(\hat{Y} = 1|A = a) = Pr(Y = 1|A = a)$$

or

$$E[\hat{Y}|A = a] = E[Y|A = a]$$

.

This seems like a minor variation on the strategy above: instead of maximizing an accuracy metric that averages over unobserved protected attributes, you could do this by maximizing accuracy separately within each category.

Unlike the unawareness strategy, predictions optimized for equal calibration will explicitly use information about protected attributes. Unfortunately, the information will be used to perpetuate historical inequities rather than to remedy them. For example, if Black clients have historically been more likely to default on loans, then risk prediction systems using the equal calibration criterion will result in lower access to capital for Black clients.

Rating: F

N.B. Auto insurance price discrimination by gender is commonplace in the US. People justify higher prices for men by pointing to higher rates of car crashes, car crash deaths, speeding, drunk driving, and not using seat belts. This is essentially an equal calibration argument. This is legal in all U.S. states except Montana (source) and, as of very recently, California (source). Here in Massachusetts, gender identity is a protected attribute for many purposes, but pricing car insurance is not one of them (source). My intepretation is that equal calibration

makes sense for some purposes, but it's got nothing to do with protecting against discrimination, and in fact it would be better called a justification for discrimination.

Setting various things equal across groups

Equal accuracy

It is often possible to adjust predictions until some performance metric is the same across groups. I call this family of criteria the *equal accuracy family*. For example, one could require raw accuracy (for binary outcomes) or mean squared errors (for quantitative outcomes) to be the same for men and women.

$$Pr(\hat{Y} = Y|A = w) = Pr(\hat{Y} = Y|A = m)$$

$$E[(\hat{Y} - Y)^2|A = w] = E[(\hat{Y} - Y)^2|A = m]$$

Prediction parity

A similar option (sometimes called *statistical parity* or *demographic parity*, but I prefer the name *prediction parity*) is to require the distribution of predictions within each group to have the same mean.

$$Pr(\hat{Y} = 1|A = w) = Pr(\hat{Y} = 1|A = m)$$

I've only seen this applied to binary outcomes, and it could generalize to quantitative outcomes in various ways, which I won't discuss much. One obvious possibility is this.

$$E[\hat{Y}|A = w] = E[\hat{Y}|A = m]$$

Equality of opportunity

This is a variant of prediction parity or equal accuracy where equality is only required in part of the population. For instance, consider recidivism prediction for men and women. Suppose Y is coded as 0 for people who don't recidivate (don't commit another crime after getting out of jail). Then equality of opportunity might be stated as equality of false positive rates.

$$Pr(\hat{Y} = 1|Y = 0, A = w) = Pr(\hat{Y} = 1|Y = 0, A = m)$$

The only difference is that the metric ignores what happens when $Y = 1$. If you use this metric, you're saying that for people who recidivate, it's okay for the prediction system to do a better job catching some than others, but among people who don't commit further crimes, men and women deserve equal treatment.

This criterion only makes sense for special situations. The use in law enforcement bothers me because one mechanism of racial discrimination by police is actually different rates of leniency towards *guilty* people, and

because in some prediction systems, the outcome is merely arrest and booking, not conviction by a jury of peers.

Ratings for criteria that set things equal across groups

In general, it's not possible to reconcile prediction parity with any type of equal accuracy criterion. This family of similar-but-incompatible metrics struggles to make sense of situations in which base rates differ between groups. If you plan to use a criterion from this family, you need a solid argument for which one it should be.

In many situations, existing outcomes Y were shaped by systems of advantage based on protected classes, so maintaining equal accuracy will not produce fair results. In fact, equal accuracy seems better suited to justifying discrimination rather than preventing it; I am hard pressed to come up with an intuitive example where equal accuracy works. Prediction parity might be a better criterion to use, but it's a blunt hammer and it can be viewed as penalizing groups with lower base rates.

I give a 'C' rating to prediction parity and a 'D' rating to equal accuracy.

Counterfactual fairness

A recent paper by James Kusner and colleages (citation below) advocates for the following definition of "Counterfactual Fairness". For all possible values of x, a, and a',

$$Pr(\hat{Y}|X = x, A = a) = Pr(\hat{Y}_{A \leftarrow a'}|X = x, A = a)$$

.

In plain English, if you alter someone's gender identity or racial/ethnic background (for example), it should not affect the predictions made about them.

This really threw me for a loop: what exactly does the intervention entail? For biological/physical sex, the intervention seems simple to imagine: at conception, replace the paternal X chromosome with the paternal Y chromosome or vice versa. For a concept like race, which is tied up with culture, history, and perceptions of others, it's not at all clear to me what this intervention would mean.

A key aspect of this definition is that you have to allow the intervention's effects to propagate. For the case of biological/physical sex, the intervention happens at conception, and a fair prediction should be invariant to all the resulting changes in either life experience or innate characteristics. If you really wouldn't have recidivated but for the testosterone, that counts as protected according to this criterion.

Regarding race, Kusner et al specify that the set of protected characteristics needs to be ancestrally closed with respect to the underlying causal graph. If race is protected, so is "mother's race", and so on. This has massive implications when applying the criterion to race as a protected attribute. If you want your prediction

system to meet Kusner et al's standard of racial equity, you should imagine the intervention happening generations ago, with all the cascading experiences of racial identity, and your predictions should still come out the same.

Here's the paper if you want to dig into it yourself.

> Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In Advances in Neural Information Processing Systems (pp. 4066-4076). link

Rating: B. This criterion allows for a lot of nuance: the analyst can structure a causal graph, specifying which variables and pathways are protected. It is correspondingly very difficult to justify models and conduct inference. I view such difficulties as a positive feature, because an honest criterion ought to reflect the obvious truth that this type of problem is difficult or impossible to solve.

Optimizing society-wide recovery from injustice

The criteria above are individualistic and symmetrical. This limits their ability to express what we know about the history behind our problems: race relations and gender roles in the US are certainly not symmetrical, and historically they are even less so. It also means these criteria encode only a mandate to be fair to directly affected people, rather than a mandate to make society as a whole more just. This presents a disconnect with broader ideas of how the justice system should work (restorative justice) or how to achieve fair access to capital and educational opportunities. Whenever possible, we should step outside the intellectual box provided by a specific prediction problem and look at the landscape afresh.