medium.com

How big data is unfair - Moritz Hardt - Medium

Moritz Hardt

10-13 minuten

As we're on the cusp of using machine learning for rendering basically all kinds of consequential decisions about human beings in domains such as education, employment, advertising, health care and policing, it is important to understand why *machine learning is not, by default, fair or just in any meaningful way.*

This runs counter to the widespread misbelief that algorithmic decisions tend to be fair, because math is about equations and not skin color. Examples of this misbelief are common and evident in a recent piece on data-driven crime fighting that appeared in the Financial Times, which Cathy O'Neil brought to my attention.

Gilian Tett, a well respected financial reporter, argues that the benefits of predictive policing outweigh potential harm with the following justification:

> "After all, as the former CPD [Chicago Police Department] computer experts point out, the algorithms in themselves are neutral. 'This program had absolutely nothing to do with race… but multi-variable equations,' argues Goldstein. Meanwhile, the potential benefits of predictive policing are profound."

Ironically, Gilian Tett is well known for reporting on the failure of such things as "multi-variable equations" in the wake of the financial crisis, but she is perplexingly quick to accept that multi-variable equations are neutral and therefore fair, because the "computer experts" (whatever that means) at the police station asserted them to be so.

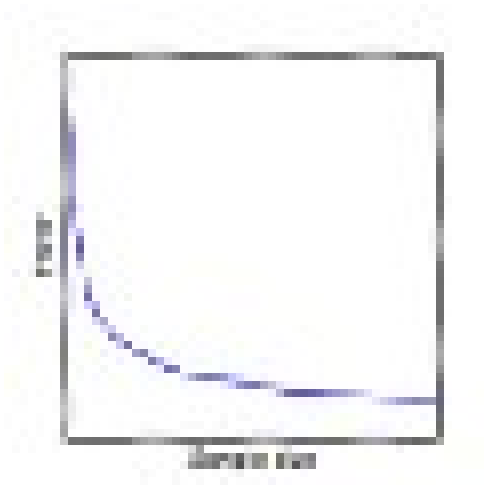
My goal is not to belabor this one example. Instead I'd like to refute the claim that "machine learning is fair by default". I don't mean to suggest that machine learning is inevitably unfair, but rather that there are powerful forces that can render decision making that depends on learning algorithms unfair. Any claim of fair decision making that does not address the technical issues that I'm about to discuss should strike you as dubious. What it means to achieve fairness in machine learning is a much more delicate question that I will return to in a future post.

Let's recap some terminology. A *learning algorithm* is loosely speaking any algorithm that takes historical instances (so-called *training data*) of a decision problem as input and produces a decision rule or *classifier* that is then used on future instances of

the problem. An attribute of the data is often called a *feature* and the set of all available attributes defines the *feature space* or *representation* of the data.

An immediate observation is that a learning algorithm is designed to pick up statistical patterns in training data. If the training data reflect existing social biases against a minority, the algorithm is likely to incorporate these biases. This can lead to less advantageous decisions for members of these minority groups. Some might object that the classifier couldn't possibly be biased if nothing in the feature space speaks of the protected attributed, e.g., race. This argument is invalid. After all, the whole appeal of machine learning is that we can infer absent attributes from those that are present. Race and gender, for example, are typically *redundantly encoded* in any sufficiently rich feature space whether they are explicitly present or not. They are latent in the observed attributes and nothing prevents the learning algorithm from discovering these encodings. In fact, when the protected attribute is correlated with a particular classification outcome, this is precisely what we should expect. There is no principled way to tell at which point such a correlation is worrisome and in what cases it is acceptable. An excellent recent work by Barocas and Selbst that goes into great detail about this aspect as well as the entire topic of this article.

Even if we had a mythical source of unbiased training data, I argue the problem would persist; machine learning can still be unfair. One reason is simple. Assuming a fixed feature space, a classifier generally improves with the number of data points used to train it. The whole spiel about big data is that we can build better classifiers largely as a result of having more data.



The error of a classifier often decreases as the inverse square root of the sample size. Four times as many samples means halving the error rate.

The contrapositive is that less data leads to worse predictions. Unfortunately, it's true by definition that there is always proportionately less data available about minorities. This means

that our models about minorities generally tend to be worse than those about the general population. Importantly, this is assuming the classifier learned on the general population does not transfer to the minority faithfully. If both groups together form one homogeneous population, then additional samples may benefit both groups.

Modeling a heterogeneous population as a gaussian mixture and learning its parameters using the EM algorithm. As expected, the estimates for the smaller group are significantly worse than for the larger. Dashed red ellipsoids describe the estimated covariance matrices. Solid green defines the correct covariance matrices. The green and red crosses indicate correct and estimated means, respectively.

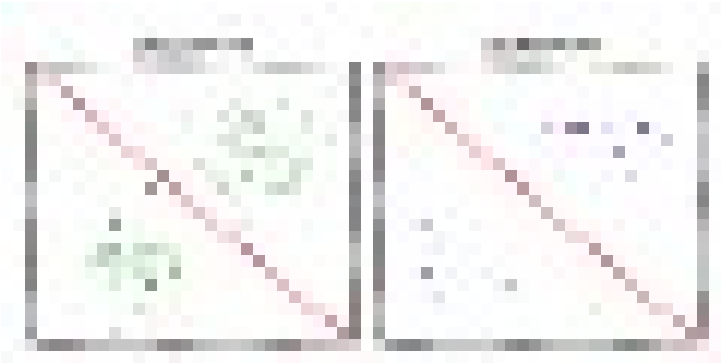
Correcting for the disparity in sample size strikes me as rather difficult. After all, if we had a way to get a better model out of less data, we'd apply the very same ideas to the general population as well for a corresponding gain. Therefore I would argue that there's a general tendency for automated decisions to favor those who belong to the statistically dominant groups.

I say "favor" because majority groups will enjoy higher rates of accuracy in decision-making. In this case, accuracy as a proxy for fairness seems appropriate. A classifier that performs no better than a coin toss when assessing minorities while accurately sorting members of the majority group should be considered blatantly unfair even if its overall prediction accuracy is extremely high. Just consider a college that tosses a coin on minority applicants regardless of their qualifications, while expending diligence to others!

Differences in *classification accuracy* between different groups is a major and underappreciated source of unfairness.

The negative effects of sample size disparities are greatly exacerbated by the existence of cultural differences. Suppose a social network attempted to classify user names into 'real' and 'fake'. Anybody still remember Nymwars? White male American names are pretty straightforward to deal with compared with ethnic

names. In some ethnic groups, names tend to be far more diverse. Fewer people (if any) carry the same name—a typical sign of a 'fake' profile among white Americans.

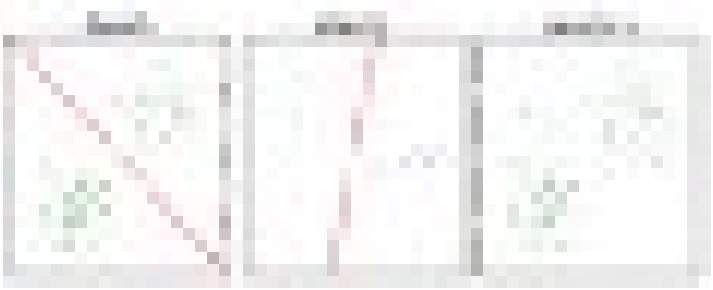


Positively labeled examples are on opposite sides of the classifier for the two groups.

The lesson is that statistical patterns that apply to the majority might be invalid within a minority group. In fact, a variable that's positively correlated with the target in the general population, might be *negatively* correlated with the target in a minority group. As we saw, a real name might be a short common name in one culture and a long unique name in another.

You might say, let's accept the sample size disparity as a fact of nature but do the best we can on dealing with cultural differences. One approach might be to learn multiple classifiers one for each group in an effort to discover those statistical patterns that may be unique to a particular group. Unfortunately, there are several impediments. First, learning and applying a separate classifier for a minority group requires testing for and acting on the protected attribute which might in itself already be considered objectionable. Even if not, the definition of minority is fuzzy and there could be many different overlapping minorities and no straightforward way to determine group membership.

A more significant roadblock has a lot to do with complexity. Here's a toy example. There might be a simple linear function that classifies the majority group correctly and there might be a (different) simple linear function that classifies the minority group correctly, but learning a (non-linear) combination of two linear classifiers is in general a computationally much harder problem. There are excellent algorithms available for learning linear classifiers (e.g., SVM), but no efficient algorithm is known for learning an arbitrary combination of two linear classifiers.

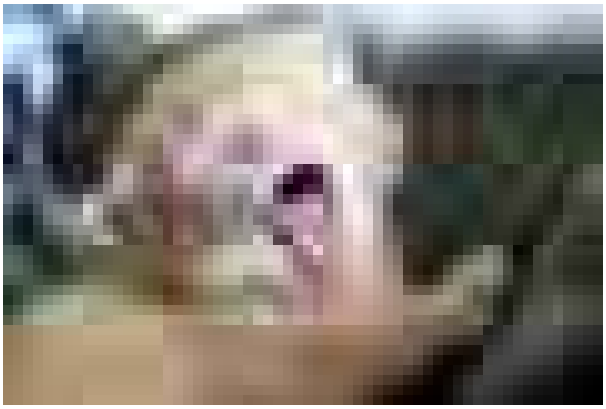


Even if two groups of the population admit simple classifiers, the whole population may not.

Achieving fairness might be computationally expensive if it forces us to look for more complex decision rules. It may also place additional demands on those that engineer the learning process. Since some of the most interesting applications of AI tend to be at the limit of what's currently computationally and humanly feasible, the additional resources necessary for achieving fairness may be limited.

Few models achieve 95% accuracy. The error rates in real-world applications are typically much higher. But even if we get lucky and achieve a 5% error, we typically have no clue what the error means. In the best case the error reflects the intrinsic uncertainty of predicting an outcome given the data and there's no way around it. But it could also be the case that there's a 5% error because we're 50% inaccurate on classifying ethnic names and 100% accurate on classifying the rest. Returning to the previous figure, if we ran soft-margin SVM—a standard learning algorithm—on the whole population, we'd essentially recover the majority solution and it would have low accuracy on the minority despite being seemingly very accurate on the population. Understanding which situation we're in requires a great deal of domain knowledge and there's no principled—let alone automated—methodology for distinguishing noise from modeling errors.

I've painted a fairly grim picture so far. Here's a picture of a cat to lighten up the mood.



A yawning kitten. Adorable.

With that I'm going to end on a more positive note. People are catching on to the fact that fairness is a pretty big problem with data driven decision making. The recent White House report on Big Data raises fairness as an important concern. An increasing number of academics from several disciplines are working on the problem. As one case in point, Solon Barocas and I are organizing a NIPS 2014 workshop called:

FAT ML: Fairness, Accountability and Transparency in Machine Learning

I'm tremendously excited about it. We have a brilliant line-up of

speakers and participants. If you're anywhere near NIPS this year, please join us and help us learn more. I hope to see you there!