Gal Yona Follow
AI / ML & social aspects
Oct 5, 2017 · 10 min read

A Gentle Introduction to the Discussion on Algorithmic Fairness

Recent years have seen the rise of machine learning algorithms: After breaking the benchmark on nearly every imaginable computer vision related tasks, machine learning algorithms are now in our homes and back-pockets, all the time. A less known fact is that they have recently begun replacing human decision makers in a number of more sensitive domains, such as the criminal justice system and the field of medical testing.

the idea of an "AI judge" is still far fetched, but in the US algorithms are already being used to determine a defendant's "risk score"; these scores are then used to inform decisions about bail, sentencing, and parole

One immediate observation that appeared when machine learning algorithms were applied to *human beings* (rather than just vectors representing images), was that the algorithms were not always behaving "fairly". It turned out that training machine learning algorithms with the standard utility-maximization objectives (e.g,

maximizing prediction accuracy on the training data) sometimes resulted in algorithms that behaved in a way in which a human observer will deem unfair, often especially towards a certain minority.

There are many potential reasons for this behavior. One natural suspect is the training data itself: Machine learning algorithms are in essence big computational machines that are trained to recognize and leverage statistical patterns in the data. If the data given to them as "ground truth" contains some bias or historical discrimination, the machines will pick up on this bias and incorporate it into their future predictions. One nice example is the work in [1], in which the authors demonstrated that the famous word embedding model word2vec actually contains implicit gender bias. I find this troubling, because numerous applications now use trained word2vec embeddings as the first pre-processing stage in their machine learning pipelines; this means this bias is being propagated on a daily basis.

Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme *he* occupations

1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss

[1] gives a **quantitative** demonstration that word-embeddings contain **gender biases** in their geometry

One way to overcome this issue is to constantly re-train machine learning models with "fresh" data, under the assumption that historical bias is on a process of correcting itself. I don't know if this is the perfect fix, but there is reason to be hopeful. One recent example I liked is by considering how the top-selling image for the search term "woman" in Getty Image's library of stock photography changed in the last decade: The 2007 most popular image is a naked woman lying on a bed, while the 2017 most popular image of a woman hiking alone on a rocky trail. See [7] for the entire evolution and some other cool examples. If you

think that stock photos—generic images that appear in places like ads, billboards, magazines and blogs—reflect anything about our society, then we're clearly on a positive path.

the top-selling image for the search term **"woman"** in Getty Image's library of stock photography: 2007 (left) versus 2017 (right)

However, a somewhat less expected result is that even with perfectly labeled training data, bias and discrimination can appear in many ways. [6] is a great blog post about the various reasons why machine learning algorithms can be not fair even with completely unbiased training data.

A famous recent example is Google Photos mistakenly labeling two black people as 'gorillas'. Obviously, it's not that the training data Google used for this task contained examples of black people labeled as 'gorillas'. But machine learning algorithms still make errors, and in this case it is a costly and unfair error, since it happened only for blacks and not for whites.

Google Photos mistakenly labels black people as 'gorillas'

Studying *fairness in classification* means studying machine-learning algorithms not only from a perspective of accuracy, but also from a perspective of fairness. Sounds easy, right? Well, surprisingly enough, most of the difficulty actually revolves around **defining what fairness** means. Over the last couple of years, many researchers defined different notions of algorithmic fairness. In many cases these definitions have trade-offs with accuracy (i.e, achieving them means necessarily paying a price in terms of the model's accuracy), but somewhat more unexpected is that many of these definitions have trade-offs within themselves.

In this post, we will work to gradually introduce different notions of algorithmic fairness by working around a concrete example: **filtering résumés for the tech industry**. Specifically, we will consider our feature space X to be some observed features that can be calculated from a person's resume (e.g: name, birth year, address, gender, programming languages, university degree and credits, etc) and the output will be either Y=1 if this person should be brought to be interviewed, Y=0 otherwise. For the purpose of this example, our **protected attribute** (the attribute we would like to guarantee fairness with respect to) is gender, which we assume to be observed and binary.

• **Personal Info**

774-987-4008
john@uptowork.com
uptowork.com/mycv/smith
linkedin.com/in/johnsmith

• **Languages**

Spanish

Intermediate

• **Key skills**

Project Management
Team Management
Budget Management
Change Management
Lean Management

• **Technical skills**

MS Windows Server 2003/2008
Linux/Unix
LAN, WAN, WLAN, SD-WAN
Active Directory
Cisco Routers
SAP

js

John
Smith

IT Project Manager

IT Professional with over **10 years** of experience specializing in **IT department management** for international logistics companies. I can implement effective **IT strategies** at local and global levels. My greatest strength is business awareness, which enables me to **permanently streamline infrastructure and applications**.

• **Experience**

2012-12 - present

IT Project Manager
Software House / San Antonio, TX, USA
Management:
• Responsible for creating, improving, and developing IT project strategies.
• Manage project teams and contractors.
• Plan and monitor IT budgets.
Key IT Project Management:
• Initiate and manage projects that provide new solutions and improvements.
• Supervise timely accomplishment of project objectives.
• Manage key project risks and change.
• Ensure the highest quality of the implemented projects.
• Manage communication between project stakeholders.
• Cross trained more than 30 employees in two months.
Key Achievements:
I managed a key project involving SAP system implementation for the region.
My commitment resulted in the successful migration of 5 new countries to the global IT infrastructure.
I reduced the costs of IT infrastructure maintenance by 5% in 2015.

we will discuss what it means for an algorithm to be "fair" by considering the task of filtering an applicant based on their resume.

First Stop: fairness of the process versus fairness of the outcome

The first distinction we introduce is between the “aware” and the “unaware” approach. An “aware” algorithm will use the information regarding the protected attribute in the process of learning; an “unaware” algorithm will not.

The motivation behind the unaware approach is that being fair with respect to a protected attribute essentially means dis-regarding it. In the resume example, this means the best way to treat men and women the same is to simply not give the algorithm access to this information.

The first observation is that removing the protected attribute alone will often not suffice. In most practical cases, the protected attribute is actually redundantly encoded in the rest of the observed features; most ML algorithms are complicated computational machine that will be able to pick up on and leverage this fact. For example, we might remove the gender feature, only to find out that our classifier gives strong importance to the length of one’s military service. Why? Well, in Israel women often serve 2 years and men serve 3 years. This means that the feature of “length of military service” is almost perfectly correlated with gender; a classifier that uses this feature essentially disobeys the unaware approach.

A quick fix which is often used in practice is to also remove all the other attribute that are highly correlated (e.g above some threshold) with the protected attribute. A more rigorous approach (and a very interesting work in itself) is “learning fair representations” (Zemel et al, 2013 [2]). It uses machine learning to replace the procedure of manually engineering a feature list that conveys no information about the protected attribute.

Let’s assume that we were successful. *Is learning in a way that is completely blind to one’s gender what we really want*? There could be inherent differences between the populations defined by the protected attribute which will in fact render this as undesirable.

In our example, suppose that majoring in Physics in high-school is highly indicative of an applicant’s future success in the tech industry,

even though it's not the actual knowledge of physics that is required. This would therefore be a feature our classifier would look for in resumes. However, it can be claimed that for sociological reasons, young girls in Israel are significantly less inclined than boys their age to study Physics. This will mean that by completely dis-regarding gender information, an algorithm that favors applicants with a background in Physics gives a seemingly unfair advantage to male applicants. This phenomena is often referred to as recovering the "majority solution" (solution here meaning the prototype of an ideal candidate), and it naturally hurts the minority. In our example, a female candidate might be equally qualified but "missed" by the algorithm.

The "aware" approach, on the other hand—that does use gender information—could overcome this by actually picking up on the fact that learning Physics is a stronger signal for future success among males than it is for females.

Another way to look at the aware vs unaware issue as a distinction between fairness of the *process* and fairness of the *outcome*. The unaware approach ensures fairness of the process: it forces the fact that during the learning phase, the algorithm does not in any way treat individuals differently based on their protected attribute. However, the final outcome of this, in what is almost an oxymoron, can actually be less fair towards the protected and non protected sub-groups. The aware approach, on the other hand, uses a process that is not fair (explicitly uses gender information, and learns different classification rules for people of different parts of the population), but can actually reach an outcome that is more fair towards the minorities.

Second Stop: Statistical Parity

One natural way to come up with definitions for algorithmic fairness is to look at how fairness is defined outside the computer science community and formulate those ideas as mathematical definitions, criteria to which we will hold our algorithms.

The US **legal theory** uses the "disparate impact" principle: a practice is considered *illegal discrimination* if it has a "disproportionately adverse" effect on members of a protected group. "disproportionately adverse" is

usually defined using the 80% Rule.

The mathematical equivalence of the disparate impact principle at its most extreme version (allowing no adverse effect on members of the protected group) for binary classification tasks is the Statistical Parity condition: it essentially equalizes the outcomes across the protected and non-protected groups.

$$Pr\left[h(x) = 1 | x \in P^C\right] = Pr\left[h(x) = 1 | x \in P\right]$$

Statistical Parity: equalize outcomes across the protected and non-protected groups

The main criticism against the notion of statistical parity and its variants (such as those that allow some parity, e.g up to 20%, between the two groups) is very natural: Do we really want to equalize the outcomes between the protected and non-protected groups?

In our resume example, this will force us to accept the same number of women and men. There has recently been very public debate on exactly this issue, in light of the computer engineer fired by Google for suggesting women are less suited to certain roles in tech and leadership [3]. Without going into my thoughts on this specific claim, there are clearly many other tasks in which enforcing statistical parity makes very little sense. Consider for example any classification task in which there is a clear causal relationship between the protected attribute and the output variable, e.g: predicting whether an individual will give birth in the next decade. Naturally, enforcing statistical parity here is ludicrous.

Step 3: Expanding to other notions of cross-group calibration

We still want to achieve some notion of equality between the protected and non-protected groups. Instead of equalizing the outcomes themselves, we can equalize some statistics of the algorithm's performance: for example, equalize the error rates across groups. A fair

algorithm for screening resumes, in this view, is one that makes *just as many mistakes on male applicants as it does on female applicants*.

A useful tool for the analysis is a confusion matrix:

> *a confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of a supervised learning algorithm*

	Failure Predicted	Success Predicted	Conditional Procedure Error
Failure – A Positive	a True Positives	b False Negatives	$b/(a+b)$ False Negative Rate
Success – A Negative	c False Positives	d True Negatives	$c/(c+d)$ False Positive Rate
Conditional Use Error	$c/(a+c)$ Failure Prediction Error	$b/(b+d)$ Success Prediction Error	$\frac{(c+b)}{(a+b+c+d)}$ Overall Procedure Error

Confusion Table: A cross-tabulation of the actual outcome by the predicted outcome

This is where the fairness literature can be a little gruesome, because many of the definitions have very subtle differences between them, and it is easy to forget what they all have in common. [5] does a great job at giving detailed examples of how comparing these matrices across the protected and non-protected can be used for defining fairness. Some examples:

> *Treatment equality* is achieved by a classifier that yields a ratio of false negatives and false positives (in the table, c/b or b/c) that is the same for both protected group categories.
>
> *Conditional procedure accuracy equality* is achieved when conditioning on the known outcome, the classifier is equally accurate across protected group categories. This is equivalent to the false negative rate and false positive rate being the same for men and women.

The undeniable, mathematical fact about confusion matrices is that there are relationships between the cell counts—e.g, they must sum to

the total number of observations. This means that the different kinds of fairness—that all share cell counts from the confusion matrix—are also related. In fact, it has already been shown that there are many trade-offs.

Tradeoffs , as their name suggest, mean that there is no *single* winner: it's a zero sum game. What is better, a small improvement in the measure of treatment equality or a small improvement in the measure of conditional use accuracy equality? [5] claims that while it is the responsibility of scientists to bring forth the discussion about the trade-offs, and possibly to design algorithms in which the tradeoffs are explicitly represented and available as turning parameters that can be easily adjusted, it is ultimately up to the stakeholders to determine the tradeoffs.

References

> Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings
>
> Learning Fair Representations

3. Google employee fired over diversity row considers legal action

4. Fairness Through Awareness

5. Fairness in Criminal Justice Risk Assessments: The State of the Art

6. How big data is unfair

7. From Sex Object to Gritty Woman: The Evolution of Women in Stock Photos