

towardsdatascience.com

A Tutorial on Fairness in Machine Learning - Towards Data Science

Ziyuan Zhong

31-39 minuten



This post will be the first post on the series. The purpose of this post is to:

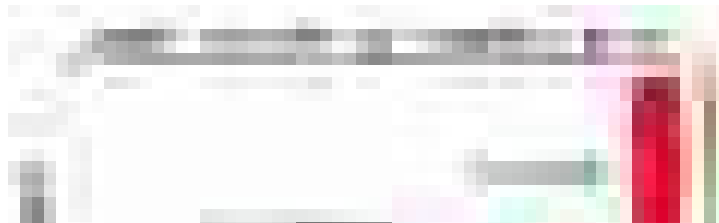
1. give a quick but relatively comprehensive survey of Fair ML.
2. provide references and resources to readers at all levels who are interested in Fair ML.

The content is based on: [the tutorial on fairness](#) given by Solon Baccrasc and Moritz Hardt at NIPS2017, day1 and day4 from [CS 294: Fairness in Machine Learning](#) taught by Moritz Hardt at UC Berkeley and my own understanding of fairness literatures. I highly encourage interested readers to check out the linked NIPS tutorial and the course website.

The current post consists of six parts:

1. Introduction
2. Motivations
3. Causes of bias in ML
4. Definitions of fairness including formulation, motivations, example, and flaws.
5. Algorithms used to achieve those fairness definitions.
6. Summary

Fairness is becoming one of the most popular topics in machine learning in recent years. Publications explode in this field (see Fig1). The research community has invested a large amount of effort in this field. At [ICML](#) 2018, two out of five best paper/runner-up award-winning papers are on fairness. There are also new conferences/workshops on fairness: [FAT/ML](#) starts in 2014, [ACM FAT](#) starts in 2018 and [FairWare](#) in 2018. There are several newly uploaded papers on Fairness on [arxiv](#) every week. Many prestigious researchers in both theoretical and practical ML communities have been involved in the field.



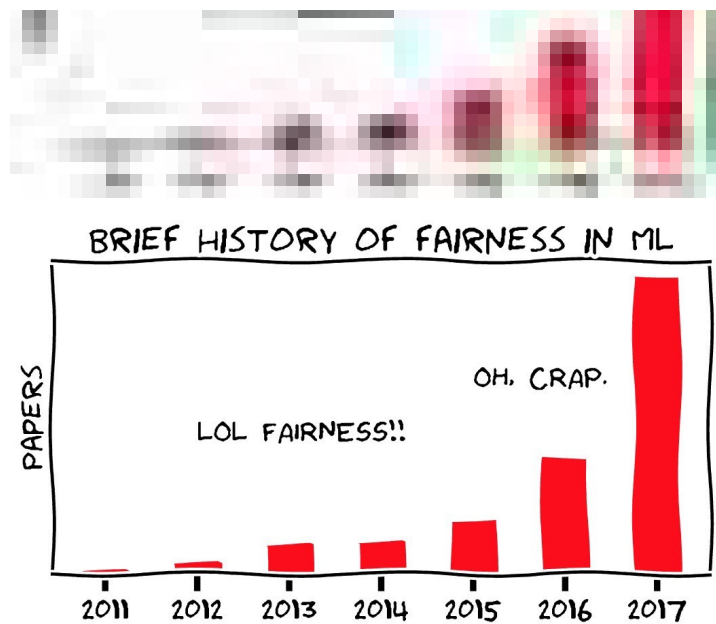



Fig1. The number of publications on fairness from 2011 to 2017

The first question to ask is that why we care about fairness? The main motivation is that it is highly related to our own benefits. We are at an age where many things have become or are becoming automated by ML systems. Self-driving cars have been around the corner and are estimated to be widely used within 5–10 years; employers use ML system to select job applicants; courts in United States use COMPAS algorithm for recidivism prediction; LinkedIn uses ML to rank job candidates queried; Amazon uses recommender system to recommend items and decide the order of items appearing on a page. Netflix uses recommender system to present customized page for every user. Machine learning systems have been an inseparable part of our daily lives. They are becoming even more widely used in the near future as more and more fields begin to integrate AI into their existing practice/products.

Artificial Intelligence is good but it can be used incorrectly. Machine Learning, the most widely used AI techniques, relies heavily on data. It is a common misconception that AI is absolutely objective. AI is objective only in the sense of learning what human teaches. The data provided by human can be highly-biased. It has been found in 2016 that COMPAS, the algorithm used for recidivism prediction produces much higher false positive rate for black people than white people(see Fig2, [Larson et al. ProPublica, 2016](#)).

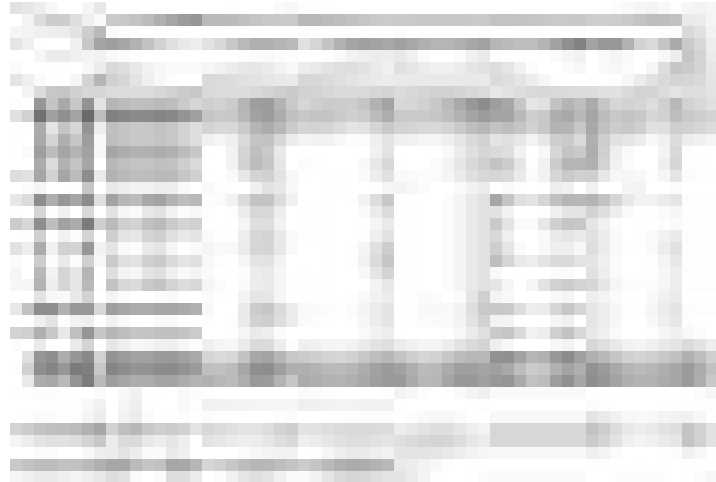


| | WHITE | AFRICAN AMERICAN |
|---|-------|------------------|
| Labeled Higher Risk, But Didn't Re-Offend | 23.5% | 44.9% |
| Labeled Lower Risk, Yet Did Re-Offend | 47.7% | 28.0% |

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. [Source: ProPublica analysis of data from Broward County, Fla.]

Fig2: The bias in COMPAS. (from [Larson et al. ProPublica, 2016](#))

XING, a job platform similar to Linked-in, was found to rank less qualified male candidates higher than more qualified female candidates(see Fig3, [Lahoti et al. 2018](#)).

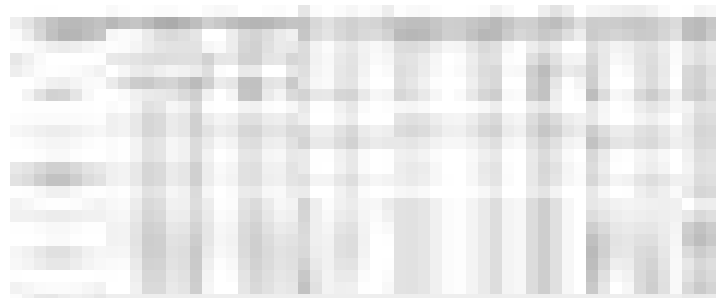


| Search query | Work experience | Education experience | Profile views | Candidate | Xing ranking |
|------------------|-----------------|----------------------|---------------|-----------|--------------|
| Brand Strategist | 146 | 57 | 12992 | male | 1 |
| Brand Strategist | 327 | 0 | 4715 | female | 2 |
| Brand Strategist | 502 | 74 | 6978 | male | 3 |
| Brand Strategist | 444 | 56 | 1504 | female | 4 |
| Brand Strategist | 139 | 25 | 63 | male | 5 |
| Brand Strategist | 110 | 65 | 3479 | female | 6 |
| Brand Strategist | 12 | 73 | 846 | male | 7 |
| Brand Strategist | 99 | 41 | 3019 | male | 8 |
| Brand Strategist | 42 | 51 | 1359 | female | 9 |
| Brand Strategist | 220 | 102 | 17186 | female | 10 |

TABLE II: Top k results on [www.xing.com](#) (Jan 2017) for the job search query "Brand Strategist".

Fig3: The bias in the query for Brand Strategist from XING(from [Lahoti et al. 2018](#)).

Publicly available commercial face recognition online services provided by Microsoft, Face++, and IBM respectively are found to suffer from achieving much lower accuracy on females with darker skin color(see Fig4, [Buolamwini and Gebru, 2018](#)).



| Classifier | Metric | All | F | M | Darker | Lighter | DF | DM | LF | LM |
|------------|---------------|------|------|------|--------|---------|------|------|------|-----|
| Microsoft | PPV(%) | 93.7 | 89.3 | 97.4 | 87.1 | 99.3 | 79.2 | 94.0 | 98.3 | 100 |
| | Error Rate(%) | 6.3 | 10.7 | 2.6 | 12.9 | 0.7 | 20.8 | 6.0 | 1.7 | 0.0 |

| | | | | | | | | | | |
|---------------|---------------|------|------|------|------|------|-------------|-------------|-------------|-------------|
| MSFT | TPR (%) | 93.7 | 96.5 | 91.7 | 87.1 | 99.3 | 92.1 | 83.7 | 100 | 98.7 |
| | FPR (%) | 6.3 | 8.3 | 3.5 | 12.9 | 0.7 | 16.3 | 7.9 | 1.3 | 0.0 |
| Face++ | PPV(%) | 90.0 | 78.7 | 99.3 | 83.5 | 95.3 | 65.5 | 99.3 | 94.0 | 99.2 |
| | Error Rate(%) | 10.0 | 21.3 | 0.7 | 16.5 | 4.7 | 34.5 | 0.7 | 6.0 | 0.8 |
| | TPR (%) | 90.0 | 98.9 | 85.1 | 83.5 | 95.3 | 98.8 | 76.6 | 98.9 | 92.9 |
| | FPR (%) | 10.0 | 14.9 | 1.1 | 16.5 | 4.7 | 23.4 | 1.2 | 7.1 | 1.1 |
| IBM | PPV(%) | 87.9 | 79.7 | 94.4 | 77.6 | 96.8 | 65.3 | 88.0 | 92.9 | 99.7 |
| | Error Rate(%) | 12.1 | 20.3 | 5.6 | 22.4 | 3.2 | 34.7 | 12.0 | 7.1 | 0.3 |
| | TPR (%) | 87.9 | 92.1 | 85.2 | 77.6 | 96.8 | 82.3 | 74.8 | 99.6 | 94.8 |
| | FPR (%) | 12.1 | 14.8 | 7.9 | 22.4 | 3.2 | 25.2 | 17.7 | 5.20 | 0.4 |

Fig4: The bias in commercial face recognition services([Buolamwini and Gebru, 2018](#)). DF, DM, LF, LM stand for: darker skin female, darker skin male, lighter skin female and lighter skin male. PPV, TPR, FPR stand for [predictive positive value](#), true positive rate and false positive rate.

Bias in ML has been almost ubiquitous when the application is involved in people and it has already hurt the benefit of people in minority groups or historically disadvantageous groups. Not only people in minority groups but everyone should care about the bias in AI. If no one cares, it is highly likely that the next person who suffers from biased treatment is one of us.

One would ask: “what causes bias in ML systems?” Essentially, the bias comes from human bias existing in training dataset due to historical reason(s). The following is a list of potential causes([Barocas and Selbst, 2016](#)):

Skewed sample: If by some chance some initial bias happens, such bias may compound over time: future observations confirm prediction and fewer opportunity to make observations that contradict prediction. One example is police record. The record of crimes only come from those crimes observed by police. The police department tends to dispatch more officers to the place that was found to have higher crime rate initially and is thus more likely to record crimes in such regions. Even if people in other regions have higher crime rate later, it is possible that due to less police attention, the police department still record that these regions have lower crime rate. The prediction system trained using data collected in this way tends to have positive bias towards regions with less police.

Tainted examples: Any ML system keeps the bias existing in the old data caused by human bias. For example, if a system uses hiring decisions made by a manager as labels to select applicants rather than the capability of a job applicants (most of time this capability is unobserved for people who are rejected). The system trained using these samples will replicate the bias existing in the manager’s decisions(if there are any). Another example is that word embeddings trained on Google News articles “exhibit female/male gender stereotypes to a disturbing extent” e.g. the relationship between “man” and “computer programmers” was found to be highly similar to that between “woman” and “homemaker” ([Bolukbasi et al. 2016](#)).

Limited features: features may be less informative or reliably collected for minority group(s). If the reliability of the label from a

minority group is much lower than the counterpart from a majority group, the system tends to have much lower accuracy for the prediction of the minority group due to these noise.

Sample size disparity: If the training data coming from the minority group is much less than those coming from the majority group, it is less likely to model perfectly the minority group.

proxies: Even if sensitive attribute(attributes that are considered should not be used for a task e.g. race/gender) is not used for training a ML system, there can always be other features that are proxies of the sensitive attribute(e.g. neighborhood). If such features are included, the bias will still happen. Sometimes, it is very hard to determine if a relevant feature is too correlated with protected features and if we should include it in training.

They can be grouped into the following three problems:

- Discovering unobserved differences in performance:skewed sample, tainted examples
- Sample Coping with observed differences in performance: limited features, sample size disparity
- Understanding the causes of disparities in predicted outcome: proxies

A natural question to ask is “how to define fairness?”, specifically, “How can we formulate fairness such that it can be considered in ML systems”. The first idea is to find legal support and check if there is any definitions that can be used to formulate fairness quantitatively. Anti-discrimination laws in many countries prohibit unfair treatment of people based on sensitive attributes, such as gender or race (Civil Rights Act. Civil Rights Act of 1964, Title VII, Equal Employment Opportunities, 1964.). These laws typically evaluate the fairness of a decision making process using two distinct notions ([Barocas and Selbst, 2016](#)): **disparate treatment** and **disparate impact**. A decision making process suffers from disparate treatment if its decisions are (partly) based on the subject’s sensitive attribute, and it has disparate impact if its outcomes disproportionately hurt (or benefit) people with certain sensitive attribute values (e.g., females, blacks). These two definitions, however, are too abstract for the purpose of computation. As a result, there is no consensus on the mathematical formulations of fairness.

There are many definitions of fairness that have been proposed in the literature (see [Gajane and Pechenizkiy, 2018](#) and [Verma and Rubin, FairWare2018](#)). However, most of them are based on the following six:

- Unawareness
- Demographic Parity
- Equalized Odds
- Predictive Rate Parity

- Individual Fairness
- Counterfactual fairness

where the Demographic Parity, Equalized Odds, and Predictive Rate Parity fall into a larger category called “group fairness”.

Note: The names I choose to use are the relatively popular ones in the literatures. There are no uniform naming conventions for these definitions.

4.0 Setup and Notation

In this section we will only consider the binary classification problem with a single sensitive attribute for simplicity. However, the formulation can be easily extended to other tasks (e.g. regression) with multiple sensitive attributes.

Imagine the problem of predicting if hiring an applicant:

- $X \in \mathbb{R}^d$: quantified features of the applicant(e.g. education, work experience, college GPA, etc.).
- $A \in \{0, 1\}$: a binary sensitive attribute(e.g. majority/minority).
- $C := c(X, A) \in \{0, 1\}$: binary predictor (e.g. hire/reject), which makes decision based on a score $R := r(x, a) \in [0, 1]$.
- $Y \in \{0, 1\}$: target variable(e.g. if the candidate is truly capable of the position).
- We assume X, A, Y are generated from an underlying distribution D i.e. $(X, A, Y) \sim D$.

We also denote $P_0[c] := P[c | A=0]$.

Initially, when we do not impose any fairness constraint, we optimize for accuracy and the best accuracy is achieved when $C(X, S) = Y \forall (X, S, Y) \sim D$.

4.1 Unawareness

This simply means we should not include the sensitive attribute as a feature in the training data. This notion is consistent with **disparate treatment**, which requires to not use the sensitive attribute.

Formulation:

$$C = c(x, A) = c(x)$$

Motivations:

Intuitive, easy to use and legal support(disparate treatment).

Flaws:

The fundamental limitation is that there can be many highly correlated features(e.g. neighborhood) that are proxies of the sensitive attribute(e.g. race). Thus, only removing the sensitive attribute is by no means enough.

4.2 Demographic Parity

Demographic Parity, also called Independence, Statistical Parity, is one of the most well-known criteria for fairness.

Formulation:

C is independent of A: $P_0 [C = c] = P_1 [C = c] \forall c \in \{0, 1\}$

In our example, this means the acceptance rates of the applicants from the two groups must be equal. In practice, there are two approximate forms that relax this equality (assume group 1 is the advantageous group i.e. has higher probability to be hired when no fairness problem is considered):

- $P_0 [C=1]/P_1 [C=1] \geq 1-\epsilon$

The p% rule is defined as satisfying this inequality when $\epsilon=p/100$ ([Zafar et al. AISTATS2017](#)).

- $|P_0 [C=1]-P_1 [C=1]| \leq \epsilon$ where $\epsilon \in [0, 1]$.

Motivations:

- Legal Support: “**four-fifth rule**” prescribes that a selection rate for any disadvantaged group that is less than four-fifths of that for the group with the highest rate. In our formulation, this is equivalent to satisfying 80% rule. If this rule is violated, justification as being job-related or a business necessity must be provided. “Business necessity means that using the procedure is essential to the safe and efficient operation of the business — and there are no alternative procedures that are substantially equally valid and would have less adverse impact” (source: [Adverse Impact Analysis / Four-Fifths Rule](#)).
- There are some papers that argue the enforcement of such criteria in short term benefits building up the reputation of the disadvantaged minority group in the labor market in the long run ([Hu and Chen, WWW2018](#)).

Flaws:

- This definition ignores any possible correlation between Y and A. In particular, it rules out perfect predictor $C=Y$ when base rates are different (i.e. $P_0 [Y=1] \neq P_1 [Y=1]$)
- laziness: if we hire the qualified from one group and random people from the other group, we can still achieve demographic parity.

4.3 Equalized odds

Equalized odds, also called Separation, Positive Rate Parity, was first proposed in Hardt, Price and Srebro, 2016 and [Zafar et al. WWW2017](#).

Formulation:

C is independent of A conditional on Y:

$$P_0 [C = r \mid Y = y] = P_1 [C = r \mid Y = y] \forall r, y$$

A weaker notion is:

$$P_0 [C \neq Y] = P_1 [C \neq Y]$$

which is called **Accuracy Parity**. The limitation of this weaker notion is that we can trade false positive rate of one group for false negative rate of another group. Such trade is not desirable sometimes (e.g. trade rejecting ($C=0$) qualified applicants ($Y=1$) from group 1 ($A=0$) for accepting ($C=1$) unqualified people ($Y=0$) from group 2 ($A=1$)).

In many applications (e.g. hiring), people care more about the true positive rate than true negative rate so many works focus on the following relaxed version:

$$P_0 [C = 1 | Y = 1] = P_1 [C = 1 | Y = 1]$$

which is called **Equality of Opportunity**.

In our example, this is to say we should hire equal proportion of individuals from the qualified fraction of each group.

Motivations:

- Optimality compatibility: $C=Y$ is allowed.
- Penalize laziness: it provides incentive to reduce errors uniformly in all groups.

Flaws:

It may not help closing the gap between two groups. For example, imagine group A has 100 applicants and 58 of them are qualified while group B also have 100 applicants but only 2 of them are qualified. If the company decides to accept 30 applicants and satisfies equality of opportunities, 29 offers will be conferred to group A while only 1 offer will be conferred to group B. If the job is a well-paid job, group A tends to have a better living condition and affords better education for their kids, and thus enable them to be qualified for such well-paid jobs when they grow up. The gap between group A and group B will tend to be enlarged over time.

4.4 Predictive Rate Parity

Predictive Rate Parity, also called Sufficiency, appeared in [Zafar et al. WWW2017](#) (I am not sure about the first literature that deals with it).

Formulation:

Y is independent of A conditional on C :

$$P_0 [Y = y | C = c] = P_1 [Y = y | C = c] \quad \forall y, c \in \{0, 1\}$$

This is equivalent to satisfying both

$$P_0 [Y = 1 | C = 1] = P_1 [Y = 1 | C = 1] \text{ and}$$

$$P_0 [Y = 0 | C = 0] = P_1 [Y = 0 | C = 0],$$

which are called **Positive Predictive Parity** and **Negative Predictive Parity** respectively.

In our example, this is to say that if the score returned from a prediction algorithm (used to determine the candidate's eligibility of the job) for a candidate should reflect the candidate's real capability

of doing this job. It is consistent with the employer's benefit.

Motivations:

- Optimality compatibility: $C=Y$ satisfies Predictive Rate Parity.
- Equal chance of success($Y=1$) given acceptance($C=1$).

Flaws:

- The flaw is similar to the flaw of equality of opportunities: it may not help closing the gap between two groups. The reasoning is similar as before.

4.5 Individual Fairness

Individual fairness is a relatively different notion. The previous three criteria are all group-based while individual fairness, as its name suggests, is individual-based. It was first proposed in [Fairness Through Awareness](#) by Cynthia Dwork et al. in 2012, which is one of the most important foundational papers in the field. The notion of individual fairness emphasizes on that: similar individuals should be treated similarly.

Formulation:

Denote O to be a measurable space and $\Delta(O)$ to be the space of the distribution over O . Denote $M: X \rightarrow \Delta(O)$ to be a map that maps each individuals to a distribution of outcomes. The formulation is then:

$$D(M(X), M(X')) \leq d(X, X')$$

where $X, X' \in R^d$ are two input feature vectors, and D and d are two [metric functions](#) on the input space and the output space respectively. See Fig5 for an illustration.



Fig5: illustration of individual fairness

Motivation:

Rather than focusing on group, as individuals, we tend to care more about the individuals. Besides, individual fairness is more fine-grained than any group-notion fairness: it imposes restriction on the treatment for each pair of individuals.

Flaws:

It is hard to determine what is an appropriate metric function to measure the similarity of two inputs ([Kim et al. FATML2018](#)). In our case, it is hard to quantify the difference between two job candidates. Imagine three job applicants, A, B and C. A has a bachelor degree and 1 year related work experience. B has a master degree and 1 year related work experience. C has a master degree but no related work experience. Is A closer to B than C? If so, by how much? Such question is hard to answer since we cannot have the performance of A, B, and C (we cannot hire all three). Otherwise we can apply techniques in a field called [metric learning](#). It becomes even worse when the sensitive attribute(s) comes into the play. If we should and how to count for the difference of group membership in our metric function?

4.6 Counterfactual fairness

Counterfactual fairness was proposed in [Russell et al., NIPS2017](#). It provides a possible way to interpret the causes of bias.

Formulation:

$$P[C_{\{A \leftarrow 0\}} = c | X, A = a] = P[C_{\{A \leftarrow 1\}} = c | X, A = a]$$

A counterfactual value replaces the original value of the sensitive attribute. The counterfactual value propagates “downstream” the [causal graph](#) (see Fig6 for an example) via some structural equations. Everything else that are not descendant of the sensitive attribute remains the same.

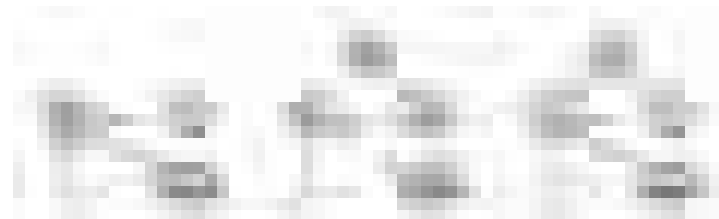


Fig6: some possible causal graphs

Motivations: 4.1 is far from being enough due to many correlated features. 4.2–4.4 are all observational fairness criteria. They cannot be used to find the cause of the unfairness phenomenon. 4.5 has fundamental limitation of finding the proper metric. Counterfactual fairness solved all these problems.

Counterfactual fairness provides a way to check the possible impact of replacing only the sensitive attribute. It provides a way of explaining the impact of bias via a causal graph. Fig6 shows several possible graphs in the scenario of applying to college. Notice that when we replace the sensitive attribute, all the other features that are correlated with it will also be influenced. In Fig6, if the sensitive attribute (race) is changed, the education score as well as work score will also change.

Flaws:

The idea is very ideal. In practice, it is hard to reach a consensus in terms of what the causal graph should look like and it is even

harder to decide which features to use even if we have such a graph (we may suffer large loss on accuracy if we eliminate all the correlated features).

4.7 The Impossibility theorem of fairness

It turns out that any two of the three criteria in 4.2–4.4 are mutually exclusive except in non-degenerate cases.

Demographic Parity VS Predictive Rate Parity

If A is dependent of Y , then either Demographic Parity holds or Predictive Rate Parity but not both.

Proof:



Demographic Parity VS Equalized Odds

If A is dependent of Y and C is dependent of Y , then either Demographic Parity holds or Equalized Odds but not both.

Proof:



Equalized Odds VS Predictive Rate Parity

Assume all events in the joint distribution of (A, C, Y) have positive probability. If A is dependent of Y , either Equalized Odds holds or Predictive Rate Parity but not both.

Proof: ([Wasserman](#) Theorem 17.2)



Literatures mainly focus on the proof for the third one. Proofs can be found in [Chouldechova, 2016](#) and [Kleinberg et al. 2016](#), where the later gives a proof given a weaker assumption.

The following is an example that illustrates the third mutual exclusion (Equalized Odds VS Predictive Rate Parity). Imagine we have the following outcomes and predictions after optimizing our classifier without any fairness constraints (see Fig7). We get the predictions for group a all correct but makes one false positive mistake on group b.



Fig7: illustration of impossibility theorem(original)

Since we want to preserve equalized odds, we decide to make two false positive mistakes on a as well. Now the true positive rates as well as true negative rates are equal: both have $1/2$ and 1 (See Fig8).





Fig8: illustration of impossibility theorem(equalized odds is preserved)

However, although positive predictive parity is also preserved, negative predictive parity is violated with this setting. It is not possible to preserve negative predictive parity without sacrificing equalized odds/positive predictive parity(see Fig9).



Fig9: illustration of impossibility theorem(PPV is satisfied but NPV is not)

The essence of COMPAS debate is similar to this toy example. ProPublica's main charge is that black defendants face higher false positive rate i.e. it violates the equality of opportunity and thus equalized odds. Northpointe's main defense is that scores satisfies predictive rate parity.

4.7 Trade-off between fairness and accuracy

The impact of imposing the above constraints on the accuracy truly depends on the dataset, the fairness definition used as well as the algorithms used. In general, however, fairness hurts accuracy because it diverts the objective from accuracy only to both accuracy and fairness. Therefore, in reality, a trade-off should be made(see Fig10 for an illustration).



Fig10: trade-off between accuracy and demographic parity on a linear classification problem ([Zafar et al. AISTATS2017](#))

There are many algorithms that claim to help improve fairness. Most of them fall into three categories: preprocessing, optimization at training time, and post-processing.

5.1 Preprocessing

The idea is that we want to learn a new representation Z such that it removes the information correlated to the sensitive attribute A and preserves the information of X as much as possible and ([Zemel et al. ICML2013](#), [Louizos et al. ICLR2016](#), [Lum and Johndrow 2016](#), [Adler et al. 2016](#), [Calmon et al. NIPS2017](#), [Barrio et al. 2018](#)). The downstream task (e.g. classification, regression, ranking) can thus use the “cleaned” data representation and thus produce results that preserve demographic parity and individual fairness(if proper metric is given). See Fig11 for an illustration.

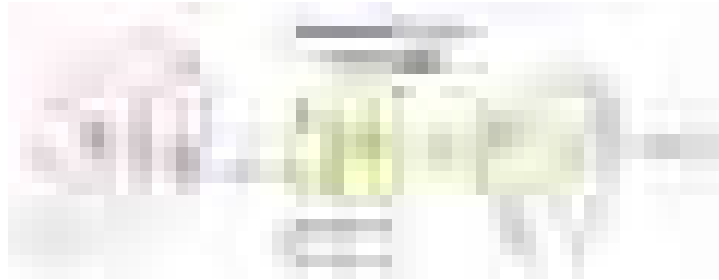


Fig11: Illustration of preprocessing

Example:

The following algorithm was proposed in [Zemel et al. ICML2013](#).

Notations(slightly different from the original paper):

$X \in \mathbb{R}^{d \times N}$: the training data.

Y : a binary random variable representing the classification decision for an individual.

$X^+ := \{x \in X \mid A = 1\}$: the positive training data.

Z : a multinomial random variable, where each of the $K \in \mathbb{Z}^+$ values represents a “prototype”. Associated with each prototype is a vector v_k in the same space as the individuals x .

The idea is to represent each data point x as a weighted linear combination of K prototypes to satisfy demographic parity, and keep original information and accuracy as much as possible. The loss function reflects this idea.

Define the softmax version of the probability of an element being a particular prototype to be

$$p_k = \frac{\exp(-d(x, v_k))}{\sum_{k=1}^K \exp(-d(x, v_k))}$$

where d is a distance measure function(e.g. l_2 distance). In this paper, it is defined as a weighted l_2 distance function:

Define the prediction for y_n , based on marginalizing over each prototype’s prediction for Y to be:

where L_z is to regularize demographic parity, L_x is the reconstruction error and L_y counts the prediction loss. The associated A_z, A_x, A_y are hyper-parameters to balance these losses.

In the training phase, v, w, α are optimized jointly via L-BFGS to

minimize the objective L . Hyper-parameters are selected via grid-search. It must be noted that the objective is non-convex and thus does not guarantee optimality.

Pros:

- Preprocessed data can be used for any downstream task.
- No need to modify classifier.
- No need to access sensitive attributes at test time.

Cons:

- In general, preprocessing can only be used for optimizing Statistical Parity or Individual Fairness(if the metric is given) because it does not have the information of label Y .
- Inferior to the other two methods in terms of performance on accuracy and fairness measure.

5.2 Optimization at Training Time

The most intuitive idea is to add a constraint or a regularization term to the existing optimization objective. Most works in literatures fall into this category. Such methods can be used to optimize for any fairness definition ([Calders et al. 2009](#), [Woodsworth et al. 2017](#), [Zafar et al. AISTATS2017](#), [Zafar et al. WWW2017](#), [Agarwal et al. ICML2018](#)). The method to optimize counterfactual fairness also falls into this category([Russell et al., NIPS2017](#)).

Example:

The following algorithm was proposed in [Zafar et al. WWW2017](#).

Notations:

$x \in \mathbb{R}^d$: non-sensitive attributes.

$y \in \{-1, 1\}$: class labels.

$\hat{y} \in \{-1, 1\}$: predicted labels.

θ : the parameters to learn.

$L(\theta)$: the original convex loss.

$d_\theta(x)$: the signed distance from the feature vector to the decision boundary.

$f_\theta(x)$: the classifier function. $f_\theta(x)=1$ if $d_\theta(x) \geq 0$ and -1 otherwise.

$z \in \{0, 1\}$: the sensitive attribute.

The paper mainly deals with the following three fairness constraints:



One or all three of them can be added as constraints to the original optimization problem.

It should be noted that the equality of false negative rates implies the equality of true positive rates so this implies the equality of opportunity. The equality of false positive rates implies the equality of true negative rates. The equality of overall misclassification rates implies the equality of classification rate, which is accuracy parity, a weaker notion of equalized odds. Here, we take false positive rate (the constraint is the same as equality of opportunity) as an example. It has the following optimization formulation:

$$\min_{\mathbf{w}, b} \frac{1}{N} \sum_{i=1}^N \ell(\mathbf{w} \cdot \mathbf{x}_i + b, y_i) \quad \text{subject to} \quad \frac{1}{N_1} \sum_{i \in \mathcal{S}_1} \ell(\mathbf{w} \cdot \mathbf{x}_i + b, y_i) = \frac{1}{N_0} \sum_{i \in \mathcal{S}_0} \ell(\mathbf{w} \cdot \mathbf{x}_i + b, y_i)$$

The problem is that the constraints make the optimization intractable. As a result, the paper relaxes the constraints. It uses the covariance between the users' sensitive attribute(s) and the signed distance between the users' feature vectors and the classifier decision boundary to act as a proxy to capture the relation between the sensitive attribute and group-level (conditional) predictions:

$$\min_{\mathbf{w}, b} \frac{1}{N} \sum_{i=1}^N \ell(\mathbf{w} \cdot \mathbf{x}_i + b, y_i) \quad \text{subject to} \quad \left| \frac{1}{N_1} \sum_{i \in \mathcal{S}_1} (\mathbf{w} \cdot \mathbf{x}_i + b) - \frac{1}{N_0} \sum_{i \in \mathcal{S}_0} (\mathbf{w} \cdot \mathbf{x}_i + b) \right| \leq c$$

After the relaxation the formulation now becomes:

$$\min_{\mathbf{w}, b} \frac{1}{N} \sum_{i=1}^N \ell(\mathbf{w} \cdot \mathbf{x}_i + b, y_i) \quad \text{subject to} \quad \left| \frac{1}{N_1} \sum_{i \in \mathcal{S}_1} (\mathbf{w} \cdot \mathbf{x}_i + b) - \frac{1}{N_0} \sum_{i \in \mathcal{S}_0} (\mathbf{w} \cdot \mathbf{x}_i + b) \right| \leq c$$

where the covariance threshold $c \in \mathbb{R}^+$ controls how adherent to equality of opportunity.

Such formulation is still non-convex so next we will convert these constraints into a Disciplined Convex-Concave Program (DCCP), which can be solved efficiently by leveraging recent advances in convex-concave programming ([Shen et al. 2016](#)).

First, we can rewrite the fairness constraints as:

$$\frac{1}{N_1} \sum_{i \in \mathcal{S}_1} (\mathbf{w} \cdot \mathbf{x}_i + b) \sim \frac{1}{N_0} \sum_{i \in \mathcal{S}_0} (\mathbf{w} \cdot \mathbf{x}_i + b)$$

where \sim denotes " \geq " or " \leq ". We drop the constant $1/N$ for simplicity. Since $z \in \{0, 1\}$, we can split the sum into two terms:

$$\sum_{i \in D_0} z_i = \sum_{i \in D_1} z_i \quad (10)$$

where D_0 and D_1 are the subsets of D s.t. $z=0$ and $z=1$ respectively. Define $N_0=|D_0|$ and $N_1=|D_1|$, then $\overline{z} = N_1/N$. The constraint now becomes:

$$\sum_{i \in D} z_i = N \overline{z} \quad (11)$$

which, given that g_θ is convex in θ (by assumption), results into a convex-concave function. Thus, the optimization now becomes:

$$\min_{\theta} L(\theta) \text{ s.t. } \sum_{i \in D} z_i = N \overline{z} \quad (12)$$

which is a Disciplined Convex-Concave Program (DCCP) for any convex loss $L(\theta)$, and can be efficiently solved using well-known heuristics such as the one proposed in [Shen et al.2016](#).

Pros:

- Good performance on accuracy and fairness measures.
- May higher flexibility to choose the trade-off between accuracy and fairness measures (This depends on specific algorithm).
- No need to access sensitive attributes at test time.

Cons:

- Method in this category is task specific.
- Need to modify classifier, which may not be possible in many scenarios.

5.3 Post-processing

Such methods attempt to edit posteriors in a way that satisfies fairness constraints. It can be used to optimize most fairness definitions(except counterfactual fairness). The basic idea is to find a proper threshold using the original score function R for each group. As preprocessing, no retraining/changes is needed for the classifier([Feldman 2015](#), [Hardt et al. NIPS2016](#)).

Example (Statistical parity/Equality of opportunities):

The following algorithm was proposed in [Hardt et al. NIPS2016](#).

Given a classifier, we can draw the corresponding ROC curves for both groups. Next, we can find threshold based on the ROC curves (See Fig12 for an illustration).



Fig12: Finding the optimal equalized odds predictor (left), and equal opportunity predictor (right) ([Hardt et al. NIPS2016](#))

Equalized odds(both true positive rates and false positive rates are equal) is only satisfied when the ROC curves of the two groups intersect, as shown in the left graph; equality of opportunity, as a weaker notion, can be satisfied by taking threshold such that the true positive rates of the two groups are equal, as shown in the right graph.

Pros:

- Can be applied after any classifiers.
- Relatively good performance especially fairness measures.
- No need to modify classifier.

Cons:

- Require test-time access to the protected attribute
- Lack the flexibility of picking any accuracy–fairness tradeoff.

5.4 Experiment

It should be noted that the experiment result in this section is from [Agarwal et al. ICML2018](#). Their method falls into the category of optimization at training time. I did not cover their method since it takes time to fully explain their method and the theoretical guarantee their method provide. I may introduce their paper as a separate post in the future.

Dataset:

The following four datasets were used for experiments. All data sets have binary protected attributes except for *adult4*, which has two sensitive attributes(gender, race) and thus four protected attribute values (both attributes were binarized for simplicity).

The adult income data set ([Lichman, 2013](#))

- size: 48,842 examples
- task: predict if someone's income is >\$50k/year
- sensitive attribute: gender(male/female) or gender(male/female) and race(white/non-white)

[ProPublica's COMPAS recidivism data](#)([Larson et al. ProPublica, 2016](#))

- size: 7,918 examples
- task: predict recidivism from someone's criminal history, jail and prison time, demographics, and COMPAS risk scores
- sensitive attribute: race(white/black)

Law School Admissions Council's National Longitudinal Bar Passage Study ([Wightman, 1998](#))

- size: 20,649 examples
- task: predict someone's eventual passage of the bar exam
- sensitive attribute: race(white/black)

The Dutch census data set ([Dutch Central Bureau for Statistics, 2001](#))

- size: 60,420 examples
- task: predict if someone has a prestigious occupation
- sensitive attribute: gender(male/female)

Experiment results:

A comparison of the performance of several methods including preprocessing(reweighting and relabeling methods proposed in [Kamiran and Calders, 2012](#)), optimization at training time(reduction: grid and reduction: EG in [Agarwal et al. ICML2018](#)), post-processing([Hardt et al. NIPS2016](#)) was conducted(See Fig13).



Fig13: *Classification error versus constraint violation on test examples with respect to Demographic Parity(DP) and Equalized Odds(EO).* The curves plot the Pareto frontiers of several methods. Markers correspond to the baselines. Vertical dashed lines are used to indicate the lowest constraint violations. ([Agarwal et al. ICML2018](#))

The result shows that all the methods were able to substantially reduce or remove disparity without much impact on classifier accuracy(except on Dutch census data set). For demographic parity, the reduction methods(belonging to optimization at training time) uniformly achieve the lowest constraint violations, outperforming or matching the baselines. The post-processing algorithm performs well for small violations. Reweighting and

relabeling(belonging to preprocessing) are the worst.

5.5 Discussions

In practice, without surprise, the method that achieves the best trade-off between accuracy and fairness is via optimization at training time. However, preprocessing and post-processing methods grant the ability to preserve fairness without modifying the classifiers. Such feature is desirable when we do not have power to modify the classifiers.

It must be noted that all the previous methods we discussed require information about the sensitive group during training time. The post-processing method even needs it at test-time. This may not be possible when people do not disclose their identities. There is a recent work along this line and optimizes accuracy parity using techniques from robust statistics without knowing the sensitive attribute(s) ([Hashimoto et al. ICML2018](#)).

- Fairness becomes a very popular topic in ML community in recent years.
- Fairness matters because it has impact on everyone's benefit.
- Unfairness in ML systems is mainly due to human bias existing in the training data.
- No consensus on "the best" definition of (un-)fairness exists
- Observational criteria can help discover discrimination, but are insufficient on their own. Causal viewpoint can help articulate problems, organize assumptions.
- Any two of the three group fairness definitions demographic parity, equalized odds, and predictive rate parity cannot be achieved at the same time except in degenerate cases.
- Trade-off between accuracy and fairness usually exists.
- There are three streams of methods: preprocessing, optimization at training time, and post-processing. Each has pros and cons.
- Most fair algorithms use the sensitive attributes to achieve certain fairness notions. However, such information may not be available in reality.

In future posts: I consider to talk about fair ML on settings beyond binary classification, recent progress of fair ML, counterfactual fair ML using causal inference in greater details or maybe greater details of theoretical results.

Thank you for your reading! This is my first post on Medium. Errors and Omissions are highly likely so please let me know and I will correct them as soon as possible! I will really appreciate your support. Please motivate me to write more, higher-quality posts in the future.