Background on Fairness Machine Learning

Contents

About

This reference introduces concepts, methods, and libraries for measuring fairness in machine learning (ML) models as it relates to problems in healthcare. This is a revamped version of the tutorial presented at the KDD 2020 Tutorial on Fairness in Machine Learning for Healthcare, the notebook for which can be found here: /docs/publications/KDD2020-FairnessInHealthcareML-TutorialNotebook.ipynb.

There are abundant other publications covering the theoretical basis for fairness metrics, and many resources both online and academic covering the details of specific fairness measures (See References (bottom) and Additional Resources (bottom), or Our Resources Page for just a few). Many of these otherwise excellent references stop short of discussing edge cases and the practical and philosophical considerations raised when evaluating real models for real customers. Here we attempt to bridge that gap.

All Models Are Biased

All machine learning models can be assumed to hold biases, just as all humans hold biases, and all humans fall ill at some point in their lives. The motivation that drives us to study and prevent the harm caused by human illness drives us to prevent the harm caused by innate biases. That means building models that provide fair representation for all demographics. This starts with measurement and evaluation.

Framing the Problem

Background

Fairness in Machine Learning

In issues of social justice, discrimination is the unjustified, differential treatment of individuals based on their sociodemographic status [Romei and Ruggieri 2014]. A "fair" model could be considered one that does not discriminate.

The "socially salient" sociodemographic groups [Speicher 2018] about which discrimination is measured are known as ***protected attributes***, *sensitive attributes*, or *protected features*.

Disparity

The term "discrimination" typically evokes direct or deliberate action to disadvantage one race, religion, or ethnicity. This kind of disparity is known as *disparate treatment*. However, a more insidious form of discrimination can occur when ostensibly unbiased practices result in the – perhaps unconscious – unfair treatment of a socially disadvantaged group. This is known as *disparate impact*.

Disparate impact in a machine learning model originates from bias in either the data or the algorithms. A popular example is the prejudicially biased data used for recidivism prediction. Due to disparate socioeconomic factors and systemic racism in the United States, blacks have historically been (and continue to be) incarcerated at higher rates than whites [NAACP]. Not coincidentally, blacks are also exonerated due to wrongful accusation at a considerably higher rate than whites [NAACP]. A recidivism model that fails to adjust for circumstances such as these will predict a higher rate of recidivism among blacks.

Machine learning models can also be a source of disparate impact in their implementation, through unconscious human biases that affect the fair interpretation or use of the model's results. This tutorial does not cover measurement of fairness at implementation. However, if you are interested in fair implementation, we recommend looking at Google's Fairness Indicators.

Harms

In evaluating the potential impact of an ML model, it can be helpful to first clarify what specific harm(s) can be caused by the model's failures. In the context of fairness, machine learning "harms" are commonly observed to fall into one of two categories.

- **Allocative Harm:** functionality promoting unfair allocation of finite resources
- **Representational Harm:** functionality promoting the continued marginalization of some groups
  - Examples include:
    - Quality of Service: allocating higher insurance payouts for males than for females
    - Stereotyping: service more likely to show advertising for bail bonds to dark skinned men
    - Under-Representation: image search for "doctor" returning mostly images of white men
    - Recognition: facial recognition mistakenly and offensively labeling a person as an animal

References: The Trouble with Bias Kate Crawford, NIPS2017

Defining and Measuring Fairness

The following section defines common fairness measures that are used elsewhere in the notebook. Skip ahead to Part 3 for an example of how these measures are applied.

Convenient Charts of Fairness Measures

Definitions of Fairness

There are six common metrics for determining whether a model is considered "fair": Equal Treatment ("**Unawareness**"), **Demographic Parity**, **Equalized Odds**, **Predictive Parity**, **Individual Fairness**, and **Counterfactual Fairness**.

Statistical Criteria for Fairness Metrics

Metric	Statistical Criteria	Definition	Description
Demographic Parity	Statistical Independence	$R \perp G$	sensitive attributes (A) are statistically independent of the prediction result (R)
Equalized Odds	Statistical Separation	$R \perp A\|Y$	sensitive attributes (A) are statistically independent of the prediction result (R) given the ground truth (Y)

Metric	Statistical Criteria	Definition	Description
Predictive Parity	Statistical Sufficiency	$Y \perp A\|R$	sensitive attributes (A) are statistically independent of the ground truth (Y) given the prediction (R)

Definitions of Fairness

Category	Metric	Definition	Weakness	References
Group Fairness	**Demographic Parity**	A model has **Demographic Parity** if the predicted positive rates (selection rates) are approximately the same for all protected attribute groups. $\frac{P(\hat{y}=1\|unprivileged)}{P(\hat{y}=1\|privileged)}$ Harms Addressed: Allocative	Historical biases present in the data are not addressed and may still bias the model.	Zafar *et al* (2017)

Category	Metric	Definition	Weakness	References
	Equalized Odds	Odds are equalized if $P(+)$ is approximately the same for all protected attribute groups. **Equal Opportunity** is a special case of equalized odds specifying that $$P(+\|y = 1)$$ is approximately the same across groups. Harms Addressed: Allocative, Representational	Historical biases present in the data are not addressed and may still bias the model.	Hardt *et al* (2016)
	Predictive Parity	This parity exists where the Positive Predictive Value is approximately the same for all protected attribute groups. Harms Addressed: Allocative, Representational	Historical biases present in the data are not addressed and may still bias the model.	Zafar *et al* (2017)
Similarity-Based Measures	**Individual Fairness**	Individual fairness exists if "similar" individuals (ignoring the protected attribute) are likely to have similar predictions. Harms Addressed: Representational	The appropriate metric for similarity may be ambiguous.	Dwork (2012), Zemel (2013), Kim *et al* (2018)

Category	Metric	Definition	Weakness	References
	Unawareness	A model is unaware if the protected attribute is not used.	Removal of a protected attribute may be ineffectual due to the presence of proxy features highly correlated with the protected attribute.	Zemel *et al* (2013), Barocas and Selbst (2016)
Causal Reasoning	**Counterfactual Fairness** *	Counterfactual fairness exists where counterfactual replacement of the protected attribute does not significantly alter predictive performance. This counterfactual change must be propogated to correlated variables. Harms Addressed: Allocative, Representational	It may be intractable to develop a counterfactual model.	Russell *et al* (2017)

** Note that this tutorial will not elaborate the details of Counterfactual Fairness since the libraries used do not have built-in functionality for it. For an example of Counterfactual Fairness, see "ThemisML" by Bantilan (2018).*

Group Fairness Measures

Demographic Parity

A model has **Demographic Parity** if the predicted positive rates (selection rates) are approximately the same for all groups of the protected attribute. Two common measures are the Statistical Parity Difference and the Disparate Impact Ratio.

The *Statistical Parity Difference* is the difference in the probability of prediction between the two groups. A difference of 0 indicates that the model is perfectly fair relative to the protected attribute (it favors neither the privileged nor the unprivileged group). Values between -0.1 and 0.1 are considered reasonably fair.

> $statistical_parity_difference = P(\hat{y} = 1\ |\ unprivileged) - P(\hat{y} = 1\ |\ privileged) $

The *Disparate Impact Ratio* is the ratio between the probability of positive prediction for the unprivileged group and the probability of positive prediction for the privileged group. A ratio of 1 indicates that the model is fair relative to the protected attribute (it favors neither the privileged nor the unprivileged group). Values between 0.8 and 1.2 are considered reasonably fair.

> $$disparate\_impact\_ratio = \frac{P(\hat{y} = 1 \mid unprivileged)}{P(\hat{y} = 1 \mid privileged)} = \frac{selection\_rate(\hat{y}_{unprivileged})}{selection\_rate(\hat{y}_{privileged})}$$

Equal Odds

Odds are equalized if P(+) is approximately the same for all groups of the protected attribute.

The *Equalized Odds Difference* is the greater between the difference in TPR and the difference in FPR. This provides a comparable measure to the Average Odds Difference found in AIF360. A value of 0 indicates that all groups have the same TPR, FPR, TNR, and FNR, and that the model is "fair" relative to the protected attribute.

> $ equalized_odds_difference = max((FPR_{unprivileged} - FPR_{privileged}), (TPR_{unprivileged} - TPR_{privileged}))$

The *Equalized Odds Ratio* is the smaller between the TPR Ratio and FPR Ratio, where the ratios are defined as the ratio of the smaller of the between-group rates vs the larger of the between-group rates. A value of 1 means that all groups have the same TPR, FPR, TNR, and FNR. This measure is comparable to the Equal Opportunity Difference (found in AIF360).

> $ equalized_odds_ratio = min(\dfrac{FPR_{smaller}}{FPR_{larger}}, \dfrac{TPR_{smaller}}{TPR_{larger}})$

Equal Opportunity Difference (or Ratio) compares the recall scores (TPR) between the unprivileged and privileged groups.

> $$equal\_opportunity\_difference = recall(\hat{y}_{unprivileged}) - recall(\hat{y}_{privileged})$$

Measures of Disparate Performance

These measures evaluate whether model performance is similar for all groups of the protected attribute.

The *Positive Predictive Parity Difference (or Ratio)* compares the Positive Predictive Value (PPV), aka. precision, between groups.

> $positive\_predictive\_parity\_difference = precision(\hat{y}_{unprivileged}) - precision(\hat{y}_{privileged})$

The *Balanced Accuracy Difference (or Ratio)* compares the Balanced Accuracy between groups, where balanced accuracy is the mean of the sensitivity and specificity. **Since many models are biased due to data imbalance, this can be an important measure.**

> $balanced\_accuacy\_difference = (Sensitivity_{unprivileged} + Specificity_{unprivileged})/2 - (Sensitivity_{privileged} + Specificity_{privileged})/2$

Comparing Group Fairness (Statistical) Measures

The highlighted row in the cell above indicates that the Disparate Impact ratio is out of range, but what is that range and how is it determined? In 1978, the United States Equal Employment Opportunity Commission adopted the "Four-Fifths Rule", a guideline stating that, "A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded... as evidence of adverse impact."EOC (1978) This rubric has since been adopted for measures of fairness in ML. This translates to a "fair" range of selection rate ratios that are between 0.8 and 1.2.

The four-fifths rule works well when comparing prediction performance metrics whose values are above 0.5. However, the rule fails when comparing small values, as is the case in this example and which is as shown in the stratified report in the cell below. The ratios between two such small values can easily be well above 1.2, even though the true difference is only a few percentage points. For this reason it's useful to compare both the ratios and the differences when evaluating group measures.

Returning to the language example in the cell above: the Disparate Impact Ratio and Statistical Parity Difference are two related measures that compare the selection rates between the protected and unprotected groups. Although the Disparate Impact Ratio in our example is outside of the "fair" range for ratios (it's above 1.2), the Statistical Parity Difference is well within range for differences. We can see why more clearly by examining the Stratified Performance Report (also above). Here we see that the selection rates (shown as: "POSITIVE PREDICTION RATES") are actually quite close. The same is true for the Equalized Odds Ratio, which also appears outside of the "fair" range. The Equalized Odds Difference is

actually quite small, which we can understand more clearly by looking at the True Positive Rates and False Positive Rates (shown as TPR and FPR) in the Stratified Report.

Group Measure Type	Examples	"Fair" Range	Favored Group
Statistical Ratio	Disparate Impact Ratio, Equalized Odds Ratio	0.8 <= "Fair" <= 1.2	< 1 favors privileged group, > 1 favors unprivileged
Statistical Difference	Equalized Odds Difference, Predictive Parity Difference	-0.1 <= "Fair" <= 0.1	< 0 favors privileged group, > 0 favors unprivileged

Problems with Group Fairness Measures

Although these statistically-based measures make intuitive sense, they are not applicable in every situation. For example, Demographic Parity is inapplicable where the base rates significantly differ between groups. Also, by evaluating protected attributes in pre-defined groups, these measures may miss certain nuance. For example, a model may perform unfairly for certain sub-groups of the unprivileged class (e.g., black females), but not for the unprivileged group as a whole.

The Impossibility Theorem of Fairness

Another drawback of these statistically-based measures is that they are mathematically incompatible. No machine learning model can be perfectly fair according to all three metrics at once. People + AI Research (PAIR) posted an excellent visual explanation of the Impossibility Theorem.

Similarity-Based Measures and Individual Fairness

Measures of individual fairness determine if "similar" individuals (ignoring the protected attribute) are likely to have similar predictions.

Consistency Scores

Consistency scores measure the similarity between specific predictions and the predictions of like individuals. They are not specific to a particular attribute, but rather they evaluate the generally equal treatment of equal individuals. In AIF360, the consistency score is calculated as the compliment of the mean distance to the

score of the mean nearest neighbor, using Scikit's Nearest Neighbors algorithm (default: 5 neighbors determined by the Ball Tree algorithm). For this measure, values closer to 1 indicate greater consistency, and those closer to zero indicate less consistency. More information about consistency scores is available in [Zemel (2013)].

> $ consistency_score = 1 - \frac{1}{n\cdot\text{n_neighbors}}\sum_{i = 1}^n |\hat{y}*i* - |*sum*{j\in\mathcal{N}_{\text{n_neighbors}}(x_i)} \hat{y}_j| $

The Generalized Entropy Index and Related Measures

The *Generalized Entropy (GE) Index* was proposed as a metric for income inequality [Shorrocks (1980)]), although it originated as a measure of redundancy in information theory. In 2018, Speicher *et al.* proposed its use for ML models. These measures are dimensionless, and therefore are most useful in comparison relative to each other. Values closer to zero indicate greater fairness, and increasing values indicating decreased fairness.

> $ GE = \mathcal{E}(\alpha) = \begin{cases} \frac{1}{n \alpha (\alpha-1)}\sum_{i = 1}^n\left[\left(\frac{b_i}{\mu}\right)^\alpha - 1\right],& \alpha \ne 0, 1,\ \frac{1}{n}\sum_{i = 1}^n\frac{b_{i}}{\mu}\ln\frac{b_{i}}{\mu},& \alpha = 1,\ -\frac{1}{n}\sum_{i = 1}^n\ln\frac{b_{i}}{\mu},& \alpha = 0. \end{cases} $

Special Cases

The *Theil Index* occurs where the GE alpha is equal to one. Although it is dimensionless like other indices of generalized entropy, it can be transformed into an Atkinson index, which has a range between 0 and 1.

> $ Theil Index = GE(\alpha = 1) $

The *Coefficient of Variation* is two times the square root of the GE where alpha is equal to 2.

> $ Coefficient of Variation = 2*\sqrt{GE(\alpha = 2)} $

Generalized Entropy of Error

Generalized Entropy Error is the Generalized Entropy Index of the prediction error. Like the Consistency Score above, this measure is dimensionless; however, it does not provide specific information to allow discernment between groups.

> $ GE(Error, \alpha = 2) = GE(\hat{y}_i - y_i + 1) $

Between Group Generalized Entropy Error is the Generalized Entropy Index for the weighted means of group-specific errors. More information is available in Speicher (2013).

> $ GE(Error_{group}, \alpha = 2) = GE([N_{unprivileged}*mean(Error_{unprivileged}), N_{privileged}*mean(Error_{privileged})]) $

Comparing Similarity-Based Measures

Some measures of Individual Fairness are dimensionless, and for that reason they are most useful when comparing multiple models as we will see in Part 3. However, some measures such as the Consistency Score and Between-Group Generalized Entropy Error exist on scales from 0 to 1. The directions of these scales can differ between measures (i.e., perfect fairness may lie at either 0 or at 1 depending upon the measure), so you will want to make a note of which applies. For example, for the Consistency Score shown above, a score of 1 is considered perfectly "fair". Adapting the four-fifths rule, we can say that a model should be consistent for at least 80% of predictions. By this measure, our example model above is out of range.

Problems with Similarity-Based Fairness Measures

Similarity-based measures are not without their own drawbacks. The Consistency Score, for example, uses scikit's standard K-Nearest Neighbors (KNN) algorithm to define similarity, which may need additional (separate) parameter tuning, can be sensitive to irrelevant features, and may not be appropriate in cases of high dimensionality, sparse data or missingness. This then begs the question: *is the Consistency Score out of range because our prediction model is unfair, or because we haven't properly tuned the KNN algorithm?* Without significant additional work we cannot rule out the latter. Even supposing that a properly fit KNN model is possible, the results still may not be the most appropriate measure of similarity. For example, slthough diseases and procedures may be predictive, can it be said that all cardiac arrest survivors who recieved an Echocardiogram should be predicted to spend the same amount of time in the ICU?

See Also

Summary Tables: Convenient Charts of Fairness Measures

Choosing the Appropriate Measure(s)

Our choice of measure is informed both by the use cases for each particular measure, and also by the problem context and by the preferences of the community(ies) affected by the model. Unfortunately this means that Unfortunately no one "correct" way to measure fairness. This also means that there is no one "correct" way to demonstarate that fairness. The burden is on the

Data Scientist to transparently document their process and prove that they've taken reasonable steps to develop and to measure a model that is as fair as reasonably possible.

Although no model can be perfectly fair according to all metrics per the Impossibility Theorem (above), ideally a model will be at least within the range of fairness across the measures. From there, it's a matter of optimization for the specific measure(s) that is most applicable to the problem at hand. Thus the process begins with a clear understanding of the stakeholders and how they will view the potential outcomes. For healthcare models, the stakeholders are typically the patients, care providers, and the community(ies) being served, although it is likely that the care providers will represent the interests of the other two. It can also be helpful to create a table of outcomes, similar to the one below, to clearly document the harms, benefits, and preferences involved.

See Also: Value Sensitive Design

Example Table of Outcomes

Prediction	Outcomes	Preference
TP	Benefit: Deserving patient recieves help	high importance
TN	Benefit: Community resources saved	less important
FP	Harm: community resources wasted on an individual without need	less important (to avoid)
FN	Harm: reduced likelihood of recovery	high importance (to avoid)

P := "long length of stay expected (refer to counseling)"

Useful Questions to Ask when Choosing the Appropriate Measure(s)

1) What ethical frameworks are held by the stakeholders? How do they weigh the costs and benefits of different outcomes?

2) Which among all available measures are out of range? **2b)** Why are they out of range? Is is it due to the data, the model, the measure, or some combination?

3) Can the sources of unfairness be sufficiently addressed through changes to either the data or the model? **3b)** If the model remains unfair, is it still more fair than the current decisionmaking process?

How fair is fair enough?

While this specific solution may not always be available, there will likely always be options for potential improvement. Yet, we know from the Impossibility Theorem that we cannot produce a model that is perfectly fair by all measures. So how do we know when to stop?

The ulitmate metric for the fairness of our model is whether our results meet the expectations of the people who are affected by it. Can we justify our results to them. Will they stand up to the standards of the community, the healthcare practitioners, and most importantly, the patients?

Final Remarks

Just as data and model performance can drift over time, so too can prediction fairness. We recommend integrating fairness evaluation with your modeling pipeline as a form of continuous process improvement. By regularly evaluating multiple measures of fairness at once you can ensure that it continues to meet the expectaions of the stakeholders.

For more examples of fairness measurement using the FairMLHealth tool, see the Example-Template-BinaryClassificationAssessment Notebook in our tutorials_and_examples section. There are also a number of additional references at the bottom of this page, as well as in our Documentation Folder.

Fairness-Aware ML Algorithms

More than a dozen fairness-aware machine learning algorithms have been developed, although as shown above they may not be necessary to improve your model. However, if you are unable to mitigate the bias in your model by adjusting the data or changing the algorithm you use, you may want to consider one of the followin fairness-aware machine learning algorithms that are readily available through the libraries used in this notebook.

Fairness-Aware Algorithms

Algorithm	AIF360	Fairlearn	Reference
Optimized Preprocessing	Y	-	Calmon et al. (2017)
Disparate Impact Remover	Y	-	Feldman et al. (2015)

Algorithm	AIF360	Fairlearn	Reference
Equalized Odds Postprocessing (Threshold Optimizer)	Y	Y	Hardt et al. (2016)
Reweighing	Y	-	Kamiran and Calders (2012)
Reject Option Classification	Y	-	Kamiran et al. (2012)
Prejudice Remover Regularizer	Y	-	Kamishima et al. (2012)
Calibrated Equalized Odds Postprocessing	Y	-	Pleiss et al. (2017)
Learning Fair Representations	Y	-	Zemel (2013)
Adversarial Debiasing	Y	-	Zhang et al. (2018)
Meta-Algorithm for Fair Classification	Y	-	Celis et al. (2018)
Rich Subgroup Fairness	Y	-	Kearns, Neel, Roth, & Wu (2018)
Exponentiated Gradient	-	Y	Agarwal, Beygelzimer, Dudik, Langford, & Wallach (2018)
Grid Search	-	Y	Agarwal, Dudik, & Wu (2019); Agarwal, Beygelzimer, Dudik, Langford, & Wallach (2018)

Other Fairness Libraries of Note

- Aequitas
- AIF360
- Awesome Fairness in AI
- Dalex
- Fairlearn
- Fairness Comparison

- FAT Forensics
- ML Fairness Gym
- Themis ML