Fairness

Last updated by | Mateusz Macias | Jun 15, 2020 at 12:22 PM GMT+1

Fairness

Definition

Before we understand Fairness, we need to cover the term Bias; in a social context, **bias** refers to a preference towards or against a group of people/an individual which is unfair. From this definition we can infer Fairness refers to non-bias. What is Fair depends on the context.

Eg: As an example, we could imagine a situation where a bank is evaluating 100 mortgage applications. 70 are from men and 30 are from women. The protected attribute is gender. One measure of fairness is demographic parity. The percentage of men and women who get approved should be the same. If 35 men (50%) get their mortgage approved, there should be 15 for women (50%). However, in some sense this is unfair if there is a higher percentage of loan-worthy individuals in either group.

Principles of fairness

1. **non-discrimination and prevention of bias** principle articulates that bias in AI, i.e. in the training data, technical design choices or the technology deployment should be mitigated to prevent discriminatory impacts.

2. **Representative of high quality data** The use of a dataset that is not representative leads to skewed representation of a group in the dataset compared to the actual composition of the target population. This introduces bias and reduces the accuracy of the system's eventual decisions. e.g. Garbage in garbage out (as colloquially said)

> > The Montreal Declaration and the European Charter on AI in judicial systems call for representative and high quality data but state that even using the gold standard in data could be detrimental if the data are used for "deterministic analyses."

3. **Fairness and Equality** AI systems should act impartial towards the data subjects. Similarly equality refers to the AI systems producing similar outcomes for similar input.

> > "all people are treated fairly without unjustified discrimination on the grounds of diverse backgrounds such as race, gender, nationality, age, political beliefs, religion, and so on." - according to Japanese AI principles

> > "Equality of human beings goes beyond non- discrimination, which tolerates the drawing of distinctions between dissimilar situations based on objective justifications. In an AI context, equality entails that the same rules should apply for everyone to access to information, data, knowledge, markets and a fair distribution of the value added being generated by technologies." -- according to European High Level Expert Group guidelines

4. **Inclusiveness in impact** AI systems are expected to create an impact in achievement of fair, inclusive and peaceful society.
   The European High Level Expert Group guidelines add some detail around what "benefits" might be shared: "AI systems can contribute to wellbeing by seeking achievement of a fair, inclusive and peaceful society, by helping to increase citizen's mental autonomy, with equal distribution of economic, social and political opportunity."

5. **Inclusiveness in design** AI systems are expected to be designed by people from diverse backgrounds including women and people with disabilities to ensure non-discriminatory AI.

> > > Europe document from the European Commission affirms that “More women and people of diverse backgrounds, including people with disabilities, need to be involved in the development of AI, starting from inclusive AI education and training,
> > > in order to ensure that AI is non-discriminatory
> > > and inclusive.”

Sources of Bias

1. **Dataset**: Unrepresentative dataset/unsampled or oversampled dataset
2. **Model building**: What features are involved in building a model is also representative of bias. How the model is used for classification further will also contribute to bias eg., setting thresholds to achieve accuracy

Types of Bias

Below are a few types of bias, however there are more. These can be explored at https://arxiv.org/pdf/1908.09635.pdf

Historical Bias: Historical bias is the already existing bias and socio-technical issues in the world and can seep into from the data generation process even given a perfect sampling and feature selection.

Representation Bias: Representation bias happens from the way we define and sample from a population, e.g. Lacking geographical diversity in datasets.

Measurement Bias: Measurement bias happens from the way we choose, utilize, and measure a particular feature, e.g. prior arrests and friend/family arrests were used as proxy variables to measure level of “riskiness” or “crime”.

Evaluation Bias: Evaluation bias happens during model evaluation, e.g. using inappropriate/disproportionate benchmarks for evaluation.

Population/sampling Bias: Population bias arises when statistics, demographics, representatives, and user characteristics are different in the user population represented in the dataset/platform from the original target population.

Popularity Bias: Items that are more popular tend to be exposed more. However, popularity metrics are subject to manipulation, e.g. fake reviews by bots on social media.

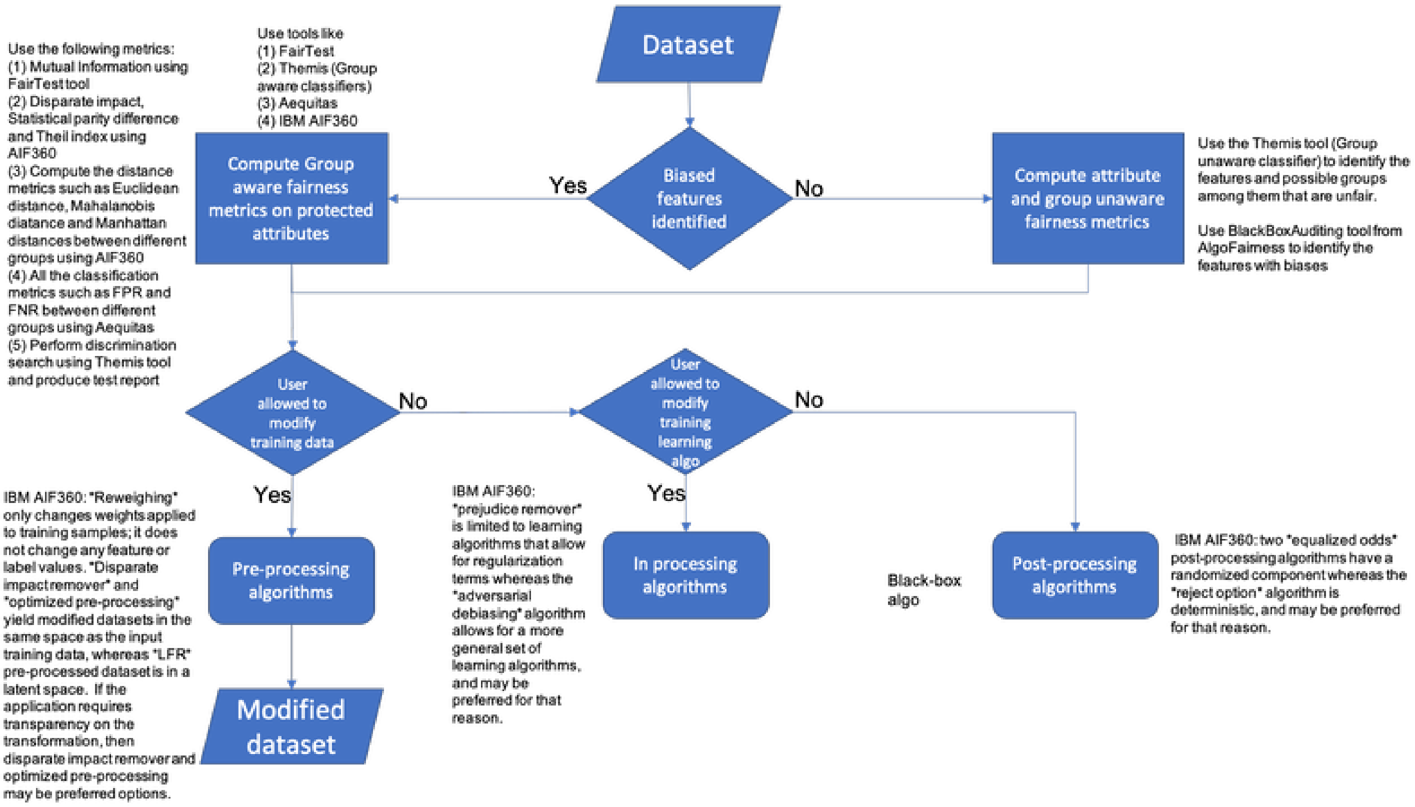
Funding Bias: Funding bias arises when biased results are reported in order to support or satisfy the funding agency or financial supporter of the research study.

Observer Bias: Observer bias happens when researchers subconsciously project their expectations onto the research.

Questions to be addressed by a Data Scientist

1. Is the dataset a balanced one? (i.e. do we have equal distribution of class samples?)
2. If the dataset is structured with a defined set of features, do we have discriminative features? (i.e. features that might easily help in assigning a class variable, but may discriminate humans. e.g. race, gender, religion and geography (based on project contexts))
3. Is the model biased towards a particular class?
4. Does the model use any form of setting thresholds? Does these thresholds bring in any form of bias?
5. What were the assumptions made while modelling? e.g. in NLP people consider words to be independent of each other (independence assumption)
6. Were the data samples distributed from various sub-groups? e.g. when we collect training data for Document Intelligence across all geographies, what is the proportion of data from each geography?

7. How can an analyst determine the relative significance of the inputs to a black-box predictive model, in order to assess the model's fairness (or the extent to which it is discriminatory)?
8. What does it mean for your predictive model to be fair?


Dataset
Biased features identified
Yes
No
Compute Group aware fairness metrics on protected attributes
Compute attribute and group unaware fairness metrics
User allowed to modify training data
No
User allowed to modify training learning algo
No
Yes
Yes
Pre-processing algorithms
In processing algorithms
Black-box algo
Post-processing algorithms
Modified dataset

Use the following metrics:
(1) Mutual Information using FairTest tool
(2) Disparate impact, Statistical parity difference and Theil index using AIF360
(3) Compute the distance metrics such as Euclidean distance, Mahalanobis diatance and Manhattan distances between different groups using AIF360
(4) All the classification metrics such as FPR and FNR between different groups using Aequitas
(5) Perform discrimination search using Themis tool and produce test report

Use tools like
(1) FairTest
(2) Themis (Group aware classifiers)
(3) Aequitas
(4) IBM AIF360

Use the Themis tool (Group unaware classifier) to identify the features and possible groups among them that are unfair.

Use BlackBoxAuditing tool from AlgoFairness to identify the features with biases

IBM AIF360: "Reweighing" only changes weights applied to training samples; it does not change any feature or label values. "Disparate impact remover" and "optimized pre-processing" yield modified datasets in the same space as the input training data, whereas "LFR" pre-processed dataset is in a latent space. If the application requires transparency on the transformation, then disparate impact remover and optimized pre-processing may be preferred options.

IBM AIF360: "prejudice remover" is limited to learning algorithms that allow for regularization terms whereas the "adversarial debiasing" algorithm allows for a more general set of learning algorithms, and may be preferred for that reason.

IBM AIF360: two "equalized odds" post-processing algorithms have a randomized component whereas the "reject option" algorithm is deterministic, and may be preferred for that reason.

Pre-processing algorithms - refer to the algorithms that are applied for the training data

OptimizedPreprocessing

This algorithm will transform the dataset to have more equity in positive outcomes on the protected attribute for the privileged and unprivileged groups. If the application requires transparency on the transformation, then disparate impact remover and optimized pre-processing may be preferred options. Please have a look at the notebook to understand how to use this tool.

Link to research work
Notebook

DisparateImpactRemover

Unintentional bias is encoded via disparate impact, which occurs when a selection process has widely different outcomes for different groups, even as it appears to be neutral. The algorithm corrects for imbalanced selection rates between unprivileged and privileged groups at various levels of repair. If the application requires transparency on the transformation, then disparate impact remover and optimized pre-processing may be preferred options.

Link to research work
Notebook

LRF (Learning fair representations)

Learning fair representations is a pre-processing technique that finds a latent representation which encodes the data well but obfuscates information about protected attributes. LFR's pre-processed dataset is in a latent space, therefore this is not recommended if the application requires transparency. This requires the protected attributes to be specified explicitly.

Link to research work ↗
Notebook ↗

Reweighing ↗

Reweighing is a preprocessing technique that Weights the examples in each (group, label) combination differently to ensure fairness before classification. Among pre-processing algorithms, reweighing only changes weights applied to training samples; it does not change any feature or label values. Therefore, it may be a preferred option in case the application does not allow for value changes. This algorithm requires the protected attributes to be specified explicitly.

Link to research work ↗
Notebook ↗

In-Processing algorithms - algorithms that are applied to models during its training.

This is an in-processing technique that learns a classifier to maximise prediction accuracy and simultaneously reduce an adversary's ability to determine the protected attribute from the predictions. This approach leads to a fair classifier as the predictions cannot carry any group discrimination information that the adversary can exploit. This could be better understood by examining the code base.

Tool ↗ (IBM AIF360)
Link to research work ↗
Notebook ↗

Prejudice Remover ↗

Prejudice involves a statistical dependence between sensitive features and other information. To remove the prejudice, they have come up with a prejudice remover regularizer function, which becomes large when a class is determined mainly based on sensitive features; thus, sensitive features become less influential in the final determination.

Among in-processing algorithms, the prejudice remover is limited to learning algorithms that allow for regularization terms whereas the adversarial debiasing algorithm allows for a more general set of learning algorithms, and may be preferred for that reason.

Link to research work ↗
In addition to the above mentioned algorithms, there are other InProcessing algorithms such as GerryFair classifier ↗, MetaFair classifier ↗ and ART classifier ↗. However they are significantly used less compared to the previous mentions.

Post-Processing algorithms - bias mitigation algorithms applied to predicted labels.

Reject Option classifier ↗

Reject option classification is a post-processing technique that gives favourable outcomes to unprivileged groups and unfavourable outcomes to privileged groups. This is done in a confidence band around the decision boundary with the highest uncertainty. This is used to fix the right thresholds -- this algorithm is recommended when there are thresholds applied at the classification stage.

Link to research work
Notebook | Implementation

Calibrated equalized odds postprocessing

Calibrated equalized odds is a post-processing technique that optimizes over calibrated classifier score outputs to find probabilities, with which to change output labels with an equalized odds objective (FNR, FPR).

Link to research work
Notebook

Tools for fairness tests/metrics

Tools that doesn't have comprehensive documentation are not listed here. Please have a look at the comprehensive of tools from here.

IBM AIF360

This is an open source fairness toolkit that has a comprehensive set of fairness metrics. They have broadly classified the metrics into

1. Dataset metric - Used for computing metrics based on one StructuredDataset.
2. Binary dataset metric - Used for computing metrics based on a single BinaryLabelDataset.
3. Classification metric - Used for computing metrics based on two BinaryLabelDatasets. The first dataset is the original one and the second is the output of the classification transformer (or similar).
4. Sample Distortion metric - Used for computing metrics based on two StructuredDatasets. Provides a number of distance functions as metrics.

The documentation for all these metrics are provided here. This is one of the most comprehensive toolkits that is recommended to be used. However the next listed tool Aequitas has nice visualization provisions.

Aequitas -- This is highly recommended for testing & plotting purposes

This is an open source bias audit toolkit developed by the Center for Data Science and Public Policy at University of Chicago. and This can be used to audit the predictions of machine learning based risk assessment tools to understand different types of biases, and make informed decisions about developing and deploying such systems.

This toolkit includes the metrics provided besides the questions to address them. Its highly recommended to have a look at the notebook in the link below. However brief possible explanations are provided here:

- What biases exist in my model?
    - What is the distribution of groups, predicted scores, and labels across my dataset? - Tool provides plotting functionality to plot the subgroups within features .
    - What are bias metrics across groups? - provides a function called **get_crosstabs** which computes the **True Positive Rate, True Negative Rate, False Omission Rate, False Discovery Rate, False Positive Rate, False Negative Rate, Negative Predictive Value, Precision, Predicted Positive Ratio, Group Prevalence** across different groups in the protected variables.
    - How do I interpret biases in my model? Please refer to this link to understand the bias metrics listed above.

    - How do I visualize biases in my model? Please refer to this link to know how to visualize the biases using this tool.

- What levels of disparity exist between population groups?

    - How does the selected reference group affect disparity calculations? refer to link
    - How do I interpret calculated disparity ratios? refer to link
    - How do I visualize disparities in my model? For all the bias metrics listed earlier, disparities an be computed and visualized. refer to link.

- How do I assess model fairness? The answers to the below 3 questions are answered in this page

    - How do I interpret parities?
    - How do I visualize bias metric parity?
    - How do I visualize parity between groups in my model?

Tool (Aequitas)
Documentation | Notebook

AlgoFairness

This tool includes a kit (BlackBoxAuditing) for blackbox testing which addresses the question: "How does a feature (or a group of features) indirectly affect the outcome?" Following is an intercept from one of the papers

> Consider trying to verify that racial considerations did not affect an automated decision to grant a housing loan. We could use a standard auditing procedure that declares that the race attribute does not have an undue influence over the results returned by the algorithm. Yet this may be insufficient. In the classic case of redlining [17], the decision-making process explicitly excluded race, but it used zipcode, which in a segregated environment is strongly linked to race. Here, race had an indirect influence on the outcome via the zipcode, which acts as a proxy.

This tool audits the black-box models for indirect influence on the outcomes. Furthermore, based on the audit outcomes it also provides a tool to repair the dataset as well. Please have a look at the materials provided here.

Tool (AlgoFairness)
Link to research work
Tutorial | Notebook

FairTest

This tool has an algorithm (Association guided tree construction), inspired by decision-tree classifiers. This algorithm recursively splits the user space into smaller subsets so as to maximize some metric of association between algorithm outputs and protected user attributes. Each step yields subpopulations of decreasing size and increasingly strong disparate effects. This efficiently searches for user subpopulations exhibiting strong unfair effects.

This tool (FairTest) also tests for the following metrics in one go, and produces a test report with scores for each metric on the sub-populations:

1. Normalized Mutual Information - For categorical protected feature and output
2. Normalized Conditional Mutual Information - For categorical protected feature and output
3. Binary Difference and Binary Ratio - For binary protected feature and output

4. Conditional Binary Difference - For binary protected feature and output
5. Pearson Correlation - For ordinal protected feature and output (Ordinals are categorical in nature but can be ordered. e.g. classes 1,2,3,4)

Tool (FairTest)
Link to research work
Notebook

Themis

This tool measures two kinds of discrimination

1. Group discrimination - done by measuring what % of each group has resulted in a class being True.
2. Causal discrimination - done by performing multiple experiments. For example, it is possible to execute the software on two individuals identical in every way except race, and verify if changing the race causes a change in the output.

They evaluate with four discrimination aware and four discrimination unaware classifiers. The evaluations are provided in the Examples link provided below.

Tool
Link to research work
Examples

Further Readings

Fairness in Machine Learning: Limitations and Opportunities https://fairmlbook.org/pdf/fairmlbook.pdf
IBM AI Fairness 360: https://arxiv.org/pdf/1810.01943.pdf