**Comparing Optimized Pre-Processing for Discrimination Prevention Across Different Models**

**AVENIR TOUWE**

**UNIVERSITY OF MARYLAND GLOBAL CAMPUS**

**DATA 450**

**PROF**

**1-Explain in your own words on up to 1 page what the following fairness metrics are: Statistical parity difference, Disparate impact, Average odds difference, Equal opportunity difference, and Theil index. Hint: run** [**https://aif360.mybluemix.net/data**](https://aif360.mybluemix.net/data) **with one dataset and look at the results (and click on “i” )**

- Statistical parity difference: Also known as disparate impact or demographic parity, measures the difference in the probability of favorable outcomes between different groups. This metric is computed as the difference of the rate of favorable outcomes received by the unprivileged group to the privileged group.

- Disparate impact: This metric is computed as the ratio of rate of favorable outcomes for the unprivileged group to that of the privileged group. The ideal value of this metric is 1.0. Is a fairness metric that quantifies the ratio of favorable outcomes between different groups. It is calculated as the ratio of the percentage of favorable outcomes for the unprivileged group to the percentage of favorable outcomes for the privileged group.

- Average odds difference: This metric is computed as average difference of false positive rate (false positives / negatives) and true positive rate (true positives / positives) between unprivileged and privileged groups. The ideal value of this metric is 0.

- Equal Opportunity Difference: It measures the difference in the true positive rate between different groups. The true positive rate is the ratio of true positives to the total number of actual positives for a given group. The ideal value is 0.

- Theil index: This metric is computed as the generalized entropy of benefit for all individuals in the dataset, with alpha = 1. It measures the inequality in benefit allocation for individuals. The Theil Index is a fairness metric that measures the overall inequality in predicted outcomes across different groups.

**2-In your own words, in up to 1 page, explain the debiasing algorithm implemented in the OptimPreproc function.**

Optimized preprocessing is a preprocessing technique that learns a probabilistic transformation that edits the features and labels in the data with group fairness, individual distortion, and data fidelity constraints and objectives.

The debiasing algorithm implemented in OptimPreproc follows a two-step process:

Preprocessing: In the first step, the algorithm preprocesses the data to identify and quantify bias. It uses a set of user-defined fairness metrics, such as statistical parity difference, disparate impact, average odds difference, equal opportunity difference, or the Theil index, to measure the extent of bias in the data. These fairness metrics are used to assess the disparities in prediction outcomes between different groups, such as privileged and unprivileged groups, based on sensitive attributes like race, gender, or age.

Optimization: In the second step, the algorithm applies optimization techniques to transform the data and achieve fairness in prediction outcomes. It formulates the problem as a constrained optimization problem, where the goal is to minimize the bias in the data while preserving the model's accuracy. The algorithm introduces fairness constraints based on the user-defined fairness metrics and optimizes the data transformation to achieve the best trade-off between fairness and accuracy.

**3- In your own words, in up to 1 page, explain the debiasing algorithm implemented in the ReWeighing function.**

The ReWeighing function is another debiasing algorithm implemented in the AIF360 library. It is a simple and intuitive method that involves reweighting the data to achieve fairness in machine learning models. The algorithm implemented in the ReWeighing function follows these steps:

Data Preprocessing: The algorithm starts by preprocessing the data to identify and quantify bias. It uses a set of user-defined sensitive attributes, such as race, gender, or age, to divide the data into different groups, typically privileged and unprivileged groups. The algorithm then computes the group-wise statistics, such as the group size and the average value of the sensitive attributes, for each group.

Bias Quantification: The algorithm calculates the bias in the data by comparing the distribution of the sensitive attributes across different groups. It computes the disparity in the group-wise statistics, such as the mean or median of the sensitive attributes, between the privileged and unprivileged groups. This disparity serves as a measure of bias in the data and indicates the extent to which the sensitive attributes are unevenly distributed among different groups.

Reweighting: The algorithm applies reweighting to the data to achieve fairness in prediction outcomes. It assigns different weights to the data samples in the privileged and unprivileged groups based on the disparity in the sensitive attributes. The samples from the group with higher disparity are downweighted, while the samples from the group with lower disparity are upweighted. This reweighting process aims to balance the distribution of the sensitive attributes across different groups and reduce bias in the data.

Model Training: After reweighting the data, the algorithm trains a machine learning model using the transformed data. The reweighted data is used to update the model's parameters, ensuring that the model is trained on a more balanced and fair dataset. The trained model can then be used for making predictions on new data, with the expectation that it will exhibit reduced bias and improved fairness in prediction outcomes.

**4-Compare and contrast the OptimPreproc algorithm to the ReWeighing algorithm. In both cases, do you think that applying these debiasing algorithms can lower the overall accuracy of the classification models being built?**

The OptimPreproc and ReWeighing algorithms are both debiasing algorithms implemented in the AIF360 library, but they differ in their approach to mitigating bias in data and achieving fairness in machine learning models.

1. Approach: The OptimPreproc algorithm follows a two-step process involving preprocessing and optimization. It uses user-defined fairness metrics to quantify bias in the data and then formulates an optimization problem to transform the data and achieve fairness in prediction outcomes. On the other hand, the ReWeighing algorithm is a simpler approach that involves reweighting the data based on the disparity in sensitive attributes to achieve a more balanced distribution of these attributes across different groups.
2. Complexity: The OptimPreproc algorithm is more complex compared to the ReWeighing algorithm as it involves an optimization step that requires solving a constrained optimization problem. It may require additional computational resources and expertise to fine-tune the optimization parameters. On the other hand, the ReWeighing algorithm is a straightforward approach that involves reweighting the data based on the disparity in sensitive attributes, making it relatively simpler to implement and interpret.
3. Flexibility: The OptimPreproc algorithm provides more flexibility in choosing different fairness metrics and allows users to customize the algorithm based on their specific requirements. It can accommodate various types of fairness constraints and optimize the data transformation accordingly. The ReWeighing algorithm, on the other hand, relies solely on the disparity in sensitive attributes for reweighting and does not offer customization options for fairness metrics or constraints.

Regarding the impact on overall accuracy of classification models, applying debiasing algorithms can potentially lower the accuracy of the models in some cases. Both the OptimPreproc and ReWeighing algorithms aim to achieve fairness by addressing bias in the data. However, in the process of mitigating bias, there may be a trade-off between fairness and accuracy. For example, reweighting may result in loss of information or overcorrection, which can affect model accuracy. Similarly, the data transformation and optimization steps in the OptimPreproc algorithm may introduce noise or distortion in the data, leading to lower accuracy. It's important to carefully evaluate the impact of these debiasing algorithms on model accuracy and make informed decisions based on the specific context and requirements of the application.

References:

<https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb>

<https://aif360.readthedocs.io/en/latest/modules/generated/aif360.algorithms.preprocessing.OptimPreproc.html>

<https://towardsdatascience.com/algorithmic-solutions-to-algorithmic-bias-aef59eaf6565>

<https://chat.openai.com/>