Fairness Metrics User Manual

1 Notations

- Let Y be the outcome variable, e.g. $Y \in \{0, 1\}$ in a binary classification problem, and $Y \in \mathbb{R}$ in a regression problem.
- Let A be a sensitive attribute, which may be categorical or continuous. For binary sensitive attribute, $A \in \{0, 1\}$, and $A = 0$ denotes the privileged group, and $A = 1$ denotes the non-privileged group.
- Let $\hat{Y}$ be the predicted outcome variable by a machine learning model.
- Let X be a vector of non-sensitive attributes.
- Let $S \in [0, 1]$ be the prediction score in binary classification problems.
- Let Y_i denote the ith instance of the outcome variable in the data set. $\hat{Y}_i, A_i, X_i, S_i$ are defined analogously.
- Let TP, FP, TN, FN denote true positive ($Y = 1, \hat{Y} = 1$), false positive ($Y = 0, \hat{Y} = 1$), true negative ($Y = 0, \hat{Y} = 0$), and false negative ($Y = 1, \hat{Y} = 0$), respectively.

2 List of Fairness Metrics

2.1 Statistical Parity

General Definition
Statistical parity requires the prediction $\hat{Y}$ is independent of the sensitive attribute A, $\hat{Y} \perp\!\!\!\perp A$.

Statistical parity for binary Classification
For binary classification with a binary sensitive attribute, the definition of statistical parity reduces to

$$\mathbb{P}(Y = y|A = 0) = \mathbb{P}(Y = y|A = 1), \quad y = 0, 1.$$

Computation of statistical parity ratio for binary classification The statistical parity ratio is computed as

$$\frac{(TP_{A=1} + FP_{A=1})/(TP_{A=1} + FP_{A=1} + TN_{A=1} + FN_{A=1})}{(TP_{A=0} + FP_{A=0})/(TP_{A=0} + FP_{A=0} + TN_{A=0} + FN_{A=0})},$$

where $TP_{A=1}$ is the true positive restricted to $A = 1$, and similarly for all the other quantities.

2.2 Equalized Odds

General Definition of Equalized Odds
Equalized odds requires that conditional on the outcome Y, the predicted outcome $\hat{Y}$ is independent of the sensitive attribute A, $\hat{Y} \perp\!\!\!\perp A|Y$.

Equalized odds for binary classification
For binary classification with a binary sensitive attribute, the definition of equalized odds reduces to

$$\mathbb{P}(\hat{Y} = 1|Y = y, A = 0) = \mathbb{P}(\hat{Y} = 1|Y = y, A = 1), \quad \forall y \in \{0, 1\}.$$

Equalized odds is rarely satisfied in practice.

2.2.1 Equal Opportunity

Equal opportunity can be considered as a relaxation of equalized odds.
Equal opportunity for binary classification
Equal opportunity requires

$$\mathbb{P}(\hat{Y} = 1|Y = 1, A = 0) = P(\hat{Y} = 1|Y = 1, A = 1).$$

Computation of equal opportunity ratio for binary classification
The equal opportunity ratio is computed as

$$\frac{TP_{A=1}/(TP_{A=1} + FN_{A=1})}{TP_{A=0}/(TP_{A=0} + FN_{A=0})}.$$

2.2.2 Predictive Equality

Predictive equality can be considered as another type of relaxation of equalized odds.
Predictive equality for binary classification
Predictive equality requires that

$$\mathbb{P}(\hat{Y} = 1|Y = 0, A = 0) = \mathbb{P}(\hat{Y} = 1|Y = 0, A = 1).$$

Computation of predictive equality ratio for binary classification
Predictive equality ratio is computed as:

$$\frac{FP_{A=1}/(FP_{A=1} + TN_{A=1})}{FP_{A=0}/(FP_{A=0} + TN_{A=0})}.$$

2.3 Overall Accuracy Equality

General definition of overall accuracy equality
Not applicable.

Overall accuracy equality for binary classification
Overall accuracy equality requires that

$$\mathbb{P}(\hat{Y} = Y|A = 0) = \mathbb{P}(\hat{Y} = Y|A = 1).$$

Computation of overall accuracy equality ratio for binary classification
The overall accuracy equality ratio is computed as

$$\frac{(TP_{A=1} + TN_{A=1})/(TP_{A=1} + FP_{A=1} + TN_{A=1} + FN_{A=1})}{(TP_{A=0} + TN_{A=0})/(TP_{A=0} + FP_{A=0} + TN_{A=0} + FN_{A=0})}.$$

2.4 Conditional Use Overall Accuracy Equality

General definition of conditional use overall accuracy
Not applicable.

Conditional use overall accuracy equality for binary classification
Conditional use overall accuracy equality requires that

$$\mathbb{P}(Y = y|\hat{Y} = y, A = 0) = \mathbb{P}(Y = y|\hat{Y} = y, A = 1), \quad \forall y \in \{0, 1\}.$$

2.4.1 Predictive Parity

Predictive parity is a relaxation of conditional use overall accuracy equality.

Predictive parity for binary classification

$$\mathbb{P}(Y = 1|\hat{Y} = 1, A = 0) = \mathbb{P}(Y = 1|\hat{Y} = 1, A = 1).$$

Computation of predictive parity ratio for binary classification

Predictive parity ratio is computed as:

$$\frac{TP_{A=1}/(TP_{A=1} + FP_{A=1})}{TP_{A=0}/(TP_{A=0} + FP_{A=0})}.$$

2.4.2 Negative Predictive Parity

Negative predictive parity is another relaxation of conditional use overall accuracy equality.

Negative predictive parity for binary classification

$$\mathbb{P}(Y = 1|\hat{Y} = 0, A = 0) = \mathbb{P}(Y = 1|\hat{Y} = 0, A = 1).$$

Computation of negative predictive parity ratio for binary classification
Negative predictive parity ratio is computed as:

$$\frac{TN_{A=1}/(FN_{A=1} + TN_{A=1})}{TN_{A=0}/(FN_{A=0} + TN_{A=0})}.$$

2.5 Balance

General definition of overall balance
Not applicable.

Overall balance for binary classification Overall balance is satisfied if

$$\mathbb{E}(S|Y = y, A = 0) = \mathbb{E}(S|Y = y, A = 1), \quad \forall y \in \{0, 1\}.$$

2.5.1 Balance for positive class

Balance for positive class is a relaxation of overall balance.

Balance for positive class for binary classification
Balance for positive class requires that

$$\mathbb{E}(S|Y = 1, A = 0) = \mathbb{E}(S|Y = 1, A = 1).$$

Computation of balance for positive class ratio for binary classification
Balance for positive class ratio is computed as

$$\frac{\frac{1}{n_1}\sum_{i:A_i=1,Y_i=1} S_i}{\frac{1}{n_2}\sum_{i:A_i=0,Y_i=1} S_i},$$

where $n_1 = |\{i : A_i = 0, Y_i = 1\}|$, $n_2 = |\{i : A_i = 1, Y_i = 1\}|$.

2.5.2 Balance for negative class

Balance for negative class is another relaxation of overall balance.

Balance for negative class for binary classification
Balance for negative class requires that

$$\mathbb{E}(S|Y = 0, A = 0) = \mathbb{E}(S|Y = 0, A = 1).$$

Computation of balance for negative class ratio for binary classification
Balance for negative class ratio is computed as

$$\frac{\frac{1}{n_1}\sum_{i:A_i=1,Y_i=0} S_i}{\frac{1}{n_2}\sum_{i:A_i=0,Y_i=0} S_i},$$

where $n_1 = |\{i : A_i = 0, Y_i = 0\}|$, $n_2 = |\{i : A_i = 1, Y_i = 0\}|$.

3 List of Mitigation Methods

3.1 Reweighing method

Type of method
The reweighing method is a *pre-processing* method for bias mitigation.

Description of method
The reweighing which involves assigning different weights to objects. For example, one may assign higher weights to objects with $A_i = 0, Y_i = 1$ than objects with $A_i = 0, Y_i = 0$, and higher weights to objects with $A_i = 1, Y_i = 0$ than $A_i = 1, Y_i = 1$.

The weights are determined by the expected and observed probabilities of each of the four classes and lower weights will be assigned to objects that have been deprived or favored.

Fairness metric
The reweighing method targets *statistical parity*.

3.2 Re-sampling method

Type of method
The re-sampling method is a *pre-processing* method for bias mitigation.

Description of method Similar as in Reweighing, the re-sampling method computes for each of the groups $A_i = 0, Y_i = 1$, $A_i = 0, Y_i = 0$, $A_i = 1, Y_i = 1$, $A_i = 1, Y_i = 0$ their expected sizes if the given dataset would have been non-discriminatory.
The uniform sampling approach involves sampling each group separately, until its expected group size is reached. The preferential sampling approach uses the idea that data objects close to the decision boundary are more likely to be discriminated or favored due to discrimination in the dataset.

Fairness metric
The reweighing method targets *statistical parity*.