Fairness Metrics

Most statistical measures of bias rely on the following metrics, which are best explained using a confusion matrix for classification task. Rows and columns of the matrix represent instances of the predicted and actual classes respectively. Cells of the confusion matrix explain the following definitions:

- True positive (TP)

A case when the predicted and actual outcomes are both in the positive class.

- False positive (FP)

A case predicted to be in the positive class when the actual outcome belongs to the negative class.

- False negative (FN)

A case predicted to be in the negative class when the actual outcome belongs to the positive class.

- True negative (TN)

A case when the predicted and actual outcomes are both in the negative class.

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** PPV = $\frac{TP}{TP+FP}$ TPR = $\frac{TP}{TP+FN}$	**False Positive (FP)** FDR = $\frac{FP}{TP+FP}$ FPR = $\frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** FOR = $\frac{FN}{TN+FN}$ FNR = $\frac{FN}{TP+FN}$	**True Negative (TN)** NPV = $\frac{TN}{TN+FN}$ TNR = $\frac{TN}{TN+FP}$

Table 2: Confusion matrix.

The confusion matrix divided by predicted and actual values builds the fundamental definition of some commonly used fairness measurement metrics. Suppose a model awaited for evaluation is **M**, and a dataset with sample denoted by **X**. The actual value is **Y** and the predicted value by model M is **Y'**. The sensitive feature of data sample is denoted by **A**, which is the applied for testing the impact of sensitive attributes over model using group-level metrics. We now explain 6 metrics and summarized them in Table 3.

- 1. Demographic Parity

Demographic parity is also known as statistical parity. A predictor Y' satisfies demographic parity if P(Y' | A = 0) = P(Y'| A = 1). The likelihood of a positive outcome should be the same regardless of whether the person is in the protected (e.g., female) group. In other words, demographic parity is achieved when the probability of a certain prediction is not dependent on sensitive group membership. In the binary classification scenario, demographic parity refers to equal selection rates across groups.

- 2. Equalized Odds

A predictor Y' satisfies equalized odds with respect to protected attribute A and outcome Y, if Y' and A are independent conditional on Y, i.e., P(Y'=1| A=0, Y=y) = P(Y'=1 | A=1,Y=y), y∈{0,1}. This means that the probability of a person in the positive class being correctly assigned a positive outcome and the probability of a person in a negative class being incorrectly assigned a positive outcome should both be the same for the protected and unprotected group members. In other words, the equalized odds definition states that the protected and unprotected groups should have equal rates for true positives and false positives.

- 3. Equalized Opportunity

A binary predictor Y' satisfies equal opportunity with respect to A and Y if P(Y'=1|A=0,Y=1) = P(Y'=1|A=1,Y=1). This means that the probability of a person in a positive class being assigned to a positive outcome should be equal for both protected and unprotected (e.g., female and male) group members. In other words, the equal opportunity definition states that the protected and unprotected groups should have equal true positive rates.

- 4. Conditional Statistical Parity

The **legitimate factors L** refers to a group of attributes that may affect the prediction outcome other than the sensitive attribute. For a set of legitimate factors L, predictor Y′ satisfies conditional statistical parity if P(Y′=1 | L=1, A=0) = P(Y′=1 | L=1, A=1). Conditional statistical parity states that people in both protected and unprotected (e.g., female and male) groups should have equal probability of being assigned to a positive outcome given a set of legitimate factors L.

- 5. Fairness through Unawareness

An algorithm is fair as long as any protected attributes A are not explicitly used in the decision-making process. In other word, the prediction outcome should be the same for applicants i and j who have the same attributes x: $x_i = x_j \rightarrow Y'_i=Y'_j$.

- 6. Fairness through Awareness

An algorithm is fair if it gives similar predictions to similar individuals. In other words, any two individuals who are similar with respect to a similarity (inverse distance) metric defined for a particular task should receive a similar outcome. The similarity of individuals is defined via distance metrics; for fairness to hold, the distance between the distributions of outputs for individuals should be at most the distance between the individuals. Formally, for a set of applicants V, a distance metric between applicants $k : V \times V \rightarrow R$, a model from a set of applicants to probability distributions over outcomes $M : V \rightarrow \delta A$, and a distance metric D between distribution of outputs, fairness is achieved iff $D(Y'_i, Y'_j) \leqslant k(x_i, x_j)$.

1. **Demographic Parity**		
$Pr(Y'=1 \mid A=0) = Pr(Y'=1 \mid A=1)$	Predicted Value	Group Level
2. **Equalized Odds**		
$Pr(Y'=1 \mid A=0, Y=y) = Pr(Y'=1 \mid A=1, Y=y)$	Predicted Value + Actual Value	Group Level
3. **Equalized Opportunity**		
$Pr(Y'=1 \mid A=0, Y=1) = Pr(Y'=1 \mid A=1, Y=1)$	Predicted Value + Actual Value	Group Level
4. **Conditional Statistical Parity**		
$Pr(Y'=1 \mid L=1, Y=y) = Pr(Y' =1 \mid L=1, Y=y)$	Predicted Value + Actual Value	Group Level
5. **Fairness Through Unawareness**		
If $x_i = x_j$, then $Y'_i=Y'_j$	Predicted Value	Individual Level
6. **Fairness through Awareness**		
$D(Y'_i, Y'_j) \leqslant k(x_i, x_j)$	Predicted Value	Individual Level

Table 3: Commonly used fairness metrics in research.

Additional Material (in <*Additional Material*> folder)

1. [Paper] ACM Computing Survey + 2019 + A Survey on Bias and Fairness in Machine Learning
https://dl.acm.org/doi/abs/10.1145/3457607

2. [Paper] ACM FairWare + 2018 + Fairness definitions explained
https://dl.acm.org/doi/10.1145/3194770.3194776

3. [Blog] Fairness in Machine Learning -> I recommend reading :)
https://dida.do/blog/fairness-in-ml

4. [Blog] An Introduction to Fairness in Machine Learning
https://medium.com/analytics-vidhya/an-introduction-to-fairness-in-machine-learning-62ef827e0020