Mutual Exclusion of Fairness Criterions

Notation

- Y: True class, where 1 is the preferred, favorable outcome.
- S: Predicted score.
- $\hat{Y}$: Decision, defined as
$$\hat{Y} := \mathbf{1}_{S>c}$$
meaning the decision is 1 if the score S exceeds a threshold c.
- A: Protected attribute (e.g., demographic group).

Definitions

1. Group Fairness (Independence)
$$P(\hat{Y}=1 \mid A=a) = P(\hat{Y}=1 \mid A=b) \iff \hat{Y} \perp A$$
2. Equalized Odds (Separation)
$$\begin{aligned} P(\hat{Y}=1 \mid A=a, Y=1) = P(\hat{Y}=1 \mid A=b, Y=1) \\ \wedge \\ P(\hat{Y}=1 \mid A=a, Y=0) = P(\hat{Y}=1 \mid A=b, Y=0) \\ \iff \hat{Y} \perp A \mid Y \end{aligned}$$
3. Positive Rate Parity (Sufficiency)
$$\begin{aligned} P(Y=1 \mid A=a, \hat{Y}=1) = P(Y=1 \mid A=b, \hat{Y}=1) \\ \wedge \\ P(Y=1 \mid A=a, \hat{Y}=0) = P(Y=1 \mid A=b, \hat{Y}=0) \\ \iff Y \perp A \mid \hat{Y} \end{aligned}$$

Premise

No two of the three fairness equalities (1) (2) (3) may be fulfilled simultaneously in nontrivial cases i.e. when $\neg(A \perp Y \vee \hat{Y} \perp Y)$.

Proof. (A) Assume (1) and (2) are satisfied.

$$P(\hat{Y}=1 \mid A=a) = \sum_{y\in\{0,1\}} P(\hat{Y}=1, Y=y \mid A=a) =$$
$$\sum_{y\in\{0,1\}} P(\hat{Y}=1, A=a \mid Y=y) \cdot \frac{P(Y=y)}{P(A=a)}$$

Using (1) and (2) we might obtain

$$P(\hat{Y}=1)=\sum_{y\in\{0,1\}} P(\hat{Y}=1\mid Y=y)P(Y=y\mid A=a).$$

Conversely, also from the law of the total probability, we obtain

$$P(\hat{Y}=1)=\sum_{y\in\{0,1\}} P(\hat{Y}=1\mid Y=y)P(Y=y).$$

Consequently:

$$\begin{aligned}
&P(\hat{Y}=1\mid Y=0)(P(Y=0\mid A=a)-P(Y=0))+\\
&P(\hat{Y}=1\mid Y=1)(P(Y=1\mid A=a)-P(Y=1))=0 \iff\\
&P(\hat{Y}=1\mid Y=0)(P(Y=0\mid A=a)-P(Y=0))+\\
&P(\hat{Y}=1\mid Y=1)(1-P(Y=0\mid A=a)-(1-P(Y=0)))=0 \iff\\
&P(\hat{Y}=1\mid Y=0)(P(Y=0\mid A=a)-P(Y=0))-\\
&P(Y=0)(P(\hat{Y}=1\mid Y=0)-P(\hat{Y}=1\mid Y=1))=0 \iff\\
&(P(\hat{Y}=1\mid Y=0)-P(\hat{Y}=1\mid Y=1))(P(Y=0)-P(Y=0\mid A=a))=0 \iff\\
&\hat{Y}\perp Y \vee A\perp Y
\end{aligned}$$

(B) Assume (1) and (3) are satisfied. From (1) and first eq of (3) we get

$$\begin{aligned}
\frac{P(\hat{Y}=1\mid A=a)}{P(\hat{Y}=1\mid A=b)}&=\frac{P(Y=1\mid A=b,\hat{Y}=1)}{P(Y=1\mid A=a,\hat{Y}=1)} \iff\\
\frac{P(\hat{Y}=1,A=a)P(A=b)}{P(\hat{Y}=1,A=b)P(A=a)}&=\frac{P(Y=1,A=b,\hat{Y}=1)P(A=a,\hat{Y}=1)}{P(Y=1,A=a,\hat{Y}=1)P(A=b,\hat{Y}=1)} \iff\\
P(Y=1,\hat{Y}=1\mid A=a)&=P(Y=1,\hat{Y}=1\mid A=b)
\end{aligned}$$

Analogous computations in case $\hat{Y}=0$ lead to $P(Y=1,\hat{Y}=0\mid A=a)=P(Y=1,\hat{Y}=0\mid A=b)$. Summing up corresponding sides of the two equalities we marginalize over $\hat{Y}$ and obtain:

$$\begin{aligned}
P(Y=1\mid A=a)=P(Y=1\mid A=b) &\iff\\
Y\perp A.
\end{aligned}$$

(C) Assume (2) and (3) are satisfied. From first eq of (2) and first eq of (3) we get:

$$\begin{aligned}
\frac{P(Y=1\mid A=a,\hat{Y}=1)}{P(Y=1\mid A=b,\hat{Y}=1)}&=\frac{P(\hat{Y}=1\mid A=a,Y=1)}{P(\hat{Y}=1\mid A=b,Y=1)} \iff\\
\frac{P(A=b,\hat{Y}=1)}{P(A=a,\hat{Y}=1)}&=\frac{P(Y=1,A=b)}{P(Y=1,A=a)} \iff\\
P(A=b,\hat{Y}=1)\cdot P(Y=1,A=a)&=P(Y=1,A=b)\cdot P(A=a,\hat{Y}=1)
\end{aligned}$$

Analogously for $\hat{Y} = 0$ one might obtain $P(Y = 1, A = a) \cdot P(A = b, \hat{Y} = 0) = P(Y = 1, A = b) \cdot P(A = a, \hat{Y} = 0)$. Now summing up coressponding sides of the equations one might marginalize over $\hat{Y}$ and get:

$$\begin{aligned} P(Y = 1, A = a) \cdot P(A = b) = P(Y = 1, A = b) \cdot P(A = a) \iff \\ P(Y = 1 \mid A = a) = P(Y = 1 \mid A = b) \iff \\ Y \perp A. \end{aligned}$$

$\square$