Pragmatic Normalized Group-Fairness and Fairness-vs-Performance Trade-Off in Binary Classifiers

Riccardo Rovatti

riccardo.rovatti@unibo.it

23/01/2026

Classical approach

- Group Fairness in binary classifiers
- Let's ask an «expert»

Name		f
statistical parity	F_1	$\frac{TP+FP}{TP+TN+FP+FN}$
conditional accuracy	F_2	$\frac{TP}{TP+FN}$
false positive/negative rate	F_3	$\frac{FP}{TN+FP}$
calibration	F_4	$\frac{TP}{TP+FP}$
false discovery/omission rate	F_5	$\frac{FP}{TP+FP}$
treatment	F_6	$\frac{FP}{FN}$

An example

- ISIC challenge 2020
- Train+Val: 29985 skin lesions – malignant(P 543)/benign(N 29442)
- Test: 9912 skin lesions – malignant (P 255)/benign (N 9657)
- Resnet-50 trained on Image net + 2 fully connected layers to generate the logits
- Unbalanced binary cross-entropy
- ADAM + Reduce learning rate on plateau + batch 32
- 15 epochs – best val loss

Classical analysis

Let’s make some calculation

$\mathcal{Q}$ = female patients

		predicted P	predicted N
true	P	69	22
	N	1076	3224

$\mathcal{R}$ = male patients

		predicted P	predicted N
true	P	126	38
	N	1487	3870

Fairness normalization

- Hp: focus functions are scale invariant → TP, TN, FP, FN are rates instead of counts

$$\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN} = 1$$

$$0 \leq \mathrm{TP}, \mathrm{TN}, \mathrm{FP}, \mathrm{FN} \leq 1$$

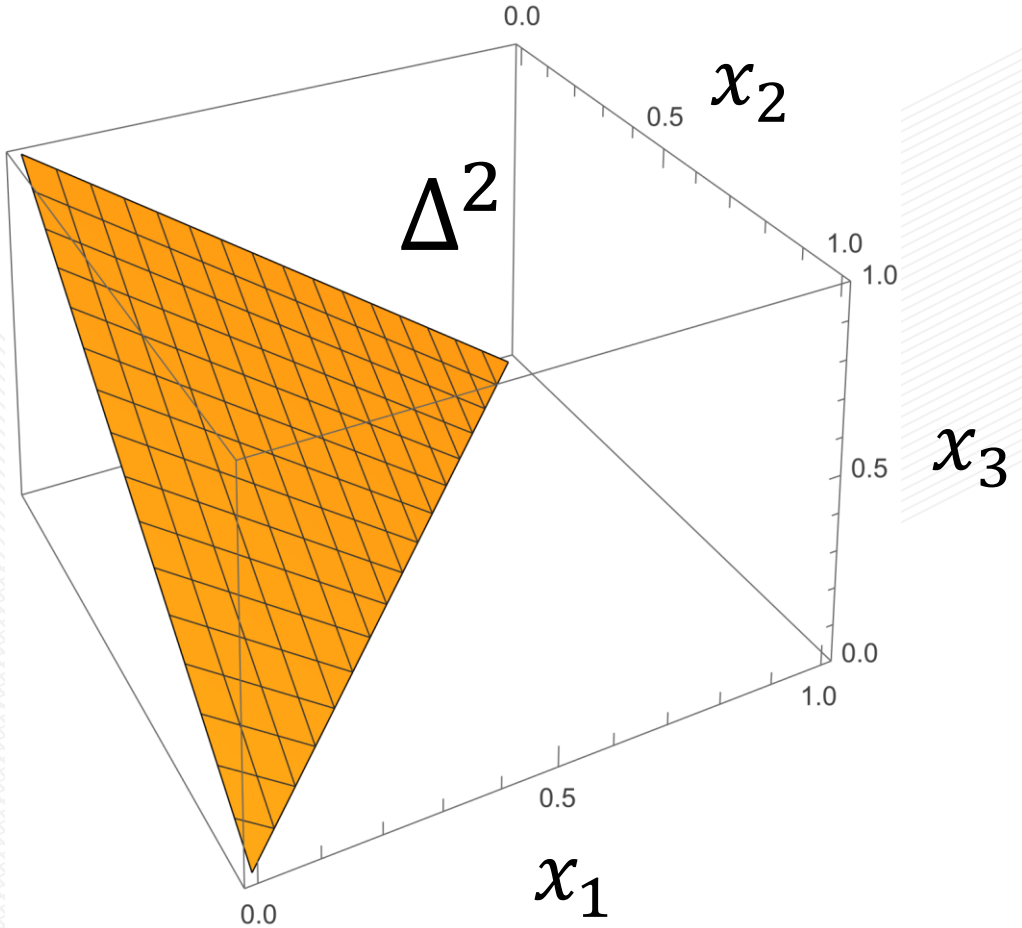
Name		f
statistical parity	F_1	$\frac{\mathrm{TP}+\mathrm{FP}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$
conditional accuracy	F_2	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$
false positive/negative rate	F_3	$\frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}}$
calibration	F_4	$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$
false discovery/omission rate	F_5	$\frac{\mathrm{FP}}{\mathrm{TP}+\mathrm{FP}}$
treatment	F_6	$\frac{\mathrm{FP}}{\mathrm{FN}}$

Fairness normalization

- Hp: focus functions are scale invariant → TP, TN, FP, FN are rates instead of counts

$$\Delta^3 \quad \mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN} = 1$$

$$0 \leq \mathrm{TP}, \mathrm{TN}, \mathrm{FP}, \mathrm{FN} \leq 1$$



Fairness normalization

$$X = (x_1, x_2, x_3, x_4)$$
$$Y = (y_1, y_2, y_3, y_4)$$

$$\upsilon(X, Y) = |f(X) - f(Y)|$$

$$Y = (\mathrm{TP}^{\mathcal{R}}, \mathrm{TN}^{\mathcal{R}}, \mathrm{FP}^{\mathcal{R}}, \mathrm{FN}^{\mathcal{R}})$$
$$X = (\mathrm{TP}^{\mathcal{Q}}, \mathrm{TN}^{\mathcal{Q}}, \mathrm{FP}^{\mathcal{Q}}, \mathrm{FN}^{\mathcal{Q}})$$

$$\upsilon(X, Y) = \underline{\upsilon}$$

$$\phi_{\mathcal{Q},\mathcal{R},f} = 9 \int_{\{X \in \Delta^3, Y \in \Delta^3 | \upsilon(X,Y) \geqslant \underline{\upsilon}\}} \mathrm{d}X \mathrm{d}Y$$

Fairness normalization

$$\phi_{\mathcal{Q},\mathcal{R},f} = 9 \int_{\{X \in \Delta^3, Y \in \Delta^3 | \upsilon(X,Y) \geqslant \underline{\upsilon}\}} \mathrm{d}X \mathrm{d}Y$$

Fairness in [0,1]

Le possibili coppie di matrici di confusione…

…che risultano meno «fair»

Special cases

$$\phi_{\mathcal{Q},\mathcal{R},f} = 9 \int_{\{X \in \Delta^3, Y \in \Delta^3 | \upsilon(X,Y) \geqslant \underline{\upsilon}\}} \mathrm{d}X \mathrm{d}Y$$

- Statistical parity $\phi = \frac{1}{5}\left(5 - 12\underline{\upsilon} + 20\underline{\upsilon}^3 - 15\underline{\upsilon}^4 + 2\underline{\upsilon}^6\right)$
- Conditional accuracy & Calibration $\phi = (1 - \underline{\upsilon})^2$
- Treatment $\phi = 2\frac{\underline{\upsilon} - \ln(1 + \underline{\upsilon})}{\underline{\upsilon}^2}$

Fairness normalization

$$f' = \alpha f + \beta$$

$$\upsilon' = \left| f'^{\mathcal{B}} - f'^{\mathcal{G}} \right| = |\alpha| \left| f^{\mathcal{B}} - f^{\mathcal{G}} \right| = |\alpha| \upsilon$$

$\upsilon' \geqslant \underline{\upsilon'}$ è equivalente a $\upsilon \geqslant \underline{\upsilon}$

Name		f
statistical parity	F_1	$\frac{\text{TP+FP}}{\text{TP+TN+FP+FN}}$
conditional accuracy	F_2	$\frac{\text{TP}}{\text{TP+FN}}$
false positive/negative rate	F_3	$\frac{\text{FP}}{\text{TN+FP}}$
calibration	F_4	$\frac{\text{TP}}{\text{TP+FP}}$
false discovery/omission rate	F_5	$\frac{\text{FP}}{\text{TP+FP}}$
treatment	F_6	$\frac{\text{FP}}{\text{FN}}$

Flipping: an abstract model for fairness improvement

- Select a subgroup in each group
- Flip the decisions in those subgroups with certain probabilities depending
  - On the group
  - On the original prediction

$$
\begin{aligned}
\hat{\mathrm{TP}}^{\mathcal{Q}} &= \mathrm{TP}^{\mathcal{Q}} - \nu_{\mathrm{P}}^{\mathcal{Q}}\mathrm{TP}^{\mathcal{Q}'} + \nu_{\mathrm{N}}^{\mathcal{Q}}\mathrm{FN}^{\mathcal{Q}'} \\
\hat{\mathrm{TN}}^{\mathcal{Q}} &= \mathrm{TN}^{\mathcal{Q}} - \nu_{\mathrm{N}}^{\mathcal{Q}}\mathrm{TN}^{\mathcal{Q}'} + \nu_{\mathrm{P}}^{\mathcal{Q}}\mathrm{FP}^{\mathcal{Q}'} \\
\hat{\mathrm{FP}}^{\mathcal{Q}} &= \mathrm{FP}^{\mathcal{Q}} - \nu_{\mathrm{P}}^{\mathcal{Q}}\mathrm{FP}^{\mathcal{Q}'} + \nu_{\mathrm{N}}^{\mathcal{Q}}\mathrm{TN}^{\mathcal{Q}'} \\
\hat{\mathrm{FN}}^{\mathcal{Q}} &= \mathrm{FN}^{\mathcal{Q}} - \nu_{\mathrm{N}}^{\mathcal{Q}}\mathrm{FN}^{\mathcal{Q}'} + \nu_{\mathrm{P}}^{\mathcal{Q}}\mathrm{TP}^{\mathcal{Q}'}
\end{aligned}
\quad \rightarrow \quad \hat{\phi}_{\mathcal{Q},\mathcal{R},f}
$$

Flipping: an abstract model for fairness improvement

Administer flipping rates to maximize fairness

$$\max_{\nu_{\mathrm{P}}^{\mathcal{Q}},\nu_{\mathrm{N}}^{\mathcal{Q}},\nu_{\mathrm{P}}^{\mathcal{R}},\nu_{\mathrm{N}}^{\mathcal{R}}} \quad \hat{\phi}_{\mathcal{Q},\mathcal{R},f}$$

$$\text{s.t.} \quad \hat{p}^{\mathcal{Q}} \geqslant p_{\min}^{\mathcal{Q}}$$

$$\text{s.t.} \quad \hat{p}^{\mathcal{R}} \geqslant p_{\min}^{\mathcal{R}}$$

Keep performance above a minimum level

Conclusion

Normalized Fairness

- Pragmatic
- Simple
- Allows (fair!) comparison between different criteria
- Allows studying trade-off between criteria
- Allows studying trade-off between fairness and performance