Post-Processing Calibrated Classifiers

Ran Canetti[1 3], Aloni Cohen[2], Nishanth Dikkala[2], Govind Ramnarayan[2], Sarah Scheffler[1], Adam Smith[1]

[1] BU; [2] MIT; [3] Tel-Aviv U. Email: {canetti, sscheff, ads22}@bu.edu, {aloni, govind, ndikkala}@mit.edu

Calibration

A **calibrated** score accurately represents the confidence of a binary classifier. A classifier $\hat{S}$ is calibrated if $\forall s \in [0,1]$ for which $\Pr[\hat{S}(X) = s] > 0$,

$$Pr[\text{true type of } X = 1 \mid \hat{S}(X) = s] = s.$$

Popular methodology when given imperfect information:

1. Construct calibrated soft classifier
2. Post-process to get a final binary decision

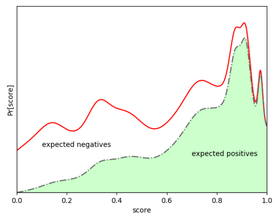
Can Calibrated Scores Be Turned into "Fair" Decisions?

When trying to post-process a calibrated score into a binary decision:

- There is **no general way to equalize even one binary fairness constraint** across two protected groups (e.g. equalizing either FPR, FNR, PPV, NPV)
- Single, global thresholds do not guarantee any fairness outcomes.
- When the sets of scores returned by the classifier for different groups share at least one element, one binary fairness constraint can be equalized.
- **Allowing the post-processor to defer** on some decisions allows all four considered binary fairness constraints to be achieved (FPR, FNR, PPV, NPV).
  - Technical condition: some score should occur with nonzero probability in all groups.
- When using deferrals, the treatment of deferred results is important for determining the fairness of the overall system.

Distributions on Calibrated Scores

We call the probability mass function of the scores output by a calibrated classifier the **DOCS**, or Distribution on Calibrated Scores. We often compare two DOCS, corresponding to a classifier being run on two different protected groups.



Partial References (see paper for further related work)

[1] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *CoRR*, abs/1703.00056, 2017.

[2] J. M. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. *CoRR*, abs/1609.05807, 2016.

[3] J. Larson, S. Mattu, L. Kirchner, and J. Angwin. How we analyzed the compas recidivism algorithm. propublica, 2016.

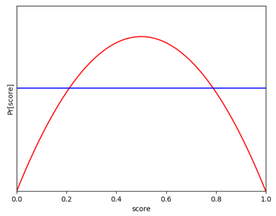
Limitations of Post-Processing

We study the following fairness notions for *binary* classifiers

$$\mathsf{PPV} = \Pr[\text{true type of } X = 1 \mid \text{predicted type of } X = 1]$$

$$\mathsf{NPV} = \Pr[\text{true type of } X = 0 \mid \text{predicted type of } X = 0]$$

We show that post-processing DOCS to equalize PPV or NPV is not possible in general, and requires extra structure. Even with extra structure, cannot equalize PPV and NPV simultaneously by thresholding, exhibited by this example:

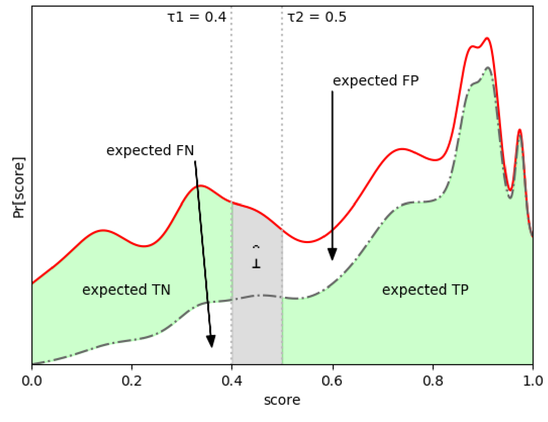


The PPV and NPV of these distributions cannot be simultaneously equalized using one threshold per group, even though the groups have equal base rates. Intuitively, this is because the "line" DOCS corresponds to a much better soft classifier.

Deferrals

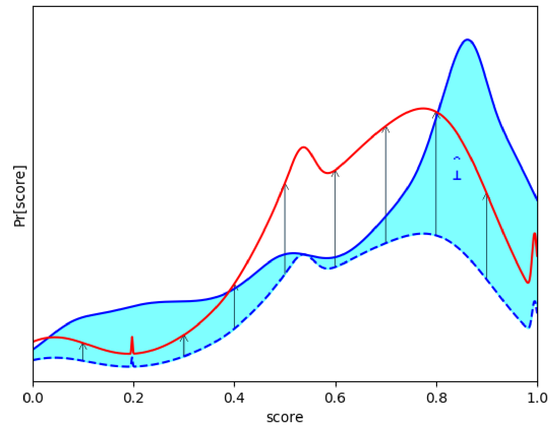
We allow classifiers to **defer** on some points, indicated by $\perp$. The distribution of $\perp$ results will be biased, but the remaining non-$\perp$ scores can bypass the impossibility results of [2, 1]. However, since the $\perp$ results are chosen in a biased way, the nature of the downstream decision-maker who must deal with the deferrals is important. For example, the downstream decision-maker may be effective but expensive, have poor performance on a specific group, or bear a high cost to the individual being classified. The nature of this decision-maker strongly influences how appropriate it is to defer.

Thresholding with Deferrals

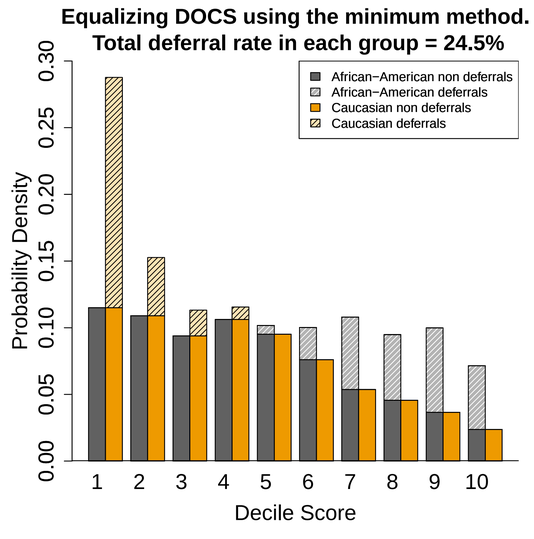


By setting two thresholds per group, two binary fairness constraints (equalized PPV and equalized NPV) can be met in expectation. The thresholds will likely differ between groups.

DOCS Transformation with Deferrals



Transform one Distribution on Calibrated Scores (DOCS) into another by deferring with with probability based on the calibrated score. Conditioned on not deferring, the new distribution remains calibrated. This can be used to set the DOCS of one group equal to the DOCS of another group, equalizing PPV, NPV, FPR, and FNR across all groups. We applied this approach to COMPAS recidivism data.



We equalize DOCS on COMPAS data [3] by taking the pointwise minimum. Different methods of equalization could be preferable in different scenarios, depending on the downstream decision maker and the societal effect of deferrals.

Acknowledgments

R.C. is supported by CPIIS, the NSF MACS project and ISF grant 1523/14. A.C. is supported by NSF award CNS-1413920. N.D. is supported by NSF CCF-1617730, CCF-1650733, and ONR N00014-12-1-0999. G.R. is supported by NSF awards CCF-1665252 and DMS-1737944. S.S. is supported by the Clare Boothe Luce Graduate Research Fellowship and NSF award 1414119. A.S. is supported by NSF awards IIS-1447700 and AF-1763665.