More Impossibilities between Calibration and Balance

Kevin Su, Amir Zur, and Jade Lintott, advised by Omer Reingold

Department of Computer Science, Stanford University

Notions of Fairness and Incompatibilities

- It is important that classifiers and predictors are fair across protected groups that humans define, denoted as S.
- **Calibration** assures that a *predictor* "says what it means" - of all individuals in S (and not in S) receiving a score of X from the predictor, X of them should in fact have a positive outcome.
- **Balance** assures that a *classifier* isn't discriminating through persistent errors - it should have the same false positive (and false negative) rate across individuals in S as those not in S.
- Both calibration and balance are important to ensure fairness: without calibration, a predictor might be less precise for one group than another; without balance, a classifier might be less accurate on one group than another.
- In impossibilities proved simultaneously by Kleinberg et al. and Chouldechova, neither a single classifier not a single predictor can be both balanced and calibrated if the groups have different base rates.

A Possibility in the Impossibility?

- It is still possible to construct a balanced classifier from a calibrated predictor; the question is, how useful is it to do so?
- One such way is to use two separate cutoff thresholds, one for S and the other for all but S. But this blatantly breaks calibration - scores don't mean the same thing across different groups.
- Reich et al. proposed a different algorithm which, using a calibrated predictor, constructs a balanced classifier given a **single threshold**.

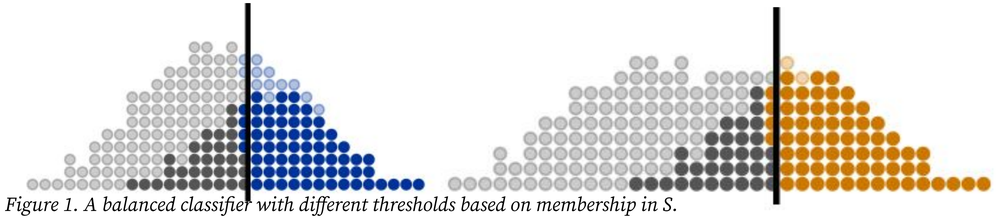

Figure 1. A balanced classifier with different thresholds based on membership in S.

A Closer Look

- Reich et al. first produce a classifier that satisfies equalized odds. They then derive a calibrated predictor *given* a single threshold by postprocessing the Bayes optimal predictor using an optimal transport method.
- Two issues:
  - Normally, a cutoff threshold is determined by the decision maker *after* the predictor is derived. But here, the predictor depends *on* the threshold. Given different thresholds, there would be different predictors and scores would mean different things across *thresholds*.
  - The algorithm can be potentially harmful in some situations. It artificially boosts the predictive scores of the group with the lower base rate and lowers the predictive scores of the group with the higher base rate.

Simpler is Better

- On the other hand, the two cutoffs thresholds holds the following advantages:
  - The predictor is independent of the threshold. Hence, it can be used to inform a cutoff.
  - It makes explicit the affirmative action implied by balance

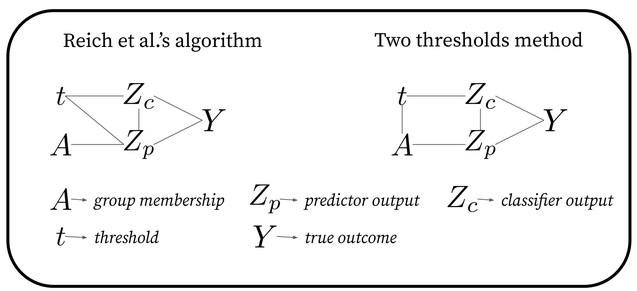

Figure 2. Graphical representations of the dependencies between random variables of the two methods.

A New Impossibility Result

- In a fashion similar to the earlier impossibility result, we showed that the following four conditions cannot be simultaneously true for a predictor, classifier, threshold, and protected group S if the groups have different base rates.
  - (i) The predictor is calibrated
  - (ii) The classifier satisfies equalized odds in a way that is independent of group membership
  - (iii) The predictor does not depend on the threshold
  - (iv) The threshold is independent of group membership
- Reich et al.'s algorithm satisfies (i), (ii), and (iv).
- The two thresholds method satisfies (i) and (iii), and satisfies equalized odds but in a way that depends on group membership