Automating Procedurally Fair Feature Selection in Machine Learning

Clara Belitz, Lan Jiang, and Nigel Bosch – University of Illinois Urbana-Champaign – cbelitz2@illinois.edu

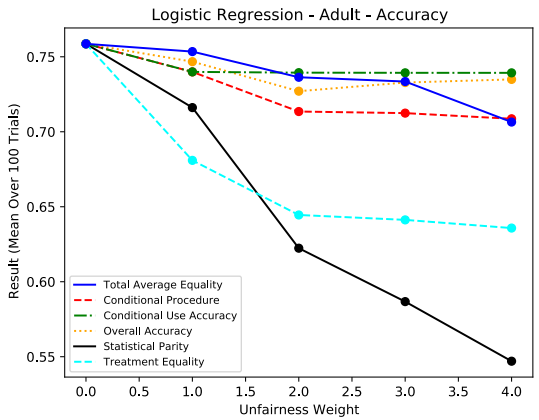
Overview

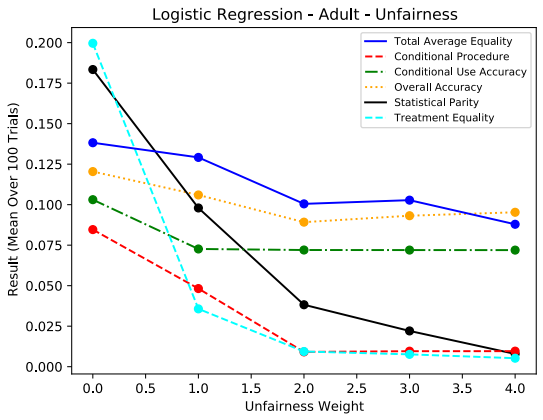
Working from the idea of procedural fairness (i.e. fair process), we propose changes to feature selection.

Weight used to penalize unfair features, making them less likely to be included in the model.

RQ1: Does this method reduce unfairness? If so, according to which unfairness definitions?

RQ2: How does the selection of unfair features affect accuracy and unfairness overall?





Approach

$$max \sum (accuracy - weight * unfairness)$$

- Accuracy measured using AUC
- Weight is a tunable hyperparameter
- Unfairness measured using 6 different statistical metrics
- Four datasets: 3 from UCI ML repository, 1 simulated
- 5 unfairness weights [0-4]

Results

RQ1: Including an unfairness penalty did reduce unfairness for the relevant fairness definitions.

RQ2: Penalizing features caused a reduction in available information, so accuracy was reduced as unfairness was reduced. Generally a proportional and acceptable reduction, with no affect on accuracy if there was no equivalent reduction in unfairness.