Zafar article

Most articles before this article focus on disparate treatment and disparate impact when discussing the notion of unfairness. The authors Zafar et al propose an alternative notion of unfairness called disparate mistreatment. Disparate mistreatment occurs when the rate of erroneous decisions (false positives and false negative) for a sensitive attribute are different and the outcome, "ground truth," is available for review.

The article described their theory by using a dataset associated NYPD Stop-question-and Frisk program and used the binary sensitive attribute of gender focusing on the misclassification measures:

Overall misclassification rate (OMR)

False positive rate (FPR)

False negative rate (FOR)

The authors walk us through how to train decision boundary-based classifiers without disparate mistreatment resulting in a fair logistic regressor can be determined by solving for the following constrained optimization problem:

$$\begin{aligned} \text{minimize} \quad & -\sum_{(\mathbf{x},y)\in\mathcal{D}} \log p(y_i|\mathbf{x}_i,\boldsymbol{\theta}) \\ \text{subject to} \quad & \tfrac{-N_1}{N}\sum_{(\mathbf{x},y)\in\mathcal{D}_0} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \\ & +\tfrac{N_0}{N}\sum_{(\mathbf{x},y)\in\mathcal{D}_1} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \leq c \qquad (18) \\ & \tfrac{-N_1}{N}\sum_{(\mathbf{x},y)\in\mathcal{D}_0} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \\ & +\tfrac{N_0}{N}\sum_{(\mathbf{x},y)\in\mathcal{D}_1} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \geq -c. \end{aligned}$$

The authors methodology removes disparate mistreatment by using fairness constraints and avoids disparate treatment by not using sensitive attribute information in decision making. They conduct experiments using their methodology on synthetic data sets and the ProPublica COMPAS dataset and were able to remove disparate mistreatment while minimally compromising accuracy.

Hardt

The method proposed in this work aims to achieve fairness of a similar notion to disparate mistreatment by "post-processing the probability estimates of an unfair classifier to learn different decision thresholds for different sensitive attribute value groups and applying these group-specific thresholds at decision making time" (Zafar 3). The method cannot be used in when there is no sensitive attribute information (whether it's unavailable or prohibited due to disparate treatment laws) as the it requires this information in decision time.

Their fairness measure first, fixed problems with the fairness of demographic parity (i.e. they were able to correct for bias in the use of fico scores along sensitive attribute of race in determining who gets a loan). Second, they were able to improve the accuracy of their classifiers.

They applied their method to FICO score data (fico scores, people with a history of not paying off prior dept, and race of people applying for a load) to improve the fairness of historically disadvantaged races being unable to receive loans. This was in attempt to make it more fair for disadvantaged groups to qualify for loans.