CIS399: The Science of Data Ethics | **Lecture 11: Feb 21, 2019**
Instructor: Michael Kearns & Ani Nenkova | **Scribes: Drew Boyette, Regina Lin**

1 Recap

Previously, we explored a constrained optimization problem that tried to find $\hat{h} \in H$ minimizing $\hat{\epsilon}(h)$ subject to $\hat{u}(h) \leq \gamma$. Now we use the same heuristics to solve fairness problems. We want to find $\hat{h} \in H$ that minimizes $\lambda\epsilon(h) + (1-\lambda)\hat{u}(h)$, with intervals for lambdas in discrete sizes. The more weight on gamma/lambda, the weaker the fairness constraint is.

In this cost sensitive classification problem, our goal is to trace out an entire convex curve (pareto distribution) of models with varying fairness and error constraints. We begin by finding the model that minimizes error only, at the top left of the curve, and then consider fairness constraints.

2 Building the Pareto Curve

Given $\hat{h} \in H$ with minimum error, we want to build a new classifier $\tilde{h}$.

	Gender($\bar{x}$) = male	Gender($\bar{x}$) = male
$\hat{h}(\bar{x}) = 0$	p	q
$\hat{h}(\bar{x}) = 1$	r	s

We have the following:

- $p, q, r, s \in [0, 1]$
- Conditional probability that $\tilde{h}(\bar{x}) = 1$
- $p = q = 0$
- $r = s = 1$
- FP rate of $\tilde{h}$ on males = $pPr[\hat{h}(\bar{x}) = 0] + rPr[\hat{h}(\bar{x}) = 1]$
- FP rate of $\tilde{h}$ on females = $qPr[\hat{h}(\bar{x}) = 0] + sPr[\hat{h}(\bar{x}) = 1]$
- FP rate of $\tilde{h}$ on males - FP rate of $\tilde{h}$ on females = $(p-q)Pr[\hat{h}(\bar{x} = 0)] + (r-s)Pr[\hat{h}(\bar{x}) = 1]$
- Letting $q = p$ and $r = s$ will satisfy constraint

Our goal is to choose p, q, r, s to minimize $\hat{\epsilon}(\tilde{h})$ such that $\hat{u}(\tilde{h}) \leq \gamma$. In the worst case scenario, the pareto curve is a straight line.

3 Definitions of Fairness

- **Group fairness:** few in number, statistical guarantees
- **Subgroup Fairness:** Many in number, statistical guarantees against ”fairness gerrymandering”

- **Individual Fairness:** one in number. It is difficult to account for individual fairness because values are discrete (e.g. if we have many groups of size one applying for a loan applicant, we will either deny or accept applications, and we therefore cannot balance some percentage of fairness across a group).

3.1 Interpolating Between Groups and Individuals

Problem: we achieve group fairness by subgroup discrimination.

- E.g. disabled Hispanic women over age 55
- N.B. Facebook hate speech policy

We cannot generally protect arbitrarily refined groups (individuals) and there is no reason to expect this practice won't happen under standard fairness notions.

Formulation:

Given a population distribution: $< x, y > \approx P$, we have a model h (in class H) that makes predictions h(x) and g(x) as a membership indicator function of some subgroup. We say h is γ-false-positive fair w.r.t. g if:

$$P[g(x) = 1 \ \& \ y = 0]x[P[h(x) = 1 \ \& \ y = 0] - [P[h(x) = 1 \ \& \ y = 0|g(x) = 1] < \gamma$$

Now we ask that h be γ-fair with respect to all g in some rich but limited class G (e.g. G = linear thresholds, or G = conjunctions, with γ as a parameter allowing us to trade off error and unfairness).

To learn a G-fair model from H, we formulate our problem as a 2-player repeated zero-sum game.

- **Learner** has pure strategy space H of hypotheses. It will minimize error and ignore fairness entirely in its first move.
- **Auditor** has pure strategy space G of subgroups and will find the violated subgroup.
- **Nash equilibrium** yields error-optimal model in H subject to G-fairness