CIS399: The Science of Data Ethics — **Lecture 10: Feb 19, 2019**
Instructor: Michael Kearns & Ani Nenkova — **Scribes: Zoe Colbert & Solomon Goldfarb**

1 Section 1: Recap

At the start of class, we recapped our discussion from last lecture on how to incorporate fairness into our model.

1.1 Subsection 1.1: The Old Problem

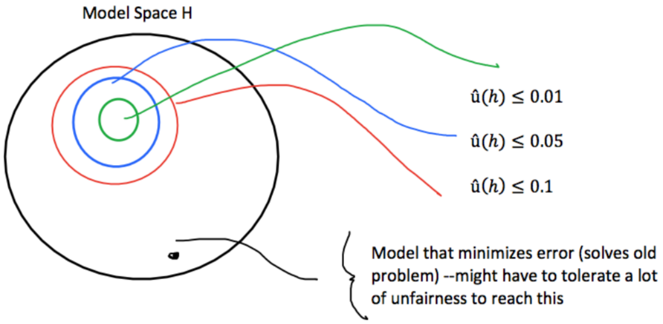
We begin with the assumption that if we have enough data, we know that our sample is a good proxy for true errors. The old problem can be summarized with the formula,

$$\hat{h} = argmin\{\hat{E}(h)\}$$

1.2 Subsection 1.2: The New Problem

The new problem includes a consideration of fairness. We want to find a model in class h that minimizes error AND is subject to constraint that observed unfairness is $\leq$ parameter gamma, $\hat{u} \leq r$.

It is unlikely that any model meets the requirement of gamma = 0.0. It becomes more likely to find a model as gamma increases. This concept is demonstrated in the image below:


Model Space H
û(h) ≤ 0.01
û(h) ≤ 0.05
û(h) ≤ 0.1
Model that minimizes error (solves old problem) --might have to tolerate a lot of unfairness to reach this

The problem turns into an error-unfairness trade off. We can plot $\hat{E}$ vs. $\hat{u}$ on a scatter plot. Models along the Pareto frontier are efficient. For any model not on the frontier, you can strictly decrease one of the criterion without harming the other by moving to a model that is on the frontier.

If we enlarge model class to allow for probabilistic mixtures (see below), we can ensure that the Pareto frontier is a piece wise linear curve.

Let's suppose that H is "closed under mixtures": For any $h_1, h_2 \in H$, for any $\alpha \in [0, 1]$,

$$h_3(\overline{x}) \triangleq \begin{cases} h_1(x) & \text{with probability} \quad \alpha \\ h_2(x) & \text{with probability} \quad 1-\alpha \end{cases}$$

2 Section 2: Constraint Optimization as a Single Criterion

2.1 Subsection 2.1

We would now like to turn the constraint problem into a single criteria. This will allow us to have a problem with a single objective and weight which parameter (error or fairness) is more important.

Consider for $\lambda \in [0, 1]$, $h \in H$:

$$argmin\{(\lambda)\hat{E}(h) + (1 - \lambda)\hat{u}(h)\}$$

If we set λ to 1 we have the original problem of minimizing error. If we set λ to 0 we are just trying to achieve fairness, ignoring error.

2.2 Subsection 2.1

We can prove that for a particular λ, a model h that minimizes $argmin\{(\lambda)\hat{E}(h) + (1 - \lambda)\hat{u}(h)\}$ is Pareto efficient.

Proof by contradiction: If it was not the case that h was on the Pareto frontier, by definition, there would be another model, h', that is southwest of h.

We can formalize this with model h' such that $\hat{E}(h') < \hat{E}(h)$ AND $\hat{u}(h') \leq \hat{u}(h)$. If this is the case, that h' is southwest of h and thus satisfies the above constraints, then h would give strictly lower values for the optimization problem and h would not be the argmin.

3 Section 3: How do we minimize $\lambda\hat{\epsilon}(h) + (1 - \lambda)\hat{u}(h)$?

Suppose we are talking about the class of models which includes all decision trees. Initially the error is very high, but at each step, the algorithm makes a local modification which minimizes the error as much as possible. Similarly, we could replace ”minimizing the error” in this algorithm with ”minimizing the objective function” and create a class of decision tree models the same way. Once we have the objective function and the data, there are many standard heuristic tricks which we can directly adapt to handle the new two part function.

3.1 Subsection 3.1

Now imagine we are given a ”black box” which, for a certain model class, returns a model h which minimizes the error. How could we use this ”black box” to minimize the objective function?
We reduce the problem to Cost Sensitive Classification (CSS); costs for predictions might be asymmetric between data points. For this problem, we consider each data point as an $(\bar{x}, y)$ pair where y can be 1 or 0.
We can calculate a vector $< \bar{x}, l_0(\bar{x}), l_1(\bar{x}) >$ for which $l_0(\bar{x})$ is the penalty for predicting 0 and $l_1(\bar{x})$ is the penalty for predicting 1.

> Whereas previously we might have calculated $\hat{u}(h)$ as a measure of unfairness such as the False Negative Rate (FNR) over females minus the FNR over males, which requires knowing the results over the full sample, this new approach lets us assign penalties on a data point-by-data point basis, such as $\frac{\lambda}{(female\&1)}$

We trick the black box into solving the problem by encoding the costs as replicas of the data points. The more important a data point is, the more copies we make, such that the fraction of the data set made up of replicas of data point a_1 is equal to $\frac{a_1}{a_1+a_2+...+a_n}$.

We call this method **"In-Processing" fairness**, as fairness is considered during the learning.

4 Section 4: An Alternative Type of Solution (Post-Processing Fairness)

4.1 Subsection 4.1: Model $\tilde{h}$

In contrast to in-processing fairness, we also began our discussion of post-processing (bolt-on processing, poor man's). In this type of processing we minimize error and then edit the model around edges to accommodate for fairness.

Given some model h, chosen only for a small $\hat{E}$, we can build $\tilde{h}$, a model that is more fair, with just $\hat{h}$ and a protected attribute.

Model $\tilde{h}$

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

4.2 Subsection 4.2: Choosing p, q, r, s

We choose $p, q, r, s \in [0, 1]$, representing the probability that $\tilde{h}(\overline{x} = 1)$. Essentially, it is a coin flip with the given probability of whether to assign 0 or 1.

We do not condition on $\overline{x}$ or y for two reasons: (1) if we included $\overline{x}$, this would essentially be in-processing fairness. We have already used $\overline{x}$ to build $\hat{h}$. (2) The values of y are not available to us as we have already used our training data to form $\hat{h}$.

Picking $r = s = 1$ and $p = q = 0$ preserves $\hat{h}$ in its entirety, while picking $r = s = p = q = 0.5$, ignores $\hat{h}$.

4.3 Subsection 4.3

In the next class we will discuss the Pareto curve associated with our new model, $\tilde{h}$.