CIS399: The Science of Data Ethics **Lecture 9: Feb 14, 2019**
Instructor: Michael Kearns & Ani Nenkova **Scribes: Monica Vyavahare, Kerry Hollis**

1 Recap from last class

Let $S = G_1 \cup G_2$, such that G_1 and G_2 are two groups that make up S, and G_1 is smaller than G_2. Since G_1 is smaller, the classifier may make more errors on G_1. How do we trade accuracy for fairness? Let H be a classifier trained on the full sample S, H_1 be a classifier trained on G_1, and H_2 be a classifier trained on G_2. H_{new} may be more accurate and fair than H, where $H_{new} =$

$$\begin{cases} H_1 & \text{if sample} \in G_1 \\ H_2 & \text{if sample} \in G_2 \end{cases}$$

H_{new} splits based on group, and uses a different classifier for each group. This may not be effective if one group is particularly small.

2 Fairness Disparity

Recall that H is our class of models, P is the true distribution we are trying to learn, and S is the sample data which we assume follows P iid. Supposed that S is data where gender is the feature we are considering. $u(h)$ is the false positive disparity. The false positive rate on its own can be calculated by

$$FPR = Pr(h(\bar{x}) = 1 | y = 0)$$

i.e. the model predicts positive given that the response variable is truly negative. The disparity can be calculated for any unfairness metric by giving the absolute value of the difference of the metric with respect to different classes. For example, we can calculate the false positive disparity with respect to gender by

$$u(h) = Pr(h(\bar{x}) = 1 | y = 0 \text{ and } male) - Pr(h(\bar{x}) = 1 | y = 0 \text{ and } female)$$

What if there are more than two classes? We can find the disparity for every pair and set $u(h)$ to the max value. This process can be applied to any metric of unfairness, not just false positive disparity.
As with the metrics calculated before, the hat implies that the quantity is calculated from the sample, so $\hat{u}(h)$ can be calculated the same as $u(h)$ by taking the difference of the fraction of false positive makes and false positive females from the sample. If we have enough data (”enough” determined from the theorem from a few lectures ago), we can accurately measure the false positive disparity. If some group is very underrepresented in the data, then our the result may not hold; we need to assume that the sample is large such that each subgrouop is well represented. As such, often the formula holds for gender, since there is usually a 50/50 split in most datasets. Even for race, generally even minorities are represented *enough* for the formula to hold.

2.1 Recap: Old Machine Learning Problem

Recall that we wanted to find

$$\hat{h} = \text{argmin}_{h \in H} \hat{\epsilon}(h)$$

i.e. the classifier that minimizes the error. The rationale was that if we have enough data, then the sample is a proxy for the real world distribution. Thus, the classifier that minimizes error on the sample will likely minimize actual error and generalize well. In the worst case dataset, this is formally intractable, but seems to work well in practice. We may not find the global optimum, but generally in practice a local optima is good enough. For example, logistic regression makes no guarantees on global optimum and can get stuck in local minima instead, but generally works well in practice.

2.2 New Problem: Fairness

Now we also want to consider fairness. We want to find the best predictive model under some fairness constraint. How do we do this?

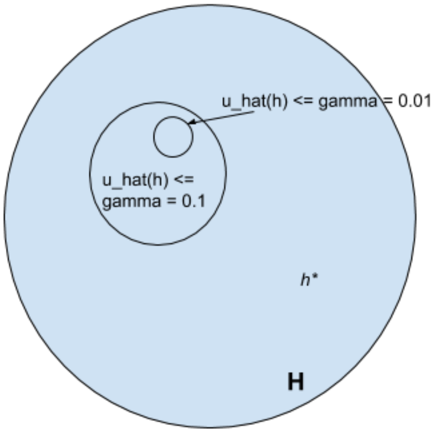
2.2.1 One Possibility

$$\hat{h} = \text{argmin}_{h \in H} \hat{u}(h)$$

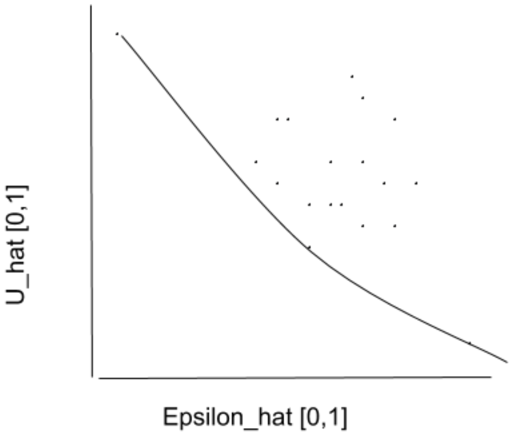
or choose the model that minimizes unfairness.
Is this also potentially intractable? Generally not. The nice property about fairness is that it can be achieved with randomness. A policy with acts the same regardless of X will achieve $\hat{u} = 0$ and $u = 0$, so this is not computationally challenging.
But is this a *good* model? Probably not.

2.2.2 Better: Constrained Optimization Problem

Find the model $\hat{h} \in H$ that minimizes $\hat{\epsilon}(h)$ subject to $\hat{u}(h)$ being smaller than or equal to some γ, for any fairness measure u.
This is also most likely intractable. The fairness consideration has not been studied as much as the old problem, but from a cursory analysis, we can see that it is probably at least as hard as the old problem.



In the figure to the left, the largest circle represents H, or the class of models considered. $h*$ is the optimal model which yields the optimal error. The smallest circle represents the models which satisfy $\hat{u}(h) \leq \gamma$ where γ is small, such as 0.01, and the medium-sized circle represents the models which satisfy $\hat{u}(h) \leq \gamma$ where γ is slightly larger, such as 0.1. This figure demonstrates that eventually when γ is relaxed and made large enough, you may select $h*$, but as a result, a lot of unfairness will have to be tolerated, too. Often, there is a tradeoff between the fairness and accuracy.



The image above demonstrates this tradeoff. The x-axis is $\hat{\epsilon}$, ranging from 0 to 1 inclusive, and the y-axis is $\hat{u}$, also ranging from 0 to 1 inclusive. Each point represents a potential model. We can see a cloud of models, but none quite close to 0 error and 0 unfairness. In fact, these are probably inversely correlated. We can draw a frontier connecting the models closest to (0,0), as can be seen with the line in the graph. This frontier probably has the best models, because any model not on the frontier can be switched to a model on the frontier that is strictly better on (at least) one axis without affecting the other axis. By better, we mean that some model h dominates h' if $\hat{\epsilon}(h) \leq \hat{\epsilon}(h')$ **and** $\hat{u}(h) \leq \hat{u}(h')$ and at least one of the less than or equal to ($\leq$) is actually a strictly less than ($<$). The question then is where on the frontier do you want to be, depending on whether you are prioritizing error or fairness first.

This concept applies to most models with two criteria; eg. stocks volatility (standard deviation of returns) vs the returns. The curve represents the tradeoffes we face. Though, it may not always look so pretty and convex. We define convex as if any two points are connected, the connecting line will be above the curve.

We suppose that our class H is closed under mixtures. This means that if we take any two models and create a new model from them then that third model is also in the class. That is, for any $h_1, h_2 \in H$, define

$$h_3(\bar{x}) = \alpha h_1(\bar{x}) + (1 - \alpha) h_2(\bar{x})$$

for any α [0,1]. Then,

$$\epsilon(\hat{h}_3) = \alpha\epsilon(\hat{h}_1) + (1 - \alpha)\epsilon(\hat{h}_2)$$

$$u(\hat{h}_3) = \alpha u(\hat{h}_1) + (1 - \alpha) u(\hat{h}_2)$$

Then, we know that we can achieve any model on the curve with a probabilistic mixture of the models on each side, assuming the sample size is big enough that we can make the model more complex. Claim without proof from class: the increasing complexity resulting from probabalistic mixtures is actually negligible to the amount of data needed.

Question: What are the ethical implications of "coin flipping", or selecting which model to use based on probability α?.