CIS399: The Science of Data Ethics | **Lecture 8: Feb 12, 2019**
Instructor: Michael Kearns & Ani Nenkova | **Scribes: Saranya Sampath, Saur Vasil**

1 From Last Time

Last time, we were looking at definitions of classifiers.

EFN: $|Pr[h(\bar{x})] = 0|y = 1$ & male$] - Pr[h(\bar{x}] = 0|y = 1[$female$]| \leq \gamma$

PPV: $|Pr[y = 1|h(\bar{x}) = 1$ & male$]Pr[y = 1|h(\bar{x}) = 1$ & female$]| \leq \gamma$

Statistical parity was also discussed. When we discuss this, we can use the example of giving out loans. The rate at which you give loans to men and women should be the same in an optimal fairness scenario, regardless of y value, as we discussed in the last lecture. This doesn't allow the lender to take into account the credit worthiness of certain individuals.

All definitions have the property that giving out loans randomly will always be fair for gamma = 0. A question was asked during lecture: if you give everyone a loan and they happen to be female, no loans for everyone and they happen to be male, then is this fair? Piazza note will be fixed to answer this question.

Almost always, fairness and accuracy will be in tension with each other, as perfect accuracy is not really achievable.

2 How unbiased inputs can actually result in more biased outputs

So far, the definitions we have talked about have constraints on the output $h(\hat{x})$. Rather than restricting the output, we can also restrict the input. Let's talk about credit score. Based on information from consumers, you can be assigned a credit score. There are laws that forbid companies to using inputs like race in making a decision in regards to this score. You can build whatever h you want, but none of the variables in $\bar{x}$ are allowed to be discriminatory aspects, such as race.

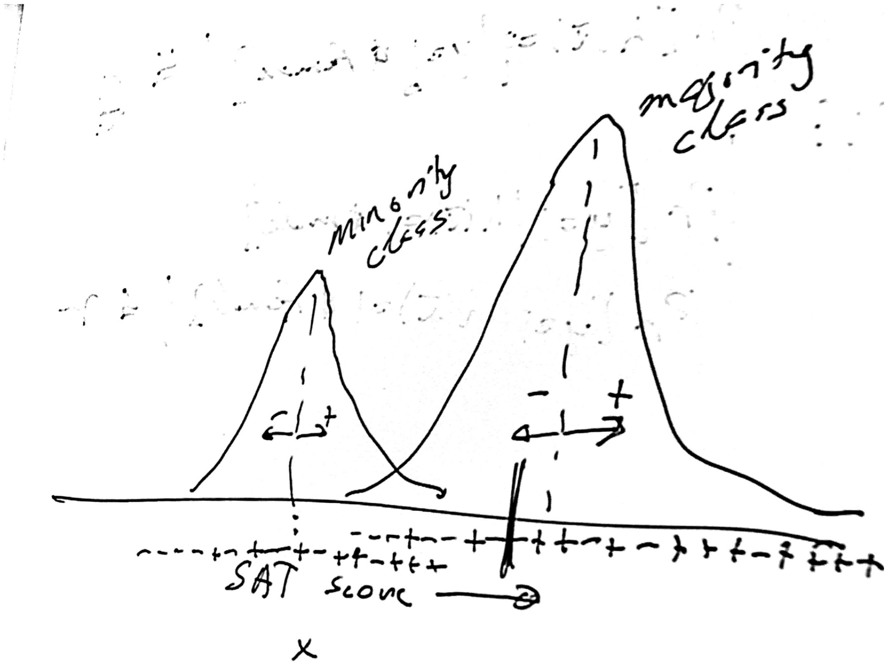
What are some flaws in this? They are sacrificing accuracy for fairness. Also, any variable that is restricted from being used is able to be estimated accurately from other variables. Therefore, you can think that you are excluding factors such as race but there is still a a factor of unfairness there that can be derived in alternate methods. This can actually makes things actively worse, because if you could accidentally harm that minority race you were trying to be fair towards by forbidding race from being a factor.

Imagine that you are trying to decide who to admit to college, and you are only using the SAT score. Let's say there are two different races applying: a majority and a minority. The majority tends to be from wealthy families, and minority tend to be from poorer families. The majority is able to take SAT multiple times and take classes, whereas the minority group don't get this. So, you would expect the majority to get better grades.

Maybe it's just as likely that the majority and minority will both succeed in college, but it is still systematically shifted down in terms of the minority's SAT score. We are using the SAT score as a proxy for college preparedness, but that is not necessarily a good assumption to make.

So if we build a race blind model and only use the SAT score to admit students to college, the combined data between the majority and the minority will allow more people from the majority class to be selected (even though there are people from the minority class that will also succeed in college). The false rejection rate on the minority rate thus becomes very high.

2.1 Example Graph of SAT Scores for Minority/Majority Class



What is the moral of this story? Trying to be fair by being "race-blind" can lead to a model that is highly discriminatory of the minority class. An alternative is to look at the data and realize that there is the same amount of people in both classes that will succeed in college. What should be done is having 2 cut offs for SAT scores for each class. This model will not be race blind but it will not be as discriminatory as the "race-blind" model. This creates a more fair and accurate model.

3 Conflicting definitions of fairness

There are many different combinations of these factors to achieve fairness, and this causes an impossibility. This can cause problems like watchdog groups saying a model is unfair by 2 definitions, but a company could state fairness by the third definition. Researchers have found it is impossible to achieve all three, unless the classifier is a perfect predictor, or the base rates between two groups is identical (ex. females and males pay back loans at perfect equality). In real world, this is not really possible. This shows us fairness can be in contention with accuracy, and fairness itself can be in contention with other definitions of fairness.

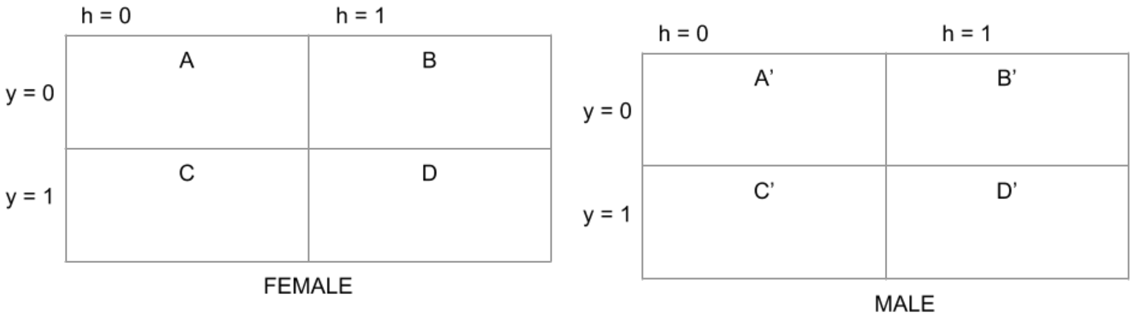
$$a + b + c + d = 100\% \text{ (of females)}$$

$$a' + b' + c' + d' = 100\% \text{ (of males)}$$

$$\text{SP} : b + d == b' + d'$$

$$\text{EFP} : b/(a + b) == b'/(a' + b')$$

3.1 Confusion Matrix



	h = 0	h = 1
y = 0	A	B
y = 1	C	D

FEMALE

	h = 0	h = 1
y = 0	A’	B’
y = 1	C’	D’

MALE