DS 102 Data, Inference, and Decisions **Spring 2020**

Lecture 5: Fairness in Decision-making

Lecturer: Moritz Hardt

Today, we'll consider what it means for decisions rule to avoid discrimination based on features or information that we consider protected or sensitive. In practice, such fairness ideas are nearly inevitable.

5.1 What is Fairness?

Fairness in decision-making is a broad topic which draws upon numerous fields. We will study it in a simple statistical setting. Although this will give only a small glimpse into the subject, we will see that some deep, interesting problems still arise. In particular, fairness is not a purely technical problem, and we will see the need to consider larger implications.

In our setting, we have features X (the data, all of the information collected about each individual), a target variable Y (in our previous notation, this was our parameter of interest θ), and a decision rule $\delta(X)$. Typically, we will think of δ as a threshold rule $\delta(X) = \mathbb{1}\{R(X) > t\}$ which performs binary classification based on some score $R(X)$.

Example 5.1. $R(X)$ could be:

- a likelihood ratio (in Neyman-Pearson framework / Likelihood Ratio Test)
- conditional expectation (which we was the Bayes optimal score under the squared loss)
- some score which is learned from the data (which will be explored in more detail in upcoming lectures)

One important complication in many practical settings: the features X may contain some sensitive A. We assume that A is a discrete random variable which partitions the data X into groups.

Example 5.2. A could be one of the legally recognized "protected classes" under U.S. law, such as: race, sex, religion, age, disability status, and genetic information. (However, there is ongoing debate about what qualifies as protected, even amongst these classes. For example, should "sex" extend to sexual orientation?)

Such protected features might arise in "regulated domains" which, under U.S. law, include: credit, education, employment, hiring, and housing. This also extends to marketing and advertising (including ranking of ads or of one's personal timeline/newsfeed), which makes questions regarding sensitive data and fairness in machine learning relevant to work done by almost every major tech company.

The key question we would like to address is, how do we avoid making decisions that discriminate based on these sensitive attributes?

5.2 Fairness through Unawareness

Can we avoid difficult questions surrounding fairness in decision-making by simply ignoring all sensitive attributes in the data?

It turns out that simply removing anything that looks like a sensitive category from the data is not an adequate solution. Some of the remaining features (or some combination of them) might be proxies for sensitive attributes.

Example 5.3. Zip code may function as a proxy for race.

For example, in the U.S., redlining maps were once used to deny financial services to individuals from certain areas. While information about race was never explicitly used in the decision-making process, redlined areas were often those with a high proportion of minority residents. Thus, discrimination based on race still effectively occurred. However, similar patterns of discrimination can still arise even unintentionally. More recently, maps of Amazon same-day delivery coverage showed that same-day delivery had lowest availability in predominantly Black areas. In this case, decisions about same-day delivery service were likely made based on predictions of the number of same-day purchases, so the route to discrimination may have been through socioeconomic status.

Fairness through unawareness fails; one will still run into issues regarding fairness and discrimination in machine learning by trying to ignore these problems. Data science and machine learning is excellent at discovering proxy variables, so if a sensitive attribute gives a boost in predictive accuracy, machine learning will likely discover it. "We didn't look at *that*" is never a valid argument to defend a model which leads to discrimination in practice.

What would be better than unawareness? In the following sections, we will examine three common statistical criteria that help us answer the question "equality of what?" when trying to design fair decision rules. In particular, discuss:

- equalizing positive rates (acceptance rates)
- equalizing error rates (false positive and false negative rates)
- equalizing false discovery and false omission rates

5.3 Equalizing positive rates

When we equalize positive rates, we require that for any two groups a and b, the probability of making a positive call in group a is the same as the probability of making a positive call in group b:

$$\mathbb{P}(\delta(X) = 1|A = a) = \mathbb{P}(\delta(X) = 1|A = b).$$

For more complex settings (for example, for multi-class classication or regression), the generalization of this rule is to require that $\delta(X)$ is independent of A.

Does enforcing equal positive rates solve all issues of fairness? It turns out that one can come up with decision rules that are indisputably unfair but still satisfy the criterion of equal positive rates.

Example 5.4. Suppose there are two groups, $A = a$ and $A = b$. Further suppose we make good, informed decisions in group a, but random decision in the other group, b. As long as we equalize the positive rate (e.g. in each group accept 20% of individuals), this overall decision rule satisfies the equal positive rate condition. However, in group a the error rates are probably quite good – that is, we are likely not making many false positives. In group b, even though the overall positive rate is what we want to satisfy the fairness criterion, the error rates will be quite poor.

Even when we are not adversarially designing unfair decision rules, decision rules which satisfy the equal positive rate condition may be unfair. For example, we may have less or lower quality data for one group.

Example 5.5. The Framingham risk score for coronary heart disease was created on a cohort of white men, then used for other patients. It performed well on other white men, but led to many false positives for other types of patients.

Equalizing positive rates cannot solve all issues of fairness because it is not just the number of positive calls that matters. In particular, true positives and false positives are fundamentally different, and, intuitively, one should not be able to match a true positive in one group with a false positive in another.

5.4 Equalizing error rates

Instead of equalizing the overall rates of positives, we can consider requiring equal false positive rates and false negative rates for any two groups:

$$\mathbb{P}(\delta(X) = 1|Y = 0, A = a) = \mathbb{P}(\delta(X) = 1|Y = 0, A = b),$$
$$\mathbb{P}(\delta(X) = 0|Y = 1, A = a) = \mathbb{P}(\delta(X) = 0|Y = 1, A = b).$$

Recall that these error rates correspond to the "row-wise" rates in the confusion matrix 5.1

		decision		
		null (0)	*non-null (1)*	
reality	*null*	n_{00}	n_{01}	$n_{00} + n_{01}$
	non-null	n_{10}	n_{11}	$n_{10} + n_{11}$
		$n_{00} + n_{10}$	$n_{01} + n_{11}$	N

Table 5.1: Different ground truth and decision relationships in multiple testing.

Similarly as for the positive rate criterion, there is a generalization of this rule: require that $\delta(X)$ is independent of A conditional on Y.

Error rate parity is a *post-hoc* criterion: at decision time, you don't know who is truly a positive/negative instance. In hindsight, somebody can collect a group of positive instances and a group of negative instances and check how well they were classified. Group differences in this kind of *post-hoc* "audit" often strike people as unfair.

This error rate parity condition has a nice interpretation in terms of the ROC curve. Suppose $\delta(X)$ is a threshold rule. It is possible to produce an ROC curve for each group, and figure out how to choose a threshold to satisfy error rate parity by inspection. In particular, error rate parity implies that the ROC curve of score conditional on group must lie under all individual curves.

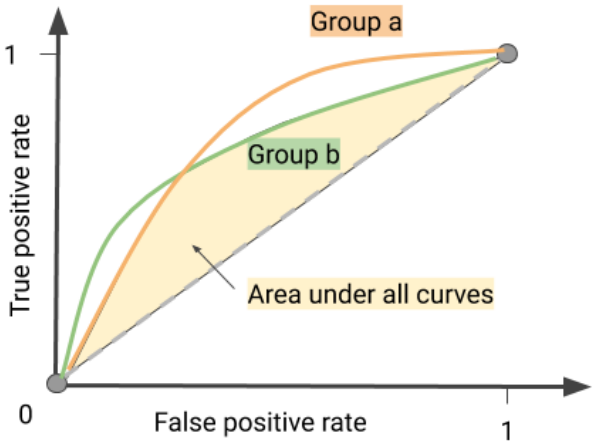

Figure 5.1: Error rate parity implies that ROC curve of score conditional on group must be under all curves.

One key criticism of this approach: in order to equalize the error rates for all groups, it will be necessary to make the predictions worse for some of the groups. Arguably, rather than worsening the predictions for some groups, it would be better to think critically about why the error rates are different between groups and try to address some of the underlying causes.

5.5 Equalizing column-wise rates

We could consider the column-wise rates in the confusion matrix 5.1 instead of the row-wise rates, and create analogous fairness criteria for false omission and false discovery rate. There is nothing wrong with this approach *per se*, but people often consider a slightly different quantity called **calibration** instead.

Suppose your decision is a threshold rule $\delta(X) = \mathbb{1}\{R(X) > t\}$. Calibration means $\mathbb{P}(Y = 1|R = r) = r$. Among all the examples that get score r, on average an r proportion of them should actually be positive. Calibration by group would be $\mathbb{P}(Y = 1|R = r, A = a) = r$.

Calibration is a fairly natural notion to consider for fairness because it is an *a priori* guarantee. The decision-maker sees the score $R(X) = r$ at decision time, and knows based on this score what the

frequency of positive outcomes is on average.

Example 5.6. A doctor uses a scoring rule for the current patient and sees a score of $R(X) = 0.6$. He know that, on average among all patients with that score, 60% have heart failure. Note that this does *not* guarantee that the current individual has a 60% chance of heart failure, but still gives interpretable information at decision time.

5.6 Incompatibility results

At this point we have seen three different fairness criteria: equal acceptance rate, equal error rate, and calibration by group. Each criterion has different strengths and weaknesses, and each makes a different assumption about what it means to be "fair" or "just." Can we have them all?

A collection of results known as "incompatibility results" prove that these three fairness criteria are mutually exclusive. Thus, we cannot simultaneously guarantee that even two of the three criteria are satisfied at once. Work by Hardt, Price, & Srebro provides further exploration of the different criteria and the trade-offs between them.

The fact that error parity and calibration are mutually exclusive has been most significant in terms of public debate.

Theorem 5.7. *Assume the groups have different base rates ($\mathbb{P}(Y = 1|A = a) \neq \mathbb{P}(Y = 1|A = b)$) and $\delta(X)$ has nonzero error rates. Then, if error rate parity holds, group calibration cannot hold.*

This trade-off was studied closely in part because of an investigative article on the COMPAS risk score, which is used by many jurisdictions in the U.S. to assess risk of recidivism. Judges can use the score in part to detain defendants. ProPublica found that Black defendants face higher false positive rate – that is, more Black individuals labeled as "high risk" end up not committing a crime upon release than among white individuals labeled "high risk." COMPAS makers Northpointe rebutted this criticism by arguing that the scores are calibrated by group and Black defendants have a higher recidivism rate. Hence, this trade-off is unavoidable.

Although incompatibility results meant that Northpointe was eventually let off the hook in this case, the situation still seems unsatisfying. When addressing issues of fairness, it is important to consider whether a non-technical solution would be better than applying these three statistical criteria.

5.7 Incentives

One further criticism of the fairness criteria we have seen thusfar is that they do not blatantly rule out unfair practices – in fact, forcing a decision-maker to satisfy them can lead to bad incentives.

Example 5.8. Suppose we choose to detain everyone who has a recidivism score about 50%. If one group has a lower average score than another, this rule will lead to very different detention rates and false positive rates for the two groups. Enforcing equal false positive rates might incentivize

us arrest more low risk individuals in the group that has a higher average score. Since those individuals are low risk, they will neither be detained nor recidivate. This is clearly a harmful practice in reality, but nonetheless has the desired effect of equalizing error rates.

Calibration faces similar issues with bad incentives. For example, one could replace the scores of a subgroup of individuals with the average score of that subgroup in order to satisfy calibration requirements, but this may lead to a decision rule that does not treat each individual fairly. See this 2017 paper by Corbett-Davies et. al for more details about these examples.

5.8 Outlook on fairness

What conclusions can we draw about fairness in decision-making from this discussion? First, fairness through unawareness fails. Working in data science or machine learning, one will inevitably run into issues of fairness, and it will be necessary to think carefully about them and confront them head-on.

Although better than unawareness, statistical fairness criteria on their own cannot be a "proof of fairness." They can provide a starting point for thinking about issues of fairness and finding useful solutions. In particular, they can help surface important normative questions about decision-making, as well as trade-offs and tensions between different potential interpretations of fairness.

For a deeper discussion of these and related ideas, see the Fairness and machine learning textbook by Barocas, Hardt, & Narayanan.