MITOCW | Fairness Criteria, Exploring Fairness in Machine Learning

[MUSIC PLAYING]

MIKE TEODORESCU: Hello, and welcome to this module on choices of fairness criteria. My name is Mike Teodorescu. I'm an assistant professor of information systems at Boston College, as was a visiting scholar at MIT D-Lab. What we'll cover in this module will be the concept of confusion matrix, as well as three popular examples of fairness criteria-- demographic parity, equalized odds, and equalized opportunity.

Some countries have laws that protect specific groups of people from discrimination based on certain attributes. As we review in the previous video, these are called protected attributes. Some examples are on this slide. Regardless of the legal landscape, machine learning has the potential to produce unfair outcomes for certain groups of people. As an algorithm designer, one should make clear choices about fairness criteria. Some criteria will be reviewed in the next few slides.

[? In the ?] previous video, we also discussed case of fairness through unawareness, which refers to leaving out the protected attributes out of your model. And we also explained why this is not a good choice. Specifically, you may end up with other attributes that correlate with protected attributes, and you may end up discriminating inadvertently nonetheless.

In order to go into additional fairness criteria, we need to discuss some additional concepts. Consider you have a binary classifier. For example, you're looking at decision of hire or not hired or to lend credit and not to lend credit. If we want to look at the predictions for a model that would do such a binary classification, we would look at the predicted values.

We could bucket them in four categories-- true positives, which would be correctly classified as positive, true negative, correctly classified as negative, false positives, which would be values that were incorrectly classified as positive by the algorithm, false negatives, which would be values incorrectly classified as negative. And, if we were to add the the true positives to the true negatives and divide that by all four, we would end up with the value of the accuracy of the model. In this example where accuracy is this fraction, an accuracy of 0.5 is the same as a random classification.

Now we should look at accuracy carefully and see that it doesn't tell us anything about the prediction of negatives. It could mislead us if two classes were imbalanced, for example, if 90% of the sample is positives, and 10% is negatives. For that, we have other additional criteria we could go into deeper, such as MCC score and AUC score.

The true positives, true negatives, false positives, and false negatives are, oftentimes, represented in a 2 by 2 matrix form called a confusion matrix. This is simply an easier presentation for us to see the behavior of the classifier.

The first additional fairness criteria is called demographic parity. It's a criterion for what's called group-level fairness. This criterion specify that the outcome, which here is denoted by a y hat, is independent of the protected attribute A. For example, the probability of being hired is independent of the gender.

There are multiple problems with demographic parity. One would be the definition that we just discussed would not hold if we had individuals who would be members of multiple protected groups. By enforcing group-level fairness for one attribute, we would still violate the group fairness for other attributes or combinations of attributes, such as subgroups of the population.

Furthermore, while enforcing group-level fairness, for example, the same hiring grade for females and males, that could still be unfair to individuals. It could force the algorithm to drop otherwise qualified individuals just to achieve independence of outcome of the attribute. Fairness at the group level could, potentially, be unfair at the individual level.

For example, if we have a high rate of false positives, we could end up-- and a low rate of false negatives, it could still end up being unfair to individuals in that we could hire people who are without merit at the disadvantage of other individuals who could be qualified and should be hired, but, due to the group-level fairness criterion, we have to hire some who are not qualified from one of the groups.

The sweet spot would be low false negatives and low false positives, which would be fair, potentially, to both the individual and the group level. We could also end up in the top right corner, which would be the worst-case scenario, low accuracy, unfair to individuals, but potentially fair for the group where we have high false negatives and high false positives.

A stricter criterion is equalized odds, which means matching both of true positive rate and the false positive rate for different values of the protected attribute. This is much harder to achieve than demographic parity, but it is one of the higher levels of algorithmic fairness.

In this case, rather than predicting the same average risk across subgroups of protected social attributes, like in demographic parity, equalized odds is stricter in that it forces equality only among individuals who reach similar outcomes. In the hiring example that we've discussed in the previous video, this implies that the probability of a qualified applicant to be hired and the probability of an unqualified applicant to be incorrectly hired should be the same across genders.

A milder version of equalized odds is equalized opportunity where we're only concerned with treating fairly those who are determined to be worthy of acceptance, i.e. dependent variable is equal to 1, or they're worthy of being hired, worthy of being admitted, and so on. Equalized opportunity is not concerned with rejecting people fairly across protected groups.

So to speak, the false positive rates and the true positive rates do not both need to be equal across the protected categories. We only need the true positive rate to be equal across protected categories. In a way, equalized opportunity is less interventionist than equalized odds, and it may be more achievable, depending on your individual situation and implementation challenges.

In the example of hiring, we only are concerned with individuals who are worthy of being hired, i.e. the actual qualified applicants. And, out of the rejected applicants, we may be sometimes rejecting unfairly.

Here's some review questions for the last two videos. What is demographic parity? What is fairness through unawareness? Is fairness at the group level always the best? What is a confusion matrix? What is the equality of odds criterion? And when might you want to use it?

This course is sponsored by the USAID grant through MIT D-Lab. And this is joint work with my faculty colleagues Lily Morse and Gerald Kane from Boston college and research assistant Yazeed Awwad from MIT D-Lab. If you would like to learn more about this, please consult the following references. Thank you so much for watching this video. We hope you find it useful and you'll continue watching the rest of the class.

[MUSIC PLAYING]