Classifier Fairness

Fraida Fund

In addition to evaluating the **error** of a classifier, we are also often concerned with the **fairness** of a classifier.

Suppose samples come from two groups: a and b. What does it mean for a classifier to treat both groups *fairly*? There are a number of different types of fairness, and like the error metrics described previously, we are often stuck in a situation where we must sacrifice on one fairness measure to improve another.

This week's case study expands on this topic.

For this case study, you will work through some online material:

1. First, read the ProPublica article, Machine Bias. Based on their description, what type of fairness is violated by the COMPAS recidivism risk prediction algorithm? (Choose one of the fairness measures described below in the section titled "Fairness metrics".) We will revisit this later in the week.
2. Next, read the article and work through the activity Can you make AI fairer than a judge? The article discusses two types of fairness regarding COMPAS - why is it not possible to reconcile them?
3. You should also interact with the activity on Measuring Fairness, which explains in greater detail why it may not be possible for an algorithm to be fair according to all metricsl.
4. Finally, read The quest to make AI less prejudiced, which discusses some other types of biases and fairness issues in machine learning.

Question

- What is the *original* source of the unfairness in the COMPAS example? Is it the risk prediction model, or is it something else?
- The COMPAS risk prediction model was meant to be used by human judges in the context of other information about the defendant. Do you think the human judges using the model understood and were aware of its overall performance? Do you think they were aware of the fairness and bias issues that the model may have? (Do you know about fairness and bias issues related to machine learning models that *you* use?)
- Suppose that despite its bias, COMPAS is still about as fair or even slightly more fair than a human decision maker. Are you comfortable with COMPAS being used under these circumstances?

Fairness metrics

This section lists some metrics related to fairness. You won't have to memorize these, but you should understand them, and given the definition of any of these metrics, you should be able say whether or not it is satisfied in a particular scenario.

- **Fairness through unawareness** is not a measure of fairness, but describes a situation in which features related to group membership are not used in classification. (This doesn't necessarily mean that the classifier produces fair outcomes! For example, if the classifier is trained on data that reflects an underlying bias in society, the classifier will be biased even if it is not trained on features related to group membership.)
- **Causal discrimination** says that two samples that are identical w.r.t all features except group membership, should have same classification.

- **Group fairness** (also called *statistical parity*) says that for groups a and b, the classifier should have equal probability of positive classification:

$$P(\hat{y} = 1|G = a) = P(\hat{y} = 1|G = b)$$

- **Conditional statistical parity** is a related metric, but now we are also controlling for factor F:

$$P(\hat{y} = 1|G = a, F = f) = P(\hat{y} = 1|G = b, F = f)$$

- **Balance for positive/negative class** is similar to *group fairness*, but it is for classifiers that produce soft output. It applies to every probability S produced by the classifier. This says that the expected value of probability assigned by the classifier should be the same for both groups -
- For **positive class balance**

$$E(S|y = 1, G = a) = E(S|y = 1, G = b)$$

- For **negative class balance**

$$E(S|y = 0, G = a) = E(S|y = 0, G = b)$$

- **Predictive parity** (also called *outcome test*) says that the groups should have equal PPV, i.e. the prediction should carry similar meaning (w.r.t. probability of positive outcome) for both groups:

$$P(y = 1|\hat{y} = 1, G = a) = P(y = 1|\hat{y} = 1, G = b)$$

- Predictive parity also implies equal FDR:

$$P(y = 0|\hat{y} = 1, G = a) = P(y = 0|\hat{y} = 1, G = b)$$

- **Calibration** (also called *test fairness*, *matching conditional frequencies*) is similar to *predictive parity*, but it is for classifiers that produce soft output. It applies to every probability S produced by the classifier:

$$P(y = 1|S = s, G = a) = P(y = 1|S = s, G = b)$$

- **Well-calibration** extends this definition to add that the probability of positive outcome should actually be s:

$$P(y = 1|S = s, G = a) = P(y = 1|S = s, G = b) = s$$

- **False positive error rate balance** (also called *predictive equality*) says that groups should have equal FPR:

$$P(\hat{y} = 1|y = 0, G = a) = P(\hat{y} = 1|y = 0, G = b)$$

- False positive error rate balance also implies equal TNR:

$$P(\hat{y} = 0|y = 0, G = a) = P(\hat{y} = 0|y = 0, G = b)$$

- **False negative error rate balance** (also called *equal opportunity*) says that groups should have have equal FNR. This is equivalent to group fairness **only** if the prevalence of positive result is the same among both groups:

$$P(\hat{y} = 0|y = 1, G = a) = P(\hat{y} = 0|y = 1, G = b)$$

- False negative error rate balance also implies equal TPR:

$$P(\hat{y} = 1|y = 1, G = a) = P(\hat{y} = 1|y = 1, G = b)$$

- **Equalized odds** (also called *disparate mistreatment*) says that both groups should have equal TPR *and* FPR:

$$P(\hat{y} = 0|y = i, G = a) = P(\hat{y} = 0|y = i, G = b), i \in 0, 1$$

- Note that if the prevalence of the (actual) positive result is *different* between groups, then it is not possible to satisfy FP and FN error rate balance *and* predictive parity at the same time!
- **Conditional use accuracy equality** says that the groups have equal PPV *and* NPV:

$$P(y = 1|\hat{y} = 1, G = a) = P(y = 1|\hat{y} = 1, G = b)$$

$$P(y = 0|\hat{y} = 0, G = a) = P(y = 0|\hat{y} = 0, G = b)$$

- **Overall accuracy equality** says that the groups have equal overall accuracy

$$P(\hat{y} = y|G = a) = P((\hat{y} = y|G = b)$$

- **Treatment equality** says that the groups have equal ratio of FN to FP, $\frac{FN}{FP}$