Worksheet

COMPAS is an algorithm that predicts the risk of recidivism.
Consider the following fictional example of classification by COMPAS (Example adapted from *Measuring Algorithmic Fairness,* Hellman, 2020, *Virginia Law Review*):

Social group A

	Recidivates	Does NOT recidivate
Predicted to recidivate	60	20
Predicted NOT to recidivate	6	14

Social group B

	Recidivates	Does NOT recidivate
Predicted to recidivate	16	5
Predicted NOT to recidivate	22	57

Questions:

1. What are the true rates of recidivism (base rate) for social group A and B, respectively?
2. What percentages of group A and B are predicted to recidivate?
3. Out of all people predicted to recidivate in group A, what percentage actually recidivated? What about group B?
4. Out of all people predicted NOT to recidivate in group A, what percentage actually did not recidivate? What about group B?
5. What percentage of group A was NOT predicted to recidivate but recidivated (false negative rate)? What about group B?
6. What percentage of group A was predicted to recidivate but did NOT recidivate (false positive rate)? What about group B?
7. **Based on the above information, do you think the algorithm is fair between the two groups? Why?**

Discussions

Here are some possible answers to question 7:

(a) Unfair – because our answers to 2 are not the same for each group. There is a higher percentage of people in group A predicted to recidivate compared to group B.
- This view is widely rejected because it does not account for difference in base rate (question 1).
- **Are there situations where this criterion is useful or appropriate?**
  Perhaps it is relevant for reparative justice? We might want to give more opportunity to historically underprivileged groups in contexts such as university admission?

(b) Fair – because our answers to 3 and 4 are about the same for the two groups.
- This is called the *calibration* criterion.
- What does it mean to be *calibrated*? Suppose we look at all the days that were forecasted with 60% chance of rain, if on 60% of these days it did rain, then the predictions are well-calibrated. Questions 3 and 4 are the analogous case for binary classification.
- **What happens when predictions are not equally calibrated?**
  The same predictions do not mean the same thing for both groups. For example, suppose we have a classifier that predicts whether a job candidate is likely to succeed in the company, and its predictions are more calibrated for women than for men. This means that a positive prediction about a female candidate is strong evidence that she will succeed, but the same prediction about a male candidate is only weak evidence that he will succeed.

(c) Unfair – because our answers to 5 and 6 are very different for the two groups.
- This is called the *equalized odds* criterion.
- We can also requires only equal false positive or only equal false negative rate.
- **What happens when false positive or false negative rates differ?**
  For example, consider, again, the classifier that predicts whether a job candidate is likely to succeed in the company. Suppose its false positive rate is higher for men than women and its false negative rate is higher for women than for men. Then, a qualified woman is more likely to be denied a job opportunity, and an unqualified man is more likely to be offered a job opportunity.

(d) Insufficient information – We cannot determine whether an algorithm is fair based on statistical criteria.
- **Why might this be?**
  One possibility is that fairness might be determined not by the *outcome* of the predictions, but the *process* by which the predictions are produced.
- **What are some pros and cons to outcome- vs process- based fairness?**

It turns out that the calibration and equalized odds criteria cannot be jointly satisfied, unless (i) the classification is 100% accurate, or (ii) the base rates are equal.

What does this imply? Is one criterion better than the other? If both are necessary for fairness, does this mean that fair classification is impossible?

Multiple choice question 1

Consider the following classification matrices:

Social group A

	Recidivates	Does NOT recidivate
Predicted to recidivate	59	13
Predicted NOT to recidivate	13	15

Social group B

	Recidivates	Does NOT recidivate
Predicted to recidivate	34	26
Predicted NOT to recidivate	26	114

Which of the following is true?

(a) This algorithm is equally calibrated between the two groups.
(b) This algorithm satisfies the equalized-odds fairness criterion.
(c) Both (a) and (b)
(d) Neither (a) nor (b)

Multiple choice question 2

Algorithm X satisfies both fairness criteria *equalized odds* and *calibration* for social groups A and B. Which of the following *must* be true?

(a) The base rates of A and B are the same.
(b) The algorithm makes no error in its predictions.
(c) Both (a) and (b) must be true.
(d) None of the above.