Fairness Primer for Contestants

Introduction - The Danger of Bias

Advances in machine learning and data enable model developers to create highly predictive models. However, these models risk reproducing or even exacerbating discriminatory practices unless they are rigorously assessed for different forms of bias. This primer explores these varying manifestations of bias by explaining several associated definitions of fairness often employed in audit and regulatory contexts.

Types of Fairness

Many notions of fairness can be split into two categories: social fairness and predictive fairness. Each captures different ideas of what fair outcomes reflect. The former take their name from the fact that they are measurements of how fairly a model treats different subgroups, while the latter refer to how accurate models are across subgroups. Deviations from these definitions of fairness can suggest different types of bias. This primer defines four ways of quantifying a model's overall fairness as particularly important: demographic parity, equalized opportunity, differential validity, and calibration. What follows is an illustrative and not an exhaustive treatment of these two types of fairness.

Common Fairness Measurements

Social Fairness	Predictive Fairness
Demographic Parity	Differential Validity
Equalized Opportunity	Calibration

Social Fairness

Demographic parity is achieved when the likelihood of being assigned a favorable outcome in a classification model is identical across subpopulations. For example, if a model is used to provide an offer of a discounted service, demographic parity by sex would mean that the percent of men who received the offer equaled the percent of women who received the offer.

Formally, a predictor satisfies demographic parity with respect to a protected attribute, if:

$$Pr(Favorable = 1 \,|\, Protected = 1) = Pr(Favorable = 1 \,|\, Protected = 0)$$

Equalized opportunity is achieved when the likelihood of being assigned a favorable outcome is similar amongst subpopulations that have the positive label, where the label is the true value of the outcome of interest. It is also called the relative true positive rate. For example, amongst those that actually have a medical condition that a model is trying to detect, equalized opportunity exists where the model is equally likely to predict the presence of the condition across subpopulations. Accuracy for this group is especially important because it concerns the extent to which a model correctly identifies what is often the most important target group for a model in each subgroup.

Formally, a predictor $\widehat{Y}$ satisfies equalized opportunity with respect to a protected attribute and true underlying outcome if:

$$P(\widehat{Y}|Protected = 0, Outcome = 1) = P(\widehat{Y}|Protected = 1, Outcome = 1)$$

Predictive Fairness

Differential validity is achieved when a model's performance is identical across subpopulations of the model. It is also referred to as classification parity. This definition of fairness is a matter of predictive accuracy rather than the similarity of scores by subgroup, and can be measured using conventional techniques like AUC or KS statistics. For example, if a model is similarly good at predicting incidence of a disease among men and women according to some quality metric like AUC, but the disease affects one sex more than another, the model would have differential validity even if it did not have demographic parity.

Formally, where Q is the quality metric used to assess a model, this means that:

$$Q(Protected = 0) = Q(Protected = 1)$$

Calibration fairness is achieved when the true outcomes do not vary by subpopulation for the same predicted values. In principle, it concerns whether a model's predictions are accurate across different subgroups. For example, under a well-calibrated model, Black patients and White patients with the same predicted risk score would have the same likelihood of actually developing the disease.

Formally, recalling that $\widehat{Y}$ is the model predictions while Y are the underlying values, this states:

$$P(Y|\widehat{Y}), Protected = 0) = P(Y|\widehat{Y}), Protected = 1)$$

These four fairness metrics are intuitive but distinct. The fact that improvements in one metric will not necessarily improve, or may even worsen, the values in another metric mean that modelers must bear in mind the context in which they are reviewing their models to know which are more appropriate. Substantial differences in social fairness metrics indicate that some

groups are being treated comparatively worse by a model, while differences in the predictive fairness metrics mean a model is not specified to be able to predict well for all groups of people. Both capture related elements of what is often meant when speaking of a model as being “fair” to those being assessed by a model.

Social and Predictive Fairness Compared

It is important to further distinguish between social and predictive fairness, and discuss some of their common contributing factors. As stated above, social fairness refers to differences in the average outcomes of the applicants by subgroup. This makes it especially important where the underlying characteristics between groups should be comparable, as this implies that individuals are being treated alike regardless of their demographic features. Where this is the case with the label in particular, demographic parity is one of the simplest metrics to consider. Where they are not alike, such as in the case where a medical condition is simply more common for one group— say women, compared to men— other metrics should also be considered. In this case, something like equalized opportunity may be more appropriate because this makes predictive accuracy a matter of how well a model predicts the true positive cases that a model needs to identify most. Ultimately, both try to capture a sense of *fairness as equitability of outcomes* being produced by a model.

By contrast, predictive fairness is not concerned with how the average scores between the groups compare, but rather how well they map on to the underlying values (the labels) of the dataset. This shifts the framework for understanding fairness away from group level comparisons of aggregate outcomes and towards the efficacy of models across each subpopulation. In this way, predictive fairness ultimately stands on the principle of *fairness as accuracy of predictions*.

Despite these differences, both forms of bias share common drivers. They can be caused by biased input data, a poor or unrepresentative selection of features used in the model for all subgroups, a failure to account for data or demographic drift, and other similar aspects. Solutions are difficult to describe without context, and for this reason it is always important to identify the most universal aspects of any corrective methodology that can generalize to other contexts. Addressing these time and structural factors feeding into social and predictive bias requires a way to detect when disparities are emerging and identify their causes. This will continue to be one of the paramount challenges shaping the equity and effectiveness of AI technologies in the 21st century.