Moral Machine Learning
Sample Lecture, Vassar College
Andrew (Andy) Ackerman
Nov. 15, 2024

Contents

1 Setting

Consider a binary classifier $\mathcal{A}$ with target variable $Y \in \{0, 1\}$. Our aim is to classify $\hat{Y}_{\mathcal{A}}$ based on some feature space $\mathcal{X}$. Moreover, let S be a *protected variable*. By this we mean a demographic variable related to both the feature space and the target variable which (as a demographic variable) has a marginalized community as an observation of said variable.

For example, in the widely discussed COMPAS[2] dataset and associated algorithm, we aim to predict the risk of repeat offense from convicted felons. This was initially intended to be used to supplement a judge's discretion in granting parole, but the algorithm (notably closed source) has been criticized as being biased against defendants of color (d.o.c). In this case, our protected variable is racial background. In which case,

$$\begin{gathered} Y \in \{0, 1\} = \{\text{no repeat offense, repeat offense}\}, \\ \hat{Y}_{\mathcal{A}} \in \{0, 1\} = \{\text{low risk, high risk}\} \\ S = \mathbb{1}_{\{d.o.c\}} \end{gathered}$$

2 Statistical Parity and Predictive Parity

We are now ready to define quantitative metrics to assess the fairness of an algorithm. We will begin with *statistical parity* which is an *independence criterion*.

Definition 1 *(Statistical Parity): $\mathcal{A}$ is said to violate ϵ-statistical parity if for $\epsilon > 0$*

$$|\mathbb{P}(\hat{Y}_{\mathcal{A}} = 1|S = 0) - \mathbb{P}(\hat{Y}_{\mathcal{A}} = 1|S = 1)| > \epsilon. \tag{1}$$

Remark: ϵ is left to user choice. It should be clear that a smaller choice of ϵ will result in a more stringent condition for 1. That said, a common legal precedent is the *four-fifths rule* [1] which sets $\epsilon = 0.2$.

A question worth considering is whether statistical parity completely encompasses what we mean by fair. On the one hand it ensure the prediction is made independently of the protected variable ($\hat{Y}_{\mathcal{A}} \perp S$) which is fair at least in the sense of equal treatment across S. However, there is also at least one sense in which statistical parity falls remarkably short. Notice that it does not take the ground truth Y into account. As a result, even an oracle classifier (one that gets the class label correct every time) could violate statistical parity if the ground truth also exhibits disparate rates of Y across S. In this case, the algorithm is deemed unfair when it is really just recovering a property of the ground truth. This is a poor property to exhibit.

To overcome this, *predictive parity* does not just require independence but conditional independence. Specifically, it conditions on the ground truth. As such, it is what is known as a *separation criterion*.

Definition 2 *(Predictive Parity): $\mathcal{A}$ is said to violate ϵ-predictive parity if for $\epsilon > 0$*

$$|\mathbb{P}(\hat{Y}_{\mathcal{A}} = 1|(S = 0 \cap Y = 0)) - \mathbb{P}(\hat{Y}_{\mathcal{A}} = 1|(S = 1 \cap Y = 0))| > \epsilon \tag{2}$$

Note, predictive parity requires $\hat{Y}_{\mathcal{A}} \perp S|Y$, and as such, an oracle classifier will not fail this condition even if the ground truth exhibits differing rates of Y across S. In essence, predictive parity is insisting upon commensurate rates of false positives. Consequently, the metric of fairness here is less about equal treatment across S and more about ensuring that the burden of misclassifications (specifically false positives) is not disproportionately borne across S.

It is worth mentioning that these are far from the only metrics for assessing fairness available in the literature. What is true of each measure of fairness though is that it will have different properties that reflect a different ideal as to what it means to be fair.

[1] EEOC Guidelines

3 Incompleteness Theorem

So is assessing fairness as subjecting an algorithm to a chosen metric of fairness and seeing if it violates the condition? Unfortunately not. These measures (and others not mentioned here) are in some fundamental way, incompatible, and this will leave user with some ambivalence that is worth further discussion. A series of incompleteness results[1] demonstrate precisely what we mean by incompatible.

Definition 3 *(Incompleteness Theorem): No single classifier, $\mathcal{A}$, will (outside of unlikely fringe cases) be able to simultaneously satisfy independence, separation and sufficiency*[2] *type criteria. These fringe cases include a perfect classifier used on a ground truth with no disparity across S.*

The implication of this result are somewhat subtle but profound. Namely, we cannot escape normative considerations such as “what do we value?" Often, the motivation behind developing statistical measures of fairness is to abstract away from such normative questions. By appealing to quantitative measures, we hope to be able to make more objective the assessment of fairness and leave it less up for debate, so to speak. However, this incompleteness theorem shows that we cannot satisfy all measures simultaneously, so we will ultimately have to make a choice as to which measure to prioritize in a given situation. The answer to this question has less to do with quantitative performance and more to do with what we value. What property of fairness do we value in the given context: equal treatment, proportionate burden of misclassifications, something else entirely? These are the questions that we historically try to avoid, but *Moral Machine Learning* aims to directly address. This result shows that, indeed, these questions are unavoidable.

[2]Examples of the first two are given above in statistical parity and predictive parity respectively. We will not have time to discuss sufficiency at length, but it requires $Y \perp S|\hat{Y}_{\mathcal{A}}$.

References

[1] Kleinberg, J., Mullainathan, S., Raghavan, M. (2022). Inherent trade-offs in the fair determination of risk scores.

[2] Larson, J., Angwin, J., Kirchner, L., Mattu, S. (2016, May 23). How we analyzed the compas recidivism algorithm. ProPublica.