A Causal Explanation for the Impossibility Theorem of Fairness

Kailash Karthik S

April 28, 2020

1 Preliminaries

Classification is the problem of predicting a target random variable when a computational model is given as evidence a number of correlated random variables. A fundamental assumption in such a problem is the fact that although the evidence variables can be continuous and from an infinite domain, the target variable have a discrete finite domain. This assumption can be relaxed, allowing the target variable also to take continuous values from an infinite domain - and this problem is called regression.

There have been many successful techniques employed for the problem of classification and regression (SVM, Decision Trees, Elastic Nets, etc.) With the increasing use of these computational models in society, there has been a spike in interest and analysis on the fairness of the predictions made.

The ProPublica report on recidivism predictions showed that the prediction models can be biased towards certain values of certain evidence variables. When such biases are in line with social notions of unfairness, then there is an adversarial bias exhibited by the model. The presence of machine bias can be either necessary or adversarial depending on the situation. For instance, a sentiment prediction model's bias towards words like "despair", "tragedy" to predict the *negative* sentiment is justified and desirable. However, a model's bias towards a social class like racial or gender category while making a decision can be undesirable, especially when the predictions are consequential.

Such social classes that need to be protected from bias in computational models are termed as *protected* or *sensitive* attributes. There are three common metrics to evaluate the presence of machine bias - demographic parity, predictive parity and equalized odds. However, it has been proved that these metrics can not be satisfied at the same time - resulting in the *impossibility theorem of fairness.* While this theorem has been proved statistically, we present here a proof based on causal machinery that provides intuitions on the impossibility using the data generating process.

2 Notation

For simplicity of notation, we consider the task of binary prediction without loss of generality. Let an instance of the data x_i used for prediction of the label y_i be drawn from the distributions $x_i \in \boldsymbol{R^d}$ and $y_i \in \boldsymbol{Y} = \boldsymbol{\{0,1\}}$. Let a be the protected attribute as defined above be drawn from the distribution $a \in \boldsymbol{A}$. The goal of a predictor is to learn a function $f : \boldsymbol{X} \rightarrow \boldsymbol{Y}$ that approximates the true underlying joint distribution $\boldsymbol{X} \times \boldsymbol{Y}$. The model is then used to make predictions $\hat{y}$ on unseen data instances x_i such that it approximates well their true labels y_i. Let $P_a[\hat{y}] := P[\hat{y} \mid A = a]$.

3 Foundations

A model is said to well calibrated or fitted if it approximates the underlying joint distribution well. In such cases, there exists a strong correlation between the true labels y and the predictions $\hat{y}$. This means that for a well calibrated classifier, $Y \not\perp\!\!\!\perp \hat{Y}$ and this will be the focus of the analysis because only if a model is well calibrated can it be used in the real world.

A sensitive attribute can be considered as a source of machine bias only if it has a correlation with the true label. If the true label was indifferent to the sensitive attribute, then the presence or absence makes no difference to both the true label and the prediction. Thus, for this analysis, we only consider sensitive attributes that are correlated with the true labels $A \not\perp\!\!\!\perp Y$ and thus can introduce machine bias if the learning algorithm is not equipped to handle possible bias.

The three popular metrics of fairness are :

1. Demographic Parity - $\hat{Y} \perp\!\!\!\perp A$

   This metric ensures that the predictions are independent of the sensitive attribute. Thus prediction probabilities are equal across all values of the sensitive attribute, thus preventing the model from having disparate prediction bias towards a certain label for any sensitive group.

2. Predictive Parity - $Y \perp\!\!\!\perp A \mid \hat{Y}$

   This metric ensures that the calibration of the model is not dependent on the sensitive attribute value. Thus, the probability of correctness of a prediction is the same for all values of the sensitive attribute. This prevents models from being biased towards making incorrect predictions for any sensitive group.

3. Equalized Odds - $\hat{Y} \perp\!\!\!\perp A \mid Y$

   This metric ensures that the accuracy of the model is not dependent on the sensitive attribute value. Thus, the probability of predictions are independent of the the sensitive attribute for each target label groups. This encourages the model to be faithful to the underlying distribution of sensitive attribute across the target groups.

   The **Impossibility Theorem** states that no more than one of the three fairness metrics of demographic parity, predictive parity and equalized odds can hold at the same time for a well calibrated classifier and a sensitive attribute capable of introducing machine bias.

4 Causal Explanation

We assume that the classifier is well calibrated and the sensitive attribute is correlated with the true labels. This entails that the causal graph corresponding to the data generation process will always have an unblocked path between $Y - \hat{Y}$ and $Y - A$.

D-Separation is a criterion for the identification of conditional independences from causal diagrams. Blocking is defined as the prevention of flow of distributional influences between two nodes in the diagram. Directed paths between any two paths in a causal diagram can be decomposed into sequences of triplets. There are three types of triplets - causal chain, common cause and common effect. While the first two types are unblocked when none of the nodes in the triplets are observed, common effects are unblocked when the common effect node or any of its descendants are observed.

If all the paths between two variables are composed of only active triplets, then the variables are dependent on each other, conditioned on any node that is observed to make the paths active. Such variables are said to be d-connected. If all the paths between the nodes are blocked, either by observed nodes or through common effect colliders, then the nodes are (conditionally) independent and are said to be d-separated.

In the following causal diagrams, let the curved lines denote an unblocked path of arbitrary length and configuration. We also ignore all the attributes in X apart from the sensitive attribute A. The presence and configuration of the other attributes are not relevant to the analysis of fairness and do not change any of the results.

1. Demographic Parity

   The causal diagram of a data generation process corresponding to a classifier that satisfies demographic parity is shown below. It can be observed that Y and $\hat{Y}$ are d-connected through the unblocked represented by the curved line. Similarly, Y and A are d-connected. Thus the assumption that the model is well calibrated and that the sensitive attributed is correlated to true label. It can also be observed that the sensitive attribute is d-separated with the predicted label since the path contains the collider through the node Y.

   This is the only data generation process configuration that satisfies the assumptions on the model as well as demographic parity.

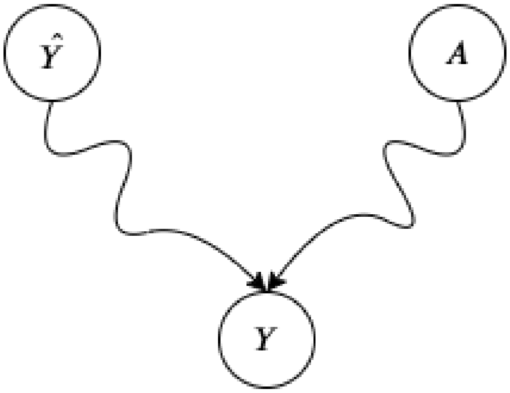

Figure 1: Causal Diagrams for Demographic Parity

   In this process, $\hat{Y}$ is not a d-separating set for A and Y. Thus, observing $\hat{Y}$ does not make them independent and thus predictive parity can not hold.

   It can also be seen that observing Y opens up the collider on the node which makes the sensitive attribute and the prediction d-connected. Thus, equalized odds can not hold true in the model corresponding to this diagram.

2. Equalized Odds

   The causal diagrams for a model of data generation that satisfies the equalized odds notion of fairness as well as the model assumptions are shown below. There are three possible configurations depending the direction of the edges entering/leaving Y. If the node Y is the center of either a causal chain or a common cause triplet, then equalized odds holds as the observation of Y blocks the path between $\hat{Y}$ and A.

   It can be observed from the causal diagram that the path between A and $\hat{Y}$ is unblocked unless Y is observed. This entails that demographic parity can not hold in such a model.

   The prediction label $\hat{Y}$ does not lie between Y and A and thus, its observation hence has no effect on their d-connectedness. Thus, predictive parity can not hold in such settings.

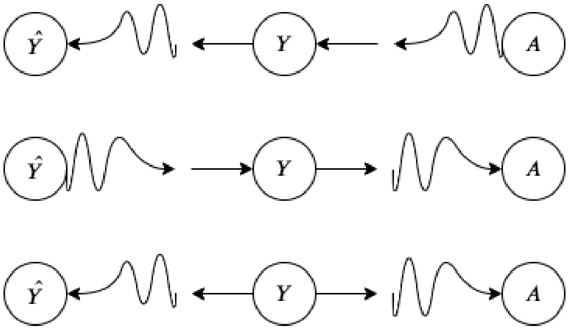

Figure 2: Causal Diagrams for Equalized Odds

3. Predictive Parity

   The data generation process for a model that satisfies both the model assumptions as well as predictive parity is shown below. As mentioned in the equalized odds scenario, there are three possible configurations depending on the directions of the causal relationships of $\hat{Y}$ on the path between A and Y.

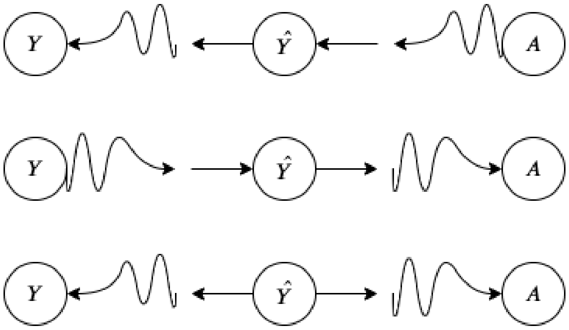

Figure 3: Causal Diagrams for Predictive Parity

   The causal diagram is very similar to the equalized odds scenario. In the same way, the sensitive attribute is always d-connected to the prediction and thus demographic parity can not be satisfied by such a model.

   In addition to this, it can be observed that Y can never d-separate A and $\hat{Y}$ as it does not lie on the unblocked causal paths between them and hence can not block them. It can thus be seen that equalized odds can not hold in such a model.

5 Inferences

From these causal diagrams, it can be concluded that a computational model that satisfies one of the three notions of fairness can not simultaneously hold any of the other two. This inability is not attributed to the lack of capability of the models but rather the restrictions of the data generation regime. We have proved by classifier model-agnostic causal techniques that the impossibility theorem is true.

Thus, a machine learning practitioner building computational models for socially impactful applications must choose one of these metrics to satisfy at the expense of the other two. However, the issue with such a selection is that any model that is deemed to be unfair by any of these metrics can be disputed by citing one of the other two metrics that can be satisfied.

Thus, a refinement of metrics for fairness is required - possibly one that involves counterfactuals so that the graphical constraints observed above can be circumvented.

6 Building Fair Classifiers and Regressors

The reason there is a lack of fairness in many data-driven classifiers is due to the discrepancy between what is deemed correct by the data and what is thought of as ”correct” by the society. Many papers in the literature have mentioned that fairness in machine learning is not a statistical but rather a sociological issue in using machine learnt classifiers in social settings.

While the objective of any learning algorithm is usually empirical risk minimization (ERM) which tries to bridge the gap between the true labels and the predicted labels, the notion of fairness is orthogonal. A fair classifier tries to negate the historical dependence of the true label on the sensitive attribute by making predictions that take into consideration the sensitivity of certain attributes - thereby deviating from the true labels in the (historic) data.

$$\begin{aligned} f_{\text{ERM}} &= \text{argmin}_f[\boldsymbol{P}_{(x_i,y_i)\sim\boldsymbol{D}}(y_i \neq \hat{y}_i)] \\ f_{\text{fair}} &= \text{argmin}_f[\boldsymbol{P}_{(x_i,y_i)\sim\boldsymbol{D}}(y_i^{\text{fair}} \neq \hat{y}_i)] \\ \text{where } y_i^{\text{fair}} &= f_c(y_i) \end{aligned}$$

The true labels from the data are determined to be unfair based on some socially acceptable correction criterion and changed using the corresponding correction function f_c.

The correction function introduce a notion of randomization that depends on the value of the sensitive attribute, trying to disentangle the dependence between the true labels found in data and the desired labels from the classifier. Thus, the goal of the learning algorithm is no longer to produce a model that maximizes the fit of the training data, but rather to fit the training data with this correction function in the picture.

$$Y_{\text{goal}} \neq Y_{\text{ERM}}$$

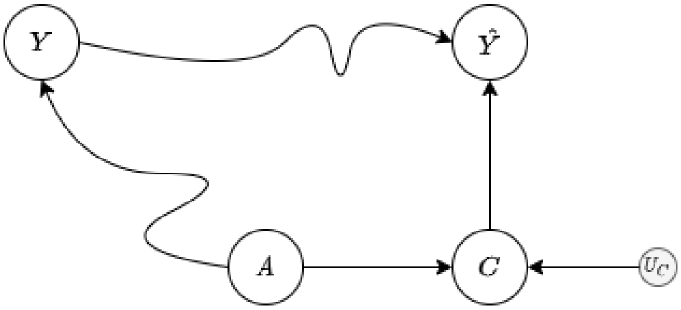

Figure 4: Causal Diagrams with Correction Variable

Thus, a desirable classifier is no longer constrained the requirement for perfect calibration ($Y \perp\!\!\!\perp ?\ \hat{Y}$). The only constrain then is the fact that the sensitive attribute has some effect of the true labels in the dataset ($Y \not\perp\!\!\!\perp A$).

Under this constraint, the causal diagram for the data generation process can then be represented by introducing a correction variable C that blocks the influence of Y on $\hat{Y}$. That is, the correction variable will determine if information from the true label is to be propagated to the prediction or not. The correction variable by itself is a function of the protected attribute ($C \not\perp\!\!\!\perp A$) but does not let information from A pass through it. While the flow of information from Y to $\hat{Y}$ doesn't have to be necessarily blocked for the advantaged classes, this is desired for the disadvantaged class to ensure fairness.

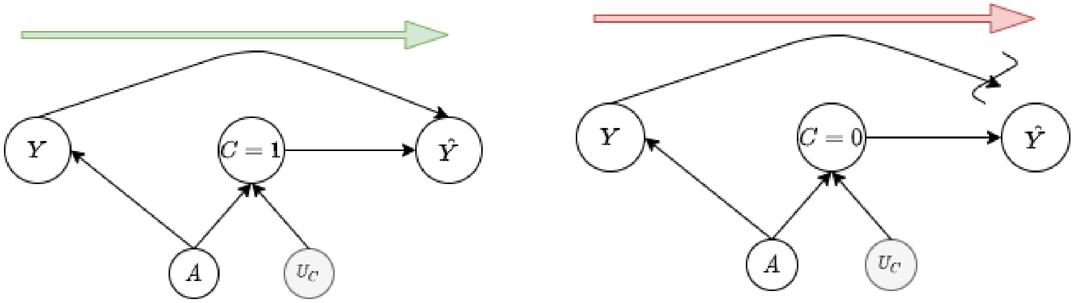

Figure 5: Effect of Correction Variable

Existing Fairness Metrics using Correction Variables

1. Demographic Parity

   In the new causal mechanism, though A is dependent on $\hat{Y}$, the influence is only through the activation or disabling of the gate C. Thus, we achieve a looser version of demographic parity where the sensitive attribute does not have any direct causal path to the prediction and all the paths are through the correction variable C.

2. Equalized Odds

   In this scenario as well, conditioned on the true label Y, all the causal paths from the sensitive attribute to the prediction are through the correction variable and there exists no other influence.

3. Predictive Parity

   In the new scenario, predictive parity can never be achieved as there is always a causal influence of the sensitive attribute on the true label Y. I argue that this metric should not be satisfied as it is against the intuition we developed earlier about the desired classifier. The introduction of the correction mechanism was to selectively offset the true labels in the dataset with the aim of negating existing bias. Thus, a fair classifier is one that would preferentially reverse the true label Y for the disadvantaged class $A = a$ that has been subject to prejudice. Thus, we would expect that the $P(Y = v \mid \hat{Y} = v, A = a) > P(Y = v \mid \hat{Y} = v, A = d)$, where v is any value in the range of Y and $\hat{Y}$, a is the advantaged group and d is the disadvantaged group.

Modified Fairness Equations

In the new regime which includes the correction variable, the notions of demographic parity and equalized odds can be satisfied together conditioned on the correction.

$$\hat{Y} \perp\!\!\!\perp A \mid C = 0$$

$$\hat{Y} \perp\!\!\!\perp A \mid Y, C$$