Bias in ML – Notes

Jonathan Meyer

1 Concepts

The aim of this work is to examine the introduction of bias into datasets through a more formal vision. As any bias is in essence an imbalance in the data, the approach will be to formalize ways datasets can be transformed (i.e. ways imbalance can be introduced, or the observed distribution can be changed). Note that we will consider datasets to be finite, which will also facilitate the concretization of these transformations. We will consider a dataset consisting of n points which are described by a set of features $X = \{X_1, X_2, ...X_k\}$, a sensitive attribute A, and a label G. For the sake of simplicity, we will assume the sensitive attribute A and label G are binary. Also, w.l.o.g. we will consider $A = 0$ to be the unprivileged group, and $G = 1$ as the desirable label. Assuming that these points are i.i.d., a statistic for any such dataset follows the form

$$\Omega_n = \prod_{i=0}^{n} X_i A_i Y_i$$

In order to transform this data, a number of operations may be applied to the datapoints. The goal of these transformations is to achieve an arbitrary target distribution. We will consider three types of operations:

- Row removal
- Row modification
- Row addition

In practice, these correspond to the actual ways a dataset could be manipulated: either by adding or removing rows (rows, points, tuples – the terms are used interchangeably here), or by changing the features of specific rows. At the core, these transformation form the basis for any manipulations to the dataset; both for introducing bias and for removing it.
We will formalize these operations by considering functions which are applied to the dataset (i.e. they are applied to each row, resulting in a change in the overal distribution of the data). Each of the above operations hinge on the principles of 'selection' and 'execution'. By selection we simply mean which points are affected by the operation (e.g. 'positively labeled tuples of the favored group'). This selection can be done according to an arbitrary distribution. 'Execution'

is taken to mean the actual result of the operation (e.g. 'remove the row from the dataset').

General form

The chosen operations can be described as probabilistic functions of the form $f^{op}(X, A, G, C)$ (abbreviated $f_C^{op}(X, A, G)$) where X, A and G are the features and label of the tuple, *op* is the type of operation, and C is a stochastic variable which describes the distribution used for selecting points. We will assume that the operations are only conditional on A and G, in accordance with the idea that the sensitive attribute encompasses all the conditions that could lead to bias.
For any point, the chosen operation is either executed or not - conditional on A, G, and C. C therefore plays the role of a 'decision process'. Although it can have an arbitrary set of outcomes, in practice we will think of this as a binary variable: either $C = +$ (the operation is executed) or $C = -$ (the row is left unchanged).

Removal

The first type of operation is the removal of rows. In practice, this can be seen as selectively sampling the data (as a sample is just a subset of the dataset, which could be obtained by removing all undesired points). When removing data point-by-point, the transformation has two possible outcomes: either the point is removed, or it remains unchanged. The outcome of course depends on the selection that is made, which depends on C. In this case, $C = +$ implies the removal of the row. In order to indicate this removal, we will introduce the outcome d, which represents a deleted tuple that can effectively be ignored. We can now think of row removal in terms of the function

$$f_C^{rem}(X, A, G) = \begin{cases} X, A, G, & P(C = -|A, G) \\ d, & P(C = +|A, G) \end{cases} \tag{1}$$

The first outcome corresponds to leaving the tuple unchanged, while the second corresponds to deleting the tuple. $P(C|A, G)$ describes which tuples are removed. For instance, say the goal is to suppress the favored group, then $P(C = +|A = 1, G = 1) > 0$ and $P(C = -|a, g) = 1$ for all other $a \neq 1 \wedge g \neq 1$.

Modification

The second type of operation involves changing the labels of selected points. This has been described as 'massaging the dataset' in order to remove discrimination, but can also be applied to introduce bias instead. In this case, selected tuples will have their label changed to another outcome in the domain of G. In this work, as G is binary, this effectively means 'switching' the label between 0 and 1. However, in general, the new label will be denoted by Y. Similarly to

row removal, the outcomes of this transformation can be described as:

$$f_C^{mod}(X, A, G) = \begin{cases} X, A, G, & P(C = -|A, G) \\ X, A, Y, & P(C = +|A, G) \end{cases} \tag{2}$$

Addition

The final type of operation is the addition of new points to the data. In practice this can be implemented in many ways (e.g. by selectively sampling with replacement, or duplicating rows). In general, applying this operation to a dataset can be considered as a re-weighting of the points. Points can be duplicated into the data multiple times, effectively giving each point a corresponding coefficient γ. The value of the coefficients is conditional on A, G, and C:

$$f_C^{add}(X, A, G) = \gamma(X, A, G) \tag{3}$$

γ can be considered a stochastic variable of its own, which maps the values of $A \times G \times C$ to $\mathbb{R}^+$.