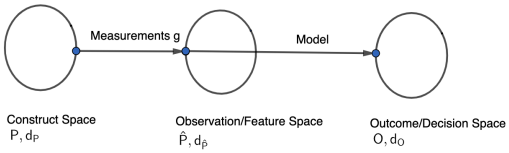
On the (Im)possibility of Fairness

January 14, 2020

This lecture focused on the work of Friedler, Scheidegger, and Venkatasubramanian, which aimed to focus on part of the ML pipeline that most algorithm designers gloss over or consider the domain of other experts: namely, that when making predictions or decisions about individuals for some task, whatever data we have to make that decision with respect is the result of some measurement of individuals, and those measurements are necessarily imperfect surrogates for some properties of individuals we'd ideally like to use directly when making decisions. Many notions of fairness ask that similar people be treated similarly, and to formalize such a statement, it is necessary to describe our representations of individuals, the features we observe about them, and how similar or dis-similar individuals are for the particular task we're trying to accomplish.

There are many examples which are appropriate to flesh out here; a few tasks we discussed in class were in the space of hiring, college admissions, and online advertisement. For each of these tasks, a decisionmaker has some abstract concept of what properties of an individual they would like to use to make a determination about that person. For example, for hiring, an employer might want to hire workers as a function of their dedication, reliability, and relevant skills or experience. These properties of an individual are abstract, difficult to define precisely, and rarely if ever directly measurable, but are nevertheless the properties a decisionmaker might want to use to govern their choices when accomplishing their task. We refer to these properties of a person as belonging to a person's representation in *construct space*.

The desirable, relevant properties of individuals for a task (such as reliability in the case of hiring) are difficult to describe precisely. One can imagine observations of individuals which may correlate with these properties (did the person graduate from high school? Do previous employers say they were always on time for their shifts? Did the person stay in their last role for more than six months?), but even with all of these observations about an individual, it may not be possible to reconstruct the properties of that individual in construct space. We refer to the process of gathering observations about individuals as taking *measurements*. A collection of measurements applied to an individual in construct space produces a collection of observations (or features) representing that individual; this feature representation of an individual belongs to another metric space we refer to as the *observation space*, or feature space.



Representations of an individual in observation space are the data decisionmakers (and machine learning models) have access to, and use to predict or decide outcomes of their task. For example, an employer has access to an applicant's resume, which contains the number of months or years of experience the candidate has in different roles. The decisionmaker then makes either a prediction (of these candidates whose resumes I have, which will show up to work every day for 3 months?), or a decision (of these candidates, which will I hire?). Both predictions and decisions are *outcomes* the decisionmaker selects amongst. We then name the metric space over outcomes the *outcome* or *decision space*, (O, d_O).

So, to summarize and add a bit of notation, for a given task, a decisionmaker wants to make choices about some n individuals as a function f of their representation in construct space $X_1, \ldots, X_n$, but doesn't have direct access to these properties and therefore doesn't have the ability to evaluate $f(X_1, \ldots, X_n) = o$, the desired outcome for this task. Instead, the decisionmaker looks at a set of measurements $Y_i = g(X_i)$ for each individual i, and makes a decision as a function of those measurements for the n individuals: $\hat{f}(Y_1, \ldots, Y_n) = \hat{o}$. The question then is how well this process works, which can be informally described as the relationship between o and $\hat{o}$, will certainly depend on the measurements employed as well as the function of those measurements used to make decisions.

1 How well do these measurements and decision process reflect our goals for a task?

One way to evaluate how well a process accomplishes the goal of the task is to consider the extent to which either measurements or models or their composition affect the *distances* between individuals in different spaces. In particular, the informal notion that "similar people should be treated similarly" can be translated into a statement about individuals' distances.

Definition 1. *The* additive distortion *of a function $f : A \to B$ over a set A is defined as*

$$\rho_f = \max_{a,a' \in A} |d_A(a, a') - d_B(f(a), f(a'))|$$

where d_A, d_B are distances over A and B, respectively.

The additive distortion of a function measures the maximum amount that function changes distances between pairs of points. So, the larger the distortion of f, the less similarly f is guaranteed to treat similar points. However, distortion is a worst-case measurement, and therefore other evaluations of spread may suit some applications better than this worst-case formulation of f's ability to change distance.

It is also useful to ask how we can evaluate the distance between sets of points, which have different frequencies of appearing, rather than just pairs of points. In order to compare sets of points, we will describe a way of identifying points in the first set with points in the second. One way to do so is to define probability measures μ_A, μ_B on sets A and B, and to consider couplings of these measures and sets:

Definition 2. *A probability measure ν over $A \times B$ is a coupling measure of $\nu(A, \cdot) = \mu_A$ and $\nu(\cdot, B) = \mu_B$. That is, ν couples the probability measures on A and B if the projections onto A and B have the corresponding probability measures those sets had to begin with.*

We can measure the distance between two sets by finding the coupling that makes the distance between the sets as small as possible:

Definition 3. *Given a metric space (X, d) and $Y, Y' \subseteq X$ with probability measures $\mu_Y, \mu_{Y'}$. The Wasserstein distance between Y and Y' is then*

$$W_d(Y, Y') = \min_{\nu\text{: coupling measures over } Y,\ Y'} \int d(y, y')\nu(y, y').$$

and it will then be useful to ask how to measure distance between sets of points in different metric spaces (for example, before and after measurement). For this reason, we introduce the Gromov-Wasserstein distance (GWD).

Definition 4. *Given $(X, d_x), (Y, d_Y)$, tow metric spaces with probability measures μ_x, μ_y, the Gromov-Wasserstein distance between X and Y is*

$$GW(X, Y) = \frac{1}{2} \inf_{\nu\text{: coupling measures over } X, Y} \int \int |d_X(x, x') - d_Y(y, y')\| d_{\mu_X} \times d_{\mu_X} d_{\mu_Y} \times d_{\mu_Y}.$$

The GWD measures something like the shape of X compared to Y, at least inasmuch as the sets of points' shape has to do with the pairwise distances these sets exhibit.

Given the machinery defined, it is now possible to define one mathematical notion of fairness that one might like to satisfy in some scenarios. This notion is a formalization of the intent that "similar people should be treated similarly", where we interpret the similarity of people to be defined as small distance in construct space (which, we should recall, has a distance function that is decidedly task-specific), and "treated similarly" to mean they should have small distance in outcome space.

Definition 5. *A mapping $f : P \to O$ is said to be (ϵ, ϵ')-fair if, for any $x, y \in P$,*

$$d_P(x, y) \leq \epsilon \Rightarrow d_O(f(x), f(y)) \leq \epsilon'.$$

This notion is weaker than asking that f have small distortion (for appropriate parameter values), which would require that all distances be (approximately) preserved.

The question, when given a particular construct space, observation space, and decision space, is how one assumes (or observes) individuals are treated under the mappings between these spaces. Are distances approximately or exactly preserved by measurement and model? Does the amount by which distances change differ for different subsets of construct space?

This perspective leads us to the following collection of worldviews, which make different assumptions about the spaces' relationships.

What you see is what you get(WYSIWYG): That there exists a set of measurements $f : P \to \hat{P}$ with distortion at most ϵ.

Structural Bias: Given a partitioning of X, Y into $\mathcal{X} = (X_1, \ldots, X_k)$ and $\mathcal{Y} = (Y_1, \ldots, Y_k)$, the group skew between $\mathcal{X}$ and $\mathcal{Y}$ is defined as

$$\sigma(\mathcal{X}, \mathcal{Y}) = \frac{GW(\mathcal{X}, \mathcal{Y})}{\binom{k}{2} \frac{1}{k} \sum_{i=1}^{k} GW(X_i, Y_i)}.$$

This, at a rough level, measures how much the distances between groups change relative to the distances within groups change.

If the group skew between construct space and observation space is large, this suggests that the set of measurements being done are biased against some groups. The paper then presents the notion of t-structural bias to be a pairing of construct space and observation space which has group skew larger than t.

These are different assumptions about the relationship between construct space and observation space, the former assuming all distances are approximately preserved, while the latter property measures the extent to which a partitioning of points into groups has different properties within groups and between groups.

When designing a decisionmaking process, one goal might be to avoid having large group skew between observation space and decision space: if there is large group skew between them, this suggests the groups are receiving quite different treatment by $\hat{f}$, the decision rule.

Definition 6. *Consider the observation space $(\hat{P}, d_{\hat{P}})$ and decision space (O, d_O). If*

$$\sigma(\hat{\mathcal{P}}, \mathcal{O}) > t$$

then we say these spaces (along with the function mapping between them) exhibit t-direct discrimination.

While one goal might be to minimize the amount of direct discrimination present in some decision rule between observation and decision space, a more demanding goal is to ask that the group skew between concept space and decision space be small, as t-non discrimination does.

Definition 7. *Consider the construct space* $(P, d_{\hat{P}})$ *and decision space* (O, d_O)*. If*

$$\sigma(\mathcal{P}, \mathcal{O}) < t$$

then we say these spaces (along with the function mapping between them) are t*-nondiscriminatory.*

Finally, the extent to which these goals ought apply to a given system may depend on the extent to which one believes that the "We are all equal (WAE)" axiom holds, which measures the distance between groups in construct space: $(W_{d_P}(X_i, X_j) \leq \epsilon)$ for all $X_i, X_j \in \mathcal{P}$.