ISYE 8813: Algorithmic Foundations of Ethical ML **Fall 2021**

Lectures 14: Formalizing ML Assumptions and Fairness

Lecturer: Juba Ziani

Note that the lecture notes below are based on [FSV16].

What does it mean for an algorithm to be "fair"? To even start to answer the question, one needs to examine the assumptions that we make about fairness in society. To do so, we are going to start by formalizing the different spaces that a typical machine learning algorithm is working with.

1 Spaces: Construct, Observed and Decision

At a high-level, what machine learning does is the following: it takes a data-set as an input, tries to learn properties of this dataset and use these properties to make decisions on future data points. I.e., we take inputs from some *feature space* and return outputs in a *decision space*. For example, if you want to predict whether someone is likely to re-pay their loans, you can take as an input a dataset that contains features about individuals' family and financial situations, and about whether they re-payed their loans in the past. You then train a machine learning algorithm that tries to find patterns in this data, and to see how certain features correlate with the desired target variable (whether someone is likely to repay their loan).

The data and features we work with Defining the feature space is not a trivial problem, and is often subjective as we saw before. For example, look at college admission decisions. Admission offices need to determine what aspects of a person they want to use in their admission decision. This might include potential, intelligence, technical ability, diligence.

But one issue is that it is not even clear in practice that such attributes can be observed directly. Generally, we do not access features such as ability or potential directly; we measure them, for example through standardized testing. This is where an important distinction arises: there are the "true" features that we want to work with, and the features we observe that are often just a proxy for these true features. A machine learning algorithm will work with *observed* features, which are usually an imperfect representation of or way to measure the desired true features.

To deal with this distinction, **CITE** introduces two distinct notions of feature spaces that a machine learning algorithm may care about. The first one is the construct space, that contains the true features we want to make a decision on:

Definition 1 (Construct Space)**.** *The construct space is a metric space $\mathcal{CS} = (P, d_P)$ consisting of individuals and a distance between them. The distance d_P is a measure of closeness of different vectors of features in the construct space.*

This is the space containing the features that we would like to make a decision based on. These are the "desired" or "true" features at the time chosen for the decision, and the ability to accurately measure similarity between people with respect to the task. Instead, we will only be able to work with the features we observe. These features come from a (generally imperfect, noisy) process that maps "real" features to observations:

Definition 2 (Observed space)**.** *The observed space is a metric space* $\mathcal{OS} = (\hat{P}, d_{\hat{P}})$*. We assume an* observation *process* $g : P \to \hat{P}$ *that generates an entity* $\hat{p} = g(p)$ *from a person* $p \in \mathcal{CS}$*.*

The decision process Finally, the whole point is that we want to feed our observed data into a machine learning algorithm, that is going to make *decisions* based on that data (what credit score to assign to person X? Should we admit student Y? Should we release Z on bail?) The *decision space* is going to be the space of possible decisions we can make on someone.

Definition 3 (Decision Space)**.** *A decision space is a metric space* $\mathcal{DS} = (O, d_O)$*, where* O *is a space of outcomes and* d_O *is a metric defined on* O*. A task* T *can be viewed as the process of finding a map from* $\hat{P}$ *to* O*.*

Some common decision spaces:

- $\{0, 1\}$ or $\{-, +\}$. Here, the goal is to assign a *binary* outcome or decision to each data point. I.e., whether we will give a loan or not to a specific individual. This is known as *classification*.
- $[0, 1]$. Here the goal is to assign a score (renormalized between 0 and 1) to rate the "quality" of a data point or agent. For example, rather than a $0-1$ decision on whether to give someone a loan, this could be a credit score. This could also be an estimate of the probability that that person repays their loan. In this case, we are in a *regression* setting.
- etc.

How the spaces interact. Algorithmic decision-making is a set of mappings between the three spaces defined above:

- The desired outcome is a mapping from $\mathcal{CS}$ to $\mathcal{DS}$ via an unknown and complex function $o = f(X_1, X_2, \ldots)$ of features that lie in the construct space. This is the relationship between true features and outcomes.
- In order to implement an algorithm that predicts the desired outcome, we must first extract usable data from $\mathcal{CS}$: this is a collection of mappings from $\mathcal{CS}$ to $\mathcal{OS}$. The features $Y_1, Y_2, \ldots, Y_\ell$ in $\mathcal{OS}$ might be:
  - noisy variants of the $X_i : Y_i = g(X_i)$ where $g(\cdot)$ is some stochastic function,

  - some unknown (and noisy) combination of $X_i : Y_i = g(X_{i_1}, X_{i_2}, \ldots)$, or
  - new attributes that are actually independent of any of the X_i.

  Further, some of the X_i might even be omitted entirely when generating Y_i.

- Our goal is (ideally) to determine o. We instead design an algorithm that learns $\tilde{o} = \tilde{f}(Y_1, Y_2, \ldots Y_\ell)$ i.e a mapping from $\mathcal{OS}$ to $\mathcal{DS}$. The hope is that $\tilde{o} \simeq o$.

2 Formalizing Fairness

The first, formal definition of fairness we will see here captures the following intuition: "elements" that are similar should be treated roughly the same way. Generally, many of the fairness metrics that we will see through the semester will have this flavor:

- When we will talk about fairness at the individual level, we will take each element to be an individual and require that similar individuals must be treated similarly.
- When talking about fairness at the level of group/sub-populations, we will take each element to be a group, and will aim to guarantee that groups will be treated similarly (we will see what this means later on).

We will formally write this notion of fairness using a notion of "distortion". At a high-level, distortion measures how much a mapping f from a space X to a space Y distorts this space. When the distortion is small, it means that if two individuals $p, q \in X$ are close in terms of features in X, $f(p)$ and $f(q)$ should also be close in Y: this intuitively aligns with our definition of fairness that we should make similar decisions on similar individuals. On the other hand, high distortion means that p, q can be close but $f(p), f(q)$ far from each other (or vice-versa).

Here, I am talking about an individual notion of distortion, but we could also imagine defining this distortion across different socio-economic groups rather than individuals. At the high-level, the idea is to have a metric $\rho(\mathcal{X}, \mathcal{Y})$ that measures the distortion between these metric spaces. One example can be the following:

Definition 4 (Additive Distortion)**.** *Let (X, d_X) and (Y, d_Y) be two metric spaces and let $f : X \to Y$ be a map from X to Y. The* distortion ρ_f *of f is defined as the smallest value such that for all $p, q \in X$*

$$|d_X(p, q) - d_Y(f(p), f(q))| \leq \rho_f$$

The distortion $\rho(X, Y)$ is then the minimum achievable distortion ρ_f over all mappings f.

CITE also looks at more general notions of distortion, in particular some that can be applied to groups rather than just individuals. We can then write our notion of fairness or *non-discrimination* as follows:

Definition 5 (Non-Discrimination)**.** *Let $\mathcal{CS} = (X, d_X)$ and $\mathcal{DS} = (Y, d_Y)$. A mapping $f : \mathcal{CS} \to \mathcal{DS}$ is t-nondiscriminatory if the group skew $\rho(\mathcal{X}, \mathcal{Y}) \leq t$.*

This allows for a more fine-grained approach to fairness (compared to the original) definition in which we "control" the level of unfairness we provide. This is where the parameter t comes in

3 Assumptions we make and how they affect fairness

The difficulty here is that we want the mapping from $\mathcal{CS}$ to $\mathcal{DS}$ to be fair, but we only observe features in $\mathcal{OS}$. Unfairness can then arise at several levels, depending on what assumptions we make about the world, and in particular about how the construct space maps to the observed space. **CITE** defines assumptions that we can make about this process.

Assumptions on the mapping from $\mathcal{CS}$ to $\mathcal{OS}$ There are different assumptions we can make on the mapping between the construct space and observed space by trying to characterize whether the construct space and observed space are *essentially the same*, or whether there is distortion between these spaces.

Assumption 6 (What You See Is What You Get (WYSIWYG))**.** *There exists a mapping $f : \mathcal{CS} \to \mathcal{OS}$ such that the distortion ρ_f is at most ϵ for some small $\epsilon > 0$. Or equivalently, the distortion ρ between $\mathcal{CS}$ and $\mathcal{OS}$ is at most ϵ.*

In practice, we can think of ϵ as a very small number. In the college admissions setting, the WYSIWYG is the assumption that features like SAT scores and high-school GPA (which are observed) correlate well with the applicant's ability to succeed (a property of the construct space). More precisely, it assumes that there is some way to use a combination of these scores to correctly compare true applicant ability.

In practice, we often believe this assumption not to be true. In many real-world societal applications, the noise in the transformation from construct to observed space is non-uniform in a societally biased way. This is what we refer to as *structural bias*.

We represent groups as a partition of individuals into sets $G_1, G_2, \ldots, G_k$. Structural bias manifests itself in unequal treatment of groups. In order to quantify this notion, [FSV16] defines a notion of *group skew*: the way in which group structure might be distorted between spaces. Structural bias can then be defined as follow:

Definition 7 (Structural Bias)**.** *The metric spaces $\mathcal{CS} = (P, d_P)$ and $\mathcal{OS} = (\hat{P}, d_{\hat{P}})$ admit* t-structural bias *if the group skew $\sigma(\mathcal{X}, \mathcal{Y}) > t$.*

For example, researchers have shown that the SAT verbal questions function differently for the African-American subgroup, so that the validity of the results as a measure of ability are in question for this subgroup. In the case where SAT scores are a feature in the observed space, this research indicates that we should consider these scores to be the result of structural bias.

How do these assumptions affect fairness? One "naive" way to think about fairness is to take the construct space totally out of the equation, and worry about what we call *direct* discrimination: i.e., even if two individuals have similar (or the same) *observed* features, we make different decisions on them. This is called *direct* discrimination (or often "disparate treatment") because the decision rule/algorithm itself that we use decides to treat people that look similar differently.

Definition 8 (Direct Discrimination)**.** *The metric spaces* $\mathcal{OS} = (X, d_X)$ *and* $\mathcal{DS} = (Y, d_Y)$ *admit* t-direct discrimination *with respect to decision rule* $\tilde{f}$ *if the distortion satisfies* $\rho(\mathcal{X}, \mathcal{Y}) > t$.

Whether we should prevent or aim for direct discrimination is actually a function of the assumptions we make and the world view we take:

- In our first view of the world (WYSIWYG), one can see pretty easily that preventing direct discrimination is sufficient to obtain non-discrimination. Indeed, in this case, the distortion comes entirely from the mapping from observed to decision spaces, and it suffices to make this distortion small. This is exactly what direct discrimination covers.

- In the second view of the world in which structural bias is present, preventing direct discrimination now can be insufficient. In particular, the mapping $\tilde{f}$ should aim to correct for the structural bias in order to obtain fair decisions. We may want, for example, to make similar decisions on individuals that look different in observed space, just because they actually look similar in the construct space. Back to the college admissions example, the idea is that if we do *not* allow direct discrimination, the fact that the same SAT score has different meaning for different groups and that we failed to account for structural bias will lead to (indirect) discrimination.

References

[FSV16] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.