Math 547: Final Report—Algorithmic Fairness

Due on Dec. 7, 2019

Instructor: Dr. Steven Heilman

Gregory Faletto

1 Notation, Problem Statement, and Motivation

Algorithmic fairness concerns the properties of predictive models. A practitioner observes a random sample of n observational units $(X^{(i)}, A^{(i)}, Y^{(i)}) \in \mathbb{R}^p \times \mathcal{Y}$, $i \in \{1, \dots, n\}$. $A = (A^{(1)}, \dots, A^{(n)})^T \in \mathbb{R}^{n \times p_A}$ is a set of p_A *protected* or *sensitive* attributes of each individual, such as race or gender. $X = (X^{(1)}, \dots, X^{(n)})^T \in \mathbb{R}^{n \times p_X}$ are the remaining $p_X = p - p_A$ unprotected attributes of each individual, and $Y = (Y^{(1)}, \dots, Y^{(n)})^T \in \mathcal{Y}^n$ is a set of observed outcomes. In this article, I will assume $Y \in \{0, 1\}^n$ (that is, the model is a binary classifier), but other possibilities exist, like $Y \in \mathbb{R}^n$. Further, for simplicity I will assume $A \in \{0, 1\}^n$. Sometimes I will distinguish between the observed outcome Y and the actual outcome of interest Y' which may not be observed (see Example 1).

The practitioner hopes to fit a model $\hat{f}$ trained on the observed data (X, A, Y) to generate predictions $\hat{Y} = \hat{f}(X, A)$. In particular, the practitioner hopes this model will generate accurate predictions on future data $(\tilde{X}, \tilde{A})$ such that $\tilde{Y} \approx \hat{f}(\tilde{X}, \tilde{A})$, where $\tilde{Y}$ are the eventual observed outcomes (only $\tilde{X}$ and $\tilde{A}$ are observed at the time the prediction $\hat{f}(\tilde{X}, \tilde{A})$ is generated). In this setting, researchers in *algorithmic fairness* hope to (a) define notions of fairness that, when satisfied, result in individuals being treated fairly by the model $\hat{f}$, (b) develop methods to test whether an existing algorithm satisfies these notions of fairness, and (c) develop procedures to create models that satisfy algorithmic fairness at a minimal loss to other considerations, like computational resources and prediction loss.

In other areas of statistics, there exist principles such as *unbiasedness*, *mininum variance*, *invariance*, and so on that are desirable for estimators or methods. Unfortunately, for fairness no such first principle is immediately clear, and much of the work of algorithmic fairness is in finding a sensible first principle for fairness.

1.1 Motivation

It is worth asking why algorithmic fairness is important in and of itself, if only to better understand what problem algorithmic fairness strives to solve. If models generate predictions that are optimized to yield the lowest possible error, one may wonder what is "fair" at all about intervening in that process with some kind of external standard.

One reason why algorithmic fairness may be necessary is that we may have good reason to believe that Y' is independent of A but Y is not; that is, something is wrong with the way that Y is observed that leads to inequitable treatment across protected groups in models trained on Y.

Example 1. Suppose Y contains indicator variables for whether individuals are arrested or convicted of a crime. The actual response of interest might be Y', whether someone commits a crime. Not everyone who commits a crime is caught, and not everyone who is arrested actually committed a crime, so $Y \neq Y'$. Further, because different races have different levels of contact with law enforcement all else equal, the distribution of Y differs from Y' across race. Y can be reliably observed, but unfortunately Y' cannot. A model

trained on Y to with the goal of predicting Y' may therefore be biased with respect to race unless some notion of fairness is enforced.

Another reason algorithmic fairness can be important is if Y' does correlate with A, but in a way that decision-makers agree is unfair and would like to compensate for. One way this can happen is that a model could rely on a proxy for a lurking feature, and the proxy is associated with A even though A does not cause Y'.

Example 2 (from Kusner et al. [2017])**.** Suppose a car insurance company develops a model to predict the probability a prospective customer will get in a car accident in the next year in order to set rates. Let $Y \in \{0, 1\}$ be an indicator variable for the customer getting in an accident (and suppose for this example $Y = Y'$). Suppose Y is entirely determined by a latent variable U corresponding to *aggression*.

The car insurance company would like to measure U and train a model from it if U were not latent. Since this is not possible, the car insurance company relies on X, whether the customer drives a red car, which can be observed. U is associated with X—more aggressive drivers are more likely to drive a red car—so this information can be used to learn about the latent aggression U of the driver. However, it turns out that X is also associated with A, the religion of the driver. For some reason unrelated to aggression, members of Religion 1 are more likely to drive a red car than members of Religion 2. (The situation is shown in a causal directed acyclic graph in Figure 1.) In this example, A is associated with Y through X and U even though there is no causal relationship between A and Y.

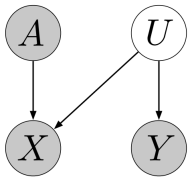

Figure 1: Causal graph of Example 2. Even though A does not cause Y, A is correlated with Y through X and U. Figure (and example) from from Kusner et al. [2017].

Even though A does not cause $Y = Y'$ (in particular, A is independent of Y conditional on X which also does not cause Y), an algorithm that uses X may discriminate with respect to A because the proxy X for U is associated with A.

2 Common Notions of Fairness

Definition 1 (Fairness through unawareness)**.** *Fairness through unawareness* is the requirement that $\hat{f}$ does not use the information from A; that is, $\hat{f} = \hat{f}(X)$.

Fairness through unawareness is a simple and appealing notion of algorithmic fairness. Proposition 1 of Kilbertus et al. [2017] proves that under some assumptions this criterion is satisfactory. However, it turns out not to be useful most of the time. For example, Lemma 3 of Kusner et al. [2017] proves that in the setting of Example 2 if the relationships between variables connected by arrows are linear, an ordinary least squares model fit using both X and A satisfies several notions of fairness (e.g. counterfactual fairness and demographic parity) while a model fit using only X violates these notions of fairness. The details are in Kusner et al. [2017], but the intuition is that including A allows for the linear model to control for different baseline levels of X in different religious groups.

Definition 2 (Demographic parity)**.** *Demographic parity* is the requirement that

$$\mathbb{P}(\hat{Y}^{(1)} = \hat{y} \mid A^{(1)} = 0) = \mathbb{P}(\hat{Y}^{(1)} = \hat{y} \mid A^{(1)} = 1), \qquad \forall \hat{y} \in \{0, 1\}, \tag{1}$$

where the probability is taken over the distribution of the random sample of the observed data and any randomness used to create the model.

Demographic parity can be thought of as analogous to affirmative action: the distribution of $\hat{Y}$ must be identical in each protected group. This notion of fairness is intuitive, and unlike several remaining notions of fairness does not require observing Y, which is often not possible (for instance, we cannot observe whether a student who was not admitted to a college would have graduated that college in under five years had they been admitted).

However, demographic parity can result in significantly different treatment of people with identical X attributes if the distribution of X differs significantly across protected groups. This could make sense if it is believed that Y' is independent of A so any difference in distributions of X across A are irrelevant (as in Example 2), or these differences are themselves reflective of discrimination decision-makers would like to mitigate or compensate for.

Definition 3 (Equalized odds)**.** *Equalized odds* is the requirement that

$$\mathbb{P}(\hat{Y}^{(1)} = \hat{y} \mid Y^{(1)} = y, A^{(1)} = 0) = \mathbb{P}(\hat{Y}^{(1)} = \hat{y} \mid Y^{(1)} = y, A^{(1)} = 1), \qquad \forall \hat{y}, y \in \{0, 1\}. \tag{2}$$

In other words, equalized odds requires that $\hat{Y}$ is independent of A conditional on Y. This could be a good notion of fairness if it is believed that Y' is independent of A conditional on Y (that is, any measurement error of Y is the same across all protected groups); in particular, it makes sense in the special case $Y' = Y$. Unlike demographic parity, equalized odds allows differentiation between groups based on their observed response Y. But this requires observing Y, and this correction may do more harm than good if Y is a biased measurement of Y' with respect to A, as in Example 1.

Definition 4 (Calibration)**.** *Calibration* is the requirement that

$$\mathbb{P}(Y^{(1)} = y \mid \hat{Y}^{(1)} = \hat{y}, A^{(1)} = 0) = \mathbb{P}(Y^{(1)} = y \mid \hat{Y}^{(1)} = \hat{y}, A^{(1)} = 1), \qquad \forall \hat{y}, y \in \{0, 1\}. \quad (3)$$

Calibration reverses equalized odds; an algorithm satisfies calibration if within the sets of people who received the same prediction $\hat{Y}$, Y is independent of A (Y is independent of A conditional on $\hat{Y}$). This standard is widely used in the educational and psychological testing literature [Chouldechova, 2017]. Like equalized odds, calibration may perpetuate bias if Y is biased relative to Y' with respect to A, but *prima facie* calibration and equalized odds both might seem like reasonable criteria.

Example 3 (Propublica/Northpointe dispute)**.** A company Northpointe created an algorithm COMPAS to predict recidivism rates in arrestees. After an arrest, law enforcement would take in information about the arrestee and run COMPAS to generate a predicted probability the arrestee would commit another crime in the next two years. This prediction was used to set bail amounts.

Angwin et al. [2016] showed empircally that COMPAS violated equalized odds and suggested that this made the algorithm unfair. However, Northpointe showed that COMPAS satisfied calibration and argued this meant it was in fact fair.

It was later proven by Chouldechova [2017] and Kleinberg et al. [2017] (independently around the same time) that no algorithm can simultaneously satisfy both equalized odds and calibration under realistic assumptions.

Theorem 1 (Theorem 1.1, Kleinberg et al. [2017])**.** Suppose an algorithm satisfies both calibration and equalized odds. Then either $Y \perp\!\!\!\perp A$ or the relationship between Y and X within each group in A is deterministic ($\mathbb{P}(Y = y \mid X = x, A = a) \in \{0, 1\}$ for all a, y, x).

3 Theoretical Results

These results raise the question of how to pick which criterion is better in a given situation. Yeom and Tschantz [2018] found a number of useful results illuminating the differences between these criteria. Before discussing the results, I will present one more notion of fairness they introduced.

Definition 5 (Output Disparity Control [Yeom and Tschantz, 2018])**.** *Output disparity control* is the requirement that

$$d_{TV}(\hat{Y}^{(1)}|A^{(1)} = 0, \hat{Y}^{(1)}|A^{(1)} = 1) \leq d_{TV}(Y'^{(1)}|A^{(1)} = 0, Y'^{(1)}|A^{(1)} = 1). \quad (4)$$

where $d_{TV}(X, Y)$ is the total variational distance between random variables X and Y:

$$d_{TV}(X, Y) = \frac{1}{2} \sum_{y \in \{0,1\}} \left| \mathbb{P}(X = y) - \mathbb{P}(Y = y) \right|.$$

The idea is that the distributions of predictions conditional on A should be no further apart than the conditional distributions of Y'. This is a weak notion of fairness; it is agnostic to whether existing disparity is benign or bad, and it doesn't say anything about whether reducing disparity would be fair or not. The idea is that controlling output disparity does not necessarily mean that a model is fair, but amplifying output disparity almost certainly makes the model unfair. With that in mind, the following results shed some light on when different criteria may be best used.

Theorem 2 (From Theorems 1, 5, and 6 in Yeom and Tschantz [2018])**.** Any model that satisfies demographic parity also controls output disparity. Further, suppose $Y' \perp\!\!\!\perp A$. Then every model satisfying $\hat{Y} = Y'$ satisfies demographic parity, and models satisfying equalized odds and calibration may allow output disparity amplification.

Theorem 3 (From Theorems 2, 4, and 6 in Yeom and Tschantz [2018])**.** Suppose $Y = Y'$. Then every model satisfying $\hat{Y} = Y'$ satisfies equalized odds, and models satisfying demographic parity and calibration may allow output disparity amplification.

In light of these results, demographic parity seems best if $Y' \perp\!\!\!\perp A$. This makes sense, since demographic parity enforces $\hat{Y} \perp\!\!\!\perp A$. Equalized odds seems best if $Y = Y'$, which also matches intuition. Perhaps surprisingly, calibration seems like the worst out of these three criteria—it turns out it always allows disparity amplification (Theorem 6).

These results are helpful in selecting which criterion makes sense under which assumptions, but they still do not provide a one-size-fits-all solution. Indeed, Theorem 1 of Kilbertus et al. [2017] suggests that under some circumstances creating a fairness criterion from observational data alone is impossible.

4 Counterfactual Fairness

The above criteria have the advantages of being estimable based only on observed data and model predictions, but none seem to strictly dominate the others. The results from Yeom and Tschantz [2018] suggest that the differences between when each criterion is effective have to do with the nature of the relationship between Y' and both Y and A. *Counterfactual fairness* [Kusner et al., 2017, Kilbertus et al., 2017] is a notion of fairness that relies on explicit models of these relationships. It requires either knowledge of or assumptions on (a) which variables have causal effects on others and (b) the functions that relate them. However, it seems to strictly dominate the other criteria when its assumptions are satisfied.

Definition 6 (Counterfactual fairness [Kusner et al., 2017, Kilbertus et al., 2017])**.** *Counterfactual fairness* is satisfied for a model $\hat{f}(A^{(i)}, U^{(i)})$ if

$$\mathbb{P}(\hat{f}(a, U) = \hat{y} \mid X = x, A = a) = \mathbb{P}(\hat{f}(a', U) = \hat{y} \mid X = x, A = a),$$
$$\forall \hat{y} \in \{0, 1\}^n, x \in \text{supp}(X), a, a' \in \{0, 1\} \quad (5)$$

where $U = (U^{(1)}, \ldots, U^{(n)})^T \in \mathbb{R}^{n \times p_U}$ is the set of all unobserved latent variables that may cause X, A, and Y.

It may seem strange that the conditional values of the variables don't change on each side of (5). This is because we are interested in the probability conditional on the observed variables after *intervening* on the variable of interest; that is, changing the value of a variable by fiat in order to understand the *counterfactual* relationships.

Consider again Example 2. It is clear that if an experimenter intervened to randomly assign drivers to different colors of cars, there would be no relationship between color of car (or religion) and probability of getting in an accident, even though the association holds in the population of observational data. In particular, the intervened distribution of the random function $\{\hat{f}(a', U) \mid X = x, A = a\}$ is different from the *observational* random function $\{\hat{f}(a', U) \mid X = x, A = a'\}$. The notation in (5) reflects that we are interested in how the model might change if we intervened on protected status, conditional on the actual observed data.

Remark 1. Kilbertus et al. [2017] point out that intervening directly on protected attributes of interest (e.g. race and gender) is often impossible. For this reason, they focus on proxies for these protected attributes, such as name or appearance, which may be subject to intervention[1]. They define *proxy discrimination* as a violation of (5) except with interventions on proxies P rather than protected attributes A. They are able to prove some results about when proxy discrimination can be avoided.

This notion of fairness follows the principle that it is unfair to treat individuals differently based on factors outside of their control (as protected attributes often are).

Both a strength and weakness of counterfactual fairness is that it requires either (a) knowledge of or assumptions on the causal graph as well as the functions relating the variables, or (b) the ability to intervene on the variables of interest (in order to make inferences about the causal graph). On the one hand, requiring more assumptions is a drawback. On the other, the results from Section 3 suggest that the relationships between A, Y, and Y' are central to the question of which notion of fairness is best in a given situation. A fairness criterion that requires determining these relationships (or making assumptions on them) may be applicable in all circumstances.

Kusner et al. [2017] prove results about settings in which counterfactual fairness is guaranteed. One way to create a counterfactually fair model is using only features that are

[1]For example, in studies like Bertrand and Mullainathan [2004], the same resume was sent to employers under a name associated with Black Americans and a name associated with white Americans to see if they received different treatment.

non-descendants of A as predictors; then clearly (5) holds. In Example 2, that would leave no predictors, since the only non-descendant of A other than Y is the latent variable U. Another way is to model the latent variables U that are *parents* of observed variables and non-descendants of A and include these in the model as well.

Kusner et al. [2017] propose the FairLearning algorithm to create counterfactually fair models. Given data (X, A, Y), a model form that can be fit with such data, a causal graph, and a prior probability model $\mathcal{M}$ for the latent variables U in (5), the posterior distribution $\mathbb{P}_{\mathcal{M}}(U \mid X, A)$ is estimated via Markov chain Monte Carlo and the augmented data set (X, A, Y, U) is formed. Each element of U corresponds to the conditional posterior expected value given the corresponding elements of X and A. Finally, the model can be trained on U as well as any non-descendants of A in X. The idea is to extract what "fair" information we can out of X while discarding the "unfair" information tainted by A.

5 Empirical Analysis

For this analysis, I analyzed the publicly available data from Angwin et al. [2016] described in Example 3 and created a counterfactually fair model using the FairLearning algorithm. The observational units were arrestees, and the data contained some covariates on the arrestees, the COMPAS prediction of their risk level for recidivating in the two years following arrest, and whether or not they actually recidivated. The sensitive covariates are race (coded in the data in six categories; for simplicity I used a binary indicator variable R equalling 1 when "race" was "hispanic" or "black" and 0 otherwise) and gender (coded in the data as "male" or "female;" I transformed this to an indicator G equalling 1 for "female"). The predictors I used were "felony" F (indicator for whether the crime for which the arrestee was detained was classified as a felony rather than a misdemeanor) and "priors" P (count of the number of prior offenses of the arrestee). Finally, the response Y is a binary indicator variable for whether the arrestee recidivated.

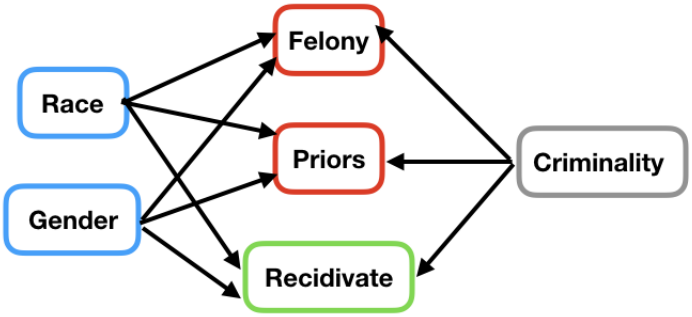

Figure 2: Assumed causal graph for empirical analysis.

I assumed that there exists a latent normally distributed variable "criminality" U, fixed over the course of an individual's life, that determines probability of committing a

crime in any given year. Then I assumed that the variables have the relationships shown in Figure 2: whether someone is arrested and how severely they are charged reflects both criminal behavior and inequitable treatment in the justice system across race and gender. In particular, I assumed the variables have the following relationships in each individual:

$$\begin{aligned} U &\sim \mathcal{N}(0,1), \\ F &\sim \text{Bernoulli}\left(\text{logit}^{-1}\left[\beta_F + \beta_F^U U + \beta_F^R R + \beta_F^G G\right]\right), \\ P &\sim \text{Poisson}\left(\exp\left\{\beta_P + \beta_P^U U + \beta_P^R R + \beta_P^G G\right\}\right), \\ Y &\sim \text{Bernoulli}\left(\text{logit}^{-1}\left[\beta_Y + \beta_Y^U U + \beta_Y^R R + \beta_Y^G G\right]\right), \end{aligned} \tag{6}$$

where $\text{logit}^{-1}(x) := \frac{1}{1+e^{-x}}$ and I modeled the coefficients in a Bayesian framework with standard Gaussian prior distributions on each coefficient.

I used 80% of the data for training and left the rest for testing. I used the training data to estimate the posterior distributions of U as well as the parameters of the models in (6) using the probabilistic programming language Stan (in particular, the R package interface `rstan`)[2]. I used these parameters to estimate U in the test data.

Following the modeling assumptions, I fit the counterfactually fair model

$$\hat{\mathbb{P}}_{\text{fair}}(Y=1) = \text{logit}^{-1}\left(\hat{\beta}_{Y,\text{fair}} + \hat{\beta}^U_{Y,\text{fair}} U\right)$$

along with the unfair full model

$$\hat{\mathbb{P}}_{\text{UF, full}}(Y=1) = \text{logit}^{-1}\left(\hat{\beta}_{Y,\text{UF}} + \hat{\beta}^F_{Y,\text{UF}} F + \hat{\beta}^P_{Y,\text{UF}} P + \hat{\beta}^R_{Y,\text{UF}} R + \hat{\beta}^G_{Y,\text{UF}} G\right)$$

and the fair through unawareness (but counterfactually unfair) model

$$\hat{\mathbb{P}}_{\text{FUA}}(Y=1) = \text{logit}^{-1}\left(\hat{\beta}_{Y,\text{FUA}} + \hat{\beta}^F_{Y,\text{UF}} F + \hat{\beta}^P_{Y,\text{UF}} P\right).$$

Following Kusner et al. [2017], I compared the loss of the fair models compared to the unfair models[3]. The results in Table 1 show that the accuracy for all of the models was very similar, though the log loss for the fair model was greater than that of the other models. This suggests that the fair model was accurate in its predictions but was less confident than the other models, hedging with predictions closer to 0.5. Given that the fair model has fewer degrees of freedom and less than full information, this seems like fairly impressive performance[4].

[2] Kusner et al. [2017] released their code on Github (https://github.com/mkusner/counterfactual-fairness), and I modified their framework for my data set and models.

[3] I used log loss and accuracy as my metrics rather than mean squared error since this is a classification problem. For the accuracy prediction, I interpreted a predicted probability over 0.5 as predicting the arrestee would recidivate, and vice versa.

[4] It is also worth noting that i was not able to get the Markov chains to converge; `rstan` kept warning me about this and suggesting that I use more iterations. The default is 2000 iterations; I used 80,000, which was the most I could do with my computational resources and seemingly still insufficient.

Table 1: Test set prediction losses for models on data from Angwin et al. [2016].

	Unfair (full)	Unfair (unaware)	Fair
Log Loss	783.776	786.122	1,002.566
Accuracy	0.644	0.647	0.646

References

J. Angwin, J. Larson, L. Kirchner, and S. Mattu. Machine bias, Mar 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

M. Bertrand and S. Mullainathan. Are Emily and Greg More Employable than Lakisha and Jamal ? A Field Experiment on Labor Market Discrimination Author (s): Marianne Bertrand and Sendhil Mullainathan Source : The American Economic Review , Vol . 94 , No . 4. *American Economic Review*, 94(4):991–1013, 2004.

A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153—-163, 2017. URL https://arxiv.org/pdf/1610.07524.pdf.

N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, B. Schölkopf, and B. M. De. Avoiding Discrimination through Causal Reasoning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. URL https://arxiv.org/pdf/1706.02744.pdf.

J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *The 8th Innovations in Theoretical Computer Science Conference*, 2017. URL https://arxiv.org/pdf/1609.05807.pdf.

M. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual Fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, pages 4069—-4079. Curran Associates, Inc., 2017. URL https://obamawhitehouse.archives.gov/blog/2016/05/04/big-risks-big-opportunities-intersection-big-data-.

S. Yeom and M. C. Tschantz. Discriminative but Not Discriminatory: A Comparison of Fairness Definitions under Different Worldviews. Technical report, 2018. URL http://arxiv.org/abs/1808.08619.