Jan 27, 2025, Lecture Note: Different Aspects of Responsible AI.

Ranit Debnath Akash, Mohammed Abdul Hadi

`CS 516: Responsible Data Science and Algorithmic Fairness; Spring 2025`
`Abolfazl Asudeh;` www.cs.uic.edu/∼asudeh/teaching/archive/cs516spring25/

1 Recap

Until now, we have discussed how feature selection, data collection, and labeling biases can happen. There are different types of attributes in a data set; some of the features of observations are used for prediction, and there are target variable attributes that are not observable during the inference time, and our goal is to predict them. There are sensitive attributes that specify the grouping of data. For example, sensitive attributes include race, gender, college degree, economic status, etc. Previous lectures have discussed how things can go wrong and introduce bias during data collection, labeling, feature definition, and transformation to some structured features.

2 Proxy attributes

The attributes used as substitutes for true target variables (e.g., GPA as a proxy for student success) or sensitive attributes (e.g., skin color as a proxy for race).

2.1 Target variable dilemma (proxy of target variable)

To learn about the proxy attributes, first, we need to pay attention to the target variable dilemma. Consider predictive policing tasks, where we may not have the appropriate data to perform the tasks. To design predictive policing, we need to know, "When is the crime happening? Who is committing the crime?" We want to know who committed the crime so we can build the dataset for our model. Unfortunately, many people commit crimes and never get caught, so we use only the data related to who committed a crime and got arrested. So, in that way, it is almost impossible to observe the true target variable. It could happen that people of a specific race were caught or arrested more for crime than people of another race. If that happens, instead of observing the true target variable of who committed the crime, you end up observing a representation of the target variable (arrest) that correlates with another feature (in this case, race). So, the arrest record is working as a proxy for the target variable. Another example is finding out which students/program are more successful at a university. But measuring actual success and comparing it is hard, so people go for the proxy of success, e.g., higher GPA, some salary threshold, or jobs at certain companies, etc. So, it can be understood that defining target variables is hard, and most of the time, data scientists have to choose from the recorded data they have. And they end up selecting a proxy of the target variable instead of choosing the actual target variables.

Now, proxy attributes become the problem when treating them as a target variable, and they are also correlated with sensitive attributes. For example, let's say someone looks into the GPA as a proxy for student success. It turns out that GPA, while highly correlated with student success, is also highly correlated with other things, such as gender or some economic status of students. So, it is not the true measure of student success; it is just a proxy that gives you an error when measuring the target variable. It also has a high correlation with sensitive attributes. And from that perspective, it is not only not the true target variable but also biased. So, that is how proxy attributes can be problematic.

2.2 Proxy of sensitive attributes

Proxy attributes, in general, can be any of the attributes. Some of the features can be representative of sensitive attributes. For example, skin color is a representative of a race. So, in general, any feature that is highly correlated with the sensitive attributes and with high probability and accuracy can predict the sensitive attributes is called a proxy for sensitive attributes.

2.3 Fairness through unawareness and its limitations

Let's say a big tech search company wants to create a tool that is not biased. So, they don't use any demographic information (race, gender, etc.) from the user to deliver the results. They respond only to the query typed in the search box without knowing the demographic information. Can they claim that their tool is fair because they are not using any demographic information while building the tool? The answer is no! Because there could be certain sections of people using certain tools in a certain way, and it can not be generalized for everyone, and that could serve as a proxy. So, the answer is that even though sensitive attributes are not directly used, their proxies could be used implicitly for modeling or building the tool. Unless it is proved that no proxy of the sensitive attribute is being used, then the fairness of the model cannot be claimed. And that is also why fairness through unawareness is impossible.

2.3.1 Bank loan example of fairness through unawareness

Consider some data scientists want to design a loan approval software because it is risky to give loan to some specific customer. It is similar to credit card approval software. If your model puts more relative weight on the income of the people to provide the loan, it will deny the loans to low-income people. Now, this model does not use any sensitive attributes. Is the model fair?

But by nature, it is biased. Specifically, it has a gender bias because we know, unfortunately, salaries for men are higher than for women, right? So, if they just use income as an important signal in their decision modeling, their algorithm is already biased! So, fairness through unawareness is not possible because of the proxy of the sensitive attribute is being used in the process of building the tool.

3 Masking and digital redlining

Masking is the process of deliberately obscuring sensitive attributes or their proxies in a model while still achieving discriminatory outcomes. This can involve carefully selecting features or generalizing features to remove detail, effectively hiding the bias.

We previously discussed redlining, which is drawing a line in the map to decide if some people will receive favorable or unfavorable decisions. For example, the insurance companies 100 years ago, let's say they wanted to decide if they wanted to provide insurance to a specific individual or not. They could just they would look into explicitly the race of those individuals and decide not to give the insurance. For example, in Chicago, there are white neighborhoods and black neighborhoods, and people live in certain neighborhoods simply because of their insurance. They couldn't get insurance in certain other regions, so they were forced to move to some regions. This is called **redlining**, which means that you are drawing these lines on the map, that you can live within those regions but not outside those regions. Redlining is illegal, no company can do that now!

But the question is that can you use the data to synthesize redlining while you can justify it? In other words, it's called **digital redlining**. Let's say a bad insurance company doesn't want to give loans to certain demographic groups. Can they play with the data such that the output of their algorithm sounds legit, but it prevents certain groups from receiving loans? The answer is yes! They can carefully handpick some features from the data or can generalize some features to remove some details from the features. Then what they get is a black box algorithm, that takes the input features which doesn't use explicitly sensitive attributes. The output automatically denies loans for some certain demographic groups with significantly higher probabilities, and they try to justify it through fairness and unawareness. However, they are using **masking** to hide the internal formulation of the model.

3.1 Recidivism score and masking

Also, in the recidivism score, we saw that the two distributions (Black vs White) were totally different from each other. But it can be justified by the accuracy; parity across the two groups is the same. One can play with the features to come up with such examples. One can cherry-pick or synthesize data in a way that, for the two groups, it seems that the accuracy is equal in parts, but only because they are hiding some important features. Which means that if they had brought those features into the picture, it would have been clear that they are not even satisfying the performance metrics properly.

3.2 US ranking score and masking

Another example is US ranking scores. Many people look at the US ranking when applying for top schools, and it is very important for the schools to be in the top 10 or top 20 of the ranking. Every year, they change the raking criteria by some measures by justifying that they are doing that with a committee of experts. They consider many things, like publications of schools and departments, their undergrad student success, GPA, placement in the big companies, the average score of the GRE people applying to those schools, how much assets they have, etc. But they never publish the actual weights of those scoring. So, it can happen that they adjust the weight of certain criteria (e.g., 0.1 to 0.123) in such a way that some schools remain in the top 20. It can also happen that adjustments of the weight are intentional (or bribed). So, their ranking function might or might not be using masking. How can we tell if US ranking is masking their ranking function or not? So, explicitly, it can not be directly said by looking at the data, but implicitly, you can guess about it. How can you fix the problem? Standardizing the weights, or by a group of experts what they are saying they are using? One answer to this is using a stable ranking algorithm. If we can get a lot of rankings, and assume that every ranking gives some information, if you look into the distribution of possible rankings. If a ranking is near the mean of that distribution, then it is reliable, but if a ranking is far from the mean, you can say that it's not reliable or something like that.

3.3 How to detect the evidence of masking?

Instead of the role of the data scientist who wants to create an algorithm that is fair, let's say here you are in the role of an audit company or audit person. How can you audit and detect cases of masking? Unfortunately, masking is very complicated to detect and prove. Also, It is an understudied problem, very few number of research papers studied this problem. It must be highlighted that masking is not very well researched, despite its importance, because it's a very hard problem. It's very difficult to detect if masking is happening, and even if you detect/guess it, it's even harder to prove that masking is happening. But again, this is a very well-motivated problem and extremely understudied, which means that it needs a lot of work if anyone is interested. It is one of the exciting areas.

So, we learned from the perspective of data how things can go wrong. During data collection, one can introduce sampling bias and measurement bias. Also, bias can be introduced during labeling, standardizing the raw data into standard features. Furthermore, developed algorithms can be biased, selecting the features or proxy attributes can cause issues, and masking can be another issue.

4 Legal and ethical frameworks: Disparate treatment and Disparate impact

Let us introduce two very important laws. These rules in the United States give us basically the direction toward what is not allowed in data-driven decisions. We previously learned that societal norms matter. Two of the societal norms we will learn about, which are also the most important ones, are disparate impact and disparate treatment.

4.1 Disparate treatment

Disparate treatment says it is illegal to *explicitly* use grouping information for decision-making (pay attention to the word *explicit*, so it doesn't mean *implicitly* you aren't allowed to use the grouping). It means that the group membership (gender, race, etc.) of individuals should not be used as an input feature to the model, and it is a rule! There could be exceptions, for example, we talked about health care and there are models that can benefit from that information to draw better predictions with those membership information.

For example, during grading students in this class, if someone separates the students based on gender or race or any other way, i.e., boys and girls will be graded two different ways. So clearly, this is wrong! That is what disparate treatment is saying that when you are making a decision, you shouldn't use the group membership of people in your decision. Unfortunately, a very large portion of the papers published in algorithmic fairness make the same basic mistake when they want to achieve fairness; then, they use race or any other sensitive attributes as one of the inputs to the models. They assume that they have access to that, and then, based on that, they draw the conclusion that it violates the spirit. But again, it doesn't mean one should never do that. Remember, it's for decision-making, not for prediction. So, for example, if you're making healthcare predictions, sometimes it makes sense to use sensitive attributes as part of the model input.

4.2 Disparate impact

Disparate impact is talking about the implicit impact, that it doesn't matter if you're using the group information as the input or not, but if the outcome that you are generating is different for different groups, we can say that we have a disparate impact. So disparate impact is looking at the output, and disparate treatment is looking at the input. Disparate impact says that you shouldn't have different distribution of outcomes for different groups. In the context of classification, whatever your class prediction is, it should be independent of the sensitive attribute. One of the examples of disparate impact that has been extensively discussed is the example of recidivism scores. In the case of the recidivism score, it wasn't right because the distribution of scores for blacks was different from the distribution score for whites.

$$h_\theta(x) \perp\!\!\!\perp S$$

Here x is the input features, $h_\theta(x)$ is the prediction function, and S is the sensitive attributes. The prediction should be independent of the sensitive attributes according to the disparate impact. On the left-hand side, we have the prediction. On the right-hand side, we have the sensitive attributes. It has nothing to do with the true label, accuracy, false positive, false negative. Whatever prediction it is making, if you look at that, you consider the random variable. That random variable should be independent of the sensitive attribute.

5 How to prevent disparate impact or ratio practically?

How can we ensure that the algorithms that we're going to design in the end are not problematic? Given a dataset with m features for each input X, target variable y , and set of sensitive attributes S, what type of observation features should we select to create a decision-making model?

$$x = (x_1, x_2, x_3, ..., x_m)$$

We want the correlation of x and y to be high because we want high accuracy, so we look for features that highly correlate with the target variable. Also, we want a set of features for predicting the target variable that does not correlate with the sensitive attributes. And at the same time, we want the correlation of x and s to be lower. Considering salary as an important feature is not a good idea because salary (with a combination of some other features) can be a proxy for gender. We have to ensure that sensitive attributes are not predictable by the features and, at the same time, your target variable is predictable by the features. But that is almost impossible in practice, so it means that in practice we have to optimize that. We have have to minimize the correlation between x and s and maximize the correlation between x and y.

$$\min(Corr(x, s)) \wedge \max(Corr(x, y))$$

If we can do this it means that we are minimizing bias in the data. And when the bias in the data is minimized the unfairness issues are minimized. So, in this way, we cam achieve close to zero predictability of s and maximum predictability of y in practice. The next steps are to carefully and responsibly curate data, design algorithms to break the cycle of unfairness and start a positive cycle of fairness.

6 Formalizing Fairness, Individual vs. Group Fairness

Our definition of fairness can be very different from that of social scientists. Social scientists looks into some phenomenon and explains with some examples why that is unfair. In the case of recidivism scores, Propublica, Northpoint, and Court each had their own justification for why the software was unfair or not. In general, when can we call an outcome fair? In summary, the answer is not straightforward, but if some outcome makes most people happy, can we call it a fair outcome? But, we want to formalize the problem and decide how to measure the goodness of a model. If we have a predictive model or algorithm, and if the outcome is conditioned on a specific group versus if you don't condition it on that specific group, it is the same, and the outcome doesn't depend on that sensitive attributes, is that fair? So, we can see that we need to use a metric to define fairness. We want the performance of your algorithm or model to be independent from the group membership.

A vague answer is **a model is fair if the performance of the model is equally good for all.** But it has three parts, and if we change one of those, we get different definitions based on that. Also, we need to define, a) what is performance? b) what is good? c) who is this all? All other definitions of fairness fall under this definition.

If we replace *"all"* with group memberships, then we want equally good performance across different groups.

$$\text{Perf}^A \mid g_i \approx \text{Perf}^A$$

where, Perf^A is performance of Algorithm A. And it is saying that performance of algorithm A should be similar when considering their demographic group or not.

For all in the definition above, we can have 3 branches:

- individual level.
- group level
- subgroup level,

6.1 Individual fairness

The definition of individual fairness is performance metric independent (which means that you put whatever performance metric you want, and it holds). Individual fairness says, **similar individuals should be treated similarly.** For example, the grades are between 0 and 100 during grading, but in the end, students have to be assigned grades among the A, B, C, D, and F. Then, individual fairness is satisfied if two students whose grades are similar should receive a similar grade letter, i.e., if one is receiving 95, the other student is getting 94; the two should receive the same grade. If one gets an A, the other one should also get an A.

In a more general perspective, in the context of classification and if you have two input with similar features i.e. two vectors of observation x_1 and x_2, are similar to each other, the model output should be similar for them. This means that for recidivism scores if one of them is predicted as high risk, the other one should also be predicted as high risk. Let's say that you have two individuals, t_1 and t_2; we have a distance metric *dist* that measures the similarity or distance between two individuals, i.e., how dissimilar two individuals are. Because no deterministic algorithm can guarantee individual fairness (more on that later), it looks at the outcomes from a probabilistic perspective. So, we have an output probability distribution O from which we are getting the outcome for individuals.

$$dist(t_1, t_2) \geq \mathcal{E}\Delta(O(t_1), O(t_2))$$

Here $\mathcal{E}$ is the normalization factor, and Δ is the output distance metric for calculating the distance between probability distribution. The distance between two probability distributions can be calculated using techniques like KL-Divergence. So it says the distance between t_1 and t_2 shouldn't be smaller than the distance between the outcomes. It means that two individuals are similar to each other since their distance is small; the outcome distance should also be small. This definition comes from the paper [1]. Based on this definition, an interesting direction is that if the Δ is equal to zero for everything, this condition is satisfied; it is fair! If the model has the same output to anybody, irrespective of their input, it is fair both at the individual and the group level. But again, it raises the question that the model is not looking at the qualifications of the people. So let's say that the model is supposed to predict if somebody is risky or not, and it predicts everyone is not risky; then it's also a useless model, isn't it? The assumption is that the entropy of outcome has to be not zero to make the model useful.

6.2 Impossibility of individual fairness vs Randomized algorithms

One of the ways to achieve individual fairness is to use randomized algorithmic decisions, but randomized decisions are not socially accepted. For example, in terms grading, if you get 95, the probability of you receiving A is 95%, but there is still a 5% possibility that even if you receive a 95, you're going to get a B. It is clearly fair! It is also possible that somebody with 97 gets a B and somebody with 30 gets an A. That's why, even though it is theoretically the only answer to achieving individual fairness, It's not socially accepted. Fairness at the individual level is impossible to achieve socially because you need a randomized algorithm, and randomized algorithms are not always socially accepted. That is a major issue we have with individual fairness! In a perfect world, we can achieve individual fairness. But unfortunately, since it's not possible in the real world, we're going to relax it from individual level to group level. We want to be fair for the group of individuals on average at least, even though we can not be fair at individual level.

References

[1] Cynthia Dwork et al. "Fairness through awareness". In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ITCS '12. Cambridge, Massachusetts: Association for Computing Machinery, 2012, pp. 214–226. ISBN: 9781450311151. DOI: 10.1145/2090236.2090255. URL: https://doi.org/10.1145/2090236.2090255.