Thesis Report: Examining Individual Unfairness in Machine Learning

Ravi Kothari[1], Pranav Ragupathy[2], Swagam Dasgupta[3], Dibyendu Misra[4]

[1,2,3,4] Department of Computer Science, Ashoka University
Sonipat, Haryana, India

Abstract: We aim to create a methodology to identify individual unfairness after the classification process is complete. Our approach involves the development of a distance metric to quantify the similarity between two individuals as well as the metric to quantify the conditions in which an individual is treated unfairly. This is in accordance with our definition of unfairness which is as follows: An individual is treated unfairly when a significant proportion of other similar individuals are categorized differently. This semester we explore the different approaches to fairness in order to create a novel definition that is not based on protecting sensitive attributes using our learning we aim to optimize the distance metric in order to create a general framework to identify unfairness and proceed with our project.

1. Introduction

Traditional notions of bias have usually been confined to the disciplines of social choice theory, game theory, economics and law; but with the advent of machine learning being adopted across multiple industries, it becomes integral for us to question the underlying ethics behind this technology. In this light, there have been significant contributions to the field of fairness in machine learning that has borrowed definitions and methodologies from other disciplines — primarily, the humanities[1]. These methodologies are primarily based on rectifying unfairness during classification. Additionally, these definitions of fairness mostly deal with deriving methods that can protect sensitive attributes.

In contrast to previous approaches, we aim to develop a strategy to identify individual unfairness post the classification process. This is done by examining the feature vectors present in the input space. Section 2 explores the different definitions of bias, their advantages and shortcomings while Section 3 summarizes our goals and contributions. Section 4 breaks down the definition of fairness in order for us to measure it effectively. Additionally, Sections 5 and 6 describe our experiment and results. Finally, in Section 7 we acknowledge our limitations and propose a methodology for the coming semester.

[1] Borrowed from Theory of Justice by John Rawls. (Dwork et. al.)

2. Related Work

As discussed above, previous approaches to fairness are broadly based on protecting sensitive attributes such as race, ethnicity, sex, caste etc. One of the earliest methods required the machine to be unaware of these sensitive attributes (Hardt et. al.). In technical terms, these attributes are removed as features in the training dataset. This method, known as fairness through unawareness, is consistent with the theory of disparate treatment (Barocas et. al.) which proposes the exclusion as a means of protecting certain features. Although it seems to be an intuitive approach, the flaw resides in the fact that removing these attributes might not lead to a reduction in bias. This is due to the presence of features that are highly correlated to the sensitive attributes, i.e, proxies. For example, a feature that specifies the neighbourhood a person resides in may be significantly correlated with her race.

Building on the notion of sensitive attributes, there exists another common conception of fairness — demographic or statistical parity. This requires a classification to be independent of the sensitive attribute. In the case of a binary classification $Y_b \in \{0,1\}$ and a binary protected attribute $A \in \{0,1\}$, this constraint can be described as $\Pr\{Y_b = 1 \mid A = 0\} = \Pr\{Y_b = 1 \mid A = 1\}$. In other words, membership in a protected class should have no correlation with the decision. Unfortunately, this notion does not ensure fairness since the classifier will enforce decisions such that percentages of acceptance are equal. Therefore, it could potentially accept qualified applicants in the demographic $A = 0$, but unqualified individuals in $A = 1$, as long as the acceptance rates are similar (Dwork et. al.). Additionally, this approach disregards any meaningful correlation between the protected attribute and the predictor.

As a response to the flaws of demographic parity, Hard et. al. introduced two notions of fairness — equalized odds and equal opportunity. The former consists of a predictor Y_b that satisfies equalized odds with respect to protected attribute A and outcome Y, if Y_b and A are independent conditional on Y. Equalized odds allows the predictor to depend on A but only through the target variable Y, unlike demographic parity. In addition, equal opportunity offers a relatively weaker approach to enforce fairness. It imposes an equal acceptance rate only if the individual is qualified, i.e, $Y = 1$. Even though equal opportunity and equalized odds make up for the lacunae in demographic parity, they fail to satisfy calibration unless the groups have identical base rates, i.e. rates of positive outcomes. This can be best explained by the following example:

Imagine the hypothetical scenario in which a company wishes to hire 30 applicants out of 200. These applicants are divided into two groups — A and B. Group A and B consist of 58 and 2 qualified applicants respectively. If we implement the equal opportunity approach, 29 people from group A would receive the job, while only 1 applicant from group A will be offered the same. Therefore, the gap between the two groups cannot decrease with this approach. Additionally, the *unidentifiability* result of Hard et. al. describes that observational criteria are unable to distinguish two intuitively very different scenarios.

A causal reasoning methodology was proposed in order to combat the problems of observational criteria. This framework introduces the resolving variable and proxy variable that are used to define the causal discrimination criteria (Kilbertus et al.). Furthermore, the approach attempts to remove proxy discrimination by constructing structural equations of the causal model. Although fairness through causal reasoning can potentially yield effective results, it is dependent on our ability to construct valid causal graphs of the model. These graphs must consist of directional causal mechanisms through which both sensitive attributes and proxy attributes are related to the predictor. This limitation dramatically reduces the scope of utilizing causal reasoning as a fairness metric.

Dwork et. al. proposes another interesting definition of non-observational fairness that assumes a similarity measure on individuals that requires two similar individuals to receive a similar distribution over possible outcomes. This approach abandons previous definitions that focussed on unfair treatment of groups and instead puts the individual in the foreground. Like previous methodologies, Dwork et. al.'s individual fairness is based on protecting sensitive attributes while computing similarity between individuals.

Most of the contributions to the field have been to create fair classifiers by altering the machine's training process, i.e, to ensure that the model fairly does not unfairly discriminate against groups or individuals during while classifying. These approaches are centred around protecting sensitive attributes (race, ethnicity etc.) such that the predictor cannot unfairly discriminate against members having particular protected features. In our view, this narrow focus has barred us from viewing this issue from a different perspective. This is to say that focusing on sensitive attributes have led us to assume that unfairness is solely a function of (mis)utilizing these attributes. Finally, although the concept of individual fairness has been expounded on in recent literature, there is scope to explore the inherent value in using this approach in order to identify and create fair predictors. In the next section, we list out our aims and potential contributions relative to the related work within the field of fairness in machine learning.

3. Our Objectives and Contributions

We assay the problem of unfair discrimination in machine learning by exploring the characteristics of individual fairness. Our aims are the following:

- To develop a methodology that allows us to identify whether decisions made by the classifier are unfairly discriminating against any individual. In contrast to immediately rectifying the classifier's learning process, our goal is primarily to check whether the classifier is making biased decisions, and against whom these decisions are being taken. Therefore, we follow a post-classification approach that can be further extended to remedy the classifier. Post classification analysis allows the predictor to be a good predictor of the real world before identifying unfairness.

- Unlike previous notions of fairness, our understanding is that each attribute has the ability to unfairly discriminate against an individual. This is to say that focussing our attention on sensitive attributes might be a narrow methodology to follow given that unfair discrimination can have diverse sources. Therefore, we aim to create a general framework that treats each attribute similarly to every individual. This skeleton can serve to be a platform on which we can utilize other methods such as causal reasoning in situations wherein valid causal graphs are possible.

- To differentiate the concepts of discrimination from that of unfairness. These two terms are used synonymously in the colloquial language, yet there is a vast difference between them in the domain of machine learning. Our definitions describe the inherent utility of an attribute to have the ability 'to discriminate' in order to mark decision boundaries. In contrast, unfairness is shown to be a subset of discrimination.

4. Approach

We follow a methodology to identify if an individual has been classified unfairly after the predictor has been trained. In order to expound on the meaning of an unfair treatment, it is essential for us to define the following terms:

4.1. Defining Discrimination

We conceptualize discrimination in machine learning to be the inherent characteristic of an attribute that allows the model to create an effective decision boundary.

Therefore, discrimination is not necessarily a negative phenomenon since the lack of it would result in inaccurate predictions by the machine. In other words, an attribute is chosen to be used in a model because it describes an aspect of the real world that we wish to replicate and

predict. This prediction is an outcome of the feature that is discriminating against subsets of data points in order for the model to effectively produce a decision.

4.2. Defining Unfairness

An individual is treated unfairly when a significant proportion of other similar individuals are categorized differently.

The above definition is an intuitive perspective to describe unfairness or unfair discrimination. In order to put this notion to practice, we are required to break down the definition and quantify two vital elements — similar individuals and significant proportion. An interesting point to note is that any attribute, not just sensitive features, can mislead the classification to produce unfair results. In this light, each attribute should be treated the same when discussing its potential to unfairly discriminate. Therefore, unfairness is a subset of discrimination, such that the ability of an attribute to discriminate may (but not always) lead to unfair discrimination.

4.3. (a) Understanding "Similar Individuals"

We consider the individual to be represented by a feature vector on the input space. In order to identify other similar feature vectors, we needed to develop an appropriate distance metric. Initially, we had considered the first 'k' nearest neighbour to be similar to the feature vector that is being assessed. This metric is used in our experiments as depicted in Section 5. Although following this methodology can lead to fruitful results, it cannot be considered to be an appropriate distance metric using Euclidean distance. This is because the metric is dependent on the dataset. To elaborate, consider an individual feature vector in which the first 10 neighbours are at a significant distance away from the feature vector. In this case, the individual is not at all similar to its neighbour since they are far apart.

In order to combat the issue with the nearest neighbour metric, we propose a new measure of similarity which considers a radius r around the given feature vector. The value of r is chosen such that all the points that lie within this radius can be considered similar to the feature vector. We aim to optimize for the value of r in the next semester. This general framework allows for a future modification wherein the distance metric can be influenced by a protected variable if required.

4.3. (b) Understanding "Significant Proportion"

In our definition of unfairness, we can use different metrics to measure a significant proportion of similar individuals. In our experiment, we use the max voting rule. In this case, if a majority of similar individuals are categorised different from the given feature vector, then the feature vector or individual is unfairly discriminated against. We have used the max voting measure since it is an intuitive approach to capture which category the individual is supposed to belong to. The limitation of using such a method is that majority voting could be considered to be an extremely utilitarian approach. Therefore numerous other metrics can be adopted such as two-thirds majority and a weighted average of distances.

5. Experiments

For the purposes of our experiment, we use the cod-rna dataset which is used for the detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. During the experiment phase, we considered two different simulations - average distance from clusters and max voting. Going forward we realized that finding average distance of a point from the cluster centres does not determine which class it is meant to be from. This approach was antithetical to our initial definition of fairness. Hence, we decided to proceed our experimentation by analyzing ‘k’ nearest neighbours of a point.

5.1. The nearest neighbours

In this experiment we consider the technical definition of unfairness to be as follows - a feature vector is unfairly discriminated against if a majority of its ‘k’ nearest neighbours belong to a different class. We picked a binary classification dataset for our experiments which had 8 features for over 80,000 instances. Running this data through a Linear SVM model we got 93 % accuracy. To begin with we considered 10 points near the decision boundary of some class **1.** Now we consider the 5 closest neighbours, in our first iteration, and check which class each of them belongs to. According to our definition, we classify a point to be unfairly discriminated if a majority of similar points belong to a different class. In our example, we consider a point to be unfairly classified if more than half of its neighbours are not from the same class. Using this approach we classified a point based on 5, 10 and 15 of its neighbours. Below are the results for each of these iterations. Here, the points selected to be judged fair or unfair are circular and a different colour based on the decision.

6. Results

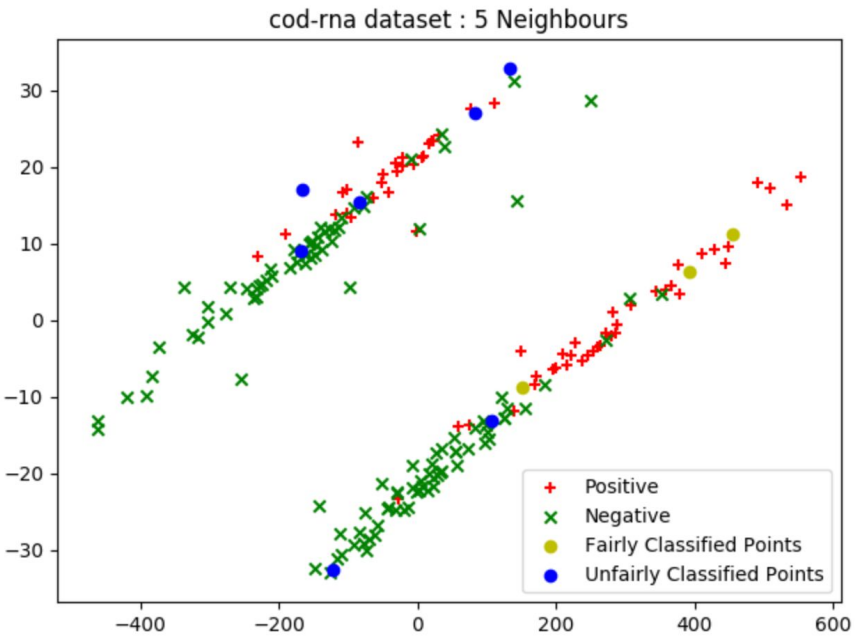

Fig. (i) Top 5 Neighbours is being taken to verify fairness of 10 points.

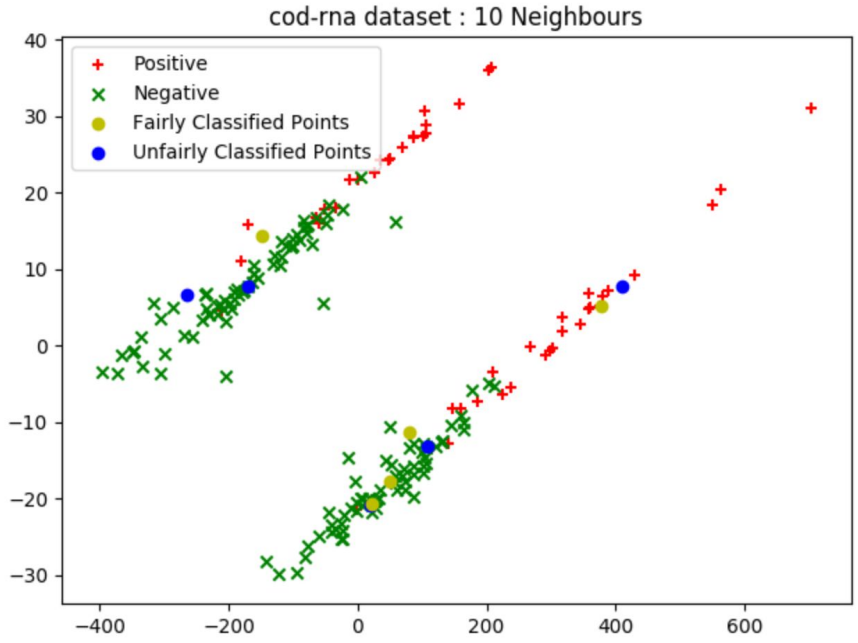

Fig. (ii) Top 10 Neighbours is being taken to verify fairness of 10 points.

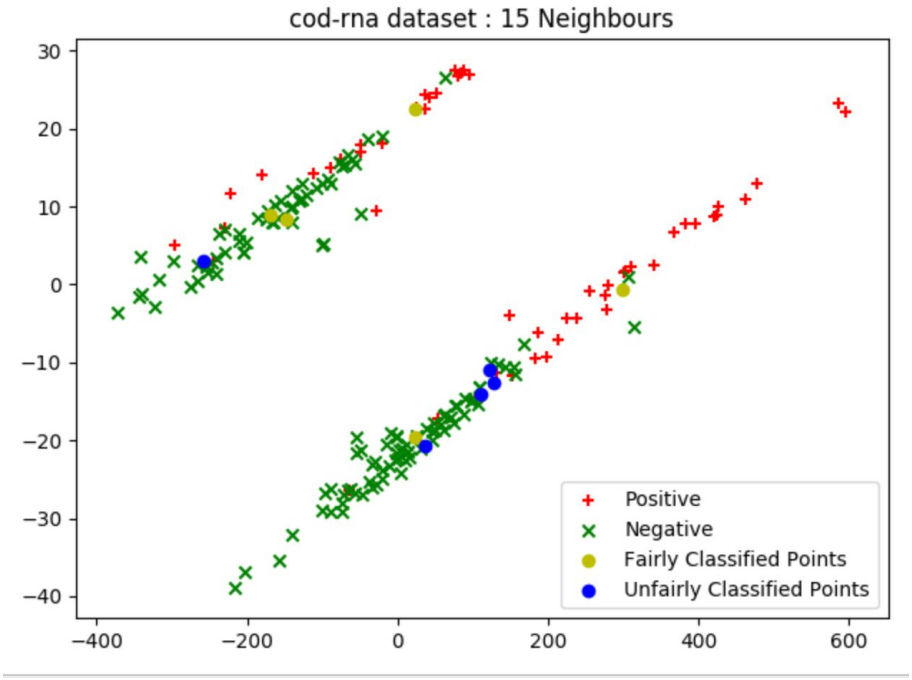

Fig. (iii) Top 15 Neighbours is being taken to verify fairness of 10 points.

Separately, we are now considering a set of 10 points that are constant for each iteration and they are classified as Fair or Unfair based on its neighbours. The results are shown for 4 different values of k .

Number of points taken from Class 1	**Number of Neighbours (*k*)**	**No. of points Fairly Classified**	**No. of points Unfairly Classified**
10	5	3	7
10	10	3	7
10	15	4	6
10	20	1	9

Table (i) Results for constant ten points verified for fairness based on its top k neighbours

7. Future Scope of Study and Conclusions

Contributions to the field of fairness in machine learning are dominated by theories involving sensitive attributes which require protection. Although this approach seems to be intuitively fair, it is not robust enough to ensure truly fair classification. Additionally, each attribute can contribute to the unfair discrimination of an individual. Therefore, we envision to

create a methodology that identifies unfairness post the classification process, while holding each attribute equally accountable. This general framework can prove to be the foundation for future modifications which consider certain sensitive attributes to be the source of unfairness through valid causal mechanisms.

According to our definition, an individual is treated unfairly if a significant proportion of other similar individuals are categorized differently. In order to quantify the two vital elements — similar individuals and significant proportion, we can use certain metrics. Distance is required to measure similarity. Our primary methodology followed that the 'k' nearest neighbours of a feature vector would be similar to that feature vector. As shown, this does not hold true in a general case since the closest points to the individual might be too far apart for them to be considered similar. Additionally, we use the maximum vote metric to quantify "significant proportion". Although maximum voting can yield effective results, it might enforce a utilitarian approach to deciding unfairness. In this light, we propose methods such as two-thirds voting or weighted distance to be metrics that measure "significant proportion".

Our experiments show the shortcomings of using the nearest neighbour approach to establish similarity. Therefore, within the future scope of the study, we aim to develop a radius metric that optimizes a value of the radius for each feature vector such that all points that lie within that radius could be considered as similar to the feature vector. Additionally, our goal is to enhance the model by using different types of distances to strengthen the concept of similarity, namely, Wasserstein (Earth Mover's) distance or sub-space distance. We conclude that our methodology provides a novel method for understanding individual unfairness after the classification process is complete.

By not using sensitive attributes as our starting point, we provide a new perspective on fairness in machine learning wherein each attribute is capable of unfair discrimination. Our aim for the next semester is to optimize the parameters for the radius metric (including the type of distance measured) we can create a model that can identify individual unfairness.

Works Cited

Barocas, Solon and Selbst, Andrew D., Big Data's Disparate Impact (2016). 104 California Law Review 671 (2016). Available at SSRN: https://ssrn.com/abstract=2477899 or http://dx.doi.org/10.2139/ssrn.2477899

Dwork, Cynthia, et al. "Fairness through Awareness." *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12*, 29 Nov. 2011, doi:10.1145/2090236.2090255.

Hardt, Moritz, et al. "Equality of Opportunity in Supervised Learning." *Conference on Neural Information Processing Systems*, 2016, arxiv.org/pdf/1610.02413.pdf.

Kilbertus, Niki, et al. "Avoiding Discrimination through Causal Reasoning." 8 June 2017, https://arxiv.org/abs/1706.02744