Legal Safeguards in Algorithmic Decision Making

Harsh Gupta

12MA20017

under the supervision of
Dr. PVSN Murthy
and
Prof. Shreya Matilal

Indian Institute of Technology
Kharagpur, India

Certificate

This is to certify that the project titled "Legal Safeguards in Algorithmic Decision Making" submitted by Harsh Gupta, Roll No: 12MA20017 during 2016-2017, in partial fulfillment of the degree of Integrated M.Sc. in Mathematics and Computing to the Indian Institute of Technology Kharagpur, is a bonafide record of the work carried out under my supervision and guidance. This report fulfills all the requirements as per regulations of the institute and has not been submitted to any other institute/university for any degree or diploma.

Dr. PVSN Murthy
Department of Mathematics
IIT Kharagpur

Prof. Shreya Matilal
RGSOIPL
IIT Kharagpur

Acknowledgement

I would like to thank **Dr. PVSN Murthy** and **Prof. Shreya Matilal** for their continued guidance, support and supervision. I thank Department of Mathematics and our faculty advisor **Dr. G P Raja Shekhar** for allowing me to carry out inter disciplinary research. I also thank Kumar Krishna Agarwal and Shivam Vats for insightful conversations.

Abstract

From face recognition to hiring, recent years have a growth of use of machine learning in large areas of human endeavor. With the rise and ubiquity of algorithms, we are also seeing that algorithms are not living up to their promise. They are found to only reflect the existing bias and discrimination found in today's society but also exaggerate it. There is a growing literature which trying to classify, detect and rectify various kinds of discriminations which can be found in machine learning applications, who have developed various formalized notions of fairness. In this work we take one such formalization and try to adapt it in the context of equality and anti-discrimination provisions in Constitution of India so that it can be give force of law.

Contents

Introduction **1**

Motivating Examples **1**
- Racial in Crime Risk Assessment 1
- Blacks being tagged as Gorillas 2
- Occupational Gender stereotypes in Google Images 2
- Staples Dynamic Pricing 2
- Job Advertisements 2

Overview of reasons of Biases **3**
- Poorly Collected or Poorly Selected Data 3
- Poor Algorithm 3

Formalizing Notions of Fairness and Non Discrimination **3**
- Construct Space, Observed Space and Decision Space 4

Fairness and anti-discrimination in The Constitution of India **6**
- Article 14 6
- Article 15(1) 7

Summary and Conclusion **8**

Discussion and Further Research **8**

References **8**

Introduction

Recent year have seen an exponential rise in the use of machine learning and algorithmic systems in almost all domains, including very "human" aspects of human society. Now algorithmic systems suggests us which music should we listen to, which movie to watch, which news article is worth reading, which resumes to forward to the hiring manager, who should be given admission and even who should be given a larger sentence for a crime. Historically some of these areas, like employment, education and crime have seen a lot of contention regarding bias and discrimination.

Often algorithmic systems are projected as something which is neutral and has the ability to fix human bias. But recent experiences have been shown that they can not only reflect human biases but also exaggeration it. Large amount of literature is emerging which brings out more of these biases, finds out why these biases occur and how to tackle them. These algorithms takes decisions in real world, affecting people real people in real ways, therefore we believe their behavior needs to regulated to prevent harm and the legal notions of equality and anti-discrimination need to built into the machine learning systems. In order to do so we need to frame these concepts in ways such that machines can understand and enforce them.

The first step towards the goal is to formulate these notions as abstract mathematical concepts so that machines can reason about them, the second step is fit into the legal framework. Our work looks at a very recent formulation of fairness in terms of mapping between metric spaces and tries to adapt it for Article 14 and Article 15 of The Constitution of India.

Motivating Examples

Racial in Crime Risk Assessment

Some jurisdictions in the United States use a software system to assess the risk of a re-offending by an offender. A study by ProPublica found that though the accuracy of the system for whites and blacks was similar, the type of error they make for different races are very different. Blacks are almost twice as likely as whites to be labeled a higher risks but not actually re-offend. The system made the opposite mistake for whites. [1]

Error type	White	Black
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

This particular reported has generated wide response, including rebuttal from

Northpoint, the company which makes the risk assessment software and rebuttal of rebuttal from ProPublica. The main argument from Northpoint's side is that rate of violent crimes is higher among blacks directly leads to asymmetry in the two error rates and correcting for the disparity in error rates will lead to fair treatment to white inmates [2]. Recent literature, has suggested that it might be impossible to achieve certain notions of fairness without compromising on others. [3]

Blacks being tagged as Gorillas

Users found that in a Machine Learning based image tagging system in Google Photos produced offensive labels, tagging black people in photos as "gorillas". [4] In a similar example, flickr was found to tag a black person as 'Ape' [5] and in 2009 it was seen that when HP's web cam wasn't able to recognize black faces when it was easily able to track white ones [6] Though there are technical explanation for these behaviors of the algorithms, Face recognition techniques are used widely to identify offenders and error or bias in these systems can lead to loss of reputation and livelihood of innocent people. [7]

Occupational Gender stereotypes in Google Images

The current society has large amount of gender bias with regards to occupations. It was found that Google Image search results exaggerate the existing gender bias in occupation. This is important because representation of gender in image search results shifts people's perception about real world distribution. [8]

Staples Dynamic Pricing

An online shopping store in US, Staples dynamically price items for customers on the basis of their browsing history, location from the nearest competitor store and other parameters. A study by The Wall Street Journal found that the pricing scheme worked negatively for low-income households as they are more likely to live outside the core of the city. This considered as an unintended side effect of the pricing scheme. [9]

Job Advertisements

A study on effects of web tracking on advertisements by Google found that setting the gender to female resulted in fewer instances of advertisements related to high paying jobs. [10]

Overview of reasons of Biases

Poorly Collected or Poorly Selected Data

Good algorithms require good data to work on it. The source of data can have inherent biases, for example the typical editor of Wikipedia is someone who is White, college educated, western and male [11] [12]. The content of wikipedia also under represents non western population of the world. [13]

The data get corrupted while handing and transformation, or though bugs in software [14].

The process of collecting data can also introduce certain bias in the data, for example if certain data is collected through online form, people with low access to the internet are likely to missing from the data, and these people are also more likely to be low income, less educated, non english speaking and from depressed race and caste. The errors introduced through selection process can be very subtle, such disparities which are often under appreciated.

Poor Algorithm

Even if the underlying data is of good quality and represents the population well, the design of the algorithm might introduce certain biases in the result. One example such effects is called uncertainty bias.

The uncertainty associated with a subpopulation in data has inverse relation with the sample. In cases the algorithmic system takes decisions conservatively, uncertainty bias can lead to inferior result for population with lower sample, even if everything else remains same.

To illustrate this consider you have a sample 1000 people, 900 white and 100 black and all of them have same likelihood of defaulting on a loan 5 %, and you give loan to any with a default rate of less 10%. Now if you give loans conservatively (if you think chances of someone defaulting is between 3 to 8 % then you'll consider defaulting rate to be 8%) you won't give loans to any of the black folks and to every white folk and that because their sample size is smaller the uncertainty associated with black sample is higher than the white sample. [15]

Formalizing Notions of Fairness and Non Discrimination

In the literature of discrimination aware machine learning, researchers have used various definitions of fairness which mutually exclusive at times. For analysis of

anti-discrimination provisions in The Constitution of India we adapt a framework developed by Friedler et. al. [16]

Because their framework is central to our discussion on Formalization of Article 14 and 15 of Constitution of India we reproduce a large section of their paper here.

For the purpose of this report we will call the framework as framework of construct space.

Construct Space, Observed Space and Decision Space

In framework of construct space we view the task of producing a decision algorithm as finding a map from the true characterises of a person at a given time, called construct space, to a decision variable in the space of possible decisions called decision space. Formally,

Construct space represents the true features of the individuals which we would like to use these features might contain attributes like intelligence, skills in the area, ability to work in team, communication ability etc. But because these attributes aren't always well defined, and very hard or impossible to observe directly we see a manifestation of these attributes in form of observable characteristics like GPA, scores on standardized tests, degrees obtained, etc. This space of observable attributes is called observe space. Formally

Definition *The construct space is a metric space $CS = (P, d_p)$ consisting of individuals and a distance between them. It is assumed that the distance d_p correctly captures closeness with respect to the task.*

Definition *The observed space with respect to the task at hand, is a metric space $OS = (\hat{P}, \hat{d})$. We assume an observation process $g : P \to \hat{P}$ that generates an entity $\hat{p} = g(p)$ from a person $p \in CS$*

And **Definition** *A decision space is a metric space $DS = (O, d_O)$, 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 P or $\hat{P}$ to O.*

Then they take the informal understanding of fairness that similarly situation people should be treated similarly. In terms present framework, it means that two persons who are close in the construct space should be close in decision space. Defined formally as

Definition *A mapping $f : CS \to DS$ is said to be fair if objects that are close in CS are also close in DS. Specifically, fix two thresholds ϵ, ϵ'. Then f is defined as* (ϵ, ϵ)- fair if for any $x, y \in P$,

$$dP(x, y) \leq \epsilon \implies d_O(f(x), f(y)) \leq \epsilon'$$

Because, the mapping from Construct Space which is largely unknown. We need to make assumptions about the mapping from Construct Space to Observation

Space. One such

Hence, assumptions is introduced about mapping from Construct Space to Observation Space. One such assumption is that Observation space is a good proxy for Construct Space, in other words two people who are close in construct space are also close in observation space. Friedler et. al. call this *what you see is what you get (WYSIWYG)* axiom.

Axiom (WYSIWYG) There exists a mapping $f : CS \to OS$ such that the distortion ρ_f is at most ϵ for some small $\epsilon > 0$. Or equivalently, the distortion ρ_f between CS and OS is at most ϵ.

Distortion ρ_f of $f : X \to Y$ for metric spaces (X, d_X) and (Y, d_Y) is defined as the smallest value such that for all $p, q \in X$

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

.

Under this axiom, Friedler et. al. defines Individual fairness mechanism as

Definition Fix a tolerance ϵ. A mechanism IFM_ϵ is a mapping $f : OS \to DS$ such that for all $\rho_f \leq \epsilon$

Unfortunately in many real world applications WYSIWYG axiom doesn't hold true. Due to historical, cultural and other reasons the mapping of CS to OS is more distorted for individuals belonging to certain groups than other individuals.

This dispositional skew is called structural bias

Definition *(Group fairness mechanism (GFM)). Let X be partitioned into groups $X_1, X_2, \ldots$ as before. A rich mapping $f : OS \to DS$ is said to be a valid group fairness mechanism if all groups are treated equally. Specifically, fix ϵ. Then f is said to be a GFM_ϵ if for any i, j and a group distance function W, $W_{d_O}(Xi, Xj) \leq \epsilon$*

As an interesting result, Friedler et. al. showed that if systemic bias is present it is impossible to achieve both Individual fairness mechanism and Group fairness mechanism.

Another important result form Friedler et. al. is that Individual Fairness Mechanism is impossible if the decision space is discreet. To understand the result intuitively, consider a test where passing marks are 30 out of 100, there is one student who gets 31 marks and is passed and there is another student who got 29 marks and is failed. In terms of marks both the students are similarly situated in observation space but are far away in decision space. So wherever there is a decision where everyone is not given the same treatment, there will exist two people who are similarly situated but gets very different results, one just inside the decision boundary and another just outside it.

Numerous real world decisions are discreet and binary, to hire or not to hire, to give admission or not to give admission, to give loan or not to give loan.

A definition of fairness where it is impossible to fair in so many application areas isn't a very useful definition. Hence, we slightly modify the Individual Fairness Mechanism for binary decisions by considering decision scores instead of actual decisions. Decision Score is real number assigned to every individual who is considered for a particular decision and then the actual decision is done on the basis of a cut off in the decision score space. Sometimes the decision score is explicitly part of the decision making process. For example Indian Institute of Technology conduct a Nationwide test called Joint Entrance Exam and students are admitted only on the basis of the rank they obtain in the test. In cases such score is not explicitly part of the system we consider decision score to the probability that a certain decision will be made for the individual.

Definition *Decision Score Space* $DSS = ([a, b], d)$ *where* $a, b \in R, b > a$ *and* d *is Euclidean distance.*

Definition Binary Individual Fairness Mechanism: $f(g(.))$ where $DS = (\{0, 1\}, d_O)$, $g : OS \to DSS$ such that $\rho_f > \epsilon$ and $f : DSS \to DS$ where for all $x \in DSS$ $f(x) = 0, x \leq z$ and $f(x) = 1, x > z, a \leq z \leq b$

Fairness and anti-discrimination in The Constitution of India

Part of III of The Constitution of India specifies the Fundamental Rights given to every citizen of India. Of then Article 14 describes right to equality, Article 15 formulates the anti-discrimination provisions and Article 16 talks about anti-discrimination in terms of public employment by the government. Though these provisions apply only to actions by State, these cover a large portion real world issues which can occur though machine discrimination.

Corporations largely owned and funded by the Government are also covered under these provisions.

Article 14

Article 14 of The Constitution gives every citizen right to equal treatment before law and it says:

> The State shall not deny to any person equality before law or equal protection of the laws withing the territory of India.

Here law is understood as any action by the State and not just statutory law. One of the widely used test to determine equality under 14 is called test of reasonable classification.

Equality here doesn't mean that every citizen should be treated in the exact same matter but is understood as similarly situated persons should be treated similarly. This notion if equality is similar to the one used by Friedler et. al. To determine when State can make different treatment, the Courts have formulated what is called the test of

reasonable classification. The State can treat two persons differently only if:

1. There is an intelligible differential between the two persons
2. There is an objective for the different treatment and the different treatment satisfies the objective.

This is not to say only the test of reasonable classification determines equality in context of Article 14. Court have used various legal and philosophical principles to determine when a particular action of State is to considered equal. But test of reasonable classification is to understood as a minimum criterion for equality.

We believe that Individual Fairness Mechanism along with Binary Individual Fairness Mechanism described above captures the test of reasonable classification well. The Case of Subramanyan Swami v. Raju illustrates the importance of using Decision Score Space instead of Decision Space. In this Subramanyan Swami contested the age at which a person attains the status of major, which is 18 years in India. The Court said even though the age 18 is arbitrary, it is not unreasonable, hence it does not violate Article 14.

Article 15(1)

Article 15(1) says:

> The State shall not discriminate against any citizen only on the grounds of religion, race, caste, sex, place of birth or any of them.

The word "only" is important here, Article 15(1) doesn't guarantee equality of outcome of these different groups but says that while considering whether two people are similarly or situated or not sex, religion, race, caste or place of birth.

In terms of the formalization we did above, Article 15(1) says that the distance function in OS should be invariant of these groups.

As it has been widely studied in the context of racial bias in United States, ignoring the race doesn't eliminate racial discrimination as wide array of useful information, from someone's pin code to the name is correlated with race. What is not clear is that to extent is Article 15(1) forces one to normalize these parameters in terms of race.

Both Article 14 and Article 15(1) works towards ensuring individual fairness mechanism but as we have discussed in the section above doing so disallows us to work towards group level fairness. To resolve this conflict provisions to Article

15(1) and 15(2), that is 15(3),15(4) and 15(5) allows state to override individual fairness mechanism to ensure group level fairness. They read as follows:

> (3) Nothing in this article shall prevent the State from making any special provision for women and children.
>
> (4) Nothing in this article or in clause (2) of article 29 shall prevent the State from making any special provision for the advancement of any socially or educationally backward classes of citizens or for the Scheduled Castes and the Scheduled Tribes.
>
> (5) Nothing in this article or in sub-clause (g) of clause (1) of article 19 shall prevent the State from making any special provision, by law, for the advancement of any socially and educationally backward classes of citizens or for the Scheduled Castes or the Scheduled Tribes in so far as such provisions relate to their admission to educational institutions including private educational institutions, whether aided or unaided by the State, other than the minority educational institutions referred to in clause (1) of article 30.

Summary and Conclusion

In Summary we state that our modified Binary Individual Fairness Mechanism along with Individual Fairness Mechanism captures the essence of test of reasonable classification under article 14 and anti-discrimination under Article 15(1).

Discussion and Further Research

Future studies should look take up a deeper study of the case laws and analyze how the formal notions of fairness and anti-discrimination applies in these circumstances.

References

[1] J. Angwin, J. Larson, S. Mattu, and K. Kirchner, "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks." *ProPublica*. May-2016 [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. [Accessed: 11-Nov-2016]

[2] "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear." *Washington Post*. Oct-2016 [Online]. Available: https://www.washingtonpost.com/news/monkey-cage/wp/

2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/. [Accessed: 12-Nov-2016]

[3] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," *arXiv:1609.05807 [cs, stat]*, Sep. 2016 [Online]. Available: http://arxiv.org/abs/1609.05807. [Accessed: 14-Nov-2016]

[4] "Google Photos labeled black people 'gorillas'," *USA TODAY*. 2015 [Online]. Available: http://www.usatoday.com/story/tech/2015/07/01/google-apologizes-after-photos-identify-black-people-as-gorillas/29567465/. [Accessed: 11-Nov-2016]

[5] "Flickr Fixing 'Racist' Auto-Tagging Feature After Black Man Mislabeled 'Ape'," *PetaPixel*. May-2015 [Online]. Available: http://petapixel.com/2015/05/20/flickr-fixing-racist-auto-tagging-feature-after-black-man-mislabeled-ape/. [Accessed: 12-Nov-2016]

[6] A. Frucci, "HP Face-Tracking Webcams Don't Recognize Black People," *Gizmodo*. 2009 [Online]. Available: http://gizmodo.com/5431190/hp-face-tracking-webcams-dont-recognize-black-people. [Accessed: 12-Nov-2016]

[7] A. Kofman, "How a Facial Recognition Mismatch Can Ruin Your Life," *The Intercept*. 2016 [Online]. Available: https://theintercept.com/2016/10/13/how-a-facial-recognition-mismatch-can-ruin-your-life/. [Accessed: 09-Nov-2016]

[8] M. Kay, C. Matuszek, and S. A. Munson, "Unequal Representation and Gender Stereotypes in Image Search Results for Occupations," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 3819–3828 [Online]. Available: http://doi.acm.org/10.1145/2702123.2702520. [Accessed: 09-Nov-2016]

[9] J. Valentino-DeVries, J. Singer-Vine, and A. Soltani, "Websites Vary Prices, Deals Based on Users' Information," *Wall Street Journal*, Dec. 2012 [Online]. Available: http://www.wsj.com/articles/SB10001424127887323777204578189391813881534. [Accessed: 11-Nov-2016]

[10] A. Datta, M. C. Tschantz, and A. Datta, "Automated Experiments on Ad Privacy Settings," *Proceedings on Privacy Enhancing Technologies*, vol. 2015, no. 1, pp. 92–112, 2015 [Online]. Available: https://www.degruyter.com/view/j/popets.2015.1.issue-1/popets-2015-0007/popets-2015-0007.xml?ncid=txtlnkusaolp00000618. [Accessed: 11-Nov-2016]

[11] "Wikipedia Survey – Overview of Results." Mar-2010 [Online]. Available: https://web.archive.org/web/20110728182835/http://www.wikipediastudy.org/docs/Wikipedia_Overview_15March2010-FINAL.pdf. [Accessed: 14-Nov-2016]

[12] S. T. K. Lam, A. Uduwage, Z. Dong, S. Sen, D. R. Musicant, L. Terveen, and J. Riedl, "WP: Clubhouse?: An exploration of Wikipedia's gender imbal-

ance," in *Proceedings of the 7th international symposium on Wikis and open collaboration*, 2011, pp. 1–10 [Online]. Available: http://dl.acm.org/citation.cfm?id=2038560. [Accessed: 14-Nov-2016]

[13] M. Graham, R. K. Straumann, and B. Hogan, "Digital Divisions of Labor and Informational Magnetism: Mapping Participation in Wikipedia," *Annals of the Association of American Geographers*, vol. 105, no. 6, pp. 1158–1178, Nov. 2015 [Online]. Available: http://dx.doi.org/10.1080/00045608.2015.1072791. [Accessed: 14-Nov-2016]

[14] J. Hrala, "Excel Is to Blame for Major Typos in 20% of Scientific Papers on Genes," *ScienceAlert*. Aug-2016 [Online]. Available: http://www.sciencealert.com/excel-is-responsible-for-20-percent-of-errors-in-genetic-scientific-papers. [Accessed: 14-Nov-2016]

[15] B. Goodman and S. Flaxman, "European Union regulations on algorithmic decision-making and a' right to explanation'," *arXiv preprint arXiv:1606.08813*, 2016 [Online]. Available: https://arxiv.org/abs/1606.08813. [Accessed: 09-Nov-2016]

[16] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, "On the (im)possibility of fairness," *arXiv preprint arXiv:1609.07236*, 2016 [Online]. Available: https://arxiv.org/abs/1609.07236. [Accessed: 09-Nov-2016]