Fairness-Aware Machine Learning: Practical Challenges and Lessons Learned

Sarah Bird
Microsoft
slbird@microsoft.com

Ben Hutchinson
Google
benhutch@google.com

Krishnaram Kenthapadi
LinkedIn
kkenthapadi@linkedin.com

Emre Kıcıman
Microsoft
emrek@microsoft.com

Margaret Mitchell
Google
mmitchellai@google.com

ABSTRACT

Researchers and practitioners from different disciplines have highlighted the ethical and legal challenges posed by the use of machine learned models and data-driven systems, and the potential for such systems to discriminate against certain population groups, due to biases in algorithmic decision-making systems. This tutorial aims to present an overview of algorithmic bias / discrimination issues observed over the last few years and the lessons learned, key regulations and laws, and evolution of techniques for achieving fairness in machine learning systems. We will motivate the need for adopting a "fairness-first" approach (as opposed to viewing algorithmic bias / fairness considerations as an afterthought), when developing machine learning based models and systems for different consumer and enterprise applications. Then, we will focus on the application of fairness-aware machine learning techniques in practice, by highlighting industry best practices and case studies from different technology companies. Based on our experiences in industry, we will identify open problems and research challenges for the data mining / machine learning community.

ACM Reference Format:
Sarah Bird, Ben Hutchinson, Krishnaram Kenthapadi, Emre Kıcıman, and Margaret Mitchell. 2019. Fairness-Aware Machine Learning: Practical Challenges and Lessons Learned. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3292500.3332280

1 OUTLINE OF THE TUTORIAL

The tutorial will consist of the following parts: an overview of algorithmic bias / discrimination issues observed in practice over the last few years and the lessons learned (1 hr), fairness notions and techniques from the perspective of different web applications (1 hr), and case studies from different technology companies, along with open problems and research directions (1 hr).

We will cover the following key topics:

- Introduction to algorithmic bias / discrimination
- Industry best practices
- Sources of biases in ML lifecycle
- Algorithmic techniques for fairness in ML
- Fairness methods in practice: Challenges and lessons learned
  - Image and related topics
  - Machine translation
  - Conversational agents
  - Web search, talent search, and other ranking domains
  - Key takeaways, reflections, and open problems

Tutorial webpage:
https://sites.google.com/view/kdd19-fairness-tutorial

1.1 Algorithmic Bias in Machine Learning

The topic of algorithmic bias and discrimination has been studied extensively across disciplines such as law, policy, and computer science (e.g., see [2, 11] and the references therein). An early work considers a computer system to be biased if it systematically and unfairly discriminates against certain individuals or groups of individuals in favor of others [9]. Systematic discrimination combined with an unfair outcome is considered to result in bias.

Many scholars have investigated two different notions of fairness: (1) *individual fairness*, which requires similar treatment for similar people [6], and (2) *group fairness*, which requires similar treatment for the disadvantaged group and the advantaged group [16]. Some studies focus on quantifying the extent of discrimination (e.g., [1, 3, 16]), while others explore fairness-aware algorithms for mitigating bias (e.g., [4–8, 12–15, 17–20]) and limitations / trade-offs in achieving different notions of fairness [5–7, 15].

In the tutorial, we will also discuss sources of data bias at various stages of the data generation, collection, and analysis pipeline, and current best practices for identifying/monitoring and mitigating these sources of bias. We will present examples of algorithmic bias / discrimination issues observed over the last few years, highlighting the pitfalls / lessons learned and strategies to address these issues.

1.2 Fairness Notions

We will discuss different definitions of fairness in machine learning (such as statistical parity, equalized odds, and equal opportunity [13]), and the relative advantages and limitations of these definitions. We will also present a broad taxonomy of web applications, and discuss appropriate fairness notions for each class.

Ranking users for a given "query": Examples include ranking individuals for credit offers and lending, ranking potential candidates in response to a recruiter's search query or job posting,

KDD '19, August 4–8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08.
https://doi.org/10.1145/3292500.3332280

ranking results about people in web or social search, and ranking students for college admissions. Here we are concerned with fairness to the users being ranked. Applicable fairness notions include statistical parity [6], equalized odds/opportunity [13], and fairness-aware ranking [4, 10, 19].

Ranking items for a given user: Several common web applications fall into this category (e.g., news article recommendations, job search and recommendations, microtargeted ads, and movie recommendations). The items could be ranked based on explicit search query, on user profile, and/or on past user behavior. Here we are concerned with fairness for the user to whom we present the ranked items. For instance, a desirable fairness notion could be to ensure that we present similar items to similar users (individual fairness). In some cases, there may be conflicts to resolve between what the user likes as judged by behavior, and fairness goals. For instance, men and women might have different preferences for specific types of news articles or ads.

Ranking other users for a given user: Examples include YouTube channel recommendations, people recommendations (*People You May Know* of Facebook and LinkedIn, *Who To Follow* of Twitter), and various dating applications. We could be concerned about fairness for both the users that receive recommendations, and the users they are recommended to or not. Also, there may be a fine balance needed between fairness and preferences – for instance, supposing that users are more likely to connect with others of similar age or experience level, one needs to ponder whether to recommend people of different ages to people of different ages.

Other considerations: There may also be fairness considerations that do not belong to the above categories. Consider for example, personalized web applications that send notifications to users (e.g., push notifications in the app, email notification). It may be desirable to ensure that different classes of users are treated similarly with respect to how often and when they are presented with such notifications, especially when the notification corresponds to an opportunity (e.g., a credit offering or a job recommendation). For instance, a model that presents men with an earlier notification of a job opening or more frequent notifications of educational or job opportunities could result in disparate impact to women. Further, different classes of users may experience significantly dissimilar treatments due to the presence of externalities. For example, a provider of educational or employment opportunity may either be not able or need to pay more to reach a class of users (e.g., women) due to the presence of other competing ads.

1.3 Case Studies from Industry

As part of the tutorial, we will also focus on the application of fairness-aware machine learning techniques in practice, by presenting case studies from different technology companies spanning applications such as image/vision, machine translation, conversational agents, and talent search (e.g., representative ranking for LinkedIn talent search [10]). We will discuss AI & ethics initiatives within and across different companies (e.g., Partnership on AI), focusing on technical solutions being adopted to address known issues. Finally, we will discuss open challenges and research directions for the community.

2 BRIEF BIOGRAPHIES OF THE PRESENTERS

Sarah Bird leads strategic projects at the intersection of AI research and products at Microsoft. Her current work focuses on AI Ethics and developing AI responsibly at scale. Ben Hutchinson is a Senior Engineer in Google's Research & Machine Intelligence group, working on AI, fairness, and ethics, in Google's Ethical AI team. His interdisciplinary research includes learning from social sciences to inform the ethical development of AI. Krishnaram Kenthapadi is part of the AI team at LinkedIn, where he leads the fairness, transparency, explainability, and privacy modeling efforts across different LinkedIn applications. He also serves as LinkedIn's representative in Microsoft's AI and Ethics in Engineering and Research (AETHER) Committee. Emre Kıcıman is a Principal Researcher at and co-leads Microsoft Research AI's efforts on AI and its implications for people and society. In addition, his research focuses on causal analysis and data bias in the context of computational social science analyses and decision support systems. Margaret Mitchell is a Senior Research Scientist in Google's Research & Machine Intelligence group, working on AI, multimodality, and ethics, and she currently leads Google's Ethical AI team. Her research involves vision-language, computer vision, and grounded language generation, focusing on how to evolve AI towards positive goals.

REFERENCES

[1] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. *ProPublica*, 2016.
[2] S. Barocas and M. Hardt. Fairness in machine learning. In *NIPS Tutorial*, 2017.
[3] A. Caliskan, J. J. Bryson, and A. Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 2017.
[4] L. E. Celis, D. Straszak, and N. K. Vishnoi. Ranking with fairness constraints. In *ICALP*, 2018.
[5] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *KDD*, 2017.
[6] C. Dwork, M. Hardt, T. Pitassi, and R. Z. Omer Reingold. Fairness through awareness. In *ITCS*, 2012.
[7] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. On the (im) possibility of fairness. *arXiv:1609.07236*, 2016.
[8] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A comparative study of fairness-enhancing interventions in machine learning. *arXiv:1802.04422*, 2018.
[9] B. Friedman and H. Nissenbaum. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, 14(3), 1996.
[10] S. C. Geyik, S. Ambler, and K. Kenthapadi. Fairness-aware ranking in search & recommendation systems with application to LinkedIn talent search. In *KDD*, 2019.
[11] S. Hajian, F. Bonchi, and C. Castillo. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *KDD Tutorial on Algorithmic Bias*, 2016.
[12] S. Hajian, J. Domingo-Ferrer, and O. Farràs. Generalization-based privacy preservation and discrimination prevention in data publishing and mining. *Data Mining and Knowledge Discovery*, 28(5-6), 2014.
[13] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *NIPS*, 2016.
[14] S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, and A. Roth. Fairness in reinforcement learning. In *ICML*, 2017.
[15] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. In *ITCS*, 2017.
[16] D. Pedreschi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *KDD*, 2008.
[17] B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro. Learning non-discriminatory predictors. In *COLT*, 2017.
[18] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *WWW*, 2017.
[19] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates. FA*IR: A fair top-k ranking algorithm. In *CIKM*, 2017.
[20] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *ICML*, 2013.