Soumya Konar

CISY 8503 – Application Database Management

Professor Enke

Research Paper – Machine Learning Bias and Fairness

November 20, 2022

We would like to think that Machine Learning and the idea of self-aware machines has developed only over the past few years – when these ages back to Greek Mythology in terms of self-aware machines. We started with Charles Babbage, the first general purpose computer and here we are now. As every second passes, there is a new model to represent innovation and technology, something that displaces the previous one in this fast society. This applies to Machine learning, a subset of Artificial Intelligence – working towards a new reality. This research paper will delve into the importance of Machine Learning Bias and Fairness models – highlighting the importance of bias ratios and fairness metrics through multiple peer-reviewed case studies.

The question in context is not about what Machine Learning Bias and Fairness is in definition but the impact the several types of ML bias and fairness have in our society. Throughout this research paper, we will cover concepts of bias and fairness, detection, and types, along with case studies pertaining to bias and fairness in machine learning systems implemented.

Bias in machine learning algorithms, also known as AI bias and Algorithm bias occurs when an error based on prejudices is caused that skews the predictions away from the accurate results. This is where the terms overfitting and underfitting arise from. “Both overfitting and underfitting lead to poor predictions on new data sets” (Medium.com, 2020). The concept of variance in a machine learning algorithm portrays the spread of our data, where the model generalizes on the data therefore resulting in errors from test data. The opposite is the concept of bias where the algorithm trains in a broader scope, which oversimplifies the model. Underfitting results when there is a high bias and low variance in the model, while overfitting occurs with low bias and high variance.

When you look at a problem to solve through a machine learning technique – a couple of questions should always be asked to cover all bases. Through research from the First Report of the Axon Artificial Intelligence and Policing Technology Ethics Board conducted in June 2019, these are a couple questions that machine learning engineers need to address:

- What is the specific problem to be solved?
- How important is the problem?
- How certain is it that the technology will address the problem
- May there be unintended or secondary benefits?
- Can the specific piece of technology be used or misused in unanticipated ways?
- Will it lead to greater criminalization or to policing in counterproductive ways?
- Will technology impact personal information privacy?
- What is the data captured, retained, owned, accessed, protected?
- Does technology implicate potential biases, especially racial or other identity factors, whether in design or use?
- Does technology create transparency-related concerns for the public?
- Does technology risk, directly or indirectly, violations of constitutional or legal rights?

Ramesh Srinivasan, an Indian American professor who pioneers in the intersection of technology innovations, information studies and the society talks about how “We encode our biases into everything we create: books, poems, and AI, and what does that means for an increasingly automated future?” (Busting the myth of “neutral” AI, 2020). This opens multiple

routes of discussion for esteemed scientists, researchers, engineers, and even common society – to decide what is best for humanity. Therefore, pointing to machine learning bias, also known as algorithm bias, where algorithms produce results that are biased or prejudices due to multiple reasons. This research paper will highlight distinct types of bias and fairness throughout various sectors, and the importance of accurate and unbiased solutions. Most of the previous research has treated algorithmic bias as a static factor, which fails to capture the iterative nature of bias that generates from both humans and algorithms.

To support the previously stated statement of bias in multiple sectors – a case study conducted with Google Translate highlights the assessment of gender bias in their machine learning algorithms and techniques. Since 2018, Google Translate is one of the leading machine translation tools available in the market – but recent studies have portrayed concerns that some word and sentence embeddings "particularly prone to yielding gender stereotypes" (Prates et al., 2020). For instance, occupations from traditionally male-dominated fields such as scholar, engineer and CEO are interpreted as male, while occupations such as nurse, baker and wedding organizer are interpreted as female. The case study dives into different solutions proposed by researchers to solve these bias scenarios in any statistical translation tool. It revolves around "mapping sentences constructed in gender neutral languages to English by the means of an automated translation tool" – then working with a number of these languages to see the trend.

This case study allowed me to break down the components while we "debias" the algorithm and the succeeding results – starting with plotting histograms to visualize the data distribution. Heatmaps, computed cell value tables, bar charts are other ways to visualize the data we are looking for – in this case female and male occupation and pronouns. To "strengthen the results, tests were run on gender translation statistics against the U.S. Bureau of Labor Statistics data on

the frequency of women participation for each job position" (Prates et al., 2020). This allowed them to verify that Google Translate failed to meet the equal distribution of male and female employees in any industry. Another case study that can assist this research paper is included in an article titled, "Bias in, bias out" – referencing to the established concept of Garbage in, garbage out. This case study displays the underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer. This revolved around research conducted on diverse skin types and populations to broaden the scope of detecting cancer – diversifying the data set.

Diving off both the case studies, the opposing and prevailing view is that there is a huge market that can benefit from machine learning and artificial intelligence – in turn helping citizens. The opposing view can highlight how even if there is a vital gap between the accurate result and predicted one – due to bias, lack of fairness(parity) and bad data, machine learning innovations have had an enormous impact on society in recent years. With advances in the healthcare industry, automation, predictive analysis algorithms and self-driving cars, humanity has taken a leap forward to tap into the multiple benefits this spectrum of Artificial Intelligence, Machine Learning, and Deep Learning entails.

In the article, "Algorithms of Machines and Law: Risks in Pattern Recognition, Machine Learning and Artificial Intelligence for Justice and Fairness," the author proceeds to discuss the law of IT and AI. It highlights how the facts of AI and predictive systems are part of a "socio-technical system for governance that embraces human decisions, machine decisions and responsibility" (Losavio, 2021). This helps us bring civic liability and governance into question – governance of these self-aware systems that can break or make an individual's life in certain instances.

After covering bias in machine learning systems, another concept that systems focus on is the concept of fairness in any algorithm. In the field of machine learning, there are many challenges and opportunities that one must overcome when dealing with the definition and types of fairness. Fairness is a broad term and constitutes different meanings in different spheres of the technology industry. The first step in addressing these challenges in the fairness context is to synthesize a definition of fairness. Different perspectives conjure different statements – but the underlying definition revolves around the understanding and rectifying biases in algorithms. To become aware of the common human biases is the prerequisite to eliminate fairness from any machine learning algorithm.

One of the examples that this research paper will further assess in the realm of machine learning fairness is sparked by the COMPAS controversy. COMPAS, one of the leading decision-support system tools is used by the court system to determine the likelihood of the offence being committed again – which takes an extreme quantitative approach. This is an example where "fairness" needs to be implemented in the algorithm so that quantitative criteria are not the only reason for an individual to be convicted.

The diverse types of fairness include unawareness, demographic parity, equalized odds, predictive rate parity, individual fairness, and counterfactual fairness. Demographic parity is one of the major types of fairness that is selected as a criterion while measuring the success of a machine learning algorithm. Why is measuring fairness and bias so essential for a machine learning engineer or researcher? Because the designed systems determine vital decisions for society – decisions that can make or break lives. From a varied range of loan applications, credit card approval applications, legal, judicial, and corporate matters – machine learning models play

an integral part in the public sector. Therefore, it is critical to mitigate through these fairness elements to achieve equitable and accurate predictions from any machine learning model.

The following are some of the tests that can be conducted to measure the fairness level and detect bias in the algorithm, dataset, or both.

- Association tests – This test allows machine learning engineers to measure the association between two sets of words for prejudice and bias.
- Perturbation tests – This test determines the correlation between input and output element in the model.
- The last step would be to use strategies like pre-processing, in-processing, and post-processing to reduce and eliminate biases in an algorithm.

What does the future of Machine Learning Bias and Fairness look like? When you run a machine learning system – the future is determined by the past inputs and elements. Thus, including bias and fairness associated with it; and resulting in inaccurate predictions. Firstly, the mindset of eliminating bias from the start will be implemented while creating any machine learning program. Fairness through awareness would be one of the major movements in the AI society in the next few years, where biases can be eliminated just by being conscious of your data and algorithm procedures.

With machine learning models being implemented in the judicial and healthcare systems for the public – it is imperative for those models to produce results with not only higher accuracy but results which are free from any type of bias. To maintain that there is a constant need to use tools like AI 360, and Google Fairness Metric tools that can allow us to detect and mitigate fairness

throughout any model. Machine learning models are set to take over future innovations – ranging from clinical systems, healthcare, robotics, automation, computer vision, and personalization.

In conclusion, artificial intelligence, machine learning, and deep learning technology innovations are going to change the world step by step. But the type of impact it has on society is in our hands, and it is important to ensure that there is an ethical boundary with every machine learning practice swept into society. As we proceed, testing on real data is imperative to judge our model's accuracy thus shifting focus on metrics as needed. This topic can be seen from multiple perspectives, and through this research paper – I have highlighted the importance of fairness metrics in an algorithm while portraying the benefits of machine learning models in society.

References

Abhishek Tiwari. (2017, September 1). *Bias and fairness in machine learning*. Abhishek Tiwari. Retrieved December 9, 2022, from https://www.abhishek-tiwari.com/bias-and-fairness-in-machine-learning/#:~:text=In%20AI%20and%20machine%20learning,algorithms%20learn%20from%20training%20data.

Guo, Lee, M. S., Kassamali, B., Mita, C., & Nambudiri, V. E. (2022). Bias in, bias out: Underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection—A scoping review. Journal of the American Academy of Dermatology, 87(1), 157–159. https://doi.org/10.1016/j.jaad.2021.06.884

Hernandez, & Crawford, K. (2019). The social implications of machine learning (Hernandez, Interviewer). Dow Jones & Company.

Losavio. (2021). Algorithms of Machines and Law: Risks in Pattern Recognition, Machine Learning and Artificial Intelligence for Justice and Fairness. *Public Governance, Administration and Finances Law Review*, *6*(2), 21–34. https://doi.org/10.53116/pgaflr.2021.2.3

Mehrabi, Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys, 54(6), 1–35. https://doi.org/10.1145/3457607

Mosteiro, Kuiper, J., Masthoff, J., Scheepers, F., & Spruit, M. (2022). Bias Discovery in Machine Learning Models for Mental Health. Information (Basel), 13(5), 237–. https://doi.org/10.3390/info13050237

Obermeyer, Ziad, and Ezekiel J. Emanuel. "Predicting the Future — Big Data, Machine Learning, and Clinical Medicine." The New England Journal of Medicine, vol. 375, no. 13, 2016, pp. 1216–19, https://doi.org/10.1056/NEJMp1606181.

Prates, Avelar, P. H., & Lamb, L. C. (2020). Assessing gender bias in machine translation: a case study with Google Translate. Neural Computing & Applications, 32(10), 6363–6381. https://doi.org/10.1007/s00521-019-04144-6

Reagan, M. (2021, April 2). *Understanding bias and fairness in AI Systems*. Medium. Retrieved December 6, 2022, from https://towardsdatascience.com/understanding-bias-and-fairness-in-ai-systems-6f7fbfe267f3

Sahani, G. R. (2020, October 29). *Elucidating bias, variance, under-fitting, and over-fitting.* Medium. Retrieved December 9, 2022, from https://medium.com/analytics-vidhya/elucidating-bias-variance-under-fitting-and-over-fitting-273846621622#:~:text=Specifically%2C%20underfitting%20occurs%20if%20the,and%20variance%20in%20simpler%20terms.

Sarkar, D., Bali, R., Sharma, T. (2018). Machine Learning Basics. In: Practical Machine Learning with Python. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-3207-1_1

Srinivasan. (2020). Busting the myth of "neutral" AI. Big Think.

Sun, Nasraoui, O., & Shafto, P. (2020). Evolution and impact of bias in human and machine learning algorithm interaction. PloS One, 15(8), e0235502–e0235502. https://doi.org/10.1371/journal.pone.0235502