Zhaowen Zhou
Boston University
EC601: Project 4

Machine Learning: Data Bias

Since the term machine learning was first coined in 1959, this concept and its application has penetrated into every aspect of people's life. Especially, with the improvement of the computer hardware and software, sharp advantages of machine learning are clearly showed: machine can work day and night without getting tired, machine learning can take huge amount of data into consideration and find out some potential patterns hidden inside the data. Due to these advantages, the machine learning algorithm plays a huge role in decision making now, even in making life-changing decisions. However, when talking about making decisions that might affect many people's lives, we must ask ourself: doesn't machine learning have any drawbacks or risks, and can it be absolutely objective? The answer is no. Because of the heavy reliance on data, the output of machine learning algorithm can be extremely influenced by the bias within the data. It becomes the responsibility of developers of data systems and anyone involved in the decision-making process to find out the data bias and address it.

The concept of data bias is that the available data is not representative of the general population or does not express the true nature of the studied phenomenon. A Biased data can lead to an unfair result, which means that prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics.

There are two sources of bias. The first one is bias of data and the second one is the bias of the algorithm. This essay will more focus on the bias from data.

[1] The main reason that there are biases within the data is because of the heterogeneities of the data. An easier way to understand the heterogeneity is that the heterogeneity can be regarded as the differences among subgroups. For example, when trying to figure out the relationship between the consumption of pasta and BMI, data from many people must be included, the fitness level of everyone from this group varies – there are people who doesn't exercise and there are people who exercise a lot. When these human data are classified according to different criteria, the results will be different. This is a simple example of where the bias within the data is come from.

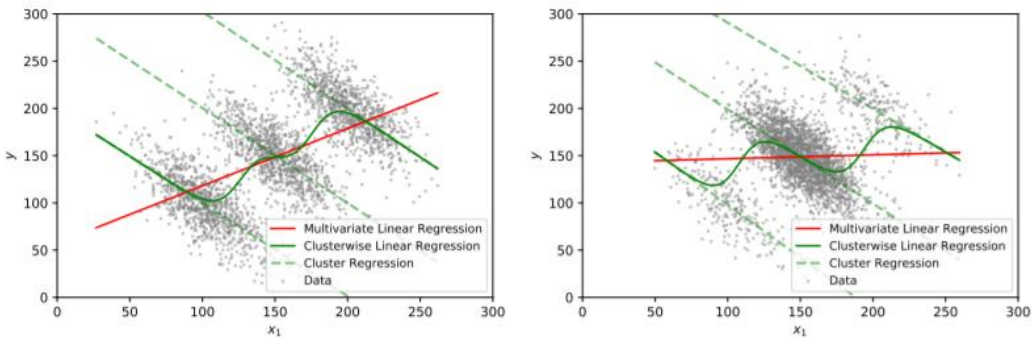

Fig.1 Illustration of bias in data

To better understand and address the bias within the data, the researchers have divided bias into several categories. Some of the bias will be discussed.

Aggregation Bias

Aggregation bias happens when false conclusions are drawn for a subgroup based on observing other different subgroups or generally when false assumptions about a population affect the model's outcome and definition.

Fig2. Tweets containing emojis

[2] A typical example of aggregation bias is the meaning of some hashtags or emojis when trying to use NLP to analyze the meaning of some tweets. In some tweets, some words will be replaced by emojis. For example, in a tweet about street life and gang, the word 'clapped' means 'to shoot with a gun' and this 'clapped' will be replaced by the emoji of clapping hands. If the NLP tool is an untrained model, it cannot understand the emoji "clapping hands" as "to shoot with a gun", and within the same tweet, the emoji "shit" also appears. In a tweet about street life, in this "shit" can be understand as in the street life, but an untrained model cannot understand this emoji as well.

So, a tweet which its original meaning is "Anybody could be shot in these streets and in this gang life, do not get caught unprepared without a gun" can be translated to "Anybody could clap their hands in this shit, do not get caught unprepared without a gun." This sentence obviously has no meaning.

Imagine this tweet is one of the input data, the false assumption (the meaning of emojis) made by the NLP model can influence the final result.

Historical Bias

Historical bias is the already existing bias and socio-technical issues in the world and can seep into from the data generation process even given a perfect sampling and feature selection. In other words, this kind of bias is originated from the society we live in and quite hard to avoid.

This kind of bias is very easy to understand. When searching the word "troops" on Google image, most of the soldiers are male, however, there are female soldiers and the female soldier is not uncommon. Although the fact that the army is mainly composed of male soldiers is undisputable, it still remains careful consideration that whether the input data should reflect that.

Population Bias

Population bias arises when statistics, demographics, representatives, and user

characteristics are different in the user population represented in the dataset or platform from the original target population. This kind of bias can easy be understood.

According to a result conclusion of paper, [3] Hispanic students are significantly more likely to use MySpace than are Whites, while Asian and Asian American students are significantly less likely to use MySpace, but Xanga and Friendster instead. From the angle of the education background of users' parents, students whose parents have lower levels of schooling are more likely to be MySpace users, while the other students whose parents have a higher educational background prefer Facebook.

Sampling bias

Sampling bias arises due to non-random sampling of subgroups.

This kind of bias means that it is possible to bring the trend of a population to the whole population. Keep considering the example of population bias. Considering researchers want to analyze the change of people's moods through social media during the lockdown. The researchers collect a lot of texts from different social media: MySpace, Facebook, Twitter, etc. Then when sample the data, researchers did not collect content equally across social media but mainly focus on the content from twitter. Then, the result that is said to be an average result of different social media is actually mainly the result of twitter's users, which means that this biases the result of analyze.

There also lots of biases, all biases can be related to the process of machine learning. Different steps in the process of machine learning produce different bias.

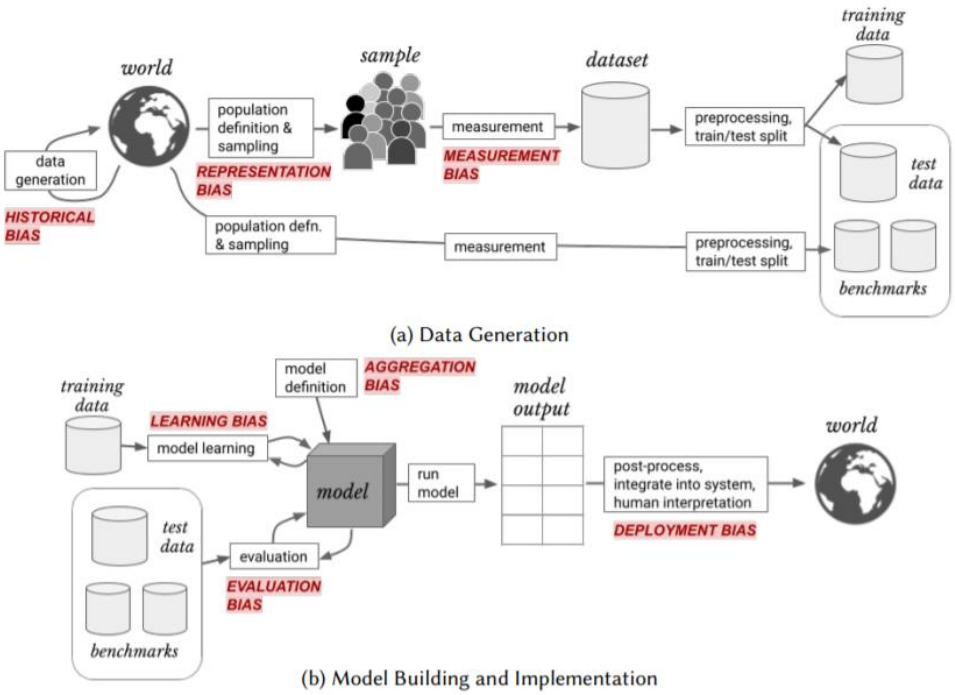

Fig3. Machine learning process and its biases

[4] This figure can show biases and their matching generating steps.

Methods for fair machine learning

The process of machine learning often has three parts: The first part is preprocessing the data, the second part is training the model or implement the algorithm, in-processing in short, the third part is validation, which can be also called post-processing. Thus, when trying to mitigate the bias, the mitigation method can also be done through these three parts.

Researchers have already proposed many methods, such as having datasheets as a supporting document, which reports the sampling method, motivations and aims for sampling this dataset, having labels in order to categorize data. Also, casual graphs can be used to visualize the data discrimination straightforwardly.

Due to the importance of bias mitigation, there are now encapsulated tools to handle this kind of problem.

The AI Fairness 360 is a tool created by IBM trusted AI team to find and address the bias with fairness metrics and bias mitigators. The running environment is python.

One of the fairness metrics is called mean_difference. To evaluate through this metric, we manually select privileged feature and unprivileged feature, then compare the percentage of favorable results for the privileged and unprivileged groups, subtracting the former percentage from the latter. If the answer is negative, then it indicates that less favorable outcomes for the unprivileged groups. This is unfavorable, which means that mitigation method is needed. The next step is to choose a mitigation method. Since this kind of bias is created with the collection of data, we need to choose a pre-processing bias mitigation method. One method is called reweighing. This algorithm will transform the dataset to have more equity in positive outcomes on the protected attribute for the privileged and unprivileged groups.

As we can see from last paragraph, there are lots of fairness metrics (disparate impact, average odds difference, statistical parity difference, equal opportunity difference, and Theil index) and bias mitigators (reweighing, prejudice remover, and disparate impact remover). In order to use these tools wisely, it is important to be problem specific and find the appropriate rubric and methodology for the data set.

For example, first, where does the bias come from needed to be figure out, then choose the matching method. Sometimes, both pre-processing methods and in-processing methods can be used to mitigate bias, and different machine learning methods can be used as well. In this situation, different evaluation metrics can be used to compare and select the best result.

Conclusion

In this essay, the most essential reason for bias is heterogeneity is concluded, and the different bias are introduced according to the process of machine learning, and specific examples of each type of bias are given. Finally, we introduce AI Fairness 360, a tool for discovering bias and mitigating bias.

Reference

[1]. Mehrabi N, Morstatter F, Saxena N, et al. A survey on bias and fairness in machine learning[J]. ACM Computing Surveys (CSUR), 2021, 54(6): 1-35.

[2]. Frey W R, Patton D U, Gaskell M B, et al. Artificial intelligence and inclusion: Formerly gang-involved youth as domain experts for analyzing unstructured twitter data[J]. Social Science Computer Review, 2020, 38(1): 42-56.

[3]. Hargittai E. Whose space? Differences among users and non-users of social network sites[J]. Journal of computer-mediated communication, 2007, 13(1): 276-297.

[4]. Suresh H, Guttag J. A framework for understanding sources of harm throughout the machine learning life cycle[M]//Equity and Access in Algorithms, Mechanisms, and Optimization. 2021: 1-9.