Data Bias

Introduction

As rapidly development of artificial intelligence, the use of machine learning continues to expand and affect our lives gradually. We would be exposed to more and more applications based on machine learning in the future that can be met. For example, the recommendation system, basing on our preferences and product browsing data, finds the products customers are most likely to buy, which increases the company's sale. For another example, in the transportation industry, the machine learning model is used to fit the past car flow data to obtain the optimal arrangement for traffic lights. Under such circumstance, the new term "data mining" is invented and even become an attractive major in most university. There are so many applications out there. It's hard for us to realize, but they are everywhere.

Definition

The question also follows, is the machine learning model reliable? Will there be prejudice that affect our judgement? The answer is no. Data bias is invented to precisely describe this phenomenon. Data bias refers to the unfair output of the machine learning model due to the unbalanced data set which is used for training. A well-known example is COMPAS, mentioned in[1], which is a widely used commercial risk assessment software. COMPAS predicts the likelihood that a defendant will re-offend and may be used by judges or parole officers to make decisions around pre-trial release. However, the software was misled by the bias dataset which includes more minority communities' cases. So, the resulting model had a significantly higher false positive rate for black defendants versus white defendants. Another example mentioned in[2] is the disability problem. Disability is difficult to quantify. It has many dimensions and people can experience multiple disabilities. Even with the data included, there may not be enough individuals with a given type and severity of disability in a dataset for the machine to identify a pattern. It would be neglect as an outlier.

Reasons

[1][4] summarizes the reason for the data bias. The problem caused by data consist of many aspects. And here are five main areas.

1) Historical Bias. Historical bias is the already existing bias and can seep into from the data generation process even given a perfect sampling and feature selection. For example, the number of women in the stem industry is obviously less than men, so fewer women data would be included such dataset. This unbalanced problem should be considered in the model.
2) Representation Bias. Representation bias occurs when the development sample under-represents fails to generalize well for a subset of the use population. Lacking geographical diversity in datasets like ImageNet is an example for this type of bias.
3) Measurement Bias. Measurement bias happens from the way we choose, utilize, and measure a particular feature. COMPAS mentioned above is a good example for this kind.
4) Evaluation Bias. This includes the use of inappropriate and disproportionate benchmarks for evaluation of applications. This happens during model evaluation and leads to unfair estimates of the model.

5) 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.

Approach

Fairness is an increasingly important concern as machine learning models are used widely. Researchers and industry developers desperately need a method that can solve the data bias. Nowadays, the most famous open-source project is the AI FAIRNESS 360[3]. This toolkit is to help facilitate the transition of fairness research algorithms to use in an industrial setting and to provide a common framework for fairness researchers to share and evaluate algorithms. In order to facilitate the operation, several classes are abstracted in AIF360.

1) Dataset Class. Dataset class which is a key abstraction that handle all forms of data. To maintain a common format, independent of what algorithm or metric is being applied, dataset class is structured so that all these relevant attributes — features, labels, protected attributes, and their respective identifiers are present and accessible from each instance of the class. So, whether it is training dataset, testing dataset or output dataset, any dataset can be represented by this class.
2) Metric Class. Metric Class computes various individual and group fairness metrics to check for bias in datasets and models, such as BinaryLabelDatasetMetric and SampleDistortionMetric.
3) Explainer Class. The Explainer class is intended to be associated with the Metric class and provide further insights about computed fairness metrics.
4) Algorithm Class. AIF360 currently contains 9 bias mitigation algorithms that span these three categories according to the stage in which they work. The first is pre-processing algorithms work before the dataset is sent into the model. And the next is the in-processing algorithms, such as adversarial debiasing and prejudice remover. The last is the post-processing algorithms.

By using the above four abstract classes, users can significantly reduce data bias and achieve better model performance.

Conclusion

In fact, data bias is the main challenge faced by machine learning methods today, directly affecting the performance of the model, and even our final judgment, such as COMPAS. There are many reasons for data bias, from data sampling, generation, to evaluation, which covers all aspects of dataset. Finally, this article sorts out a method for reducing data bias, AI fairness 360, which has a very positive impact on researchers and industry developers.

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] Trewin S. AI fairness for people with disabilities: Point of view[J]. arXiv preprint arXiv:1811.10670, 2018.

[3] Bellamy R K E, Dey K, Hind M, et al. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias[J]. arXiv preprint arXiv:1810.01943, 2018.

[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.