Paper review

Paper title: A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle

Paper link: <https://arxiv.org/pdf/1603.02754v3.pdf>

Paper abstract: As machine learning (ML) increasingly affects people and society, awareness of its potential unwanted consequences has also grown. To anticipate, prevent, and mitigate undesirable downstream consequences, it is critical that we understand when and how harm might be introduced throughout the ML life cycle. In this paper, we provide a framework that identifies seven distinct potential sources of downstream harm in machine learning, spanning data collection, development, and deployment. In doing so, we aim to facilitate more productive and precise communication around these issues, as well as more direct, application-grounded ways to mitigate them.

Review:

1. Introduction: The introduction does a good job of outlining the key areas where harm could be formed when creating ML algorithms - data collection, model development, and deployment processes. It also gives examples of cases where algorithms trained on public data caused issues related to racial bias. The second paragraph emphasizes the complexity of data gathering and the oversimplification when people talk about “biased training data” as the sole cause for the poor performance of a model. The next paragraph expands on the possible sources of bias – different from the data itself. The paper then provides examples of at first glance similar cases which are indicative of different issues with ML algorithms. The rest of the introduction talks about the structure of the paper and some of the key findings.
2. MACHINE LEARNING OVERVIEW – this section starts with some basic introduction to where ML is applied. It then proceeds to outline the different sections of the ML project namely:
   1. Data collection – it starts with identifying a “Target population” also what “Features” and “Labels” we can have from it. Worth noting that we do not use the whole data – just a sample for it. There is also an example given with a loan company and what data they might be interested in.
   2. Data preparation – it relates to the preprocessing of data – filling in missing data, simplifying the features space etc. This is also the step where the data is split into testing and training data. Part of the training data can be used for validation.
   3. Model development – Here the paper talks about what the goal of the model is – optimizing specific objective. Hyperparameter optimizations are also part of this step.
   4. Model evaluation – After choosing the model, this step comes into play, where we test the model using the test data. It is also suggested that we can use benchmark datasets to further demonstrate the model’s robustness.
   5. Model Postprocessing – after training and testing, it is sometimes necessary to do additional steps, like converting % certainty of a binary classifier to categories depending on certain thresholds.
3. Model Deployment – The final step also holds nuances. For example, the need for explainability may change the already prepared model, or the data the model was trained on may be different from the real-world data in some way.
4. SEVEN SOURCES OF HARM IN ML – this section builds on the previous section with the different stages of ML development to identify where and how sources of harm or different types of “Bias” could occur. The paper also aims at breaking down biases into more concrete terms like “measurement bias” or “historical bias”.
   1. Historical bias – this bias relates to the issue of reinforcement of stereotypes, which, although could produce accurate algorithms, may be harmful to particular groups. The section then gives examples with embedding algorithms trained on data from specific periods where terms like “nurse” and “engineer” might be associated with a particular gender more.
   2. Representation bias – this type of bias occurs when the training and testing data do not accurately represent all parts of the population and fail to generalize for the real-world data. The section then breaks down where this Representation bias could occur:
      1. When defining the target population, if it does not reflect the use population.
      2. When defining the target population, if contains underrepresented groups.
      3. When sampling from the target population, if the sampling method is limited or uneven.

IMPORTANT: The paper then gives an example with ImageNet, which is intended to be used widely but predominantly contains natural images from United States, North America or Western Europe. Further studies have shown that algorithms trained on this dataset struggle with tasks related to classifying images from underrepresented regions like Pakistan or India.

* 1. Measurement Bias – This bias refers to the mismatch between the target concept and how it is measured. An example is given with “creditworthiness” as abstract construct, measured with proxies like “credit score”. The bias can occur in these areas:
     1. The proxy is an oversimplification of a more complex construct.
     2. The method of measurement varies across groups.
     3. The accuracy of measurement varies across groups.
  2. Aggregation bias – This type of bias relates to cases where varying data is being clumped together, when it should be considered differently. It works on the assumption that mapping from inputs to labels is consistent across subsets of the data.
  3. Learning Bias – This type of bias arises during model training and is related to the specific objective, optimized by the model. It is important because it shows how optimizing for accuracy could have a negative impact on the “fairness” of the model (disparate impact). In the given example it is mentioned that techniques like “Pruning” force the model to preserve information about the most frequent features and amplifies the disparities in the unrepresented attributes.
  4. Evaluation Bias – This bias relates to the evaluation of the model on benchmark data (which may not relate to real world data). It arises from the desire to quantifiably compare models with one another. This type of bias can lead to overfitting for a specific benchmark. This is problematic if the benchmark itself has some of the previously mentioned biases. Lastly, another issue arises when inappropriate evaluation metrics are chosen for a given model.
  5. Deployment Bias – This bias occurs when a system designed for one specific task is placed in a larger frame which does tasks it was not originally designed for. The risks here come from phenomena like “confirmation bias” and “automation”.
  6. Identifying sources of Harm – the last subsection goes over the previously mentioned biases and provides contexts for all of them. It also suggests that knowledge of a model’s intended use and context is required to assess the possible sources of harm.

1. FORMALIZATION AND MITIGATIONS
   1. Formalization of the framework – this section is slightly more technical and just puts labels on the features and labels datasets and subsets.
   2. Designing Mitigations – The section starts with a recognition of a growing trend to design more “fair” algorithms but also acknowledges that “fair” or the notion of “fairness” can be different depending on how it is defined in different literature. The rest of the section goes over each bias and places it as part of the previously established framework.
2. Conclusion – just a conclusion of the framework and the previously discussed topics. It does not provide areas for further research, although there are plenty of references for further reading.