MIT 6.871/HST.956 Spring 2021
Problem Set 4

- Release: April 8, 2021 (Thursday)
- Due: April 21, 2021 (Wednesday) 11:59pm EST
- Please submit your final write-up via Canvas/Assignments/Pset 4/Gradescope

Collaboration Policy

- Students must write up their problem sets individually. Students should not share their code or solutions (i.e., the write up) with anyone inside or outside of the class, nor should it be posted publicly to GitHub or any other website. You are asked on problem sets to identify your collaborators. If you did not discuss the problem set with anyone, you should write ”Collaborators: none.” If in writing up your solution you make use of any external reference (e.g. a paper, Wikipedia, a website), both acknowledge your source and write up the solution in your own words. It is a violation of this policy to submit a problem solution that you cannot orally explain to a member of the course staff.
- Plagiarism and other dishonest behavior cannot be tolerated in any academic environment that prides itself on individual accomplishment. If you have any questions about the collaboration policy, or if you feel that you may have violated the policy, please talk to one of the course staff.

1 Interpretability for Chest X-Ray Diagnosis Models [30 points]

Overview

In this problem, we will work with a subset of the MIMIC-CXR dataset, distributed through PhysioNet as in previous PSets. We will present you with a simulated experience of evaluating a potential model for deployment. In this case, the model in question (which we will provide to you, fully trained) is being evaluated for its use in the detection of Pneumothorax (PTX); a potentially dangerous condition often colloquially known as a collapsed lung. Ultimately, your desired use case for this model is to use it as a diagnostic aid for new patients to the ED.

However, being an experienced MLHC researcher, you're concerned that perhaps this model is not actually leveraging viable diagnostic information to make its PTX decision–instead, it may be confounded by other factors in the data. You'll leverage 4 interpretability techniques to assess this risk:

1. Error Auditing, in which you'll examine different kinds of errors the model makes.
2. Interpretability Visualizations, in which you'll implement two algorithms for visualizing the salient regions of the image for a CNN model, then use those to examine a subset of the images.
3. Multimodal Analysis, in which you'll use the associated free-text reports to do a targeted investigation of image confounders, and,
4. Error Stratification, in which you'll implement a (very simple) classifier over the reports in order to split images according to our confounder of concern, then examine the model's performance when the confounder is absent.

Ultimately, in each of these sections you'll be asked to comment on any strategies you see, and, at the end, answer several concluding questions.

Learning Goals

Interpretability is often touted as being critical in machine learning for healthcare for its use in improving the trust of downstream end-users in the model; however, this overshadows what may be an even more important role of interpretability / explainability in ML4H – namely, its use for creators of models to gain confidence in their model's internal and external validity. Here, we task you with using interpretability methods for exactly this purpose, in a real-world setting of examining whether or not a trained model that performs well according to traditional metrics is really ready for deployment. In addition, other learning goals include:

- To think through possible confounders of a model designed for clinical use.
- To gain familiarity with two basic forms of CNN visualization.
- To demonstrate the power of leverarging multimodal information for interpretability.
- To question the accuracy of automatic labels / gain exposure to label noise.
- To gain an introduction to CNNs/CV/radiograph data.

Code and Write-up Questions

Code and write-up questions can be found here:
https://colab.research.google.com/drive/1_Uv8c7AqbeFYnwnKgyeZRCVt4-wSEnGR.

You will be required to turn in both your final .ipynb colab file and include in your single separate writeup the answers to all questions explicitly asked in the ”Write-Up Questions” sections.

2 Fairness in Machine Learning Algorithms for Health [30 points]

Overview

Fairness in machine learning is a recently established area that studies how to ensure that biases in the data and machine learning model do not lead to algorithms that treat individuals unfavorably on the basis of characteristics such as race, gender, disabilities, and sexual or political orientation. It is particularly concerning when biases arise from machine learning applications in medicine due to its consequential impact on human health and life.

In this problem, you will read a research paper that formulates and analyzes racial bias in commercial prediction algorithms for health care. You are expected to appreciate their mathematical formulation of this problem and come up with your solution to address the bias in machine learning for health.

Write-up Questions

Please read this article: https://science.sciencemag.org/content/366/6464/447 and answer the following questions.

1. In the *Data and Analytic Strategy* section, the article proposes that $\mathbb{E}[Y|R,W] = \mathbb{E}[Y|R,B]$ indicates the absence of racial bias, where Y is a variable of interest, R is the risk score predicted by a machine learning algorithm, W means being racially white, and B means being racially black.

   (a) To examine the presence/absence of racial bias, please give three examples of what Y can be. [5 points]

   (b) Can $\mathbb{E}[R|W] = \mathbb{E}[R|B]$ indicate the absence of racial bias? Briefly explain your answer. [5 points]

2. In the *Mechanism of Bias* section, when the variable of interest is healthcare cost C, we have $\mathbb{E}[C|R,W] = \mathbb{E}[C|R,B]$ based on the authors' empirical analysis, where C is the health care cost, R is the risk score predicted by a machine learning algorithm, W means being racially white, and B means being racially black.

   (a) Why does solely $\mathbb{E}[C|R,W] = \mathbb{E}[C|R,B]$ NOT indicate the absence of racial bias? [5 points]

   (b) The manufacturer of the algorithm examined in this article chose to predict future costs as part of their machine learning training objective. How may it *bias* the resulting algorithms? [5 points]

3. This paper demonstrates a series of experiments to gain insight into how label choice affects both predictive performance and racial bias. How would you design a machine learning approach that minimizes racial bias? Please briefly describe the data (including labels) you would like to collect to

train the machine learning model, the model, the objective functions. Please also discuss if and/or what you may have to compromise (e.g., predictive performance) in order to minimize racial bias. Please refer to and use the mathematical formulation in the *Obermeyer et al.* article when you describe your solution. [10 points]