Problem Set 4 (46 Points)

MLHC 2023

April 7, 2023

Submission Instructions

Please submit your write-up and code to Gradescope. Please print your notebook and append to the end of your report When you write up your report, put all writing into one file, furthermore make sure your report follows the same structure as the problem set. You must write up their problem sets individually. You should not share your 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.

Late Day policy:

- (4 "slack" days) We understand that sometimes things outside one's control prevent submitting by the deadline. As such, each student is given 4 "slack" days that they can use throughout the semester (e.g. you could submit two psets one day late each or you could submit one pset two days late) without a late penalty. The days do not subdivide into sub-day units: 2 hours late would spend one of the slack days without 22 hours of "rollover". In your pdf writeup, specify how many slack days you are using (they cannot be used retroactively).
- (10% off per unexcused late day.) If you submit a pset a day late without slack days,
- (max 4 days late.) Homework will not be accepted after 4 days late, regardless of how many slack days are used, absent communication from S3 or OGE.
- (write on homework) In order to use a slack day, students must include it in writing on their submission pdf. Otherwise, TAs will assume no slack days used and deduct 10% for each late day.

In this problem set, 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.

Colab notebook https://colab.research.google.com/drive/14ngqTUsjgIt3FrRcKhBzmHVi8HKSDrmy?usp=sharing

1 Error Auditing [12 points]

In this first problem, we will try to understand the data and the performance of our model on a high level.

1.1 Compute Accuracy and AUC on the evaluation set. Do you think this AUROC is good, or bad? If you saw this on a model used to treat patients, how would you feel about that? (3 points)

In the notebook, a pandas dataframe named 'evaluation_set' contains a column 'Pneumothorax' with the labels, use that and the output of "eval_model".

1.2 Describe but do not implement one way to compute confidence intervals for each of AUC and accuracy (2 points). Bonus (+2) points: if you implement and compute the confidence intervals correctly.

Feel free to look up different methods online and briefly summarize them.

1.3 Compute the confusion matrix of our classifier (FP,FN,TP,TN) and the class balance (% of labels that are 1). Compute rates instead of counts (i.e. normalized confusion matrix) (2 points)

1.4 Investigate the chest xrays and reports in the evaluation set, the "display_study" function is helpful here. Report some of your observations on the variation of images. Report some of your observations on the notes with pneumothorax. Do you notice any patients who were labeled as pneumothorax but who might not be so? Can you say why the label is incorrect and contradicts the report? (5 points).

In this question we are not looking for specific observations, but want to see an effort in you looking at the data.

2 Interpretability Visualizations [15 points]

One serious concern is that the model may not be diagnosing Pneumothorax at all, but instead largely relying on a treatment-based confounder, such as the presence of a "pigtail cathater", a small-bore tube used in the case of pneumothorax to allow reinflation of the lung or escape of air from the pleural space (note that pigtail cathaters have numerous other uses as well, but this is the use case of interest to us here), or a chest tube. Both of these are very evident on an x-ray, and are strong indicators that the patient did at one time have Pneumothorax; however, predicting them offers no diagnostic benefit to us as they are only seen if the Pneumothorax is already known and under treatment. For the rest of our analyses, we'll focus on trying to determine if this is the case or not. First, we'll look at interpretability analyses via Saliency maps and Class Activation Maps.

2.1 Implement Saliency Maps. In the code, the body of the saliency map is empty but there are TODOs we have written for you to follow. Please also copy the code here for your solution (10 points)

2.2 Visualize different images. Use the 'plot_interpretability_measures' to plot images with both intepretability viz methods for images that are true positives and images that are true negatives. Do regions highlighted by the saliency map and CAM suggest that the model may be largely leveraging treatment devices, such as pigtail catheters and chest tubes, rather than diagnostic criteria for predicting pneumothorax? You may also look at FN and FP cases. [5 points]

3 Leveraging Radiology Reports [19 points]

Another oft-understated tool to understand a model's decision making process is to leverage additional, related modalities of data to better explore the underlying data and determine if there are any unexpected confounders in the data. Here, we'll leverage the radiology reports co-released with these images in MIMIC-CXR to do just that.

3.1 Using the reports in evaluation set, report the AUC and accuracy of the model by gender and the count by gender (how many in each gender). You will need to extract gender from the reports, if for a report you are unsure of the gender, label it as unknown and evaluate as its own category. [4 points]

3.2 read below for 3 parts [10 points]

1) Find the top 10 words that appear the most in the reports that have pneumothorax compared to the reports that are labeled 0.
To do this, for each word that exists in the positive reports, find how many times it appears in total, then subtract how many times it appears in the negative reports.
2) Do the same thing but now for the top 10 words that appear the most in reports without pneumothorax compared to reports with.
The CountVectorizer from sklearn with unigrams will be useful: "cv = CountVectorizer(ngram_range=(1, 1), binary = True)".
3) What do you note? Do these terms appear to be diagnostically relevant to Pneumothorax? Or are they more about other medical concepts? Are there any relationships related to treatments, as we fear?

3.3 Evaluate the AUC of the model on the reports that include either of the words "drain", "pigtail", "catheter" or "tube". Then evaluate the AUC of the model on reports that don't include all of the four words. What does this suggest about what is guiding our model's performance? [5 points]