MIT HST.953 HW 3

Rohan Doshi, Yanran Li, Walt Williams, Rob Lewis, Kexuan Liang

Assignment: https://canvas.mit.edu/courses/16659/assignments/206986

Problem 1. Dealing With Data [6 points]

(a) [2 pts] The data you are working with for the project has already been collected and processed. However, in a clinical setting, there is often a delay between when the data is collected and when it is entered into the system. How does your model deal with that delay? Does it still achieve its purpose even with the delay, or would you need to work around the delay for the model to be effective? If you will need to work around it, how will you do so?

The model needs the X-ray scans and certain metadata in order to be implemented in a real-time fashion. We agree with the phenomena that a delay between data being collected and entered into the system in a clinical setting. Our model dealt with that delay by trying some pre-trained model and utilized data augmentation during the data cleaning part. Since we haven't utilized our model in an interactive application yet, we still need more test data to check whether our model still achieves its purpose even with the delay. If we need to work around it, we may try training our model sequentially and we may also try some other ensemble models to ensure its reliability and robustness.

(b) [2 pts] The data you are working with has its own distribution, but in a clinical setting you may have outlier data points that fall far beyond the tails of that distribution. How do you detect and handle outliers in such a setting?

One could apply statistical measures to check for outliers and warn the clinician accordingly so they can adjust their confidence in the model predictions. Suppose your input features are real values and you assume the data follows a normal distribution. From the training data, you could calculate the mean and standard deviation. During inference in a clinical context, after you normalize the data and check whether it is more than 3 standard deviations from the norm. If so, the clinician should get a warning about the prediction and know they should not be as confident in the prediction since they are dealing with an outlier the model may be unable to handle.

Also, if there are many outliers during inference, one could improve model robustness with data augmentation during training. For example, our machine learning model expects X-rays with a vertical orientation during training. But during inference, it is possible that an X-ray could be inputted with a slight tilt - an outlier. In that case, we'd want to make sure our model knows how to handle tilts in inputs, so we could apply data augmentation. In our case, we randomly apply a tilt permutation during training with some probability to improve model robustness to outliers.

(c) [2 pts] The distribution of your data might change over time, and if your model is not calibrated for this change, it might start to perform below expectations. How do you expect to handle this data drift? Briefly explain the process you would follow.

Periodic model retraining is the best way to handle model drift over time. I'd start by implementing a dashboard that measures input and prediction statistics (mean / standard deviation) over time. As these metrics drift and hit certain trigger thresholds (e.g. 5% performance drop), it should retrigger a model-retraining on newly collected data. Older data could either be discarded from the new training dataset or down-weighted in the loss function. Ultimately, continuous model retraining, say every week, ensures the model learns a data distribution that is consistent with the latest population it is serving at the time.

Problem 2. Model Evaluation [6 points]

(a) [2 pts] What metrics will you use to evaluate your model? Will you use the same metrics to evaluate the model after it is deployed?

The metrics we're using are precision, recall, f1-score, accuracy, and AUC. Given the model's intentional use as a diagnostic aid for detecting congestive heart failure (CHF) with reduced or preserved ejection fraction we believe that all of these metrics help paint a holistic view of the model's performance during training. We're placing extra emphasis on recall, f1-scores and AUC given that the both classes in the dataset are deadly if not caught. Our goal is to minimize the amount of false negatives since in practice a false negative could result in deadly consequences for a person who is misdiagnosed.

(b) [2 pts] Who will be using the output of your model? How will you display the model's output to them?

This model is intended to be used by medical practitioners in a hospital setting. Since the end users for the model will most likely be unfamiliar with machine learning we intend to display the model's probabilistic predictions along with an image of the x-ray and heatmap visualization of the output. This way the doctors will be able to get a simple and intuitive explanation behind the model's predictions and hopefully build trust and confidence in the model.

(c) [2 pts] How do you evaluate whether your model is more effective than the currently deployed system? Do you expect to run some kind of A/B testing? If yes, describe how you will evaluate which deployed system is more effective. If not, describe the process you will use to decide on the better system.

We could evaluate our model against a currently deployed system in 2 stages. To get an initial estimate of how our model performs we could sample a set of x-ray images from patients previously seen by the hospital and collect predictions for the sample. We would then simply compute our test metrics for each set and see which model performed better. Afterwards we could deploy our model for a trial period, say 3 months, alongside the already deployed system and at the end of the trial period collect qualitative feedback from practitioners about the model's

trustability, explainability, and usefulness in their day-to-day practice compared to the already deployed system.

Problem 3. Social Implications of Model Deployment [4 points]

(a) [2 pts] Who are the stakeholders for your project, and how will they be affected by the deployment of your model?

Stakeholders are any group of people that can have an influence or be influenced by our project. The internal stakeholders for our project are our project team and collaborators, external stakeholders for our projects are clinicians and customers (such as patients), health executives, researchers and regulators.

Influence on project team and collaborators: deliver AI model into production of clinical values.

Influence on patients and clinicians: enhance the experience of healthcare practitioners, facilitate direct patient care and reduce burnout and lead to better care outcomes in patients.

Influence on health executives and regulators: increase productivity and efficiency of health care delivery and allow healthcare systems to improve quality care to more people.

(b) [2 pts] Will you need to involve the IRB to deploy your project? If yes, briefly explain what you will need to do to get IRB approval. If not, briefly explain why your project does not need IRB approval.

We do not need to involve the IRB to deploy our project, because the data sets (MIMIC-CXR-JPG & MIMIC_IV) for our project are publicly available and have been pre-approved by the IRB. IRB review and approval is not needed if the source of the data is public and analysis of the data will not make the data individually identifiable.