CPH 200C: Computational Precision Health Cornerstone Tues Apr 15, 2025

Problem Set 1: Dataset Shift and FDA Regulatory Actions

Instructor: Prof Irene Chen *Due date: Thurs May 1, 2025*

Risk stratification allows clinicians to separate patients into high and low risk patients. The primary event of interest may include patient mortality, onset of a new disease, or hospital readmission. To analyze this problem, we typically use a supervised learning model to predict future events. The main goal of this problem set is to develop your ability to conduct risk stratification from electronic health records and examine the effects of dataset shift. We will be examining a dataset of 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. Each row concerns hospital records of patients diagnosed with diabetes, who underwent laboratory, medications, and stayed up to 14 days. Our goal is to predict which diabetic patients will be readmitted within 30 days of discharge.

Instructions: This is not a group project and students will be graded individually. This project has two deliverables, which should be submitted at `https://bcourses.berkeley.edu/courses/1544328` by 11:59pm PT Thurs May 1:

- **A report summarizing your results**. The report should include point-by-point answers to the questions below.
- **A zip file with your code**. Please submit both your code and report using the Gradescope. You will get feedback on both your report and code via Gradescope.

1 Diabetes Risk Stratification and Dataset Shift (21 pts)

We will be examining a dataset of 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. Each row concerns hospital records of patients diagnosed with diabetes, who underwent laboratory, medications, and stayed up to 14 days. Our goal is to predict which diabetic patients will be readmitted within 30 days of discharge.

1.1 Data Exploration

Download the diabetes dataset and review the corresponding publication [SDG+14]. Graph the 30-day readmission rates by age, gender, and race. Which groups correlate with higher readmission rates?

1.2 Model Development

To build our risk stratification model, develop machine learning models in three model classes: linear, tree-based, and neural network. Across multiple train-validation-test splits, **we expect performance to be within or higher than 0.65-0.70.**

Describe your selected models and resulting performances, the hyperparameters searched over and chosen, and the best overall model. Graph your model performances in AUC with appropriate confidence intervals [CM04]. You may find the `confidenceinterval` Python package helpful.

Some factors to consider in your model development:

- Categorical features can be turned into one-hot encoding dummy variables. Beware multi-collinearity, which may require dropping one category.
- Some variables will have missing input data. How should you handle this missing data?
- According to [SDG+14], the interactions between variables may have strong predictive power for readmission rates.

1.3 Distribution Shift

Another area of interest is the performance of a model as the population changes. Because we do not have time stamps in this dataset, we will simulate using two different patient populations divided by age. Train a model on only patients younger than 50. You may find the `age` category here helpful. What is the performance of the best linear model trained and tested on patients younger than 50? Using patients older than 50 as a target distribution, what is the performance of the younger-than-50 model on the older than 50 population?

Take some time to explore why that might be. What are potential differences in features and outcomes that may explain this? Consider computing the feature importances of the model or differences between the two populations.

2 FDA Request for Comments (3 pt)

On January 7, 2025, the Food and Drug Administration (FDA) issued draft guidance that includes recommendations to support development and marketing of safe and effective AI-enabled devices. Please read over the draft guidance. What are 1-2 notable parts in the draft guidance?

The FDA requested public comments with a deadline of April 7, 2025 — although the website is still accepting late comments. Of the 107 public comments, choose one and read carefully. Who is writing this public comment? What are they advocating for, and how does that align with their position in either industry, academia, or another sector?

3 MIMIC-IV Warm-up (1 pt)

Looking ahead, Problem Set 2 will use the MIMIC-IV dataset. Let's do a little warm-up exercise now.

3.1 Getting Access

Do not wait on this part because getting MIMIC-IV access can take 2-4 days to receive approval.

To get access, please complete these steps:

- Create a Physionet account
- Complete the CITI training
- Submit evidence of your training
- Sign the Data Use Agreement

3.2 Patient Anchor Groups

For this portion, we recommend using AWS or Google BigQuery to access the dataset because downloading the CSV files can be time-consuming. Using your MIMIC-IV data access, create a table with the number of unique patients for each `anchor_year_group` from the `patients` table. What are anchor year groups and why are they necessary?

4 Feedback (1 pt)

How long did this problem set take you (in hours)? The 1 pt will be given for any response to this question, and the value will only be used for calibration of future problem sets. Please separate out the MIMIC-IV data access from your response, i.e., report time on problem set outside of getting MIMIC-IV data access and report time spent getting MIMIC-IV access.

References

[CM04] Corinna Cortes and Mehryar Mohri. Confidence intervals for the area under the roc curve. *Advances in neural information processing systems*, 17, 2004.

[KZ19] Nathan Kallus and Angela Zhou. The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. *Advances in neural information processing systems*, 32, 2019.

[SDG+14] Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, John N Clore, et al. Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014, 2014.