6.S897/HST.S53
Problem Set 1
Due: March 7th by midnight through Gradescope

Section 0: Instructions

Clone the github repository found at:
https://github.com/mlhc17mit/pset1materials/
Start by reading the Readme.md file there.

Dr. Daniela Thorin, the Chief of Medicine at Mirkwood General Hospital (MGH) is very enthusiastic about leveraging powerful machine learning algorithms for healthcare and personalized medicine. She heard that you excelled at 6.S897 and immediately invited you to visit MGH to talk with the attending physicians about some problems that MGH is facing and suggest data-driven solutions. She asked that you give her a single report of your analyses with relevant plots, interpretation of findings and recommendations.

Section 1: Short staffed

Your first meeting is with Dr. Dwalin, the chief of internal medicine. He explains that one of the most pressing issues that the hospital is facing is that the ICU, particularly the adult ICUs, meaning ICU's other than the neonatal ICU (NICU), are short staffed. Right now, physicians rely on the Simplified Acute Physiology Score (SAPS) to prioritize which patients to direct their attention to. The SAPS score takes into account a handful of patient characteristics to estimate patient severity and risk of death.

Dr. Dwalin believes you could do better with a learned risk estimate. He tells you that they have demographic information, vitals, lab results and date of death if available for all their ICU patients during their entire stay[1]. His main question boils down to:

"How do we leverage EHR data to make optimal use of the MD's time and effort? What outcome should we choose? Which data should we include in our risk estimation?"

Q1: Discuss the advantages and disadvantages of the following setups. Discuss whether or not these setups are well-suited to answer Dr Dwalin's question, and if not, what are situations in which these setups would be useful? Discuss any other problems with these setups from a medical or algorithmic perspective.

- Outcome = 1 year mortality; data = data collected in the first 48 hours
- Outcome = 1 year mortality; data = data collected throughout the entire visit.

[1] Further details about these variables are included in the codebook

- Outcome = In-ICU mortality; data = data collected in the first 48 hours

Q2: You decided to go with in-ICU mortality with data collected in the first 48 hours. Run an L2 regularized logistic regression to predict the in-ICU mortality using the data collected on admission and the data collected during the first 48 hours of the ICU stay. The adult_icu.gz has already been split into training and testing data: use the subset of the data where the train variable = 1 as your training data and the rest as a held out testing dataset. You may further partition the the training data into a training and validation dataset as you see fit. Compute the AUROC on the test data and comment on the model's performance. Look at the top 5 risk factors of mortality and the lowest 5 and explain what they mean.
Hint: In order to compare the coefficients of the features included here regardless of their scale, it is useful to standardize the non-binary variables.

Q3a: Dr. Dwalin is pretty excited about the results. He regrets asking you to train your model only for the adult ICU patients. Because he doesn't want to take too much of your time, he asks you to give him the trained weights from your model. He will use them to create risk estimates for patients in the NICU. What can possibly go wrong? Justify your answer.

Q3b: You assure Dr Dwalin that your schoolwork takes up a trivial amount of time and that you have enough time to retrain your model using the NICU data. Rerun the same analysis as in Q2 but this time using the NICU data. Compare the weights learned in the NICU model to the model from Q2. Comment on your findings.
Note: There are some features that are never measured in the NICU. These variables are highlighted in orange in the codebook.

Q4 and 5 are only to be done for the adult population. Note that because of space limitations on the physionet workspace, we're only providing you with a subset of the adult_notes.gz data. We are working on it, but it should not affect your results.

Q4: Right when the meeting was about to end, Dr. Dwalin went off on a rant about how long it takes him to fill out detailed patient notes. These are notes that physicians, nurses and other healthcare workers fill out describing the health state of the patient. Halfway through his rant, he has a bright idea! "Can you use those detailed notes to predict in-ICU mortality?", he asks.
The adult_notes.gz data has text from all the clinical notes that the patient had in the first 48 hours of his ICU stay. Use the text data from the adult notes to create a bag of words for each patient. Use the bag of words to predict mortality using an L1 regularized logistic regression. Report the top 5 words associated with a high risk of mortality and the lowest 5 as well as the AUROC of that model.

Q5: Combine the structured EHR data (the data you used to predict mortality in Q2) and the unstructured, free text data you used in Q5 to predict in-ICU mortality. Compute the AUROC on the test data. Briefly comment about the two different sources of data in light of your findings.

Note about preprocessing: The datasets that you worked with were preprocessed such that missing values were imputed. Specifically, we stratified the data by gender and age groups and replaced missing values with their group means. The topic of missing data is important and will be discussed further in class.

Section 2: Low birth weight

After a quick lunch break, Dr. Thorin introduces you to Dr. Bombur, an attending pediatrician at MGH. Dr. Bombur has studied that low birth weight (lbw) might have a negative effect on a baby's chances of one-year survival. He is considering starting an initiative that helps pregnant women take better care of their own health so that their babies are born at a healthy birth weight. Before embarking on this $6.5897 million dollar initiative, Dr. Bombur wants to make sure that lbw does indeed have an effect on infants' health, particularly their chances of survival. He asks you:

"Does low birth weight really decrease a baby's chance of survival beyond his/her first birthday?"

For this task, you will use the lbw data. This is a publicly available dataset of linked birth and infant death data collected and maintained by the Center for Disease Control. Each line represents an infant, and has data about the infant's date of birth and death if he/she died within a year. It also has information about the mother, her education, race, smoking habits, alcohol consumption and other as well as some information about the father.

Note that in this part of the question (Q1x), we are looking at birthweight under 2700 grams as the treatment = 1 and 0 otherwise, and the outcome is mortality.

Q1a: Dr. Bombur mentioned that his previous research assistant suggested studying the effect of low birth weight in the population of twins. Why did the previous RA think that the twin population is particularly well suited for the task? Which causality assumptions does the twin population likely satisfy? Use the lbw data to estimate the ATE of low birth weight on one year mortality. To do so, you need to use twins where only one of them is below the threshold. Does this ATE generalize to the whole population including singletons? Meaning: can we assume that the ATE of lbw in the singletons population is roughly the same?
Hint: compute the mortality rates among the twins and the singletons population.

Q2: Dr. Bombur realized he can refine his question a bit. He explains that his initiative would directly target smoking cessation among pregnant women, since it has been shown that babies born to women who smoke during their pregnancy are more likely to have a lbw. If a mother's smoking habits truly do cause babies to be born at a lower weight, Dr. Bombur's initiative would be useful. For the remainder of this question, use the singleton population for your analyses.

Do a covariate adjustment (i.e., regression) to estimate the effect of mother's smoking habits on the baby's birthweight. Clearly state what your outcome is and which variables you include in the regression. Report the weight of the treatment. Give an example of a feature that you should not include in your model.

Q3: An alternative to covariate adjustment is propensity score weighting. Compute the ATE after reweighing the distribution using propensity scores. Comment on the distribution of propensity scores using plots or summary statistics as you see fit. Compute and report the ATE.

After your meeting with Dr. Bombur, you thank all the physicians for their time and decide to stop by the Hospital cafeteria for some seed cake. It's been a long day.