6.790 Mini-Project 2: Fighting Sepsis

Please hand in your work via Gradescope via the link at https://gradml.mit.edu/info/homeworks/. If you were not added to the course automatically, please use Entry Code R7RGGX to add yourself to Gradescope.

1. Please type your solution. We highly encourage LaTeX although it is not required.
2. Project is due on **Friday December 6 at 11PM**.
3. Lateness and extension policies are described at `https://gradml.mit.edu/info/class_policy/`.

1 Background

This text is taken verbatim from
`https://physionet.org/content/challenge-2019/1.0.0/`

> Sepsis is a life-threatening condition that occurs when the body's response to infection causes tissue damage, organ failure, or death (Singer et al., 2016). In the U.S., nearly 1.7 million people develop sepsis and 270,000 people die from sepsis each year; over one third of people who die in U.S. hospitals have sepsis (CDC). Internationally, an estimated 30 million people develop sepsis and 6 million people die from sepsis each year; an estimated 4.2 million newborns and children are affected (WHO). Sepsis costs U.S. hospitals more than any other health condition at $24 billion (13% of U.S. healthcare expenses) a year, and a majority of these costs are for sepsis patients that were not diagnosed at admission (Paoli et al., 2018). Sepsis costs are even greater globally with the developing world at most risk. Altogether, sepsis is a major public health issue responsible for significant morbidity, mortality, and healthcare expenses.
>
> Early detection and antibiotic treatment of sepsis are critical for improving sepsis outcomes, where each hour of delayed treatment has been associated with roughly an 4-8% increase in mortality (Kumar et al., 2006; Seymour et al., 2017). To help address this problem, clinicians have proposed new definitions for sepsis (Singer et al., 2016), but the fundamental need to detect and treat sepsis early still remains, and basic questions about the limits of early detection remain unanswered. The PhysioNet/Computing in Cardiology Challenge 2019 provides an opportunity to address these questions.

The following part of the site describes the data and utility function. Please read it in detail! Note that

- For each patient, there is a time series of hourly data.
- In each row of that data, there are values associated with vital signs and lab results.
- *Much of the data is missing* because not every test is performed on every patient at every hour!
- Your *predictor* should operate on a patient's record, time-step by time-step, and classify the patient's history *up until time* t as indicating sepsis or not.
  - If the patient does have sepsis, then
    - You get the best score for predicting that 6 hours before the label in the data says that sepsis was detected, but a good score for predicting somewhat earlier or later than that.
    - You get a very bad score for not detecting it.
  - If the patient does not have sepsis, then
    - You get a score of 0 for not predicting they have sepsis.
    - You get a small negative score for predicting that they have sepsis.

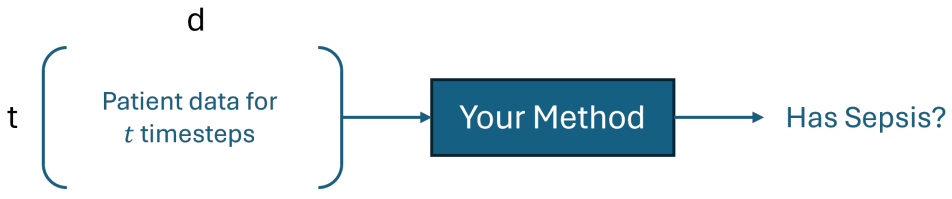
  These scores are summed over the time-course for the patient and normalized (via a piece of code we'll give you.)

All of our data is a list of (x, y) pairs for each patient, where (x, y) encodes the patient's entire history. The data for an invidual patient indicates whether they had sepsis, like this:

Record for an Example Patient

t	F1	F2	...	Fd	Has Sepsis?
1					0
2					0
3					1
...					1
k					1

Your predictor, when predicting whether the patient has sepsis at time t, can take into account all the data for that patient up to and including time t, as illustrated here:



d is fixed, but t may vary based on how long the patient has been in the hospital!

2 Machine-learning considerations

Your goal is to make a prediction method that scores well on new patients, by predicting sepsis earlier than it was originally detected (so that it can be treated more effectively) and not predicting sepsis for patients that don't have it (to avoid the side-effects and costs of unnecessary treatment.)

There are several aspects of this problem that make it not completely straightforward:

- More data values are missing than are present.
- The loss function is not a simple 0-1 loss, where we simply classify the prediction on each time step as correct or not, and sum or average over the time steps.
- There is a significant class imbalance (substantially more negative than positive examples).
- The number of data points (hourly records) varies per patient (some were in the hospital longer than others).

3 Implementation

The colab notebook here
`https://colab.research.google.com/drive/1gXR9gGh1uOSFtslvACl4G0jeiT9_VpBx`
contains code for loading the data and for evaluating a predictor.

Please use datasets train and val for all of your algorithm development (including model-selection, parameter-tuning, etc.). Run your final classifier *once* on dataset test as a simulation of actually deploying your algorithm in practice.

You may use any existing ML algorithm implementations (including sciki methods).

4 Questions

4.1 Conceptual questions

1. Are the missing values in this data set *missing completely at random*? Discuss briefly a good model for the missing data.
2. Each patient's entire record is a *time series* that encodes information not only about their most recent state but how it has been changing. Perhaps it's important that their temperature has been increasing recently or that their blood pressure is highly variable, for example. Describe a way to encode this type of historical information into a fixed-size feature vector that we could perform standard supervised learning on.
3. In the next section, we will ask you to implement several relatively simple approaches that are based on classifying each hour of each patient record (potentially including historical information). But maybe it's better to take a more holistic view: how could we address this problem with a transformer?
    - Describe a plausible input and output encoding.
    - Would it be important to have positional encoding?

- Would it be important for the transformer to have a causal attention mask?
- Would it be possible to use the actual utility function directly, to train the transformer?

If you would like to try this, or another complex model, it's completely optional, but you could do it instead of H2 below.

4. How were the "ground truth" labels in this data determined? What problems might they have? What would be another strategy for getting labels?

4.2 Implementation

5. Implement two baselines: always predict sepsis at every time step, and never predict sepsis at any time step (call these B1 and B2).
6. Implement at least two strategies for handling missing data (call these M1 and M2). Explain the likely advantages and disadvantages of these two strategies.
7. Implement at least one strategy for taking the provided utility function into account (call this L2, and call optimizing simple 0-1 loss L1). Explain the likely advantages and disadvantages of these two strategies.
8. Implement at least one strategy for handling the class imbalance (call this method C2, and call using the data as it is C1). Note that it is fine for it to look at the class balance in datasets train and val (but not dataset test)! Explain the likely advantages and disadvantages of these two strategies.
9. Try at least two underlying machine-learning methods (e.g., logistic regression and something with a non-linear hypothesis space). Call these H1 and H2. Explain the likely advantages and disadvantages of these two strategies.

4.3 Results

10. Use a validation method (with the dataset val) to compare methods B1, B2, at least all 4 of M2-L1-C1-H1 ... M1-L1-C1-H2, and any other combinations you experiment with beyond that. Explain the validation process you used and what predictor you decide is best.
11. Provide numerical bounds on the risk of using this predictor in novel settings; explaining and justifying the process you used to construct them.
12. Now! run your predictor in dataset test. How does its performance compare to your prediction?

5 What to hand in and how it will be evaluated

Please submit a single pdf document. It should be typed, not handwritten, but does not need to be in latex. It should have clearly labeled answers to the questions above, written in good English paragraphs.

This is not a contest! We won't grade on your overall prediction error or cost. We will grade on your ability to formulate an ML problem, execute the formulation, and evaluate the result.

You may not submit code. Do please submit data, tables or graphs as necessary to make your points.

Grading rubric. For each question:

- 50%: Did your decisions make basic sense given the overall problem setting and the choices you had already made?
- 50%: Were your explanations clear and well thought out?