# Problem Set 1 (67 Points)

MLHC 2023

February 14, 2023

## 1 Submission Instructions

This problem set is due 2-27-23 at 11:59pm EST. Please submit your write-up and code (* for the meantime use Canvas to upload, we might change to Gradescope later on). When you write up your report, put all writing into one file and name it `mit_email_username_pset1.pdf` (e.g. `lehmer16_pset1.pdf`), 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.

## 2 How to Access PSET Files

A key part of this homework will be accessing certain files that are only available under MIMIC-access. Here are the following steps:

1. Create a free Google Cloud Platform (GCP) account connected to a Gmail.
2. You must set your Google email on PhysioNet: https://physionet.org/settings/cloud/. If you have any trouble with this step, there are more detailed instructions here.
3. Request access to MIMIC on GCP: Go the PhysioNet MIMIC-III page, Files section, and click "Request access the files using Google Cloud Storage Browser" and 'Request access using Google BigQuery'.
4. Do the same steps as above, but for MIMIC-IV.
5. Check whether you can run the steps in the following Colab notebook. You may need to make a copy of it to run it.

## 3 MIMIC DUA (3 Points)

In this section, we hope to ensure that you have a strong understanding of MIMIC's data use agreement.

### 3.1 Google Drive

Lynnea wants to share some preprocessed MIMIC files with Josh, her homework partner who also has MIMIC access, so she uploads them to her personal, private, Google Drive and shares them with Josh. Does this violate any policies from MIMIC? If so, what precautions, if any, could she take to fix these violations? Answer in 1-2 sentences.

### 3.2 Public Clusters

Adam and Eli are working on their MLHC project. They want to use MIT's Satori Cluster, which everyone at MIT has access to. By default, files are readable by other users. Would this violate any policies from MIMIC? If so, what precautions, if any, could Adam and Eli take to fix these violations? Answer in 1-2 sentences.

### 3.3 ChatGPT

Eric is lazy. He doesn't want to do the homework, so he copy-pastes notes from MIMIC and asks Chat-GPT to answer the homework questions about the notes. Would this violate any policies from MIMIC? If so, what precautions, if any, could he take to fix these violations? Answer in 1-2 sentences.

## 4 Differential Diagnosis (20 Points)

### 4.1 Background

In clinical reasoning, one critical component is to understand the iterative nature of the diagnostic process. Clinicians make the iterative diagnostic process through the cycle of gathering data and updating the differential diagnosis via Bayes' rule. With Bayes' rule, we can reach the most possible diagnosis by updating the posterior probability after observing a sequence of symptoms or test results. In this question set, we will go through a simple clinical case and use the Bayes' model to decide which lab test to use for diagnosing the patient.

### 4.2 Problem Setup

A 65-year-old male patient without medical history is visiting your clinic and complaining that he has the problems of cough and slight shortness of breath for three days. No other significant symptoms such as high fever, sputum production, heart rate/respiratory rate/blood pressure change, are mentioned. The problem of cough also does not aggravate or relieve by time (day or night), weather, and exercise. As a clinician, you are thinking that the patient's problem can be COVID-19, tuberculosis (TB), or influenza (FLU). For each possible diagnosis, there is a corresponding lab test you can order in the clinic. However, only one lab test can be executed at a time since our clinic's lab is small.

To make the best assessment of the most likely single diagnosis that can explain the symptoms (Occam's razor), we can use the Naive Bayes model to make a diagnosis given the results of lab tests. We assume that the results of these three lab tests, COVID test, TB test, and FLU test, are conditionally independent given the hidden true diagnosis, and the cost (weight) of the test is not considered a problem.

We can use the Naive Bayes classifier, the model that picks up a diagnosis (class) $D \in \mathbf{D}$, where $\mathbf{D}$ is the set of all possible diagnoses (COVID, FLU and TB), that maximizes the joint probability of lab test results (features) L and class. We can formulate the problem as follows:

$$\hat{y} = \underset{j \in \{1..J\}}{\operatorname{argmax}}\, p(D)p(\mathbf{L}|D) \tag{1}$$

and the probability of n different lab test results given the diagnosis, $p(\mathbf{L}|D)$, is the product of the observed outcomes under the naive independence assumption:

$$p(L|D) = \prod_{i=1}^{n} p^{l_i}(1-p)^{(1-l_i)} \tag{2}$$

A diffuse distribution of possible diagnoses means we are not very certain about which disease is affecting the patient. We can quantify how uncertain a given distribution of possible diagnoses is using the idea of **entropy**:

$$H(D) = -\sum_{D \in \mathbf{D}} p(D) \log_2 p(D) \tag{3}$$

If we have the parameters for the Naive Bayes model as below (supposing that the patient must have one of the following diagnoses):

| Diagnosis D | p(D) |
|---|---|
| COVID | 0.5 |
| FLU | 0.3 |
| TB | 0.2 |

| | COVID Test | FLu Test | TB Test |
|---|---|---|---|
| p(test positive \| D = COVID) | 0.8 | 0.1 | 0.3 |
| p(test positive \| D = Flu) | 0.8 | 0.5 | 0.1 |
| p(test positive \| D = TB) | 0.1 | 0.3 | 0.9 |

## 4.3 Questions

1. (1 Point) What is the distribution of possible diagnoses before we do any tests?

2. (2 Points) What is the entropy of the distribution of diagnoses before we order any test?

3. (2 Points) What is the probability of a positive COVID test?

4. (3 Points) What is $H(\mathbf{D}|\text{COVID test positive})$?

5. (3 Points) What is $H(\mathbf{D}|\text{COVID test negative})$?

6. (9 Points) Now we are able to order either COVID or FLU test. For instance, if we choose COVID test, we could learn either COVID test is negative (COVID=0) or positive (COVID=1) but we will not know which is true until we order the lab. Between COVID test and FLU test, which lab should we order first in order to maximize the amount of expected information that we learn (i.e. reduce the entropy of the distribution)? Remember that if you order the lab test, it means that you can compute the posterior distribution for the case where we learn the result is positive or negative.

# 5 Understanding ICD Codes (9 Points)

For the rest of the sections, you will need to use the notebook provided to you, which is titled `mlhc_pset1_starter_code.ipynb` also available online at https://colab.research.google.com/drive/1BG_qbu5oYGJSwLttIuXdlTFn4QtBtvZW?usp=sharing (create your own copy before writing code). The starter code provides the basic items needed to explore the files.

1. (1 Point) What are the top 10 ICD codes in MIMIC-III and what percent of total ICD codes does each ICD-9 code in the top 10 make up (e.g, Hypertension, 1%)? Please provide the description of the ICD code and the code itself.

2. (1 Point) What is the average number of ICD codes per visit in MIMIC-III?

3. (1 Point) What are the top 10 ICD-10 codes in MIMIC-**IV** and what percent of total ICD codes does each ICD-10 code in the top 10 make up? Please provide the description of the ICD code, not the code itself.

4. (1 Point) What is the average number of ICD-10 codes per visit in MIMIC-IV?

5. (4 Points) Your friend, Ian, wants to automate ICD coding in every hospital in America. He plans to use MIMIC-IV as the basis for his machine learning model. However, you notice that the top 10 outpatient diagnoses in 2021 list looks significantly different from the top 10 ICD-10 codes in MIMIC. Cite two reasons why this list looks so different from the list you found earlier in question 3.3.

6. (1 Point) What is your favorite ICD code?

# 6 Exploring Patient Notes (15 Points)

Use the notebook provided to you, which is titled `mlhc_pset1_starter_code.ipynb`. Let's examine patient with `SUBJECT_ID` of `80110`. Everything you need is in this CSV: `patient_80110_notes.csv`. First let's look at their discharge summary, with a `ROW_ID` of 36482.

1. (1 Point) How old is this patient and what is their sex?

2. (1 Point) How long were in they in the hospital for?

3. (2 Points) Why was this patient initially admitted to the ICU?

4. (2 Points) List all of this patient's ICD diagnosis billing codes and their descriptions. Not all of these codes diagnoses are from this particular ICU visit. List at least one diagnosis that the patient was billed for, but was not made during this ICU stay.

Now, let's examine one of this patient's nursing notes (read `ROW_ID` of `570974`). Read the first two paragraphs (up until 'Significant Events' section) of the nursing note.

5. (1 Points) What section in the discharge summary do we see the most overlap with?

6. (2 Points) Are there any detail(s) mentioned in the first two paragraphs that are not mentioned in the discharge summary?

Let's quickly look at one more nursing note (`ROW_ID` value of `571028`).

7. (3 Points) How does this note differ from the other two notes? Give one possible reason why this note differs so much.

Finally, after reading these notes, please answer the following question:

8. (3 Points) These notes look very different from typical text. List 3 differences between hospital notes and typical text from the web (e.g., Wikipedia) that may present additional challenges to apply machine learning models to.

# 7 Length of Stay Prediction (20 Points)

Use the notebook provided to you, which is titled `mlhc_pset1_starter_code.ipynb`. As we saw during the COVID pandemic, hospital bed allocation is a challenging problem. One possible tool that might aid this allocation process is a machine learning model that can predict a patient's *length of stay*. Use the same notebook as before and answer the following questions. Please use the following file: `length_of_stay.csv`. First let's examine the data:

1. (1 Point) What is the average and median length of stay?

2. (2 Point) Plot a histogram of the length of stays, with the following bins: `[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]`. Which of the bin has the highest frequency?

3. (2 Point) Why does this not align with the mean and median that you found previously?

Now, let's finally do some machine learning!

1. (11 Points) Run a logistic regression using `L2` loss to predict if the patient's length of stay will be greater than 7 days. Perform a 70/30 split of the data, ensuring that you do not introduce any bias into your dataset. Then, calculate and report the AUC and accuracy of your model.

2. (1 Point) What is the accuracy of the majority class?

3. (3 Points) Find the top 3 most predictive features. Do these make sense to you? Walk through each of these and give a 1-2 sentence response about why this variable might be helpful.