Problem Set 1 (76 Points)

6.7930/HST.956 Machine Learning for Healthcare

February 11, 2025

Submission Guidelines

Due 2/24/2025, 23:59 on Gradescope

Please make a submission only if you are registered as a regular student. Submit your write-up as `[mit_email]_pset1.pdf` (*e.g.,* `demirel_pset1.pdf`), with all written work in a single PDF file following the problem set structure. Please append your code at the end of the report. For those who prefer to typeset in LaTeX, the source code of this file is here .

Submit your own work and note any collaborators - if none, state "Collaborators: none." If using external sources, cite them properly. You should be able to explain your solutions verbally. Please do not share your code or report with anyone inside or outside of the class, nor post them publicly online. Course staff welcomes any questions about these policies.

Prerequisites

Make sure you setup Vertex AI API on Google Cloud Platform to use Gemini by following the instructions outlined in the Canvas (check your inbox for the message entitled "Setting Up Google Gemini via Vertex AI on Google Cloud Platform"). You will need this for Question 3.2.

A key part of this homework will be accessing certain files that are only available under MIMIC-access, which you applied for in Problem Set 0.

1. Set your Google email on PhysioNet (use the one you redeemed the educational credit coupons with): https://physionet.org/settings/cloud/. If you have any trouble with this step, there are more detailed instructions here .
2. Request access to MIMIC on GCP: Go to MIMIC-III and click "Request access the files using Google Cloud Storage Browser".
3. Do the same steps as above, but for MIMIC-IV .
4. We created a starter Colab Notebook that you can use. You may need to make a copy of it to run. Make sure that you are able to load the MIMIC files into the Colab notebook environment from Google Cloud.

MIMIC DUA

Before starting, we hope to ensure that you have a strong understanding of MIMIC's data use agreement.

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.

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.

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 would you suggest Eric do instead to use generative AI tools with sensitive health data responsibly? Answer in 1-2 sentences.

1 Differential Diagnosis (27 pts)

1.1 Bayesian Reasoning (17 pts)

1.1.1 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, such as sequence of symptoms or test results, and updating their "beliefs" via Bayes' rule.

In this question, we will go through a simple clinical case and use the Bayes' model to decide which lab test is the most useful/informative for guiding differential diagnosis.

1.1.2 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 lab test you can order in the clinic.

Note that based on the result of some lab test, L, we can make the diagnosis with the highest *posterior probability* as follows:

$$\hat{y} = \underset{d \in \mathbf{D}}{\text{argmax}}\ p(D = d \mid L) \tag{1}$$

$$= \underset{d \in \mathbf{D}}{\text{argmax}}\ p(D = d)p(L \mid D = d) \tag{2}$$

where

$$\mathbf{D} = \{\text{COVID,TB,FLU}\}. \tag{3}$$

Since our clinic's lab is small, only one lab test can be executed and we want to pick the test that will be the most informative in guiding diagnosis. Hence, we will pick the lab test that reduces the uncertainty among the possible set of diagnoses, as we explain next.

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 with **entropy**:

$$H(D) = -\sum_{d \in \mathbf{D}} p(D = d)\log_2 p(D = d) \tag{4}$$

Similarly, we can define conditional entropy for the distribution of diagnoses based on the result of some test, $L = l$, as,

$$H(D \mid L = l) = -\sum_{d \in \mathbf{D}} p(D = d \mid L = l)\log_2 p(D = d \mid L = l) \tag{5}$$

Finally, expected conditional entropy, based on some test L, is defined as

$$H(D \mid L) = -\sum_{l \in \{-,+\}} p(L = l)H(D \mid L = l) \tag{6}$$

where $-$ means the test results was negative and $+$ means that it was positive.

We consider the *prior* disease probabilities and test behaviors given below:

Table 1: Prior disease probabilities and the diagnostic behavior of different tests

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

1.1.3 Questions

Let L_C denote the result of a COVID test and L_F that of a FLU test.

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

2. (1 pt) What is the entropy of the distribution of diagnoses before we order any test?

3. (2 pts) What is the probability of a positive COVID test, $p(L_C = +)$?

4. (4 pts) Compute $H(D \mid L_C = +)$ and $H(D \mid L_C = -)$, and finally $H(D \mid L_C)$.

5. (9 pts) 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)?

   Hint: Can you compute $H(D \mid L_C)$ and $H(D \mid L_F)$? (see Eq. 6). What do these quantities mean?

1.2 Large Language Models for Assisting Differential Diagnosis (10 pts)

1.2.1 Background

In previous parts, we explored the concept of "entropy," and used it to decide which test was the "most informative" for diagnosis. While this approach is mathematically grounded, it relies heavily on explicit probabilistic specifications that may not fully capture patient-specific nuances (Table 1). Two key challenges emerge:

1. The criticality of accurately modeling both prior disease probabilities and test characteristics
2. The difficulty of specifying these probabilities for heterogeneous patient populations

In practice, these probabilities often exhibit complex dependencies on patient characteristics. Let X denote the vector of relevant patient features (e.g., demographics, medical history, geographic location). The proper specification would require:

- Prior probabilities conditional on patient features: $P(D \mid X)$

- Test characteristics that may vary by patient: $P(\text{test positive} \mid X, D)$

Explicitly modeling these conditional probabilities for every possible patient profile quickly becomes intractable. Large Language Models (LLMs), with their extensive training on medical literature and ability to process complex contextual information, present an intriguing alternative approach.

1.2.2 Questions

1. (5 pts) Pick your favorite LLM (you can also use Gemini on Vertex AI). Prompt it as follows:

> **Prompt**
>
> You are a helpful medical expert. Consider a patient presenting in the clinic with the symptoms and history given below.
>
> 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 lab test you can order in the clinic.
>
> Now, help me with the following clinical decision.
>
> Since our clinic's lab is small, only one lab test can be executed and we want to pick the test that will be the most informative in guiding diagnosis. Which test should we pick?
>
> Think step by step and explain your reasoning.

   - Does it consider/mention disease prevalence (*i.e.*, $P(D)$)? If so, briefly discuss the arguments and how do they factor into the LLM's suggestion.

   - Does it consider/mention test characteristics (*i.e.*, $P(\text{test positive} \mid D)$)? If so, briefly discuss the arguments and how do they factor into the LLM's suggestion.

   - How does its recommendation compare to the solution in the previous part? What seems to be the main reason behind LLM's suggestion?

- Update the LLM with the following prompt:

  Prompt
  Imagine if we had a FLU test that always detected FLU and had zero false positives. Which test would you suggest in that case? Explain your reasoning clearly.

  Does the suggested test change? Is the new suggestion mainly due to the prevalence of the diseases or test characteristics? Compare the reason behind the new suggestion to the reason behind the previous suggestion.

- Provide the LLM with Table 1's probability specifications (should be as easy as taking a screenshot and giving it to the model, then asking to consider those prior disease probabilities and test specifications!).
  How does LLM use/react to this information? Briefly explain and evaluate its reasoning based on this information. Does its suggestions change?

2. (3 pts) Update the LLM with the following prompt:

   Prompt
   Modify your suggestions to accommodate a 25 years old patient presenting with the same symptoms, and has recently traveled internationally.

   - How does the LLM adjust its assessments, does it request specific travel information?

   - If yes, test the LLM with different travel destinations (e.g., Southeast Asia vs. Western Europe). Analyze how travel context influences its reasoning and recommendations.

   - Compare this adaptive reasoning with our fixed-probability approach in Section 1.1.3

3. (2 pts) Briefly discuss both the potential advantages of LLMs in incorporating complex contextual information and their limitations in providing explicit probabilistic reasoning.

2 Understanding ICD Codes (9 pts)

Use the starter Colab Notebook after creating your own copy.

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

3. (1 pt) 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 pt) What is the average number of ICD-10 codes per visit in MIMIC-IV?

5. (4 pts) 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 pt) What is your favorite ICD code?

3 Analyzing Patient Notes (22 pts)

Use the starter Colab Notebook after creating your own copy.

3.1 Exploring the Notes (12 pts)

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 pt) How old is this patient and what is their sex?

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

3. (1 pt) Why was this patient initially admitted to the ICU?

4. (2 pts) 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 pt) What section in the discharge summary do we see the most overlap with?

6. (2 pts) 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. (2 pts) 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. (2 pts) 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.

3.2 Using LLMs for Clinical Documentation (10 pts)

Do not use any other LLM for this question other than Gemini through Vertex AI API on Google Cloud Platform.

1. (4 pts) Ask Gemini to generate a structured summary of the discharge summary note from Section 3.1.

Prompt

You are a helpful medical expert. Consider the discharge summary below and generate a structured summary that would be useful for chart review by a clinician.

>>INSERT DISCHARGE SUMMARY

Compare this summary to the original note.

- Was it quick to generate? How is the output generally structured into different sections?

- What are some advantages of the LLM-generated structured summary over the original note for chart review by clinicians?

- What key information is accurately captured or potentially missing (*e.g.,* about medications)? Does Gemini seem to have hallucinated anything? Does it repeat same information multiple times?

2. (3 pts) Clinical summaries often need to be generated for different audiences (patients, specialists, primary care physicians). Ask Gemini to generate a different version of the discharge summary for a patient using the following prompt:

Prompt

Can you modify your summary to accommodate the patient as the reader? Before giving the summary, briefly describe the main changes you made.

- How does the technical detail and length change? What does it say that to have changed? Do you observe those chagens yourself?

- Does the new summary include detailed lab results? Compare the level of detail about usage instructions for discharge medications in this version to that in the clinician version.

3. (3 pts) Ask Gemini to generate a list of follow-up questions based on the nursing note from ROW_ID 570974.

> **Prompt**
>
> Raise a list of follow-up questions based on the following nursing note to help clinicians improve the quality of care.
>
> >>INSERT NURSING NOTE

- Give an example of a question Gemini asks that you like. Why do you think this is a good question?

- Do Gemini question certain decisions and ask for additional information for justifying them?

- Do Gemini offer additional procedural insights that were not performed but still be useful to the physicians?

4 Length of Stay Prediction (18 pts)

Use the starter Colab Notebook after creating your own copy.

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*.

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

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

Now, let's finally do some machine learning!

1. (10 pts) Perform a random 70/30 train/test split of the data. Run a logistic regression using `L2` loss to predict if the patient's length of stay will be greater than 7 days using the train split.

   Report the AUC, accuracy, recall, and specificity of your model in the train and test splits.

   How do these metrics change when you balance each instance by inverse probabilities of category frequencies (set `class_weight='balanced'` in `sklearn.LogisticRegression`). Comment on your findings.

2. (3 pts) 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.