CMU 10-742 (Fall 2024) - Machine Learning in Healthcare

Assignment 1: Healthcare Data, Codes, APIs...and Bayesian Inference

Out: Thur Aug 29 2024

Due: Thurs Sep 12 2024

This assignment counts for 8 points out of the 35 total points allocated to the course problem sets.

In this assignment, we're going to get familiar with a few popular healthcare datasets, codesets, and APIs. We'll do some EDA (exploratory data analysis) on a few popular datasets. We'll investigate the discrepancy between how much male and female providers make. And we'll get our feet wet with Bayesian inference for clinical diagnosis.

Notes for this and subsequent assignments:

- Make a copy of this colab notebook and provide your code/answers in the marked sections. To hand in your assignment, download the `.ipynb` file and submit it via the course Canvas site.
- You must only hand it one file - do not decompose your questions into multiple notebooks.
- We assume you have all necessary libraries already installed in your colab environment. If you get a runtime error from colab about an unrecognized import, just install it, e.g. `!pip install numpy`.
- Refer to the course syllabus for detailed policies on collaboration, using external tools, late policy, etc. Your assignment will be considered on time if the last revision to the notebook is before the assignment deadline.
- The datasets live on various GCP buckets. We've configured these buckets to allow read access from all CMU accounts. Of course, you must be authenticated to your CMU account. We have included (in a cell below) the required code to authenticate this notebook to your CMU account.
- Show your work. In some cases, this means clear, documented source code. In other cases, it means showing how you arrived at a numeric answer.
- To receive full credit for a problem, your solution must be correct and intuitive and succinct. Reproducability is critical in ML research, and we expect your code to be clean and well documented.

- Do not store your answers anywhere that others can easily access them. Your answers should not be accessible from the public internet, or any file system or cloud repository where other students (today or in the future) may be able to access them.

```
# Some preliminaries
import locale
import pandas as pd
import requests
import matplotlib.pyplot as plt
from scipy.stats import chi2_contingency

locale.setlocale(locale.LC_ALL, '')
from google.colab import auth
auth.authenticate_user()
```

PART 1: Handling healthcare data (1 point)

In this section, we will validate and reinforce your understanding of Physionet's data use agreement, which covers MIMIC and other datasets.

1.1

Gabriela and Eli are working on their class project, which involves MIMIC data. They want to use a shared folder on CMU's Andrew Cluster. By default, files in this folder are readable by other CMU account holders. Would this violate any policies from MIMIC? If so, what precautions, if any, could Gabriela and Eli take to fix these violations? Answer in 1-2 sentences.

YOUR ANSWER HERE

1.2

Eric is lazy. He doesn't want to do the homework, so he copy-pastes notes from MIMIC into ChatGPT to answer a homework question. 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. (Hint: you may wish to review this)

YOUR ANSWER HERE

PART 2: Playing with FHIR (2 points)

FHIR is a government-supported healthcare API that is becoming an industry standard for exchanging healthcare data. This set of questions will familiarize you with FHIR. For this set of questions, we rely on a publicly available FHIR server, called HAPI FHIR. This server contains a collection of synthetic patient data.

Before jumping into these questions, we recommend skimming the following:

https://fhir-drills.github.io/index.html: a helpful intro to the FHIR API.

https://www.hl7.org/fhir/references.html: the authoritative online reference.

https://build.fhir.org/resourcelist.html: list of resources; refer back to this as you go through these questions.

We have set up a custom FHIR server on Google Cloud that you have access to, ***provided you are signed into your CMU Andrew account***. The following code blocks set up the necessary project and authentication configurations such that you will be able to easily access this server, which contains the patient data that you will need to complete this question.

```
!gcloud config set project fhir-10742
import subprocess
token = subprocess.check_output(["gcloud", "auth", "print-access-token"]).strip().de
headers = {
    "Authorization": f"Bearer {token}",
}
```

```
base_url = "https://healthcare.googleapis.com/v1/projects/fhir-10742/locations/us-ce
```

```
base_url = "http://hapi.fhir.org/baseR4"
```

If you would prefer to work with the data directly on the public FHIR server, you may replace the base_url above with "http://hapi.fhir.org/baseR4". However, note that this server has known accessibility issues, particularly with high user traffic volumes, and so you are encouraged to interact with the custom server provided, private to CMU-affiliated students.

2.1

For patient ID `9312817`, how many medications were prescribed? List all these medications (e.g. `Clopidogrel 75 MG Oral Tablet`), along with the date of prescription.

Based just on this list of medications, can you surmise what is the likely medical condition for this patient?

YOUR ANSWER HERE

2.2

For this same patient, how many observations were made? List all these observations.

YOUR ANSWER HERE

2.3

Plot the glucose values over time for this patient. Note that the LOINC code for glucose is `2339-0`, so your FHIR query should look something like:

```
query_url = {base_url}/Observation?code=2339-0&patient={patient_id}
```

Two consecutive readings above 125 mg/DL is suggestive of diabetes. Does this patient appear to be diabetic, based only on the reported glucose levels?

YOUR ANSWER HERE

PART 3: CMS Data (3 points)

The Centers for Medicare & Medicaid Services (CMS) is the U.S. federal agency that administers the Medicare program, which provides health insurance to Americans aged 65 and older. Besides overseeing Medicare, CMS also jointly administers Medicaid with state governments, providing health insurance to low-income individuals and families, and also manages few other programs (e.g. CHIP, ACA).

CMS began publishing open datasets in 2010 as part of President Obama's Open Government Initiative. It's remarkable what kind of data you can download, for free and without credentials, from https://data.cms.gov. Have a look!

This set of questions will familiarize you with the kind of data that payers (CMS and private insurers like Cigna) collect and manage. This is often called “administrative” data, to contrast it with the kind of data (mostly clinical) that hospitals and doctor offices collect and manage.

We’ve downloaded the Medicare Physician & Other Practitioners - by Provider database from CMS for you, and taken a random sampling of 10% of the data, to make it more manageable.

```
path="Medicare_Physician_Other_Practitioners_by_Provider_2021_processed.csv"
!gsutil cp gs://10-742/assignment_1/{path} ./
df = pd.read_csv(path, low_memory=False)

# This file was processed from the original CMS file, available at
# https://data.cms.gov/provider-summary-by-type-of-service/medicare-physician-other-
#
# In case you're curious, here's how we processed this file for you:
#
# prune out 90% of the rows away (to make it a more manageable size)
# prune out rows which correspond to a *facility*, not an individual provider.
# remove low-frequency provider types and states
# remove columns with more than 20% missing data
# replace remaining missing cells with 0 (for numerical columns) or 'nothing' (for c
```

3.1

Plot a histogram of the values for `Tot_Mdcr_Pymt_Amt`, which is the total amount during 2021 that each provider was paid by CMS for treating Medicare members. Why might it be preferable to use a logarithmic scale on the y-axis?

YOUR ANSWER HERE

3.2

What are the top ten specialties (i.e. `Rndrng_Prvdr_Type`), ranked by decreasing average per-provider total Medicare payment? The bottom ten specialties?

YOUR ANSWER HERE

3.3

Show a bar graph of total payment by state, with the states shown in decreasing order of average payment. Note that the `Rndrng_Prvdr_State_Abrvtn` column stores the state for the provider.

YOUR ANSWER HERE

3.4

So far we've observed that Medicare payments to providers vary widely by specialty, and they also vary significantly by geography. Let's now look at the gender of the provider. Produce a bar graph of the average total Medicare payment, by gender.

YOUR ANSWER HERE

3.5

It sure looks like CMS pays male providers a lot more than female providers! That is concerning. But before we jump to conclusions, let's take a closer look at our data.

Where might there be confounding variables?

For example, nurses get paid less than doctors, and perhaps there are more female nurses, thus skewing the overall payment distribution? Could it be that the genders are not equally represented in certain (high or low paying) specialties?

To start our investigation, let's measure the association between gender and provider type. More specifically, we'll use the chi-squared test to determine if there's a statistically significant association between `Rndrng_Prvdr_Gndr` and `Rndrng_Prvdr_Type`. If the p-value for the test is below 0.05, it indicates that the association observed in the data is unlikely to be due to chance.

Report the p-value of the chi-squared test. Is there a statistically significant association between gender and provider type?

YOUR ANSWER HERE

3.6

Now let's try to remove one suspected confounding variable, which is the licensure level. Do this by only looking at rows where `Rndrng_Prvdr_Crdntls` is "MD" or "M.D." For these remaining rows,

plot the average total payment by gender.

Does this even out the results?

YOUR ANSWER HERE

3.7

Let's now remove the confounding variable of provider type. One way to accomplish this is using matched-pair analysis.

The idea is this. We will subsample the rows of our dataset so that, for each provider type, we have the same number of samples from each gender. With this new dataset, we can re-investigate the association between gender and payment, but this time, we will have removed the confounding factor of provider type.

More specifically, your task here is to:

- Consider only rows where `Rndrng_Prvdr_Crdntls` is "MD" or "M.D."
- Prune all rows where `Rndrng_Prvdr_Type` occurs less than 100 times in the dataset.
- For each remaining `Rndrng_Prvdr_Type`, select an equal number of each gender.
- Create a new bar graph, as above, showing `Tot_Mdcr_Pymt_Amt` against Rndrng_Prvdr_Gndr

What are your observations? Is there still a discrepancy? If yes, what might be the reason for that?

YOUR ANSWER HERE

Part 4: Bayesian Inference (2 points)

Congratulations! You have just been appointed Chief of Springfield General Hospital's DRG - the Diagnostic Referral Group. This is the elite team of expert clinicians who consult on the trickiest cases in the hospital.

As an expert diagnostician, you rely on Bayesian inference as a core part of your toolkit.

You have been called in to help assess whether a patient, Harry Q Bovik, has the rare disease called "Yinzer syndrome."

Yinzer syndrome affects about 1 in 10,000 people. A new test has been developed to detect this disorder, which has the following characteristics:

- Sensitivity (True Positive Rate): 99%
- Specificity (True Negative Rate): 98%

Mr. Bovik has tested positive for Yinzer syndrome. Your job is to determine whether Mr. Bovik actually has Yinzer syndrome.

4.1

Explain succinctly what is meant by sensitivity and specificity in this context. How do these terms relate to the terms recall and precision?

Let's define some terms:

$P(D)$: probability of the disease (Yinzer syndrome, in this case)

$P(T)$: probability of a positive test

$p(D|T)$ and $p(T|D)$ should hopefully be obvious

YOUR ANSWER HERE

4.2

Given this terminology and using Bayes Theorem, calculate the posterior probability $p(D|T)$ as a function of known numeric quantities. Show your work.

YOUR ANSWER HERE

4.3

It is interesting to observe that $p(T|D)$ is so large and $p(D|T)$ is so small. Provide an intuitive explanation for that. Give some real-world examples of diseases D and tests T where this same discrepancy between $p(T|D)$ and $p(D|T)$ holds.

YOUR ANSWER HERE

4.4

Now suppose new genetic research determines that individuals with certain characteristics (e.g., family history, ancestry from a particular region) have a higher base rate of Yinzer syndrome, at 1 in

1,000. Recalculate the posterior probability for these individuals. Show your work.

YOUR ANSWER HERE