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

Assignment 3: Using ML for Clinical Operations

Out: Thurs Sep 26 2024

Due: Thurs Oct 8 2024

This assignment counts for 10 points out of the 35 total points allocated to the course problem sets.

In this assignment, we'll explore various clinical datasets and apply them towards a few different important problems in healthcare operations: (1) how long we expect an admitted patient to stay in the hospital (2) where we expect to send the patient after they are discharged (3) forecasting hospital admissions.

Part 1: Length-of-Stay Prediction (3 points)

In this part, we'll build a model that predicts the length-of-stay (LOS) for patients admitted to the hospital.

LOS prediction is important for many reasons, including planning resource needs (beds, nurses, etc.) and to allow care management teams to coordinate discharge activities when patients leave the hospital.

We'll use the MIMIC-III and MIMIC-IV dataset. The MIMIC-III dataset contains contains de-identified clinical data of patients admitted to the intensive care units (ICU) at the Beth Israel Deaconess Medical Center in Boston from 2001-2012. MIMIC-IV has a somewhat different schema, richer data, and covers the years 2008 to 2019.

```
# collecting (most) of the imports for this assignment in one place. You may not end up using all of these, and
# you may need others not listed here.

from google.colab import auth
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, precision_score, recall_score, roc_auc_score, roc_curve, confusion_matrix, f1_score
from sklearn.linear_model import LogisticRegression
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.preprocessing import StandardScaler

from sklearn.feature_extraction.text import CountVectorizer


auth.authenticate_user()


# for this section, we'll use this file only. This data, originally used in MIT's 6.7930 ML in Healthcare course,
# was derived from the MIMIC-III dataset and includes lab values taken when the patient was admitted to the hospital.

!gsutil cp gs://10-742/assignment_3/length_of_stay.csv ./
```

1.1

What is the average and median length of stay?

YOUR ANSWER HERE

1.2

Plot a histogram of the length of stays. What observations can you make about this distribution? Select what you consider to be the best visualization for this data, given the distribution.

YOUR ANSWER HERE

1.3

We're now going to train a binary classifier model to predict whether a patient stays more or less than 10 days. Follow these steps:

Step 1: Create a binary 'length of stay is 10 or more days' field - this will be the target variable.

Step 2: Remove outliers -- i.e. those encounters with a length of stay exceeding 50 days.

Step 3: Scale all fields to the range [0,1]

Step 4: Split the data 80/20 into training and test. There's lots of ways to perform this split, so try to split the data in a way that does not introduce bias.

Step 5: Run a logistic regression using L2 loss to predict a patient's length of stay. Note that there are no missing values in this dataset, and all the fields are numerical.

YOUR ANSWER HERE

1.4

Calculate and report the accuracy and AUROC of your model. Compare it against the naive (baseline) classifier, which always assigns the label (0 or 1) with higher frequency in the training data.

YOUR ANSWER HERE

1.5

Show the top 10 most important features. Remember that with logistic regression models, the absolute value of a feature weight is a measure of its predictive importance.

Do these features make intuitive sense?

YOUR ANSWER HERE

1.6 (no credit - just for fun)

In step 3 above, we applied min-max scaling. There are other scaling approaches, e.g. z-scaling, which are more appropriate for data with various distributions.

Inspect the various columns in this dataframe to understand the shape of the data in that column, and apply a scaling technique that is appropriate for that column.

Can you develop a better-performing model?

YOUR ANSWER HERE

Part 2: EDA for Clinical Notes (1 point)

In this set of problems, we'll familiarize ourselves with clinical notes. We'll use the de-identified notes for one MIMIC patient.

```
!gsutil cp gs://10-742/assignment_3/patient_80110_notes.csv ./
```

2.1

Let's examine patient with SUBJECT ID of 80110. Everything you need is in this file: `patient 80110 notes.csv` in the `assignment_3` folder in the GCP bucket named `10-742`.

First let's look at their discharge summary, with a ROW ID of 36482.

2.2

How old is this patient and what is their sex?

YOUR ANSWER HERE

2.3

How long were in they in the hospital for?

YOUR ANSWER HERE

2.4

Why was this patient initially admitted to the ICU?

YOUR ANSWER HERE

2.5

THIS PROBLEM HAS BEEN REMOVED.

YOUR ANSWER HERE

2.6

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. What section in the discharge summary do we see the most overlap with? Are there any detail(s) mentioned in the first two paragraphs that are not mentioned in the discharge summary?

YOUR ANSWER HERE

2.7

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.

YOUR ANSWER HERE

Part 3: Using Text to Predict Discharge Location (3 points)

Now we'll use the text in a patient's admission note (the first note that's written for a patient when they are admitted to the hospital) to predict where the patient will end up after they leave the hospital.

Predicting where the hospital will discharge the patient - to their home, a post-acute care facility, hospice - is as important as the LOS, sincemany of these facilities require lead time to prepare for a new patient.

```
!gsutil cp gs://10-742/assignment_3/notes_pruned.csv ./
df = pd.read_csv('notes_pruned.csv')
```

3.1

When building a model to predict discharge location, what is the practical benefit of using only admissions notes and not other notes during a patient's stay in the hospital? (Hint: why do we not also use discharge summaries?)

YOUR ANSWER HERE

3.2

Clean the data as follows:

- Add a `CLEAN_NOTE` field, which contains the `NOTE` text after converting the text to lowercase, removing punctuation, and filtering out stopwords. We suggest using the `stopwords` library from `nltk.corpus`.
- Add a `LONG_STAY` boolean field (which takes the value 1 when los > 6, and 0 otherwise)
- Remove those records with a discharge location that appears fewer than 500 times in the data.

You should have five remaining discharge locations. Show a bar graph of the distribution.

YOUR ANSWER HERE

3.3

Train a Naive Bayes classifier to predict the discharge location from the text.

Hint: we observed that using the `TfidfVectorizer` is preferable to the 'vanilla' Vectorizer. (Why do you think this is?) You may wish to instantiate the vectorizer as follows:

```
vectorizer = TfidfVectorizer(stop_words='english', min_df=3, max_df=0.9)
```

We trained on 70% of the supplied data and evaluated on the remaining 30%. On that test set, we got an overall accuracy of 0.40. How did you do?

YOUR ANSWER HERE

3.4

For each class (i.e. each value of `DISCHARGE LOCATION`), list the top 8 most discriminating features for predicting that class.

We'll do this two different ways.

First, for each class, show the 8 tokens t for which `p(t|class)` is highest. What do you notice about these lists?

YOUR ANSWER HERE

3.5

Those are not very informative features. Let's try a different approach. Find the values of `t` for which `p(t | class)` is unusually high for one class ``c, compared to all the others. In other words, find the top 8 values of `t` which maximize

$$\frac{p(t|c)}{p(t)}$$

YOUR ANSWER HERE

3.6

Why do you think the second approach yield more intuitive results?

YOUR ANSWER HERE

Part 4: Demand Forecasting (3 points)

Congratulations! You have just been appointed VP of Operations at Springfield General Hospital. It's your responsibility to ensure the hospital is properly staffed for the patient load at all times during the day.

An understaffed hospital can be dangerous, as patients are forced to wait for care. An overstaffed hospital is financially unsustainable, as the hospital is spending more on staff costs than necessary.

In this set of questions, we're going to inspect some real-world hospital admission data and attempt to build a model from it, so that we can better predict future demand and thus proactively "right-size" the staffing for our hospital.

Here we're using MIMIC data. For deidentification purposes, MIMIC dates are shifted. While there's no guarantee that hours and minutes in the MIMIC data are unmasked, their values do appear to follow expected patterns and we will take the hour and minute data at face value for this assigment.

Demand forecasting is an entire field of study within operations research---see here if you're interested in learning more.

4.0

Download the following file from Physionet to your computer, uncompress it, and then upload the resulting file here (colab). There is a file icon on the left panel of colab - you can use this to upload files from your computer to colab.

```
https://physionet.org/content/mimiciii/1.4/ADMISSIONS.csv.gz
```

```
# In case physionet server is unresponsive, we've placed the required MIMIC file
# in a private GCP bucket. We will only grant access to this bucket (a) in case
# the physionet server is down, and (b) only to those students who have proven
# to the course staff that they have been granted access to the MIMIC III and IV
# file repositories on physionet.

#!gsutil cp gs://10-742-mimic/ADMISSIONS.csv ./


admissions = pd.read_csv('ADMISSIONS.csv')
admissions['ADMITTIME'] = pd.to_datetime(admissions['ADMITTIME'])
```

4.1

Have a close look at the first 20 rows of this data, and make sure you understand all the columns.

NO ANSWER REQUIRED

4.2

Plot the admission data by hour of day. What do you observe about the shape of the data?

YOUR ANSWER HERE

4.3

To us, the data looks like a mixture of Gaussians: one distribution with a peak at 5PM and a broad variance, the other distribution with a peak (really, a spike) at 7am and very small variance.

The probability density function $f(x)$ of this mixture model can be expressed as:

$f(x) = \alpha \cdot N_b(x|\mu_1, \sigma_1) + (1 - \alpha) \cdot N_n(x|\mu_2, \sigma_2)$

where:

- α is the mixing parameter representing the proportion of the two Gaussian distributions;
- μ_1, σ_1 are the mean and standard deviation of the broad distribution N_b;
- μ_2, σ_2 are the mean and standard deviation of the narrow distribution N_n.

Given the observed data, we can find a maximum-likelihood estimate for these parameters using the Expectation-Maximization (EM) algorithm. For more on the EM algorithm, see: https://www.cs.toronto.edu/~jlucas/teaching/csc411/lectures/lec15_16_handout.pdf

Use the EM algorithm and the admission data to determine maximum likelihood values for $\alpha, \mu_1, \mu_2, \sigma_1, \sigma_2$. Plot the function $f(x)$ above, using the ML parameters you calculated, on top of the observed admission data.

Report the log-likelihood of the data, given your model. For reference, we got -3.2.

4.4

Can you think of any plausible reason why the data has this bimodal shape?

YOUR ANSWER HERE

4.5

In your role as VP Operations, how might you use this newly-developed model to optimize staffing and resources at the hospital?

List up to 3 factors that might complicate your attempt to optimize staffing to match the demand pattern. We'll give you one to start: hospital staff tend to work in shifts of 8 continuous hours.

YOUR ANSWER HERE