

Homework 1

Deadline: Friday, Jan 31, at 11:59pm

Submission: Please submit all code that you wrote and output of your scripts (including diagrams) as a zip file (`code.zip`) and a written report (`report.pdf`) to Markus. If you use Jupyter notebooks, submit both the ipynb file and generate an html file using nbconvert.

Late Submission: Late assignments will not be accepted without a valid medical certificate or other documentation of an emergency.

Collaboration: What you turn in must be your own work. You may not work with anyone else on any of the problems in this assignment. If you need assistance, ask in the Piazza board for this course, or contact the instructor or TA.

1. **[0pts] Preliminaries.** In this assignment, you will be working with the MIMIC dataset (<https://mimic.physionet.org/>) which consists of about 40,000 patients admitted to ICUs in the Beth Israel Deaconess Medical Center in Boston between 2001 and 2012.
 - (a) Before you can download the dataset, you need to complete an online ethics course, “Data or Specimens Only Research” and register an account on PhysioNet. It will take several days to approve your request. Detailed instructions are given here: <https://mimic.physionet.org/gettingstarted/access/>
 - (b) Next, download the dataset and import it into a PostgreSQL database. This takes several hours; make sure you have at least 100GB of free disk space. The data should be access restricted. I.e. a personal computer with a password, or in a folder on a server that only you can access. Detailed instructions: <https://mimic.physionet.org/tutorials/install-mimic-locally-ubuntu/>
2. **[4pts] Getting Started with MIMIC.** To get warmed up with MIMIC, explore the hospital experience of patient ID 40080. For each question, give the SQL query you used.
 - (a) **[1pt]** Give a brief summary of the patient’s demographics (race, age, marital status, etc).
 - (b) **[1pt]** What was the patient’s primary diagnosis (`seq`num = 1`) and the ICD-9 code?
 - (c) **[1pt]** How long did the patient stay in the ICU? According to the discharge report, what was her condition when she was discharged?
 - (d) **[1pt]** What was the patient’s highest and lowest heart rates during the stay?
3. **[6pts] More Data Exploration.** Now that we can query some basic details about individual patients, let’s take a deeper look at the database. This will get you familiar with accessing information in different tables and creating graphs that will be useful in determining the best way to preprocess your data for your own experiments. Remember to label your axes and plots!
 - (a) **[2pts]** Create a histogram plot of the distribution of patient ages in years. Note that you may have to merge tables. You may notice some patients with abnormal or nonsensical ages due to the neonatal ICU and the HIPAA-compliant obscuring of ages for people over 90. What does this plot tell us about our patients?
 - (b) **[2pts]** Create a histogram plot of the distribution of patient heart rates. Use only the first occurrence of each patient’s heart rate. What does this plot tell us about our patients?
 - (c) **[2pts]** Create a scatter plot of patient heart rate vs age. Use cutoff values of [0-400) for age and [30-250) for heart rate to remove outliers. What does this plot tell us about our patients?

4. [10pts] **Predicting Mortality in the ICU.** In this part, you will train models to predict mortality in ICUs using data from the first 48 hours of the ICU stay and clinical notes. First, run the script to preprocess data from MIMIC (https://cs2541-ml4h2019.github.io/problem_set/mort_icu_script.py) which generates the following files:

- `adult_icu.gz`: tabular data from adult ICUs
- `adult_notes.gz`: clinical notes from adult ICUs

The binary target variable to predict is `mort_icu`, which is 1 if the patient died in the ICU. Use the last column `train=1` for training data, and `train=0` for testing. The rest of the data contains variables such as the heart rate, blood pressure, respiratory rate, and other variables, measured in the first 48 hours.

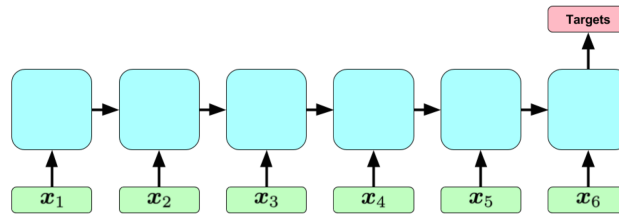
Note about preprocessing: The datasets that you worked with were preprocessed such that missing values were imputed. Specifically, we stratified the data by gender and age groups and replaced missing values with their group means. The topic of missing data is important and will be discussed further in class.

- (a) [4pts] Train a logistic regression model (using scikit-learn) to predict in-ICU mortality from all of the variables. Use L2 regularization.
- Plot the ROC graph and compute the AUC score.
 - Comment on the model performance.
 - Look at the top 5 risk factors of mortality and the lowest 5 and explain what they mean.
- (Hint: in order to compare the coefficients of the features included here regardless of their scale, it is useful to standardize the non-binary variables).*
- (b) [4pts] Use the clinical notes text data to train a bag-of-words model to predict in-ICU mortality. First, use NLTK to tokenize and remove stopwords and punctuation. Then use TF-IDF to transform bag-of-words counts into numerical features and train a L1 regularized logistic regression to predict mortality. Report the top 5 words associated with a high risk of mortality and the lowest 5. Plot the ROC graph and compute the AUC score.
- (c) [2pts] Combine the models in parts 2a and 2b to predict mortality using both clinical notes and tabular data. Plot the ROC graph and compute the AUC score. Comment on how the two sources of data affect the model.

5. [10pts] **Predicting Hypertension with LSTMs.** In this part, you will train models to predict hypertension using sequential chart data. First, run the script to preprocess data from MIMIC (https://cs2541-ml4h2019.github.io/problem_set/make_hypertension_data.py) which generates the following files:

- `hypertension_patients.gz`: contains for each `subject_id` and `hadm_id`, a column `hypertension` which is 1 if the patient has hypertension. As with before, use the last column `train=1` for training data, and `train=0` for testing.
- `hypertension_charts.gz`: contains measurements of various instruments. Each row is one measurement. The `itemid` determines which type of measurement it is, and can be one of the following types:
 - Heart rate in bpm (`itemid = 220045`)
 - Respiratory rate in breaths / minute (`itemid = 220210`)
 - O2 saturation in % (`itemid = 220277`)
 - Blood pressure in mmHg (`itemid = 220181`)

- (a) [4pts] As a baseline, train a logistic regression to predict hypertension, using the min, max, and mean heart rate as features. Remove patients with fewer than two heart rate measurements. Report the AUC and F1 scores on the test set. Then, do the same for respiratory rate, O2 saturation, and blood pressure.
- (b) [4pts] Train an LSTM (using Keras or your favorite machine learning library) to predict hypertension, using the heart rate input sequence as input, and a dense layer at the last step to produce a binary classification (see diagram). Remove patients with fewer than two heart rate measurements. Report the AUC and F1 scores on the test set. Then, do the same for respiratory rate, O2 saturation, and blood pressure.



- (c) [2pts] Comment on the performance of your models. Which measurements are the most useful for predicting hypertension? Do the results make sense?