6.S897/HST.S56 Problem Set 4
Released: Tues Mar 12
Due: Tues Mar 19 by 11:59pm through Stellar

Instructions

You will find a *psets*/*pset4* directory in the class Github. Note this is different from prior problem sets. This repo contains the write-up, starter notebook, and the code used to extract the data. The data is available on Physionet after logging in through "Authorized users may enter here.". Specifically you will need **demo.csv** and **vitals.csv** from Physionet and **train_ids.pk**, **valid_ids.pk**, and **test_ids.pk** from the Github. Please let us know ASAP if you have any issues accessing the data.

In the body of this exercise, blue text describes what specifically you should be submitting. If you have any questions about what the wording of the questions mean, ask us on Piazza!

For your final assessment, you should a standalone report of your analyses with relevant plots, interpretation of findings, and recommendations named ${mit_user}.pdf (e.g. *iychen.pdf*) and any code used for the write-up.

In your write-up, please make your answers as easy to identify as possible (e.g. highlight with colors, label questions with numbers, etc). The faster we can identify your answers, the faster we can grade 70 submissions! :)

Part 1: Physiological signals can have signal [10 points]

We have provided vitals information from the MIMIC-III dataset for adult patients. Data was gathered using **make_vitals.py** which is included in the Github for those curious. In particular, we have provided vitals for patients from the first 24 hours for the following vitals
- Heart rate in bpm
- Respiratory rate in breaths / minute
- Mean arterial pressure
- Pulse oximetry
- Systolic blood pressure in mmHg
- Diastolic blood pressure in mmHg
- Glucose reading
- Temperature

We have aggregated each vital signal into the min/max/mean of that hour. Because not all vital signs are measured at the same rate, missing values are designated as NaN. If no vital signs were measured in that hour, the patient ICU stay will have no row for that hour.

1.1 As with any new dataset, let's get some summary statistics.

- How many rows are in vitals_24h.csv?
- For each of the vital types, how many hours of vital signs are there per patient ICU stay on average (out of total possible 24 hours)? If every patient ICU stay had 24 hours of heart rates, we would say that the average number of hours of heart rates per patient ICU stay is 24.
- Plot the distribution of number of hours of vital measurements across patient ICU stays for each vital sign using histograms with 24 bins. Note that because max, min, and mean are either all measured or all unmeasured, you should provide 8 histograms for the 8 vital signs.

1.2 We want to model patients with ICU stays where we have hourly readings for the first 24 hours. Patients with ICU stays where measures aren't recorded have been marked as null.

- How many patient ICU stays are included in this dataset in any form?
- How many patients ICU stays show up with 24 hours of measurements, even if some of the hours have some missing values?
- How many patients ICU stays have 24 hours measurements with all present measurements?

1.3 In order to increase the number of include patient ICU stays, we want to keep vital readings that appear >80% of the time over all rows in vitals_24h.csv. Which vitals would we keep?

1.4 For the rest of the problem set, we will keep vitals that appear >80% of the time across all rows and then remove any patient ICU stays that have any missing values. This is only one way to handle missing data. Besides throwing out rows that are incomplete, what are three other ways we can impute missing data for multivariate time series modeling?

Part 2: Long short term memory [20 points]

2.1 We are interested in using this time series information to model patient care. Keep vitals that appear in >80% the rows. (As a sanity check, this should include 18 columns.) Remove any patients ICU stays that have any missing values with these remaining vitals. How many patient ICU stays are remaining in the dataset?

We want to analyze patient ICU stays that last longer than 48 hours by predicting from the first 24 hours, but vitals_24h.csv includes vitals from all patients in the MIMIC dataset. Join with demo.csv, which only includes patients ICU stays that last longer than 48 hours. How many patients are left in the dataset now?

Create a data tensor of shape: [n_samples, n_timesteps, n_features] for data labeled train, valid, and test. Data should be scaled using MinMaxScaler from sklearn using feature_range=(0,1) across the entire dataset before being separated into train, valid, and test.

You may find it helpful to create column names of the form: t00_heartrate_min, t00_heartrate_max, …, t23_heartrate_max, …

We often find that sex and age are predictive features. Encode sex as a “is_female” feature. Include age and sex in the tensor: t00_is_female, t00_age, …, t23_is_female, t23_age. Note that t00_age, …, t23_age will all be the same and similar with the sex features. These features can be found in demo.csv. Age should also be scaled using MinMaxScaler.

We have given you the icustay_ids corresponding to the train, valid, and test sets in **train_ids.pk**, **valid_ids.pk**, and **test_ids.pk** , which you can load using the pickle package in Python. If you are having trouble loading these IDs, or they do not match your dataframe, make a note in your write-up and describe your process of randomly selecting training/validation/testing data in a 60/20/20 split.

2.2 Build an LSTM using Keras for multivariate time series classification, predicting hospital mortality or mort_hosp. Use an LSTM with:

- 20 units
- a dense layer with sigmoid activation
- binary cross-entropy loss
- ADAM optimizer
- Dropout of 0.2
- 100 epochs
- Batch size of 128

You may find these guides helpful:

- Multivariate time series forecasting for regression
- Sequence classification using LSTMs

What is the train loss, accuracy, and AUC? Experiment with Dropout = [0.0,0.2,0.5, 0.8] by comparing the AUC on the validation data. What Dropout value yields the best AUC on the validation data? Using this best Dropout value, what is the corresponding test AUC?

2.3 Select two patient ICU stays for whom the LSTM predicts the patient ICU stay will end in patient mortality and two patient ICU stays for whom the LSTM predicts the patient ICU stay will NOT end in patient mortality. For each vital feature, plot the mean values over 24 hours for all four selected patient ICU stays. This will create 6 plots with 4 lines on each plot. Describe any meaningful differences between the patient ICU stays that are predicted to have HIGH likelihood of patient mortality and the patient ICU stays that are predicted to have LOW likelihood of patient mortality.

Part 3: Compare to linear baseline [10 points]

Whenever we use more sophisticated deep learning models, we want to compare against a simpler baseline to make sure the computational effort is worth it.

3.1 What is Logistic Regression with just the raw values from 2.2? You should use a matrix of shape [n_shapes, n_timesteps * n_features], which is an unrolled version of the data tensor from earlier. You can keep age and sex in their time-multiplied form or only include them once.

3.2 Hyperparameter tune on C=[0.01,0.1,0.5,1.0, 5.0] and penalty=['l1','l2'] by comparing the AUC on the validation data. What are the best hyperparameters on the validation data? What is the test AUC using these best hyperparameters? How does this compare to the results from 2.2?

Part 4: Not graded by please answer [0 points]

4.1 How many hours did you spend on this problem set?
4.2 This past week (Tues March 12 - Tues March 19), how many hours did you spend (attending lectures, reading papers, problem sets) on the class?