AIA Machine Learning Capstone

Results: Due 11:45 PM EST on Dec 9th;
Final Report Due: Due 11:45 PM EST on Dec 15th, 2022
No extension

Instructions:

1. You MUST clearly list **your all team members' names for those who have contributed to the Capstone project** at the top of your submission.

2. Submit your results in Jupyter notebook (Due Due 11:45 PM EST on Dec 9th; no extension).

3. Submit your results and conclusions in Jupiter Notebook (Due 11:45 PM EST on Dec 15th; no extension).

4. By uploading your submission, you agree that you have not violated any university policies related to the student code of conduct (https://policies.ncsu.edu/policy/pol-11-35-01/), and you are signing the Pack Pledge: **"I have neither given nor received unauthorized aid on this test or assignment"**.

1 Background & Task:

Sepsis is the leading cause of mortality in the United States and the most expensive condition associated with in-hospital stay, accounting for 6.2% (nearly $24 billion) of total hospital costs. In particular, *Septic shock*, the most advanced complication of sepsis due to severe abnormalities of circulation and/or cellular metabolism, reaches a mortality rate as high as 50% and the annual incidence keeps rising. It is estimated that as many as 80% of sepsis deaths could be prevented with early diagnosis and intervention; indeed prior studies have demonstrated that *early diagnosis* and treatment of septic shock can significantly decrease patients' mortality and shorten their length of stay. In this capstone, your task is to build a machine learning model for accurate early diagnosis of septic shock.

2   Methods:

You will explore two categories of Machine Learning methods: one is **Recent Temporal Patterns (RTPs) mining** which will be used in conjunction with a Support Vector Machine (SVM) model and a Logistic Regression (LR) model (i.e., RTP-SVM and RTP-LR); the other is a powerful deep learning model named **Long Short Term Memory (LSTM)**.

3   MIMIC-III Datasets:

Our dataset is a subset of **MIMIC-III** dataset. [1] The original MIMIC-III dataset contains admissions of adult patients (i.e., age > 16) who are admitted to intensive care units (ICU) at a tertiary referral hospital between 2001 and 2012, corresponding to 53,423 visits containing ∼11 million events.

3.1   Handling the Missing Data

The events in each visit were collected with irregular intervals, ranging from 0.94 seconds to 28.19 hours. Specifically, different features are measured with varying frequencies, which causes some features to be unavailable and missing from certain events. On average, the missing rate of our data is ∼80.37%. Herein, we fill the missing values using expert-designed rules: the last value within a fixed-length window (i.e., 24 hours for lab analyses and 8 hours for vital signs). Then it fills the remaining with the mean value of the corresponding features.

3.2   Selecting Features

Five features related to the sepsis progression were selected as suggested by clinicians:

1. Systolic Blood Pressure (SystolicBP)

2. Heart Rate (HeartRate)

3. Respiratory Rate (RespiratoryRate)

4. Temperature

5. White Blood Cell(WBC)

3.3   Tagging the Septic Shock Visits

Identifying the septic shock visits is a challenging task. Though the diagnosis codes, e.g., ICD-9, are widely used for clinical labeling, solely relying on the codes can be problematic: they have proven to be limited in reliability since the coding practice is mainly used for administrative and billing purposes. Based on the Third International Consensus Definitions for Sepsis and Septic Shock, our domain experts identified septic shock when either of the following two conditions was

---

[1]Dataset is available at http://mimic.physionet.org.

met: 1) Persistent hypertension through two consecutive readings (≤ 30 minutes apart), including SBP < 90 mmHg, MAP < 65mmHg, and decrease in SBP ≥ 40 mmHg within an 8-hour period; or 2) Any vasopressor administration. By combing both ICD-9 and the domain experts' rules, we identify 4,918 visits (2,459 shock-positive visits and 2,459 negative visits) containing ~ 1 million events from MIMIC-III.

3.4 Prediction Task and Final Dataset

Our goal for this Capstone project is: based on observation of a patient's visit until 24 hours before an endpoint, we will predict whether or not a patient will develop septic shock 24 hours later. For septic shock visits, the endpoint is the onset time of septic shock while for non-shock visits, the endpoint is the end of sequences.

To conduct this task, we *right aligned* all the shock sequences by septic shock onset and all non-shock by the end of their sequences and include all the EHRs until 24 hours before the end of sequences (see Fig. 1).

As shown in Figure 1, the 24 hours leading up to the endpoint is denoted as *prediction window* and before that, is denoted as *observation window*. So the visits that contain no events in the observation window are excluded from the septic shock prediction task. As a result, the final datasets, **MIMIC_III_shock.csv** and **MIMIC_III_nonshock.csv**, have 796 visits (398 shock-positive visits and 398 negative visits) containing 37,246 events. Notably, MIMIC-III is ICU-specific and sepsis onset can present itself early, thus the number of remaining visits in MIMIC-III is relatively small after applying a 24-hour cut before the onset.
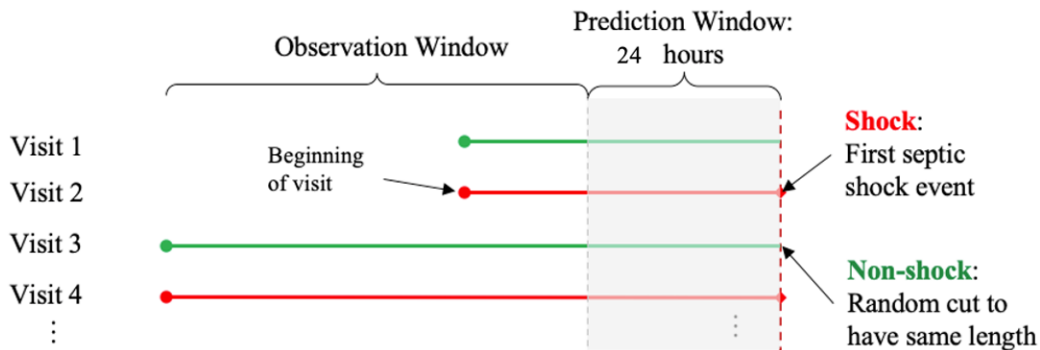

Figure 1: Illustration of septic shock early prediction setting.

4 Experiment Setup Settings

You will employ RTP and long short-term memory (LSTM) as the prediction model since extensive previous works have demonstrated their robust and superior performance in EHRs modeling. **In the following, seed the Random Number Generator with "2022" for reproducibility (you should get the exact same results each time).**

4.1 RTP Mining

For RTP Mining, please explore the following parameters:

1. Maximum gap: explore between 4 and 10 hours, in increments of 1;

2. Support for the positive cases (shock): explore between 0.1 and 0.3, in increments of 0.05

3. Support for negative cases(non-shock): explore between 0.1 and 0.3, in increments of 0.05.

After retrieving the binary matrix for the train and test data, use the train data to fit a Support Vector Machine model and a Logistic Regression model (two independent models).

4.1.1 RTP-SVM

For the Support Vector Machine, please use the following parameters:

1. $C = 1.0$

2. $\gamma = 0.1$

3. Try different kernels: 'linear', 'poly', 'rbf', 'sigmoid', but report the one you used and the final results with that chosen kernel.

4.1.2 RTP-LR

For the Logistic Regression, please use the default parameters in sklearn:

- $l2$ penalty

- dual = False

- tol = $1e - 4$

- $C = 1.0$

- LBFGS solver

- Max iterations = 100

4.2 Long Short Term Memory (LSTM)

For LSTM, please do the following:

1. Apply both padding and masking to the sequences. For masking, please use -10 as the fill value.

2. Create the model (using keras.layers) by first defining a Sequential model, then add a Masking layer, followed by a LSTM layer and lastly a Dense layer with 1 output neuron using sigmoid activation function. The LSTM layer should have 200 neurons within it. Use an Adam optimizer with a learning rate of $1e - 2$. Compile the (Sequential) model with a Binary Cross-entropy loss function, pass in the Adam optimizer.

3. Fit the Sequential model (should contain 3 layers: Masking $\longrightarrow$ LSTM [200 units] $\longrightarrow$ Dense [1 unit, the output neuron]) with 10 epochs and report the performance when the batch size is 64, 128, and 256 respectively. In other words, there should be a different LSTM model for each possible batch size (i.e., LSTM-64, LSTM-128, LSTM-256).

4.3 Evaluation Metrics

All models were evaluated by a **5-fold cross-validation**. The evaluation metrics of Accuracy, Recall, Precision, F1-score, and AUC under ROC will be employed.

5 Results (Dec 9th) & Final Reports (Dec 15th): Evaluation Metrics

Submit your results (due 11:45PM EST Dec 9th): In your Jupyter notebook, please report

- The best final RTP-LR model's settings and/or hyper-parameters.

- The best final RTP-SVM model's settings and/or hyper-parameters.

- The best final LSTM model's settings and/or hyper-parameters.

- Present and compare the RTP-LR, RTP-SVM, LSTM's results using the evaluation metrics: Accuracy, Recall, Precision, F1-score, and AUC under ROC.

Final Report (Due 11:45PM EST Dec 15th): In your Jupyter notebook, please include 1) all the results above, and 2) **write one or two paragraphs (total words number ≤ 250)** to critically evaluate the results of the experiments and describe the lessons learned: What specific things did you learn from doing this project? Did you learn anything about the problem itself, the approaches you tried, or the experiments you conducted? If there are any good ideas you came up with in the end but did not have time to pursue, this would be a good place to mention them.