STATS 503: Data Challenge

Yupeng Yang

1 Model Assumption

1. The physiological status of the patients over consecutive periods does not change significantly.
2. The physiological status of the patients is centered on mean values.
3. The original sample patients is a representative of the population.
4. The later physiological status is more important to model than the former physiological status of the patients.

2 Model Structure

The model mainly consists of a Long-Short Term Memory (LSTM) neural network and several auxiliary methods. This modified LSTM model consists of several parts (see Fig 1). At first, the pre-processing part converts the raw data into structured data and does forward filling. Then, the whole data set will be split into the training and validation sets for cross-validation. Later, the imputation, scaling, and PCA parts improve the usability of the data so that the neural network can properly operate the algorithm. Finally, the LSTM will take a 3-dimensional tensor (number of patients $\times$ sequence length $\times$ number of features) and provides the result with the best hyperparameter set given by the cross-validation. See the next section for more information about why I constructed the model this way.

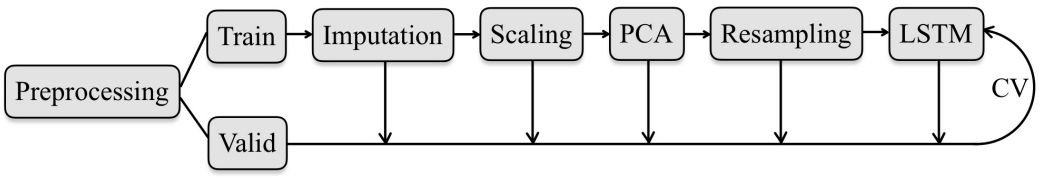

Figure 1: Model Structure

3 Model Details

Before the model fitting, a pre-processing procedure that imputes the time series value is applied. Based on assumption 1, I use forward fill (fill NA with the last valid non-NA) to impute missing values. After that, I apply

backward fill when the column has started missing values. Then, I split the whole data set into the training, validation, and test sets because all the following steps are considered part of the model.

In the formal modeling part. I first do imputation with "training set means" based on assumptions 2 and 3. Then, I use the standard scaling method and the principle component analysis (PCA). However, the fully connected layer (FC layer) can capture the linear relationship of the data. Doing PCA before the neural network will simplify the neural network structure, add numerical stability and reduce the running time potentially. Since the outcomes of the training set are highly unbalanced, I apply the oversampling method to the minority class to ensure the model absorbs both classes' information well. Before the LSTM model, I used padding and packing to the time series data to ensure all the patients had data with the same dimension. If the patient has fewer records than needed, I will add 0 at the end of the sequence (padding) because 0 will not affect the weights in neural network training. If the patient has more records than needed, I will drop all but the latest records (packing) based on assumption 4.

Table 1: Top 5 model in 10-fold Cross Validation with Grid Search (BAR: Balanced Accuracy Rate)

Seq_Len	Hidden_Dim	Layer_Dim	Batch_Size	Learning_Rate	BAR	AUC
12	250	2	32	10e-3	0.8243	0.8964
12	200	2	32	10e-3	0.8224	0.8991
12	300	4	32	10e-3	0.8237	0.8959
16	300	3	64	10e-2	0.8207	0.8959
16	300	4	64	10e-2	0.8205	0.8956

I use the Pytorch library to construct the LSTM model. Besides the typical setup, I apply dropout and early stopping to avoid overfitting. In addition, a manual grid search together with 10-fold cross-validation (see Table 1) is used to select the best hyperparameter set. To prevent potential overfitting and reduce the variance, I choose the simplest model among the top 5. After plugging in the best hyperparameters, I reuse the 10-fold to fit ten different models. Finally, I take the average of 10 models as my final prediction result to reduce the prediction variance.