Biomedical Data Science Lab
Fall 2019

Prediction of the Impending Onset of Septic Shock in Patients with Sepsis

Benjamín Béjar Haro
Assistant Research Professor
Department of Biomedical Engineering
319B Clark Hall, Johns Hopkins University
E-mail: bbejar@jhu.edu

Final Project Description

Sepsis is a life-threatening condition caused by an inflammatory immune response to an infection. It is the leading cause of death in hospitals and has a greater risk of mortality in its advanced state, also called *Septic Shock*. Early treatment of Septic Shock can dramatically increase the survival rate. Therefore, a prediction system capable of foreseeing Septic Shock onset would provide an early intervention window that has the potential to translate into improved patient outcomes. In this project we look at the problem of early prediction of Septic Shock following the approach described in:

> R. Liu, J. L. Greenstein, S. J. Granite, J. C. Fackler, M. M. Bembea, S. V. Sarma, and R. L. Winslow. "Data-driven discovery of a novel sepsis pre-shock state predicts impending septic shock in the ICU". *Scientific Reports*, 9(1):6145, 2019.

Your goal in this project is to reproduce some of the results in above paper. In particular, you are asked to fit a logistic regression model (referred to as GLM in the paper) to the provided dataset in order to predict the onset of Septic Shock. You should compute the Receiver Operating Characteristics (ROC) curve of your predictor as well as a histogram of the Early Warning Time (EWT) (*i.e.*, similar to those in Fig. 4 and 5 of (Liu et al., 2019), respectively). You should implement the prediction method in Python. You must turn in your code along with a short report of two to three pages describing the problem at hand, the implemented solution, as well as the results obtained.

Dataset

You will be provided with the curated data used in (Liu et al., 2019) which is a subset of the MIMIC-III database (Johnson et al., 2016). The data corresponds to electronic health record data of a large population of patients, and consists of measured values over time for 28 different indicators such as, heart rate, blood pressure, respiratory rate, temperature, etc. Each data point represents a particular measurement in time and for a particular patient. The data has been split into training and testing as described in (Liu et al., 2019) and is provided to you in the form of `.csv` files. Inside those files `'x'` columns correspond to feature values while the `'y'` column represents the associated label of a particular row of feature values (*i.e.*, $y = 0$ means that the patient didn't go into Septic Shock, while a label $y = 1$ indicates that the patient eventually went into Septic Shock).

Tasks

(a) Read the paper (Liu et al., 2019) to familiarize yourself with the problem and the approach to be followed.

(b) *[30 points]* Train a generalized linear model (*i.e.,* logistic regression) with lasso regularization (Friedman et al., 2010) on the provided training data. You might want to use `'pandas'` Python module and DataFrames for loading data from large `.csv` files. Follow the same pre-processing (if any) as in (Liu et al., 2019). You should use a stochastic optimization strategy by splitting your dataset into smaller chunks that you feed to the optimizer. You are allowed to use any libraries you might consider appropriate (*e.g.*, scikit-learn (Pedregosa et al., 2011)) for solving the regression task.

(c) *[20 points]* Determine the optimal regularization value $\lambda \geq 0$ via cross-validation as described in (Liu et al., 2019).

(d) *[20 points]* Evaluate the ROC curve on the test data and compute the operating point corresponding to the optimal threshold as defined in (Liu et al., 2019). Plot the obtained ROC curve overlaying the operating point and its associated probability of detection (True Positive Rate) and probability of false alarm (False Positive Rate). Add also to the figure the Area Under the Curve (AUC) value.

(e) *[10 points]* Generate a histogram of EWT values on the test data.

(f) *[20 points]* Write a short 2 to 3 pages report using the provided templates where you introduce the problem, describe the training procedure and methodology used, as well as a description of the obtained results. We will be evaluating your clarity in exposition and presentation of the research problem.

(g) *[10 points] (Optional)* Repeat the training and estimation procedure using data balancing and compare the results.

Submission Instructions

Submit through Blackboard your code and report as a single `.zip` file with the following naming convention `'final-project-JHED1-JHED2.zip'`. Your zipped files can only be IPython Notebooks `.ipynb` for code and `.pdf` documents for the report. No other file formats will be accepted. **Due date** for submission is:

October 17, 2019 at 11:59pm

References

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, Articles*, 33(1):1–22, 2010. ISSN 1548-7660. doi: 10.18637/jss.v033.i01. URL https://www.jstatsoft.org/v033/i01.

A. E. W. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):160035, 2016. doi: 10.1038/sdata.2016.35. URL https://doi.org/10.1038/sdata.2016.35.

R. Liu, J. L. Greenstein, S. J. Granite, J. C. Fackler, M. M. Bembea, S. V. Sarma, and R. L. Winslow. Data-driven discovery of a novel sepsis pre-shock state predicts impending septic shock in the icu. *Scientific Reports*, 9(1):6145, 2019. doi: 10.1038/s41598-019-42637-5. URL https://doi.org/10.1038/s41598-019-42637-5.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.