CSE8803: Project - Septic shock prediction

Jeffrey Skonhovd

Georgia Institute of Technology

December 4, 2016

Outline

Abstract

- Sepsis is an extremely serious medical condition
  - that occurs in emergency rooms across the country.
- Developed a machine learning application
  - attemps to help predict the onset of septic shock.
  - 24 hours

Introduction

- Read paper and implement algorithms associated
  - A targeted real-time early warning score (TREWScore) for septic shock.
    - by K. E. Henry
- Best Attempt at replicatation of the results

Background

Background on my Self

- A formal background in Apache Spark
  - obtained in this course
- Have taken other machine learning courses

Background on Septic Shock

- Sepsis is an extremely serious condition
  - Seven hundred and Fifty thousand patients
  - Can quickly become fatal
- Easy motivation
  - Early prevention is key
  - Help prevent organ damage

Project Execution

- Gather Data
- Study Design
- Modeling Pipeline
  - Predictive Target
  - Cohort Construction
  - Feature Construction
  - Feature Selection
  - Predictive Model
  - Performance Evaluation

Gather Data

- MIMIC III
  - Medical Information Mart for Intensive Care III
  - Beth Israel Deaconess Medical Center
  - 2001 and 2012

Cohort Construction

- Cohort Construction Criteria
  - Age 15 or older based on dob at ICU staytime
  - Must have at least one occurrence of GCS
  - Must have at least one occurrence of BUN
  - Must have at least one occurrence of Hematocrit
  - Must have at least one occurrence of Heart rate

Prediction Target

- Determine if a patient
  - will develop septic shock!
- Positive Prediction Criteria
  - ICD-9 codes indicating info
  - 2 SIRS Indicators
  - Sepsis Related Organ Disfunction

Feature Construction

- Index Date
  - The first entry of Sepsis Related Organ Disfunction
  - if Control, just the date of the last entry.

Feature Selection

- Goal
  - discover the most truly predictive features in the model.
- TREWScore
  - Foundation
  - Averages

Results

- Model
  - Logistic Regression
    - 0.91037402964
  - SVM
    - 0.908357697349
  - Decision Tree
    - 0.912591995161

Conclusion

- How well do our results match the Paper?
  - Models preformed well
  - But Kaggle wasn't very good
  - Did not have features for every patient
- Any pitfalls observed trying to replicate the experiment from the paper?
  - Learning Curve to get started
  - Took a long time to get comfortable