Predicting Intensive Care Unit (ICU) Length of Stay (LOS) Via Support Vector Machine (SVM) Regression

Morgan Cheatham
CLPS1211: Human and Machine Learning
December 15, 2015
INTRODUCTION

Extended inpatient hospital length of stay (LOS) is associated with increased cost, higher readmission rates, and increased mortality, as well as greater probabilities of contracting a hospital-acquired infection (HAI). Each of these metrics represents significant indicators for hospital performance. Thus, LOS serves as a critical proxy for these measures and can provide insight ranging from cost analysis to patient outcomes. This project leverages hospital inpatient data from the MIMIC Critical Care Database to explore support vector machine (SVM) regression models for predicting LOS. In this paper, I will begin with a literature review of research centered on the MIMIC Critical Care Database, establishing a background understanding of the current state of research at the intersection of predictive modeling and hospital inpatient data. Next, I will review the methods utilized in this project, key findings, and recommendations for further investigation.

MIMIC Critical Care Database

The Massachusetts Institute for Technology (MIT) Lab for Computational Physiology created the Multiparameter Intelligent Monitoring in Intensive Care III (MIMIC-III) database. The database is an openly available dataset of health data for over 58,000 hospital admissions in the Critical Care Unit at the Beth Israel Deaconess Medical Center from 2001-2013. This project utilized Version 1.2 of the database, which includes 26 tables that contain information pertaining to patient demographics, vital signs, laboratory tests, medications, diagnoses, and outcomes. The aggregation of the dataset has catalyzed research on developing machine learning methods and employing predictive modeling to healthcare data.

MIMIC Literature Review

A significant portion of the research centered on the MIMIC-III database has focused on developing models for patient mortality, the likelihood of septicemia, and readmission prediction. Health services and policy research pertaining to hospital costs has fueled the popularity of these questions. For instance, septicemia is the primary driver of HAIs and represents the most expensive inpatient hospital cost at $20.3 billion in 2011. The essential question of the research documented in this paper is distinguished in its investigation of LOS, a proxy for many of the models in the existing literature.

The models described in the following literature review have been developed for comparison with existing, healthcare industry-standard predictive models – namely, the Acute Physiology and Chronic Health Evaluation (APACHE) and the Simplified Acute Physiology Score (SAPS) predictive scoring systems in the Intensive Care Unit (ICU). APACHE is a risk assessment and predictive severity of disease classification system that utilizes a point score comprised of physiologic measurements, age, chronic health status, and other variables to provide risk estimates for hospital mortality. Similarly, SAPS is a point score for ICU patient severity based on physiological variables, age, type of admission, and disease-related variables that calculates mortality probability estimates.

A recent study by Mayaud et al. (2013) utilized the MIMIC-III database to develop a prediction model for hospital mortality among sepsis and hypotension patients. The study extracted relevant clinical insights through non-linear transforms of the raw physiological data. In order to prevent over-fitting in the mortality model, the researchers employed a general rule, which asserted that, “the maximum number of predictors to include in a model should be no greater than the number of events (i.e. deaths) in a sample divided by ten.” Through a genetic algorithm (GA), which mimics DNA replication and natural selection, the researchers selected
high performing variables for the model. The model outperformed APACHE and SAPS and was distinguished in its use of dynamic data, such as frequently entered test values and updated vital sign statistics, as opposed to static data.

Additional work at Johns Hopkins University by Henry et al. (2015)\(^{iii}\) leveraged the MIMIC-III database to establish a real-time early warning score (TREWScore) for septic shock. The model implemented a Cox proportional hazards model, where the time until septic shock served as a signal. After employing the supervised learning method over 54 potential features, variables were selected and weighted according to their relation to the advent of septic shock. Thus, the risk of shock at a given time is dependent upon the probability of shock given the time and the weight of the physiological variables. Multiple imputation-based approaches accounted for interval censoring, which may have affected the model parameter estimation. The features were further characterized by time until the onset of septic shock and TREWScores were calculated over time. The model performed well in identifying patients before the onset of septic shock compared to APACHE and SAPS.

Finally, additional research by Lehman et al. (2014)\(^{vi}\) has leveraged the physiologic streams of dynamic data provided by the bedside monitoring technology. Specifically, the study utilized time series data of vital signs to investigate physiologic feedback loops and develop a switching vector autoregressive (SVAR) framework. This framework employs linear dynamical systems that learn time-dependent rules in order to predict the future state of the time series within the context of a system equilibrium point. This model enabled inter-patient comparison of the amount of time spent in given dynamics (modes) of the time series, which facilitated the investigation of physiological control system response to both internal disturbances, such as infection, or external perturbations, such as drug intake. The SVAR framework calculates a vector that contains the values of vital signs at a given time. This framework utilizes a Markovian dynamic process in switching between \( K \) potential modes of physiological status. This library \( K \) of potential modes is calculated via an expectation-maximization process. The model also utilizes a fluctuation term characterized by a Gaussian distribution. Upon tuning the model, researchers analyzed relationships between the dynamic mode composition and hospital mortality through logistic regression and compared values to APACHE and SAPS scores.

These three recent studies have informed this project by providing a specific target population over which to develop the predictive model for LOS, namely, patients with sepsis. Additionally, this project employs a similar conceptual framework to the Henry et al. (2015) research, and essentially strives to provide a “warning score” in terms of LOS. Finally, this project employs dynamic physiological data similar to the work of Lehman et al. (2014), which provides time-series-based insight.

**METHODS**

*Exploratory Methods*

Exploratory tools were applied initially to define the feature space for this project. The starting feature space consisted of ten out of the twelve parameters utilized in the APACHE II classification system, including: oxygen saturation (%), temperature (degrees Celsius), mean arterial blood pressure (mmHg), arterial pH, heart rate (beats per minute), respiratory rate (inspirations per minute), sodium (mEq/L), creatinine (mg/dL), hematocrit (%), and white blood cells (K/\( uL \)). This list excluded serum potassium and the Glasgow coma metric from the APACHE II system, due to insufficient sample sizes for these tests in the MIMIC-III database.
The data was filtered for patients who were diagnosed with Severe Sepsis, a "syndrome of infection complicated by acute organ dysfunction," characterized by the ICD9 code: 99592. Patients who died during their hospital stay were not included. Next, dynamic vital sign and lab test data for the aforementioned ten features were selected for patients who met the established criteria. The first recorded value within a patient’s ICU stay for each feature was captured, representing a given patient’s initial, baseline value upon admission. LOS data was regressed on each of the features (see Appendix Figure A).

The feature space was further narrowed by the sample size of each parameter. Ultimately, four features were selected, due to optimal sample size after pruning (n=1997). Significant data cleaning was required to process the manually entered, error-prone clinical data, in order for analysis with relevant machine learning libraries. As a result of the significant amount of data cleaning required to run the models, imputation of missing values was not performed. Thus, measurement continuity across patients resulted in a mild sample size reduction (i.e. each feature required a value for a given patient in order to comprise a single datum). The four features utilized in this research were: Na, Creatinine, Hematocrit, and White Blood Cells. All sample data was normalized before running the models and outliers were removed.

The sodium and hematocrit plots appeared to follow normal distributions, whereas the creatinine and white blood cell plots seemed to follow γ distributions, where the shape parameter (k) was less than the scale parameter (θ). Thus, the sodium and hematocrit plots suggest that there are initial sodium and creatinine values, respectively, around the mean of the sample population that yield the greatest LOS. The creatinine and white blood cells plots suggest that low initial creatinine and white blood cell values initially will result in longer LOS.

**Model Selection**

The SVM regression models were built via the Scikit-Learn Machine Learning Library and included: Bayesian Ridge Regression, Support Vector Machine (SVM) with a Linear Kernel, and SVM with a Radial Basis Function (RBF) Kernel, or the Gaussian Kernel. For each model, K-Fold Cross Validation was employed for 5 and 10 folds (yielding sub-sample sizes of roughly 400 and 200). The sample size was 1997 (n=1997).

**Bayesian Ridge**

Bayesian Ridge regression builds off of the traditional Ridge regression model, where the ordinary least squares model is penalized via the $l_2$-norm of the parameters. In the Bayesian Ridge regression, uninformative priors were employed over the model hyper-
parameters in order to estimate these values from the training data itself. In this model, the parameter coefficients are driven towards zero in order to stabilize the regression. Parameters \( w, \alpha \) (precision of noise), and \( \lambda \) (precision of weights) were characterized by \( \gamma \) distributions and were estimated by the model. The \( \lambda \) parameter is the coefficient of the \( l^2 \)-norm penalty, where a large value indicates a prior more closely clustered around zero, which increases the extent to which parameters are driven towards zero. The prior parameter, \( w \), followed a spherical Gaussian: \( p(w|\lambda) = N(w|0, \lambda^{-1}I) \). The \( \alpha \) in this model was varied in order to subject different levels of constraints on the model. Variation of this parameter, when small and positive, can reduce the variance of the estimates. Generally, high \( \alpha \) values constrain the model and yield a smooth and relatively flat interpolant. For smaller \( \alpha \) values, the interpolant fits more closely to the data. This model worked by maximizing the marginal log likelihood, in order to estimate the parameters.

SVM

The primary tuning parameter in the linear SVM model was \( C \), while the RBF SVM model parameters included were the \( C \) and \( \gamma \). The \( C \) parameter indicates the penalty of the error term, which represents the trade-off between a simple surface with incorrect classification versus over-fitting the model by increasing the number of possible support vectors. The \( \gamma \) parameter represents the coefficient on the RBF kernel function. These \( \gamma \) values signal the extent to which a single training sample selected as a support vector influences the model. Higher values for this parameter indicate greater influence, with lower values indicating less influence. Finally, the \( \epsilon \) loss function value, which defines the \( \epsilon \) tube in which no penalty is accredited in the model, was kept constant at 0.1.

RESULTS

Bayesian Ridge

<table>
<thead>
<tr>
<th>K-Folds</th>
<th>( \alpha, \lambda )</th>
<th>Regression Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>( 1 \times 10^{-3} )</td>
<td>[-12.63339309 3.36316145 30.08280949 -23.6426708 ]</td>
</tr>
<tr>
<td>5</td>
<td>( 1 \times 10^{-6} )</td>
<td>[-12.63662653 3.3654465 30.0873828 -23.64510533]</td>
</tr>
<tr>
<td>5</td>
<td>( 1 \times 10^{-10} )</td>
<td>[-12.63662976 3.36544879 30.08738738 -23.64510777]</td>
</tr>
<tr>
<td>10</td>
<td>( 1 \times 10^{-3} )</td>
<td>[-13.01415794 7.8135445 29.07650283 -23.63745793]</td>
</tr>
<tr>
<td>10</td>
<td>( 1 \times 10^{-6} )</td>
<td>[-13.01735813 7.81829167 29.08031228 -23.6396198 ]</td>
</tr>
<tr>
<td>10</td>
<td>( 1 \times 10^{-10} )</td>
<td>[-13.01736134 7.81829642 29.08031609 -23.63962196]</td>
</tr>
</tbody>
</table>

Table 1. Bayesian Ridge Regression

Figure 2. Bayesian Ridge Regression, \( K=5, C=1e^{-10} \) (left)
Figure 3. Bayesian Ridge Regression, \( K=10, C=1e^{-10} \) (right)
(see Appendix Figure B for additional Bayesian Ridge Regressions)
Table 2. Support Vector Regression with Linear Kernel

<table>
<thead>
<tr>
<th>K-Folds</th>
<th>C, y</th>
<th>Regression Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>100</td>
<td>[-19.12165738, 26.26932117, 26.42808729, -19.74966858]</td>
</tr>
<tr>
<td>5</td>
<td>1000</td>
<td>[-19.39918408, 47.10911909, 26.84166909, -20.42646531]</td>
</tr>
<tr>
<td>5</td>
<td>5000</td>
<td>[-19.4859157, 50.17760322, 27.04258787, -20.56637206]</td>
</tr>
<tr>
<td>5</td>
<td>10000</td>
<td>[-19.48989888, 50.22050203, 27.03178035, -20.55024612]</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
<td>[-18.13499518, 29.36010426, 20.74691827, -18.49518604]</td>
</tr>
<tr>
<td>10</td>
<td>1000</td>
<td>[-18.73722337, 52.17085618, 22.55484471, -18.6660438]</td>
</tr>
<tr>
<td>10</td>
<td>5000</td>
<td>[-18.70820731, 57.11034529, 22.12730002, -18.59889022]</td>
</tr>
<tr>
<td>10</td>
<td>10000</td>
<td>[-18.68509439, 57.28653079, 22.00407099, -18.58423732]</td>
</tr>
</tbody>
</table>

**DISCUSSION**

Bayesian Ridge

Generally, the Ridge regression performed poorly across this dataset. According to the coefficients listed in Table 1, which are the coefficients for sodium, creatinine, hematocrit, and white blood cells, respectively, none of the parameters were shifted towards zero, nor do they reflect clinical intuition of decreased values for all parameters for a sepsis patient. Smaller $\alpha$ and $\lambda$ values provided a smoother regression than larger values for these parameters, which is intuitive as smaller values for these parameters decrease the variance of the model. The model performed slightly better with 10 folds as opposed to 5, however, the coefficients and plotted regression do not indicate predictive capability of this model.
SVM with Linear Kernel

Linear SVM generally performs better with a large number of features and a small training set. Despite the fact that the data in this project was characterized by few parameters and many data points, linear kernels were included due to their runtime when compared to non-linear kernels, as well as to serve as a comparison for the RBF kernel model. Like the Bayesian Ridge model, SVM with Linear Kernel performed poorly across the data set. The model performed better with more folds, given that there were thus fewer sample points in each fold from which to generate support vectors. The 10-fold linear SVM with C=1000 performed best across various parameter combinations implemented with this model (Figure 4). The model was able to predict the relative position of a target LOS given the initial parameters, which can be visualized as the directionality of the graph in Figure 4. However, the model was unable to predict the extent to which the magnitude of these initial values would determine the magnitude of LOS, likely because there was too much variance between patients for these magnitudes. When a C value of 5000 was implemented, the model seemed to estimate larger magnitudes for length of stay, but this was a clear instance of over-fitting, as the model even predicted negative values for LOS (Figure 6).

SVM with RBF Kernel

SVM with a RBF Kernel is typically able to handle data with fewer features and larger sample sizes better than the linear kernel model. The SVM with a RBF Kernel performed best with 5 folds, C=1000, $\gamma = 500$ (Figure 7). In this case, the model demonstrated both the relative position and magnitude of LOS better than the Linear SVM and Bayesian Ridge Regression, however, was still insufficient. Overall, the $\gamma$ parameter in the SVM model with the RBF Kernel played a critical role in determining the shape of the model, as higher $\gamma$ values allowed more influence for selected support vectors in the model, and thus fit individual sample data patterns more closely. Small $\gamma$ values constrained the model and yielded smooth curves with few recognizable trends. The optimal case (Figure 7) was obtained with intermediate $\gamma$ and C values, which reflects the nature of the RBF kernel regularization process.

CONCLUSIONS

In conclusion, the models in this project were largely inconclusive and unable to predict hospital LOS reliably. Nevertheless, the project served as an informative process for understanding data cleaning and preparation for machine learning methods, K-fold cross validation, Bayesian Ridge, and SVM methods. The project also facilitated a better understanding of how to tune parameters in machine learning models.

Recommendations

A limitation of this model occurred during the data pre-processing phase. Due to both the manual entry process as well as the clinical complexity of the laboratory and chart data, there were missing values in the data. Additionally, the models required that data points for each feature were from the same patient. Given that not all patients received the same tests within the first 24 hours of their ICU stay, some patients had "missing" features. Imputation of these missing values was not performed, as the sample size of patients who had all four tests done in the first 24 hours of their ICU stay was sufficient for the models developed in this project. Nevertheless, sample size reduction did occur as the four features were selected. Future research with this database should attempt missing value imputation, in order to maximize the patient sample size. It is likely that such imputation would require a more robust clinical understanding of the specific physiological tests. Finally, after the K-fold validation
procedure, the samples sizes were fairly small. Future research should leverage imputation methods to fill in missing clinical data, in order to preserve a larger subset of sepsis patients.

Additionally, this model can be extended to applications in novelty and anomaly detection. Specifically, this project establishes a framework for developing a novelty detector for dynamic patient data. Such a novelty detector could leverage the initial physiological values for a patient in order to predict statistically significant change in various features as data points are added to the electronic health record by caregivers over the ICU stay. Previous literature has demonstrated that changes in these features can signal the advent of clinical events, such as septic shock, would thus be valuable from a preventative medicine perspective. Thus, further exploration regarding novelty and anomaly detection for dynamic patient data would complement this initial project.

APPENDIX

Figure A. Scatter plots for the other 6 features.
Figure B.
Bayesian Ridge Regression, \( k=5, C=10^{-3} \) (top left)
Bayesian Ridge Regression, \( k=5, C=10^{-6} \) (top right)
Bayesian Ridge Regression, \( k=10, C=10^{-3} \) (bottom left)
Bayesian Ridge Regression, \( k=10, C=10^{-6} \) (bottom right)

Figure C.
SVM With Linear Kernel, \( k=5, C=100 \) (top left)
SVM With Linear Kernel, \( k=10, C=100 \) (top right)

Figure D.
SVM With RBF Kernel, \( k=5, C=1000, \gamma = 1/(\text{n}_\text{samples}) \) (top left)
SVM With RBF Kernel, \( k=10, C=1000, \gamma = 1/(\text{n}_\text{samples}) \) (top right)
REFERENCES


ix API design for machine learning software: experiences from the scikit-learn project, Buitinck et al., 2013.