Leveraging Machine Learning to Model Hospital Patient Readmittance

Eric Peterson, Connor Stout, Joseph Datz, Myles Onosko, Brian Moum, Uriah West, Michael Franusich, Jeffrey Wheeler (Faculty)

Math 1103 - BIG Problems, Dept. of Mathematics, University of Pittsburgh

Problem Statement

With anonymized data about a patient's health information, is it possible to determine which factors predict a patient's readmittance within 30 days using statistical and supervised learning techniques.

- **Readmitted Patient** – Defined as a patient coming back to the hospital within 30 days of first entry.
- **Supervised Learning** – Given an input set X and an output set Y, derive a function

$$H(X) \to Y$$

which outputs to the correct Y value at a high percentage of accuracy.
- **Feature Selection** – Using statistical techniques to determine which parameters of patient admissions data (such as race, religion, reason for admittance, drugs prescribed/given, etc.) are the most important in determining if a patient is at risk for re-admittance within 30 days.

About CraneWare:

CraneWare is a company based in Edinborough that develops insurance software supporting quality patient outcomes for hospitals. Their most well-known product, the Chargemaster Toolkit, has earned the KLAS No. 1 ranking in Chargemaster Management every year since 2006.

- Founded by Keith Neilson and Gordon Craig in 1999
- Cost Analytics, Charge Capture & Pricing, Claims analysis
- Publicly Traded
- Available in the US and UK
- Leader in Chargemaster Toolkits

Craneware®

Our Approach

The original data started off in a series of text files, and so we performed feature selection and wrote a python script to convert our data into a format more accessible for data analysis. After which, we used ℓ_1 **Lasso Regression** to isolate the most valuable features of data for analyzing. The data was then used to train a **Support Vector Machine** to draw a high-dimensional hyperplane in our data. This allowed the separation of the data into two categories - those who were readmitted to a hospital within 30 days and those who did not. The ℓ_1 regularization is more effective than the ℓ_2 in this case because the large number of features need to be reduced to only the best predictors.

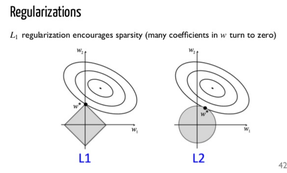

Figure 1: ℓ_1 Regularization vs. ℓ_2 regularization

How It Works

For **Lasso Regression** we are trying to optimize a function $Y(x_1, \cdots, x_n)$ by changing their respective coefficients, $\beta_0, \beta_1, \cdots, \beta_n$, where the equation is

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

Where the β_n is determined by the regularization of the ℓ_1-norm. In order to determine each β_n the algorithm solves the following minimization problem:

$$\min_{\beta_0,\beta} \left(\frac{1}{2N} \sum_{i=1}^{n} (y_i - \beta_0 - x_i^T \beta) + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

Where N is number of observations, and λ is a nonnegative regularization parameter.

Once β_n is minimized, it can be used to solve the following optimization equation for the Support Vector Machine:

$$\max W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j \langle x^i, x^j \rangle$$

$$s.t \quad 0 \le \alpha_i \le C \quad i = 1, ..., m, \quad \sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

Which is the dual form of our optimization problem. In this format, our problem is now solvable using the SMO algorithm.

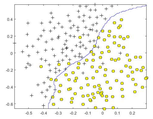

Figure 2: Visualization of a 2D SVM

Results

When setting a λ value to 1, in order to extract the 5 most important predictors, we get the following, along with their corresponding β value:

Predictor	**β-value**
Hypertension	.0211
Atrial Fibrillation	.0100
Congestive Heart Failure	.0473
Respiratory Failure	.0124
Acute Renal Failure	.0148

With $\lambda = .1710$, we were able to identify 98 different categories that were most significant in our dataset. This allowed us to clean our dataset of the weakest predictors of readmittance.

A Support Vector Machine model of the remaining data allowed us to predict readmittance with 96% accuracy on the remaining information.

Conclusion

The **Lasso Regression** model determined that the largest factors influencing patient readmittance are Congestive Heart Failure, Acute Renal Failure, Respiratory Failure, Atrial Fibrillation and Hypertension. With filtered training data, excluding extraneous information and focusing on the 98 most relevant features, the Support Vector Machine was able to predict patient re-admittance with 96% accuracy.

Technologies Used/Skills Learned

• Python/Data Cleaning • Vectorization of Categorical Data • MatLab/Support Vector Machine • Lasso Regression • SQL • Excel • R

Alternative Methods

Several other Machine Learning methods were explored before deciding on the combination of Lasso Regression and SVM. Initially principle component analysis was considered to reduce the dimensionality of the data, but since the data had such a large number of sparse features, utilizing Lasso regression was more effective since it actually removes inconsequential features. For building the model using a neural net appeared as an obvious choice, but it would be ineffective compared to a SVM due to the need for a larger data set to achieve the accuracy necessary to effectively predict patient readmittance.

Acknowledgements

Special thanks to Dr. Jeffrey Wheeler, Dr. Andrew Walsh of Health Monitoring Systems, Eric Bentley of Conduent, and David Hunter, Jourdain Lamperski and Chris McCord of MIT for their assistance and guidance. Additionally we extend our appreciation to the University of Pittsburgh Department of Mathematics for offering this course and providing the resources that aided in this project's success.

University of Pittsburgh