BIG Problems - Craneware*

Eric Peterson, Brian Moum, Uriah West, Connor Stout,
Joseph Datz, Myles Onosko, Michael Franusich

Jeffrey Paul Wheeler (Faculty)

Department of Mathematics, University of Pittsburgh
Pittsburgh, PA 15260, USA
jwheeler@pitt.edu

Spring Semester 2018

Abstract

Utilizing health-care data to track the prevalence of syndromes and diseases that could affect public health is paramount in medical preparation. Therefore, it is necessary to maintain records of accurate and complete data to properly diagnose outbreaks throughout hospital networks. While these data feeds are intended to be continuous and accurate, many times human error coupled with technological issues result in inaccurate and delayed reporting. Our deliverable is a script in the programming language Python that distinguishes good and bad data and retroactively fills in the gaps of bad data by using analytical techniques and forecasting. This paper explores the best programming methods to analyze healthcare data, whether it be ICA, PCA, Denoising Autoencoders, or Bayesian Optimization, for our given dataset.

*Funding for this project is provided by University of Pittsburgh

Contents

1 Introduction to Craneware

Craneware is a multi-national company the produces business IT software for health care companies. Headquartered in Scotland, UK, with its American Headquarters located in Atlanta, GA, Craneware produces a vast amount of new technologies that implement mathematics and computer science. Founded by Keith Nielson and Gordon Craig in 1999, it now is publicly traded with 250+ employees.

1.1 Products

Craneware is the market leader in software and supporting services that help healthcare providers improve margins so they can invest in quality patient outcomes. The company's flagship solution, Charge Master Toolkit®, has earned the KLAS No.1 ranking in Revenue Cycle – Charge Master Management since 2006 and is part of our value cycle management suite, which includes Patient Engagement, Charge Capture and Pricing, Coding Integrity, Revenue Recovery and Retention, and Cost Analytics solutions.

Patient Charge Estimator

Helps estimate the cost to a patient for a procedure as well as allows for more consistent cost estimates across departments. Additionally, it takes into account many factors to gain more accurate picture of cost.

Pharmacy Charge Link®

Tracks medications and drugs being used in the hospital. Also, detects when a medication has not been accurately recorded. It allows for accurate reimbursements from healthcare companies or Medicare. It Collects data for future analysis of spending and prescription needs

InSight Audit®

Reduces paper uses and moves most forms online by centralizing audit information on one platform for company wide consistency. It also automates data entry

Charge Master Toolkit

Serves as the initial listing of items billable to a hospital patient or insurance provider. Prices tend to be inflated and process is error prone. It has solutions that help optimize reimbursement, compliance, and pricing, which increases operational efficiency by automating chargemaster management. It was ranked #1 by KLAS in Revenue Cycle

2 Project 1

2.1 Assignment Given

Potential applications of Machine Learning include determining the risk of a patient being readmitted to a hospital within 30 days of discharge, determining the likelihood of developing potentially life threatening illnesses, and many more. Relevant healthcare data that can be leveraged in Machine Learning applications can include (but is certainly not limited to) administrative data (claims, registration data, etc), electronic medical record data, genomics data, imaging data, clinical notes, and much more.

Given the high dimensionality of such data, a few ways to improve the workflow of model development are:

- **Feature Selection**: Determining which data elements should be included in the model - examples include **Principal Component Analysis (PCA)**, **Independent Component Analysis (ICA)**, and **Deniosing Autoencoders**.
- **Model Selection**: Choosing the most appropriate model for a particular application and selecting appropriate parameters for each candidate model - examples include **grid search**, **random search**, and **Bayesian Optimization Techniques**.

Thus, this project seeks to explore various automated feature selection methods. Project deliverables are as follows:

1. Research and explore various feature selection methods, and apply to relevant healthcare problems.
2. Evaluate models and make best practices recommendations for particular use cases taking into consideration both model accuracy and computational complexity/training time.
3. Explore various automated hyperparameter tuning schemes and apply to cases important to the health care industry.
4. Evaluate models and make best practices recommendations for particular use cases, taking into consideration both model accuracy and computational complexity/training time.

With our given dataset of MIMIC-II, we have chosen to explore building a model which would predict the chance of readmittance within 30 days to a hospital with the guidelines suggested for Project #1 and Project #2.

2.2 Comments on ICA and Denoising Autoencoders in our Problem Statement

The goal of ICA is to solve the problem:

$$x = As$$

Where A is an unknown matrix of data we are attempting to solve, given only a source $s \in \mathbb{R}^n$ and $x \in \mathbb{R}^n$ as an output. The typical example is the "cocktail party problem", where we are given an output frequency vector x, a set of microphones s, and the goal is to find the matrix A which correctly separates each frequency in x with it's associated microphone source.

In order to solve this type of problem, three key assumptions are made that we cannot assume in our dataset:

1. An output dataset x is a multivariate signal
2. A signal vector s is composed of sources that are statistically independent from one another
3. Our signals are non-gaussian in nature

However, our output vector and input matrix are sets of categorical data. We also cannot state for certain that our matrix is statistically independent. Whether or not a person is given painkillers is entirely dependent on what conditions they come into the hospital with. Given these limitations in our dataset, we did not consider ICA an appropriate unsupervised learning technique to apply to our dataset.

2.3 Preliminaries

The data was made available via *Mimic-II* and *Mimic-III* which are public datasets developed by the MIT Lab for Computational Physiology. It comprises health data associated with ∼40,000 anonymous critical care patients, including demographics, vital signs, laboratory tests, medications and more.

2.3.1 Data Obtained

Data for the heatmap can be found at *data.medicare.gov* in the *Hospital Readmissions Reduction Program* subsection.

Craneware recommended the use of MIMIC-III, but there were problems with that dataset. Dates in the MIMIC-III datase had been redacted. Because the predictive models rely on the ability to identify how long after hospital discharge a patient was readmitted, the MIMIC-III datasets were practically unusable. However, MIMIC-II made available the dates redacted in MIMIC-III, and so MIMIC-II is the dataset the models are based on.

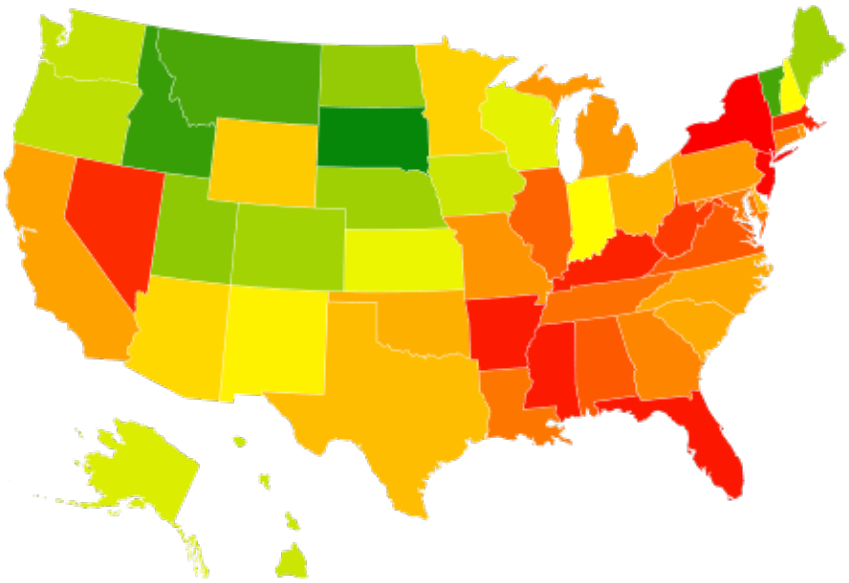


Figure 1: The graphic above shows a heatmap of Average Excess Readmission Ratio per state, with Red being very high Excess Readmission Ratios and Green being low Excess Readmission Ratios.

The MIMIC-II data is composed of multiple text files with lists of information about each patient. Before the data could be used it had to be reorganized into a usable form. Categorical pieces of information were organized into binary vectors with a "1" indicating which group it belongs to. Information like treatment, pre-existing conditions, race, gender, whether a patient was readmitted in 30 days, and other factors were treated in this way. Factors like age, height and weight were not treated as binary vectors but rather as vectors which contain numerical information. The data scraping and organization was done in Python and exported into Excel for ease of use.

2.4 Importance

In October 2012, CMS began reducing Medicare payments for Inpatient Prospective Payment System hospitals with excess readmissions. Excess readmissions are measured by a ratio, by dividing a hospital's number of "predicted" 30-day readmissions for heart attack, heart failure, pneumonia, chronic obstructive pulmonary disease, hip/knee replacement, and coronary artery bypass graft surgery by the number that would be "expected," based on an average hospital with similar patients. A ratio greater than 1.0000 indicates excess readmissions.

The Affordable Care Act (ACA), which established the Hospital Readmission Reduction Program, sought to reduce readmittance because hospital readmissions are associated with unfavorable patient outcomes and high financial costs.

The Medicare Payment Advisory Commission (MedPAC) has estimated that 12% of readmissions are potentially avoidable. Preventing even 10% of these readmissions could save Medicare $1 billion.

2.5 Our Initial Approach

Take patient ID numbers that show up more than 1 time within a 30-day period and show a correlation to care received and indicators in patient health history.

2.5.1 Lasso Regression

Our dataset was a matrix that had over 30,000 patients and over 10,000 different possible elements for each patient. Many predictors, such as Gun Shot Wound admittance, only appeared once throughout the whole dataset and were therefore of low importance on the overall prediction of a patient's re-admittance. In order to design a model that accurately predicted patient readmission, parameters such as these has to be eliminated.

One such way to select certain features is through **Regression Analysis**, which is a set of statistical processes for estimating the relationship among variables. Since the data was organized in a binary form for each feature, i.e. a 1 indicating a 'Yes' and 0 indicating a 'No', the most effect form of regression is linear regression. The two types for sparse data, as is the case for the data given, are **Ridge** and **Lasso Regression**. Ridge regression focuses on regularizing the L_2-norm, whereas **L**east **A**bsolute **S**hrinkage and **S**election **O**perator Regression (or Lasso-Regression for short) regularizes the L_1-norm. Figure 2 demonstrates the improved total accuracy of selecting the appropriate features by regularizing the L_1-norm as a opposed to the L_2-norm.

The general role for the Lasso model is 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

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \tag{1}$$

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:

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

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

When running this technique in MATLAB with a λ value of 1 to obtain the 5 most important predictors of re-admission the following β values were obtained in the following table.

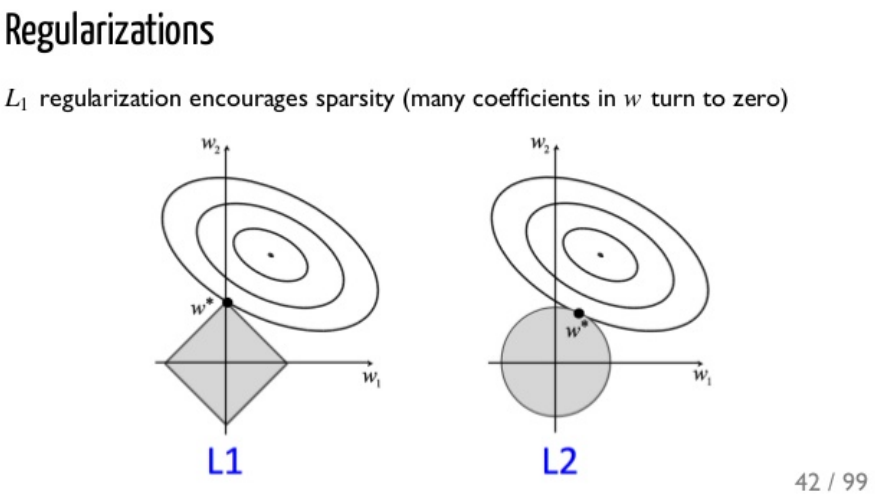


Figure 2: The graphic above shows the difference between regularizing the L_1-norm and that of the L_2-norm

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

2.5.2 Elastic Net Regression

The elastic net model overcomes the limitations of the lasso regression model, which overcomes a penalty term. For instance, a model that contains more dimensions than observation. The Lasso model will only the amount of dimensions equivalent to the amount of observations. Furthermore, the Lasso regression model tends to select one variable from a group of highly correlated variables it tends to select one variable. This is extremely pertinent to our problem given patients could be admitted for slight variations of the same illness.

The Elastic model is identical to the Lasso, however a quadratic β term is added to the model. An α of .5 was set, and R was ran to find the ideal λ, this λ was then used to find the the beta coefficients. The top 5 resulting coefficients are below.

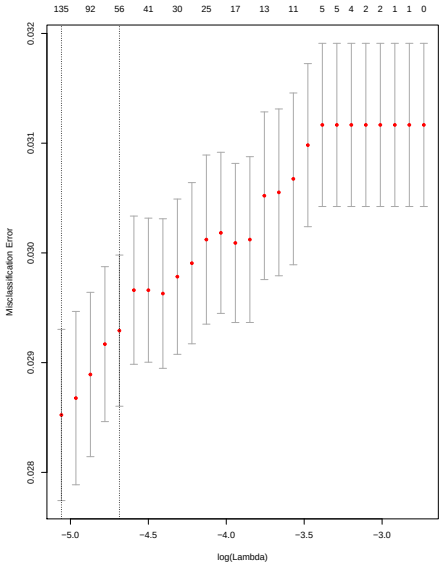


Figure 3: The graphic above displays the lambda which minimizes the MSE

Predictor	β**-value**
Fetal neonatal Jaundice	3.830281
Prematurity	3.5187507
New born Feeding problems	3.503618
Tracheostomy	1.58025026
Post Operation Infection	1.3874598

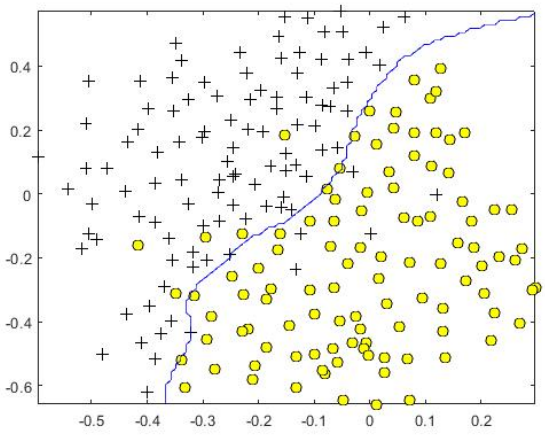
2.5.3 Support Vector Machine Approach

The approach attempted to make a predictive model using a Support Vector Machine (SVM), which establishes a curve between two or more different groups we wish to separate. The figure below, shows an SVM separate categorical data points (illustrated as plus signs or yellow dots for distinguishing the separate classes). The figure can be interpreted to mean that the model will predict that the points above the line are plus signs and the points below the line are yellow dots. In a dataset of readmitted and non-readmitted patients, we can use this information as separate classes and construct a model that separates the two.

Our first model attempted to create an SVM without the aid of lasso or ridge regression. This resulted in a model which did not predict readmitted patients accuractely as our dataset is skewed with 97% nonreadmitted patients. Thus, our total accuracy was 96.16%, but virtually 0% of any readmitted patients in our cross-validation set were correctly predicted. With the use of lasso regres-

sion and ridge regresson coefficients on a reduced dataset, we acheived 97.20% (17.59% readmitted) and 97.36% (38.31% readmitted) respectively.

This resulted in a working model, but not an effectively useful model for prediction. With such a skewed dataset, using an SVM would always have an overfitting error. Our data set has a few readmitted patients with specific conditions in the Cross Validation set that could not be factored into training. Thus we must use another model which is better suited for dealing with almost anomalous cases.



This figure shows a Support Vector Machine where the blue line separates two different types of data points into two different classifiers.

2.5.4 Bernoulli Naïve Bayes Classifier

The Naïve Bayes Classifier attempts to achieve the same results as the SVM, applying a probabilistic method of classification utilizing Bayes Theorem:

$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)} \tag{3}$$

where C_k is the vector of readmitted patients and $x = \{x_1, x_2, ..., x_n\}$ is the set of binary parameters. The probability we want to calculate, $P(C_k|x)$ is the probability that a patient will be readmitted given the parameters.
Unlike the SVM, which considers statistical dependence between parameters, 'Naïve' Bayes classifiers assume strong statistical independence between parameters. While this can affect accuracy, as some parameters likely are dependent, it can allow for calculations across a much broader spectrum of parameters. Therefore this classifier was ran on the entire dataset, rather than the regularized set used by the SVM.
The classifier which was implemented, called a Bernoulli NB, is specifically used

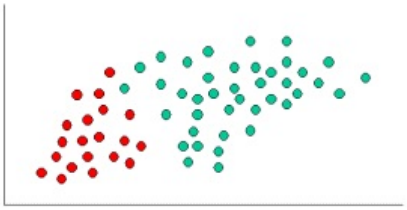


Figure 4: Classified objects

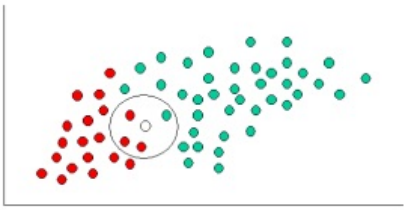


Figure 5: Unidefined object

for binary datasets, where the data points correspond to instances of occurrence rather than frequency. Implementing the classifier on the dataset returned the following results:

- Average Total Accuracy: 92.4%
- Average Readmittance Prediction Accuracy: 69.4%
- Average True Positive to False Positive Ratio: 1:3.5

2.6 Programs Used and Skills Learned

MATLAB, Microsoft Excel (For data analysis), Python, for data compilation, GitHub, R

3 References

Theobald, Oliver. Machine Learning for Absolute Beginners. Second Edition ed., Oliver Theobald, 2017.
Hagan, Martin T., et al. Neural Network Design. Second ed., s. n., 2016.