Comparing Machine Learning Methods on Heart Disease Data

Daniel Bortolussi
Dept. of Computer Science
UNC Chapel Hill
Chapel Hill, NC, USA
730123648
bortolussi@unc.edu

Kyle Moran
Dept. of Computer Science
UNC Chapel Hill
Chapel Hill, NC, USA
730162841
kylebm@live.unc.edu

Hemanth Swarna
Dept. of Computer Science
UNC Chapel Hill
Chapel Hill, NC, USA
730094141
hemanths@live.unc.edu

***Abstract*—The leading killer in the United States for many years has been heart disease [1]. With our machine learning project we set out to predict the chance of heart disease in adults. The causes of heart disease have been studied thoroughly for many years, and several different parameters can be tied to resulting complications. With this data set we were given several symptoms of possible heart disease and using that to make a prediction if the patient had underlying heart problems. Making medical predictions can mean life or death for patients so we tuned all of our models to prefer accuracy over time and false positives over false negatives. In order to make these predictions we created three different models, a logistic regression, a PCA, and a Neural Network.**

I. Introduction

In recent history all over the world the processing of foods and advent and subsequent ubiquity of fast food chains has lead to a rise in unhealthy and detrimental eating habits. This has been accompanied with a rise in heart disease seen particularly in America where heart disease is the leading cause of death [1]. The many different symptoms that result from heart disease and the differences in how it manifests in different genders can lead to many misdiagnoses of heart disease [2]. This is one of the reasons we took on this project.

A. Goals

With this project our goal is to try multiple machine learning techniques to create the best predictor of heart disease. Ideally, since this is a prediction for the medical field, and lives are at risk we will skew the result toward false positives, preferring to catch all and more cases rather than miss a positive case.

B. Methods

1) Logistic Regression: This is the simplest kind of classification that can be done, which means it will likely run quickly and can give some preliminary results that can be used to inform more complex models. It will not perform optimally on this data set since many parameters have a small, discrete input range. However, it can still tell us whether or not the parameters have a significant impact on the target classes.

2) Principal Component Analysis:

3) Neural Network: For the design of the Neural Network, the goal was to keep the model fairly simple no more than three layers since the data only has 13 dimensions, a model of that size should be able to train fairly effectively. Also, for comparison's sake we want the model to be able to output accuracy as part of its training rather than showing the amount of loss. The model is however still trained by minimizing loss (more on choosing the loss function later).

II. Data Set

The data given has 1025 patients all with complete data in each of the 13 parameters and the 1 result. The result of the data, labeled as 'target' is a binary value, 0 if the patient does not have heart disease, and a 1 if the patient is positive for heart disease. The 13 parameters are described in more detail below:

1) Age: The age measured as an integer in number of years.

2) Sex: A Boolean value, 1 if the patient is male and 0 if the patient is female.

3) Chest Pain Type: Measured on an integer scale from 0 to 3. 0 if the patient is feeling no chest pain and 3 if the chest pain is unbearable.

4) Resting Blood Pressure: Blood pressure upon admittance to the hospital, measured as an integer in millimeter of Mercury.

5) Serum Cholesterol: Serum cholesterol is a way of measuring total cholesterol. It is calculated by taking the sum of the HDL cholesterol and LDL cholesterol levels, then adding 20% of the triglyceride levels.

6) Fasting Blood Sugar: Fasting blood sugar is a blood sugar measurement taken without eating for some time. A patient with more than 120 mg of glucose per deciliter of blood suggests that the patient has diabetes. In out data this is a Boolean value, 1 if the patient test above 120 mg/dL and 0 if he or she does not.

7) Electrocardiograph Results: In three integer values, this is given as 0 if the results are normal, is if there are ST-T irregularities, and 2 is if there is left ventricle hypertrophy

8) Maximum Heart Rate: An integer value of the highest heart rate achieved by the patient.

9) Exercised induced Angina: A boolean value 1 if the patient has severe, spreading exercise induced pain in the chest area, 0 if they do not.

10) ST Depression: A decimal value that is the measure of the depression of ST waves induced by exercise divided by rest time.

11) ST Slope: An integer value of the slope of the peak exercise ST segment.

12) Number of Major Vessels: An integer value of the number of vessels colored by flourosopy from 0 - 3.

13) Thalassemia: A scale of the amount hemoglobin in the blood 1 is normal levels of hemoglobin, 2 is a history of high hemoglobin, and 3 is a currently high but fixable level of hemoglobin.

III. Logistic Regression

A. Setup

The simplest way to analyze categorical data, such as whether or not a set of attributes lead to heart disease or not, is to use a classifier that can separate data functionally. One such function capable of this is the sigmoid function:

$$\sigma(z) = \frac{1}{1 + \exp(-z)} \tag{1}$$

This function will always return a value between 0 and 1, so a combination of the input parameters can be passed through the sigmoid function to yield a value that can be compared to a threshold number, like 0.5, to either predict a 1, heart disease, or a 0, no heart disease. For N sets of p attributes $\mathbf{x}_i$ with parameter vector $\boldsymbol{\beta}$, which includes a bias term β_0, yielding a class $y \in 0, 1$, the likelihood function is:

$$\mathcal{L}(\boldsymbol{\beta}|y, X) = \prod_{i}^{N} \frac{1}{1 + \exp[y_i(\mathbf{x}_i \cdot \boldsymbol{\beta}) + (y_i - 1)(\mathbf{x}_i \cdot \boldsymbol{\beta})]} \tag{2}$$

Since logistic regression is prone to overfitting if the data is entirely separable by a hyperplane in the space of the attributes, a penalty term is appended to the likelihood to prevent the parameters from growing indefinitely. The penalty term used in this analysis is the L_2 ridge penalty.

From there, learning the model proceeds by testing a set of parameters, comparing the predictions to a test set, and modifying the parameters according to a numerical gradient descent algorithm.

B. Data Manipulation

Parameters, 2, 7, and 9 are categorical parameters, which means they are not suitable for logistic regression. The input variables $\mathbf{x}$ have to be able to be interpreted as an increasing or decreasing quantity. These columns were ignored during the regression.

C. Analysis

The coefficients of the parameters for the likelihood were

Parameter	Coefficient
Age	0.0132
Chest Pain Type	1.41795
Resting Blood Pressure	-0.0294
Serum Cholesterol	0.0024
Fasting Blood Sugar	-0.5140
Maximum Heart Rate	0.0215
ST Depression	-0.9436
ST Slope	0.9954
Number of Major Vessels	-1.0758
Thalassemia	-1.0138

The bias term was determined to be 0.3949. The model achieved a mean squared error of 0.2072 on the test data set. In the health industry, the models are chosen to prefer false positives over false negatives. This model had a false positive rate of 0.2135. Adjusting the penalty strength can vary the false positive rate at the expense of some accuracy of the model. A large false positive rate of 0.2412 was found for a large regularization strength of value 10. The graph of the trade-off between false positive rate and accuracy against increasing regularization strength is given in figure 1. The model was able to achieve accuracies of 0.8±0.01 regardless of regularization strength.

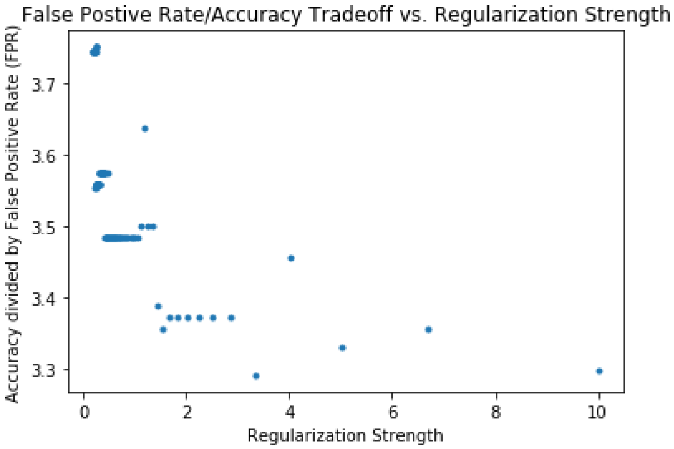

Fig. 1. Comparison between false positive rate and mean squared error

The coefficients that were obtained using this model do have some intuitive correspondence to their respective parameters, e.g. likelihood of heart disease is positively correlated with age and chest pain. However, many of the coefficients are small or do not correspond well with their parameters, which means the data the parameters supplied were not sufficient to affect the model, or they are unrelated to the probability of heart disease.

IV. Principal Component Analysis

A. Setup

For our next machine learning method we chose to use a PCA or a principal component analysis, because we wanted to reduce the effect of dimensionality while retaining most of our data. It also is the logical progression from logistic regression so we can dig deeper on that aspect.

$$\hat{\Sigma} = \frac{1}{N}\sum_{t=1}^{N} \mathbf{x}_t \cdot \mathbf{x}_t \cdot^{T} \tag{3}$$

This function allows us to take the covariance of our dataset as a matrix. This will allow us to compare our errors, so we can figure out how big they are for each particular instance.

We then project this onto an orthonormal basis using the following function:

$$z_t = W^T x_t \tag{4}$$

The projection allows us to simulate a ”test error” of sorts which is generally not possible due to the subspace modelling being unsupervised.

B. Data Manipulation

For the data I split it into a test and train data with the general 20/80 split respectively. The dimensions I reduced were every variable used aside from our response variable, ’target’. I then was able to get my answer using the formulas I listed above.

C. Analysis

The results of the reconstruction error are pretty good. For the first 100 values in the data set, for the first 10 we get somewhat large errors but after that the errors are negligible and all the same number of 7.150195862004584e-27.

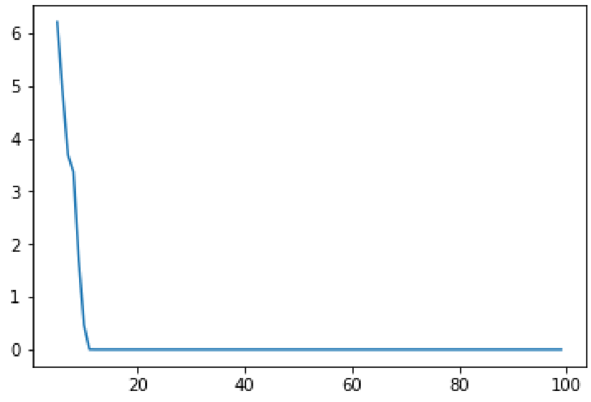

Fig. 2. Reconstruction error for the first hundred values

V. Neural Network

A. Setup

For our last machine learning method we chose to use a deep neural network, because of its relative complexity, but also relative ability to receive great results. The goal in choosing a neural network was to see how it compared to more classical methods and see if the recent push toward neural networks could be applied with success to our data as well.

B. Data Manipulation

In the heart disease data set we have many of the numbers have different ranges, i.e. the serum cholesterol values range from 126 to 564 where the maximum of the gender values is only 1. This means that inherently the values may have different weights when starting the training. This would eventually be ironed out in training, but for ease of use and speed purposes, standardization will speed up the process of training by a lot.

C. Parameter Choices

The target value of our data is either 0 meaning the patient is not expected to have heart disease or 1 meaning the patient is expected to have heart disease. This allows us to use binary cross-entropy as a loss function instead of mean squared error, but after running tests it is clear that mean squared error is faster as shown in Fig. 3.

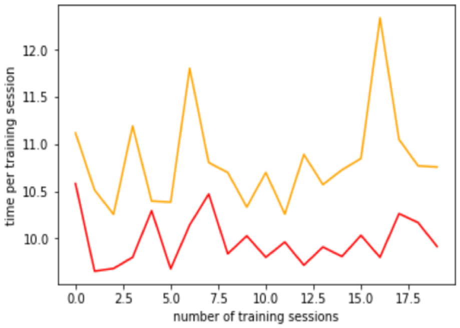

Fig. 3. Time per training session between loss functions, mean squared error(red) and binary cross-entropy (orange)

Mean squared error's improvement in accuracy over binary cross-entropy can also be seen in Fig. 4.

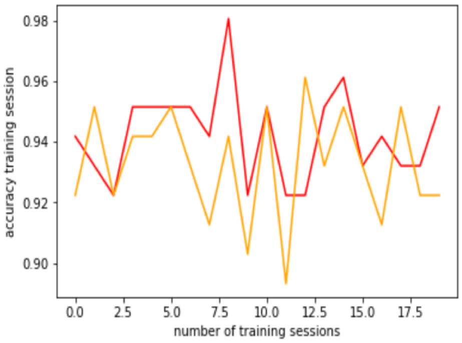

Fig. 4. Accuracy per training session between loss functions, mean squared error (red) and binary cross-entropy (orange)

D. Analysis

For a random sample of patients in the test set and the results of the neural networks prediction see the following table.

age	sex	Chest Pain	Resting BP	Serum Chol	FBS	EKG Results	Max HR	Exercised induced Angina	ST Depression	ST Slope	Number of Major Vessels	Thalassemia	Heart Disease Actual	Heart Disease Predicted
59	M	2	150mmHg	212mg/dL	>120mg/dL	ST-T	157bpm	0	1.6	2	0	history	1	1
57	M	2	152mmHg	212mg/dL	<120mg/dL	normal	150bpm	0	0.8	1	0	high	0	0
67	M	0	110mmHg	201mg/dL	<120mg/dL	ST-T	126bpm	1	1.5	1	0	normal	1	1
63	F	0	150mmHg	207mg/dL	<120mg/dL	normal	154bpm	0	4.0	1	3	high	0	0
54	M	0	140mmHg	239mg/dL	<120mg/dL	ST-T	160bpm	0	1.2	2	0	history	1	1

Over 20 trials the accuracy and loss results can be seen in the following Fig. 5 and Fig. 6

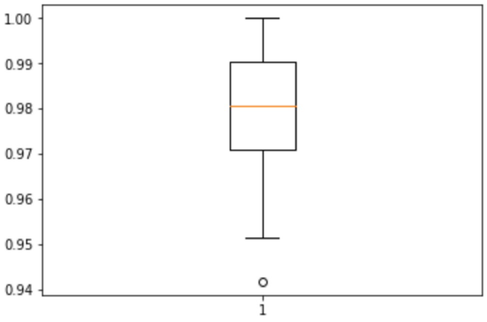

Fig. 5. Box plot of accuracy against a never seen test set after 20 different models trained with the same methods

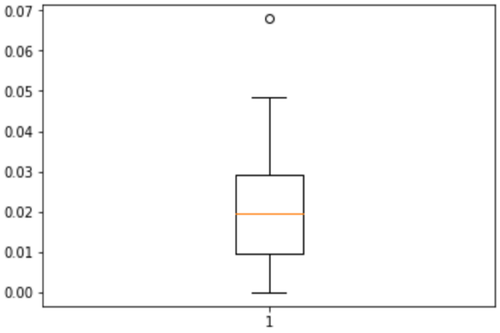

Fig. 6. Box plot of mean squared error of a never seen test set after 20 different models trained with the same methods

After only 10 epochs of training the model is able to reach test accuracy almost always above 95%, and approaching 100% test accuracy a quarter of the time. This model is very accurate, but sometimes mis-classifies certain patients. Take Patient 238, she has the parameters:

Parameter	Value
Age	64
Sex	F
Chest Pain Type	2
Resting Blood Pressure	140mmHg
Serum Cholesterol	313mg/dL
Fasting Blood Sugar	<120mg/dL
EKG Results	ST-T
Maximum Heart Rate	133bpm
Exercised induced Angina	0
ST Depression	0.2
ST Slope	2
Number of Major Vessels	0
Thalassemia	high

This patient is labeled as having heart disease, but the model, on some runs predicted she did not. Overall her parameters lead to the fact that she should have heart disease, and her most extreme parameter, her Serum Cholesterol is very high again indicating she likely has heart disease. The fact that she is female might be a factor in the model predicting she might not have heart disease, as stated in the introduction there are difficulties in predicting and diagnosing if a woman might have heart disease. It tends to be that women have statistically less chance than men to have heart disease, but it can also be misleading to consider that as a factor when creating a model to predict if a patient is positive for heart disease. This convolutional neural network is good however at predicting the chance of a person having heart disease male or female most of the time, often topping out at 100% accuracy on the test set thus a MSE of 0.

VI. Conclusion

After testing the three machine learning techniques, we found that each one was capable of modeling the data well with high accuracy without sacrificing too much time and data manipulation. The approach using logistic regression was able to produce accuracy of around 80%, and could be modified to yield a greater false positive rate if desired. The PCA produced a low test error with a small amount of training, and the neural network was able to predict heart disease with near 100% accuracy with only 2 layers deep neurons. These three models show that even a 13 dimensional data set can be boiled down fairly easily with the powerful machine learning tools at our disposal.

References

[1] Melonie Heron, ”Deaths: Leading causes for 2017.” National Vital Statistics Reports, Hyattsville, MD: National Center for Health Statistics. vol 68 no 6. 2019.

[2] ”Women and Heart Disease.” Centers for Disease Control and Prevention, 14 May 2019, www.cdc.gov/heartdisease/women.htm.