Machine Learning Engineer Nanodegree

Capstone Proposal

Urvi Patel Yi

November 30, 2018

Domain Background

"Every year, over 100,000 people die from a heart attack or related stroke in the UK alone, and heart disease and stroke remain the two biggest overall causes of death worldwide." This was quoted in an August 2018 article published in MedExpress, a medical blog published by the University of Oxford. (https://medicalxpress.com/news/2018-08-technology-fatal-heart.html) The focus of the article was on how new technologies may be used to predict heart attacks in patients. Of course, it is impossible to talk about heart attacks, without also talking about heart disease.

Heart disease is a condition in which plaque builds up on the walls of arteries. This can potentially cause blood clots that restrict the flow of blood, which in turn can cause heart attacks. (Source: http://www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease). Preventing heart disease is, therefore, a key factor in preventing heart attacks.

I am personally invested in investigating instances of heart disease because my father passed away from his first heart attack when I was seventeen years old. He had not been diagnosed with heart disease at the time, which meant that we were completely unaware of the risk when the heart attack struck. I believe a prediction system for heart disease will enable patients to better understand their risk for heart attacks.

Problem Statement

I will attempt to predict an instance of heart disease given a set of health measurements from a patient. The health measurements will be the features that will be the input into the machine learning model. The target will be to predict the presence of heart disease or the absence of heart disease from the patient. This will be treated as a binary classification problem.

Datasets and Inputs

I will use the Heart Disease Datasets from the UCI Machine Learning Repository to train and test my machine learning models. (Source: https://archive.ics.uci.edu/ml/datasets/heart+Disease)

The original datasets contain patient data from three sources:

1. Cleveland Clinic Foundation
2. Hungarian Institute of Cardiology, Budapest
3. University Hospital, Zurich, Switzerland

Cleveland UCI Heart Disease Dataset contains 303 samples with 14 features each.
Hungarian UCI Heart Disease Dataset contains 294 samples with 14 features each.
Swiss UCI Heart Disease Dataset contains 123 samples with 14 features each.

The Cleveland dataset has been most widely used by Machine Learning researchers.

Originally, 76 different attributes were collected from each patient, but the datasets only include 14 of these attributes.

1. age: age in years
2. sex: sex (1 = male; 0 = female)
3. 9 cp: chest pain type
    a. 1 = typical angina
    b. 2 = atypical angina
    c. 3 = non-anginal pain
    d. 4 = asymptomatic
4. trestbps: resting blood pressure (in mm Hg on admission to the hospital)
5. chol: serum cholesterol in mg/dl
6. fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7. restecg: resting electrocardiographic results
    a. 0 = normal
    b. 1 = having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
    c. 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria
8. thalach: maximum heart rate achieved
9. exang: exercise induced angina (1 = yes; 0 = no)
10. oldpeak = ST depression induced by exercise relative to rest

11. slope: the slope of the peak exercise ST segment
    a. 1 = upsloping
    b. 2 = flat
    c. 3 = downsloping
12. ca: number of major vessels (0-3) colored by flourosopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
14. num (goal): diagnosis of heart disease (angiographic disease status)
    a. 0 = < 50% diameter narrowing
    b. 1 = 50% diameter narrowing
       (in any major vessel: attributes 59 through 68 are vessels)

The first 13 attributes will be the features into my machine learning model. The 14th attribute (num/goal) will be my target variable.

One interesting note: The "**num**" or "**goal**" field is the diagnosis of heart disease in the patient. In the Hungarian and Swiss datasets, this attribute is integer valued from 0 (no presence) to 4. In the Cleveland dataset, this is simply 0 (no presence of heart disease) or 1 (presence of heart disease). The Cleveland dataset contains almost half of all data available. Because of this, for the Hungarian and Swiss datasets, I will convert all nonzero values of "**goal**" to a value of 1 for my analysis.

In the raw combined dataset, the distribution of the goal values is as follows:

Goal	**Count**
0	360
1	209
2	68
3	65
4	18

As I suspected, this dataset is imbalanced because values 2 - 4 were not used to collect the Cleveland data.

I will be converting values 1-4 to "1" to indicate the presence of heart disease. A value of "0" will indicate the absence of heart disease.

Heart Disease	Count
0	360
1	360

After the conversion, the dataset appears very balanced.

To compare this to the real world: I was unable to find an estimated percentage of the population that has heart disease. I did, however, find a statistic from the American Heart Association that states that an estimated 17.3% of the global population die from heart disease every year. (Source: https://www.heart.org/idc/groups/ahamah-public/@wcm/@sop/@smd/documents/downloadable/ucm_470704.pdf) This number may not account for people who have heart disease but have not died from it.

Solution Statement

I propose using supervised learning algorithms to classify the data. The classification will be binary: 0 indicates the absence of heart disease, and 1 indicates the presence of heart disease. I will split the data into a training set and a testing set so that I can use the testing set to evaluate the model and understand if my model is overfitting to the dataset. I will explore using a deep neural network to perform the classification and prediction.

I will use Python 3.6, numpy, pandas, scikit-learn, and (but not limited to) keras for this project.

Benchmark Model

Previously, the UCI Heart Disease datasets have been explored using an SVM to an accuracy of 64%. This model is presented in the Packt Publishing's Real World Machine Learning Projects with Scikit Learn Workshop (Source: https://github.com/PacktPublishing/Real-World-Machine-Learning-Projects-with-Scikit-Learn/blob/master/Section%203%20-%20Code.zip)

For the Packt SVM model, the goal values were not converted to values 0 or 1. I will use the ideas from the Packt SVM model to build my own SVM model that trains on the converted dataset. I will use this as my benchmark model.

Then, I will build a neural network and see how my results compare to the SVM approach.

I will also try to fit the data to a simple logistic regression model to get a baseline for comparing both the SVM and the neural network solutions.

Evaluation Metrics

Although this is a balanced dataset, I will use an F-Beta score to evaluate this model, with Beta = 2.

I chose Beta = 2 because I want this to be a high recall model, meaning that false positives are preferred over false negatives. (It is acceptable if a patient is falsely diagnosed with heart disease because they can be referred to other tests to gain confidence in the result. It is unacceptable if a patient with heart disease is misdiagnosed with a false negative--because then the disease will go unchecked and may result in death.) This is a Type 2 error, where unhealthy patients are misdiagnosed as healthy, and it should be avoided.

Because the F-Beta score takes into account both precision and recall, I will use this as an evaluation metric instead of the accuracy score.

Project Design

First, I will preprocess the data. I will remove any features that have a lot of missing data points. I will also convert the goal values so that any nonzero values will be set to a value of 1. Value 0 will indicate the presence of heart disease, and Value 1 will indicate the absence of heart disease. The reason behind this is that the Cleveland dataset uses only values 0 (absence) or 1 (presence), and the Cleveland dataset composes half of all available data. As a final step in preprocessing, I will split the dataset into training and testing sets using train_test_split().

I will get a naive baseline using a logistic regression approach.

Then, I will build the SVM benchmark model. This model will be based on the Packt Publishing SVM model, but I will train and test it on my preprocessed UCI Heart Disease dataset.

After that, I will perform my own analysis on the dataset using a neural network to see if I can comparable or better results. I will build a deep neural network model using keras. I plan to begin by using relu and tanh activation functions and at least one dropout layers to improve accuracy and reduce overfitting to the training set. Finally, I will compile the model, trying various optimizers.

I will train my model on the training set with various epochs to (starting from epoch=20). I plan to iteratively add layers and fine-tune the model.

Finally, I will evaluate my models using an F-Beta score, with Beta = 2.