Machine Learning Engineer Nanodegree

Capstone Proposal

Tatiana Zolhof Panisset
December 20th, 2018

Proposal

Domain Background

Heart Disease is a global problem that takes lives of millions of people every year. According to the World Health Organization - WHO[1], 17.9 million people die every year from cardiovascular diseases, corresponding to 31% of all deaths worldwide. In the United States, data taken from the Center for Disease Control and Prevention – CDC website[2] show that about 610,000 people die of heart disease in the country every year, corresponding to 1 in every 4 deaths. Also according to the WHO, triggering these diseases are tobacco smoking, unhealthy diet, physical inactivity, and the harmful use of alcohol. These in turn show up in people as raised blood pressure, elevated blood glucose, and overweight/obesity.

This is a subject of constant research, in order to identify actions that might reduce the overwhelming numbers aforementioned. One interesting academic research reference can be found in the UCI – Machine Learning Repository[3], which contains several papers on the theme.

My personal motivation for investigating this problem is the fact that my father passed away from a heart attack when he was only 49 years old. Hence, understanding this universe, and being able to make some contributions to it, would be a meaningful way to apply the knowledge I have gained in the course so far.

Problem Statement

Using a dataset of people with several information about their characteristics and health condition, it is necessary to predict the presence of heart disease in them,

creating a model that can be used in new data, not contained in the dataset. This would allow to take action to prevent the evolution of the condition, when it is already present, but also maximize the awareness of the problem as a whole, encouraging prevention efforts. This is a Supervised Learning Classification problem, since we have a defined set of features and corresponding labeled outputs; for which there are many algorithms and metrics that can be applied – for example, the Decision Tree algorithm and accuracy metric. It is also a replicable problem by using the same data and process adopted in this project.

Datasets and Inputs

The data set to be used is the Heart Disease Data Set from the UCI – Machine Learning Repository[4]. It contains 14 numerical attributes for a group of people, as follows:

- age (numerical);
- sex (categorical);
- chest pain type (categorical);
- resting blood pressure (numerical);
- serum cholesterol in mg/dl (numerical);
- fasting blood sugar (categorical);
- resting electrocardiographic results (categorical);
- maximum heart rate achieved (numerical);
- exercise induced angina (categorical);
- ST depression induced by exercise relative to rest (numerical);
- the slope of the peak exercise ST segment (categorical);
- number of major vessels colored by fluoroscopy (categorical);
- thallium stress test (categorical);
- predicted attribute (categorical : 0 – absent heart disease; 1 – present heart disease)

It is important to mention that the original dataset contains 76 attributes; however, 14 of them were found most helpful and adopted in past studies - including those that will be taken as benchmarks here. For this reason, and also considering that the 14 variables cover the characteristics and health conditions related to heart disease, this is the dataset I have chosen to work with.

Solution Statement

The solution to this problem will be applying several Machine Learning methods to the provided set of features, and create a model able to predict the presence of heart disease in each case, and also appropriately generalize to new data. More specifically, after exploring and preprocessing the data, I will be trying different Supervised Learning techniques (like Logistic Regression, Ensemble Methods, and Support Vector Machines), verify the one that solves this classification problem with the best results - according to the accuracy metric -, and compare the final model with different benchmark references. The solution is also replicable by using the same data and process adopted in this project .

Benchmark Model

The “Past Usage” section of the “heart-disease.names” file contained on the UCI - Machine Learning Repository[5] shows results of studies made with this dataset, which obtained an accuracy between 74.8 and 78.9%. Hence, those will be the benchmarks I will be comparing my results to.

Evaluation Metrics

Since this is a classification problem, and also considering the benchmarks indicated in the section above, I will be using accuracy as the evaluation metric of my model. This is defined by:

accuracy = (# of correctly classified examples/ total # of examples) * 100

Project Design

My first approach will be to work on extensively **exploring the dataset**, to gather insights from the data and make any **preprocessing steps** that are needed. That includes:

- Getting data summaries;
- Checking and adjusting missing values;
- Creating visualizations to check the features’ distributions;
- Applying feature scaling and transformations;

- One-hot encoding categorical variables;
- Checking and adjusting imbalanced classes;
- Splitting the data in training and testing sets.

Then I will work on **building the Machine Learning models**, trying several options to see the one that brings up the best results. That includes:

- Decision Trees;
- Ensemble methods, like Adaboost;
- Support Vector Machines;
- Logistic Regression.

Having chosen the model, then I will work on **tuning it**, using grid search for different parameters. I will finally gather the results according to the metric detailed in the section above, and **compare them to the benchmarks** obtained by previous studies.

References

1- https://www.who.int/cardiovascular_diseases/en/

2- https://www.cdc.gov/heartdisease/facts.htm

3- https://archive.ics.uci.edu/ml/datasets/Heart+Disease

4- Dataset source:

> Creators:
> 1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
> 2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
> 3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
> 4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.
> Donor:
> David W. Aha (aha'@'ics.uci.edu) (714) 856-8779

5- https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/heart-disease.names