Capstone Project 1 Proposal

Brian Lubeck

June 4, 2017

Introduction

Heart disease is the leading cause of death in the United States. More than 600,000 Americans die from it each year. One out of every four deaths every year is from heart disease. Heart disease is the build up of plaque in a person's arteries. As the plaque builds up, the arteries narrow reducing, blood flow to the heart. A doctor can perform several tests to diagnose heart disease, including chest X-rays, coronary angiograms, electrocardiograms (ECG or EKG) and exercise stress tests. Fortunately, heart disease is treatable. Eating a healthy diet, exercising regularly, maintaining a healthy weight and taking medications are four ways to reduce the risk of developing heart disease. The goal of this analysis is to build a classifier that accurately diagnoses if a person has heart disease based on his or her personal characteristics, symptoms and test results. The classifier will inform doctors which tests to perform on a patient and what personal information to collect. Doctors can then input the test results and personal information into the classifier to accurately determine the probability a patient has heart disease.

Data and Methodology

The Cleveland heart disease dataset from the University of California, Irvine's Machine Learning Data repository[1] will be used to build and test the classifier. The dataset consists of medical information collected from patients at the Cleveland Clinic Foundation. The dataset has 14 fields and 303 rows. Each row represents a different patient. The field of primary interest is the field labeled *num*, which we relabel as *disease_ind*. The possible values of *disease_ind* are of 0 and 1, with 0 meaning the patient does not have heart disease and 1 meaning the patient has heart disease.[2] The other 13 variables in the dataset will be used to predict the value of *disease_ind*.

To build the classifier, we will use common machine learning binary classification models and evaluation metrics. The four models we will use are logistic regression, neural networks, support vector machines and random forests. We will evaluate the predictive accuracy of the models based on their F-scores and classification accuracy. The F-score is a commonly used evaluation metric for binary classification problems when the data contains very few zeros or ones.

Deliverables

We will write a paper that explains the problem, the approach taken and our findings in technical detail. The paper will contain an introduction explaining the problem, the data and methodology used, the results - including tables and graphs and a discussion of the results. The entire Python code used to perform the analysis will be included as an appendix. In addition, we will create a slide deck that presents the analysis in a non-technical, easy-to-understand but compelling way.

[1] http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[2] Technically, a value of 0 means less than 50% narrowing of the diameter of an artery, and a value of 1 means greater than 50% narrowing of the diameter of an artery.