Predicting Heart Disease with Supervised Machine Learning

Evan Chang, Nicholas Seidl

Abstract

Every year, about 610,000 people die of heart disease in the United States - that's one out of every four deaths. Determining if a person is at risk of having heart disease is difficult, as it often depends on multiple variables such as age, gender, weight, etc. Even though certain doctors may be very good at predicting heart disease, patients would still have to physically go in for a visit. This restricts the number of people able to get evaluated to those who can afford a doctor's visit. If there were a way to predict whether someone was at risk of getting (or has) heart disease with a machine, then many more people would be able to be evaluated. Therefore, hopefully decreasing the number of deaths.

Introduction

The difficulty in predicting the risk of heart disease lies in the fact that the previous data any single doctor or organization has access to is sparse. This is due to both physical variations in humans as well as doctor-patient confidentiality agreements. This makes it difficult for a doctor to predict whether or not a patient is at risk of having heart disease, given previous patients that they've observed, assuming that their data is accessible. Predicting heart disease early is important because medication and treatment can be administered before it is too late.

Proposed Project

First, we plan to use the Cleveland Heart Disease Dataset, which contains 303 data points, each with 14 attributes. Each of these data points represents a patient, with information such as age, sex, cholesterol amount, etc., as well as if they have heart disease or not. Originally, it was part of a larger study with four hospitals participating, for a total of 920 data points with 76 attributes. But due to difficulties in normalizing which data would be collected at each location, only the data from Cleveland is used in practice: some attributes are simply missing data at other locations. There are binary features (fasting blood sugar concentration given a threshold, exercise-induced angina), ternary features (resting electrocardiographic results), as well as continuous features, such as age and heart rate.

Due to the diversity in types of features (binary, ternary, continuous), we expect some sort of data normalization to be required when classifying samples. Normalization adjusts the range of all the variables to be the same so that certain algorithms will not naively weight some features more than others.

For all classification algorithms, we will also perform k-fold cross-validation. It is a good idea to use k-fold cross-validation with this dataset due to the small total sample size.

The first classification algorithm that we will use is Logistic Regression. Logistic Regression allows us to take into account all 14 features when predicting whether the sample displays heart disease or not. The sigmoid function is used to generate the final prediction. Even though our features are not all binary, Logistic Regression will still work well. Each feature will only ever be multiplied by its own weight, and not another feature's weight. The output of Logistic Regression is a 14-dimensional hyperplane, wherein each dimension, it attempts to perfectly separate positive (1, heart disease) and negative (0, no heart disease) samples.

The second classification algorithm that we will use is the Naive Bayes Classifier. In a Naive Bayes Classifier, the assumption that all the features are independent of each other allows the Bayes' Rule to be used. The conditional probabilities of each feature given the outcome class (in this case, presence of heart disease) are modeled; therefore creating a way to calculate whether a new sample (a combination of features) does or does not display heart disease. However, because some of the features we are working with are non-binary (ternary and continuous), we will need to employ some sort of technique to convert these features into binary features. One way is to simply use the average as the threshold: if the 303 samples have an average blood pressure of 222, then any samples with less than or equal to 222 are considered 0 (low blood pressure), and any samples with greater than 222 are considered 1 (high blood pressure). This way, each feature can be modeled as a binary feature.

The third classification algorithm that we will use is the Support Vector Machine. Similar to Logistic Regression, it allows us to take into account all 14 features. However, rather than simply attempting to use any hyperplane that splits the samples, it seeks to find *the* hyperplane that maximally splits the samples. This hyperplane is the farthest distance away from the closest samples from each class. Since the distances used to influence the hyperplane evaluation are calculated geometrically, we must normalize our data. For example, a difference of 1 in blood pressure without normalization is not significant, but a difference of 1 in exercise-induced angina (which is a 0 or 1 binary feature) without normalization is significant.