Adia Ainsworth

Math 179: Mathematics of Big Data

April 27, 2024

Final Project Report

In this paper, I examined how to use Naive Bayes Classifier, logistic regression, and binary classification to examine cardiovascular health data from Cardiovascular Health Analysis and Heart Health Insights for Comprehensive Cardiovascular Risk Assessment. In using the Naive Bayes Classifier, I tested two different approaches to analyzing the data set including a Gaussian model and converting continuous features to categorical features to see which was preferable for predicting whether the lifestyle and physical attributes of each patient could determine if they had a history of heart attack, as well as using logistic regression to predict whether patients' experience with cardiovascular pain could indicate the presence of heart disease.

1. Methods Used

a. Naive Bayes Classifier

The Naive Bayes Classifier is an algorithm that utilizes Bayes' Theorem, which describes the probability of an event based on prior knowledge of the conditions that could possibly be relevant.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Figure 1: Bayes’ Theorem

The left side of the equation, P(A|B), is known as the posterior, and describes the probability that the class is A given that the data B. The right side of the equation, P(B|A)P(A)/P(B), is known as the likelihood, and essentially describes the probability that given the data B, the class is A.

The way that Naive Bayes algorithm in particular works is by applying Bayes’ Theorem under the assumption that the predictor variables conditional on the response will be independent and that all of the features contribute equally to the outcome.

b. Logistic regression

Logistic regression is a process which takes in a data set of independent variables as an input and uses them to estimate the probability of a certain discrete outcome or event occurring.

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

Figure 2: Logistic regression formula

In the real world, logistic regression finds many applications in fields such as manufacturing, where, for example, it can be utilized to calculate the probability of failure of each component used in machinery. Logistic regression models a binary outcome, which is something that can only take two values, such as true or false. In this particular case, the two binary outcomes possible were the presence or absence of exercise-induced angina in patients.

2. Implementation

For this project, I decided that I wanted to examine medical data sets listing health information about patients and see if it was possible to find patterns in this data that corresponded to whether these patients had a history of heart attack, with the eventual goal of developing a model that can accurately predict whether a patient has a history with or will experience a heart attack based on the input of these attributes. I was able to accomplish this goal by using Naive Bayes Classifier on the Heart Health Insights for Comprehensive Cardiovascular Risk Assessment data set [1]. I found that using a Gaussian model produced the most accurate predictions of whether the patient had experienced a history of heart attack, as opposed to a model which utilized a categorical approach.

In order to expand my analysis to include heart disease prediction, I decided to specifically examine a medical data set listing a variety of health and lifestyle information about patients in order to develop a model which could make an educated guess as to whether each patient was likely to either have experienced cardiovascular health complications in the past or be experiencing them currently based on this data alone. In order to do so, I recreated the implementation of Siddhardhan S.'s Heart Disease Prediction System [2]. The data set that I used was called Cardiovascular Health Analysis, which came from Kaggle [3]. This data set contained 13 columns detailing the physical, medical, and lifestyle attributes of various patients. The columns specifically recorded Patient Identification Number (0), Age (1), Gender (2), Resting blood pressure (in mmHg) (3), Serum cholesterol (in mg/dL) (4), Fasting blood sugar (5), Chest pain type (6), Resting electrocardiogram results (7), Maximum heart rate achieved (8), Exercise induced angina (9), Oldpeak = ST (10), Slope of the peak exercise ST segment (11), Number of major vessels, (12), and Classification (target) (13).

The first step in working with the data was to determine what each of the values represented and what units each was measured in so that I could determine which would be the most useful for analysis. Examining the information about the data set, I found that the fasting blood sugar variable had two possible values, with a 0 indicating a fasting blood sugar equal to or less than 120 mg/dL and a 1 indicating a fasting blood sugar greater than 120 mg/dL; the resting electrocardiogram results variable has three possible values, with a 0 indicating a normal ECG reading, a 1 indicating the presence of ST-T wave abnormality (meaning T wave inversions and/or ST elevation or depression of greater than 0.05 mV), and a 2 indicating probable or definite ventricular hypertrophy by Estes' criteria (defined as greater than 35 mm); and the chest pain type variable had three possible values, with a 0 indicating typical angina, 1 indicating atypical angina, a 2 indicating non-anginal pain, and a 3 indicating asymptomatic. The final column, Classification (target), is a binary variable with two possible values, with a 0 indicating the absence of heart disease and a 1 indicating the presence of heart disease. For this reason, the Classification (target) column was the main focus of my analysis as I attempted to use the previous columns' data to predict whether heart disease was present or absent in each individual patient.

Once the data was loaded in, the next step was to separate it into training and testing sets, and subsequently to set up the logistic regression model. Once this was complete, I calculated the accuracy score for each of the training and testing sets, finding that they had the same accuracy score. Using these components, I was able to use this logistic regression model to take input from one patient in the data set and predict whether they would be diagnosed with heart disease or not.

3. Conclusion

One difference from the example upon which this implementation was modeled was that the Heart Disease Prediction System was only equipped to handle a single entry, whereas I modified my implementation so as to be able to sort through the entire data set and predict the heart disease diagnosis of each individual. After comparing and contrasting the different methods of predicting cardiovascular health complications based on health and lifestyle data, the Naive Bayes Classifier using a Gaussian model had the best accuracy score when correctly predicting which patients had a history of heart attack, but the second model was able to handle a larger data set with less changes needed to be made to the data to make calculations possible.

References

[1] https://www.kaggle.com/datasets/mahad049/heart-health-stats-dataset/data

[2] https://www.youtube.com/watch?v=qmqCYC-MBQo

[3] https://www.kaggle.com/datasets/mahad049/heart-health-stats-dataset/data