CSC 529

Sarah Cummings

Winter 2016

Case Study 2: Diagnosing Heart Disease with Support Vector Machines

Introduction:

According to the CDC, about 610,000 people die of heart disease in the United States every year—making in the leading cause of death for both men and women. Heart disease accounts for 1 in every 4 deaths. Because of the heart disease mortality rate, it's important that doctors fully understand the factors that contribute to heart disease. Thus, hospitals and medical centers often study the relationship between heart disease diagnosis and related heart statistics. The Cleveland Clinic has allowed their heart disease dataset to be posted to the UCI machine learning repository.

In this paper, I aim to analyze the relationships between heart disease diagnosis and the other 13 variables and health statistics listed in the dataset. While chest X-rays and blood tests are commonly used to diagnose heart disease, our final goal is to predict heart disease diagnosis using a machine learning algorithm. I will start with analyzing the Cleveland Clinic data that has 303 observations of 14 variables, noting the differences among the heart diseased and non-heart diseased patients, and then form Support Vector Machine models with various kernels to classify the data. My research questions are thus as follows: 1. Which factors are the most influential on heart disease diagnosis? 2. Can we use SVM to accurately classify heart disease diagnosis? 3. If so, which SVM kernel will be most successful? Understanding the risk factors for heart disease is important for both doctors and the general population alike. We should do all we can to prevent the occurrence of this terrible disease.

Data Description:

In a dataset released by the Cleveland Clinic in 1988, there is information for 303 patients concerning 14 variables. Their goal was to predict whether the patient has heart disease. The 14 variables are as follows:

1) Age: age in years
2) Sex: Binary, 1 if male, 0 for female
3) Cp: Chest pain type, categorical. Value 1 for typical angina, 2 for atypical angina, and 3 for non-anginal pain, and 4 for asymptomatic chest pain.
4) Trerestbps: resting blood pressure (in mm Hg on admission to the hospital)
5) Chol: serum cholestoral in mg/dl.
6) Fbs: fasting blood sugar > 120 mg/dl. Value 1 for true, and 0 for false.
7) Restecg: resting electrocardiographic results. Value 0 for normal, 1 for having ST-T wave abnormality, and 2 for showing probable or definite left ventricular hypertrophy by Estes' criteria.
8) Thalach: maximum heart rate achieved.
9) Exang: indicator variable for exercise induced angina. Value 1 for yes, and 0 for no.
10) Oldpeak: ST depression induced by exercise relative to rest.
11) Slope: the slope of the peak exercise ST segment. Value 1 for upsloping, 2 for flat, and 3 for downsloping.
12) Ca: number of major vessels (0-3) colored by flourosopy.
13) Thal: 3 for normal, 6 for fixed defect, and 7 reversible defect.
14) Num: diagnosis of heart disease (angiographic disease status), with values 0 through 4. The 0 represents < 50% diameter narrowing (no heart disease). Values 1-4 for > 50% diameter narrowing, with each being different levels of severity of heart disease.

Data Cleaning:

In preparing the data for analysis, I first created a binary diagnosis variable from the original num variable, since the researchers were originally more interested in predicting whether the patient had heart disease or not, rather than predicting the different levels of heart disease. After creating the variable, we find that 139 people have heart disease and 164 do not. I then recode the following

categorical variables with R's as.factor function, to ensure their numerical categories are in fact processed as categories: sex, cp, fbs, restecg, exang, thal, and slope. Next, I created an 80-20 training testing split of the data. Fortunately, the data has no missing values and no other preprocessing was necessary.

Data Analysis:

In an initial analysis, we will examine the variance of the different variable. In the tables below, we see the distribution of the categorical variables.

The diagnosis variable, num:

```
> table(df$num)

  0   1   2   3   4
164  55  36  35  13
```

Sex:

```
> table(df$sex)

  0   1
 97 206
```

Chest pain:

```
> table(df$cp)

  1   2   3   4
 23  50  86 144
```

Fasting blood sugar> 120 mg/dl:

```
> table(df$fbs)

  0   1
258  45
```

Resting electrocardiographic results:

```
> table(df$restecg)

  0   1   2
151   4 148
```

Exercise induced angina:

```
> table(df$exang)

  0   1
204  99
```

Slope of peak exercise:

```
> table(df$slope)

  1   2   3
142 140  21
```

Thal:

```
> table(df$thal)

  3   6   7
166  18 117
```

As seen in the tables, we have 164 cases with no heart disease, and 139 across the values 1-4 indicative of heart disease. Most of the heart disease diagnoses are level 1. Note that the data is unbalanced for gender, with twice as many men as women. Also note that most patients have asymptomic chest pain, a fasting blood sugar less than 120 mg/dl, and no exercised induced angina.

A summary of the numeric values can be seen below:

Age:

Resting Blood Pressure:

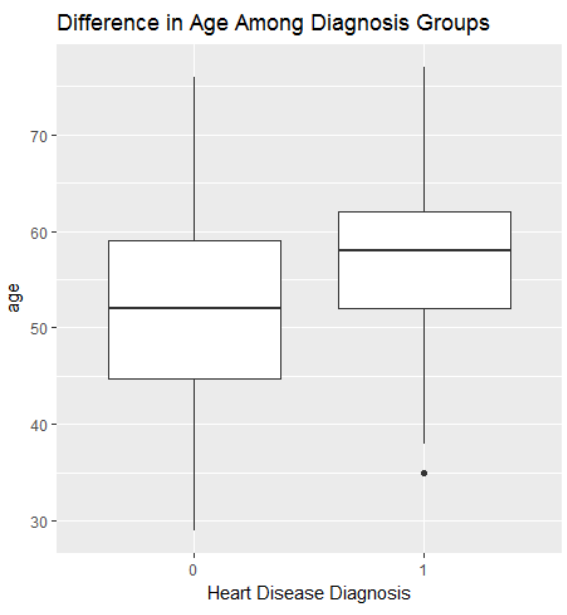
Cholesterol:

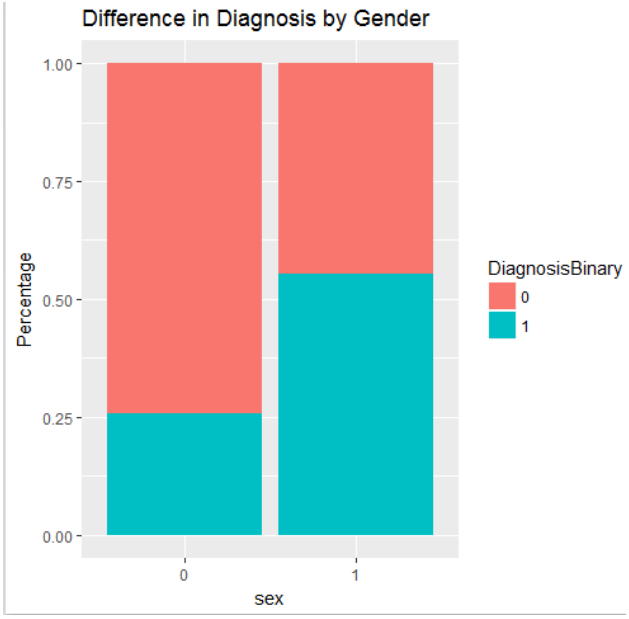
Maximum heart rate achieved:

ST depression induced by exercise relative to rest:

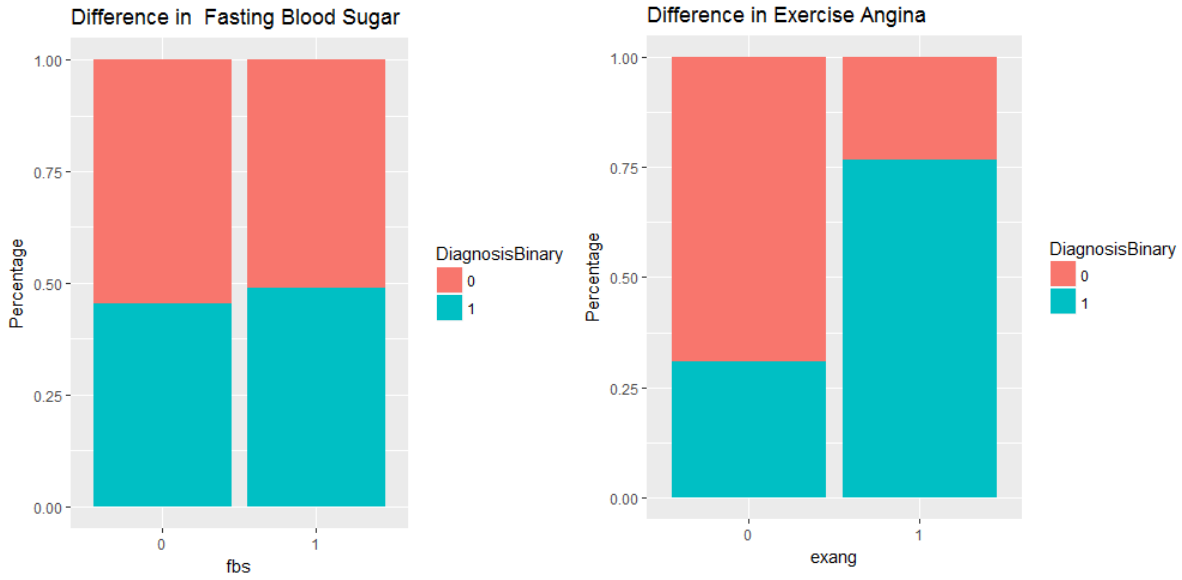
```
> summary(df$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  29.00   48.00   56.00   54.44   61.00   77.00
> summary(df$trestbps)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   94.0   120.0   130.0   131.7   140.0   200.0
> summary(df$chol)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  126.0   211.0   241.0   246.7   275.0   564.0
> summary(df$thalach)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   71.0   133.5   153.0   149.6   166.0   202.0
> summary(df$oldpeak)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    0.00    0.80    1.04    1.60    6.20
```

As seen in the summaries above, the average age of the documented patients is mid-50s. Their average resting blood pressure is around 130, cholesterol around 245, and maximum heart rate around 150. Next, I examined the differences in these variables among the heart disease and non-heart disease patients. In doing so, we can see some relationships between heart disease and the other health statistics. I begin with examining the difference in age and gender among the groups, which is visualized below.

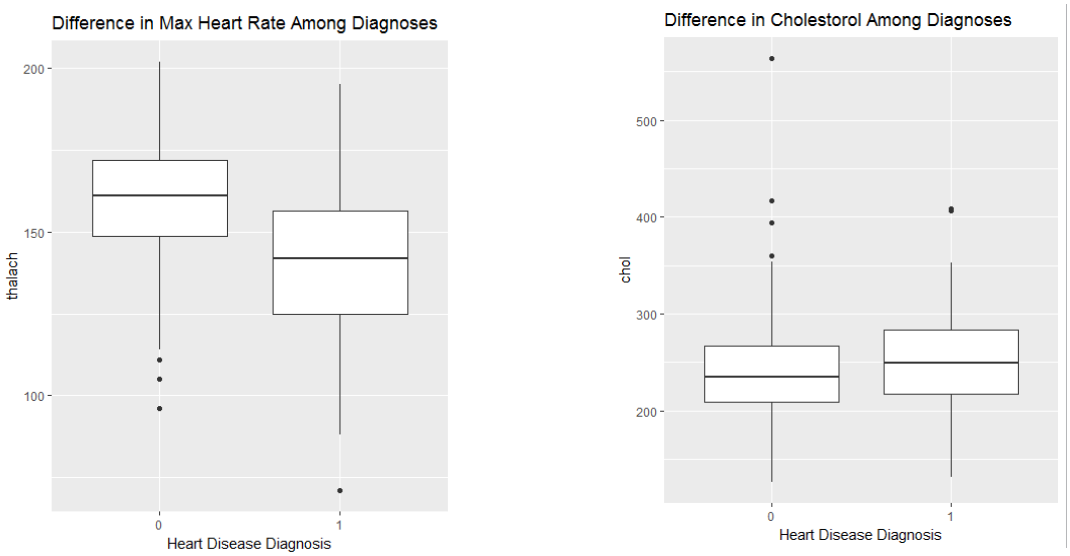




As seen in the boxplot above, the patients with heart disease are older on average than those without. As displayed in the barplot above, proportionally more men have heart disease than women. While about 25% of women in the dataset have heart disease, the number increases to 50% for men, which is significantly different.



As seen in the bar plot above, we see that slightly more heart diseased patients have a fasting blood pressure above 120 ml/dg, but it doesn't appear to be significant. The difference in exercise induced Angina, however, does appear to be significant. About 75% of heart diseased patients experience exercise induced angina, whereas approximately 30% of non-heart diseased patients experience the symptom.

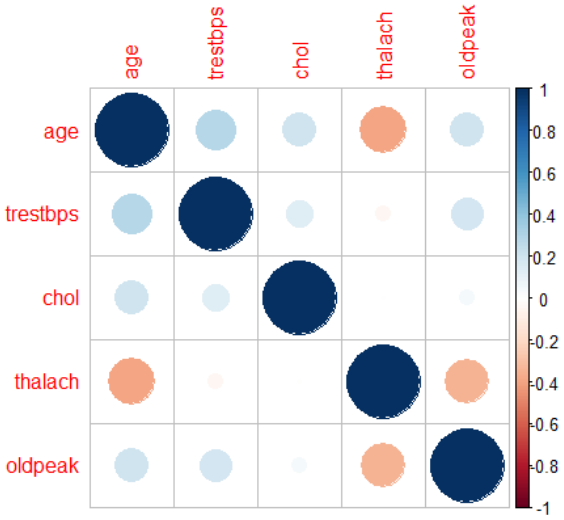


As visualized in the boxplots above, note that those diagnosed with heart disease have a significantly lower maximum heart rate achieved than those without-- and their choloestoral is higher. Their resting blood pressure is also slightly higher on average, but not significantly . Overall, based on the visualizations and rudimentary data anlysis, it appears as though gender, age, exercise angina, maximum heart rate, and cholestoral all can be important factors in the heart disease diagnosis.

Finally, we will check for correlation among the numeric variables.

	age	trestbps	chol	thalach	oldpeak
age	1.0000000	0.28494592	0.208950270	-0.393805806	0.20380548
trestbps	0.2849459	1.00000000	0.130120108	-0.045350879	0.18917097
chol	0.2089503	0.13012011	1.000000000	-0.003431832	0.04656399
thalach	-0.3938058	-0.04535088	-0.003431832	1.000000000	-0.34308539
oldpeak	0.2038055	0.18917097	0.046563989	-0.343085392	1.00000000

As seen in the correlation (above) and the respective visualization (right), age has a significant correlation with resting blood pressure, cholesterol, maximum heart rate, and ST depression induced by exercise relative to rest. These correlations may affect our model, and potentially explain some of the differences we found among the heart disease positive and heart disease negative groups.



Experimental Results:

For this analysis, I focused on Support Vector Machines with various kernel options. With SVM, the algorithm can find a hyperplane that divides the class groups. If needed, SVM can translate the data into a higher dimension and thus find the optimal hyperplane separation. To begin, I created a radial SVM in R. Its accuracy was 0.8302 (with 95% confidence interval of 0.702- 0.9193). The sensitivity was 0.8438 and specificity was 0.8095. Next, I created a linear SVM. Its accuracy was 0.8868 (with 95%

confidence interval of 0.7697- 0.9573). The sensitivity was 0.9062 and specificity was 0.8571. It was much more successful than its radial counterpart.

Then, I created polynomial and sigmoid svm models. The polynomial svm was the least successful of the models. Its accuracy was 0.7736 (with 95% confidence interval 0.6379- 0.8772). The sensitivity was 0.9688, which was impressive, but the specificity was 0.4762. Finally, the sigmoid svm model had an accuracy of 0.8679, sensitivity of 0.8095, and a specificity of 0.9062. Overall, the linear SVM model was the most successful. The confusion matrices for the aforementioned models are presented below. Note that using the svm tuning function with the various kernels did not produce any better model than the standard linear kernel svm I had already created.

Radial:

```
pred  0  1
   0 27  4
   1  5 17
```

Linear:

```
pred  0  1
   0 29  3
   1  3 18
```

Polynomial:

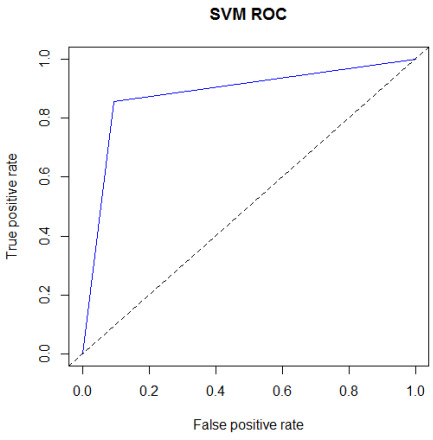
```
pred  0  1
   0 31 11
   1  1 10
```

Sigmoid:

```
pred  0  1
   0 29  4
   1  3 17
```

Experimental Analysis:

With our final model, a linear svm, we correctly identified heart disease in 18 of the 21 patients who have heart disease in Cleveland Clinic's data set. We obtained a high accuracy, with only three heart-diseased patients being misdiagnosed. Additionally, we have an ROC curve shown right. Note that the area under the curve is 0.8817, which is good. For the test data set from the Cleveland clinic, I was also interested, however, in seeing if the model formed from the Cleveland Clinic data could also correctly classify the heart disease status for patients from the other hospitals.



Conclusion:

In conclusion, I found that SVM models are successful in identifying heart disease among patients. Our model had a high accuracy, and could be used by hospitals or medical centers to identify patients who most likely have the disease. In the future, it would be great to test the model against additionally data to confirm its abilities. It's also important to note that our data analysis confirmed several factors may increase risk of heart disease or be signifiers of the disease, notably gender, age, exercise induced angina, cholesterol, and maximum heart rate achieved. It is important to make note of symptoms indicative of heart disease, such as exercise induced angina, and discuss these symptoms with your doctor to hopefully prevent or slow the effects of heart disease.