Heart Disease Classifier

Kevin Dang

Introduction

The heart disease data set (or Cleveland database) is part of a larger collection of databases on heart disease diagnosis. The heart disease data set was downloaded from the UCI Machine Learning Repository and used in the following analysis to classify whether or not an individual has heart disease, which could have important implications for classification and predictive modelling in the field of medicine.

The main objective of this analysis is to assess the performance of three different classification methods: logistic regression with regularization, k-nearest neighbours and classification trees. The secondary objective of this analysis is to train a model that can be used for classification of future observations, using the best of the three methods.

In the Methods section, three classification methods are discussed. Each classifier uses the diagnosis of heart disease as the response variable, with the remaining thirteen variables as the predictors. In the results section, the performance of the three methods are compared and a final classification model is trained for future predictions. Finally, the methods and results are further interpreted in the Discussion section and dealing with heart disease misdiagnosis is discussed.

Methods

The heart disease data set contains 303 observations with some missing values, so complete case analysis was used to reduce the number of observations to 297. Table 1 contains a written description of the fourteen variables and Table 2 contains a detailed summary of the data set. The response variable is `num` - the diagnosis of heart disease (yes/no), while the remaining thirteen variables are the predictors. The response variable along with the thirteen predictors are used in each of the classification methods.

To compare the classifiers while accounting for randomness, the data was randomly split into training and test sets five times, and the accuracy rates were recorded for each run. 80% of the data was used to train the models, and the remaining 20% was used to evaluate their performance. Accuracy is the chosen criterion for assessing the performance of the classifiers since it is straightforward and easy to understand for non-technical readers.

Logistic Regression with regularization

The first classification model is logistic regression with regularization. In the logistic regression model, the probability of heart disease given the predictors X_i is calculated as follows:

$$P(Y=1|X_i) = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik})}$$

There are several regularization methods for regression, including ridge regression and LASSO. For this analysis we used LASSO (least absolute shrinkage and selection operator), as this form of regularization shrinks predictors to zero and is consequently a feature selection method. This makes the penalized model much more simpler as it will contain fewer predictors. The LASSO coefficients minimize the negative log likelihood subject to $\sum_{j=1}^{k} |\beta_j| \leq \lambda$, where λ is the penalty term or tuning parameter. Ten-fold cross-validation was used to find the minimum λ (Appendix: Table 4 & Figure 2).

K-Nearest Neighbours

The second classifier is k-nearest neighbours (KNN). This method takes the k nearest points to the new point, then classifies the point by a majority vote of its k neighbours. In simpler terms, this algorithm takes unlabeled points and assigns them to the class that contains similar labeled examples. K-nearest neighbours uses a Euclidean distance by default, with classification decided by majority vote and ties broken at random. Due to the nature of KNN as a distance based algorithm, is it important to scale the continuous variables so that the distance is not biased towards the variables with larger values. There are a few types of scaling methods including normalization which was used in this analysis. The algorithm also requires that we specify k, the number of neighbours considered. One method to find the optimal k is leave-one-out cross-validation (Appendix: Table 5 & Figure 3).

Classification Tree

The third method for classification is a classification tree (also known as a decision tree). Classification trees use binary recursive partitioning, where the data is partitioned in an iterative manner at each node. A node is where the branch of a tree splits into two parts, and the terminal nodes are called leaves. These leaves are assigned class membership probabilities which can be used to classify new data. After the classification tree was trained, a technique called pruning was used to remove the least important branches as the model can be quite complex which leads to overfitting. When pruning the tree, a parameter called `best` (size or number of terminal nodes of a subtree to be returned) was passed into the function in order to specify the size of pruned tree. The best size was determined via ten-fold cross-validation (Appendix: Table 6 & Figure 4).

Table 1: Heart Disease Data Set

Variable	Description
age	age in years
sex	sex (1 = male; 0 = female)
cp	chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic)
trestbps	resting blood pressure (in mm Hg on admission to the hospital)
chol	serum cholestoral in mg/dl
fbs	fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
restecg	resting electrocardiographic results (0 = normal; 1 = having ST-T wave abnormality; 2 = showing probable or definite left ventricular hypertrophy)
thalach	maximum heart rate achieved
exang	exercise induced angina (1 = yes; 0 = no)
oldpeak	ST depression induced by exercise relative to rest
slope	the slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = downsloping)
ca	number of major vessels (0-3) colored by flourosopy
thal	thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect)
num	diagnosis of heart disease (1 = yes; 0 = no)

Table 2: Detailed Summary of the Heart Disease Data Set

	Total (N=297)	No Heart Disease	Has Heart Disease
Age			
Median (IQR)	56 (48, 61)	52 (45, 59)	58 (53, 62)
Sex			
Male	201 (68%)	89 (56%)	112 (82%)
Female	96 (32%)	71 (44%)	25 (18%)
Chest pain type			
Typical Angina	23 (8%)	16 (10%)	7 (5%)
Atypical angina	49 (16%)	40 (25%)	9 (7%)
Non-anginal pain	83 (28%)	65 (41%)	18 (13%)
Asymptomatic	142 (48%)	39 (24%)	103 (75%)
Resting blood pressure (mm Hg)			
Median (IQR)	130 (120, 140)	130 (120, 140)	130 (120, 145)
Serum cholestoral (mg/dl)			
Median (IQR)	243 (211, 276)	236 (209, 268)	253 (218, 284)
Fasting blood sugar (mg/dl)			
>120 mg/dl	43 (14%)	23 (14%)	20 (15%)
<= 120 mg/dl	254 (86%)	137 (86%)	117 (85%)
Resting electrocardiographic results			
Normal	147 (49%)	92 (58%)	55 (40%)
ST-T wave abnormality	4 (1%)	1 (1%)	3 (2%)
Left ventricular hypertrophy	146 (49%)	67 (42%)	79 (58%)
Maximum heart rate achieved			
Median (IQR)	153 (133, 166)	161 (149, 172)	142 (125, 157)
Exercise induced angina			
Yes	97 (33%)	23 (14%)	74 (54%)
No	200 (67%)	137 (86%)	63 (46%)
ST depression induced by exercise			
Median (IQR)	0.8 (0.0, 1.6)	0.2 (0.0, 1.1)	1.4 (0.6, 2.5)
Slope of the peak exercise ST segment			
Normal	139 (47%)	103 (64%)	36 (26%)
ST-T wave abnormality	137 (46%)	48 (30%)	89 (65%)
Left ventricular hypertrophy	21 (7%)	9 (6%)	12 (9%)
# of major vessels colored by flourosopy			
0	174 (59%)	129 (81%)	45 (33%)
1	65 (22%)	21 (13%)	44 (32%)
2	38 (13%)	7 (4%)	31 (23%)
3	20 (7%)	3 (2%)	17 (12%)
Thalassemia			
Normal	164 (55%)	127 (79%)	37 (27%)
Fixed Defect	18 (6%)	6 (4%)	12 (9%)
Reversable Defect	115 (39%)	27 (17%)	88 (64%)
Diagnosis of Heart Disease			
No	160 (54%)	160 (100%)	0 (0%)
Yes	137 (46%)	0 (0%)	137 (100%)

Results

In Figure 1, the accuracy rates of the three classification methods for the five independent runs are shown. In four out of five runs, LASSO-penalized logistic regression had the highest accuracy rate, making it the preferred classifier among the three options. The k-nearest neighbour classifier had the joint highest accuracy in Run 3 and had the second highest accuracy in three of five runs. The classification tree scored the highest in Run 1 but had the lowest accuracy in the remaining four runs. In the Appendix, Table 7 contains the accuracy rates for each run and Table 8 summarizes the overall results by taking the average accuracy rates of the three methods. Logistic regression with LASSO regularization is the clear winner, followed by k-nearest neighbour and finally classification tree in last place.

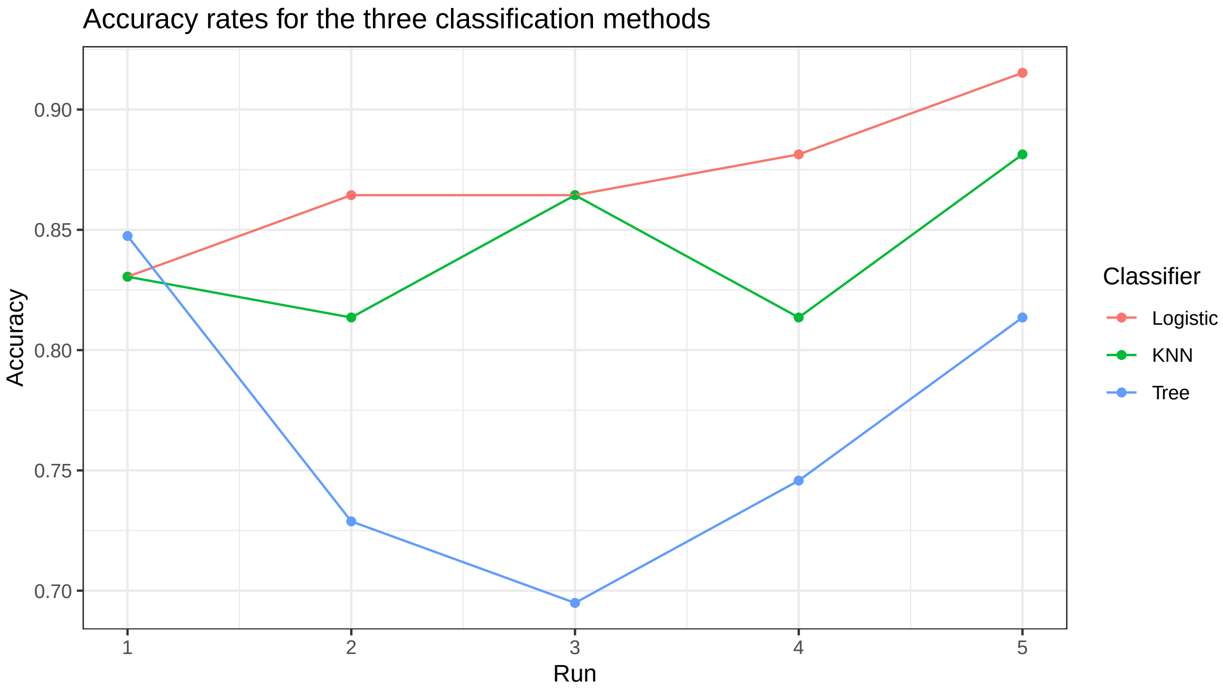

Figure 1: Accuracy rates for the three classification methods

Based on the assessment of the performance of the three classification methods, a LASSO-penalized logistic regression model was then trained to be used for classification of future observations. Previously, the data was split into training and test sets, however this model was trained on the entire data set because it will be used to predict on future observations which can be treated as our test set, as the model has not seen this new data. The coefficients of the predictors are shown in Table 3 below. It is interesting to note that the coefficients of the predictors `age`, `cp3`, `slope3` and `thal6` have been shrunk to zero. There are also several coefficients that are close to zero, most notably `trestbps`, `chol`, `restecg1`, `thalach`. The descriptions for these variables can be found in Table 1.

Table 3: LASSO-penalized logistic regression model to predict future observations

Predictor	Coefficient
(Intercept)	-4.624
age	0.000
sex1	1.132
cp2	0.480
cp3	0.000
cp4	1.656
trestbps	0.016
chol	0.002
fbs1	-0.173
restecg1	0.007
restecg2	0.307
thalach	-0.014
exang1	0.526
oldpeak	0.384
slope2	0.823
slope3	0.000
ca1	1.629
ca2	2.184
ca3	1.511
thal6	0.000
thal7	1.262

Discussion

Using the heart disease data set, the five repeated runs for our specific analysis came to a consensus that the LASSO-penalized logistic regression classifier was the best performer, followed by k-nearest neighbour and lastly classification tree. Due to the randomness of the splits of the training and test data as well as cross-validation to tune the parameters, a seed was set to ensure reproducibility. Without setting a seed, the results would vary each time the algorithms are trained and used for prediction - the performance of the three classifiers would change. This means that we cannot necessarily conclude that one classifier is better than the others in general based on these results. However, it is comforting knowing that the results achieved in this analysis are quite promising and that these are just three of the numerous possible classification methods that can be used for heart disease diagnosis.

In the field of medicine, there are important considerations that may affect how we approach a classification problem. For instance, if a false negative (i.e. individuals with heart disease incorrectly identified as healthy) is more costly than a false positive (i.e. healthy individuals incorrectly identified as having heart disease), then the classifier must be adjusted appropriately. One strategy is to set the predicted probability threshold lower, so that more individuals will be classified as having heart disease. This will increase the false positive rate and decrease the false negative rate. As an example, one can first obtain the predicted probabilities from the `predict()` function, then instead of assigning the predictions to class 0 or class 1 based on the threshold of 0.5 we can set the threshold to a smaller value such as 0.4. That means that if the predicted probability is greater than 0.4, the individual would be assigned to class 1 (having heart disease). A lower predicted probability threshold such as 0.4 ensures more individuals are classified as having heart disease compared to a threshold of 0.5.

Appendix

Tables

Table 4: Minimum values of lambda across 5 runs

Run	1	2	3	4	5
Lambda	0.009	0.015	0.013	0.012	0.018

Table 5: Optimal values of k across 5 runs

Run	1	2	3	4	5
k	3	9	12	5	9

Table 6: Best tree size across 5 runs

Run	1	2	3	4	5
Size	6	7	2	9	9

Table 7: Accuracy rates for the three classification methods

Run	Logistic Regression	K-Nearest Neighbour	Classification Tree
1	0.831	0.831	0.847
2	0.864	0.814	0.729
3	0.864	0.864	0.695
4	0.881	0.814	0.746
5	0.915	0.881	0.814

Table 8: Average accuracy rates for the three classification methods

Logistic Regression	K-Nearest Neighbour	Classification Tree
0.871	0.841	0.766

Figures

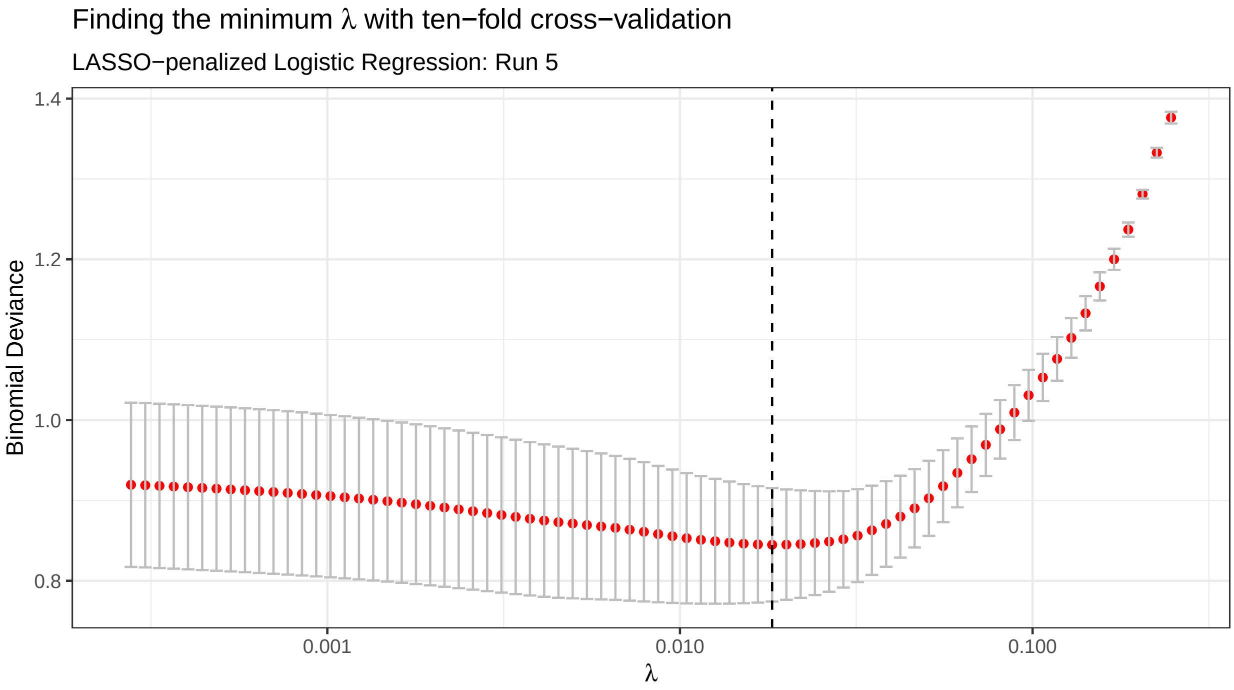

Figure 2: Finding the minimum λ with ten-fold cross-validation (Run 5)

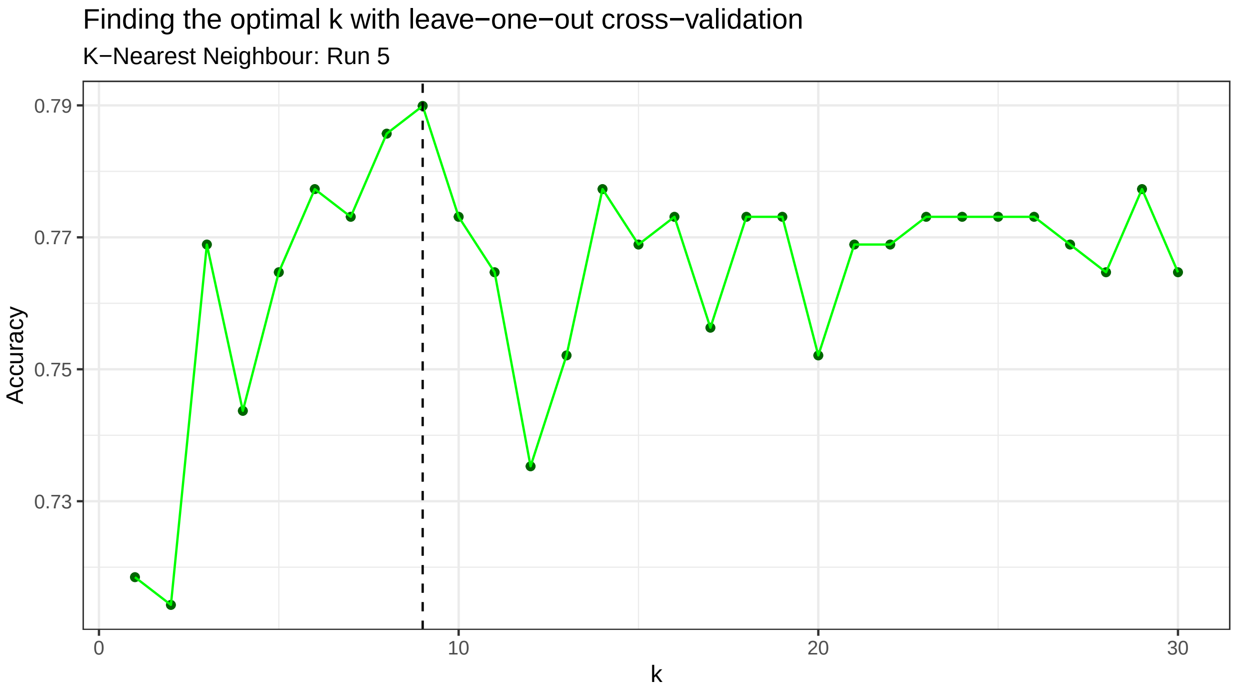

Figure 3: Finding the optimal k with leave-one-out cross-validation (Run 5)

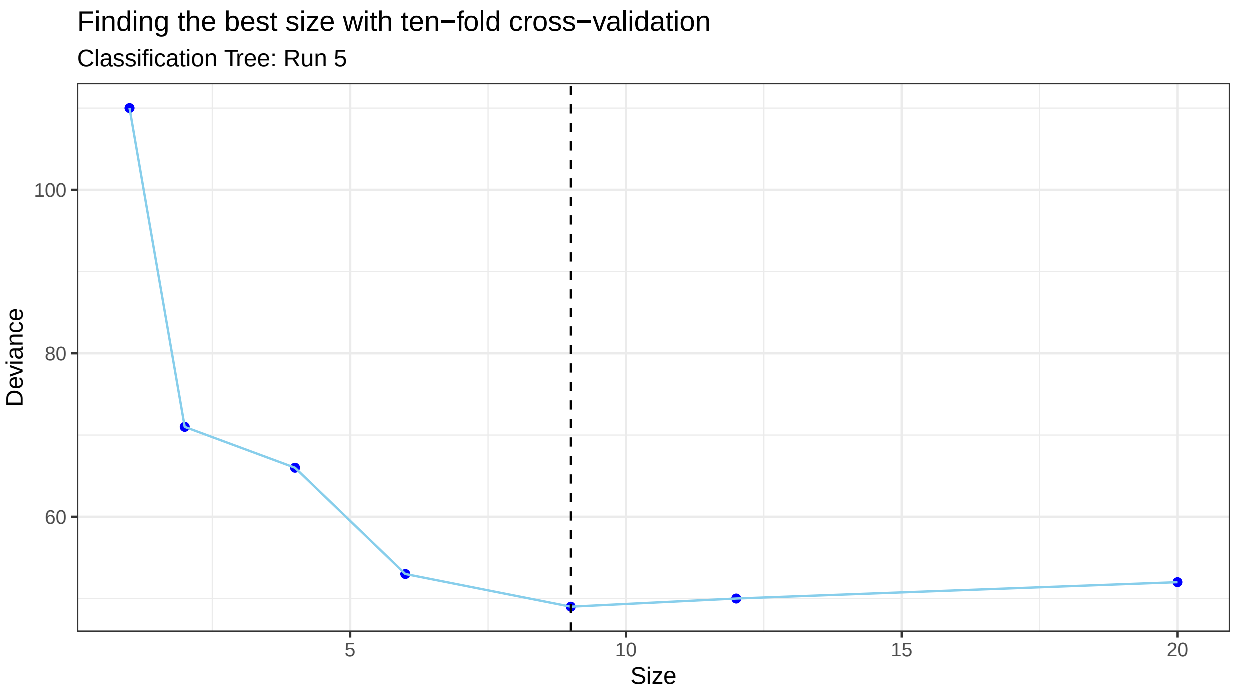

Figure 4: Finding the best size with ten-fold cross-validation (Run 5)

References

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: Springer.