Yiling Chen

Dec. 14, 2020

Hw#7 Write-up

**Question 1**

The purpose of this question is to compare the prediction results of neural network, CART and random forest models using the Cleveland heart disease data. This dataset contains 296 observations on 13 predictive variables, where the response variable is diag1. The 13 predictive variables include a range of biological measurements and diag1 shows whether the observant has heart disease (sick) or not (buff).

Graphical user interface, text, application

Description automatically generated Graphical user interface, text, application, email

Description automatically generated

My first step is to download some necessary packages such as nnet for neural network, randomforest for randomforest and rpart for CART model. Then, in order to do neural network, I converted all the data into numerical values. Hence, 13 predictive variables turns into 18 where diag1.sick becomes the response variable.

Text

Description automatically generated

My next step is to divide the new data into test, training and set the y\_true in order to make a prediction and look at the error rate. I have divided my training and testing as 80% and 20%.

Text

Description automatically generated

Now, train the neural network putting all the variables in the parameter and using the training dataset, as well as set hidden = 3. I have tried setting hidden to other numbers, but it seems like when neurons are 3, the performance is better. A visualization of the neural network is shown below.

Diagram, schematic

Description automatically generated

Next, I compute the training and testing error using our prediction function. The training error is 11% and the test error is 20%. This is a good prediction because the error rate is relatively low.

Graphical user interface, text, application, email

Description automatically generated

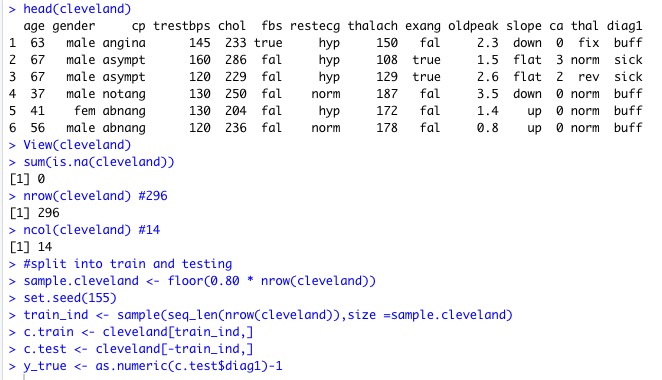
Finally, we want to tune the model to see how many neurons perform better (lowest error rate). The result shows that when hidden is at 3, the neural network performs the best. From tunning, we see that neurons is optimal with train = .05, and test = .16.

Text

Description automatically generated

Now, I want to compare the error rate of nnet with CART model. Since the variables doesn’t need to be numeric for CART, I use the original dataset because there are fewer variables.

I first set the test and training as 20% and 80% for the Cleveland dataset, then I check to see if there are any missing values, there is none.

Graphical user interface, text, application

Description automatically generated

Next, I built CART model by using rpart() function. Here we use training data and the model control we want to set the xval to 10 and cp to 0 because we want to see everything.

Then, I found out that the min\_cp is 2 there that is where I will prune my tree base on.

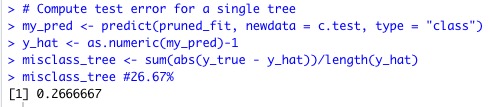
Graphical user interface, text, application

Description automatically generated

Diagram, engineering drawing

Description automatically generated

After getting the result of pruned tree, we want to look at the error rate for our single tree. Using our equation, we found out that the error rate of the CART model is 26.6%



Similarly, using the randomforest function, I have use diag1 as the respond variable and the rest as predict variables, where data is the training data with 1000 trees to be grown. Then we compute the test error for random forest and the result we got back is 13.33% for the testing.

Text

Description automatically generated

Graphical user interface, application

Description automatically generated