Yiling Chen

11/15/2020

HW#5 Write-up

**Question 1**

The purpose for this problem is to construct a classification tree and prune it to the correct size for the vehicle data. The vehicle data set contain 564 observations and 20 variables. There are 4 types of vehicles as a variable along with a set of features of the cars as the other 18 variables.

Graphical user interface, text, application

Description automatically generated

My first step is to load the vehicle data and download some packages (rpart) to be able to perform tree growing and pruning. After a brief understanding of the vehicle dataset and see there are no missing data within the vehicle dataset, I notice that in order to grow a classification tree and prune it to the correct size, I need to eliminate my class column and use classdigit column for tree growing and pruning.

Text

Description automatically generated

I then, set the parameter using rpart.control and set cp to 0 because I want to see every improvement. Then I set the classdigit as the variable and the rest as the predictors and use classification as the method.

Text

Description automatically generated

Next, I take a closer look at the different splits if we grow a tree later, as well as the perfect place to prune the tree, which is the minimum point of the cp table, in this case, 11.

Chart, scatter chart

Description automatically generated

Finally, I plot the full tree and pruned tree after manipulating the branch and cex value and turn the trees to a more friendly readable size. As we can see in the pruned tree, the splits between each variables of the column result in the classdigit label.

A picture containing text

Description automatically generated

Diagram, engineering drawing

Description automatically generated

Diagram

Description automatically generated

In conclusion, using the rpart function, we successfully grow and prune the tree of the vehicle dataset. From this tree we can see that the type of vehicles can be identified base on the values of other variables within the dataset. By pruning the tree, we eliminated unnecessary splits in the tree.

**Question 2**

The purpose for this question is to use the prostate dataset to find the best subset and to compute the AIC, BIC, five-and ten-fold cross-validation, and bootstrap .632 to estimates the prediction error of each one of these methods. The prostate dataset contains 9 variables and 97 observations on the correlation between the level of prostate specific antigen and other measures.

Graphical user interface, text, application, chat or text message

Description automatically generated

After load the data into RStudio, I first take a closer look at the prostate dataset, look at the summary, and check to see if there are any missing data. My next step is to divide the prostate into training and testing sets so I can do further predictions base on that.

Graphical user interface, text, application

Description automatically generated

Next, I use lpsa as the predictor and use exhaustive search method to find the minimal error would be at the lowest point of AIC and BIC table. The result came back as the 3 is the best for both AIC and BIC.

A picture containing text

Description automatically generated

As we can see better in the graph down below, both BIC and AIC are at their lowest when using three variables, which are lcavol,lweight and svi.

Chart, line chart

Description automatically generated

The next step is to see test and training error on the exhaustive method base on the minimum error rate when BIC and AIC are at 3.

Graphical user interface, text, application

Description automatically generated

We can see the error rate of test and training in a clear detail base on the graph I generated using quartz(). As we can see, the train error is not the best at its 3 variable which is 45.21% and test is best at its 3 variable as well which is 62.83% however since the difference in error is not that big, we can use 3variables instead of 8 where the error is the lowest for test and train.

Chart, line chart

Description automatically generated

After finding the error rate for AIC and BIC, the next step is to compute five and tenfold cross-validation and estimate their prediction error. The result is when k = 5 the lowest error rate is at 8and when k=10 the lowest error rate is at 3.

Graphical user interface, text, application, email

Description automatically generated

Put the result of error rate of k=5,10 into a graph for connivence, when k=5,10 the error rates are the lowest at the eighth variable when k = 5 which is 50.57% and at third variable when k=10 which is 51.21%. But since the difference when k=5 at third variable is not a big difference when at eighth variable, we can use the third variable for k=5 as well.

Chart, line chart

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Finally, using the bootstrap.632 to predict the error rate with the best subset which is at 3. The result I got back is 51.7%, which is the lowest error rate as shown in the graph.

In conclusion, by using the best subset exhaustive method, for AIC, BIC, five-fold and ten-fold cross-validation, and bootstrap.632 all suggests that when k=3 the error rate is at its minimum. The error rate for test and training error for AIC and BIC are 62.83%, 45.21%. The error rate for k=5 is 50.57% k=10 is 51.21%, and for bootstrap.632 is 51.7%. Similar error rates between these methods make sense because we are using the best subset method to predict the error.

**Question 3**

This data set is the result of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. This dataset contains 178 observations of 11 variables. I will first construct a classification tree, then apply random forest, finally construct an LDA model and compare their performances by looking at their error rates.

Graphical user interface, text, application, chat or text message

Description automatically generated

My first step is to split the wine data into test and train (80% training and 20%testing), as well as set parameter for the tree modeling, using Y as a predictor.

Graphical user interface, text, application

Description automatically generated

Secondly, I want to get a brief understanding of where the tree will be split on, and the best place to prune tree. I found that when cp = 6 we are likely to get the best and no extra information when growing a tree. Therefore, I constructed a pruned tree base on when cp = 6.

Text

Description automatically generatedChart, scatter chart

Description automatically generated

Next, I plot the tree and compute the error to see the accuracy of predictor using our tree model by using the test dataset.

A picture containing chart

Description automatically generatedGraphical user interface, text, application

Description automatically generated

After plotting the pruned tree, I calculated the error rate for using tree to predict the category of Y. The result is an 8% error rate.

Graphical user interface, text, application, email

Description automatically generated

Table

Description automatically generated

Furthermore, I use randomForest method to look at the misclassification rate using Y as the predictor again. The result is a 9.71% error rate, which is acceptable and similar with the pruned tree method.

Finally, I generated an LDA model, and using the regsubset to compute and predict the result as shown in the table down below. As we can see, the lda.pred has a significant amount of error when predicting Y=3. But for when Y=1,2 the error rate is actually low.

Text

Description automatically generated

In conclusion, based on mean square error calculation, the performance error rate for LDA is 58.33%, random forest is 9.71%, tree model is 8.33%. Base on the error rate, the classification tree model and random forests are better predictor than LDA model.

**Question 4**

The purpose of this problem is to use the covertype dataset and perform a prediction of the forest cover type, which is the predictor of this dataset. The covetype dataset contains 581,012 observations on 54 input variables. This dataset was obtained from the US Forest Service and are concerned with seven different type of forest cover.

Graphical user interface, text, application

Description automatically generated

My first step is to load the data and check to see if there are any missing data alone with get a deeper understanding of the dataset. In order to do further prediction, I divide the dataset into 80% training data, and 20% testing data.

Graphical user interface, text

Description automatically generated

I then set the parameter to grow a prune classification tree. First, I find the minimum cp value which will determine the split of the prune tree ( the result is 367), which means it will be a very thick tree. Therefore, instead of plot the tree out, we could look at each splitting plot base on the function fit.covertype$splits and fit.covertype$cptable as shown below.

Finally, I compute the misclassification error rate base on our tree using the test set that I previous set. The result I got back is 15% error rate using tree model to predict.

Text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

In conclusion, the question ask us to divide the dataset into test and training and make a prediction using any prediction model we can, therefore we could use the classification tree model to predict the type of forest (V55) from the dataset, even though the error rate is relatively high.