ST4061 – Computer Intensive Statistical Analytics II

2022-23

Continuous Assessment 2

Question 1

Figure 1 below shows the output of a random forest model fit to a sample of data points. This dataset comprises of 5 variables: Length, Width, Leaf, Curve and Age (Young; Intermediate; Mature).

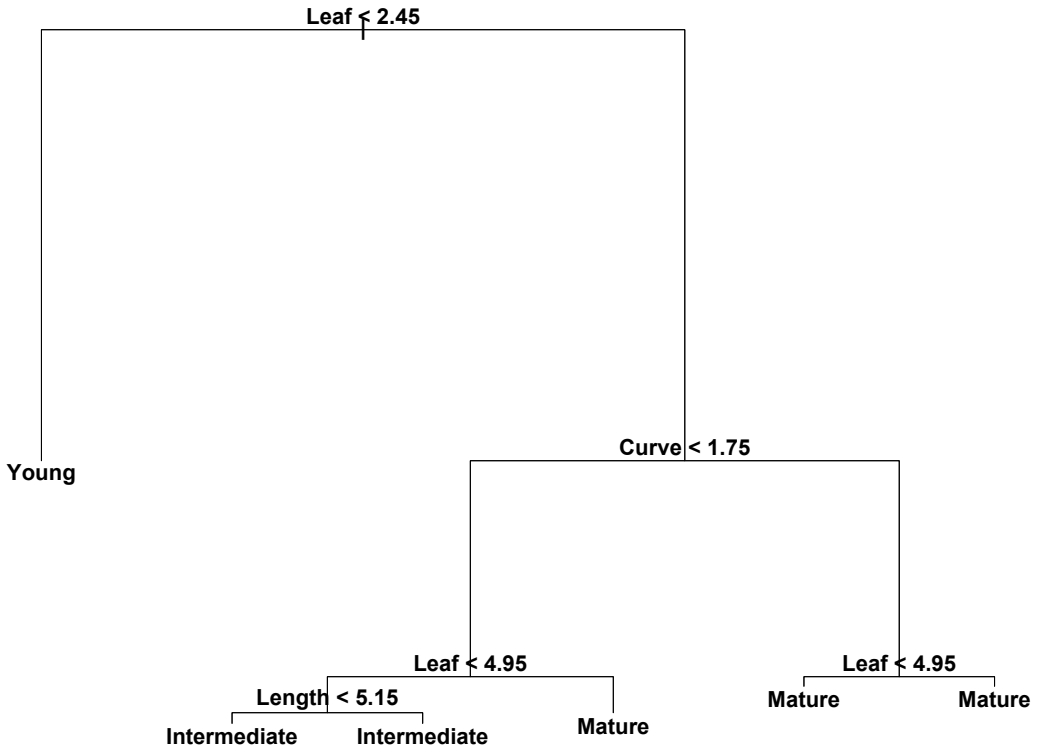

Figure 1 – Diagram for Question 1(a).

(1) Indicate which of the 5 variables is the response variable Y. Briefly justify your answer.

(2) Quote the number of nodes that would be required in each layer of the architecture of a fully connected, single-hidden-layer, feed-forward neural network with 5 hidden neurons, for that model to be applicable to this dataset. Briefly justify your answer.

(3) How many model coefficients would need to be estimated in total, in order to fit the neural network described in (2) to this dataset? Briefly justify your answer.

Answers to Question 1

Question	Your answer
1	
2	
3	

Question 2

You are required to use the **caret** package for the entire question.

Load the following datasets into your R session:

```
X = read.csv(file="Q2_X.csv", header=TRUE)
Y = read.csv(file="Q2_Y.csv", header=FALSE, stringsAsFactors=TRUE)[,1]
X.valid = read.csv(file="Q2_Xvalid.csv", header=TRUE)
Y.valid = read.csv(file="Q2_Yvalid.csv", header=FALSE,
                                    stringsAsFactors=TRUE)[,1]
```

The eight measurements in **X** and **X.valid** correspond to percentages by weight of oxides contained in a sample of glass fragments. These glass fragments were classed into six types. The aim here is to build a methodology for classification of glass fragments on the basis of concentration of these eight oxides (Na: sodium, Mg: manganese, Al: aluminium, Si: silicon, K: potassium, Ca: calcium, Ba: barium, Fe: iron).

Note: do not apply scaling to the set of predictors or the response variable **Y**.

(1) Train a kNN classifier on the data (**X**,**Y**) via single 5-fold cross-validation using the appropriate function from the **caret** package with default settings (other than those specified here for cross-validation). Run set.seed(6041) before generating your output. Quote:

   (a) the optimal value of *k* found via CV;
   (b) average test-set prediction accuracy for this classifier.

(2) Train a random forest with recursive feature elimination via single 5-fold cross-validation as in (a) on the (**X**,**Y**) dataset. Run set.seed(6041) before generating your output. Quote:

   (a) The set of optimal predictors;
   (b) the cross-validated accuracy corresponding to the best predictor subset found for the random forest.

(3) Generate predictions for the validation set by applying the random forest trained in (2) on **X.valid**. Quote the validation set prediction accuracy for this classifier.

(4) Train a single hidden-layer neural network using method **nnet** on the (**X**,**Y**) dataset via single 5-fold cross-validation using the appropriate function from the **caret** package. Run set.seed(6041) before generating your output. Generate predictions for the test data **X.valid** from that fit. Quote:

   (a) the size of the input layer for the optimal architecture for this classifier;

(b) the number of neurons in the hidden layer for the optimal architecture for this classifier;
(c) the validation set prediction accuracy for this classifier.

Answers to Question 2

Question	Your answer
1(a)	
1(b)	
2(a)	
2(b)	
3	
4(a)	
4(b)	
4(c)	

R code for Question 2:

Question 3

Consider the problem of performing Principle Components Analysis on the R dataset **mtcars**.

(1) Should one perform a PCA on the scaled or unscaled data? Briefly justify your answer

(2) Remove variable **mpg** from the dataset, and perform PCA on the rest of the data (without any data splitting). Compare the first eigenvector from each analysis, and briefly comment on the values contained in these vectors in terms of the role of each feature. What causes the difference, and why?

(3) Based on the PCA of the scaled data, indicate which 5 variables seem to capture most of the information. Briefly justify your answer.

(4) Fit a linear regression model using these 5 variables as predictors X, and **mpg** as the dependent variable Y. Are any of these variables significant predictors? Briefly explain why they are or are not.

Answers to Question 3

Question	Your answer
1	
2	
3	
4	

R code for Question 3: