STAT330/430 - Statistical Learning

Assignment 5

Due date - 30th May

Note: You must submit two files for this assignment:

- a pdf document of your solutions, and
- a complete and concisely annotated *R Script* file of all **R** analysis that was undertaken to produce your results.

In addition, to receive marks for Question 1 you are required to engage in the Topic 5 moodle discussion forum.

Please refer to the Assignment assessment criteria document for additional guidance.

Question 1 [5 marks]

In each assignment you will be able to earn up to 5 marks based on your engagement on the moodle Discussion forums from the topics associated with the given assignment. See the Assignment assessment criteria document for more details.

Question 2 [40 marks]

The *Toy* data-set is a simulated data-set that contains a response variable, Y, along with nine predictor variables. The following tasks require you to use the *neuralnet* package in **R**. Use a 75:25 train:test split for parts (a) and (b). Use a set.seed value of 430 for all questions.

(a) Train a neural network with only one hidden unit to predict the response variable using all predictor variables. Produce a graph of your trained neural network. Provide an overview of your analysis and results, including the correlation coefficient of actual and predicted Y in the data-set.

(b) Train 7 neural networks, each time adding one extra hidden node. Produce a summary graph of the correlation coefficient of each network vs the number of hidden units. Provide an overview of your analysis and results.

(c) Train 4 neural networks with differing training and testing splits as follows: [20% training; 80% testing], [40% training; 60% testing],[60% training; 40% testing], [80% training; 20% testing]. Produce a summary graph of the correlation coefficient of each network vs the percentage of training set data used. Provide an overview of your analysis and results.

Hint: for parts (b) and (c) use a *for loop* in your code to automate the process.