# Neural Networks, Regression Task, Exercise. AMS 580

# Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_SBU ID:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Dear all, please learn how to perform the regression task for a continuous outcome measure using neural networks through this exercise. You will then be ready to do Quiz 7, which is a take home quiz due Monday, March 27, before class (8:30am).

# Please include (1) RMD code; (2) Output from RMD.

#### Neural Network with the Boston Housing price Data – Regression Task

The Boston housing price data we will use for our homework is taken from the Kaggle competition site (<https://www.kaggle.com/vikrishnan/boston-house-prices>). **We will treat this dataset as our entire data.**

Each record in the database describes a Boston suburb or town. The data was drawn from the Boston Standard Metropolitan Statistical Area (SMSA) in 1970. Our data has 506 observations and 14 variables:

Attribute Information (in order):

**CRIM:** per capita crime rate by town

**ZN:** proportion of residential land zoned for lots over 25,000 sq.ft.

**INDUS:** proportion of non-retail business acres per town

**CHAS:** Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)

**NOX:** nitric oxides concentration (parts per 10 million)

**RM:** average number of rooms per dwelling

**AGE:** proportion of owner-occupied units built before 1940

**DIS:** weighted distances to five Boston employment centers

**RAD:** index of accessibility to radial highways

**TAX:** full-value property-tax rate per 10,000 dollars

**PTRATIO:** pupil-teacher ratio by town

**Black:** The result of the equation B=1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town

**LSTAT:** %lower status of the population

**MEDV:** Median value of owner-occupied homes in 1000 dollars

First, one must clean the data and decide which variables to exclude from our analysis. My recommendation is that we exclude *RAD* in the ensuing analysis. Now after the data cleaning step, your task is to split the data randomly into training (75%) and testing (25%), and normalize the data, first build the best neural network models to predict median house value using the training data, and then use these models to predict the house price in the testing data. Please use the *neuralnet* package in R to build the various neural network models.

**Note:** In the past, we did not perform data standardization (normalization) – just to make it easier for you. However, for real world application, we do need to consider normalization for all machine learning methods, including the neural network. So from this homework assignment, we have added the normalization part using the R **scale()** function, which is for each variable to subtract its sample mean, and then divide by its sample standard deviation (<https://www.r-bloggers.com/2021/12/how-to-use-the-scale-function-in-r/>).

Please also note that at the end of the solutions, Ian has added some instructions on how to convert a binary string variable first to a numerical variable, and then you can normalize it – which will not affect the results because it will still be represented by two numerical values.

Please review the following websites for related methods and concepts:

1. **neuralnet**:

<https://journal.r-project.org/archive/2010/RJ-2010-006/RJ-2010-006.pdf>

<https://datascienceplus.com/fitting-neural-network-in-r/>

<https://rpubs.com/shyambv/linear_neural_network>

2. **multiple linear regression**:

<http://www.sthda.com/english/articles/40-regression-analysis/165-linear-regression-essentials-in-r/>

1. For the entire dataset, please perform the data cleaning as instructed before; namely, exclude the variables *rad*. Then please use the random seed 123 to divide the cleaned data into 75% training and 25% testing. Please normalize the data using the R scale() function.

1. Please first build the predictive model to predict the house price using the training data and the NN model with (i) no hidden layer, (ii) the default loss function of “sse”, and (iii) the default activation function of “identity”. Please plot the model obtained using the training data. Please compute the Test MSE using the testing data.
2. Please first build the predictive model to predict the house price using the training data and the NN model with (i) one hidden layer with 3 neurons, (ii) the default loss function of “sse”, and (iii) the default activation function of “identity”. Please plot the model obtained using the training data. Please compute the Test MSE using the testing data.
3. Now we shall build the predictive model to predict the house price using the training data and the Multiple Linear Regression model. Please compute the Test MSE using the testing data. Please report the fitted linear regression model obtained using the training data – and compare to the NN model in Question 2. Please plot the predicted house price using the method in Question 2 (NN with no hidden layer), Question 3 (NN with one hidden layer, linear activation function), and Question 4 (MLR: multiple linear regression) (all three on the y-axis), against the true values of the house price, using the testing data. Please also plot the predicted house price using the neural network method in Question 2 (NN with no hidden layer) (on the y-axis) versus the predicted house price using the multiple linear regression method in Question 4 (MLR) (on the x-axis), using the testing data.
4. Please first build the predictive model to predict the house price using the training data and the NN model with (i) one hidden layer with 3 neurons, (ii) the default loss function of “sse”, and (iii) the output layer with the default activation function of 'identity', but the hidden layer with the activation function of 'relu'. Please compute the Test MSE using the testing data. Please plot the predicted house price using the method in Question 5, and the predicted house price using the multiple linear regression in Question 4 (both on the y-axis), against the true values of the house price, using the testing data.

