National University of Singapore
School of Computing

SWS3018 Predictive Analytics
Lab 6

Learning Objectives

- Perform regression using regression tree on R
- Perform classification using classification tree on R

1. In this exercise, we will be using the `Boston` dataset from the MASS R package. The data set has 506 rows and 14 columns. The variables are:
   - **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** : nitrogen oxides concentration (parts per 10 million)
   - **rm** : average number of rooms per dwelling
   - **age** : proportion of owner-occupied units built prior to 1940
   - **dis** : weighted mean of distances to 5 Boston employment centres
   - **rad** : index of accessibility to radial highways
   - **tax** : full-value property rate per $10,000
   - **ptratio** : pupil-teacher ratio by town
   - **black** : 1000(Bk – 0.63)^2 , where Bk is the proportion of blacks by town
   - **lstat** : % of lower status of the population
   - **medv** : (response variable) median value of owner-occupied homes in $1000s

   a) Install the MASS R package

   b) Use the sample() function to provide the dataset into 2 parts. 50% for training, 50% for testing data

   c) Using all the predictors and the training data (i.e. 50% of the full dataset), generate a regression tree for predicting medv.

   d) Based on the generated decision tree, what can you say about the relationship between lstat and medv?

   e) Using the `predict()` function and your decision tree, determine the MSE for your testing data. How much does the prediction deviate from the actual median value (response)?

   f) Try performing tree pruning. Plot the pruned tree. What is the MSE of your pruned tree? Does it make a difference depending on your training/testing dataset?

   g) Generate a linear regression model using the same training data and compare the MSE of the linear regression model with the regression tree model. Does it

make a difference depending on your training/testing dataset? (Try running a simulation of 10 runs and calculate the average MSE)