CS 5751 – Spring 2018 – Homework 6
Assigned: 03/20/2018
Due: 03/27/2018
Total points: 100 pts.
Submit a soft copy to canvas. Remember to write your name at the top of each file you submit.

Objectives: The objectives of this homework are the following:

- Learn how to compare different regression models with validation sets, LOOCV and k-fold cross validation.
- Learn how to generate decision trees using CART, and how to estimate its error.

Notes:

- This homework is to be done individually. You may discuss with your classmates, but the work that you write must be your own.

Activity 1: (50 pts.) (Comparison of regression models) Using either R or Python, do the following:

a) (0 pts.) Download the auto-mpg dataset from the UCI website.
b) (0 pts.) Do any pre-processing that may be needed.
c) (5 pts.) Write down the equations for two different linear models such that one of the models can be derived from the other by adding an additional term. That is, if the first of your models is $mpg \approx \beta_0 + \beta_1 disp,$ then the other model can be of the form $mpg \approx \beta_0 + \beta_1 disp + \beta_2 disp^2$ or $mpg \approx \beta_0 + \beta_1 disp + \beta_2 weight$ or $mpg \approx \beta_0 + \beta_1 disp + \beta_2 disp \cdot weight.$
d) (15 pts.) You will now compare your two models using the MSE and three different strategies:
    a. Use a random subsampling approach 10 times so that each time, you will split your *whole set* (from part 1b) into two parts (80% and 20%): a training set, and a test set. Then train both models on the new training set, and evaluate on the test set. Each time you will obtain a different MSE. Output each of the 10 MSE values that you obtain for both models (10 values each).
    b. Use k-fold cross validation with k = 10. For each fold you will obtain a different MSE. Output each of these 10 MSE values that you obtain for both models (10 values each).
    c. Use LOOCV.
e) (5 pts.) Explain what these three approaches: random subsampling, k-fold CV and LOOCV are trying to accomplish.
f) (10 pts.) Plot a figure containing two boxplots: one with the 10 MSEs you obtained for one of the models in part (1da), and the other with the 10 MSEs that you obtained for the other model in part (1da).
g) (10 pts.) Plot a figure containing two boxplots: one with the 10 MSEs you obtained for one of the models in part (1db), and the other with the 10 MSEs that you obtained for the other model in part (1db). In each of these three approaches, output the MSE for both models.

h) (5 pts.) Based on the figures and results you obtained in parts (1d) to (1f), which of the models would you pick? Why? You need to provide enough details in your answer. It cannot simply be: “Choose model two. Less error.”

For this activity, write a Jupyter notebook named yourLastName_hw6_q1.ipynb that implements 1a through 1h.

Activity 2: (50 pts.) (Decision trees) Using either R or Python, do the following:

1. Write a program to perform the following tasks (a)-(g) on the Hepatitis dataset from the UCI Machine Learning repository (http://archive.ics.uci.edu/ml/datasets/Hepatitis):
   a. (5 pts.) Output the boxplots of the attributes bilirubin, alk phosphate, and sgot in one figure.
   b. (5 pts.) From the boxplots of the three attributes of Task (a), identify which attributes have outliers, and justify your answers. If there are outliers, write code to remove the entire tuples containing the outliers from the dataset and print the dataset after those tuples have been removed.
   c. (5 pts.) Replace the missing values of each categorical attribute in the dataset with its mode, and replace the missing values of each continuous attribute in the dataset with its mean, and print the means and modes (according to the nature of the attribute) of each attribute after the missing values have been replaced.
   d. (35 pts.) Using the preprocessed dataset obtained from Tasks (b) and (c) and using the Cart algorithm, build a decision tree that classifies the tuples based on the class attribute ‘class’ in the dataset. Output the resulting decision tree in textual format. Then evaluate the error rate of the tree using k-fold cross validation sampling with k=10 folds. For each iteration of k-fold CV, print the confusion matrix to standard output, then calculate, print and store the error rate. For this section, you may use any package that implements Cart; you do not need to implement it by yourself.

For this activity, write a Jupyter notebook named yourLastName_hw6_q2.ipynb.