16-791 Applied Data Science – Spring 2018, Section A

Homework #1

Reports are due at 12:00pm (noon) on Monday, February 12.

Please upload your report and code on canvas.

Please remember to properly format for printing all documents submitted electronically.

Individual work will be strictly enforced!

You may choose to use either python or R for this homework, but you have to stick to your choice for the entire homework. Please abide by the following R/Python code guidelines for homework submissions:

1. Do not use any custom or additional packages apart from the ones specified in the assignment or recommended in class.
2. Please include comments explaining the purpose and applicability of each function.
    a. R: When you are done codling, clear your workspace (in RStudio: Workspace → "Clear All") and hit 'Source'. The code should run in its entirety. If it breaks in the middle, you WILL lose credit.
    b. Python: you should save your script to file and it should run when you execute it from command line.
3. Use relative paths (not absolute paths!) and assume your files and data are in your current working directory. We will test your code using different data that will reside in the current working directory during testing.
4. NEVER hard-code for a specific dataset/file, unless specifically requested to do so. For instance, do not assume that attribute names or cell values are fixed as they may change in subsequent tests. We will execute your code to test it, perhaps with different data than used in this assignment. Hard coded scripts that do not work on compatible but different data will receive zero credit.

R packages permitted for this homework: gplots, ggplot2, FNN, plyr. Please refer to Homework 0 on how to download and install packages.
Python packages permitted for this homework: pandas, matplotlib, numpy, scipy, scikit-learn.

In this assignment, you will assume the role of a data mining consultant for a real estate agency. They consider offering an automated valuation service for prospective customers who consider selling their homes. The current valuation method is based on historical prices of nearby properties, which are often outdated and the process requires a substantial amount of work to manually adjust values. The agency seeks improvements by using the available historical transaction data more creatively. You have been asked to construct a predictive model to estimate selling price using known characteristics of properties. They have provided reference data. You are required to write code in R or python to conduct your analysis, and assemble the results of the analysis in a set of Power Point slides to present at the next client meeting.

Problem 1

a) Please write an R or python function named “brief” which produces a high-level description of a set of data. The input of the function is a data frame, the output is a text in the following format (note that the values shown below are just for reference purposes, the actual values for your data may vary):

```
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
brief function output for house_no_missing.csv
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
This dataset has  506  Rows   9  Attributes

real valued attributes
----------------------
  Attribute_ID            Attribute_Name Missing      Mean    Median      Sdev      Min       Max
1            1               house_value       0 225328.06 212000.00 91971.04 50000.00 500000.00
2            2                Crime_Rate       0      3.61      0.26     8.60     0.01     88.98
3            4              num_of_rooms       0      6.27      6.00     0.73     4.00      9.00
4            5 dist_to_employment_center       0      3.80      3.21     2.11     1.13     12.13
5            6         property_tax_rate       0    408.24    330.00   168.54   187.00    711.00
6            7     student_teacher_ratio       0     18.46     19.05     2.16    12.60     22.00
7            8             Nitric_Oxides       0      0.55      0.54     0.12     0.38      0.87
8            9   accessiblity_to_highway       0      9.55      5.00     8.71     1.00     24.00
symbolic attributes
-------------------
  Attribute_ID      Attribute_Name Missing arity      MCVs_counts
1            3 Charles_river_bound       0     2  No(471) Yes(35)


~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
brief function output for house_with_missing.csv
 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
This dataset has  506  Rows   9  Attributes

real valued attributes
----------------------
  Attribute_ID            Attribute_Name Missing      Mean    Median      Sdev      Min       Max
1            1               house_value       0 225328.06 212000.00 91971.04 50000.00 500000.00
2            2                Crime_Rate       0      3.61      0.26     8.60     0.01     88.98
3            4              num_of_rooms       5      6.27      6.00     0.73     4.00      9.00
4            5 dist_to_employment_center       2      3.79      3.19     2.11     1.13     12.13
5            6         property_tax_rate       0    408.24    330.00   168.54   187.00    711.00
6            7     student_teacher_ratio       0     18.46     19.05     2.16    12.60     22.00
7            8             Nitric_Oxides       2      0.55      0.54     0.12     0.38      0.87
8            9   accessiblity_to_highway       1      9.56      5.00     8.71     1.00     24.00
symbolic attributes
-------------------
  Attribute_ID      Attribute_Name Missing arity      MCVs_counts
1            3 Charles_river_bound       5     2  No(466) Yes(35)
```

Notes:

Arity: The number of unique values for a symbolic attribute;

MCVs_counts: Most common values for an attribute and the number of records in which they appear. Please report up to three MCVs_counts per attribute, in a sorted order;

Missing: The number of missing entries for the attribute; missing values are indicated by blanks in data (when computing statistics, arity and MCV counts, missing values should be excluded from consideration).

Deliverable: The R or python source code of the function.

b) Use your new 'brief' function as well as R or python plotting tools to explore and summarize data provided in housing_no_mssing.csv. Compose a set of Power Point slides to give your customers an overview of their data set. The types of possible summary characteristics to consider for your first customer presentation include but are not limited to:

- Distribution of a single variable (histogram and/or density plot);
- Distribution of a single variable conditioned on another variable;
- Relationship among a pair of variables.

Your presentation should also include a few interesting analytical questions you would like to explore further given your current understanding of the data.

Deliverable: Professional quality Power Point presentation slides.

Problem 2

a) Write an R/python function named "do_cv" implementing k-fold cross-validation for linear regression, connect-the-dots, and default predictor (i.e. 0^{th} order polynomial regression).

The inputs of the function are:

df: data frame containing the data;

output: name of the output variable;

k: number of cross-validation folds of data;

model: name of a function that builds a regression model from training data and makes predictions for testing data (see explanations below).

Output of the function:

A k-element vector whose elements are Mean Squared Error scores for the subsequent data folds.

You need to write three R/python functions corresponding to three different models to be evaluated. The connect-the-dots function (get_pred_dots) is provided for you to use in the accompanying template file. You need to write similar functions for linear regression (get_pred_lr) and default model (get_pred_default).

An invocation of the do_cv function would look like this for 10-fold cross validation:

do_cv(df, output, 10, get_pred_default)

It should produce a result of 10-fold cross-validation of the default predictor using data from the frame named "df", assuming that the name of the output dimension in "df" is stored under variable "output".

A template of code is provided for your reference (homework1_2_template.R or homework1_2_template.py).

Deliverable: Your R or python source code.

b) For data set named house_no_missing.csv, build models using log(Crime_Rate) to estimate house_value.
   i. Applying the function you wrote in 2(a) execute leave-one-out cross-validation for the three different models;
   ii. Compute 95% confidence intervals of each score;
   Plot the results obtained in (i and ii) as a bar chart using R. The chart must be professionally formatted (have a title, clearly labeled axes, a legend, etc.) for inclusion in your presentation. Hint: you may consider using e.g. barplot2 function in gplots library; or matplotlib.pyplot.bar in python.
   iii. Present, interpret, and discuss the results in your presentation slides.

   Deliverable: Power Point presentation (combined with the result of your work on the Problem 1(b)).

Submission Guidelines

Your submission should include two files:

I. All R/python code (include the code for generating graphs) should be placed in the file named HW1_YourAndrewID.R or HW1_YourAndrewID.py respectively.
II. Professional quality Power Point file named HW1_YourAndrewID.pptx with required contents (please limit the number of slides to a maximum of 10).