`

CST8390 - Business Intelligence and Data Analytics

Lab 8 - *Regression*

Name: Min Li - Id: 040930563

Due Date: Week 11 in own lab sessions.

Introduction

The goal of this lab is to perform **linear regression** on housing file.

Steps for Linear Regression:

1. Open the housing.arff file (uploaded in **Brightspace**) in a text editor to read about the data. Fill in the following questions:

a. Number of instances: 513. b. Number of attributes: 14. c. Attribute Information:

1. CRIM per capita crime rate by town

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

3. INDUS proportion of non-retail business acres per town

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

5. NOX nitric oxides concentration (parts per 10 million)

6. RM average number of rooms per dwelling

7. AGE proportion of owner-occupied units built prior to 1940

8. DIS weighted distances to five Boston employment centres

9. RAD index of accessibility to radial highways

10. TAX full-value property-tax rate per $10,000

11. PTRATIO pupil-teacher ratio by town

12. B 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town

13. LSTAT % lower status of the population

14. MEDV Median value of owner-occupied homes in $1000's

2. Start **Weka** and open the file housing.arff. Find the following information from the preprocess tab. The **median** is the middle value of a sorted list, so **click** on the edit tab, and **sort** the columns and find the middle element:

    a) Median House Value (class) x $1000: 21.1.

    b) Median number of rooms per dwelling: 8.704.

    c) Median per capita crime rate: 0.33169.

3. Click on the Classify tab and choose "LinearRegression" from Functions. Modify the algorithm parameters so that outputAdditionalStats is "**true**". Ensure that "class" is set for what value is being computed. Run the algorithm to output the **weights** of the regression. (***Answer should be typed in. Snippet or screenshot not permitted.***)

a. What is the linear regression **model** for this set?

```
Class =

      -0.0914 * CRIM +

        0.0577 * ZN +

       -0.0931 * INDUS +

        2.8323 * CHAS=1 +

      -72.568  * NOX +

        2.5705 * RM +

       -1.2806 * DIS +

        0.2532 * RAD +

       -0.0132 * TAX +

       -0.7959 * PTRATIO +

        0.0094 * B +

       -0.6428 * LSTAT +

      65.9273
```

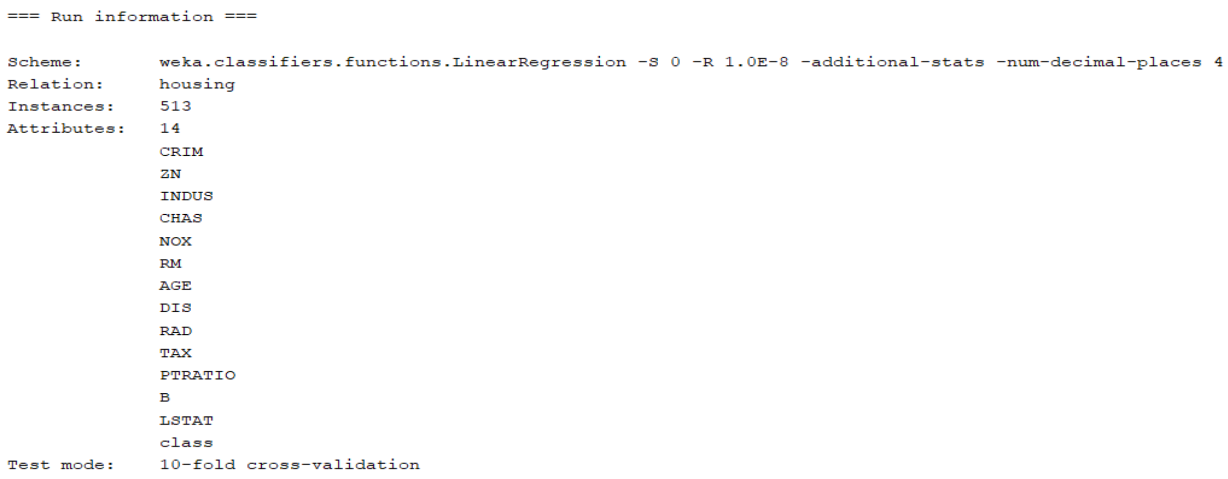
b. Which are the **two highest** factors which have a **positive influence** on the housing price?
CHAS=1, RM.

c. Which are the **two highest** factors that have a **negative influence** on housing price?
NOX, DIS.

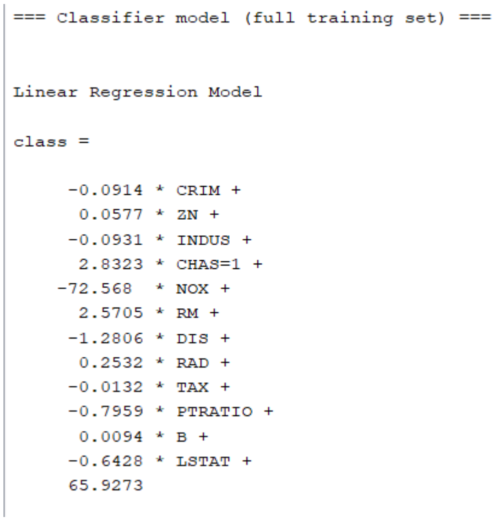
REMEMBER:

Show your **answers** to the lab professor when you are done.

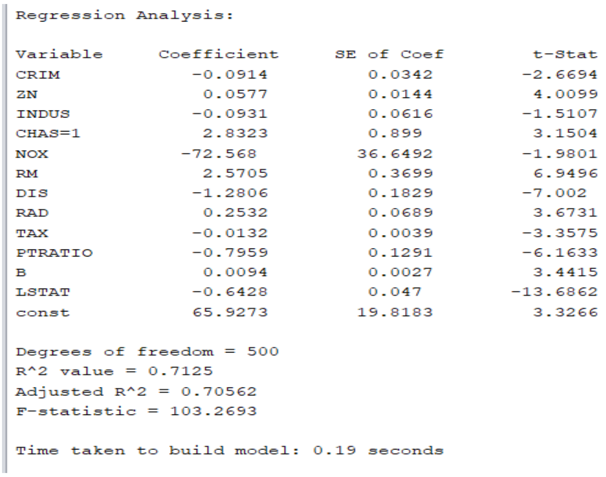
You should be ready with your results in the result pane and housing file opened in **Notepad++**.



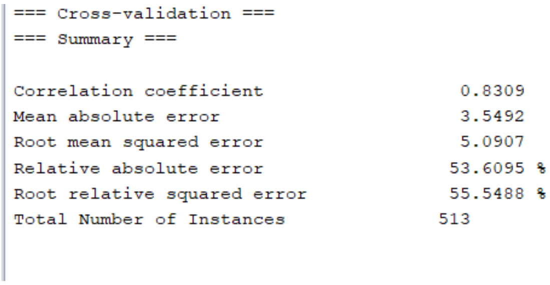
```
=== Run information ===

Scheme:       weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8 -additional-stats -num-decimal-places 4
Relation:     housing
Instances:    513
Attributes:   14
              CRIM
              ZN
              INDUS
              CHAS
              NOX
              RM
              AGE
              DIS
              RAD
              TAX
              PTRATIO
              B
              LSTAT
              class
Test mode:    10-fold cross-validation
```



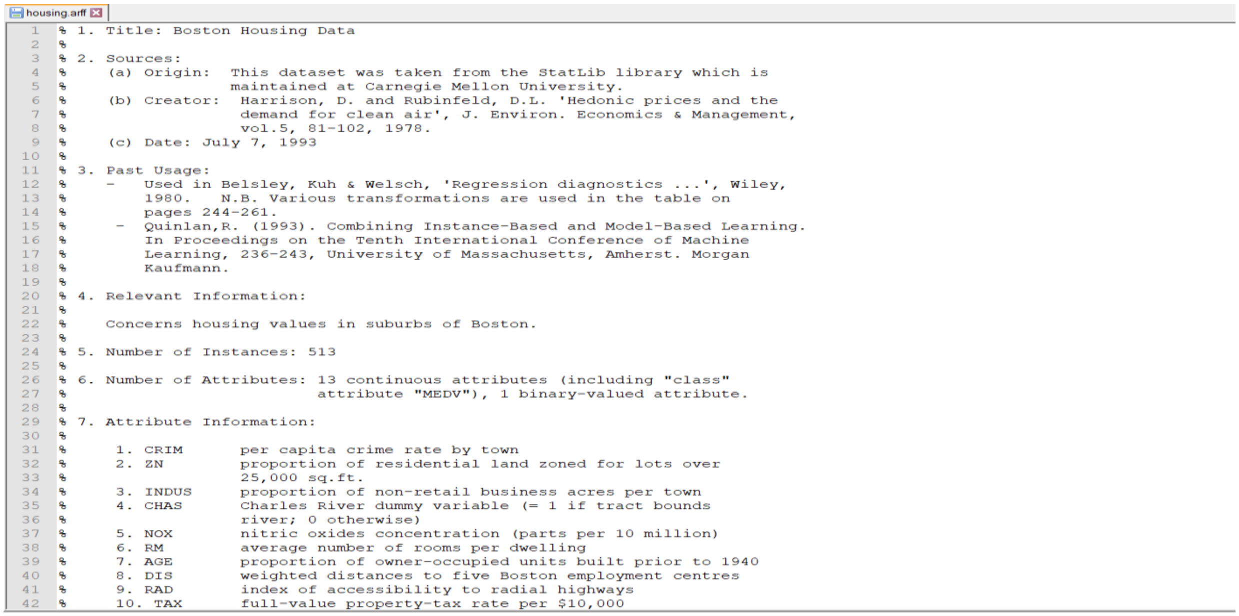
```
=== Classifier model (full training set) ===


Linear Regression Model

class =

     -0.0914 * CRIM +
      0.0577 * ZN +
     -0.0931 * INDUS +
      2.8323 * CHAS=1 +
    -72.568  * NOX +
      2.5705 * RM +
     -1.2806 * DIS +
      0.2532 * RAD +
     -0.0132 * TAX +
     -0.7959 * PTRATIO +
      0.0094 * B +
     -0.6428 * LSTAT +
     65.9273
```



```
Regression Analysis:

Variable      Coefficient     SE of Coef        t-Stat
CRIM             -0.0914         0.0342        -2.6694
ZN                0.0577         0.0144         4.0099
INDUS            -0.0931         0.0616        -1.5107
CHAS=1            2.8323         0.899          3.1504
NOX             -72.568         36.6492        -1.9801
RM                2.5705         0.3699         6.9496
DIS              -1.2806         0.1829        -7.002
RAD               0.2532         0.0689         3.6731
TAX              -0.0132         0.0039        -3.3575
PTRATIO          -0.7959         0.1291        -6.1633
B                 0.0094         0.0027         3.4415
LSTAT            -0.6428         0.047        -13.6862
const            65.9273        19.8183         3.3266

Degrees of freedom = 500
R^2 value = 0.7125
Adjusted R^2 = 0.70562
F-statistic = 103.2693

Time taken to build model: 0.19 seconds
```



```
=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.8309
Mean absolute error                      3.5492
Root mean squared error                  5.0907
Relative absolute error                 53.6095 %
Root relative squared error             55.5488 %
Total Number of Instances              513
```



```
% 1. Title: Boston Housing Data
%
% 2. Sources:
%    (a) Origin:  This dataset was taken from the StatLib library which is
%                 maintained at Carnegie Mellon University.
%    (b) Creator:  Harrison, D. and Rubinfeld, D.L. 'Hedonic prices and the
%                  demand for clean air', J. Environ. Economics & Management,
%                  vol.5, 81-102, 1978.
%    (c) Date: July 7, 1993
%
% 3. Past Usage:
%    -   Used in Belsley, Kuh & Welsch, 'Regression diagnostics ...', Wiley,
%        1980.   N.B. Various transformations are used in the table on
%        pages 244-261.
%     -  Quinlan,R. (1993). Combining Instance-Based and Model-Based Learning.
%        In Proceedings on the Tenth International Conference of Machine
%        Learning, 236-243, University of Massachusetts, Amherst. Morgan
%        Kaufmann.
%
% 4. Relevant Information:
%
%    Concerns housing values in suburbs of Boston.
%
% 5. Number of Instances: 513
%
% 6. Number of Attributes: 13 continuous attributes (including "class"
%                          attribute "MEDV"), 1 binary-valued attribute.
%
% 7. Attribute Information:
%
%     1. CRIM      per capita crime rate by town
%     2. ZN        proportion of residential land zoned for lots over
%                  25,000 sq.ft.
%     3. INDUS     proportion of non-retail business acres per town
%     4. CHAS      Charles River dummy variable (= 1 if tract bounds
%                  river; 0 otherwise)
%     5. NOX       nitric oxides concentration (parts per 10 million)
%     6. RM        average number of rooms per dwelling
%     7. AGE       proportion of owner-occupied units built prior to 1940
%     8. DIS       weighted distances to five Boston employment centres
%     9. RAD       index of accessibility to radial highways
%     10. TAX      full-value property-tax rate per $10,000
```

FOR YOUR ANALYSIS:

**** Option 1**: Explain what a **Regression** is and where to use it.*

**** Option 2**: Explain how to determine the **factors** and their **impact** (positive or negative) to the analysis.*

Option1: **Linear regression** is one of the most basic predictive process. As the name suggests, linear regression is used to find out linear relation between dependent variable and an independent variable. In a prediction, **dependent variable** means the variable which is dependent on other factors and **independent variable** refers to the mutually independent variables which effect the value of the target variables. In case of linear regression, the relation between dependent and independent variables are assumed to be linear. We can use Linear regression result to make predictions. For example, there is a **linear relationship** between miles driven and total paid for gas. Because this relationship is linear, if you spend less/more money — e.g. half vs full tank — you'll be able to drive fewer/more miles. And because that relationship is linear and you know how long is your drive from San Francisco to Las Vegas, using a **linear model** will help you **predict** how much you are going to budget for gas.

Ottawa, Mar 2020.