COMP40370 Practical 1

DATA EXPLORATION AND PREPROCESSING

Prof. Tahar KECHADI

This practical aims to get familiar with some tools and methods of data exploration and pre-processing, which were discussed in the lectures so far. Python is the chosen programming language for this module. For this practical, you need to use Python programming language with its scikit-learn, seaborn, pandas libraries, etc., to answer the questions. The required datasets are included in the practical files.

Assignment Files

• ./Practical-01.pdf	assignment questions (this file).
• ./auto-mpg.data:	data file for the questions.

Expected output files

• ./Prcatical-01.ipynb	Python notebook programs.
• ./Prcatical-01.html	Notebook in HTML showing the outputs.
• ./Practical-01-Report.pdf	Report in PDF format with answers.

Requirements

- Python 3.8+, pandas 1.3+, numpy 1.20+, sklearn 0.24+.
- tensorflow 2.0+, seaborn 0.11+, matplotlib 3.5+, scipy 1.9+.

Part A: Data Cleaning (Date: 20/09/2022)

1) The space-separated file "*auto-mpg.data*" contains fuel consumption in *mpg* with other related data of a set of cars. The original dataset, downloaded from a public domain, has been modified for the purpose of this assignment. Write a Python program to answer the following:

a. Read the data file into a pandas data frame.
b. Identify any duplicate record (s).
c. By keeping one duplicated record delete the other record (s) from the dataset.
d. What is the dimension of the data frame after removing the duplicates?

2) Write a Python program to answer the following:

a. How many missing values are in the horsepower column?
b. Remove the records having the missing values in the horsepower column.
c. Take 10% of the available records as a test set and set the horsepower to null for those records.
d. Fill in the missing values of the test set based on the mean and median of the horsepower of the training set (90%). Calculate the RMSEs for the imputed values of the test set.
e. Using the same way find the RMSEs, if scikit-learn KNNImputer (for n_neighbors 1, 3 and 5) is used with weight, acceleration, displacement and mpg features. Decide whether you need to standardise data.

f. Use the best solution to fill the missing values in the horsepower column. What are the filled values?

3) Write a Python program to answer the following:

a. What are the kurtosis and skewness values of the mpg attribute? Draw the histogram using the seaborn distplot function.
b. Identify outliers of mpg using Inter Quartile Range (IQR) approach and impute them with min and max values appropriately.
c. Transform mpg column using $\log_e$ (x+1) formula to make the mpg values follow the normal distribution.
d. Use a QQ-plot to show that $\log_e$ (x+1) is a better transformation for mpg. Find the kurtosis and skewness of mpg after the transformation.
e. Similarly detect and correct outliers in the weight, displacement, horsepower and acceleration columns.
f. Display the correlation matrix using the seaborn heatmap function between continuous variables; mpg, horsepower, weight, displacement, and acceleration.

4) Write a Python program to answer the following:

a. Identify the outliers in cylinders as a categorical variable with three main classes.
b. Correct them with kNN imputation using weight, acceleration, horsepower, displacement and mpg as features.
c. Do all cylinder 3 values assign to 4 and all cylinder 5 values assign to 6?
d. Plot a scatter diagram to visualise the relationship of mpg vs weight with the presence of number of cylinders (4, 6 and 8).

5) Write a Python program to answer the following

a. Convert *model_year* into pandas datetime format and make the data frame as an indexed time-series.
b. Resample the time-series into 3-year grouped samples and analyse mpg improvements over the groups.
c. In the origin column, encode origin 1 as Europe, 2 as USA and 3 as Japan. Using a box plot, discuss the behaviour of mpg based on origin. Do you see a trend?

6) Write a Python program to answer the following:

a. Create a new column called brand and extract the brand name from the *car_name* column.
b. Correct spelling mistakes and some short names used.
c. Group any brand less than or equal to 5 as Other.
d. What is the minimum number of brands you have in your cleaned dataset?

Part B: (Date: 27/09/2022)

The final deadline for the submission of Practical 01 (Part A and B) is Thursday, 29th of September at 23:00. Submissions should be in a single file with FirstName_LastName-P1.zip (or tar.gz) format. All submissions must be done in Brightspace.