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.
• ./auto-mpg.data:	Original data file for the questions.

Requirements

- Python 3.8+, pandas 1.3+, numpy 1.20+, sklearn 0.24+.
- 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.

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

1) Write a Python program to answer the following:

a. Transform categorical variables `cylinders' and `origin' using one-hot encoding.

b. Calculate correlation matrices for 1) one-hot encoded `cylinders' with mpg, and 2) one-hot encoded `origin' with mpg.

c. Discuss the correlation coefficient values in part b.

d. Use the label encoder technique to find the correlations between `cylinders' and mpg, and `origin' and mpg.

e. Which encoder is better (label encoder or one-hot encoder)?

2) Write a Python program to answer the following:

a. Categorize cars into three classes based on fuel efficiency (mpg): low, medium, and high. Use equal frequency (i.e. number of cars) categorization.

b. Use PCA to reduce the dimensionality of correlated features; weight, acceleration, displacement, and horsepower. (Hint: use Python library PCA from `sklearn.decomposition`)

c. Using a scatter plot, display the differences of three fuel efficiency classes with the first two principal components (PCs).

The final deadline for the submission of Practical 01 (Part A and B) is Monday, 3rd 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.