COMP40370 Practical 1

DATA PREPROCESSING

Prof. Tahar KECHADI

Academic year 2021-2022

The aim of this practical is to get familiar with some basic tools of data preprocessing and exploration, and also use some of the concepts discussed in the lectures so far. Python is the programming language to use to complete this practical. The datasets to are needed to complete the practical are described below.

Assignment Files

- ./practical1.pdf: this PDF file
- ./specs/AutoMpg question1.csv: data file
- ./specs/AutoMpg question2 a.csv: data file
- ./specs/AutoMpg question2 b.csv: data file
- ./specs/test practical1.py: Python test file

Expected output files

- ./run.py: main Python script
- ./output/question1 out.csv: data file for first question
- ./output/question2 out.csv: data file for second question

Requirements

_ Python 3.8+

_ pandas 1.3+

_ numpy 1.20+

Question 1: Data Exploration

An analyst collects surveys from different participants about their likes and dislikes. Subsequently, the analyst corrects erroneous or missing entries, uploads the data to a data warehouse, and designs a recommendation algorithm on this basis. Which of the following actions represent data collection, data preprocessing, and data analysis?

1. Conducting surveys and uploading to a database,
2. correcting missing entries,
3. designing a recommendation algorithm.

Question 2: Data Exploration

One of the important aspects of data collections is that they contain a wide variety of data types, which should be taken into account during the analysis. From the analysis point of view, we can distinguish two broad categories of data: Non-dependency-oriented data, and Dependency-oriented data.

- **Non-dependency-oriented data**: The data records do not have any specific dependencies between either the data items or the attributes. A record is referred to as data point, instance, example, transaction, entity, tuple, object, or feature- vector. Each record contains a set of fields, which are also referred to as attributes, dimensions, and features. Non-dependency-oriented data is the simplest form of data and typically refers to multidimensional data.

  We can have attributes of different types of data. These include numeric data, categorical data, binary data, text data.

- **Dependency-oriented data**: the data contains implicit or explicit relationships. In the case of implicit dependencies, the relationship is not expressed in the data. For example consecutive reading of temperature values, that are close in time, are more likely to be similar. Explicit dependencies refer to graph or network data in which edges are used to specify explicit relationships.

Different types of dependency-oriented data are: Times-Series Data, Discrete Sequences and Strings, Spatial Data, Spatio-temporal Data, and Network and Graph Data.

The same analyst obtains medical notes from a physician for data mining purposes, and then transforms them into a table containing the medicines prescribed for each patient. What is the data type of

1. The original data,
2. The transformed data?
3. What is the process of transforming the data to the new format called?

Question 3: Data Exploration

Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order):

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

1- What is the mean of the data? What is the median?
2- What is the mode of the data? Comment on the data's modality (i.e., bimodal, trimodal, etc.).
3- What is the midrange of the data?
4- Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?
5- Give the five-number summary of the data.
6- Show a boxplot of the data.

7- How is a quantile-quantile plot different from a quantile plot?

Question 4: Data Cleaning

The file AutoMpg question1.csv contains data related to cars, such as horsepower, weight, car name, and so on. Unfortunately, some of the values for the horsepower and origin columns were not properly recorded. Can you tell how many missing values are there for each one of these columns? Write the answer in your report.

1. Replace the missing horsepower values with the average of this column.
2. Replace the missing origin values with the minimum of this column
3. Save the generated data file to ./output/question1 out.csv

When saving the generated data, pay extra attention to the columns included in the file (hint: if you are using pandas, take a look at the arguments of the to_csv function).

Question 5: Data Integration

The files AutoMpg question2 a.csv and AutoMpg question2 b.csv contain similar pieces of information about car models. There are some differences between the 2 files. What you need to do is:

1. The dataset A has an attribute called car name, whereas the dataset B has an attribute called name. Rename the name attribute to car name (unintended tongue twister!).
2. The dataset B has an attribute called other, which is not present in the dataset A. Create an attribute called other in the dataset A and assign it a default value of 1.
3. Concatenate dataset A and B together, and just like in question 1, save the resulting file to ./output/question2 out.csv.