COMP40370 Practical 2

DATA PREPROCESSING

Prof. Tahar KECHADI

Academic year 2021-2022

The aim of this practical is to extend familiarity with some advanced tools of data preprocessing and exploration, and also use some of the concepts discussed in the lectures so far. Python scikit-learn library with previously used libraries is used to complete this practical. The datasets needed to complete the practical are described below.

Assignment Files

Data files

• ./practical2.pdf	assignment questions (this file)
• ./specs/Students_Results.csv	data file for Question 1
• ./specs/Sensor_Data.csv	data file for Question 2
• ./specs/DNA_Data.csv	data file for Question3
• ./test_practical2.py	Python test file to validate your solutions

Expected output and submission data

Your submission should be a single archive file (zip, tar, tgz, . . .) containing one folder called output and the following files:

• ./run.py	main Python script (solutions)
• ./report.pdf	single page PDF report
• ./output/question1 out.csv	data file for first question
• ./output/question2 out.csv	data file for second question
• ./output/question3 out.csv	data file for third question
• ./test_practical2.py	Python test file to validate your solutions

Programming requirements and tools

The assignment should be solved in Python, version 3.8 or above (3.9 is recommended). You can use the following packages for this assignment:

- pandas 1.3+
- sklearn 0.24+

We suggest you to use the **sklearn PCA** utility to reduce the dimensionality of the data. When generating the bins, you may want to take a look at the **cut** and **qcut** methods available in pandas.

Question 1: Advanced Data Exploration

A module coordinator has just completed the module assessments, and s/he would like to perform a quick analysis on the students results in various components of the module. The main objective is to see if there is any correlation between the assessment components. The students' results are given in the file "*Students_Results.csv*". Using Python script, answer the following questions:

1. Find the minimum, maximum, mean and standard deviation for each Homework column and the exam column.
2. Add an additional named as 'Homework Avg' for the average homework mark for each student. Assume that the weighting of the homework average is 25% and that of the examination is 75%, add an additional column named 'Overall Mark' for the overall folded mark.
3. Construct a correlation matrix of homework and exam variables. What can you conclude from the matrix?
4. Discuss various ways of treating the missing values in the dataset.
5. Use UCD grading system to convert the final mark into a grade (column named 'Grade'). Produce a histogram for the grades.
6. Save the newly generated dataset to "*./output/question1_out.csv*".

Question 2: Data Transformation

The *file "Sensor_Data.csv*" contains data obtained from a sensory system. Some of the attributes in the file need to be normalised, but you don't want to lose the original values.

1. Generate a new attribute called "*Original Input3*" which is a copy of the attribute "*Input3"*. Do the same with the attribute "*Input12"* and copy it into Original "*Input12"*.
2. Normalise the attribute "*Input3*" using the z-score transformation method.
3. Normalise the attribute "*Input12*" in the range [0:0; 1:0].
4. Generate a new attribute called "*Average Input*", which is the average of all the attributes from "*Input1*" to "*Input12*". This average should include the normalised attributes values but not the copies that were made of these.
5. Save the newly generated dataset to "*./output/question2_out.csv*".

Question 3: Data Reduction and Discretisation

The files "*DNA_Data.csv*" contains biological data arranged into multiple columns. We need to compress the information contained in the data.

1. Reduce the number of attributes using Principal Component Analysis (PCA), making sure at least 95% of all the variance is explained.

2. Discretise the PCA-generated attribute subset into 10 bins, using bins of equal width. For each component X that you discretise, generate a new column in the original dataset named "*pcaX_width*". For example, the first discretised principal component will correspond to a new column called "*pca1_width*".
3. Discretise PCA-generated attribute subset into 10 bins, using bins of equal frequency (they should all the same number of points). For each component X that you discretise, generate a new column in the original dataset named "*pcaX_freq*". For example, the first discretised principal component will correspond to a new column called "*pca1_freq*".
4. Save the generated dataset to "*./output/question3_out.csv*".

The final deadline for the submission is Thursday, 30th of September at 23:00. All submissions must be done in Brightspace.