COMP40370 Practical 2

Data Preprocessing

Prof. Tahar Kechadi

Academic year 2019-2020

Question 1: Data Transformation

The file `SensorData_question1.csv` contains data obtained from a sensory system. Some of the attributes in the file need to be normalised, but you do not want to loose 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/question1_out.csv`.

Question 2: Data Reduction and Discretisation

The files `DNAData_question2_a.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 contain 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_width`.

4. Save the generated dataset

Data files

- `./specs/SensorData_question1.csv`: data file
- `./specs/DNAData_question2.csv`: data file
- `./specs/test_practical2.py`: Python test file to check 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
- `./report.pdf`: single page PDF report
- `./output/question1_out.csv`: data file for first question
- `./output/question2_out.csv`: data file for second question

The final deadline for the submission is **Friday, 27th of September** at **17:00**. You can submit your solution on Brightspace.

Programming requirements and tools

The assignment should be solved in Python, version 3.5 or above (3.7 is recommended). You can use the following packages for this assignment:

- `pandas` 0.25+
- `sklearn` 0.21+

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`.