University of Wollongong
School of Computing and Information Technology
CSCI446/946 Big Data Analytics
Spring 2022

Assignment 1

(10 marks)

Due: 21:00 AEST, 26 August 2022 Friday

Aim

This assignment aims to provide students with essential experience conducting data analytics experiments with the R or the Python programming language. After completing this assignment, you should know how to

- load and save data and workspace; and

as part of data analysis:

- analyze a problem and preprocess raw dataset,
- perform clustering,
- perform classification, and
- discuss experiment results in an introductory way.

Group work: You are to work on this assignment as a group. Each group is to work independently from other groups on this assignment. Groups and group memberships are as specified on Moodle. All group members are expected to contribute to this assignment. Group members may use communications tools (e.g., UOW Zoom, UOW Webex, UOW Teams, Slack, Discord, WhatsApp, etc.) and online collaboration workspace (e.g., UOW OneDrive, Google Drive, GitHub, ZenHub, etc.) to complete the assignment. Please plan before starting the assignment, then keep a detail digital work log and timesheet for each group member. A justification and/or explanations must accompany all your answers to this assignment. One submission per group only.

Penalties: If a group member fails to make a minimum contribution, the member will be awarded zero marks. Claims of less or no contribution should provide evidence like a work log. Plagiarism of any part in this assignment will result in zero marks being awarded to the whole group.

Preliminaries

Read through the lecture slides, lab instructions and the recommended readings in Weeks 1 – 4. Conduct relevant background studies. You should use either R or Python for the tasks in this assignment. You can use any publicly accessible toolbox of library for R and Python. Your submission must include the source code file(s) which, when run, would re-create all your results.

Task 1: Problem Analysis and Data Preprocess (4 marks)

An e-commerce website NewChic.com keeps records of its products. Records of instances are restored by categories in accessories.csv, bags.csv, beauty.csv, house.csv, jewelry.csv, kids.csv, men.csv, shoes.csv, and women.csv. A product is represented by one instance (i.e., a row). The 9 CSV data sheets form the NewChic dataset. The data dictionary (.pdf) summarizes all attributes and their types in the dataset. In this assignment, you need to focus on integer and decimal type data (i.e., columns) where id in integer type is excluded. The rest columns will support your data analytics design and discussion.

The analytics into NewChic dataset aims to find

- top 10 products from your selected categories, and
- the best category among your selected categories.

For example, if you choose to analyze products in beauty, jewelry and shoes, 10 best products from these three categories are going to be reported, as well as the best category out of these three. Please use as much data (i.e., categories) as you can. Using one category will be marked zero for all tasks in this assignment.

To answer these two questions, you need to think about the following parts. A figure to illustrate your analytics plan is preferred.

1. **Design your experiment (Task 1) and report:** why would you choose all or part of data from the NewChic dataset; how would you define “top 10” and “the best”; why some columns are picked for clustering and classification algorithms and some columns are for result discussion.
2. **Program data preprocess (Task 1)** by combining CSVs in one sheet **and report**: matched, removed columns and detail explanations.
3. **Program at least two clustering algorithms (Task 2)** on preprocessed data **and report**: detail steps of each algorithm, result of all algorithms in a table, algorithm comparison and best result.
4. **Program at least two classification algorithms (Task 3)** on preprocessed data **and report**: detail steps of each algorithm, result of all algorithms in a table, algorithm comparison and best result.
5. **Discuss results (Task 4) and report:** 10 best products, the best category and your suggestions to NewChic.

Task 1 is expected to be answered in two sections in your report, under sections “Problem Analysis” and “Data Preprocess”. Please accordingly cite referred articles and programming resources in your writing. Task 1 also needs to submit the code. Add the code of data preprocess to the ZIP file for your submission if your code is saved in .R or .py.

If you use Python, Jupyter notebook is another option to submit code and report in one. Then, answers to Task 1 are a section in the Jupyter notebook.

Task 2: Clustering (2 marks)

You are to analyze the data you preprocessed in Task1. You need to perform at least two clustering algorithms and explain your selection: K-means, hierarchical clustering and more are available in reading materials. Please practice the lab - clustering first, then complete this task. Task 2 requests a report for detailed explanations of the steps of each algorithm, the result of each algorithm in a table, algorithm comparison and best result.

Task 2 is expected to be answered in the section “Clustering” in your report. Please accordingly cite referred articles and programming resources in your writing. Task 2 also needs to submit the code. Add the code of clustering algorithms to the ZIP file for your submission if your code is saved in .R or .py.

If you use Python, Jupyter notebook is another option to submit code and report in one. Then, answers to Task 2 are a section in the Jupyter notebook.

Task 3: Classification (2 marks)

You are to analyze the data you preprocessed in Task1. You need to perform at least two classification algorithms and explain your selection: KNN, Naïve Bayes and more are available in reading materials. Please practice the lab - classification first, then complete this task. Task 3 requests a report for detailed explanations of the steps of each algorithm, the result of each algorithm in a table, algorithm comparison and best result.

Task 3 is expected to be answered in the section “Classification” in your report. Please accordingly cite referred articles and programming resources in your writing. Task 2 also needs to submit the code. Add the code of classification algorithms to the ZIP file for your submission if your code is saved in .R or .py.

If you use Python, Jupyter notebook is another option to submit code and report in one. Then, answers to Task 3 are a section in the Jupyter notebook.

Task 4: Result Discussion (2 marks)

Task 4 can answer the following questions and more:

- Are the clusters well separated from each other?
- Did the classifiers well separate products from each other into different classes?
- Do any of the clusters/classes have only a few points?
- Are there meaningful and non-meaningful clusters/classes to the analytics problems questioned in Task 1?
- What are the advantages, shortages for clustering and classification algorithms in this analytics case? Which one provides results of greater value?
- Are the examined algorithms suitable for Big Data analytics? and why in your opinion?
- Will data preprocess affect clustering and classification results? and why in your opinion?
- More you can report …

Finally, please report 10 best products, the best category and your suggestions to NewChic.

Task 4 is expected to be answered in the section "Result Discussion" in your report.

Submission:

The submission link for assignment 1 is on the subject's Moodle site, under the Week 5 Toggle. Only one submission per group. **The submission must be a zip file named "`A1.zip`", under 200 MB, and contains a report (mandatory), code (mandatory) and video presentation (optional).** Either following way is acceptable:

1. a report in `.pdf` format, and code files in `.R` or `.py`; or
2. a Jupyter document in `.ipynb`, combining report and code (if your group uses Python language only).

A video presentation is optional in `.mp4` format or a shared link saved in a `.txt` file.

Important:

- The report must be in a single file and in `.pdf` or `.ipynb` format. The title page must list the full name and student ID of all members in the group. Clearly indicate members who did not make a minimum in contributions.
- The report must answer the questions in their order as given in the assignment specification. There is no page limit.
- The report must have a clear heading for each part of each task.
- Sufficient description, explanation, justification, and discussion are essential parts of your answers. Marks will be deducted for incomplete or vague answers.
- Sufficient, suitable, and legible annotation shall be provided in your code to make it easy to understand. Marks will be deducted for untidy code, code that is difficult to read, code that does not run, or code that does not reproduce the results in your report.

Note: Failure of your code to run may attract zero marks. Plagiarism of any part in your code, or any part in your report will attract zero marks for this assignment. It is the responsibility of the group to ensure that your submission does not contain plagiarized material. You may be requested to demonstrate and explain your program or explain your answer in the report. Marks are deducted if you are unable to offer an explanation. Marks will be awarded for correct design, implementation, style, completeness, and justification.

-- *END*--