SIT720 Machine Learning
Assessment Task 2: Problem solving task.

This document supplies detailed information on Assessment Task 2 for this unit.

Key information

- Due: **Week 7, Monday 29 August 2022** by 8.00 pm (AEST),
    - Weighting: 25%

Learning Outcomes

This assessment assesses the following Unit Learning Outcomes (ULO) and related Graduate Learning Outcomes (GLO):

Unit Learning Outcome (ULO)	Graduate Learning Outcome (GLO)
ULO2 - Perform unsupervised learning of data such as clustering and dimensionality reduction.	**GLO1 -** through the assessment of student ability to use data acquisition techniques to obtain, manipulate and represent data. **GLO3 -** through student ability to use specific programming language and modules to obtain, pre-process, transform and analyse data. **GLO4** -through assessment of student ability to make decisions to obtain data, use appropriate techniques to represent and visualise complex relationships in the data. **GLO5 -** through assessment of student ability to solve problems relates to ill-defined data.

Purpose

This assessment task is for student to apply skills for data clustering and dimensionality reduction. Students will be required to demonstrate ability in data representation, and competency in applying suitable clustering/dimensionality reduction techniques in a real-world scenario.

Assessment 2 **Total marks = 30**

Submission Instructions

a) Submit your solution codes into a **notebook file with ".ipynb"** extension. Write discussions and explanations including outputs and figures into a separate file and **submit as a PDF file**.
b) Submission other than the above-mentioned file formats will not be assessed and given **zero** for the entire submission.
c) Insert your Python code responses into the cell of your submitted ".ipynb" file **followed by the question** i.e., copy the question by adding a cell before the solution cell. If you need multiple cells for better presentation of the code, add question only before the first solution cell.
d) Your submitted code should be executable. If your **code does not generate** the submitted solution, then you will **get zero** for that part of the marks.
e) Answers must be **relevant and precise**.
f) No **hard coding** is allowed. Avoid using specific value that can be calculated from the dataprovided.
g) Use **topics covered till week 6** for answering this assignment.
h) Submit your assignment **after running each cell individually**.
i) The submitted notebook **file name** should be of this form "SIT720_A2_studentID.ipynb". For example, if your student ID is 1234, then the submitted file name should be "SIT720_A2_1234.ipynb".

Background

In end user perspective, travel and tourism is mostly explorative in nature and repetitive travels to same locations are minimal. So, travellers have to take decisions regarding their destinations and associated facilities to be consumed without adequate prior or personal knowledge. The best option available is to leverage social media and internet. Tourism recommenders are the best solutions in this scenario.

Dataset

Dataset file name: tripadvisor_review.csv

Dataset description: User's average feedback/rating information on 10 categories of attractions in East Asia captured from tripadvisor.com. Dataset contains 980 user records with 10 feedback attributes inferred from numerous destination reviews.

Questions

---

1. In this dataset (tripadvisor_review.csv), we have traveller's average feedback/rating information on 10 different categories of attraction. We are interested in finding optimal number of traveller groups based on their attraction ratings.
    a. What method shall we use for solving this problem and why? (**1 mark**)
    b. Does this data suffer from curse of dimensionality? Explain. (**1 mark**)
    c. Find out optimal number of traveller groups, report the outcome and justify your findings. **(2 marks)**

2. Implement two alternative solutions of Q1 (c). Compare and report the findings. **(2+1=3 marks)**

3. Evaluate quality of the groupings that you have reported as a solution of Q1 (c) and Q2. Based on the evaluation outcomes, report the best solution and explain the results. **(3+2=5 marks)**

4. Quantify and print the relationship among independent variables of this dataset (tripadvisor_review.csv). Calculate two collective variables that represent the same dataset. Create a two-dimensional plot to display the relationship between these new variables and explain the plot. **(1+2+2=5 marks)**

5. Is there any loss of information due to the transformation performed in Q4? Explain your answer with evidence. **(3 marks)**

6. Principal component analysis applied on a given dataset, and the percentage of variance for the first N components is X%. How is this percentage of variance computed? **(2 marks)**

Following questions are D & HD level tasks. You have to do your own research explore current literatures and solutions for answering this question.

7. Apply component factor- and projection-based dimensionality reduction approaches on the given dataset (tripadvisor_review.csv) for creating three collective variables. Does this new feature space improve the grouping of travellers compared to original dataset? Present your results with appropriate evidences. **(3 marks)**

8. Let's consider the data shown in the Figure 1 (see next page).
    a) Is it possible to obtain the cluster shown in the figure by k-means clustering (k = 6)? Provide evidence including code and explanation to justify your findings. **(2 marks)**

b) Explore state-of-the-art clustering methods (explore recent research articles) that can produce better results than k-means for this problem? Describe the selected approach, evaluate performance and report your findings. **(3 marks)**

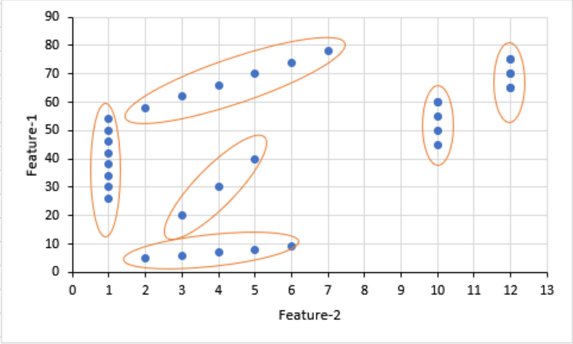

Figure 1: Scatter plot with expected clusters (elliptical shapes)

Submission details

Deakin University has a strict standard on plagiarism as a part of Academic Integrity. To avoid any issues with plagiarism, students are strongly encouraged to run the similarity check with the Turnitin system, which is available through Unistart. A Similarity score MUST NOT exceed 39% in any case. Late submission penalty is 5% per each 24 hours from- Week 7, Monday 29 August 2022 by 8.00 pm (AEST), No marking on any submission after 5 days (24 hours X 5 days from- Week 7, Monday 29 August 2022 by 8.00 pm (AEST)).

Extension requests

Requests for extensions should be made to Unit/Campus Chairs well in advance of the assessment due date. If you wish to seek an extension for an assignment, you will need to submit a request using the “Extension Request” link of the “Assessment” menu in the unit site, as soon as you become aware that you will have difficulty in meeting the scheduled deadline, but at least 3 days before the due date. When you make your request, you must include appropriate documentation (medical certificate, death notice) and a copy of your draft assignment. Conditions under which an extension will normally be approved include:

Medical To cover medical conditions of a serious nature, e.g. hospitalisation, serious injury or chronic illness. Note: Temporary minor ailments such as headaches, colds and minor gastric upsets are not serious medical conditions and are unlikely to be accepted. However, serious cases of these may be considered.

Compassionate e.g. death of close family member, significant family and relationship problems.

Hardship/Trauma e.g. sudden loss or gain of employment, severe disruption to domestic arrangements, victim of crime. Note: Misreading the timetable, exam anxiety or returning home will not be accepted as grounds for consideration.

Special consideration

You may be eligible for special consideration if circumstances beyond your control prevent you from undertaking or completing an assessment task at the scheduled time. See the following link for advice on the application process: http://www.deakin.edu.au/students/studying/assessment-and-results/special-consideration.

Assessment feedback

The results with comments will be released within 15 business days from the due date.

Referencing

You must correctly use the Harvard method in this assessment. See the Deakin referencing guide.

Academic integrity, plagiarism, and collusion

Plagiarism and collusion constitute extremely serious breaches of academic integrity. They are forms of cheating, and severe penalties are associated with them, including cancellation of marks for a specific assignment, for a specific unit or even exclusion from the course. If you are ever in doubt about how to properly use and cite a source of information refer to the referencing site above.

Plagiarism occurs when a student passes off as the student's own work, or copies without acknowledgement as to its authorship, the work of any other person or resubmits their own work from a previous assessment task.

Collusion occurs when a student obtains the agreement of another person for a fraudulent purpose, with the intent of obtaining an advantage in submitting an assignment or other work.

Work submitted may be reproduced and/or communicated by the university for the purpose of assuring academic integrity of submissions: https://www.deakin.edu.au/students/study-support/referencing/academic-integrity.