Assessment Information

SIT720: Machine Learning

Assessment 1: Individual Problem solving task

This document supplies detailed information on assessment tasks for this unit.

Key information

- Due: 22 August by 11.30pm AEST
- Weighting: 25%
- Word count: max 20 pages including all relevant material, graphs, images and tables

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)
ULO 1: Apply suitable clustering/dimensionality reduction techniques to perform unsupervised learning of data in a real-world	**GLO 1:** Discipline knowledge and capabilities **GLO 3:** Digital literacy **GLO 4:** Critical thinking **GLO 5:** Problem solving

Purpose

In this assignment, you need to demonstrate your skills for data clustering and dimensionality reduction. There are two parts of this assignment

Instructions

This is an individual assessment task of maximum 20 pages including all relevant material, graphs, images and tables. Students will be required to provide responses for series of problem situations related to their analysis techniques. They are also required to provide evidence through articulation of the scenario, application of programming skills, analysis techniques and provide a rationale for their response

Task A - Clustering

Download BBC sports dataset from the Cloud. This dataset consists of 737 documents from the BBC Sport website corresponding to sports news articles in five topical areas from 2004-2005. There are 5 class labels: athletics, cricket, football, rugby, tennis. The original dataset and raw text files can be downloaded from here

1. There are 3 files in the dataset corresponding to the feature matrix, the class labels and the term dictionary. You need to read these files in Python notebook and store in variables X, trueLabels, and terms.

   (5 marks)

2. Next perform K-means clustering with 5 clusters using Euclidean distance as similarity measure. Evaluate the clustering performance using adjusted rand index and adjusted mutual information. Report the clustering performance averaged over 50 random initializations of K-means

   (5 marks)

3. Repeat K-means clustering with 5 clusters using a similarity measure other than Euclidean distance. Evaluate the clustering performance over 50 random initializations of K-means using adjusted rand index and adjusted mutual information. Report the clustering performance and compare it with the results obtained in step 2

   (5 marks)

4. For clustering cases (Euclidean distance and the other similarity measure), visualize the cluster centres using Tag cloud using Python package WordCloud.

   (5 marks)

Assessment Information

Task B - (Dimensionality Reduction using PCA/SVD

For the provided BBC sports dataset, perform PCA and plot the captured variance with respect to increasing latent dimensionality. What is the minimum dimension that captures (a) at least 95% variance and (b) at least 98% variance?

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 11.30pm, 22nd of August. No marking on any submission after 5 days (24 hours X 5 days from 4pm 10th of August)
Be sure to downsize the photos in your report before your submission in order to have your file uploaded in time.

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 apply by email directly to Chandan Karmakar (karmakar@deakin.edu.au), 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

SIT720 Machine Learning
Assessment Task 1: Individual problem-solving rubric

Criteria	Excellent	Good	Fair	Unsatisfactory
Criteria 1: Reading files corresponding to the feature matrix, class labels and the term dictionary and store them in variables X, true Labels and terms using Python notebook.	**5 marks** Successfully read all files and stored in corresponding variables using Python notebook.	**3 marks** Partially achieved the goal by missing reading or storing one file or variable.	**2 marks** Only able to either reading files or creating variables in Python to store any value.	**0 mark** Fail to read and store using Python notebook.
Criteria 2: * Perform K-means clustering with 5 clusters using Euclidean distance as similarity measure. * Evaluate the clustering performance using adjusted rand index and adjusted mutual information. * Report the clustering performance averaged over 50 random initializations of K-means.	**5 marks** Successfully completed all three tasks.	**3 marks** Successfully completed any two of the three tasks.	**2 marks** Successfully completed only one of the three tasks.	**0 mark** Failed to complete any given task.
Criteria 3: * Repeat K-means clustering with 5 clusters using a similarity measure other than Euclidean distance. * Evaluate the clustering performance over 50 random initializations of K-means using adjusted rand index and adjusted mutual information. * Report the clustering performance and compare it with the results obtained in step 2.	**5 marks** Successfully completed all three tasks.	**3 marks** Successfully completed any two of the three tasks.	**2 marks** Successfully completed any one of the three tasks.	**0 mark** Failed to complete any given task.
For clustering cases (Euclidean distance and the other similarity measure reported in previous two tasks), visualise the cluster centres using Tag cloud using Python package WordCloud	**5 marks** Successfully used the WordCloud Package to visualise the cluster centres using at least two different similarity measures.	**3 marks** Successfully used the WordCloud Package to visualise the cluster centres using at least one similarity measure.	**2 marks** Demonstrated knowledge in WordCloud Package and visualisation, but cannot use them successfully.	**0 mark** Failed to show any evidence of knowledge in WordCloud Package and visualisation.
PART 2	**Excellent**	**Good**	**Fair**	**Unsatisfactory**
For the provided BBC sports dataset: * Perform PCA * Plot the captured variance with respect to increasing latent dimensionality. * What is the minimum dimension that captures (a) at least 95% variance and (b) at least 98% variance?	**5 marks** Successfully completed all three tasks.	**3 marks** Successfully completed any two of the three tasks.	**2 marks** Successfully completed any one of the three tasks.	**0 mark** Failed to complete any given task.