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

Assessment 1: Problem solving task

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

Key information

- Due: Wednesday 21 August 2019 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: Perform unsupervised learning of data such as clustering and dimensionality reduction.	**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 Python programming skills, analysis techniques and provide a rationale for their response.

Part-1 Clustering:

Download the digit dataset from the unit site. This dataset contains 8x8 pixel images of digits 0-9.

Instructions: there are five different files where each file contains a different number and types of digit images. The file name ends with a digit between 0 to 4. Please compute the modulus operation (fID=SID % 5), where SID is your own student ID number. Now select the data file, name of which ends with the same fID value. For example, if your student id is 218201419, then you should compute fID=218201419%5. This result is fID=4 so in this case you should work with the file named "digitData4.csv'. If the result was fID=2 you must work with the file named “digitData2.csv”.

1- Read the downloaded file into a matrix M(mXn). Create an empty numpy array X with m rows and n-1 columns. Assign all m rows and first n-1 columns of M into X. Create a numpy vector trueLabels and assign n-th column of M into that. Print dimensions of M, X and trueLabels. **(1+1+1+1+1=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 (ARI) and adjusted mutual information. Report the clustering performance averaged over 50 random initializations of K-means. **(1+1+3=5 marks)**

3- If we have an ARI value of 0.7 after a single run of K-means clustering with 'Kmeans++' initializaton for any data set then what will be the value of averaged ARI over 20 repeatations. Explain why? **(1+1=2 marks)**

4- Repeat K-means clustering with 5 clusters using a similarity measure other than **Euclidean distance** (you are free to use other libraries). 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. **(2+1+2=5 marks)**

SIT720 Machine Learning
Assessment Task 1: Individual problem-solving task
Part-2 Dimensionality Reduction using PCA/SVD:

For the provided digits dataset:

1- Perform PCA. Plot the captured variance with respect to increasing latent dimensionality. What is the minimum dimension that captures at least 95% variance? **(1+2+2=5 marks)**

2- Create a scatter plot with each of the total rows of X projected onto the first two principal components. In other words, the horizontal axis should be v1, the vertical axis v2, and each individual should be projected onto the subspace spanned by v1 and v2. Your plot must use a different color for each digit and include a legend. **(2+1=3 marks)**

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, 21st of August. No marking on any submission after 5 days (24 hours X 5 days from 11.30pm 21st 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

Part 1	**Excellent**	**Good**	**Fair**	**Unsatisfactory**
***Read the downloaded file into a matrix M(mXn).** ***Create an empty numpy array X with m rows and n-1 columns.** ***Assign all m rows and first n-1 columns of M into X.** ***Create a numpy vector trueLabels and assign n-th column of M into that.** *** Print dimensions of M, X and trueLabels.**	Successfully completed all five tasks.	Successfully completed any three of five tasks.	Successfully completed any two of five tasks.	Fail to complete any given task.
*** Perform K-means clustering with 5 clusters using Euclidean distance as similarity measure.** *** Evaluate the clustering performance using adjusted rand index (ARI) and adjusted mutual information.** *** Report the clustering performance averaged over 50 random initializations of K-means.**	Successfully completed all three tasks.	Successfully completed any two of the three tasks.	Successfully completed only one of the three tasks.	Failed to complete any given task.
*** If we have an ARI value of 0.7 after a single run of K-means clustering with 'Kmeans++' initializaton for any data set then what will be the value of averaged ARI over 20 repetitions *** *** Explain why?**	Successfully answered both of them with appropriate reasoning.	Answers are correct but reasoning is not appropriate.	Answer is correct for only the first part and explanation is missing.	Failed to complete any given task.
*** 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.**	Successfully completed all three tasks.	Successfully completed any two of the three tasks.	Successfully completed any one of the three tasks.	Failed to complete any given task.

Part 2	**Excellent**	**Good**	**Fair**	**Unsatisfactory**
For the provided digits dataset: *** Perform PCA** *** Plot the captured variance with respect to increasing latent dimensionality.** *** What is the minimum dimension that captures at least 95% variance?**	Successfully completed all three tasks.	Successfully completed any two of the three tasks.	Successfully completed any one of the three tasks.	Failed to complete any given task.
*** Create a scatter plot with each of the total rows of X projected onto the first two principal components. In other words, the horizontal axis should be v1, the vertical axis v2, and each individual should be projected onto the subspace spanned by v1 and v2.** *** Your plot must use a different color for each digit and include a legend.**	**3 marks** Successfully completed both tasks.	**2 marks** Successfully completed the first task.		**0 mark** Failed to complete any given task.