Xi'an Jiaotong-Liverpool University
西交利物浦大學

INT104 – Artificial Intelligence

Coursework – Unsupervised Learning Exercise

Introduction

In this coursework, a spreadsheet has been provided to perform a set of data analysis. The spreadsheet contains the following information: the index of student, gender of student, the programme that a student is enrolled, the grade that the student is in, total marks that a student is awarded and the mark of 5 exam questions (indexed as Q1, Q2, Q3, Q4 and Q5).

The first column is the ID of the student. The gender of the student is represented as “1” and “2”. The grade of the student is either “2” or “3”. The programme of the student is represented as “1”, “2”, “3” and “4”. The full mark for 5 exam questions are 8 marks (Q1), 8 marks (Q2), 14 marks (Q3), 10 marks (Q4) and 6 marks (Q5) respectively.

The coursework requires students to cluster the samples in the dataset with the best dedicated metric.

Tasks

1. With at least three sets of features, divide the provided dataset into four clusters. For each set of features, there should be at least three clustering process applied. The code of the experiment should be uploaded on Learning Mall. (60 Marks)
2. A lab session for live demonstration (Week 12) will be organised. Over the session, the students are given a new set of data. The student should adapt the Python script to the new dataset within the 4 hours. The time of adaptation will be recorded by your TA in a timely manner whereas the metric of performance will be recorded by the end of lab session. (40 Marks)

Marking Criteria

Task 1:

Please upload your source code to Learning Mall as a Python script. The implementation of the following functions counts towards full marks of the task:

- Apply transforms to raw data and obtain three sets of features. (1 marks each transform, 1 merk for each feature being applied to a clustering method, in total 12 marks)
- Implementation of three clustering models (4 marks each, 12 marks in total): GMM, hierarchical clustering (HC) and k-means (KM).
- An attempt to try different settings for each single clustering method (3 marks for each setting, 9 cases for 3 clustering methods lead to 27 marks in total)
- A table showing the best performed clustering results for each set of features and each clustering method (possibly with the best configuration) (9 marks)

Task 2:

You will be given a new set of data and you need to adopt your Python script with the new data. The lab session will last for 4 hours. During the lab session, you should call your TA to record the time that you have successfully adapted your Python script to the new dataset for the first time as marks are awarded accordingly. You then may keep on tuning the configuration of Python script. Once you are satisfied with your output, you should call your TA again to record the best metric you have obtained. The submission must be made before the end of lab session. Once you have uploaded your output file, you may leave the lab.

Time	Marks	GMM	HC	KM	Marks
1 hour	16 marks				8 marks
1 hour 30 mins	15 marks				7 marks
2 hours	14 marks				6 marks
2 hours 15 mins	13 marks				5 marks
2 hours 30 mins	12 marks				4 marks
2 hours 45 mins	11 marks				3 marks
3 hours	10 marks	Metric Submitted			2 marks
3 hours 10 mins	9 marks	Metric Not Submitted			0 marks
3 hours 20 mins	8 marks	Metric: ratio between intra-cluster distance and inter-cluster distance Specific marking for metrics will be released before the start of live demonstration due to different datasets.			
3 hours 30 mins	7 marks				
3 hours 40 mins	6 marks				
3 hours 50 mins	5 marks				
4 hours	4 marks				

Submission

1. Task 1 should be submitted in a format of Python script (*.py or *.ipynb) or a package of files (*.zip).
2. The timing and metric information for task 2 will be directly documented by a TA.
3. Please name your submission file as ID_FirstName_LastName_C2.zip or ID_FirstName_LastName_C2.py etc.
4. Late submission policy of XJTLU applies.