

INT104 – Artificial Intelligence

Coursework – Supervised 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 classify the samples to the group that the student belonged to.

Tasks

1. Use at least three sets of features to train a classifier for programme with the provided dataset. 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. (50 Marks)
2. A lab session for live demonstration (Week 13) 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. (50 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:

- Implementation of three classification models (2 marks each, 6 marks in total): decision tree (DT), naïve Bayes (NB) and kNN.
- An experiment that compares the parameters in all classification methods. (3 marks for each parameter, students should try 2 parameters for each classification method, which leads to 18 marks in total).
- Present a table that compare the best performance of the basic classification systems (DT, NB and kNN). (2 marks)
- Build a system that uses ensemble learning (3 marks) and compare different ensemble strategies (5 marks). (8 marks in total)
- All classification methods and the ensemble learning must apply cross validation. (4 marks each system, 16 marks in total).

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 to pursue a higher accuracy. 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.

The marking scheme regarding to accuracy will be released before live demonstration due to different datasets. The marks are awarded according to the following table:

Time	DT	kNN	NB	Ensamble	Marks
1 hour					10 marks
1 hour 30 mins					9 marks
2 hours					8 marks
2 hours 20 mins					7 marks
2 hours 40 mins					6 marks
3 hours					5 marks
3 hours 15 mins					4 marks
3 hours 30 mins					3 marks
Submitted					2 marks
No Submission					0 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_C3.zip or ID_FirstName_LastName_C3.py etc.
4. Late submission policy of XJTLU applies.