Durham University

Computer Science

Summative Assignment

Module code and title	COMP4157 Learning Analytics
Academic year	2023-24
Coursework title	Learning Analytics coursework
Coursework credits	10 credits
% of module's final mark	100%
Lecturer	Stamos Katsigiannis, Nelly Bencomo
Submission date*	Thursday, January 25, 2024 14:00
Estimated hours of work	20 hours
Submission method	Ultra

Additional coursework files	*dataset.csv, "LMS Case Study-POMDP.zip", springer_template.docx*
Required submission items and formats	***Part 1:*** *Zip file named "[CIS_USERNAME_CourseWorkPart1].zip" containing the Word/PDF document for the individual.* ***Part 2:*** *A Jupyter Notebook with the required Python code, named "[CIS_USERNAME].ipynb" (e.g. "abcd12.ipynb")*

* This is the deadline for all submissions except where an approved extension is in place.

Late submissions received within 5 working days of the deadline will be capped at 40%.
Late submissions received later than 5 days after the deadline will receive a mark of 0.
It is your responsibility to check that your submission has uploaded successfully and obtain a submission receipt.

Your work must be done by yourself (or your group, if there is an assigned groupwork component) and comply with the university rules about plagiarism and collusion. Students suspected of plagiarism, either of published or unpublished sources, including the work of other students, or of collusion will be dealt with according to University guidelines (https://www.dur.ac.uk/learningandteaching.handbook/6/2/4/).

Coursework Part 1 (Dr Nelly Bencomo) **[50 marks]**

Requirements

Students are expected to work on the coursework **individually**.

This assignment requires you to study and further understand the use of Markov Decision Processes (MDP) and, specifically, Partially Observable Markov Decision Processes (POMDP) as techniques used for Descriptive and Predictive Analysis in Learning Analytics (LA) and Learning Management Systems (LMS). The concepts related to POMDP and the connections with LA were covered by Week 4.

Download the file "LMS Case Study-POMDP.zip" which contains the implementation of a pre-designed POMDP model (*SolvePOMDP-master*) applied to a given LMS Scenario. The file *README*.txt has the instructions for configuring of the *SolvePOMDP-master* java project. The initial setup of the POMDP model for the LMS case study is provided in "*LMS-Prob.xlsx*". The scenario is a simplification of the scenario presented in [1]. You will receive by email the two files *Question2_TestRun.txt* and *Question4_TestRun.txt*. The files contain the outputs of 2 test runs, which are assigned to you only. You will be required to:

1. Read the paper [1] to further understand the context of POMDPs in Learning Analytics and the Scenario given as an example of the application discussed in the paper.
    a. Understand the POMDP that has been provided, which was tailored to the simplified scenario. You should execute the POMDP implementation to run tests (based on the README file provided). Use LMS.POMDP domain for execution. Similar examples have been explained during the lectures of Week 4 [you may want to revisit the lecture slides and recordings].
    b. Understand the format of the configuration file for the POMDP model (LMS-Prob). Answer these questions:
        I. How do you interpret the behavioural probabilities of the model provided, specifically the transition probabilities: P(PS=T | AC = *T*, PS = F, LA = *PD*)) and P(PS=T | AC = *T*, PS = T, LA = *LS*)). [**2 Marks]**
        II. Would you provide other probability values for these situations? Provide 2 different values and explain your choices. [**4 Marks**]
2. Consider the output of the test run (Question2_TestRun.txt) that has been provided. For the test, answer the following:
    a. What do you conclude about the learning progress of the student concerning the level of acquisition of the skills Problem Solving (PS) and Assertive Communication (AC)? [**10 Marks**]
    b. What do you conclude about the learning activities prompted by the POMDP as the activities (interventions) to be followed by the student? [**10 Marks**]

3. Read Chapter 2 "The Elicitation Context", in the book [2]. Write a critical analysis of the elicitation process in general terms (briefly), highlighting the differing perspectives of statistical and psychological research in eliciting expert knowledge and the link to Learning Analytics and Learning Management Systems (limit 1.5 pages). [**8 Marks**]
4. Consider a scenario where the initial transition probabilities provided for the model are different from the initial set-up used in question 2 shown above (and, perhaps, far from ideal). These probabilities are given in "*Transition_probabilities_ver2"* in "*LMS-Prob.xlsx*". Consider the "*Question4_TestRun.txt*" file provided that was generated as a result of using these probabilities. You are required to critically analyse the

behaviour by answering the following questions: What went wrong with the learning progress of the student? What was the impact on the satisfaction of the skills? Suggest how it can be corrected. [**10 Marks]**

5. How is the correct elicitation of probabilities important for POMDP-based decision-making, and what would be the impact of incorrect initial probabilities? Feel free to provide brief examples to support the explanation [**6 Marks]**

References

[1] "*Decision-Making Support for Adaptive Learning Management Systems based on Bayesian Inference*", Bencomo N. Samin H., Pavlich-Mariscal J, Workshop Causal Inference in Conference in Educational Data Mining and Conference on Artificial Intelligence in Education and (EDM & AIED 2022), Durham, UK, July 2022. (the pdf file can be found on Blackboard- reading section)

[2] "*Uncertain Judgements: Eliciting Experts' Probabilities*", Anthony O'Hagan, Caitlin E. Buck, Alireza Daneshkhah, J. Richard Eiser, Paul H. Garthwaite, David J. Jenkinson, Jeremy E. Oakley, Tim Rakow, ISBN: 978-0-470-02999-2 July 2006

Written Report

Use the Springer Computer Science Proceedings template for your report. The template for MS Word or LaTeX can be found here: https://www.springer.com/gp/computer-science/lncs/conference-proceedings-guidelines or in the "springer_template.docx" file.
The report should be divided according to the questions shown above.

(Continue to next page for Part 2 of the coursework)

Coursework Part 2 (Dr Stamos Katsigiannis) **[50 marks]**

In this assignment, you are going to use an anonymised student performance dataset which contains data obtained via a survey of a students' math course in school. The aim is to use the dataset in order to train machine learning models for predicting the performance of students in this course. The dataset consists of one Comma Separated Values (CSV) file, named "dataset.csv". The structure of the dataset is as follows:

	Column	Description
Attributes	sex	Student's sex
	age	Student's age
	address	Student's address (U: Urban, R: Rural)
	family_size	Size of family (GT3: >3, LE3: ≤3)
	parents_cohabitation	Parents' cohabitation status
	mothers_education	Mother's education level
	fathers_education	Father's education level
	mothers_job	Job sector of the mother
	fathers_job	Job sector of the father
	guardian	Who is the guardian of the student
	traveltime	Time allocated to travelling to/from school
	studytime	Time allocated to studying
	failures	Prior failures to pass the course
	activities	Engagement with out-of-school activities
	internet	Access to the internet
	romantic_relationship	Currently involved in a romantic relationship
	family_relations	Rating of relations within the family
	freetime	Level of free time
	going_out	Frequency of going out
	alcohol_workday	Alcohol consumption during working days
	alcohol_weekend	Alcohol consumption during the weekend
	health	Student's health
	absences	Total number of absences from the course
Responses	grade_term1	Course grade for the 1st term
	grade_term2	Course grade for the 2nd term
	grade_final	Final course grade

You have to use the above learning analytics dataset for the following tasks:

1. **Data curation and preparation**
   a. Pre-process the dataset to ensure that it is consistent, and free of noise and artefacts. **[9 marks]**
   b. Examine the attributes in the dataset and encode them to a suitable numerical representation if needed. **[6 marks]**

2. **Conduct exploratory data analysis**. Examine the dataset. Explore the distribution and other statistics of the values of the attributes. Do you notice any relationships between any attributes? You may compute histograms, correlation matrices, etc. Use visualisations to better demonstrate your findings. **[15 marks]**

3. **Train three machine learning models** to predict the students' **performance** based on all or a few of the available attributes. The trained models should be able to predict the "grade_term1", "grade_term2", and "grade_final" respectively. Report the regression performance of each model in terms of the Mean Squared Error (MSE), Mean Absolut Error (MAE), and R^2 metrics, as well as by a scatter plot depicting the real values vs. the predicted values. Alternatively, convert the problem into a multi-class classification problem (n_{class}>2) following an appropriate strategy and report the classification performance in terms of accuracy, precision, recall, F1-score, and the confusion matrix. Make sure that proper cross-validation is used to evaluate your models. **[15 marks]**

4. **Written discussion (1,500 words max)**.
    a. Provide justification for the data preparation steps and the attribute encoding approaches that you applied. **[5 marks]**
    b. Critically discuss the performance of the trained models. **[5 marks]**
    c. Critically discuss the suitability of the provided dataset for the task at hand. You may refer to the results of the exploratory data analysis and the performance of the trained models to justify your arguments. **[5 marks]**

For Tasks 1-3 you should use the Python 3.x programming language and prepare and submit a single Jupyter Notebook with your solutions. Make sure that the code is properly commented and that the solution for each task is easily distinguishable. The answers to Task 4 should be included at the end of the Jupyter Notebook using markdown cells.

Notes:

- You are strongly advised to use the Pandas Python library for loading and manipulating the data.
- The submitted Jupyter Notebook should be named "[CIS_USERNAME].ipynb" (e.g. "abcd12.ipynb")

General Instructions

Students are expected to work on the coursework **individually**.

Marking Criteria

The following marking scheme is followed for assessing coursework:

90-100: Perfect
80-89: Outstanding
70-79: Excellent
65-69: Very Good
60-64: Good
55-59: Very Satisfactory
50-54: Satisfactory
40-49: Adequate
0-40: Fail

Further information on each criterion can be found here.

Examiners' expectations

What the examiners expect from the report in **Part 1** and the written discussion in **Part 2**:

- Your report needs to be professional. The language should be scientific.
- Your report should provide justification for the answers you propose (when it applies)

What the examiners expect from your code implementation in **Part 2**:

- **Your program must be runnable – a program that partially works or does not run at all will receive no mark.**
- You are asked to use Python 3.x.
- Your source code should be documented with comments, making it to be as easily followed as possible.

Page/Word Limit policy

Examiners will stop reading once the page/word limit has been reached, and work beyond this point will not be assessed.

Formatting

You can use formatting (e.g. bold font, underlying font, bullets) to highlight answers to the various questions. It can help you to convey your answers and the examiner to digest and understand your work.

Plagiarism and collusion

Your assignment will be put through the plagiarism detection service and the submitted Python code will be checked using a programming plagiarism detection tool.

Students suspected of plagiarism, either of published work or work from unpublished sources, including the work of other students, or of collusion will be dealt with according to the Computer Science Department and University guidelines.