Autumn Examinations 2019/2020

Exam Code(s)	1MAO, 1MAI, 1CSD, 4BCT1, 4BP1, 4BS2, SPE
Exam(s)	MSc in Computer Science - Artificial Intelligence - Online MSc in Computer Science - Artificial Intelligence MSc in Computer Science - Data Analytics Fourth BSc in Computer Science & Information Technology Fourth BE in Electronic & Computer Engineering Fourth Bachelor of Science (Honours) (CS Pathway) Structured PhD
Module Code(s)	CT4101, CT5143
Module(s)	Machine Learning
Paper No.	1
Repeat Paper	No
External Examiner(s)	Dr Jacob Howe
Internal Examiner(s)	*Prof. Michael Madden Dr Patrick Mannion
Instructions:	Answer two questions from Section A and two questions from Section B. All questions carry equal marks. This is an open book exam.
Duration	2 hours
No. of Pages	6
Discipline(s)	Computer Science
Course Co-ordinator(s)	Dr Desmond Chambers (BCT), Dr Enda Howley (CSD), Dr Michael Schukat (MAI), Dr James McDermott (MAO)

Requirements:

MCQ	Release to Library: Yes
Handout	None
Statistical/ Log Tables	None
Cambridge Tables	None
Graph Paper	None
Log Graph Paper	None
Other Materials	None
Graphic material in colour	No

[PTO]

SECTION A

1. (a)

Super Mario Bros. (shown above) is a single player platformer video game. The goal of the game is to move the player character (Mario) to the right of the screen, navigating through each level safely until the flagpole at the end of the level is reached. During each level, Mario must avoid hazards such as pits and enemies that can result in death. Mario may also collect items such as coins and powerups that will increase the player's score. Mario may move to the left or right, jump, and attack enemies with a projectile if the player has collected the correct powerup. Each level has a time limit for completion (typically 300 seconds).

You are required to complete the initial design for an agent to play Super Mario Bros. You should therefore specify an MDP that models an agent's interactions with the Super Mario Bros. game. Your design should include the following:

- State space
- Action space
- Reward function

For each of the items above, you should provide a detailed explanation in your own words of the reasons for your design choices. [9]

(b) Explain in your own words the difference between **deterministic** and **stochastic** environments. In your answer you should discuss the importance of the **transition function T** in the MDP formulation.

In your view is the transition function T for the Super Mario Bros. MDP deterministic or stochastic? You should support your view on the nature of T for the Super Mario Bros. MDP with a detailed explanation of your reasoning. [6]

(c) Explain in your own words what is meant by the **Markov property**. Use the game Super Mario Bros. as an example in your answer. [4]

(d) Which of the below algorithms (as covered in the lecture notes) would be best suited to generate a policy to solve the Super Mario Bros. MDP? Justify your answer with an explanation in your own words of the key differences between each algorithm.

- Value iteration
- Q-learning

[6]

2. (a) Explain the following concepts in your own words:

- Entropy [2]
- Information gain [2]

[4]

(b) Consider the following set of 7 letters: SCIENCE

i. Calculate the entropy (in bits) of the letters in this set. Hint: you should treat each unique letter as a different class in your calculations. [4]

ii. Calculate the information gain (in bits) if the set of letters is split into two subsets: a subset containing the consonants, and another subset containing the vowels. [6]

iii. Calculate the maximum possible entropy (in bits) of a set of 7 letters. Hint: assume that each letter in the set is unique. [4]

For all parts i – iii above, provide complete calculations, along with detailed comments explaining your reasoning for each calculation step.

[14]

(c) ID3 is a widely used algorithm for inductive learning of decision trees. Explain in your own words and with simple original examples how the concepts of entropy and information gain are used by the ID3 process when building a decision tree.

[7]

[PTO]

3. (a)

ID	Births Live Young	Lays Eggs	Feeds Offspring Own Milk	Warm-blooded	Cold-Blooded	Land and Water Based	Has Hair	Has Feathers	Class
1	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	mammal
2	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	amphibian
3	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	mammal
4	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	bird
Q	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	???

The dataset above contains four examples of animals that have been classified as mammals, amphibians or birds, based on their features. You are required to apply similarity-based learning to this dataset to develop a model that can be used to classify newly discovered species (such as the unclassified species ID = Q above).

i. Propose a suitable method to measure similarity between the instances in this dataset (i.e. a distance metric or a similarity index covered in the lecture notes). Provide a detailed description of your chosen method in your own words, and justify your choice of method by explaining why it is more suitable than any other two methods covered in the lecture notes. [8]

ii. Using your chosen method from part i. above, compute the similarity/distance between the query Q and each of the four instances in the training set. [8]

iii. If you chose a 1-NN model with uniform weighting, what class would be assigned to the query Q? [2]

iv. If you chose a 4-NN model with uniform weighting, what class would be assigned to the query Q? [2]

v. In your opinion, is the 1-NN or the 4-NN model more suitable for this dataset? Provide an explanation in your own words to support your opinion. [5]

[25]

[END OF SECTION A]

[PTO]

SECTION B

4. (a) An immunologist researching a new virus has sent you the following message. Prepare a comprehensive reply, in your own words.
"I am applying machine learning to categorise cells in culture into one of three classes (healthy, unhealthy, dead), and I would appreciate your advice in analysing my results. Specifically, I am already calculating the accuracy of the machine learning method that I use, but a reviewer of my work has advised me to use a confusion matrix instead. What exactly does a confusion matrix show, how can I calculate one, and how does it relate to accuracy? Why did the reviewer recommend it, anyway – why would it be better? Another reviewer asked about training set coverage, and how many different training cases I should use in creating my classifier. Do you have any advice about this?" [10]

(b) Next, the immunologist would like to construct a ROC curve and compute AUROC. Can they do that for the task described in Part (a), or if not, how would they have to modify the task? Describe in detail in your own words the process of constructing a ROC curve for this. Provide an example of a ROC curve for this, with axes and all other parts of the curve clearly labelled. Also, explain in your own words what AUROC is and how it is computed. [8]

(c) In your own words, explain for the immunologist what can be determined from comparing two ROC curves. [3]

(d) Describe in your own words what the key differences are between a learning curve and a ROC curve. [4]

5. (a) A research group is testing the idea that societal factors might increase the probability of COVID-19 outbreaks. The believe that an outbreak probability depends on how densely populated a region is (high, average or low), the general level of air pollution (high, average, low, very low), and what kind of area it is (tourism area, shopping area, industrial area, residential area). In your own words, describe in detail a Bayesian algorithm to tackle this, including a description of the calculations required, the assumptions made and the probabilities that would have to be specified for your model. As part of your answer, include a diagram of the Bayesian classifier. [13]

(b) The two main ways of generating a Bayesian network structure are: (1) construct it by hand; (2) learn it from data. Outline the basic idea of these two approaches, in your own words. [6]

(c) Explain in your own words how to determine probabilities associated with a Bayesian network structure, after the structure has been determined. As part of your answer, explain what a conditional probability table (CPT) is. Referring back to the earlier parts of this question, provide an example of a single CPT, with probability values that you have made up. [6]

[PTO]

6. (a) You have received the following message from a biomedical scientist. Prepare a comprehensive reply in your own words, making reference to concepts including bias, variance, and hypothesis complexity.
"*I am trying to predict whether or not a drug will be successful for treating the symptoms of COVID19, based on the characteristics of the drug. I have done an initial study with data divided into a training set and a testing set. First I used linear regression and found that the model's root mean squared error was very high, even on the training data. Then I used a polynomial regression with terms to the power of 3 and 4, and found that the model worked better on the training data, but very poorly on the test data. Can you explain why I got these two results? Following on from this, can you give me at least two recommendations on what I should try next, with explanations for what you propose?* [10]

(b) *"The cost function for multiple linear regression has a global optimum."* Explain in your own words what this means. What are the implications of this for applying techniques such as gradient descent to linear regression? [4]

(c) In your own words, describe in detail the batch gradient descent algorithm, as applied to multiple linear regression, including how to initialise values, details of how they are updated, and how to check for convergence. As part of your answer, provide a definition of the empirical error cost function used in this algorithm. [8]

(d) What are the key differences between batch gradient descent and stochastic gradient descent, applied to linear regression? [3]

[END]