NUI Galway
OÉ Gaillimh

Semester 1 Examinations 2020/2021

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:	This is an open book exam. During the first 30 minutes of the exam period, you must complete an MCQ (worth 25%) on Blackboard. During the remaining 90 minutes of the exam period, you must answer any three questions on this exam paper. All questions on this exam paper carry equal marks (25% per question).
Duration	2 hours
No. of Pages	4
Discipline(s)	Computer Science
Course Co-ordinator(s)	Dr Desmond Chambers (BCT), Dr Frank Glavin (CSD), Dr Matthias Nickles (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]

1. (a)

The Legend of Zelda (shown above) is a single player role-playing video game. The player character (Link) must progress through various levels of the game, fighting enemies along the way using weapons such as a sword and a bow. During each level, Link must avoid hazards such as pits and enemies that can result in death. Link may move to the north, south, east, west, and attack enemies. Link may also collect helpful items such as health potions, keys to unlock new areas, and additional weapons. Levels in *The Legend of Zelda* typically do not have a time limit for completion.

i. Propose a suitable reward function that could encourage an agent to play *The Legend of Zelda* successfully. [3]

ii. Which of the below algorithms (as covered in the lecture notes) would be best suited to generate a policy for an agent to play *The Legend of Zelda*? Justify your answer with an explanation in your own words of the key differences between each algorithm.

- Value iteration
- Q-learning

[7]

iii. Explain in your own words what a Q-value (as learned by the Q-learning algorithm) represents. Also describe in your own words how a policy can be derived from Q-values. Use *The Legend of Zelda* **and** another original example (i.e. one not covered in the lecture notes) to aid your explanation. [8]

(b) i. Describe in your own words how you would detect overfitting when applying *k*-NN to a regression task. [3]

ii. Describe in your own words how z-normalisation may be applied to a dataset. Also explain with the aid of a simple original example (i.e. one not covered in the lecture notes) why it is often necessary to use normalisation with the *k*-NN algorithm. [4]

[25]

[PTO]

2. (a)

ID	CORE-TEMP	STABLE-TEMP	GENDER	DECISION
1	high	TRUE	male	gen
2	low	TRUE	female	gen
3	high	FALSE	female	icu
4	high	FALSE	male	icu
5	low	FALSE	female	icu
6	low	TRUE	male	icu

The training dataset above contains data about whether post-operative patients were sent to an intensive care unit (icu) or a general ward (gen) for recovery. You are required to develop a 1R model that can be used to predict the target DECISION for future patients.

When answering each part below, you should provide detailed comments explaining your calculations.

i. Calculate the entropy of the dataset with respect to the target variable DECISION. [2]

ii. Calculate the information gain when the dataset is partitioned using the CORE-TEMP feature. [4]

iii. Calculate the information gain when the dataset is partitioned using the STABLE-TEMP feature. [4]

iv. Calculate the information gain when the dataset is partitioned using the GENDER feature. [4]

v. Based on your answers to parts ii, iii and iv above, draw a diagram showing the completed 1R model. [2]

vi. In your opinion, would a decision tree be more appropriate than a 1R model for this dataset? [2]

(b) Briefly describe in your own words how you would apply the *k*-NN algorithm to the dataset in Q3 (a) above. As part of your answer, you should provide 2 examples of methods to measure similarity that would be suitable for this dataset, and 2 examples of methods to measure similarity that would not be suitable for this dataset. You should also discuss how you would choose a suitable value of *k*.

[7]

[25]

[PTO]

3. (a) You are collaborating with a group of vaccine development researchers. They are using a machine learning package for a binary classification task and wish to analyse their results using a confusion matrix, with a test set of 200 cases. Provide an example of a confusion matrix consistent with this, and make up values of your own for each entry (numbers, not percentages). Explain how every entry would be obtained in practice. For your confusion matrix, compute the accuracy and True Positive rate. [5]

(b) The research group next wish to compute a ROC curve for the problem. Can they do this for their task, or do they have to modify the task first? In your own words, describe in detail the process of constructing the ROC curve, including an illustration of one with axes and all other parts clearly labelled. [5]

(c) The research team have sent you a new message. They have a new training set of 500 cases, and they would like to perform repeated 5-fold cross-validation. In your own words, explain the process, including details of how many cases will be in each fold, how the data in each fold is used, and what it means to do it repeatedly. Also, how many classifiers will they build in total, and how do they estimate the performance of the final classifier? [8]

(d) In your own words, describe in detail an algorithm to build a linear classifier with multiple inputs, including an explanation of the perceptron learning rule. [7]

4. (a) A research group is testing the idea that geographic 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. [9]

(b) In your own words, describe in detail the stochastic 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, including the meaning of all terms in it. [7]

(c) In your own words, explain each of the following terms, in each case providing an example of an application and an appropriate technique: (i) feature extraction; (ii) feature transformation; (iii) automatic feature selection. [6]

(d) Automatic feature selection approaches may be divided into filter approaches and wrapper approaches. In your opinion, what are the key strengths and weaknesses of these two categories of approach? [3]

[END]