NUI Galway
OÉ Gaillimh

Autumn 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	5
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]

Please write answers on paper and scan the pages to upload a single pdf on Blackboard.

This is an open book exam. You may read textbooks, notes, and existing internet resources. You may not communicate with anyone, in person, via phone or internet, or otherwise at any time during the exam. You may not post questions concerning the exam, or post text or images from the exam paper, on internet sites or elsewhere, during or after the exam.

You are required to adhere to the highest standards of integrity and honesty when completing assessments and examinations at NUI Galway.

In submitting your work, you confirm that it is entirely your own. You acknowledge that you may be invited to online interview if there is any concern in relation to the integrity of my exam, and you are aware that any breach will be subject to the University's Procedures for dealing with breaches of Exam Regulations.

1. (a)

i. Using a simple original example of your choice, explain what is meant by the term **Markov property**. Briefly discuss how the Markov property may be leveraged to simplify the process of computing a policy for an MDP when applying a reinforcement learning algorithm. [5]

ii. Explain in your own words the difference between an MDP with a **discrete state space**, and an MDP with a **continuous state space**. As part of your answer, provide an original example of a problem with a discrete state space, and an original example of a problem with a continuous state space. [4]

iii. Explain in your own words what is meant by the term **reward function** in the context of an MDP. Provide an example of a reward function using an original problem domain of your choice to support your answer. [4]

(b)

i. Describe in your own words how you would detect overfitting when applying the ID3/C4.5 algorithm to a classification task. [3]

ii. Briefly explain in your own words what is meant by the term **pure inductive learning**. [2]

iii. Explain in your own words and with simple original examples how the concepts of entropy and information gain are used by the ID3/C4.5 algorithm when building a decision tree. [7]

[25]

[PTO]

2. (a)

property_id	num_beds	num_baths	floor_area	sale_price
275	1	1	40	123000
314	2	1	50	150000
2212	3	4	130	265000
3390	3	2	90	205000

The training dataset above contains data about the prices achieved at a recent auction for various properties. You are required to develop a *k*-nearest neighbours model using this training data; this model will be used to predict the target variable `sale_price` for properties to be sold at future auctions.

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

i. In preparation for applying 0-1 normalisation, choose appropriate minimum and maximum values for each of the independent variables (`num_beds`, `num_baths`, `floor_area`) that will be used to make predictions. [3]

ii. Using the minimum and maximum values that you chose for part i. above, compute the 0-1 normalised values for all independent variables for each data point in the training set. You should present the normalised data in tabular format. [4]

iii. Choose an appropriate distance metric (or similarity index) to use when applying k-NN to this dataset, and briefly justify your choice. [2]

iv. Using your chosen distance metric (or similarity index) from part iii. above, compute the distance (or similarity) between each pair of points in the training dataset using the normalised data from part ii. [5]

v. Using a 3-NN model with uniform weighting and your chosen distance metric (or similarity index), compute the predicted `sale_price` for each instance in the training dataset. [6]

vi. Using the predicted `sale_price` values from part v. above, compute the RMSE on the training set for the 3-NN model that you have developed. [3]

vii. Briefly comment on the RMSE value achieved by your 3-NN model – does this RMSE value indicate that the model performs well on the training data? Briefly describe one modification that could be made to the algorithm to improve the RMSE. [2]

[25]

[PTO]

3. (a) You are helping a startup company called CatPics, for sending your friends photos of cats and accompanying captions. To avoid spam, they wish to automatically approve or reject messages, based on the caption text. Here is a training set that you have been given:

Reject:	*Approve:*
"click this link"	"kitten sleeping today"
"weight drugs link"	"good luck kitten"
"good drugs news"	"smells good"
	"tiger smells daisies"
	"tiger link news"

Using a Naïve Bayes classifier without smoothing, compute the probability of the following two messages being **approved**: (i) "good drugs link"; (ii) "drugs news". Show all steps in your computation and explain any assumptions you make. [10]

(b) Explain in your own words what the purpose of Laplace smoothing is for Naïve Bayes classifiers. Recompute the probably of approving the message "drugs news", using Laplace smoothing with k=1. [4]

(c) The two main ways of generating a Bayesian network structure are: (1) construct it by hand; (2) learn it from data. In your own words, summarise the basic ideas behind both of these. [4]

(d) In your own words, explain in detail how to apply the gradient descent algorithm to Logistic Regression. Could you apply this algorithm to the data from Part (a)? Justify your answer. [7]

4. (a) An aeronautical engineer has sent you the following email message. Prepare a detailed reply in your own words.

"I am trying to predict the strength for some new composite materials in which we embed fibres in resin, based on some quantities that we can control such as the percentage by volume of fibres, the orientation of fibres, how much cooling we apply, and the overall mass of material. Since this is a numerical quantity, I believe I could use a regression algorithm for this, is that right? From some initial reading, I have heard of linear regression, polynomial regression, and logistic regression, but I am not clear about the differences between them. Can you explain the main distinctions, and let me know what kind of data I need for them? Also, can you recommend at least two other algorithms I should look into, in addition to those I have just mentioned, and explain why you would recommend them?" [10]

(b) The engineer has a dataset with 1000 cases, and initially intended to train a model with all of this data, and then test the model on all of the data also. In your own words, explain why this would not be a good approach. Continuing from this, explain what the distinctions are between data used for training and testing, and how they should divide the data. Is there a third category they should consider? If so, explain what it is and how they should handle it. [8]

(c) The engineer asks how many cases they need to fully train a model. Provide a clear response in your own words, making reference to learning curves. [4]

(d) The engineer has built two models, A and B, and wishes to determine which is best. Describe in your own words how to perform a suitable statistical test for this. [3]

[END]