TAKE-HOME ONLINE EXAMINATIONS APRIL – JUNE 2021

Please read through all instructions carefully before you start your exam

In most cases, the take-home online exam will be available for 23 hours and your answer(s) must be submitted before 9.00am UK time the day after your take-home online exam opens, unless you have been given a different deadline from your School.

Completing your take-home online exam

1. You need to download the question paper and any accompanying documents from Blackboard and your answers can be written in a Word document as you would for coursework (unless you have been given specific instructions from your School).
2. You are advised only to work on your answers for the duration of the time stated on the front page of the take-home online exam paper. You are not expected to work for 23 hours on your answer(s).
3. For some exams a timer will be applied so that, once started, you must submit within the specified time limit. Exams with a time limit will require you to submit either to a Blackboard test or to a Gradescope assignment. These will be clearly identified in Blackboard.
4. For exams which are time-restricted, and where you are required to upload your answers in a file, an extra 30 minutes will be added. This additional time is to allow you to scan any handwritten work and upload your file. Work submitted beyond this 30 minute period will be treated as late, and will not be marked.
5. Please read your exam paper thoroughly before you start to ensure that you understand what you need to do.
6. Do not exceed the specified word limits where they are stated.
7. You should use 12pt font size, Arial and 1.5 line spacing for word processed submissions.
8. Please write your 5-digit anonymous candidate number (from your RISIS exam timetable), module code and the number(s) of the question(s) answered on the top of each piece of work that you submit.
9. Save your work regularly as you are working on it.
10. You are responsible for the content of the work you upload and for the academic integrity of your answer(s).
11. Complete and upload your answer(s) to the submission point(s) in your Blackboard course.
12. You are responsible for organising your time and should aim to submit your answer(s) as early as possible to ensure your work is submitted prior to the deadline specified for your take-home online exam paper.

Full guidance can be found in the Take Home Exams area of this Blackboard course.

Submission

1. Please check the front of your take-home online exam paper and Blackboard for any specific instructions for uploading your work to submission point(s) for each paper.
2. You can submit multiple times for most exams (unless you have been told there is only one submission possible by your School), but you should ensure that your final version is uploaded before the deadline.
3. When submitting to Turnitin the ‘submission title’ must start with your 5-digit anonymous candidate number, followed by the module code. An example of a submission title is 12345 HS3DR.
4. You are responsible for ensuring that you have uploaded the correct document to the correct submission point. Some exams require submission of answers for different questions to different submission points.
5. You are responsible for ensuring that your file has been uploaded successfully. When you submit to a Turnitin, Blackboard or Gradescope ‘variable-length’ assignment you will receive an email receipt which you must keep. If you do not get this email receipt your work has not been submitted (check your spam folder).
   You will not receive an email receipt when you submit to a Blackboard Test or Gradescope Online Assignment.
   Please note that the final stage of a Turnitin submission requires you to confirm your submission by clicking ‘Confirm’.
6. Please allow yourself plenty of time to upload and submit your answer(s) by the deadline. If you have problems submitting your answer(s) please email your work to take-home-exam@reading.ac.uk as soon as possible.
7. We will not be emailing reminders from the University ahead of exam papers and submission points opening, or non-submissions after the take-home exam submission has closed.
8. Guidance on how to submit files for your exams can be found at https://rdg.ac/takehomeexam

Full guidance can be found in the Take Home Exams area of this Blackboard course.

Where to get support

If you need support during your exam you can contact us on +44 (0) 118 378 7049

For technical issues (Blackboard and IT) you can also raise a ticket via the DTS Self Service Portal.

For other non-technical queries, please check the exams FAQs on Essentials or email take-home-exam@reading.ac.uk. Emails need to be sent from your University email account and you should provide your 5-digit candidate number.

Please note that ‘live’ support is available from 8:00am-5:00pm UK time Monday to Friday and will be available for limited hours 8:00am-9.30am UK time on Saturday morning during the exam period.

DAS registered students

1. If you have been provided with a green sticker, please attach this to the front page of your answer(s), as you would do for coursework submissions.
2. If you have any additional arrangements including extra time, you should consider the necessary requirements prior to the start of your exam and read the advice in the exam FAQs on Essentials.
3. If you still have queries please contact the Disability Advisory Service (DAS) or email take-home-exam@reading.ac.uk. Emails need to be sent from your University email account and you should provide your 5-digit candidate and student number.

IMPORTANT - You must read this before you start your exam

Academic Integrity

We are treating this online examination as a time-limited open assessment. This means that:

1. You are permitted to refer to published materials to aid you in your answers.
2. Published sources must be referenced. This includes all on-line sources.
3. Over-reliance on published sources is considered to be poor academic practice.

Apart from appropriate referencing, you must ensure that:

a. the work you submit is entirely your own;
b. you do not communicate with other students on the topic of this assessment for the whole time the assessment is live;
c. you do not obtain advice or contribution from any third party, including proof-readers, friends, or family members.
d. For advice on academic integrity, you can see the University Library’s Academic Integrity Toolkit.

You should note that:

1. Failure to adhere to these requirements will be considered a breach of the Academic Misconduct regulations (available here), where the offences of cheating, plagiarism, collusion, copying, and commissioning are particularly relevant.
2. Your exam answers will be run through Turnitin, and the usual similarity reports will be available to markers.

Please read and note this statement of originality:

By submitting this work I certify that:

1. **it is my own unaided work;**
2. **the use of material from other sources has been properly and fully acknowledged in the text;**
3. **neither this piece of work nor any part of it has been submitted in connection with another assessment;**
4. **I have read the University's definition of plagiarism, guidance on good academic practice and the guidelines set out above; and**
5. **I will comply with the requirements these place on me.**

I acknowledge the University may use appropriate software to detect similarities with other third-party material, in order to ensure the integrity of the assessment.

I understand that if I do not comply with these requirements the University will take action against me, which if proven and following the proper process may result in failure of the year or part and/or my removal from membership of the University.

With best wishes and good luck for your take-home online exams over the coming period.

Please read the instructions below before you start the exam.

April/May 2021 **CS3DS19 2020/1 A 800**

UNIVERSITY OF READING

DATA SCIENCE ALGORITHMS AND TOOLS (CS3DS19)

One and a half hours

Answer any **TWO** out of THREE Questions.

If a word limit is not specified next to a QUESTION then EACH QUESTION (e.g. Q1, Q2, Q3, etc.) has a word limit of 1000 words.
This limit excludes scanned images of diagrams or hand-written formulas but includes images with hand-written text.

Submit your answers to **EACH QUESTION** SEPARATELY to the relevant submission point on Blackboard.

EACH Question is worth 20 marks.

1. Given a Classification problem and a dataset, where each record has several attributes and a class label, a learning algorithm can be applied to the data in order to determine a classification model. The model is then used to classify previously unseen data (data without a class label) to predict the class label.

(a) Hunt's algorithm is the general approach to learn a classification model in the form of a decision tree. Provide its pseudocode. What are the three main design choices in any 'specific' decision tree induction algorithm? Provide the definition of the GINI index for a single node and the GINI index for a binary split.

(8 marks)

(b) How do you measure the performance of a Decision Tree? What are the generalisation error and the resubstitution error?

(3 marks)

(c) A golf player keeps a record of the weather condition of days in which they went to play. Consider the set of records with four features (O, T, H, W) and a class ("play") shown in Table Q1-1. What data type (nominal, ordinal, binary) are the attributes and the class?

Compare the two decision trees shown in Figure Q1-1 by computing the two estimates of the generalization error based on the re-substitution error:

- the optimistic estimate and
- the pessimistic estimate with penalty term of 0.9.

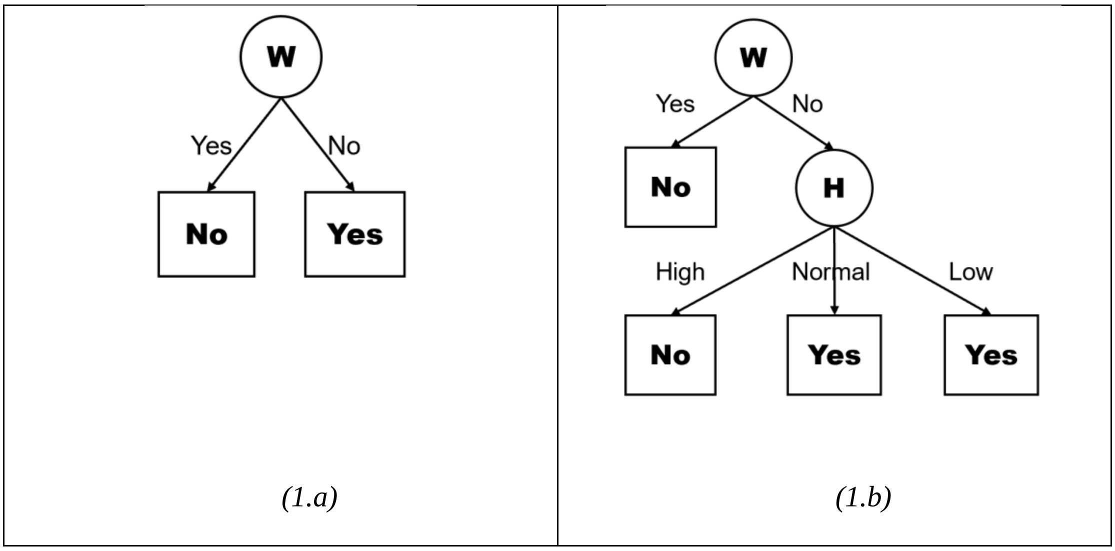
(5 marks)

(d) What is the meaning of the penalty term in estimating the generalisation error? For which value of the penalty term in the decision tree in Figure Q1-1.a would have a smaller pessimistic estimate of the generalisation error than the one in Figure Q1-1.b?

(4 marks)

ID	Outlook (O)	Temperature (T)	Humidity (H)	Windy (W)	play
1	Overcast	Cool	High	Yes	No
2	Overcast	Cool	Low	No	Yes
3	Overcast	Cool	Low	Yes	No
4	Overcast	Cool	Normal	Yes	No
5	Overcast	Hot	High	No	No
6	Overcast	Hot	Normal	No	Yes
7	Overcast	Mild	Low	Yes	Yes
8	Rainy	Cool	High	No	No
9	Rainy	Hot	High	No	No
10	Rainy	Hot	High	Yes	No
11	Rainy	Mild	Normal	No	Yes
12	Rainy	Mild	Normal	Yes	No
13	Sunny	Cool	Normal	No	Yes
14	Sunny	Cool	Normal	Yes	No
15	Sunny	Mild	High	No	Yes
16	Sunny	Mild	High	Yes	No

Table Q1-1. Golf data



(1.a)

(1.b)

Figure Q1-1. Decision Trees

2. (a) Briefly discuss Cluster Analysis in general and, in particular, two types of clustering: partitional and hierarchical.

(4 marks)

(b) Compare and contrast one algorithm for partitional clustering and one for hierarchical clustering in terms of advantages (at least two) and disadvantages (at least two), including their computational complexity.

(6 marks)

(c) Consider the set of 6 data points in 2 dimensions (x,y) in the table in Figure Q2-1. Apply one iteration of the k-means algorithm (for k=2) to find the cluster allocation (C0 or C1) of each data point and the values of the centroids at the end of the iteration (it-1).

Which of the two alternative initialisations of the centroids (c_0 and c_1 at it-0) given below produced the best clustering according to the cost function (SSE) optimised by k-means?

Provide the results in the following page as well as a worked solution (formulas and your arithmetic calculations) to compute the values of the centroids at the end of the iteration (it-1), the cluster allocations and the values of the cost function before and after the iteration (it-0 and it-1).

data ID	*x*	*y*
p_0	0.2	0.3
p_1	0.2	0.5
p_2	0.4	0.4
p_3	0.6	0.4
p_4	0.65	0.3
p_5	0.65	0.5

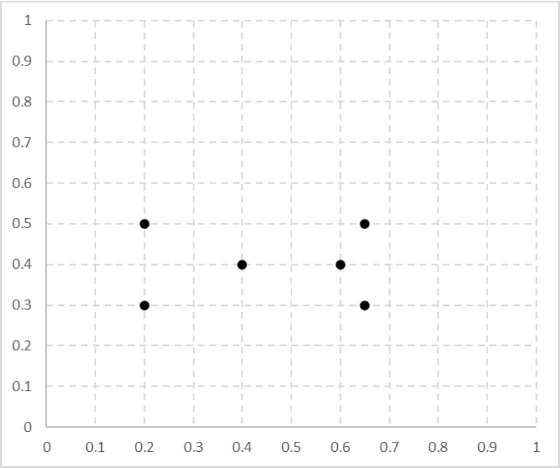


Figure Q2-1. The input data points

Case 1:

Centroid c_0 for Cluster 0:

c_0	x	y
it-0	0.2	1
it-1		

Centroid c_1 for Cluster 1:

c_1	x	y
it-0	0.5	1
it-1		

Cluster allocations:

data ID	*Cluster ID (0\1) after it-1*
p_0	
p_1	
p_2	
p_3	
p_4	
p_5	

	cost function (SSE) value
it-0	
it-1	

Case 2:

Centroid c_0 for Cluster 0:

c_0	x	y
it-0	0.3	1
it-1		

Centroid c_1 for Cluster 1:

c_1	x	y
it-0	0.6	1
it-1		

Cluster allocations:

data ID	*Cluster ID (0\1) after it-1*
p_0	
p_1	
p_2	
p_3	
p_4	
p_5	

	cost function (SSE) value
it-0	
it-1	

The best clustering solution is given by:

- □ Case 1
- □ Case 2

(10 marks)

3. An Association Rule is an implication expression of the form X →Y, where X and Y are disjoint itemsets.

(a) How many possible non-empty itemsets can be generated from a list of 12 unique items? How many non-redundant association rules can be generated from them? (5 marks)

(b) Is Association Rule Mining a descriptive or predictive data mining approach? Explain ARM and the meaning of the data model it provides. (3 marks)

(c) What are the support and the confidence of an association rule? Describe these measures and provide their formula. Given the transactions in Table Q3-1, what are the support and the confidence of the following four rules?

1. {steak} → {wine}
2. {bread, eggs} → cheese}
3. {cod, potatoes} → {peas}
4. {peas, potatoes} → {cod} (12 marks)

TID	***itemset***			
1	potatoes	onions	sausages	peas
2	cod	peas	potatoes	
3	bread	steak	crisps	wine
4	eggs	bread	oranges	
5	crisps	cola	sausages	beer
6	onions	potatoes	eggs	
7	cod	eggs	peas	wine
8	crisps	chocolate	cola	
9	crisps	beer		
10	steak	wine	lettuce	cheese
11	cheese	eggs	bread	
12	onions	potatoes	cod	
13	chocolate	crisps	cola	beer
14	oranges	peas	lettuce	potatoes
15	bread	cheese		
16	eggs	sausages	potatoes	
17	steak	cod	eggs	wine
18	crisps	chocolate		
19	bread	eggs	cheese	
20	bread	sausages	wine	

Table Q3-1. A list of transactions

(End of Question Paper)