A36904

Calculators may be used in this examination provided they are not capable of being used to store alphabetical information other than hexadecimal numbers



UNIVERSITY OF BIRMINGHAM

School of Computer Science

Fourth Year Undergraduate/Postgraduate

06-32258

LM Algorithms for Data Science

Main Summer Examinations 2025

Time allowed: 2 hours

[Answer all questions]

Note

Answer ALL questions.

Question 1

(a) What is cross validation used for? **[8 marks]**

(b) Briefly describe the steps of k-fold cross validation **[10 marks]**

(c) In model selection, under the same train/test split, different data you choose for training and testing may lead to different results. For example, under the 80/20 train/test split, if the first 80% data are used for training and the last 20% for testing, the linear model may have the smallest testing error; if the last 80% data are used for training and the first 20% for testing, the quadratic model may have the smallest testing error. It is known that k-fold cross validation can mitigate this issue. Do you think that k-fold cross validation can 100% rule out this happening? That is, under the same k value, different runs of k-fold cross validation never lead to different results; e.g., it is impossible that in one run the linear model has the smallest error, while in another run the quadratic model has the smallest error. Justify your answer. **[15 marks]**

Question 2

(a) Tim wants to model the relationship between temperature and the number of bikes hired, i.e., to build a model that can predict the number of bikes hired based on the temperature. Which one of the following models is the most suitable for? Briefly justify your answer. **[10 marks]**

- Linear regression
- Logistic regression
- Regression tree
- Naive Bayes

(b) Scientists need to predict whether planets in our galaxy are habitable for humans or not, based on several observable features. They collect some data as shown in the table. They want to create a decision tree based on the data in the table by using Gini impurity.

Planet	Size	Orbit	Rings	Habitable
A	big	near	yes	yes
B	small	far	no	no
C	big	far	yes	yes
D	big	far	no	no
E	small	far	no	no
F	small	near	no	yes
G	big	near	yes	no
H	big	near	no	yes

(i) Calculate Gini impurity values after using the feature Size to split the data (including its left child node, right child node and the overall split). **[6 marks]**

(ii) Which feature should you choose as the root of the decision tree when using Gini impurity to construct the tree? **[6 marks]**

(iii) Now suppose we have got a new feature, Colour, which has multiple choices. It can be blue, grey, brown, or red. Now we want to build a decision tree based on the four features. How many possible splits in total we need to consider when creating the root node? Briefly explain your answer. **[12 marks]**

Question 3

(a) Explain how AdaBoost differs from bagging-based ensemble methods like Random Forest. **[10 marks]**

(b) A bank needs to use AdaBoost to predict whether a transaction is fraudulent. The following table shows the initial predictions made by a weak classifier $h(x)$.

Transaction	Amount	Fraudulent (Label)	Weak Learner $h(x)$	Initial Weight
A	£600	no	yes	1/6
B	£880	yes	yes	1/6
C	£20	no	no	1/6
D	£800	yes	yes	1/6
E	£1000	no	yes	1/6
F	£450	no	no	1/6

(Note that you do not need to compute exact decimal values but should simplify where possible and show step-by-step calculations.)

(i) Calculate the Total Error of the weak learner $h(x)$. **[3 marks]**

(ii) Calculate the importance (Amount of Say) of the weak learner $h(x)$. **[4 marks]**

(iii) Update the weights for each sample for the next iteration. **[4 marks]**

(c) A bank is using AdaBoost to detect fraudulent transactions. After training, they observe that some transactions with extremely high amounts (£10,000 or more) are always misclassified, and the model assigns them increasingly high weights in each iteration. As a result, overall model accuracy starts to decrease.

(i) Explain why this happens in AdaBoost. **[4 marks]**

(ii) Discuss how AdaBoost can be improved to avoid this from happening. **[8 marks]**

This page intentionally left blank.

Do not complete the attendance slip, fill in the front of the answer book or turn over the question paper until you are told to do so

Important Reminders

- Coats/outwear should be placed in the designated area.
- Unauthorised materials (e.g. notes or Tippex) must be placed in the designated area.
- Check that you do not have any unauthorised materials with you (e.g. in your pockets, pencil case).
- Mobile phones and smart watches **must** be switched off and placed in the designated area or under your desk. They must not be left on your person or in your pockets.
- You are not permitted to use a mobile phone as a clock. If you have difficulty seeing a clock, please alert an Invigilator.
- You are not permitted to have writing on your hand, arm or other body part.
- Check that you do not have writing on your hand, arm or other body part – if you do, you must inform an Invigilator immediately
- Alert an Invigilator immediately if you find any unauthorised item upon you during the examination.

Any students found with non-permitted items upon their person during the examination, or who fail to comply with Examination rules may be subject to Student Conduct procedures.