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

Algorithms for Data Science

Main Summer Examinations 2024

Time allowed: 2 hours

[Answer all questions]

Turn Over

Note

Answer ALL questions. Each question will be marked out of 20. The paper will be marked out of 60, which will be rescaled to a mark out of 100.

Question 1 Linear Regression

Suppose John wants to find a linear model $f(x) = \theta_0 + \theta_1 x$ to fit a dataset shown in the following table. He considers $J(\theta) = \frac{1}{m}\sum_{j=1}^{m}(f(x^{(j)}) - y^{(j)})^2$ as the cost function, where m is the number of data items in the dataset.

y	x
1	1
3	2

(a) John uses the gradient descent to iteratively update the parameters θ_0 and θ_1 starting from $\theta_0 = \theta_1 = 1$. Answer the following 2 questions.

(i) Calculate the cost $J(\theta)$ when $\theta_0 = \theta_1 = 1$. **[5 marks]**

(ii) Calculate θ_0 and θ_1 after the first iteration of the update process in the gradient descent, where the learning rate is set to 0.1. **[5 marks]**

(b) Suppose that to avoid overfitting John now considers the regularisation, where the regularisation parameter is set to $\lambda = 1$.

(i) Calculate the cost $J(\theta)$ when $\theta_0 = \theta_1 = 1$. **[4 marks]**

(ii) Calculate θ_0 and θ_1 after the first iteration of the update process in the gradient descent, where the learning rate is set to 0.1. **[3 marks]**

(iii) Assume that after the training John finds the model function $f(x) = x$. Now he wants to test it on a new dataset in the following table. Calculate the error (i.e., cost). **[3 marks]**

y	x
1	0
2	2

Question 2 AdaBoost

Consider a dataset for predicting customer satisfaction in an e-commerce platform. Each data point represents a customer, along with a binary target variable indicating whether the customer is satisfied or not.

Customer ID	Gender	Premium Account	Returns	Satisfied (Label)
001	F	yes	no	yes
002	M	yes	no	no
003	F	no	yes	no
004	F	no	yes	yes
005	M	yes	yes	no
006	F	yes	no	no

(a) Discuss how AdaBoost combines weak learners to form a strong classifier. **[5 marks]**

(b) Assign an initial weight to each sample and build the first stump using Gini index. **[5 marks]**

(c) Calculate the 'Total Error' and 'Amount of Say' for the first stump. **[2 marks]**

(d) Assume that the new sample weight for the 6 Customers are [0.125, 0.125, 0.25, 0.125, 0.125, 0.25], create a new collection of samples base on the given sample weight. Briefly explain your answer. **[2 marks]**

(e) Instead of creating a new collection of samples, build the second stump by weighted Gini index with new sample weight provided in (d). **[6 marks]**

Question 3 : Clustering and Dimensionality Reduction

(a) Given the following data distributions, which clustering algorithm is most appropriate to use: **[2 marks]**

(i) Irregularly shaped or sized clusters (not overlapping)

(ii) Ellipsoidal-shaped clusters

(b) Explain whether the k-means method is sensitive to outliers. **[2 marks]**

(c) You are tasked with analysing a large dataset containing information about customers' purchasing behaviour in an e-commerce platform, where the data is collected from two different platforms such as Amazon and eBay. Design a solution that incorporates K-means clustering, the Elbow Method, and Principal Component Analysis (PCA) to segment customers into distinct groups based on their purchasing patterns. In your solution to this question, address the following points: **[16 marks]**

- Describe the steps you would take to preprocess the data.
- How would you select the number of PCs? Explain the logic behind using PCA for dimensionality reduction.
- Explain how to determine the optimal number of clusters using the Elbow Method.
- Segment the customers using K-means clustering, outlining the steps in order.
- Based on your understanding of k-means and PCA, why do you think applying them is useful for this kind of dataset (their contribution to the overall solution)? what are the potential insights that could be derived from the customer segmentation?

End of Paper

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.