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

LM Algorithms for Data Science

Main Summer Examinations 2023

Time allowed: 2 hours

[Answer all questions]

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 Regression and Model Selection

The following procedure is designed by someone to find a predictor (i.e., a regression model function for prediction) of house price in an area, based on the size of the house and the number of its bedrooms. However, there may exist some problems in the procedure. Please list all of them and briefly explain why they are a problem. **[20 marks]**

- Step 1. Split the given data for training and testing. For all the data, arbitrarily split them into two portions, one major portion (e.g., 80%) for training and the rest (e.g., 20%) for testing.
- Step 2. Evaluate all the models. For each of the considered models (e.g., linear and quadratic), firstly train it on the training data with the regularisation term (e.g., try the regularisation parameter $0, 0.01, 0.02, 0.04, 0.08, 0.16, \ldots, 10.24, \ldots$); secondly evaluate the trained model (i.e., calculate its error) on the testing data with the corresponding regularisation parameter used in training; and lastly choose the regularisation parameter that leads to the smallest error in the evaluation for the model.
- Step 3: Select the model and return its function and error. Among all the models, the model with the smallest error (denoted by ϵ) in the evaluation is selected. Its function (i.e., the trained model in Step 2), along with the regularisation parameter, is returned as the predictor, and the error ϵ is returned as the predicted error for new data in future.

Question 2 Naive Bayes

In a medical study, a hospital uses naive Bayes classifier to predict if a patient with some symptoms is healthy or not. The symptoms that they consider are "running nose", "coughing", "sore throat", and "headache", each of which takes the value true ("+") or false ("-"). They did some trials and collected the data as shown in the table.

examples	running nose (N)	coughing (C)	sore throat (T)	headache (H)	output
p1	+	−	+	−	− (healthy)
p2	−	+	−	+	+ (ill)
p3	+	−	−	+	− (healthy)
p4	−	+	−	−	− (healthy)
p5	+	+	−	−	+ (ill)
p6	+	+	+	−	+ (ill)
p7	−	+	−	+	+ (ill)
p8	−	−	−	+	+ (ill)

Table 1: Dataset of Question 1

(a) Predict if a person is ill or not if they have a running nose and coughing but no sore throat and headache. Please show the **step-by-step** calculations. **[8 marks]**

(b) There is another patient who has a running nose and sore throat, but we don't know if he/she is coughing or has headache. Predict if this person is ill or not. Please show the **step-by-step** calculations. **[6 marks]**

(c) Now suppose that we can further classify the people who are ill into two classes, having flu and having pneumonia. So now we have three classes. The corresponding target labels are provided in the table below. Predict which class the people who have the same symptoms as in Question (a) (i.e., have a running nose and coughing but no sore throat and headache) belongs to. **[6 marks]**

examples	running nose (N)	coughing (C)	sore throat (T)	headache (H)	output
p1	+	−	+	−	C1 (healthy)
p2	−	+	−	+	C2 (flu)
p3	+	−	−	+	C1 (healthy)
p4	−	+	−	−	C1 (healthy)
p5	+	+	−	−	C2 (flu)
p6	+	+	+	−	C3 (pneumonia)
p7	−	+	−	+	C3 (pneumonia)
p8	−	−	−	+	C2 (flu)

Question 3 Principal Component Analysis and Document Mining

(a) We are given a 6-dimensional data set $\mathcal{D} = \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^N\}$, where individual data points $\mathbf{x}^i = (x_1^i, x_2^i, ..., x_6^i)^T \in \mathbb{R}^6$, $i = 1, 2, ..., N$, characterise human subjects through 6 numerical features: x_1 – age (in years), x_2 – weight (in kilograms), x_3 – height (in millimetres), x_4 – number years of formal education, x_5 – average distance walked in a day (in miles), x_6 – average number of hours at work on a working day. By applying Principal Component Analysis (PCA) to $\mathcal{D}$, we obtain the following eigenvalues of the data covariance matrix, $\lambda_1 = 100, \lambda_2, = 0.1, \lambda_3 = 0.055, \lambda_4 = 0.05, \lambda_5 = 0.03, \lambda_6 = 0.01$. We thus naturally conclude that the data distribution is inherently one-dimensional and the data can be safely projected onto the dominant eigenvector of the covariance matrix, corresponding to λ_1.

(i) Explain why this is a wrong conclusion. **[3 marks]**

(ii) What do you expect the dominant eigenvector to look like and what can you say about the nature of the data projections onto it? **[4 marks]**

(iii) What modifications need to be performed to remedy the problem? **[3 marks]**

(b) Explain the following schemes for vector representations of document data in the bag-of-words framework. Provide advantages and disadvantages for each scheme.

(i) binary representation of terms **[2 marks]**

(ii) frequency representation of terms **[3 marks]**

(iii) TFIDF - term frequency inverse document frequency representation **[5 marks]**

This page intentionally left blank.

LM Algorithms for Data Science **Non-alpha only**

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.