UNIVERSITY OF LONDON

CM3005

BSc EXAMINATION

COMPUTER SCIENCE

Data Science

Release date: TBD

Submission date: TBD

Time allowed: 24 hours to submit

INSTRUCTIONS TO CANDIDATES:

Section A of this assessment paper consists of a set of **10** Multiple Choice Questions (MCQs) which you will take separately from this paper. You should attempt to answer **ALL** the questions in Section A. The maximum mark for Section A is **40**.

Section A will be completed online on the VLE. You may choose to access the MCQs at any time following the release of the paper, but once you have accessed the MCQs you must submit your answers before the deadline or within **4 hours** of starting whichever occurs first.

Section B of this assessment paper is an online assessment to be completed within the same 24-hour window as Section A. We anticipate that approximately **1 hour** is sufficient for you to answer Section B. Candidates must answer **2** out of the 3 questions in Section B. The maximum mark for Section B is **60**.

Calculators are not permitted in this examination. Credit will only be given if all workings are shown.

You should complete **Section B** of this paper and submit your answers as **one document**, if possible, in Microsoft Word or a PDF to the appropriate area on the VLE. You are permitted to upload 30 documents. However, we advise you to upload as few documents as possible. Each file uploaded must be accompanied by a coversheet containing your **candidate number**. In addition, your answers must have your candidate number written clearly at the top of the page before you upload your work. Do not write your name anywhere in your answers.

SECTION A

Candidates should answer the **10** Multiple Choice Questions (MCQs) quiz, **Question 1** in Section A on the VLE.

Question 1

(a) With respect to matrices, which of the following are scalar values?
Select ALL statements that apply.

i. Rank
ii. Determinant
iii. Trace
iv. Cross product

(b) Which of the following are measures of central tendency?
Select ALL statements that apply.

i. Mean
ii. Median
iii. Mode
iv. Range

(c) Which of the following are unsupervised learning tasks?
Select ALL statements that apply.

i. classification
ii. regression
iii. clustering
iv. dimensionality reduction

(d) How is precision defined?
Choose ONE option.

i. The proportion of predicted positive values that are actually positive
ii. The proportion of predicted negative values that are actually negative
iii. The proportion of actual positive values that are predicted positive
iv. Accuracy divided by recall

(e) Which of the following is true of high complexity machine learning models?
Select ALL statements that apply.

i. They exhibit higher bias than low complexity models
ii. The learning curves on training and validation data converge sooner than low complexity models
iii. Adding more training data may improve the fit
iv. They exhibit higher variance than low complexity models

(f) How does lemmatization differ from stemming?
Select ALL statements that apply.

i. Stemming only works for regular verbs
ii. Lemmatization is informed by the linguistic context
iii. Stemming is a more crude, heuristic process
iv. Stemming requires access to a lexical database

(g) What are some of the shortcomings of traditional ‘one-hot’ vector encodings’?
Select ALL statements that apply.

i. They tend to be very short
ii. They tend to be relatively sparse
iii. They tend to be very dense
iv. They tend to contain many zero elements

(h) Which of the following would you use to show the correlation between true positive labels and predicted positive labels?
Choose ONE option.

i. Confusion matrix
ii. Probability density function
iii. Network diagram
iv. Bar chart

(i). A key idea behind support vector machines is maximizing which of the following?
Choose ONE option.

i. The margin
ii. The loss function
iii. The learning rate
iv. The gradient

(j) What would you use to measure the proportion of spam messages that are correctly predicted as spam by your classifier?
Choose ONE option.

i. Negative predictive value
ii. Precision
iii. Recall
iv. Accuracy

SECTION B

Candidates should answer any **2** questions from Section B.

Question 2

(a) What is the difference between machine learning, artificial intelligence, and data science?

(4 marks)

(b) What are feature vectors?

(4 marks)

(c) What is linear regression, and what is meant by multiple linear regression? Give an example of the difference between regression and classification in machine learning.

(8 marks)

(d) What is the difference between linear regression and logistic regression? Give an example of a logistic regression problem.

(6 marks)

(e) What is ‘Naive’ in a Naive Bayes? Explain the assumptions that it makes, and give examples where these may not hold.

(8 marks)

Question 3

(a) What are differences between univariate, bivariate, and multivariate data analysis? Give examples of each, and suggest suitable visualization techniques for each type of data.

(12 marks)

(b) What do you understand by the term Normal Distribution? State the typical properties of a normal distribution. What proportion of the data would be within, 1, 2 and 3 standard deviations from the mean?

(10 marks)

(c) What is TF.IDF vectorization, and what is it used for? Explain the benefit it provides over more basic representation schemes.

(8 marks)

Question 4

(a) What is cross-validation? Give examples of different cross-validation techniques.
(6 marks)

(b) What cross-validation technique would you use on a time series data set? Give an example of how you would apply it to a data set with 5 time periods.
(6 marks)

(c) What is the difference between a Validation Set and a Test Set?
(4 marks)

(d) You are given a data set consisting of samples with missing values. What methods might you apply to fill in missing data? What are the potential negative consequences of doing so?
(8 marks)

(e) You have built a classification model for tumour detection that achieves an accuracy of 96 percent. Why might this figure be misleading, and what might be a better measure of performance?
(6 marks)

END OF PAPER