Queen Mary
University of London

Main Examination Period 2019

ECS784P Data Analytics **Duration: 2 hours 30 minutes**

YOU ARE NOT PERMITTED TO READ THE CONTENTS OF THIS QUESTION PAPER UNTIL INSTRUCTED TO DO SO BY AN INVIGILATOR

Answer ALL FOUR questions

Calculators **are** permitted in this examination. Please state on your answer book the name and type of machine used.

Complete all rough workings in the answer book and cross through any work that is not to be assessed.

Possession of unauthorised material at any time when under examination conditions is an assessment offence and can lead to expulsion from QMUL. Check now to ensure you do not have any notes, mobile phones or unauthorised electronic devices on your person. If you do, raise your hand and give them to an invigilator immediately.

It is also an offence to have any writing of any kind on your person, including on your body. If you are found to have hidden unauthorised material elsewhere, including toilets and cloakrooms it will be treated as being found in your possession. Unauthorised material found on your mobile phone or other electronic device will be considered the same as being in possession of paper notes. A mobile phone that causes a disruption in the exam is also an assessment offence.

EXAM PAPERS MUST NOT BE REMOVED FROM THE EXAM ROOM

Examiners: Dr. Anthony Constantinou, Mr Bhusan Chettri

Question 1

a) Explain the difference between the following terms:

   i) Data vs big data
   ii) Structured vs unstructured data
   iii) Qualitative vs quantitative data

[6 marks]

b) Data analytic projects involve expertise from different disciplines. What are the three main desirable disciplines in any data analytic project? Briefly explain the contribution to a data analytic project by each discipline.

[4 marks]

c) What do you understand by data analytics? Briefly explain the various phases involved in a data analytic project.

[12 marks]

d) Convert the given block of statements below into an equivalent list comprehension.

```
list_of_numbers = np.arange(100)
even_numbers = list ()

for number in list_of_numbers:
    if number % 2 == 0:
       even_numbers.append(number)
```

[3 marks]

Question 2

a) Define the following terms:

i) Overfitting
ii) Data wrangling
iii) Underfitting
iv) Deep learning
v) Machine learning

[5 marks]

b) Discuss the two efficient data-structures that Pandas library offer for data analysis.

[5 marks]

c) There can be *N* different algorithms to solve a given problem *'P'*. As a data scientist you must ensure that the approach you chose is justifiable and performs good enough in real time. Discuss the different criteria you would consider choosing one over the other.

[4 marks]

d) What do you understand by worst case, best case and average case running time of an algorithm? Discuss the best case and worst case running time of the 'Binary Search' algorithm.

[6 marks]

e) What is a confusion matrix and why do we use it? Discuss with a suitable example.

[5 marks]

Question 3

a) It is well-known that structure learning algorithms perform better on simulated scenarios than they do on real-world scenarios. Suppose we experiment with a structure learning algorithm by testing the algorithm on simulated data. We can manipulate simulated data as we wish.

For each statement enumerated below, which represents a type of data manipulation, indicate whether the statement is TRUE or FALSE, and explain why.

To make the results obtained from the structure learning algorithm more realistic in terms of real-world performance, we should manipulate simulated data by:

i. Making some values missing.
ii. Making some variables missing.
iii. Increasing the number of states per variable.
iv. Decreasing the number of states per variable.
v. Randomly changing some of the data values.
vi. Increasing the number of variables.

[12 marks]

b) Explain the difference between a conditional arc and an informational arc, and then describe two examples where each arc becomes useful.

[5 marks]

c) Discuss two main steps of the PC algorithm and state two limitations of the PC algorithm.

[8 marks]

Question 4

a) Give two examples of knowledge-based constraints which may be incorporated into the structure learning process of a causal discovery algorithm. Name two advantages and two disadvantages of knowledge-based constraints.

[6 marks]

b) Figure 1 shows a simple Bayesian Network.

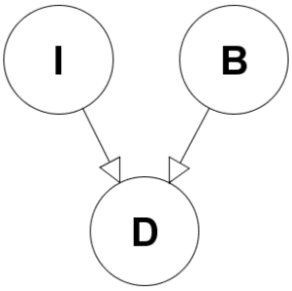

Figure 1. A Bayesian Network.

You are given the following information:

$P(d_1|b_1, i_1) = 0.25$
$P(d_1|b_1, i_2) = 0.1$
$P(d_2|b_2, i_1) = 0.7$
$P(d_1|b_2, i_2) = 0.2$
$P(i_2) = 0.12$
$P(b_1) = 0.65$

i. Complete the Conditional Probability Table of node D shown in Table 1.

Table 1. CPT of node D.

I	i_1		i_2	
B	b_1	b_2	b_1	b_2
d_1				
d_2				

ii. Find $P(d_2)$ and show the full calculation.

[6 marks]

c) What are the current issues with evaluating Bayesian Network structure learning algorithms? List and discuss three of them.

[6 marks]

d) Explain Simpson’s paradox. Describe a real-world scenario in which it can occur and further explain how it can be avoided in that particular scenario.

[7 marks]

End of Paper