MONASH University

2022 Semester Two (October 2022)(Sample) Examination Period

Faculty of Information Technology

EXAM CODES: FIT1043
TITLE OF PAPER: Introduction to data science
EXAM DURATION: 2 hours 10 mins

Rules

During your eExam, you must not have in your possession any item/material that has not been authorised for your exam. This includes books, notes, paper, electronic device/s, smart watch/device, or writing on any part of your body. Authorised items are listed above. Items/materials on your device, desk, chair, in your clothing or otherwise on your person will be deemed to be in your possession. Mobile phones must be switched off and placed face-down on your desk during your exam attempt.

You must not retain, copy, memorise or note down any exam content for personal use or to share with any other person by any means during or following your exam. You are not allowed to copy/paste text to or from external sources unless this has been authorised by your Chief Examiner.

You must comply with any instructions given to you by Monash exam staff.

As a student, and under Monash University's Student Academic Integrity procedure, you must undertake all your assessments with honesty and integrity. You must not allow anyone else to do work for you and you must not do any work for others. You must not contact, or attempt to contact, another person in an attempt to gain unfair advantage during your assessment. Assessors may take reasonable steps to check that your work displays the expected standards of academic integrity.

Failure to comply with the above instructions, or attempting to cheat or cheating in an assessment may constitute a breach of instructions under regulation 23 of the Monash University (Academic Board) Regulations or may constitute an act of academic misconduct under Part 7 of the Monash University (Council) Regulations.

Closed Book
This is a closed book assessment. You're not permitted to use any notes, texts, websites or other reference material to assist you in answering the questions.

If your assessment contains file download/upload questions, you are permitted to use the application required to execute the downloaded file.

Instructions

- This is a closed-book exam.
- Please answer ALL questions. Marks are indicated next to each question. The total marks for the exam are 65 marks.
- This exam is divided into 2 parts and the total marks for the exam is 65 marks:

  Part 1 consists of 15 Multiple Choice Questions and is worth 15 marks. Each question is worth 1 mark.

  Part 2 consists of 25 Short Answer Questions and is worth 50 marks. Each question is worth 2 marks.
- Once the exam duration is finished, your exam will automatically submit. Please ensure you finalise your answers before the end
  of the allocated exam time

Multiple Choice Questions

Information

Part 1: Multiple Choice Questions (15 marks in total)
This part consists of 15 Multiple Choice Questions and is worth 15 marks. Each question is worth 1 mark. Please answer ALL questions. Identify the choice that best completes the statement or answers the question. There is only one best answer for each question. Sometimes two answers may appear feasible, but you are to pick the one you believe is the best.

Marking Scheme for Multiple Choice Questions:
1 mark for a correct answer
0 marks for a wrong
0 marks for no answer

Please pay attention that this is a sample exam, and we provided some questions to give you an insight into the type of questions which you would have in your final exam. The number of questions is less than what you will see in your final exam. You will have 15 multiple choice question and 25 short answer questions in your final exam.

Question 1
Drew Conway Venn Diagram: Which of the following best explains the "Danger Zone" intersection in the Drew Conway Venn Diagram?

A. It describes people who are well versed in conducting end to end machine learning and report coefficients, but without understanding what they mean.
B. It describes people who are not sure what they are doing, although they can explain the meaning of the output of the coefficients.
C. It is an area that we should not enter as it is dangerous and can result in harmful analysis.
D. It is an area where people conduct trial and error experiments with data and report the best results.

Question 2
Machine learning is useful when:

A. human expertise is not available
B. humans cannot explain their expertise (as a set of rules)
C. humans are expensive to use for the work
D. ALL of the other cases

Question 3
What is the proper explanation for the following Python code:
(No answer given, you try on your own Jupyter. Read the titanic dataset, etc. :)

```
titanic.groupby(['sex','class'])['age']
```

A. It groups the data by the data based on Sex and Class and returns the average

B. It shows the average of age in each class and sex
C. It groups the data based on Sex and Class
D. It first groups the data by sex. Then shows the average age in different classes

Question 4

What is Hadoop?

A. An abbreviation for “Hadrian's Loop”, a firewall management system
B. A programming language designed for agile development
C. An encryption system used extensively at Google
D. A system for partitioning computation across a compute cluster

Question 5

The 3Vs of big data are important because:

A. they are an industry standard.
B. they are the basis for the development of more Vs (e.g. Value).
C. they are used to describe in what way a dataset may be too big to handle.
D. they are from the influential Gartner Inc.

Question 6

Privacy: What is the technological reason for the continued increase in lack of privacy?

A. the flow of technology makes surveillance easier unless particular measures are set in place.
B. the increase in cybercrime and terrorism makes it a necessity.
C. the open internet and the cloud remove privacy.
D. it follows from Koomey's Law.

Short Answer Questions

Information

Part 2: Short Answer Questions (50 marks in total)
This section consists of 25 Short Answer Questions and is worth 50 marks. Each question is worth 2 marks. Your answer should be written in clear, simple English and should be complete enough in addressing the question. Extensive prose is not required. Structured bullet points are acceptable.
The questions are numbered from 16 onwards as in the actual exams (actual exams will have Questions 16 – 40).

Question 16
Explain what big data is. Consider the four V's of big data and explain veracity in a few words.

Sample answer:
BIG DATA is any attribute (among the V's) that challenges constraints of a system capability or business need and veracity is uncertainty of data.

- Volume is size of data.
- Velocity is the frequency/Pace of incoming data that needs to be processed.
- Variety refers to different types of data.
- Veracity refers to the fact that how accurate or truthful a data set may be. More specifically, how accurate and reliable the data is?

Question 17
Assume you are collecting data about traffic accidents in Melbourne to develop a predictive model. Would it be better to collect "more data" (e.g., the locations of accidents over many years) or "more types of data" (e.g., the types of vehicles involved, the weather conditions, etc)? Give a brief justification.

Sample answer:
Assuming there is sufficient data for building a predictive model, usually more types of data helps a predictive model more than just collecting the more data.
However, if there are insufficient amount of data, then it would be better to ensure that there is sufficient data for the model building.

Question 18
Explain the differences between a classification and a regression. Which one can be used to predict a salary based on age and job title of a person?

Sample answer:
- Classification: The depended variable is a categorical variable. i.e., discrete values (categorical), such as Spam or not Spam.
- Regression: The depended variable is a continuous value such as price.

We can use regression to predict the salary based on the age and job as salary is a continuous value.

Question 19
Would you consider user's emails as to be sensitive information? Why or why not?

Sample answer:
Yes, emails should be considered as sensitive information. They may contain different types of private information such as addressed, cell phone numbers, vacation notices, financial information and so on. This information should be confidential.

Question 20
Explain the k-means algorithm.

Sample answer:
K-Means is an unsupervised clustering machine learning algorithm which groups the similar data points together to help us discover the underlying patterns by looking at the fixed number of clusters (k). The algorithm is as follows:

1. Define the K
2. Initialize the centroids
3. Assigns the data points to the centroids 4- Update the centroids
4. If the new values of the centroid changed significantly from the previous values, return to step 3; otherwise, stop the algorithm.

(In the actual exams, there will be questions until question 40)

END OF EXAMS