MONASH University

2024 Semester One (June 2024)(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 choicequestion 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.

Question 7
Which of the following statements about Python is TRUE?
   A. The first element of an array in Python has the index 1.
   B. Python was designed by statisticians.
   C. Python is a scripting language.
   D. Python is a proprietary programming language.

Question 8
Unix shell commands like "less" and "grep":
   A. can be used to manipulate large data files easily
   B. are poorly documented
   C. are used to fit regression tree models
   D. are examples of technology that is too old to be useful to a modern data scientist

Question 9
Over the years, disk capacity is generally growing:
   A. exponentially
   B. quadratically
   C. linearly
   D. logarithmically

Question 10
Which of the following is real world applications of Machine Learning?
   A. Self-driving cars
   B. Spam filtering
   C. Weather forecasting
   D. All of the options.

Part 2 : Short Answer Questions (SAQ)

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

Sample answer:
Big data refers to the large, diverse sets of information that grow at ever-increasing rates. It encompasses the volume of information, the velocity or speed at which it is created and collected, the variety or scope of the data points being covered, and the quality and availability of data (veracity).

Question 17
List some types of the metadata might be associated with an image.

Sample answer:
Descriptive metadata: title of image, the owner, the number of versions…
Structural metadata: keywords, collections…
Administrative metadata: rights management information, restrictions on use, licensing rights, and expiration date…

Question 18
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 19
Explain "sensitivity" as one of the classification metrics.

Sample answer:
Sensitivity describes (True Positive Rate) the ratio of how many predictions were correctly labelled positive out of all predictions that were actually positive. TP / TP + FN

Question 20

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 21

Name two different data science roles (jobs) and explain their responsibilities.

Sample answer (any of the two):
- Data developers: people focused on the technical problem of managing data—how to get it, store it, and learn from it
- Data researchers: people with an academic research background, using their training to "understand complex processes"
- Data businesspersons: most focused on the organization and how data projects yield and profit
- Data creatives: the broadest of data scientists, those who excel at applying a wide range of tools and technologies to a problem

(You could have used role names such as data scientist, data analyst, business analyst to answer this question as well.)

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

END OF EXAMS