Two hours

UNIVERSITY OF MANCHESTER
DEPARTMENT OF COMPUTER SCIENCE

Data Engineering

Date: Wednesday 15th January 2020

Time: 09:45 - 11:45

Please answer BOTH Questions
Each Question is worth 30 marks

Use a SEPARATE answerbook for each SECTION

This is a CLOSED book examination

Electronic calculators may be used in accordance with the University regulations

Section A

1. a) In the context of classification:

(i) Outline a decision tree classification algorithm; discuss how the attribute used at each node is chosen and what effect different training sets may have. (4 marks)

(ii) Tests on a classifier give the following confusion matrix:

		Predicted		
		win-yes	lose-no	Total
Actual	win-yes	6	3	9
	lose=no	2	9	11
	Total	8	12	20

Calculate the precision, recall, and F1-Measure for the classifier based on this table. Provide the formulae you use for calculation. (3 marks)

(iii) Explain the potential effects of unbalanced data on the usefulness of a classifier. Give an example to support your answer. (2 marks)

b) In the context of association rule (itemset) mining:

(i) Outline the working of the Apriori algorithm. What is the effect of the subset property on itemset generation? (4 marks)

(ii) Using Apriori, suppose that L3 is the list:

{ {1, 2, 3}, {1, 2, 4}, {1, 3, 4}, 2, 5}, {1 ,3, 5}, {2, 3, 4} }

a. At the join step of the algorithm, which itemsets are placed in C4 (the candidate set)? (3 marks)

b. Which itemsets are discarded by the prune step of the algorithm? (2 marks)

Provide your working as appropriate.

(Question 1 continues on the following page)

(Question 1 continues from the previous page)

c) (i) Discuss the trade-offs between privacy of the individual hand and the utility of data – use case studies as appropriate to illustrate your answer. How do we guarantee that data is private and that methods might be used to “fuzz” data and increase anonymity and diversity? (4 marks)

(ii) Compare and contrast statistics and data analytics/mining. Explain how their different approaches to “learning from data” are complementary and can proceed side-by-side. (4 marks)

(iii) Discuss the current state of data warehousing and how it has adapted to the emerging “big and complex data” revolution - including several, often dynamic, data sources and their requirements for analysis. (4 marks)

Section B

2. a) Discuss, using examples to help illustrate your answer, what constitutes data of high value, including discussions and comparisons of both data of high value that does not degrade over time, and data of high value that degrades over time. (3 marks)

b) Describe the Data Life Cycle, giving, for each phase, one example of an issue that must be addressed during that phase. (8 marks)

c) (i) Discuss the main differences between a Data Warehouse and a Data lake. (4 marks)

(ii) Describe an example business scenario where a Data Warehouse would be appropriate and a business scenario where a Data lake would be appropriate. In each example, discuss the rationale for why it is appropriate. (2 marks)

(iii) Briefly discuss how an organisation can look to ensure that its Data Lake does not turn into a Data Swamp. (2 marks)

d) Discuss, using examples to help illustrate your answer, possible reasons why data might be dirty. (4 marks)

e) Discuss the various ways that the issue of missing data might be tackled, and discuss when each way might be most appropriate. (4 marks)

(Question 2 continues on the following page)

(Question 2 continues from the previous page)

f) Consider the following relational database schemas that store information about airline flights, passengers and bookings, and which need to be integrated. Aiming at reconciling the conflicts that need to be resolved before integration is possible, answer the following questions.

Use the notation Schema_i.TableName.AttributeName, to refer to an attribute in any of the two schemas; Schema_i.TableName, to refer to a table in any of the two schemas.

Schema_1:

Passenger (passengerID, givenNames, surname, address, age, emailAddress)
Booking (*passengerID*, *flightID*, bookingDate)
HoldBag (*passengerID*, *flightID*, weightLimit, weight)
Flight (flightID, from, departureDate, departureTime, to, arrivalDate, arrivalTime, *registrationNumber*)
Airplane (registrationNumber, modelNumber, capacity)

Schema_2:

Flight (flightID, origin, destination, timeAndDateOfDeparture, timeAndDateOfArrival, *registrationNumber*)
Aeroplane (registrationNumber, numberOfSeats, modelNumber)
Customer (passengerID, emailAddress, firstName, middleNames, lastName, age, houseNo, street, city, postcode)
FlightBooking (*passengerID, flightID*)
Luggage (*passengerID*, *flightID*, luggageKg, maxKgAllowance)

(i) Describe two different one-to-one table name conflicts. (1 mark)

(ii) Describe two different one-to-one attribute vs. attribute conflicts. (1 mark)

(iii) Describe two different many-to-many attribute conflicts. (1 mark)

END OF EXAMINATION