DCU

Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

AUGUST/RESIT EXAMINATIONS 2022/2023

MODULE: CA270 - Data Warehousing and OLAP

PROGRAMME(S):

DS	BSc in Data Science
ECSA	Study Abroad (Engineering & Computing)

YEAR OF STUDY: 2,X

EXAMINER(S):

Mark Roantree	(Internal)	(Ext:5636)
Dr. Ziqi Zhang	(External)	External

TIME ALLOWED: 2 Hours

INSTRUCTIONS: **Answer Question 1 (40 marks) and 2 other questions (equal 30 marks each)**

PLEASE DO NOT TURN OVER THIS PAGE UNTIL YOU ARE INSTRUCTED TO DO SO.
The use of programmable or text storing calculators is expressly forbidden.
Please note that where a candidate answers more than the required number of questions, the examiner will mark all questions attempted and then select the highest scoring ones.

There are no additional requirements for this paper.

QUESTION 1 (Clustering) [TOTAL MARKS: 40]

Q 1(a) [8 Marks]

Explain the difference between a data matrix and a dissimilarity matrix. Provide an example of both in your answer.

Q 1(b) [12 Marks]

(i) How does a data scientist calculate a dissimilarity matrix for binary variables? Explain your answer through the use of the contingency table.

(ii) Write and explain the formula for *symmetric* dissimilarity. In what circumstances would you use this function?

(iii) Write the function for *asymmetric* dissimilarity. Why might you use the asymmetric dissimilarity function?

Q 1(c) [20 Marks]

Given a data set with 1-dimensional points D = {1, 3, 6, 10, 20, 100}, and using the k-medoids approach, where k = 2 and the initial medoids are 1 and 100.

(i) Define and explain the absolute-error function used in *k*-medoids.

(ii) Calculate the cost of the 2 clusters and show how each point is assigned to its respective medoid.

(iii) Assume the first iteration of the algorithm tests whether 1 could be replaced with a new medoid with value 3.
What is the result of testing in this step? Show the calculations needed to arrive at your answer.

[End of Question 1]

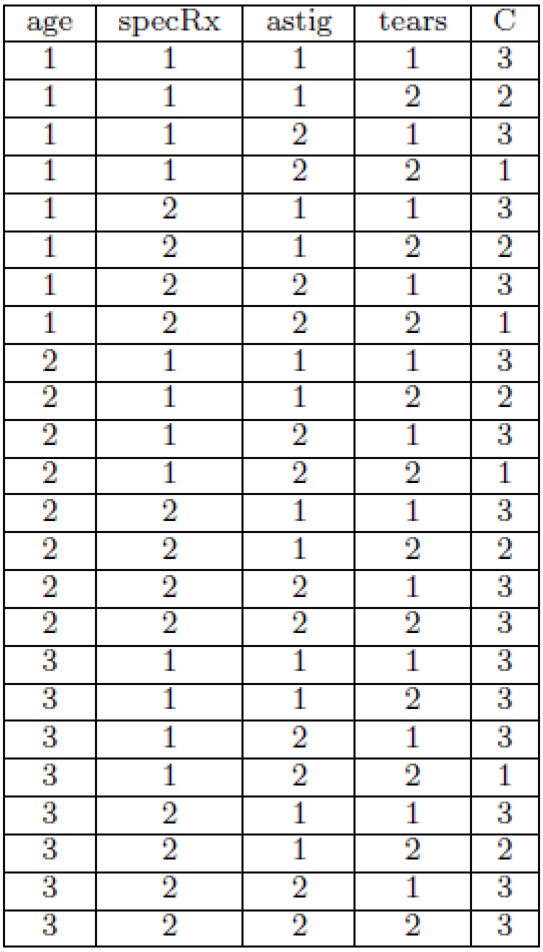
QUESTION 2 Classification *[TOTAL MARKS: 30]*

Q 2(a) [10 Marks]

Describe the 3 attribute selection approaches covered in the course, in relation to constructing a decision tree. In your answer, just comment on these aspects:

(i) How these approaches work;
(ii) Any weakness or bias they may have;
(iii) How they differ from each other.

Q 2(b) [20 Marks]



age	specRx	astig	tears	C
1	1	1	1	3
1	1	1	2	2
1	1	2	1	3
1	1	2	2	1
1	2	1	1	3
1	2	1	2	2
1	2	2	1	3
1	2	2	2	1
2	1	1	1	3
2	1	1	2	2
2	1	2	1	3
2	1	2	2	1
2	2	1	1	3
2	2	1	2	2
2	2	2	1	3
2	2	2	2	3
3	1	1	1	3
3	1	1	2	3
3	1	2	1	3
3	1	2	2	1
3	2	1	1	3
3	2	1	2	2
3	2	2	1	3
3	2	2	2	3

Use Information Gain to rank the attribute set (age, specRx, astig, tears) for selection for the first branch in the decision tree, from best to worst.

Important: you are required to provide the **initial** attribute selection only.

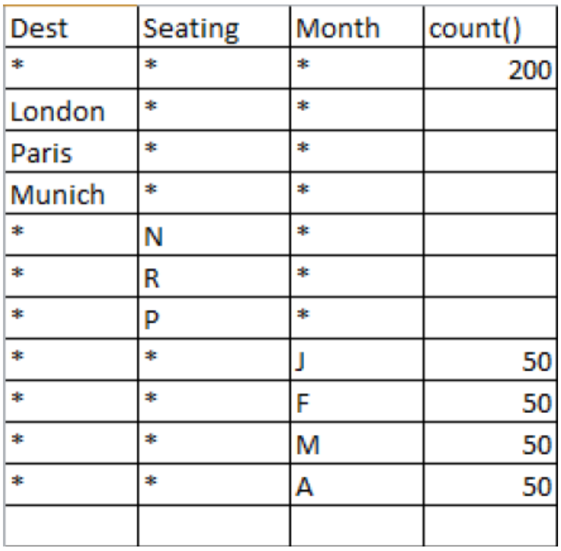
[End of Question 2]

QUESTION 3 (Data Warehousing & OLAP) [TOTAL MARKS: 30]

You are hired to create a data cube for analysis by an international airline containing **Passenger** demographics (passenger data entered when booking a flight), **Flight** data, and **date** with ticket price as the measure. The goal is to analyse the price by bookings, destination and season (dates).

Q 3(a) [8 Marks]
Draw the lattice for the data Cube and additionally, write a separate list of all cuboids.

Q 3(b) [14 Marks]
Take a small dataset with 3 city destinations (London, Paris and Munich), 3 passenger seating requirements (None, Reserved seat, Priority boarding) and date as (Jan,Feb,Mar,Apr). In total, there are 200 flights in the sample.



Dest	Seating	Month	count()
*	*	*	200
London	*	*	
Paris	*	*	
Munich	*	*	
*	N	*	
*	R	*	
*	P	*	
*	*	J	50
*	*	F	50
*	*	M	50
*	*	A	50

A partially complete sample dataset is shown in the table above. Use your own sample data to complete the table, taking care to ensure that all aggregations are correct. You are required to supply the 6 missing counts.

Q 3(c) [8 Marks]

(i) Define an *ancestor* cell. As part of your answer, explain what is a descendant cell.

(ii) Use the sample table from part (b) of this question to provide examples which clearly illustrate the difference between each type of cell.

[End of Question 3]

QUESTION 4 (Association Rule Mining) *[TOTAL MARKS: 30]*

T001	A,C,H
T004	A,B,E,F,H
T005	A,B,C,D
T008	A,B,C,E

The above table shows 4 transactions, each as a set of items in a shopping basket. For this set of transactions, minimum support, **minsup** is 50% and minimum confidence, **minconf** is 60%.

Q 4(a) [6 Marks]
List all frequent itemsets together with their support.

Q 4(b) [12 Marks]

(i) List those itemsets from part 4a) that are **closed**.

(ii) List those itemsets that are **maximal**.

(iii) For all frequent itemsets of maximal length, list all corresponding association rules (ie. including subsets) satisfying the requirements for minimum support and minimum confidence together with their confidence. (You are being asked to list each rule and confidence measure)

Q 4(c) [12 Marks]
Compute lift for every association rule you provided in 4(b) part iii.

[End of Question 4]

[END OF EXAM]