Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

AUGUST/RESIT EXAMINATIONS 2021/2022

MODULE: CA270 - Data Warehousing and OLAP

PROGRAMME(S):
DS BSc in Data Science

YEAR OF STUDY: 2

EXAMINER(S):
Mark Roantree (Internal) (Ext:5636)

TIME ALLOWED: 2 Hours

INSTRUCTIONS: **Answer 3 questions. All questions carry equal marks.**

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 (Data Warehousing) *[TOTAL MARKS: 40]*

Q 1(a) **[9 Marks]**

What is Inmon's definition of a Data Warehouse?
For each characteristic, explain why these are NOT supported in typical database (OLTP) systems.

Q 1(b) **[9 Marks]**

What is meant by an "ETL" architecture or system?
Name and explain each component. Be sure to describe the role or function of each component.

Q 1(c) **[12 Marks]**

Consider an airline company like Ryanair and an analysis on the number of seats occupied on each of their planes.
Their data warehouse is based on a multidimensional data model which views data in the form of a data cube.

Identify 4 dimensions upon which they could perform their analyses.
Draw a lattice to represent the entire set of cuboids within the data cube.

Q 1(d) **[10 Marks]**

Draw a simple star schema (3 columns for each dimension) to represent the warehouse schema.

[End of Question 1]

QUESTION 2 (Classification) *[TOTAL MARKS: 30]*

Q 2(a) **[7 Marks]**

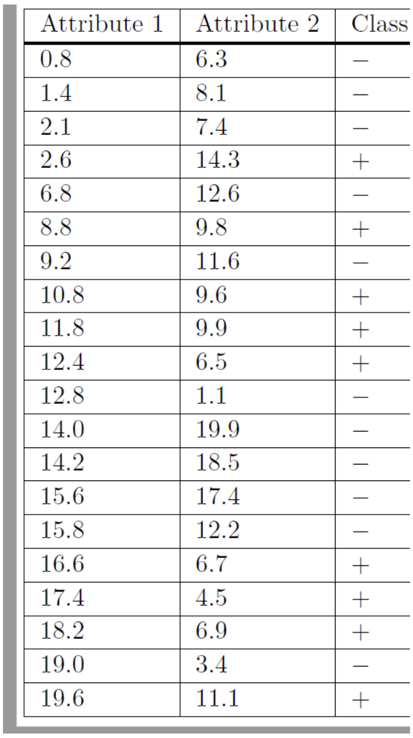
Describe nearest neighbour (*k*-NN) classification in terms of how the method is used to classify unseen instances. Explain the purpose of *k* in your answer.

Q 2(b) **[8 Marks]**

What is the purpose of a distance function in *k*-NN classification and for the data in table 2, what distance function would you use?

Q 2(c) **[15 Marks]**

Using a 7-NN classifier, classify the unseen instance (12.5, 17.5).
Be clear to show precisely how the algorithm makes its prediction.



Attribute 1	Attribute 2	Class
0.8	6.3	−
1.4	8.1	−
2.1	7.4	−
2.6	14.3	+
6.8	12.6	−
8.8	9.8	+
9.2	11.6	−
10.8	9.6	+
11.8	9.9	+
12.4	6.5	+
12.8	1.1	−
14.0	19.9	−
14.2	18.5	−
15.6	17.4	−
15.8	12.2	−
16.6	6.7	+
17.4	4.5	+
18.2	6.9	+
19.0	3.4	−
19.6	11.1	+

Table 2: k-nearest neighbour data

[End of Question 2]

QUESTION 3 (Association Rule Mining) **[TOTAL MARKS: 30]**

Table 3 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%.

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

Table 3. Shopping Basket Transactions

Q 3(a) **[6 Marks]**

List all frequent itemsets together with their support.

Q 3(b) **[12 Marks]**

i. List those itemsets from part 3a) that are **closed**.

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

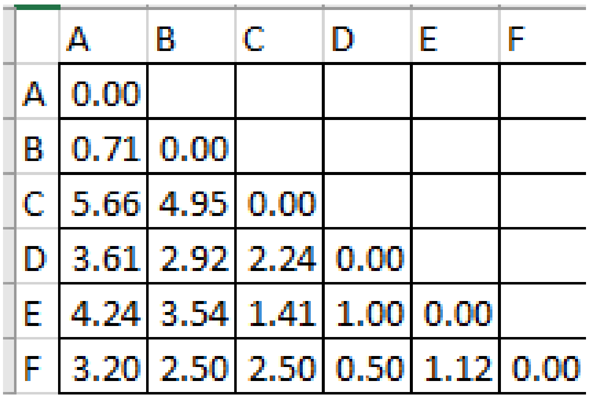
iii. For all frequent itemsets of maximal length, list all corresponding association rules (ie. including subsets) satisfying the requirements for *minimum support* and *minimum confidence* together with their confidence. (You are being asked to list *each* rule and confidence measure)

Q 3(c) **[12 Marks]**

Compute lift for *every* association rule you provided in 3(b) part iii.

[End of Question 3]

QUESTION 4 (Hierarchical Clustering) ***[TOTAL MARKS: 30]***



	A	B	C	D	E	F
A	0.00					
B	0.71	0.00				
C	5.66	4.95	0.00			
D	3.61	2.92	2.24	0.00		
E	4.24	3.54	1.41	1.00	0.00	
F	3.20	2.50	2.50	0.50	1.12	0.00

Table 4: Distance Measures

Q 4(a) **[4 marks]**

What is the difference between a Data Matrix and a Dissimilarity Matrix?

Q 4(b) **[8 marks]**

What 2 Distance functions could be used for the data in Table 4? Describe how both of these functions perform their calculations.

Q 4(a) **[18 Marks]**

Cluster the 6 points A,B,C,D,E,F in table 4 using an Agglomerative Hierarchical Clustering approach. At each step, show the current state of the graph and the new matrix of distance measures.

[End of Question 4]

[END OF EXAM]