Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

SEMESTER 1 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:5404)
Dr. Ziqi Zhang	(External)	External

TIME ALLOWED: 2 Hours

INSTRUCTIONS: **Answer 4 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 (Association Rule Mining) *[TOTAL MARKS: 40]*

This question uses the dataset in **Appendix C** which is a set of transactions, each with a selection from 5 items that were purchased together.

Q 1(a) [6 Marks]

If `min_sup` is set at 0.2 (which is a support *count* of 2), which of the 5 items is supported? In your answer, show the calculations for support for each item.

Q 1(b) [8 Marks]

What are the 2 main steps in the Apriori-Gen algorithm?
Describe the role of both steps in generating *Candidate* Itemsets and *Supported* Itemsets.

Q 1(c) [14 Marks]

In applying the Apriori algorithm, you have generated already L_1 using part (a) of this question. Now generate C_2 (candidate itemsets of size 2) and L_2 (supported itemsets).

Write C_2 and L_2 into your answer book, together with the logic used to derive L_2 from C_2.

Q 1(d) [6 Marks]

Generate C_3, the candidate itemsets of size 3.

Q 1(e) [6 Marks]

Are either of the following rules supported? Explain why.

(i) I1,I2→I3
(ii) I5,I2→I1

[End of Question 1]

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

This question uses the dataset in **Appendix A** which shows a set of results for Covid-19 tests. The attributes captured are Gender, Age, Setting (where they believed the exposure took place), TripsMade (in the last 5 days), and Home (their living arrangements).

Using a **Bayes Classification** approach, answer the following questions.

Q 2(a) [6 Marks]

Which columns require transformation before you can begin? Be precise as to why each transformation is required and *precisely* what you plan to do.
Create and write the transformed dataset in your answer book (copy all columns).

Q 2(b) [6 Marks]
What is meant by *prior probability*?
Calculate the prior probability for each of the classes in this dataset.
What is the weakness with making predictions using only prior probability?

Q 2(c) [14 Marks]
What is meant by conditional or *posterior probability*?
Create a table showing the full set of conditional probabilities.

Q 2(d) [4 Marks]
What is the result of testing the new instance: {F, 45, Sports, 3, Alone}.

[End of Question 2]

QUESTION 3 (Clustering) [TOTAL MARKS: 30]

This question uses the dataset in **Appendix B** which represents values from 3 different temperature sensors in the range 0 to 50, at 8 different locations A to H.

Using a *k*-means clustering (where *k*=3), the task is to determine which set of sensors are most similar. The initial centroid selections are A, B and C.

Q 3(a) [10 Marks]

(i) Using a Euclidean distance function, compute the distance from each location to the 3 centroid locations.
(ii) Which centroid has the largest cluster?
(iii) Which centroid has the smallest cluster?

Q 3(b) [6 Marks]
Using the initial cluster allocation, compute the new centroids.

Q 3(c) [4 Marks]
After 1 further iteration, what if any, are the changes to clusters?

Q 3(d) [10 Marks]
Let us assume a situation where the accuracy of the objective function is poor and you believe this is caused by the high values for sensor t3.

(i) What function would you use to normalise these values? Explain how the function works (performs its calculations).

(ii) Recompute the values for t3 so they are on a normalised scale.

[End of Question 3]

QUESTION 4 (Data Warehousing & OLAP) *[TOTAL MARKS: 30]*

Q 4(a) [8 Marks]

(i) In Data Warehousing, what is considered to be the primary goal of data integration? What is the task required to achieve this goal?

(ii) What is the benefit to Summarised Data in a Warehouse?

(iii) How is it represented (or stored) and what are the typical SQL commands which create these summaries?

(iv) Provide 5 examples of metadata as found in a Data Warehouse.

Q 4(b) [8 Marks]

Consider a data warehouse in operation for a national airline.

(i) Construct a 4-dimensional star schema useful for queries or analyses related to flight capacities (seat availability). Explain what each dimension captures.

(ii) Explain the structure of what you have created in terms of components and connectivity.

Q 4(c) [6 Marks]

In plain language/text:

(i) Provide an example of a 1-dimensional query from your schema in 5(b).

(ii) Provide an example of a 4-dimensional query from the same schema.

Q 4(d) [8 Marks]

Construct a Data Warehouse Bus Matrix which corresponds to your star schema in *part (b)* and addresses the requirements your provided in *part (c).*

[End of Question 4]

Appendix A

Gender	Age	Setting	TripsMade	Home	Result
F	18	Pub	1	Family	Positive
F	18	College	1	Shared	Negative
F	22	Sports	2	Alone	Negative
F	23	Concert	2	Family	Positive
F	23	Pub	3	Shared	Positive
M	31	College	3	Alone	Negative
M	14	Sports	4	Family	Positive
M	67	Concert	4	Shared	Negative
M	81	Pub	4	Alone	Negative
F	66	College	4	Family	Negative
F	50	Sports	4	Shared	Positive
F	45	Concert	5	Alone	Negative
M	42	Sports	5	Family	Negative
F	51	Concert	5	Shared	Negative
M	66	Pub	5	Alone	Negative
M	70	Pub	5	Family	Negative
F	35	Pub	6	Family	Negative
M	34	Pub	6	Family	Positive
M	33	College	7	Family	Positive
F	20	Sports	8	Alone	Positive
F	19	Sports	9	Shared	Positive

Dataset (D): Covid-19 test results

Appendix B

	t1	t2	t3
A	12	13	48
B	14	14	22
C	15	19	41
D	29	14	47
E	16	13	45
F	16	19	43
G	16	12	44
H	15	11	39

Dataset (D): Temperature values (for sensors t1,t2,t3) at locations A-H.

Appendix C

TID	List
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

Dataset D with itemlist I1, I2,I3,I4,I5.

[END OF EXAM]