Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

SEMESTER 1 EXAMINATIONS 2023/2024

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 Question 1 (40 marks) and 2 (30 marks) other questions.**

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 (Cluster Analysis) **[TOTAL MARKS: 40]**

TABLE 1: Test Variables

COL1	COL2	COL3	COL4	COL5	COL6	COL7	COL8	COL9
TRUE	10	P	TRUE	15	HYP1	HEALTHY	10	P
FALSE	55	N	FALSE	40	HYP2	DIAG1	55	N
					HYP3	DIAG2		

4 patients are tested using 9 different testing protocols. Test labels are recorded as columns 1 to 9 (Col1-Col9) where: columns 2,5 and 8 have continuous scores; columns 1,3,4 and 9 have Boolean values; and columns 6 and 7 are categorial values. Table 1 shows the number of categorical variables as 3 for both columns 6 and 7, and the ranges for the continuous variables. As the continuous variables are similar, assume that no form of normalisation is required for these values.

TABLE 2: Patient Test Results

Patient_A	TRUE	35	N	TRUE	22	HYP2	DIAG1	22	N
Patient_B	TRUE	16	P	FALSE	23	HYP2	DIAG2	55	N
Patient_C	FALSE	17	P	FALSE	24	HYP1	HEALTHY	31	P
Patient_D	FALSE	32	N	FALSE	25	HYP3	HEALTHY	32	P

Table 2 shows the test results for 4 anonymous patients listed as A,B,C and D. Assume the binary variables are treated as asymmetric. You are required to determine which 2 patients are most likely to have the same medical condition. Answer parts a) to d) to demonstrate your approach to this problem.

Q 1(a) *[10 Marks]*
Using only continuous variables, which 2 patients are most similar in terms of their medical tests?
Use the manhattan function and explain how you compute the distance scores.

Q 1(b) *[10 Marks]*
Using only boolean variables, which 2 patients are most similar in terms of their medical tests?
Explain the distance function you used in this part of the question and show the results.

Q 1(c) *[10 Marks]*
Using only discrete variables, which 2 patients are most similar in terms of their medical tests?
Explain the distance function you used in this part and show the results.

Q 1(d) *[10 Marks]*
When comparing the entire record of all patients, which 2 are most similar and which two are least similar?
How did you compute the final result?

[End of Question 1]

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

TABLE 3: Customer Loans and Defaults

Customer ID	Loan	Age	Years-In-Job	Default
1	25000	41	2	Y
2	30000	32	4	N
3	22000	21	1	N
4	100000	43	16	N
5	120000	44	8	N
6	33000	50	3	Y
7	25000	27	2	Y
8	25000	53	4	N
9	40000	25	2	Y
10	48000	44	4	N
11	100000	23	2	?

Copy Table 3 to your exam paper and add 3 extra columns: **D1** for question 2a; **D2** for question 2c; and **L-Normal** for question 2b.

Important: Use 2 decimal places in your distance calculations.

Q 2(a) *[12 Marks]*
Using *k*-Nearest Neighbour where *k=3*, calculate the prediction for the new customer with ID=11. In your answer, place all distance computations into **D1** and explain how you came to your decision.

Q 2(b) *[10 Marks]*
Using the *min-max* function, normalise the **Loan** column and write the new values into the new **L-Normal** column.

Q 2(c) *[8 Marks]*
Does the normalised Loan column affect the prediction? Fill column **D2** as part of answering this question.

[End of Question 2]

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

Climate scientists are creating a new experiment using air quality sensors in Dublin City and expect to generate large amounts of data. They require a data warehouse design for their storage to facilitate the extraction of multi-dimensional datasets.

The different dimensions are the **Experiment (3 types); Sensor (3 types); Date (daily);** and **Location (4 locations)** for experiments. The facts are sensors readings (numbers).

Q 3(a) *[5 Marks]*
In data warehousing define what is meant by *time variant*.
For the climate case study, provide an example of data storage that is *not* time variant and then show and explain how a data warehouse facilitates this concept.

Q 3(b) *[5 Marks]*
Identify 5 types of metadata (what does each type describe?) generally found in a data warehouse.

Q 3(c) *[10 Marks]*
Draw a lattice which represents all cuboids for this data warehouse.

Q 3(d) *[10 Marks]*
How does a fact constellation differ from a star schema? Which would you use for this type of application (and explain why)?
Construct a schema for the climate application using a small number of attributes for each dimension. Explain the usage of keys in your design.

[End of Question 3]

QUESTION 4 (Classification) **[TOTAL MARKS: 25]**

Table 4: Golfers and Observed Decisions to Play

Outlook	Temperature	Humidity	Wind	Play
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

Q 4(a) *[8 Marks]*
Using **Information Gain**, calculate the expected information required to classify a tuple in the dataset in Table 4.

Q 4(b) *[8 Marks]*
How much more information is required if we partition on *Outlook*?

Q 4(c) *[8 Marks]*
How much more information is required if we partition on *Wind*?
Between Outlook and Wind, which should be chosen as the partitioning variable?

Q 4(d) *[6 Marks]*
How much more information is required if we partition on *Humidity*?
How does a partition on Humidity rate? Why do you think this is?

[End of Question 4]

[END OF EXAM]