DCU

DUBLIN CITY UNIVERSITY

SEMESTER 1 EXAMINATIONS 2017/2018

MODULE: CA4010 - Data Warehousing and Data Mining

PROGRAMME(S):

CASE - BSc in Computer Applications (Sft.Eng.)
ECSAO - Study Abroad (Engineering and Computing)
CAPT - PhD Track

YEAR OF STUDY: 4,O,X

EXAMINERS: Dr. Mark Roantree (Ph:5636)
Dr. Hitesh Tewari

TIME ALLOWED: 3 hours

INSTRUCTIONS: Answer 4 questions. All questions carry equal marks.

PLEASE DO NOT TURN OVER THIS PAGE UNTIL 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.

Requirements for this paper (Please mark (X) as appropriate)

□ *Log Tables*
□ *Graph Paper*
□ *Dictionaries*
□ *Statistical Tables*

□ *Thermodynamic Tables*
□ *Actuarial Tables*
□ *MCQ Only - Do not publish*
□ *Attached Answer Sheet*

QUESTION 1 *[Total marks: 25]*

Clustering

1(a) [4 Marks]

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

1(b) [6 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 cases is it necessary to use this function?
iii) Write the function for *asymmetric dissimilarity*. Why might we use the asymmetric dissimilarity function? Provide an example in your answer.

1(c) [15 Marks]

Given a data set with 1-dimensional points D = $\{$1,3,6,10,20,100$\}$, consider 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 why each point is assigned to its respective medoid.
(iii) Assume the first step of the first iteration of the k-medoid algorithm tests whether 1 could be replaced with a new medoid with value 3.
What is the result of the testing in this step? Show the calculations needed to arrive at your answer.

[End Question 1]

QUESTION 2 *[Total marks: 25]*

Classification

2(a) [5 Marks]

Define the term *Prior Probability*.
Calculate the *Prior Probability* for all 3 classes (Hard, Soft, None) for the Lens dataset in figure 1.

2(b) [12 Marks]

(i) Define the term *Conditional Probability*.
(ii) Calculate every conditional probability for all 3 classes and use a matrix to display these values.
(iii) What is the probability of class C=”hard contacts” given:
age=1, specRx=1, tears =1, and astig = 2.

Value of attribute				Class
age	specRx	astig	tears	
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

classes
1: hard contact lenses
2: soft contact lenses
3: no contact lenses

age
1: young
2: pre-presbyopic
3: presbyopic

specRx
(spectacle prescription)
1: myopia
2: high hypermetropia

astig
(whether astigmatic)
1: no
2: yes

tears
(tear production rate)
1: reduced
2: normal

Figure 1: Lens Dataset for Q2

2(c) [8 Marks]

(i) Construct the decision tree for a *takefirst* algorithm.
(ii) Do you think this might lead to a good decision tree? Explain your answer.

[End Question 2]

QUESTION 3 ***[Total marks: 25]***

Cube Computation

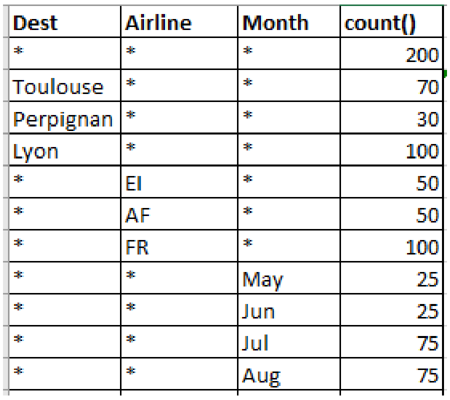
You are a data scientist working at Dublin airport and have access to all the data captured when a boarding card is scanned. This includes airline, destination and date. Your lowest level of granularity is ”flight” and thus, your measure is the number of passengers per flight. Your goal is to analyse passengers by airline, destination and season (dates).

3(a) [6 Marks]

Draw the lattice for the data Cube. Also, separately list all of the cuboids.

3(b) [6 Marks]

What is meant by an *ancestor* cell? As part of your answer, describe a *descendant* cell in relation to an ancestor cell.
Provide examples which clearly illustrate the difference between both types of cell. Use the data from 3(c) in your answer if you wish.



Dest	Airline	Month	count()
*	*	*	200
Toulouse	*	*	70
Perpignan	*	*	30
Lyon	*	*	100
*	EI	*	50
*	AF	*	50
*	FR	*	100
*	*	May	25
*	*	Jun	25
*	*	Jul	75
*	*	Aug	75

Figure 2: Boarding Card Dataset

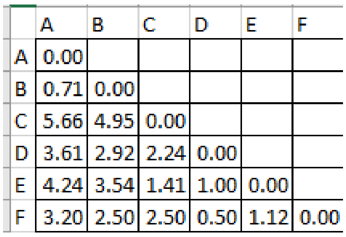
3(c) [13 Marks]

Assume we have a small dataset with 3 destinations in France: Toulouse, Perpignan and Lyon; 3 airlines: Aer Lingus (EI), Air France (AF) and Ryanair (FR); and date as (May,Jun,Jul,Aug). All 3 airlines can fly to all 3 destinations.
In total, we have 200 flights in the sample with a partially complete dataset shown in figure 2, containing all 1-D cuboids.
(i) Write the dimensional values for all 3-D cuboids (Hint: there will be 36 in total). Your answer should not include any measures.
(ii) Provide sample measures for all 3-D cuboids for Toulouse. Hint: You must ensure that the aggregates are valid.

[End Question 3]

QUESTION 4 ***[Total marks: 25]***

Clustering



	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

Figure 3: Similarity Matrix (Q4)

4(a) [15 Marks]

Cluster the 6 points A,B,C,D,E,F in figure 3 using an Agglomerative Hierarchical Clustering approach. At each step, show the current state of the graph and the new similarity matrix.

4(b) [10 Marks]

You are testing a new machine which records the speed of cars outside the university. Your job is to detect outlier values recorded by the device. The machine is considered to have 90% accuracy and it is suspected that 10% of the values will be outliers, or very unlikely to be correct.

(i) In terms of normal distribution, describe what working hypothesis you use. Briefly discuss an appropriate discordancy test and significance probability function.
(ii) Describe the different approaches taken in block procedures and sequential procedures. Use examples from the speed dataset in giving your answer.
In general, which approach is better?

[End Question 4]

QUESTION 5 ***[Total marks: 25]***

Data Warehousing

5(a) [13 Marks]

A clothing store requires a data mart to monitor to effect of a series of promotions (discounts) for products. They have full demographic information (gender, age, address) for most of their customers (through their loyalty card scheme); a database of all products; and all transactions at point of sales. They are keen to understand:

- Volume of sales by gender, by product brand
- A comparison of products before and during the promotion
- Most popular products during the promotion

Construct a single data mart (star schema) to meet these requirements. In your answer, draw your multi-dimensional schema and comment on the role of *every* attribute in dimension and fact tables.

5(b) [8 Marks]

Conformance of dimensions is regarded as a critical feature of warehouse development. Provide an example of both *conforming by rollup* and *conformance by subsets*. Be sure to explain how conformance is achieved in each case. Use the case study from part 5(a) to provide examples.

5(c) [4 Marks]

Explain the difference between an *independent* data mart and a *dependent* data mart? Give an example of each from the case study in 5(a).

[End Question 5]

[END OF EXAM]