DUBLIN CITY UNIVERSITY

AUGUST/RESIT 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*	☐	*Thermodynamic Tables*
☐	*Graph Paper*	☐	*Actuarial Tables*
☐	*Dictionaries*	☐	*MCQ Only - Do not publish*
☐	*Statistical Tables*	☐	*Attached Answer Sheet*

QUESTION 1 ***[Total marks: 25]***

Cluster Analysis

Use the *k*-means algorithm and Euclidean distance to cluster the following 8 instances into 3 clusters: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9). The distance matrix based on the Euclidean distance is provided in figure 1.

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

Figure 1: Euclidean Distance Matrix

1(a) [10 Marks]

Assume that the initial seeds (centers of each cluster) are A1, A4 and A7 and you run the *k*-means algorithm for the first iteration. At the end of this iteration, show the new clusters (i.e. the instances that belong to each cluster)

1(b) [5 Marks]

Calculate the centers of the new clusters.

1(c) [10 Marks]

Calculate the number of iterations that are needed for the process to terminate. In order to do this, you should compute the required similarity matrices to clearly show the clusters and centers for *every* iteration.

[End Question 1]

QUESTION 2 ***[Total marks: 25]***

Association Rule Mining

The dataset in figure 2 has five transactions with min_sup = 60% and min_conf = 80%.

2(a) [17 Marks]

Find all frequent itemsets using the Apriori algorithm. Be sure to show the output datasets from each step.

TID	*items_bought*
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y }
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I ,E}

Figure 2: Sample Transactions

2(b) [8 Marks]

List all of the strong association rules (with support *s* and confidence *c*) matching the following metarule *a,b* $\rightarrow$ *c*. In other words, those rules with 2 elements on the left hand side and a single right hand side element.

[End Question 2]

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

Cube Computation

DCU use a data cube to meet the following requirements:
1. They need to determine which counties in Ireland provide students with the highest points from the leaving certificate.
2. They require an analysis of modules which score the highest Firsts and those with the highest number of Fails.
3. They need to analyse module grades over a number of years.

3(a) [6 Marks]

Selecting appropriate dimensions, draw the lattice for the data Cube and provide a separate list of all cuboids.

3(b) [13 Marks]

Assume we are looking at a small dataset with 350 students. Using 3 dimensions where each dimension has just 3 values, list all 1-D and 2-D instances where all have a measure value > 0; and list the 3-D cuboids for one of the dimensions. Your goal is to ensure that all aggregations are correct.

3(c) [6 Marks]

Define an *ancestor* cell. As part of your answer, define a *descendant* cell in relation to an ancestor cell.
Use the sample table from part (b) of this question to provide examples which clearly illustrate the difference between each type of cell.

[End Question 3]

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

Warehousing and OLAP

Dublin airport are undertaking a major study to understand the movement of passengers out of Dublin airport and are using the scans of every boarding card as the source of data for their study. In addition, they have been provided with some passenger details for each boarding card (town or city, county, gender, age) for each customer who's boarding card is scanned. The analyses require the following datasets:

- Destination across dates by gender
- Preferred airline by destination by country of residence
- Most popular days for departure to specific destinations, based on the age of the traveller.

4(a) [5 Marks]

Describe what is meant by a Star Schema. Use a diagram to show a sample star schema

4(b) [14 Marks]

Construct the warehouse with the appropriate data mart(s) to meet the information requirements for analysts. In your answer, draw your multi-dimensional schema and comment on the role of *every* attribute in dimension and fact tables.

4(c) [6 Marks]

Provide a definition of **Drill-down**, **pivot** and **slice** functions. Use sample data from the Dublin airport case study to help explain your answer.

[End Question 4]

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

Classification

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 3: Lens Dataset

5(a) [7 Marks]

Explain the term *Information Gain* with respect to decision tree induction. In your answer, describe the method behind both algorithms (formulae) used in the process.

5(b) [18 Marks]

Using *Information Gain*, rank the four attributes in figure 3 (age, specRx, astig, tears) for selection for the *first* split only, from best to worst. In other words, you are required to provide the initial classification only.

[End Question 5]

[END OF EXAM]