DUBLIN CITY UNIVERSITY

SEMESTER 1 EXAMINATIONS 2015/2016

MODULE: CA4010 - Data Warehousing and Data Mining

PROGRAMME(S):

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

YEAR OF STUDY: 1,4,X

EXAMINERS: Dr Mark Roantree (Ph:5636)

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☐	*Dictionaries*	☐	*MCQ Only - Do not publish*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

Assume you are using 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 examples belonging to each cluster)

1(b) [5 Marks]

Calculate the centers of the new clusters.

1(c) [10 Marks]

How many more iterations are needed for the process to terminate? As part of your answer, compute the required similarity matrices to illustrate the clusters and centers for each iteration.

[End Question 1]

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

Association Rule Mining

The database shown in table below has five transactions. Let 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}

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]***

Classification

3(a) [5 Marks]

What is the definition of Prior Probability?

Calculate the Prior Probability for each of the 3 classes for the Lens dataset.

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

3(b) [14 Marks]

i. What is meant by Conditional Probability?

ii. Construct a table and calculate the conditional probability for each attribute, for each class. For example, the probability that *class = C1, given age = 1*, might be your first entry in that table.

iii. If we are only interested in Class C1 and focus solely on the attribute tears, what prediction can we make?

iv. What is the probability of class C3 given the set of values $\{2,2,2,2\}$ in the order in which attributes appear in the dataset?

3(c) [6 Marks]

Construct the decision tree for a *takefirst* algorithm.

[End Question 3]

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

Warehousing and OLAP

DCU are building a data warehouse to meet the following requirements:
1. They need to determine which counties in Ireland provide students with the highest points from the leaving certificate (High-Points-Req);
2. They require an analysis of modules which score the highest Firsts and those with the highest number of Fails (First-Fails-Req);
3. They want to schedule rooms with the optimum number of students, based on the last 5 years of trends in popular modules (Popular-Modules-Req).

4(a) [3 Marks]

Describe what is meant by a Star Schema.

4(b) [10 Marks]

Draw and describe the fact tables you would use construct to meet DCU's requirements. Comment on the grain in each case.

4(c) [6 Marks]

Draw and describe 3 of the dimensions you would employ.
Build your schema in the form of a fact constellation.

4(d) [6 Marks]

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

[End Question 4]

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

Cluster Analysis

5(a) [7 Marks]

What is the main idea behind Agglomerative Hierarchical Clustering? Describe each step in the AHC algorithm.

5(b) [18 Marks]

Use *single link* agglomerative clustering to group the data described by the following distance matrix. Show the formation of the dendrogram in your answer.

	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

How would *complete link* clustering differ from this approach?

[End Question 5]

[END OF EXAM]