# DUBLIN CITY UNIVERSITY

## AUGUST/RESIT 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)
Dr Ian Pitt

**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.

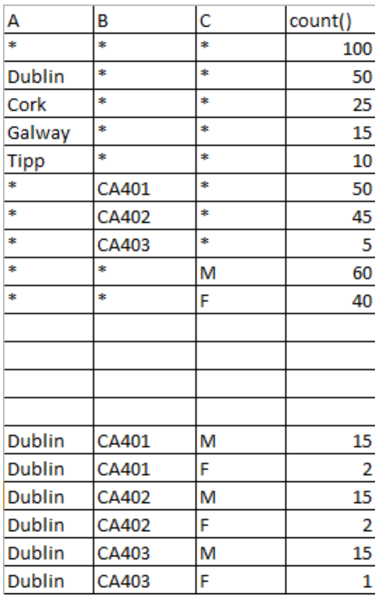
*Requirements for this paper (Please mark (X) as appropriate)*

- □ *Log Tables*
- □ *Graph Paper*
- □ *Dictionaries*
- □ *Statistical Tables*
- □ *Thermodynamic Tables*
- □ *Actuarial Tables*
- □ *MCQ Only - Do not publish*
- □ *Attached Answer Sheet*

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

*Cube Computation*

Assume you are required to create a data cube for analysis on DCU student data containing Student address (county), Module code and gender with student count as the measure. Its purpose is to analyse the numbers of students taking one of three *optional* modules.



| A | B | C | count() |
|---|---|---|---|
| * | * | * | 100 |
| Dublin | * | * | 50 |
| Cork | * | * | 25 |
| Galway | * | * | 15 |
| Tipp | * | * | 10 |
| * | CA401 | * | 50 |
| * | CA402 | * | 45 |
| * | CA403 | * | 5 |
| * | * | M | 60 |
| * | * | F | 40 |
| | | | |
| | | | |
| | | | |
| | | | |
| Dublin | CA401 | M | 15 |
| Dublin | CA401 | F | 2 |
| Dublin | CA402 | M | 15 |
| Dublin | CA402 | F | 2 |
| Dublin | CA403 | M | 15 |
| Dublin | CA403 | F | 1 |

Figure 1: Sample cuboids for module analysis

1(a) [6 Marks]

Draw the lattice for this Cube and also write down the list of all cuboids.

1(b) [11 Marks]

Assume we are looking at a dataset with 3 modules (CA401, CA402 and CA403), 4 counties where students live (Cork, Dublin, Galway, Tipperary) and Gender as (M,F). In total we have 100 students in the sample and students are allowed to take **only one** of the 3 modules;

A partially complete sample table is shown in Fig. 1. Use your own sample data to complete the table, taking care to ensure that **all aggregations are correct**.

(Hint: While copying the table to your answer sheet, leave the 3D cuboids for Dublin until last as you do not know how many cuboids you will need to complete the Cube).

1(c) [8 Marks]

Define an *ancestor* cell. Also define a *descendant* cell. Use the sample table from part (b) of this question to provide examples which clearly illustrate how these definitions are formed.

***[End Question 1]***

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

*Classification*

2(a) [7 Marks]

Explain what is meant by a *Dendogram* and then describe the basic algorithm for Agglomerative Hierarchical Clustering.

2(b) [8 Marks]

Using a small number of objects (10 or 11 for example), show how a Dendogram is constructed and the steps involved in reaching the final cluster. Your answer should contain an illustration of the final Dendogram.

2(c) [10 Marks]

(i) Using the *k*-Means algorithm, the 16 points represented by (x,y) in Table 1 have been clustered using the initial centroids C1:(2.2, 5.5), C2:(7.8, 12.2), and C3:(8.2, 12.1). Distances *d1*, *d2*, and *d3* have been calculated for centroids *C1*, *C2* and *C3* respectively. Compute the missing distances for all points from centroid C3 and make your cluster assignments.

For your answer, copy the table into your exam paper and complete the missing distance values and cluster assignments.

(ii) Finally, calculate the new values for all three centroids.

Table 1: Distances d1,d2,d3 from Centroids C1, C2, C3

| x | y | d1 | d2 | d3 | cluster |
|---|---|---|---|---|---|
| 6.8 | 12.6 | 8.5 | 1.1 | | |
| 0.8 | 9.8 | 4.5 | 7.4 | | |
| 1.2 | 11.6 | 6.2 | 6.6 | | |
| 2.8 | 9.6 | 4.1 | 5.6 | | |
| 3.8 | 9.9 | 4.7 | 4.6 | | |
| 4.4 | 6.5 | 2.4 | 6.6 | | |
| 4.8 | 1.1 | 5.1 | 11.5 | | |
| 6.0 | 19.9 | 14.9 | 7.9 | | |
| 6.2 | 18.5 | 13.6 | 6.5 | | |
| 7.6 | 17.4 | 13.1 | 5.2 | | |
| 7.8 | 12.2 | 8.7 | 0 | | |
| 5.8 | 7.7 | 4.2 | 4.9 | | |
| 8.2 | 4.5 | 6.1 | 7.7 | | |
| 8.4 | 6.9 | 6.4 | 5.3 | | |
| 9.0 | 3.4 | 7.1 | 8.9 | | |
| 9.6 | 11.1 | 9.3 | 2.1 | | |

***[End Question 2]***

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

*Association Rule Mining*

3(a) [8 Marks]

A survey asked university students to list their hobbies from the following set: Cinema (cin), Music listening (mus), Piano (pia), Guitar (gui), photography (pho), theatre (the), books (boo), football (foo), athletics (ath), and chess (che).
Thus, we have itemset *I* = {ath, boo, che, cin, foo, gui, mus, pho, pia, the}; the number of items *m = 10*; and the number of students in our sample (transactions) *n = 9*..

| Txn | Itemsets |
|---|---|
| 1 | {ath, boo, cin} |
| 2 | {cin, mus, gui, the} |
| 3 | {cin, mus, pho, the} |
| 4 | {che, cin, pho, pia} |
| 5 | {ath, cin, foo, mus, the} |
| 6 | {foo, gui, mus} |
| 7 | {che, cin, foo, mus} |
| 8 | {cin, foo, mus} |
| 9 | {cin, foo, mus, pia, the} |

i. What is the support for {cin, mus}? Explain your answer using the equation for calculating support.

ii. What is meant by the rule {cin,mus} → {foo}?

iii. What is the support for the rule in (ii)? Once again, explain the equation used to calculate this support.

iv. What is the difference between *support* and *confidence*? Calculate confidence for this rule.

3(b) [6 Marks]

(i) What is **minsup** and how is it used in generating rules? Provide an example.
(ii) What is **minconf** and how does it differ from **minsup**? Provide an example.

3(c) [11 Marks]

Assume we have a database with 6500 transactions and a rule L → R with the following support counts:
count(L) = 3400; count(R) = 4000; count(L $\cup$ R) = 3000

a) Calculate support for L → R
b) Calculate confidence for L → R
c) Calculate lift for L → R
d) What does the lift function tell us about a rule?
e) Calculate leverage for L → R

***[End Question 3]***

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

*Classification*

| age | specRx | astig | tears | C |
|---|---|---|---|---|
| 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 |

4(a) [7 Marks]

Explain the term *Information Gain* with respect to decision tree induction.

4(b) [18 Marks]

Using *Information Gain*, rank the four attributes above (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 4]***

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

*Data Warehousing and OLAP*

5(a) [6 Marks]

(i) What is meant by a *Fact Table* in Data Warehousing? In your answer, be clear on what is a *measurement*.
(ii) What is the role of *Dimension Tables*? Describe the relationship between Dimension and Fact tables and explain why Dimension tables are necessary.

5(b) [13 Marks]

Consider a data warehouse for the Harvey Norman store, selling electronic goods where a *star schema* has been used to warehouse and analyse all sales and supplier data. Use a diagram to describe the schema with an appropriate Fact Table and at least three Dimension Tables. In your diagram, be sure to highlight those attributes that are keys.
Also, provide a brief discussion on each table.

5(c) [6 Marks]

(i) What is meant by a Roll-up operation?
(ii) Using your diagram in part 3b), provide 2 examples of Roll-up operations using 2 separate dimensions.
Use your fact table, and incorporate sample values to demonstrate the rollup computations in each case.

***[End Question 5]***

***[END OF EXAM]***