DUBLIN CITY UNIVERSITY

SEMESTER 1 EXAMINATIONS 2016/2017

MODULE: CA4010 - Data Warehousing and Data Mining

PROGRAMME(S):

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

YEAR OF STUDY: 4,O,X

EXAMINERS: 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□ *MCQ Only - Do not publish*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]***

Data Warehousing

1(a) [8 Marks]

Discuss the differences between traditional databases and OLAP-oriented data warehouse systems under the following headings:
i) user types
ii) database design
iii) data
iv) summarisation

In each case, use a real world example of both types of data management systems.

1(b) [7 Marks]

Provide an illustration of an ETL architecture. Ensure you have a description and goal of each layer (or component), and describe what takes place.

1(c) [10 Marks]

i) Explain the difference between an *independent* data mart and a *dependent* data mart? Give an example of each from the real world.
ii) Explain the design concept that is used to manage and control the development of multiple data marts by (possibly separate) teams.
iii) Draw a sample Bus Architecture. Use the organisation/company of your choice and list 4 requirements and show how they appear in the bus architecture and how their overlap is represented.

[End Question 1]

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

Classification
The golf dataset (overleaf) represents whether or not a golfer will decide to play golf, depending on the attributes provided. You are required to use *Information Gain* to determine the *first* attribute on which to branch.

2(a) [17 Marks]

i) Calculate Entropy for the entire dataset.
ii) Calculate the expected entropy *for every attribute*, for the *first* attribute selection.
iii) Show which attribute is selected.
iii) Describe the process for managing continuous variables

2(b) [8 Marks]

i) Why was the *Gain Ratio* approach adopted? In your answer explain how it differs from *Information Gain*.

Outlook	Temp (°F)	Humidity (%)	Windy	Class
sunny	75	70	true	play
sunny	80	90	true	don't play
sunny	85	85	false	don't play
sunny	72	95	false	don't play
sunny	69	70	false	play
overcast	72	90	true	play
overcast	83	78	false	play
overcast	64	65	true	play
overcast	81	75	false	play
rain	71	80	true	don't play
rain	65	70	true	don't play
rain	75	80	false	play
rain	68	80	false	play
rain	70	96	false	play

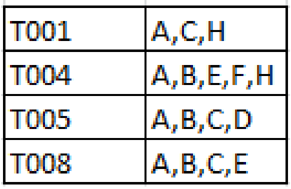
Classes
play, don't play
Outlook
sunny, overcast, rain
Temperature
numerical value
Humidity
numerical value
Windy
true, false

ii) Explain the *SplitInfo* concept, in terms of its approach.
iii) Write and explain the *Gain Ratio* function
iv) What is the difference between decision trees created by Gain Ratio and those created using the *Gini index*?

[End Question 2]

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

Association Rule Mining



T001	A,C,H
T004	A,B,E,F,H
T005	A,B,C,D
T008	A,B,C,E

The above table shows 4 transactions, each with a set of items in a shopping basket. Assume that minimum support, **minsup** = 50% and minimum confidence, **minconf** = 60%.

3(a) [6 Marks]

List all frequent itemsets together with their support.

3(b) [9 Marks]

i) List those itemsets from part a) that are closed.
ii) List those itemsets that are maximal.
iii) For all frequent itemsets of maximal length, list all corresponding association rules (ie. including subsets) satisfying the requirements for minimum support *and* minimum confidence together with their confidence.
In other words, list each rule and confidence measure.

3(c) [10 Marks]

Compute *lift* for every association rule from b)iii).

[End Question 3]

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

Clustering

4(a) [4 Marks]

For interval scaled variables, what function is used to standardise measurement units? Write the function (2 steps) and explain how it works.

4(b) [8 Marks]

i) How can we 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*. Provide an example of where this function might be used.
iii) Write the function for *asymmetric dissimilarity*. Why might we use the asymmetric dissimilarity function? Provide an example in your answer.

4(c) [13 Marks]

i) Describe the 4 cases for object reassignment in the *k*-medoids algorithm (4 marks).
ii) Cluster the following objects:
A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9), using the *k*-medoids algorithm (9 marks).

[End Question 4]

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

Cube Computation

Assume a case study for a university where there are dimensions: **Student** (student demographics such as address, age, gender), **Course** (you choose the attributes for drill down, and **Date**.

5(a) [5 Marks]

Draw the lattice for student grades which clearly shows every dimension and level.

5(b) [4 Marks]

Using your illustration in part (a):
i) Provide an example of a *base* cell.
ii) Provide an example of an *aggregate* cell.
iii) Explain the terms *ancestor* and *descendant*. Provide an example of each, again using the university case study.

5(c) [6 Marks]

i) What is meant by an *Iceberg* Cube? Provide an example of an iceberg cube using the university case study
ii) What is meant by an *Iceberg condition*? Provide an example.

5(d) [10 Marks]

Define the BUC Algorithm for Cube Computation. In your answer:
i) State the input parameters and why they are required.
ii) Describe the output.
iii) List the (pseudo)code for the algorithm.

[End Question 5]

[END OF EXAM]