Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

SEMESTER 1 EXAMINATIONS 2018/2019

MODULE: CA4010 - Data Warehousing & Data Mining

PROGRAMME(S):

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

YEAR OF STUDY: 4,O,X

EXAMINER(S):

Dr. Mark Roantree	(Internal)	Ext:5636
Dr. Hitesh Tewari	(External)	

TIME ALLOWED: 3 Hours

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

PLEASE DO NOT TURN OVER THIS PAGE UNTIL YOU ARE 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.

There are no additional requirements for this paper.

QUESTION 1 **[TOTAL MARKS: 25]**

(Data Warehousing)

This question uses **figure 1** in **Appendix 1**.

Q 1(a) **[4 Marks]**

Describe the schema you see in figure 1. Ensure you explain the overall layout; its components; and how those components are joined.

Q 1(b) **[6 Marks]**

Extend this schema to convert to a Constellation. Draw the new schema and explain what you did.

Q 1(c) **[4 Marks]**

Inman's definition of a data warehouse contains 4 main characteristics. Describe each characteristic in detail, using an example from **figure 1**, where appropriate.

Q 1(d) **[6 Marks]**

Convert the schema in **figure 1** to a lattice. Do not draw the lattice but provide a list of all cuboids that comprise the lattice.

Hint: attributes are not required.

Q 1(e) **[5 Marks]**

Describe (you do *not* need to provide SQL code) a 3-D OLAP query for this schema. Then describe a RollUp operation for your query.

[End of Question1]

QUESTION 2 **[TOTAL MARKS: 25]**

(Classification)

This question uses **figure 2** in **Appendix 2**.

Q 2(a) **[4 Marks]**

How does *classification* differ from *prediction*? Provide an example of each in your answer. How do they differ in terms of determining their accuracy?

Q 2(b) **[6 Marks]**

Using the Degrees dataset (*D*) in **figure 2**, calculate the expected information needed to classify a tuple in *D*. In other words, calculate *Info(D).*

Q 2(c) **[10 Marks]**

There are 5 possible attributes that could be used to split the initial dataset. Calculate the expected information required to classify a tuple after splitting for *every* attribute.

Q 2(d) [5 Marks]

What is meant by *Information Gain*? Which of the attributes in **figure 2** delivers the highest information gain? Explain your answer by showing the information gain for each of the 5 attributes.

[End of Question2]

QUESTION 3 [TOTAL MARKS: 25]

(Association Rule Mining)
This question uses **figure 3** in Appendix 3.

A survey asked a group of 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).
In **figure 3**, we can see the **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, **n = 9**.

Q 3(a) [8 Marks]

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 the rule {cin,mus} → {foo}.

Q 3(b) [6 Marks]

i. What is **minsup** and how is it used for the generation of rules? Provide an example rule in your answer.

ii. What is **minconf** and how does it differ from minsup? Provide an example rule in your answer.

Q 3(c) [11 Marks]

Assume we have a database with 6500 transactions and a rule L → R with the following: count(L) = 3400; count(R) = 4000; count(L ∪ 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 of Question3]

QUESTION 4 **[TOTAL MARKS: 25]**

(Clustering)
This question does not use the Appendix.

Q 4(a) **[5 Marks]**

Describe the steps in the *Partitioning Around Medoids* algorithm. Be clear as to the inputs and outputs of the algorithm.

Q 4(b) **[12 Marks]**

Cluster the following objects using the *k*-medoids algorithm, where *k*=2 and the initial representative objects are **A1** and **A8**.
Show the cost (absolute error) of each cluster.

A1 = (2,10)
A2 = (2,5)
A3 = (8,4)
A4 = (5,8)
A5 = (7,5)
A6 = (6,4)
A7 = (1,2)
A8 = (4,9)

Q 4(c) **[8 Marks]**

Assume the next step is to replace **A8** with **A4** as the new representative object.

i. How does this affect cluster 1 (centroid=A1)?
ii. How does this affect cluster 2 (new centroid A4)?
iii. Are these clusters better than the previous clusters? Why?

[End of Question4]

QUESTION 5 **[TOTAL MARKS: 25]**

(Classification)

This question uses **figure 2** in **Appendix 2**.

Q 5(a) **[4 Marks]**

Use the Bayesian approach to classification, what is meant by the terms *mutually exclusive* and *exhaustive*? Does the dataset in figure 2 conform to these conditions? Ensure that you explain why the dataset does or does not meet these requirements when giving your answer.

Q 5(b) **[4 Marks]**

Define what is meant by *Prior* Probability. Calculate the prior probability for each of the classes in the dataset.

Q 5(c) **[12 Marks]**

Define what is meant by *Conditional* Probability. Calculate all conditional probabilities for this dataset. Place the values into a table or matrix which clearly shows the conditional probabilities in relation to their classes.

Q 5(d) **[5 Marks]**

Using your matrix of prior and conditional probabilities, determine the classification for the unseen instance **(B,B,A,B,B)** and show how the formula is used in your classification.

[End of Question5]

Appendix 1 (Figure 1)

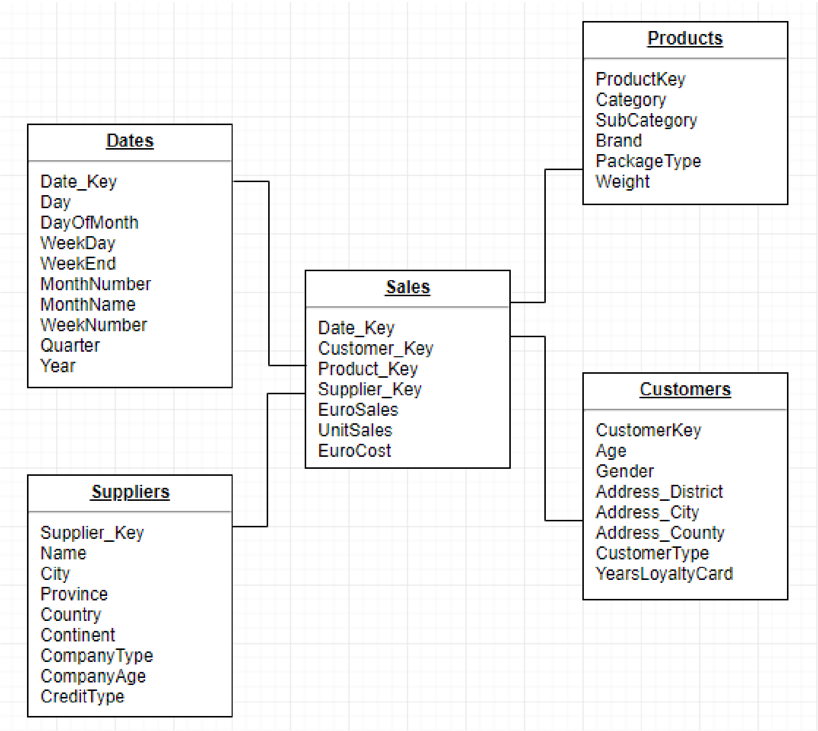
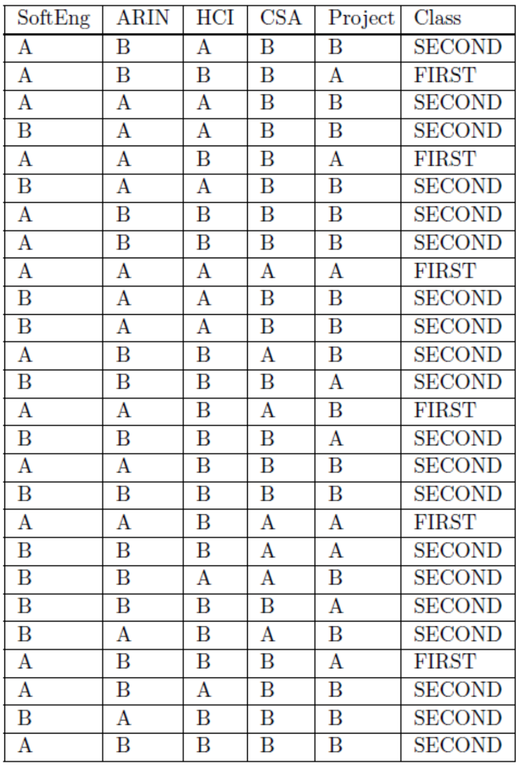


Figure 1: Sales Analysis

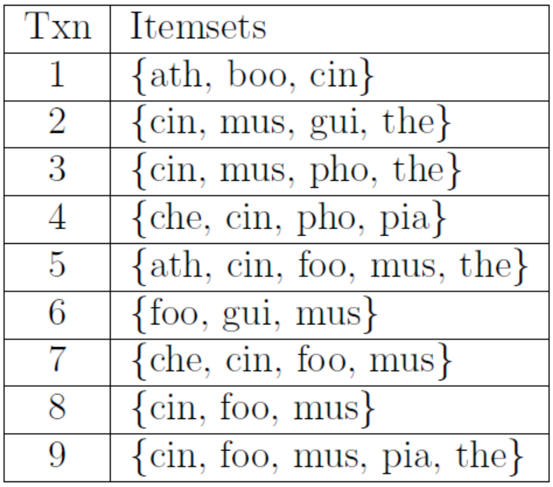
Appendix 2 (Figure 2)



SoftEng	ARIN	HCI	CSA	Project	Class
A	B	A	B	B	SECOND
A	B	B	B	A	FIRST
A	A	A	B	B	SECOND
B	A	A	B	B	SECOND
A	A	B	B	A	FIRST
B	A	A	B	B	SECOND
A	B	B	B	B	SECOND
A	B	B	B	B	SECOND
A	A	A	A	A	FIRST
B	A	A	B	B	SECOND
B	A	A	B	B	SECOND
A	B	B	A	B	SECOND
B	B	B	B	A	SECOND
A	A	B	A	B	FIRST
B	B	B	B	A	SECOND
A	A	B	B	B	SECOND
B	B	B	B	B	SECOND
A	A	B	A	A	FIRST
B	B	B	A	A	SECOND
B	B	A	A	B	SECOND
B	B	B	B	A	SECOND
B	A	B	A	B	SECOND
A	B	B	B	A	FIRST
A	B	A	B	B	SECOND
B	A	B	B	B	SECOND
A	B	B	B	B	SECOND

Figure 2: Degrees dataset D

Appendix 3



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}

Figure 3: Students 1-9 choose from the Hobby Dataset

[END OF EXAM]