Semester Two of Academic Year (2017---2018) of BDIC
《 Data Mining 》
Module Code: COMP3027J
Exam Paper Sample

Exam Instructions：Answer Question 1 (worth maximum of 40 points), and any other two questions (worth maximum of 30 points each).

Total marks available is 100.

Honesty Pledge：

I have read and clearly understand the Examination Rules of Beijing University of Technology and University College Dublin and am aware of the Punishment for Violating the Rules of Beijing University of Technology and University College Dublin. I hereby promise to abide by the relevant rules and regulations by not giving or receiving any help during the exam. If caught violating the rules, I would accept the punishment thereof.

Pledger：_____ Class No：_____

BJUT Student ID：_____ UCD Student ID______

。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。
。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。。

Notes：

The exam paper has 4 questions on 5 pages, with a full score of 100 points. Answer Question 1 (worth maximum of 40 points), and any other two questions (worth maximum of 30 points each). You are required to use the given Examination Book only.

Total score of exam paper (for teacher use only)

Item	**Part 1**	**Part 2**	**Part 3**	**Total Score**
Full Score	**40**	**30**	**30**	
Obtained Score				

Question 1 (40 marks):

a) Explain briefly what is meant by **Unsupervised Learning** and **Supervised Learning**. Give an example of each. (5 marks)

b) Explain briefly **Feature Selection**. Why would you use feature selection?
(5 marks)

c) Briefly explain **Underfitting** and **Overfitting**. Draw a simple plot demonstrating each.
(5 marks)

d) What does it mean when a prediction model is said to **Generalise** well?
(5 marks)

e) List the three types of **data quality issues** and briefly explain each.
(5 marks)

f) What is the **Inductive Bias** of a machine learning algorithm. Give examples.
(5 marks)

g) What is the difference between a **Continuous feature** and a **Categorical feature**? Give examples.
(5 marks)

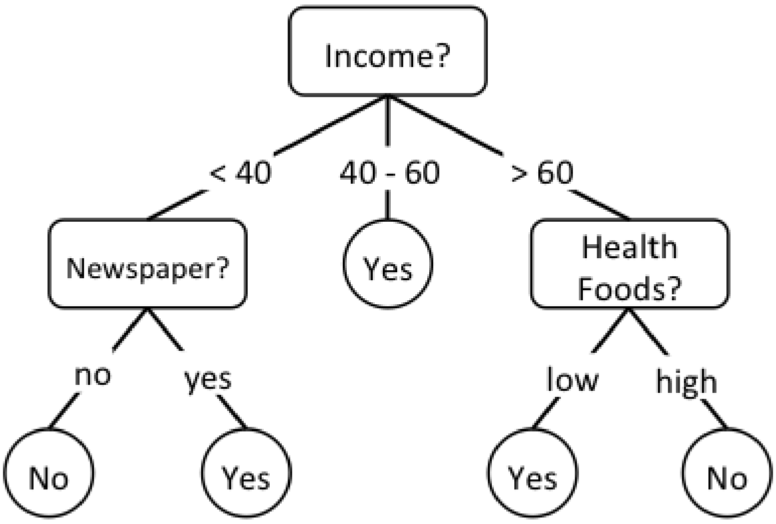
h) What is the difference between a **Raw feature** and a **Derived feature**? Give examples.
(5 marks)

Question 2 (30 marks):

The following table presents a dataset collected by a retail company capturing historical details of which of their customers have responded to promotions the company has run. The information captured covers customer income bracket, customer age, whether or not the customer regularly buys a newspaper, the proportion of health foods typically included in the customer's shopping, and, finally, whether or not they responded to previous promotional mailings.

ID	Income	Age	Newspaper	Health Foods	Respond
C-01	<40	81	no	low	No
C-02	<40	76	no	high	No
C-03	40-60	86	no	low	Yes
C-04	>60	84	no	low	Yes
C-05	>60	45	yes	low	Yes
C-06	>60	66	yes	high	No
C-07	40-60	41	yes	high	Yes
C-08	<40	68	no	low	No
C-09	<40	32	yes	high	Yes
C-10	>60	56	yes	low	Yes
C-11	<40	58	yes	high	Yes
C-12	40-60	52	no	high	Yes
C-13	40-60	90	yes	low	Yes
C-14	>60	69	no	high	No

This dataset has been used to induce a **decision tree** that can predict whether or not new customers will respond to promotional mailings. This decision tree is shown below.



(i) What is **information gain?** Describe the three step process for calculating information gain using the following equations:

(10 marks)

$$H(t,\mathcal{D}) = -\sum_{l \in levels(t)} (P(t=l) \times log_2(P(t=l)))$$

$$rem(d,\mathcal{D}) = \sum_{l \in levels(d)} \underbrace{\frac{|\mathcal{D}_{d=l}|}{|\mathcal{D}|}}_{\text{weighting}} \times \underbrace{H(t,\mathcal{D}_{d=l})}_{\text{entropy of partition } \mathcal{D}_{d=l}}$$

$$IG(d,\mathcal{D}) = H(t,\mathcal{D}) - rem(d,\mathcal{D})$$

(10 marks)

(ii) The **information gain** of the feature *Income* at the root node of the tree is 0.247. A colleague has suggested that *Newspaper* would be the best feature to query at the root node of the tree. Demonstrate whether or not this is the case. Show all workings.

(10 marks)

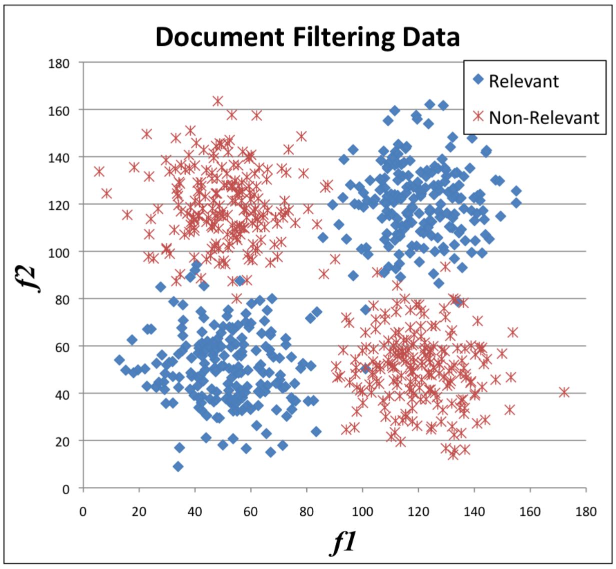
(iii) Decision trees are often used as the basis for **ensemble models**. The key to training effective ensemble models is to introduce diversity into the models in the ensemble. Compare the ways that *diversity* is introduced into ensembles in the **bagging** and **random forest** ensemble methods.

Question 3 (30 marks):

A confused client has come to you with three different customer marketing response prediction applications, each of which uses a particular classification algorithm to perform the response prediction (except for the classification algorithm used, all other aspects of the applications are identical).
Describe the evaluation criteria and experiments you would recommend so as these applications could be ranked from best to worst.

(15 marks)

The image below shows a scatter plot of a dataset from a simple document filtering problem. There are just two continuous features in this dataset, *f1* and *f2*, and two classes, *Relevant* and *Non-Relevant*. In the scatter plot *f1* is shown on the horizontal axis, *f2* is shown on the vertical axis and the shapes of the points represent class.



Discuss the difficulties associated with building classification models from datasets with characteristics similar to those shown in the scatter plot. In your answer comment on the suitability of specific classification approaches.

(15 marks)

Question 4 (30 marks):

There will also be a fourth question. It will most likely be on either **Similarity-based Learning** or **Error-based Learning.** The format will be similar to question 2 above.