Xi'an Jiaotong-Liverpool University
西交利物浦大学

Module Code	Examiner	Department	Tel
INT402		INT	

1st SEMESTER 2020/21 Remote Open-Book Resit Exam

Postgraduate

Data Mining and Big Data Analytics

Exam Duration: *2 Hours*

Crash Time Allowed: *15 Minutes* (for online exam)

INSTRUCTIONS TO CANDIDATES

1、 **This is a blended open-book exam. Please tick the integrity disclaimer *when uploading your answers on LEARNING MALL* and complete the assessment independently and honestly.**

2、 **Total marks available are 100. This accounts for 100% of the final mark.**

3、 **Answer all questions. Relevant and clear steps should be included in the answers.**

4、 **Only English solutions are accepted. Answers need to be handwritten and fully and clearly scanned or photographed for submission as one single PDF file via LEARNING MALL.**

5、 **Students should use the format "Module Code-Student ID.filetype" to name their files before submitting to ICE. For example, "INT402-18181881.pdf".**

6、 **The duration is 2 hours, and an additional 15-minute crash time beyond the exam duration will be allowed for you to report and resolve minor technical issues which may be encountered during the exam. Where there are any major problems preventing you from continuing the exam or submitting your answers in time, please do not hesitate to email the Module Examiner or Assessment Team of Registry (assessment@xjtlu.edu.cn).**

Notes:

- To obtain full marks for each question, relevant and clear steps should be included in the answers.
- Partial marks may be awarded depending on the degree of completeness and clarity.

Question 1

(1) Explain the Association algorithm in Data mining and why is the Association Rule necessary in Data Mining?

[5 points]

(2) Describe different data mining tasks in short paragraphs and diagrams, and give the primary objectives of data mining tasks?

[5 points]

(3) Given a support threshold s, the sets of items that appear in at least s baskets are called frequent itemsets. How do you choose a support threshold s?

[7 points]

(4) Confidence of association rule is the probability of J given $\mathrm{I} = \{\mathrm{i}_1, \mathrm{i}_2, \dots, \mathrm{i}_\mathrm{k}\}$. Consider the following expression:

$$\mathrm{conf}(\mathrm{I} \rightarrow \mathrm{j}) = \frac{\mathrm{support}(\mathrm{I} \cup \mathrm{j})}{\mathrm{support}(\mathrm{I})}.$$

So what happens if j is always bought?

[8 points]

[Total 25 points]

Question 2

The following table gives a training data set of 14 class labeled tuples. The class label attribute, buys a computer, has two distinct values (namely {yes, no}). Given an unknown tuple X = (age <= 30, income = medium, student = yes, credit_rating = fair), the task is to classify the tuple X by the following approach.
[Decision tree]: select attributes by information gain, compute ID3 decision tree, and classify the tuple X.

[15 points]

age	income	student	credit_rating	Buys_computer
<=30	high	no	Fair	No
<=30	high	no	excellent	No
31…40	high	no	fair	Yes
>40	low	no	fair	Yes
>40	low	yes	fair	Yes
>40	low	yes	excellent	No
<=30	medium	no	fair	No
<=30	low	yes	fair	Yes
>40	medium	yes	fair	Yes
<=30	medium	yes	excellent	Yes
31…40	medium	no	excellent	Yes
31…40	high	yes	fair	Yes
>40	medium	no	excellent	No

[Total 15 points]

Question 3

Clustering is the process of examining a collection of "points," and grouping the points into "clusters" according to some distance measurements. The goal is that points in the same cluster have a small distance from one another, while points in different clusters are at a large distance from one another. The "curse of dimensionality" makes clustering in high-dimensional spaces difficult, but enables some simplifications if used correctly in clustering algorithms.
Design a high-dimensional data set and provide a clustering method which requires the integration of several clustering techniques.
Describe your clustering method in short paragraphs (or diagrams) and justify your design. You may declare more assumptions, if necessary, to ease your design. In addition, provide reasoning explaining why the integration of clustering methods may sometimes improve the quality and efficiency of clustering.

(Hint: use one clustering algorithm as a pre-processing step for another clustering algorithm.)

[Total 20 points]

Question 4

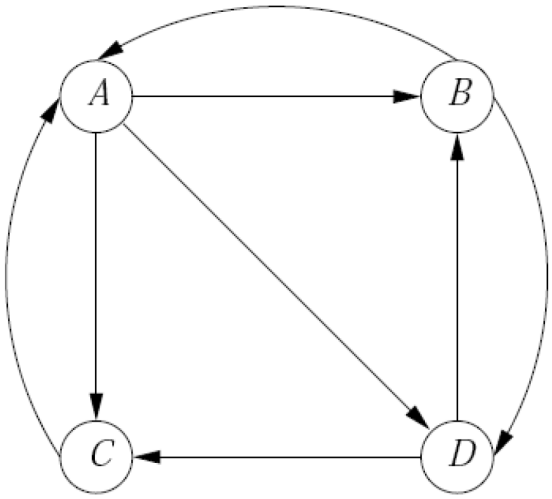
Use three iterations to compute the topic-sensitive PageRank for the graph of the following figure, assuming the teleport set is (suppose β=0.8):

(1) A only.

[12 points]

(2) A and C.

[8 points]



[Total 20 points]

Question 5

Three computers, A, B, and C, have the numerical features listed below:

Feature	A	B	C
Processor Speed (GHz)	3.06	2.68	2.92
Disk Size (GB)	500	320	640
Main-Memory Size (GB)	6	4	6

We may imagine these values as defining a vector for each computer; for instance, A's vector is [3.06, 500, 6]. We can compute the cosine distance between any two of the vectors, but if we do not scale the components, then the disk size will dominate and make differences in the other components essentially invisible. Let us use 1 as the scale factor for processor speed, α for the disk size, and β for the main memory size.

(1) In terms of α and β, compute the cosines of the angles between the vectors for each pair of the three computers.

[7 points]

(2) A certain user has rated the three computers as follows: A: 4 stars, B: 2 stars, C: 5 stars. You need to first normalize the ratings for this user,

[6 points]

(3) and next to compute a user profile for the user, with components for processor speed, disk size, and main memory size.

[7 points]

[Total 20 points]

END OF RESIT EXAM