HONG KONG BAPTIST UNIVERSITY

SEMESTER 2 EXAMINATION, 2005-2006

Subject Code: COMP 4650 Section Number: 1 Time Allowed: 2 Hour(s)

Subject Title: Data Mining and Knowledge Discovery Total No. of Pages: 2

Question 1 (15 minutes) (10 points)

Data warehousing, OLAP and data mining are three core technologies in business intelligence. What is the major role of each technology? What are the major differences between OLAP and data mining?

Question 2 (15 minutes) (10 points)

Describe two data mining tasks. For each task, give a real world example and explain what algorithm you can use to carry out the task.

Question 3 (15 minutes) (10 points)

Association rule algorithms are widely used in retails. Given the following transaction data, what is the total number of transactions? What are the support and confidence of the association rule (beer=>peanut)?

Transaction ID	Purchased Item	Amount $
X001	beer	9
X001	peanut	15
X001	bread	3
X002	beer	3
X003	peanut	15
X003	beer	9
X004	peanut	15
X004	bread	3

Question 4 (15 minutes) (10 points)

Assume the transaction database in Question 3 is very large and you use the Apriori algorithm to discover interesting product associations. Given two threshold values of support and confidence, you only obtain a few association rules which are not very interesting to you. If you want to obtain more association rules, what will you do? What will be the effects of your results?

Question 5 (10 minutes) (5 points)

What are the additional data required in sequential association rule mining in retail stores? Through what method do retail companies obtain the additional data?

Question 6 (10 minutes) (5 points)

You are given a task to use data mining to select target customers for promoting a new investment fund. In this task, customer income is very important. However, after exploring the customer database, you have found that about 40% of the customers have their income missing in the database. Describe a data mining method that you can use to fill in the missing customer income values before you build the campaign model.

Question 7 (10 minutes) (10 points)

Both decision trees and neural networks can be used to solve classification problems. Describe under what situations decision trees have advantages and under what situations neural networks have advantages.

Question 8 (5 minutes) (10 points)

You are given a task to use a clustering algorithm to discover fraudulent claims for an insurance company. Which clustering algorithm can be used and why?

Question 9 (5 minutes) (10 points)

Sketch the k-means clustering process. If you use the k-means algorithm to segment a customer database and you want to understand the characteristics of the segments, which data mining algorithm can you use to solve this problem?

Question 10 (20 minutes) (20 points)

Describe the data mining process in marketing campaigns in a bank. If no past campaign result is available, describe how the marketing campaign model can be built.

End