UNIVERSITY OF ABERDEEN SESSION 2009/2010

Examination in CS4031 Data Mining and Visualization

26 January 2010 12.00 – 14.00

Candidates are not permitted to leave the Examination Room during the first or last half hours of the examination.

Calculators Allowed
Answer any TWO questions.
Each question is worth 25 marks; the marks for each part of a question are shown in brackets

1. (a) Briefly explain what you understand by the term 'Exploratory Data Analysis'. (3)

(b) Consider the following table which shows a sample of data collected by a small business about their customers. The attribute Repeat_customer records a value of 'Yes' if the customer repeatedly buys goods from the business and a value of 'No' if the customer purchases only once. Imagine that you have been asked to build a decision tree using the data to help the business predict whether new customers would be Repeat_customers or not.

City	Gender	Education	Repeat_customer
London	F	College	Yes
Edinburgh	M	Graduate	Yes
London	F	College	Yes
London	F	College	No
Glasgow	M	High school	No
London	F	College	Yes
London	F	Graduate	Yes
Glasgow	M	College	Yes
London	F	High school	No
London	F	College	Yes

i. Using Information Gain select an attribute at the root node of the decision tree. (8)
ii. Using the root node selected in part (i) build the decision tree completely (4)
iii. Explain the decisions you had to make to build the complete tree. (2)

(c) Consider the following transactions in a local store. Imagine that you have been asked to produce association rules for the items using Apriori algorithm.

Transaction_id	Items_bought
101	milk,bread,cookies,juice
792	milk,juice
1130	milk,eggs
1735	bread,cookies,coffee

i. Using a minimum support of 0.5, generate the frequent itemsets for the above data showing clearly the application of apriori principle in pruning infrequent itemsets. (5)
ii. Using a minimum confidence of 0.5, generate the association rules generated from the frequent itemsets computed in part (i) showing clearly the application of apriori principle in pruning low confidence rules. (3)

PLEASE TURN OVER

2. (a) While performing shape based similarity matching over a database of raw time series, describe how you would remove different types of distortions. (4)

(b) Consider a time series represented by Piecewise Aggregate Approximation (PAA) of six segments as shown below:

Segment	PAA Value
1	0.64
2	0.34
3	0.12
4	-0.14
5	-0.50
6	-0.76

i. Compute the Symbolic Aggregate Approximation (SAX) representation for the above time series using the breakpoint information given below:

Alphabet	Breakpoint 1	Breakpoint 2
a	Negative Infinity	< -0.67
b	>= -0.67	< 0
c	>=0	<0.67
d	>= 0.67	Positive Infinity

(3)

ii. Derive the bitmap visualization of the SAX representation computed in part (i) showing simple sketches of the bitmap at each of the derivation steps. (5)

(c) Imagine you are in charge of the junior football team training at the Aberdeen Sports Village. In order to improve your bench strength you are constantly recruiting new players from local schools. Each school regularly sends you statistics of goals scored by players in the weekly school matches. You collect these statistics at regular time intervals so that you can build a time series of goals for each player. You explore this time series data collection using TimeSearcher, the time series visualization and data mining tool you have used in the practicals. With the help of sketches, specify the following queries visually using *timeboxes* and *variable time timeboxes* (VTTs) to be run on TimeSearcher:

i. List all the players who have scored goals in the last four weeks. (2)
ii. List all the players who have scored goals in successive weeks in the last one month. (3)

(d) Consider once again the time series of the goals scored by school children described in part (c). Instead of using TimeSearcher, you decide to use a data mining tool (such as Weka) for separating good players from the rest. For each player you create a feature based representation of the time series by extracting the average goals scored by the player in the last one month and also in their whole career. The following table shows data for six players:

Career Avg	Avg for Last Month
19.9	6.0
18.5	6.2
17.4	2.6
12.2	7.8
12.6	6.8
11.6	1.2
11.1	0.8

i. With the help of simple sketches describe how the k-means algorithm would compute two clusters for the above data. (5)
ii. Comment on the effectiveness of the k-means algorithm if we add a few rows of data to the above table with zero values in both the columns which means the players never scored goals. (2)
iii. What clustering method would you use in the situation described in part (ii). (1)

3. (a) Why can general data mining algorithms not be directly used for mining spatial data? Describe two practical approaches for dealing with the special properties of spatial data.(4)

(b) What are fisheye views? Explain how they are unique compared to the other visualization techniques you learnt about in this course. (3)

(c) You want to develop a new IDE (Integrated Development Environment such as Eclipse) for helping programmers visualize their source code more effectively. You decide upon using Fisheye Views for displaying source code files.

i. By choosing appropriate functions for a priori importance (API) and distance (D) for computing degree of interest (DOI) in this context, describe your design for generating fisheye displays of source code. (5)
ii. Source code files belonging to a particular project might be organized into hierarchies of folders which can be represented using Treemaps. Imagine you use treemaps to present folder hierarchies and fisheye views to display individual source code files. What are the merits and demerits of such a design? (4)

(d) Imagine you are in charge of the junior football team training at the Aberdeen Sports Village. In order to improve your bench strength you are constantly recruiting new players from local schools. Each school sends you data for several attributes such as age, weight, height, family history of playing football at professional level, a sibling already playing for the junior team and average goals scored both over their whole career and also over the last season.

i. Design an appropriate visualization tool for exploring the multi-dimensional player data, stating clearly any assumptions you make. (5)
ii. Describe how you would evaluate your visualization tool. (4)

END OF PAPER