UNIVERSITY OF ABERDEEN SESSION 2010/2011

Examination in CS4031 Data Mining and Visualization

19 January 2011 09.00am – 11.00am

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) Consider the following table which shows 10 instances of data for binary attributes A, B and C of an entity that are classified as either + or –.

A	B	C	Class
0	0	0	+
0	0	1	-
0	1	1	-
0	1	1	-
0	0	1	+
1	0	1	+
1	0	1	-
1	0	1	-
1	1	1	+
1	0	1	+

You have been asked to manually construct a decision tree that captures the classification knowledge from the above training data.

i. Draw the tree stumps for each of the attributes A, B and C marking clearly the class values for each of the instances in the leaf nodes of the tree stumps. (2)

ii. By using the notion of 'purity of leaf node' on the tree stumps from 1.(a).(i), select the attribute at the root of the decision tree. (2)

iii. Extend the tree stump selected in 1.(a).(ii) to create partial trees with each of the remaining attributes marking clearly the class values for each of the remaining instances in the leaf nodes of the trees. (2)

iv. By using the notion of 'purity of leaf node' on the partial trees from 1.(a).(iii), select the attribute for the second level of the decision tree. (2)

(b) Describe briefly the naive Bayes approach to classification stating clearly the assumptions the method makes about the data. (3)

(c) Consider the data from question 1.(a). Using the naive Bayes approach, compute the probabilities of class labels + and – for a test instance with A=0, B=1 and C=0. (7)

(d) Describe the Task by Type Taxonomy (TTT) framework for organizing visualizations. (7)

2. (a) What is PAA (Piecewise Aggregate Approximation) representation of a time series? (2)

(b) Consider the following time series data of goals scored by Diego Maradona and Pele playing for their respective national teams. The first column shows the years and the second column shows the corresponding goals for Maradona. The third column shows the data from the second column (goals data) in the normalized (scaled) form. The fourth column shows Pele's goals while the fifth column shows Pele's goals in the normalized form.

Year	Maradona Goals	Maradona Goals Normalized	Pele Goals	Pele Goals Normalized
1	0	-0.70	2	-0.83
2	0	-0.70	9	1.06
3	3	-0.08	11	1.60
4	7	0.73	4	-0.29
5	1	-0.49	0	-1.37
6	2	-0.29	8	0.79
7	0	-0.70	7	0.52
8	0	-0.70	2	-0.83
9	6	0.53	9	1.06
10	17	2.77	5	-0.02
11	4	0.12	0	-1.37
12	1	-0.49	4	-0.29

i. Compute a 3 segment PAA representations of the goals time series data (i.e. the number of segments in the PAA equal to 3) for each player showing all the major steps. (8)

ii. What is SAX (Symbolic Aggregate Approximation) representation of a time series? Write two advantages of SAX representation over other structural representations of time series. (3)

iii. For the two PAAs from 2.(b).(i) compute the SAX representations using the following break point data. (3)

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

(c) In the context of similarity matching of time series, briefly describe the term 'lower bounding distance' and state its importance for time series data mining. (3)

(d) Consider the original time series and their SAX representations computed above.

i. Inspect the SAX representations from 2.(b).(iii) above and comment on the similarity between the performance of the two players over the given time span. (2)

ii. Inspect the original times series for each player in the table above and comment on the similarity between the performance of the two players over the given time span. (2)

iii. Based on the comments from 2.(d).(i) and 2.(d).(ii) above, and the notion of lower bounding distance from 2.(c) above, justify the use of SAX representations for time series similarity matching. (2)

PLEASE TURN OVER

3. (a) Describe what you understand by the term 'spatial auto-correlation'. (2)

(b) Consider the following transactions involving five items. Imagine that you have been asked to produce association rules for the items using Apriori algorithm

Transaction_id	Item Lists
1	{a,b,d,e}
2	{b,c,d}
3	{a,b,d,e}
4	{a,c,d,e}

i. Using a minimum support of 0.75, 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 0.75, generate the association rules generated from the frequent itemsets computed in part 3.(b).(i) showing clearly the application of Apriori principle in pruning low confidence rules. (3)

(c) Imagine you are in charge of a project to create a user interface for visualizing a recipe collection. The recipe collection contains recipes of different cuisines such as Chinese, Indian and Thai. The collection contains recipes for starters, main courses, and desserts. The recipe collection contains data about popularity and simplicity ratings collected from users. Assume any other data relevant to this application.

i. Design an appropriate visualization tool for exploring the multi-dimensional recipe data, stating clearly any assumptions you make. (5)

ii. Explain how your visualization designs from 3.(c).(i) follow the design guidelines recommended by Edward Tufte. (2)

iii. Explain how your visualization designs from 3.(c).(i) follow the design guidelines recommended by Ben Shneiderman. (2)

(d) Shown below is a table of example data related blood donors. Assume that more rows exist in the real database than shown below:

Donor	Months Since Last Donation	Frequency	Amount of Blood Donated (cc)	Months Since First Donation
A	2	50	12500	98
B	0	13	3250	28
C	1	16	4000	35
D	2	20	5000	45
E	1	24	6000	77

i. What do you understand by the term 'rank-by- feature framework'? (2)

ii. With the help of an example from the above data explain how rank-by-feature framework helps in understanding the above data better. (4)

END OF PAPER