UNIVERSITY OF ABERDEEN SESSION 2011/2012

Examination in CS4031 Data Mining and Visualization

18 January 2012 15:00 – 17:00

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) Consider the following table which shows 12 instances of data related to two classes of customers C0 and C1 with attributes ID, Gender, Car Type and Shirt Size.

ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Large	C0
4	M	Sports	Extra Large	C0
5	F	Sports	Small	C0
6	F	Sports	Medium	C0
7	F	Luxury	Large	C0
8	M	Family	Extra Large	C1
9	M	Family	Medium	C1
10	F	Luxury	Small	C1
11	F	Luxury	Medium	C1
12	F	Luxury	Large	C1

i. Compute the Information Gain for the ID attribute selected as the root node. (3)

ii. Compute the Information Gain for the Gender attribute selected as the root node. (3)

iii. Compute the Information Gain for the Car Type attribute selected as the root node. (3)

iv. Compute the Information Gain for the Shirt Size attribute selected as the root node. (3)

v. Which attribute would you select for the root node of the decision tree? Explain the reasons for your choice. (3)

vi. Explain why you would not select ID as the root node of the decision tree. (2)

Note: The following table shows the log to base 2 values for the first 12 numbers which are useful for calculating information gain in question 1.a.

Number	1	2	3	4	5	6	7	8	9	10	11	12
Log to base 2	0	1	1.58	2	2.32	2.58	2.80	3	3.16	3.32	3.45	3.58

(b) Consider the one dimensional data set shown below where X is the attribute of an instance and Y is its class label:

X	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
Y	-	-	+	+	+	-	-	+	-	-

Classify the data point X=5.0 according to its 1-, 3-, 5- and 9-nearest neighbours using majority vote. (4)

(c) Confusion matrices for two classification models, M1 and M2 are given below:

Model M1		Predicted Class	
		Class Yes	Class No
Actual Class	Class Yes	126	74
	Class No	56	144

Model M2		Predicted Class	
		Class Yes	Class No
Actual Class	Class Yes	152	48
	Class No	24	176

i. Draw an ROC graph showing the performances of the two models given above. (2)

ii. Select the model that out-performs the other and explain your selection based on the ROC graph from above. (2)

2. (a) The following table shows two time series shapes.

Time Stamp	Time series 1	Time series 2
0	3	6
1	4	8
2	6	12
3	4	8
4	3	6

i. If you want to perform shape based similarity on these series, explain why you cannot apply Euclidean distance directly on these time series. (2)

ii. Transform the series appropriately and compute the Euclidean distance between them assuming that the means of the two series are 4 and 8 and the standard deviations are 1 and 2 respectively. (3)

(b) Consider the following data recorded by a SCUBA (Self Contained Underwater Breathing Apparatus) dive computer. Instead of recording the raw dive profiles (a time series of depth values recorded at regular time intervals) assume that the computer recorded a Piecewise Linear Approximation (PLA) of the dive profile using three segments (Descent, Bottom and Ascent). The table below shows each dive using a DiveId and for each DiveId data about the three segments are shown.

DiveId	Segment Type	Segment Start Time (Minutes from the start)	Segment Start Depth (Meters)	Segment End Time (Minutes from the start)	Segment End Depth (Meters)
1	Descent	0	0	2	14
1	Bottom	2	14	8	14
1	Ascent	8	14	15	0
2	Descent	0	0	2	32

2	Bottom	2	32	8	32
2	Ascent	8	32	16	0
3	Descent	0	0	3	12
3	Bottom	3	12	7	12
3	Ascent	7	12	13	0
4	Descent	0	0	3	30
4	Bottom	3	30	8	30
4	Ascent	8	30	18	0
5	Descent	0	0	3	36
5	Bottom	3	36	8	36
5	Ascent	8	36	20	0

Using the K-means algorithm, compute the clusters of similar dives showing the important intermediate stages of the K-means algorithm. You can assume that the only features important in a dive profile are 'bottom depth' (the depth value of the bottom segment) and 'ascent speed' (bottom depth divided by the duration of the ascent segment). (15)

(c) Consider a collection of dive profiles recorded by a SCUBA (Self Contained Underwater Breathing Apparatus) dive computer where each dive profile is a time series of depth values (in meters) measured in 20 second time intervals (similar to the dive profile data you used in the practicals). With the help of sketches, specify (draw) the following queries visually using *timeboxes* and *variable time timeboxes* (VTTs) to be used in TimeSearcher, the time series visualization tool you used in the practicals:

i) List all the dive profiles that have an ascent (rise in height) of 12 meters in 2 minutes starting from the 10^{th} minute of the dive. (2)
ii) List all the dive profiles that have an ascent (rise in height) of 12 meters in 2 minutes anywhere during the period between the start and end of the dive. (3)

3. (a) 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	{newspaper,beer,pen,water}
2	{beer,magazine,pen}
3	{newspaper,beer,pen,water}
4	{newspaper,magazine,pen,water}

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. (4)

(b) An environmental agency collects river level data on a daily basis generating a time series of river levels using a fixed number of gauging stations along a river. The database contains large numbers of such time series each corresponding to a different gauging station and a different river. Imagine you are in charge of a project to create a user interface for visualizing these time series of river levels. Design a visualization tool for exploring the multiple time series from the database, stating clearly any assumptions you make. Your design should allow data analysts to view each individual time series completely on one screen without scrolling. (5)

(c) Shown below are two geometries representing two plots of land selected for creating sports facilities in a school:

Geometry 1 = POLYGON((0 0, 20 0, 25 10, 0 10, 0 0))
Geometry 2 = POLYGON((20 0, 30 0, 25 10, 20 0))

Designing the sports facilities depends upon how these two geometries are positioned topologically.

i. Construct a DE-9IM (Dimensionally Extended Nine Intersection Matrix) for the given geometries? (9)

ii. Using the above matrix check if the topological relation between the given geometries is 'touches' which is specified by the codes 'FT*******' or 'F**T*****' or 'F***T****'. (2)

END OF PAPER