UNIVERSITY OF ABERDEEN SESSION 2023/2024

Aberdeen Institute of Data Science and Artificial Intelligence, South China Normal University

Examination in JC3503 Data Mining and Visualisation

June 2024

Part A (25 marks)

Answer ALL questions. Each part is worth 25 marks; the marks for each question are shown in brackets.

---

1. The table below shows a confusion matrix.

		Actual	
		+	-
Predicted	+	140	60
	-	40	160

Use the confusion matrix to calculate the following measures:

(a) Overall accuracy **[1 mark]**

(b) Sensitivity (True Positive Rate) **[1 mark]**

(c) Specificity (True Negative Rate) **[1 mark]**

2. Statistically we define four levels of measurement for attribute values of data: **Nominal**, **Ordinal**, **Interval**, and **Ratio**.
Classify the following attribute values into these four levels of measurement:

(a) Age **[1 mark]**

(b) Gender: {male, female} **[1 mark]**

(c) Assignment grade: {A, B, C, D, F } **[1 mark]**

(d) Degree program: {AI, BMIS, CS} **[1 mark]**

(e) Temperature in Celsius **[1 mark]**

3. 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 understand and predict whether new customers would be Repeat_customers or not. **Note:** there are a few tables with some useful calculations below that will help you with this.

ID	City	Gender	Education	Repeat_customer
1	London	F	College	Yes
2	Edinburgh	M	Graduate	Yes
3	London	F	College	Yes
4	London	F	College	No
5	Glasgow	M	High school	No
6	London	F	College	Yes
7	London	F	Graduate	Yes
8	Glasgow	M	College	Yes
9	London	F	High school	No
10	London	F	College	Yes

Useful Fractions		
$2/7 = 0.29$	$2/3 = 0.67$	$1/6 = 0.17$
$5/7 = 0.71$	$5/6 = 0.83$	$1/3 = 0.33$

Useful Multiplication		
$0.29 * -1.79 = -0.52$	$0.6 * 0.66 = 0.4$	$0.83 * -0.26 = -0.22$
$0.7 * 0.87 = 0.61$	$0.7 * -0.51 = -0.36$	$0.67 * -0.58 = -0.39$
$0.71 * -0.49 = -0.35$	$0.17 * -2.58 = -0.44$	$0.3 * 0.92 = 0.28$
$0.33 * -1.60 = -0.53$	$0.3 * -1.74 = -0.52$	

Useful Log_2 Calculations		
$log_2(0.1) = -3.32$	$log_2(0.17) = -2.58$	$log_2(0.29) = -1.79$
$log_2(0.3) = -1.74$	$log_2(0.33) = -1.60$	$log_2(0.67) = -0.58$
$log_2(0.7) = -0.51$	$log_2(0.71) = -0.49$	$log_2(0.83) = -0.26$

(a) Compute the Information Gain for the ID attribute selected as the root node. **[3 mark]**

(b) Compute the Information Gain for the City attribute selected as the root node. **[3 mark]**

(c) Compute the Information Gain for the Gender attribute selected as the root node. **[4 mark]**

(d) Compute the Information Gain for the Education attribute selected as the root node. **[4 mark]**

(e) Which *attribute* would you select for the root node of the decision tree? Explain the reasons for your choice. **[2 mark]**

(f) Explain why you would not select ID as the root node of the decision tree. **[1 mark]**

Part B (25 marks)

Answer ALL questions. Each part is worth 25 marks; the marks for each question are shown in brackets.

---

4. 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

(a) 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 mark]**

(b) Using a minimum confidence of 0.5, generate the association rules generated from the frequent itemsets computed in part (a) showing clearly the application of Apriori principle in pruning low confidence rules. **[3 mark]**

5. Consider the following six data objects (points) a—f in the two-dimensional Euclidean space (x1 and x2 are their coordinates):

Point	X_1	X_2
a	1	1
b	3	1
c	1	3
d	3	3
e	5	3
f	5	1

We are going to use the k-means algorithm to begin clustering the above data objects into **two clusters**. Assume Euclidean distance is used to measure the dissimilarity between data points.

Note: Please use $dist(i, j)$ to represent the distance between i and j, where i and j could be any points or cluster centers. Similarly, you can use $dist^2(i, j)$ to represent the squared distance between i and j. You can round any fractions to two decimal points.

Useful Calculations		
$0.67^2 = 0.45$	$1.33^2 = 1.77$	$3.33^2 = 11.09$

(a) When Objects a and c are selected as the initial cluster centers, give detailed steps of the algorithm when processing the above data and the value of SSE (Sum of Squared Error) after convergence.

[5 mark]

(b) When Objects a and e are selected as the initial cluster centers, give detailed steps of the algorithm when processing the above data and the value of SSE (Sum of Squared Error) after convergence.

[5 mark]

(c) What conclusion(s) can be drawn from the SSE values you calculated in (a) and (b)?

[1 mark]

6. 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

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

(a) Compute the Symbolic Aggregate Approximation (SAX) representation for the above time series using the breakpoint information given below: **[3 mark]**

(b) Explain what Dynamic Time Warping is. Give one advantage and one limitation of this method. **[3 mark]**

END OF EXAM PAPER