MD2

54

THE UNIVERSITY OF THE WEST INDIES
ST. AUGUSTINE

EXAMINATIONS OF DECEMBER 2019

Code and Name of Course: **COMP 3605 - Introduction to Data Analytics** Paper: 1

Date and Time: Monday 2nd December 2019 1pm Duration: 2 hours

INSTRUCTIONS TO CANDIDATES: This paper has 3 pages and 4 questions

The use of non-programmable scientific calculators is allowed.
Answer ALL Questions

PLEASE TURN TO THE NEXT PAGE

Total Mark = 50

Question 1

You are given a training data set *D* shown in the table below for a binary classification problem. The class label attribute *Play* has two different values {*Yes*, *No*}.

The Class-Labeled Training Data Set *D*

TID	*Outlook*	*Temperature*	*Humidity*	*Wind*	*Class: Play*
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rainy	Mild	High	Weak	Yes
5	Rainy	Cool	Normal	Weak	Yes
6	Rainy	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rainy	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rainy	Mild	High	Strong	No

a. Compute the information gain for the attribute ***Outlook***. [4 marks]

b. Compute the gain ratio for the attribute ***Temperature*** using *Gain*(***Temperature***) = 0.064. [3 marks]

c. Compute the Gini index for the attribute ***Temperature*** and the splitting subset {*Cool*, *Mild*}. [3 marks]

[Total mark: 10]

Question 2

a. You are given the transactional database *D* shown in the table below. The database has four transactions. Let the minimum support *min_sup* be 2.

Transactional data set *D*

TID	**Items**
001	*A, C, D*
002	*B, C, E*
003	*A, B, C, E*
004	*B, E*

Find all frequent itemsets in *D* using the Apriori algorithm. [7 marks]

PLEASE TURN TO THE NEXT PAGE

b. The following contingency table summarizes supermarket transaction data, where A refers to the transactions containing an item A, $\overline{A}$ refers to the transactions that do not contain A, B refers to the transactions containing an item B, and $\overline{B}$ refers to the transactions that do not contain B.

	A	$\overline{A}$	Σ_{row}
B	65	35	100
$\overline{B}$	40	10	50
Σ_{col}	105	35	150

i. Given a minimum support threshold *min_sup* of 40% and a minimum confidence threshold *min_conf* of 60%, is the association rule $A \rightarrow B$ strong? [3 marks]

ii. What is the correlation between the two items A and B? [5 marks]

[Total mark: 15]

Question 3

a. Write the nested loop algorithm for the $DB(r, \pi)$-outlier detection. [7 marks]

b. Briefly describe the Hopkins statistic that is used to measure the clustering tendency of a given data set D. [8 marks]

[Total mark: 15]

Question 4: Consider the linearly separable data set D in a two-dimensional space, as shown below, which contains eight training instances $\mathbf{x}_i$, class labels $y_i \in \{-1, 1\}$, and Lagrange multipliers λ_i for $i = 1, 2, ..., 8$.

Instances	x_1	x_2	y_i	λ_i
$\mathbf{x}_1$	2	2.5	1	2.7027
$\mathbf{x}_2$	2.5	3.2	−1	2.7027
$\mathbf{x}_3$	4	2.5	−1	0
$\mathbf{x}_4$	3.5	4	−1	0
$\mathbf{x}_5$	1	2	1	0
$\mathbf{x}_6$	2.2	1.5	1	0
$\mathbf{x}_7$	4.5	3.3	−1	0
$\mathbf{x}_8$	1.5	0.5	1	0

a. Specify support vectors from the given data set D. [2 marks]

b. Determine a decision boundary of a linear SVM (support vector machine). [6 marks]

c. Describe how to use the trained linear SVM to classify a test instance $\mathbf{z}$. [2 marks]

[Total mark: 10]

End of Question Paper