THE UNIVERSITY OF THE WEST INDIES
ST. AUGUSTINE

EXAMINATIONS OF DECEMBER 2018

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

Date and Time: 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 *buys_computer* has two different values {*yes*, *no*}.

The Class-Labeled Training Data Set D

TID	***age***	***income***	***student***	***credit_rating***	***Class: buys_computer***
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

a. Compute the information gain for the attribute ***age***. [4 marks]
b. Compute the gain ratio for the attribute ***income*** using $Gain(\textbf{\textit{income}}) = 0.029$. [3 marks]
c. Compute the Gini index for the attribute ***income*** and the splitting subset {*low*, *medium*}. [3 marks]
[Total mark: 10]

Question 2

a. Write the nested loop algorithm for the $DB(r, \pi)$-outlier detection. **[5 marks]**
b. Suppose that a city's average temperature values in July in the last 14 years are, in value-ascending order, 20.0°C, 24.0°C, 28.1°C, 28.2°C, 28.3°C, 28.4°C, 28.5°C, 29.1°C, 29.2°C, 29.3°C, 29.4°C, 29.5°C, 29.6°C, and 29.7°C. Let's assume that the average temperature values follow a normal distribution $\mathcal{N}(\mu, \sigma^2)$.

Use the maximum likelihood method to detect an outlier from the given univariate data set. [4 marks]
c. Describe how to use Mahalanobis distance to detect outliers in a multivariate data set. [6 marks]
[Total mark: 15]

PLEASE TURN TO THE NEXT PAGE

Question 3

A transactional data set D is given below, where $|D| = 9$, $min\_sup = 2$ (i.e., $2/9 = 0.22 = 22\%$).

Transactional data set D

TID	*List of item_IDs*
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

Find frequent itemsets in D using the Apriori algorithm. **[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
$\mathbf{x}_1$	0.3858	0.4687	1	65.5261
$\mathbf{x}_2$	0.4871	0.611	−1	65.5261
$\mathbf{x}_3$	0.9218	0.4103	−1	0
$\mathbf{x}_4$	0.7382	0.8936	−1	0
$\mathbf{x}_5$	0.1763	0.0579	1	0
$\mathbf{x}_6$	0.4057	0.3529	1	0
$\mathbf{x}_7$	0.9355	0.8132	−1	0
$\mathbf{x}_8$	0.2146	0.0099	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