SIM GLOBAL EDUCATION

UNIVERSITY OF WOLLONGONG AUSTRALIA

School of Computing and Information Technology

Student to complete:

Family name	
Other names	
Student number	
Table number	

CSCI316
Big Data Mining Techniques and Implementation

Final Examination Paper
Session 1 2021

Exam duration	3 hours
Weighting	50% of the subject assessment
Marks available	50 marks
Directions to students	6 questions to be answered. Marks for each question are shown beside the question.

Question 1 (7 marks)

(1.1) Implement from scratch a Python function for simple numerical encoding. This function takes a list of string values as input and returns a vector of integers as output. Write down the Python code.

(3 marks)

(1.2) Implement from scratch a Python function to compute the Gini index of a list. This function takes a list of categorical values as input and returns the Gini index as output. Write down the Python code.

(4 marks)

Question 2 (9 marks)

(2.1) Explain why pre-processing is important in big data.

(3 marks)

(2.2) Explain the advantages and disadvantages of data aggregation.

(3 marks)

(2.3) Explain undersampling and oversampling, and when you will apply them.

(3 marks)

Question 3 (9 marks)

(3.1) Assume that you are given a set of records as shown in the following table, where the last column contains the target variable. Present the procedure of using Gain Ratio to identify which attribute should be split. You need to show all steps of your calculation in detail.

(6 marks)

Case	Lecturer experience	Programming Subject?	Student satisfaction
1	Strong	No	Low
2	Weak	No	Low
3	Weak	Yes	Low
4	Weak	Yes	Low
5	Strong	No	High
6	Strong	No	High
7	Strong	Yes	High
8	Weak	Yes	High

(3.2) Why an ensemble classifier (such as a Random Forest) can enhance the performance of individual classifiers?

(3 marks)

Question 4 (8 marks)

(4.1) Use an example to illustrate the conditional independence assumption, and explain why it is important to the Naïve Bayes classifier.

(3 marks)

(4.2) Assume that a Bayesian classifier returns the following outcomes for a binary classification problem, which are sorted by decreasing probability values. P (resp., N) refers to a record belonging to a positive (resp., negative) class.

Tuple #	Class	Probability
1	P	0.90
2	P	0.80
3	N	0.70
4	P	0.60
5	P	0.55
6	N	0.54
7	N	0.53
8	N	0.51
9	P	0.50
10	N	0.40

Answer the following questions for the above example, and present all steps of calculation in detail.

a) What are the true positive (recognition) rate and false negative (recognition) rate if setting the probabilistic classification threshold to ≥ 0.70?
b) What is the smallest probabilistic classification threshold such that the precision is at least 60%?

(5 marks)

Question 5 (10 marks)

(5.1) Use an example to explain how the MapReduce model can process the outer join operation.

(3 marks)

(5.2) Why Apache Spark is suitable for large-scale machine learning? Use an example to support your answer.

(3 marks)

(5.3) Assume that a DataFrame named `FlightsDF` of flight statistics is defined in PySpark, with the following code processed.

```
FlightsDF.printSchema()
Out:
root
 |-- DEST_CITY: string (nullable = true)
 |-- DEST_COUNTRY_NAME: string (nullable = true)
 |-- ORIGIN_CITY: string (nullable = true)
 |-- ORIGIN_COUNTRY_NAME: string (nullable = true)

DF.show(2)
Out:
+---------------------+-----------+--------------+
|DEST_CITY|DEST_COUNTRY|ORIGIN_CITY|ORIGIN_COUNTRY|
+---------+------------+-----------+--------------+
|Sydney   |Australia   |Melbourne  |Australia     |
|Auckland |New Zealand |Singapore  |Singapore     |
+---------+------------+-----------+--------------+
only showing top 2 rows
```

Based on `FlightsDF`, write down the code in PySpark to implement the following operation: Find the country or countries with most international flights. (Note. An international flights has different original and destination countries.)

(4 marks)

Question 6 (7 marks)

(6.1) Why a classical Perceptron (i.e., a single layer of linear threshold units) is not preferable to use?

(2 marks)

(6.2) Implement a feedforward neural network by using the Keras API in TensorFlow for a multi-class classification problem. Assume that the data set has four numerical features and one target variable whose values are 1, 2 and 3. The network has one hidden layer with the sigmoid activation function. Present the Python code.

(5 marks)

End of Examination