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 3 2021

Exam duration	3 hours
Weighting	50% of the subject assessment
Marks available	50 marks
Directions to students	6 questions to be answered. Each question contains multiple sub-questions. Marks for each sub-question are indicated.

Question 1 (7 marks)

(1.1) Given a list named X which contains words (as strings), implement a Python function to compute the frequencies of words in X. Write down the Python code.

(3.5 marks)

(1.2) A list named Y contains numerical values (in float type). But some elements of Y have the value “None” (i.e., a missing value). Implement a Python function which replaces each missing value with the mean value of Y. Write down the Python code.

(3.5 marks)

Question 2 (9 marks)

(2.1) Explain why we cannot reuse the training data for testing in data mining.

(2 marks)

(2.2) Explain the concept of feature selection and feature generation, and in what situation to use each method.

(3 marks)

(2.3) Explain why we need to convert strings to numerical values in data mining. Describe a concrete example to demonstrate the advantage(s) of one-hot encoding compared with the direct conversion of strings to numerical values.

(4 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 information gain to identify which attribute should be split. You need to show all steps of your calculation in detail.

(6 marks)

Case	Income	Age	Buy?
1	High	Youth	Yes
2	High	Middle age	Yes
3	High	Senior	Yes
4	Medium	Senior	Yes
5	Medium	Youth	No
6	Medium	Youth	No
7	Low	Middle age	No
8	Low	Senior	No

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

(3 marks)

Question 4 (9 marks)

(4.1) In Naïve Bayesian classifiers, the numerical underflow and the zero count are two important issues. Explain these two issues and describe at least one common technique to overcome each issue.

(4 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	P	0.70
4	N	0.60
5	P	0.55
6	P	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 is the F score (i.e. F1 score) if setting the probabilistic classification threshold to 0.59?
(b) What is the highest probabilistic classification threshold such that the recall (sensitivity) is at least 80%?

(5 marks)

Question 5 (9 marks)

(5.2) Why is Apache Spark more suitable for data-parallel computation than for model-parallel computation? Also use an example to support your answer.

(4 marks)

(5.3) Assume that two Spark data frames named PATENT and CITATION are defined in PySpark, with the following code processed.

```
PATENT.printSchema()
Out:
root
 |-- PATENT_ID: long (nullable = true)
 |-- PATENT_NAME: string (nullable = true)
 |-- CLAIM_NUMBER: integer (nullable = true)
 |-- COUNTRY: string (nullable = true)

CITATION.printSchema()
Out:
root
 |-- CITING: long (nullable = true)
 |-- CITED: long (nullable = true)
```

Each row in PATENT indicates the ID, name, claim number (or number of claims) and country for a patent. Each row in CITATION indicates that a patent whose ID is in the CITING column cites a patent whose ID is in the CITED column. (Assume that the same patent ID is not in both columns for each row; in other words, a patient does not cite itself.)

Based on the two data frames PATENT and CITATION, write down the code in PySpark to implement the following operations.

(a) Find the country or counties with the highest *average number of claims per patent*.
(b) The *citation count* of a patent means “the total number of patents which cite that patent”. Find the patent name(s) with the highest citation count.

(5 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 regression problem. Assume that the data set has four numerical features and one numerical target variable. The network has one hidden layer with the sigmoid activation function. The number of neurons in the hidden layer is considered as a hyperparameter which you need to fine-tune. Write down the Python code of the implementation.

(5 marks)

End of Examination