UNIVERSITY OF WESTMINSTER

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
TIMED ASSESSMENT SEMESTER 2 2020/21

Module Code:	**5DATA002W**
Module Title:	**MACHINE LEARNING & DATA MINING**
Module Leader:	**DR VASSILIS S. KONTOGIANNIS**
Release Time:	**9 April 2021, 11:00 (UK time)**
Submission Deadline:	**9 April 2021, 14:30 (UK time) /** **[16:15 (UK time) only for the RAF students]**

Instructions to Candidates:

Please read the instructions below before starting the paper

- The time of exam is 3h and 30 min (total: 210 minutes). Only for those (RAF students) who are eligible for extra time, the time of that adjusted exam is 3h & 30 min + 105 min (total: 315 minutes).
- You are allowed to use scientific calculators.
- You need to answer all questions (both in part A and B)
- Show all steps of your work
- The Module Leader will be available during the exam release time to respond to any queries via the Discussion Board of the module's Blackboard site
- This is an individual piece of work so do not collude with others on your answers as this is an academic offence
- Plagiarism detection software will be in use
- Where the University believes that academic misconduct has taken place the University will investigate the case and apply academic penalties as published in Section 10 Academic Misconduct regulations.
- ***Once completed please submit your paper via the Assignment content. In case of problems with submission, you will have TWO opportunities to upload your answers and the last uploaded attempt will be marked.***
- ***Work submitted after the deadline will not be marked and will automatically be given a mark of zero***

Module Specific Information

This assessment is based on the following learning outcomes:

1. Effectively implement, apply and contrast unsupervised/supervised machine learning / data mining algorithms for simple data sets (L5.4)
2. Demonstrate an awareness of the issues of general machine learning / data mining problems and compare and contrast different solutions. (L5.3)
3. Be aware of ethical issues and personal responsibility in the preparation of data and the implementation and application of machine learning / data mining algorithms (L5.8)

Part A

1. **What are the different types of Learning/Training models in Machine Learning (ML)? Briefly explain their principles (4 marks)**

2. **What is the essential difference between classification and clustering? (4 marks)**

3. **Summarise the strength(s) and the weakness(es) of K-means clustering. (4 marks)**

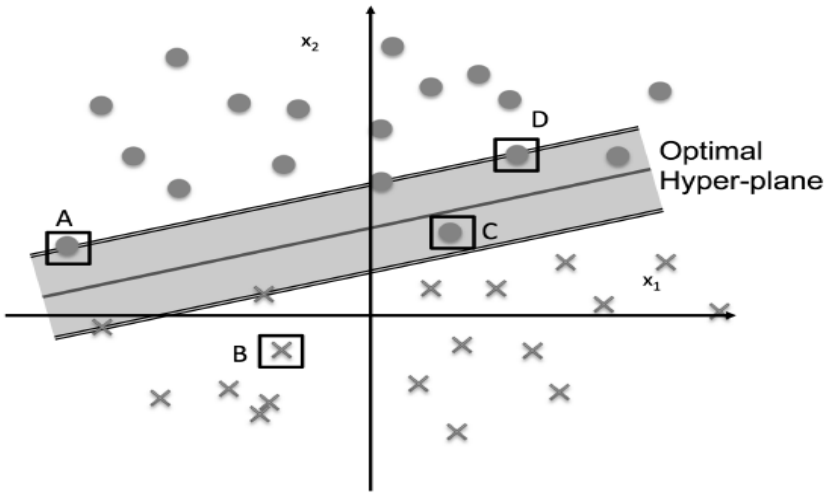
4. **Suppose you are training a decision tree on a dataset that contains k binary features (i.e. attributes). The dataset contains a very large number of examples/samples ($N \gg k$). What is the maximum possible depth of the decision tree? How likely is to have such situation in real decision trees cases? Justify your response. (4 marks)**

5. **You are training a Multilayer Perceptron (MLP) neural network for a particular classification task. After, some investigation, your neural network is constructed with 5 input variables, one hidden layer with 12 nodes and one output layer with 3 nodes (the classes). How many network parameters are required to be tuned/trained? Show your detailed calculations. (4 marks)**

6. **Briefly describe the general objective of Association Rules mining. What is the "Apriori Principle"? (4 marks)**

7. **The ethics of how a Machine Learning system is to function is a common thought that arises when we read about all these advancements in this domain. To build however an ML system, we need, among others, lots of data. Unfortunately, the selection/utilisation of data for using it in our ML system generates some ethical and biased issues. Briefly describe some of them. (6 marks)**

8. **The figure below displays the training samples and the learned SVM hyperplane. Which of the four highlighted samples is NOT considered as a support vector? Justify your response. (4 marks)**



9. **Briefly describe some of the advantages and disadvantages of Principal Component Analysis (PCA), a classic method used for data analysis. (4 marks)**

10. **Consider the following confusion table summarising the testing results for iris classification. As you are aware, the iris data is a classic multi-class benchmark dataset (12 marks in total)**

Confusion Matrix for IRIS dataset		Actual Class		
		Setosa	Versicolor	Virginica
Predicted Class	Setosa	20	0	0
	Versicolor	1	1	1
	Virginica	0	4	16

- **What is the overall classification accuracy? (2 marks)**
- **What is the sensitivity and specificity for each class? (6 marks)**
- **Use the table as an example to explain why confusion matrix is a better way to assess the performance of a classifier than the overall classification accuracy (4 marks)**

Part B

Question B-1

Market Basket Analysis is one of the key techniques used by large retailers to uncover associations between items. It works by looking for combinations of items that occur together frequently in transactions. Association Rules are widely used to analyse retail basket or transaction data. You have been given the following transaction database that consists of items (a, b, c, d &e) bought in a store by customers.

TID	Items
1	a,b,d,e
2	b,c,d
3	a,b,d,e
4	a,c,d,e
5	b,c,d,e
6	b,d,e
7	c,d
8	a,b,c
9	b,d
10	a,d,e

Find all the closed frequent itemsets which are not maximal, along with their support, for a minsupp threshold of 0.3. **Procedure:** Define first, all the frequent itemsets (10 marks), then all the closed frequent itemsets (10 marks) and finally all the closed frequent itemsets which are not maximal (10 marks). Show all steps/results of your work and justify any decision you have taken in your analysis.

Marks: 30

Question B-2

Complete the tasks based on provided observation table.

row	Weather in Athens	Mood
1	Overcast	good
2	Rain	good
3	Rain	bad
4	Sunny	good
5	Sunny	good
6	Overcast	bad
7	Overcast	good
8	Rain	bad
9	Sunny	good
10	Overcast	bad
11	Sunny	bad
12	Rain	good
13	Sunny	good
14	Sunny	bad
15	Rain	bad
16	Rain	bad

Frequency Table			
Class Weather	**Good**	**Bad**	
Overcast			4/16=0.25
Rain			
Sunny			
Total	8		
	8/16=0.5		

Likelihood Table – P(weather| mood)		
Class Weather	**Good**	**Bad**
Overcast		
Rain		
Sunny		

- Complete the Frequency and Likelihood tables (5 marks for each table)
- What is Naive Bayes Prediction for the mood when weather is Rain and Sunny respectively? (5 marks for each case)

Perform and show all calculations.

Marks: 20

END OF THE TEST