UNIVERSITY OF WESTMINSTER

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
TIMED ASSESSMENT SEMESTER 2 2020/21 (Referral/Deferral) – Summer 2021

Module Code:	**5DATA002W**
Module Title:	**MACHINE LEARNING & DATA MINING**
Module Leader:	**DR VASSILIS S. KONTOGIANNIS**
Release Time:	**30 June 2021, 10:00 (UK time)**
Submission Deadline:	**30 June 2021, 13:30 (UK time)** **[15:15 (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. **Hierarchical clustering can be divided into two main types. Name these two types and provide a brief description for both of them (5 marks).**

2. **Support Vector Machine (SVM) is a supervised algorithm used for classification tasks. As, all machine learning (ML) algorithms, SVM has a number of issues affecting its performance. Please indicate which of the following case/s are responsible when SVM's performance is deteriorated. Justify your response (4 marks).**
   A) The data is linearly separable
   B) The data is clean and ready to use
   C) The data is noisy and contains overlapping points

3. **Overfitting is one important cause for the poor performance of machine learning algorithms. Briefly describe what it is and mention possible reasons for its presence (4 marks).**

4. **You are training a Multilayer Perceptron (MLP) neural network for a particular time-series regression task. After, some investigation, your neural network is constructed with 7 input variables, two hidden layers with 16 and 8 nodes for the first and second hidden layers respectively and one output layer with 1 node (the regression output). How many network parameters are required to be tuned/trained? Show your detailed calculations. (4 marks)**

5. **We have been provided with the dataset shown below, to learn a decision tree which predicts if students pass the machine learning module (false/true), based on their previous A-Levels results and whether or not studied.**

A-Levels	Studied	Passed
C	F	F
C	T	T
B	F	F
B	T	T
A	F	T
A	T	T

   What is the entropy H(Passed)? For the calculation of log$_2$, remember from maths, that: $\log_a b = \dfrac{\log_{10} b}{\log_{10} a}$ **. Show all calculations (4 marks).**

6. **Hierarchical clustering is an alternative approach to *k*-means clustering for identifying groups in a data set. In contrast to *k*-means, hierarchical clustering creates a hierarchy of clusters and therefore does not require us to pre-specify the number of clusters. However, a fundamental question in hierarchical clustering is: *How do we measure the dissimilarity between two clusters of observations?* A number of different linkage methods have been developed, among them, the complete, the single, the average and the centroid linkage scheme. Provide a brief description for these four linkage schemes, focusing also on their differences (8 marks).**

7. **Many business enterprises accumulate large quantities of data from their day-to-day operations. For example, huge amounts of customer purchase data are collected daily**

at the checkout counters of grocery stores. Here, we have a very small grocery store that sells 10 unique products through a number of transactions. Calculate the total number of itemsets that can be created by those 10 products (2 marks). What is the number of possible association rules that can be created by those unique products (before performing any pruning technique)? Show all steps of calculations (4 marks).

8. **Briefly describe the term "black box" models we usually encounter in machine learning applications. Provide an example of such a black box model (3 marks).**

9. **Consider the following confusion table summarising the testing results for a Fruit dataset classification (12 marks in total)**

Confusion Matrix for Fruit dataset		Predicted Class		
		Banana	Orange	Strawberries
Actual Class	Banana	6	0	2
	Orange	3	9	1
	Strawberries	1	0	10

- **What is the overall classification accuracy? (3 marks)**
- **What is the sensitivity, specificity and precision for each class? (9 marks)**

Part B

Question B-1

Clustering is a data mining technique to group a set of objects in a way such that objects in the same cluster are more similar to each other than to those in other clusters. In this hierarchical clustering scheme, we assign initially each object (sample) to a separate cluster. Then we compute the distance (similarity) between each of the clusters and join the two most similar clusters.

Let's have the following 6 samples:

- **A: (7,0)**
- **B: (10,0)**
- **C: (17,0)**
- **D: (24,0)**
- **E: (35, 0)**
- **F: (43,0)**

Use the Euclidean Distance to compute the distance matrix related to these samples. Notice, this is a symmetric matrix (5 marks). Perform a hierarchical clustering, using the complete linkage approach and show all steps of this process (four stages/matrices) (20 marks, i.e. 5 marks for each stage). Finally, create a plot of the final dendrogram (5 marks).

Total marks: 30

Question B-2

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

Find all the frequent itemsets along with their support, for a minsupp threshold of 0.3 (10 marks). Find all frequent maximal itemsets (10 marks). Show all steps/results of your work and justify any decision you have taken in your analysis.

Marks: 20

END OF THE TEST