UNIVERSITY OF WESTMINSTER

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
In Class Test Semester 2 2021/22

Module Code:	**5DATA002W**
Module Title:	**MACHINE LEARNING & DATA MINING**
Module Leader:	**DR VASSILIS S. KONTOGIANNIS**
Date:	**~~12 April 2022, 16:00-18:00 (UK time)~~**
Location:	**~~Copland Building: 1.109, 1.108 and 1.112~~** **~~Also in LG.102 (only for RAF students)~~**

Instructions to Candidates:

Please read the instructions below before starting the paper

- ~~The time of exam is 2h (total: 120 minutes). Only for those (RAF students) who are eligible for extra time, the time will be 2h and 30mins (total: 150 minutes)~~
- You are allowed to use scientific calculators. You are not allowed to use the calculator in your smartphone.
- You need to answer all questions (both in part A and B)
- Show all steps of your work
- This is an individual piece of work, so do not collude with others on your answers as this is an academic offence
- This is a closed book test.
- ~~**When you finish the test, you need to return both the answer booklet AND the paper with the specifications of the test.**~~

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, discussing and addressing issues arising from their use, in work-based scenarios from industry and wider community.
2. Demonstrate an awareness of ethical issues and personal responsibility in the preparation of data and the implementation, application and communication of machine learning / data mining algorithms in a professional working environment

Part A

1. **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 (i.e. the regression output). How**

many network parameters are required to be tuned/trained? Show your detailed calculations (2 marks).

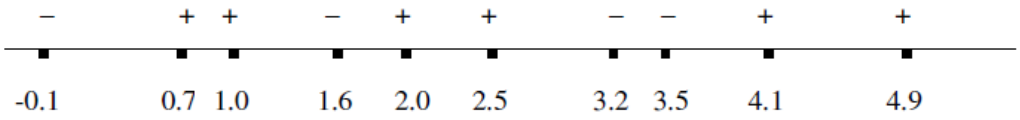
2. **Briefly describe the term "black box" models we usually encounter in machine learning applications (2 marks). Provide an example of such a black box model and briefly discuss why you consider it as such (2 marks).**
3. **Given three clusters, X, *Y* and *Z*, containing a total of six points, where each point is defined by an integer value in one dimension, X = {0, 2, 6}, Y = {3, 9} and *Z* = {11}, which two clusters will be merged at the next iteration stage of Hierarchical Agglomerative Clustering when using the standard Euclidean distance and (i) Single Linkage (4 marks), (ii) Complete Linkage (4 marks). Justify your response for these two cases, by showing all steps of your work.**
4. **Consider the following confusion matrix (CM) summarising the testing results for a Fruit dataset classification (14 marks in total).**

Confusion Matrix for Fruit dataset		Predicted Class		
		Apple	Pears	Grapes
Actual Class	Apple	6	0	2
	Pears	3	9	1
	Grapes	1	0	10

- **What is the overall classification accuracy of this CM? (2 marks)**
- **Decompose the above 3x3 CM into three individual (per fruit) 2x2 CMs (3x2= 6 marks)**
- **What is the sensitivity and specificity for each class? (3x2 = 6 marks)**

Show all steps of your work.

5. **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 (2 marks).**
6. **Consider the following dataset with one real-valued input x and one binary output y. You are going to use k-NN with standard Euclidean distance to predict y for x. What is the leave-one-out cross-validation error of 1-NN on this dataset? Fill the empty column in the provided table and provide your answer as the number of misclassifications. Obviously, this number must be consistent with the information you will provide in the table (2 marks).**



X	Y	Predicted Y
-0.1	-	
0.7	+	
1.0	+	
1.6	-	
2.0	+	
2.5	+	
3.2	-	
3.5	-	
4.1	+	
4.9	+	

7. **We have undertaken an environmental research study and we have collected 1500 observations for the following three animals: Cat, Parrot and Turtle. The input variables**

(or attributes) are categorical in nature i.e., they store two values, either True or False, and they are: Swim, Wings, Green Colour, Sharp Teeth. The following table summarises our observations.

Animal	Swim	Wings	Green	Sharp Teeth	Total
Cat	450	0	0	500	500
Parrot	50	500	400	0	500
Turtle	500	0	100	50	500
Total	1000	500	500	550	1500

Using, this data, we would like to classify the following observation into one of the output classes (Cats, Parrot or Turtle) by using the Naive Bayes (NB) Classifier.

Animal	Swim	Wings	Green	Sharp Teeth
unknown	True	False	True	False

Your task, is to predict whether this animal is a Cat, Parrot or a Turtle based on the defined input variables (swim, wings, green, sharp teeth) (10 marks).

- ***Probabilities calculations: 2 marks***
- ***Application of NB for each animal case (3 x 2 = 6 marks)***
- ***Final Decision (2 marks)***

Show all steps of your work. Justify your response.

Part B

Question B-1

The following table is a set of three-course menus from a famous restaurant. The idea here, is to create a decision tree (using the ID3 algorithm) to correctly classify similar examples. As the aim is to classify menus regarding whether they are good or not, calculate whether this algorithm would use "Starter" or "Main course" as the root of the decision tree. In this specific question you will use the entropy and information gain concepts. For the calculation of $\log_2$, remember, from maths, that $\log_a b = \frac{\log_{10} b}{\log_{10} a}$.

Starter	Main Course	Dessert	Good Menu
salad	steak	cheesecake	yes
soup	salmon	profiteroles	yes
salad	variety-roast	Fruit-salad	no
salad	surprise-bake	cheesecake	no
soup	variety-roast	Fruit-salad	yes
salad	salmon	Fruit-salad	yes
salad	variety-roast	Fruit-salad	no

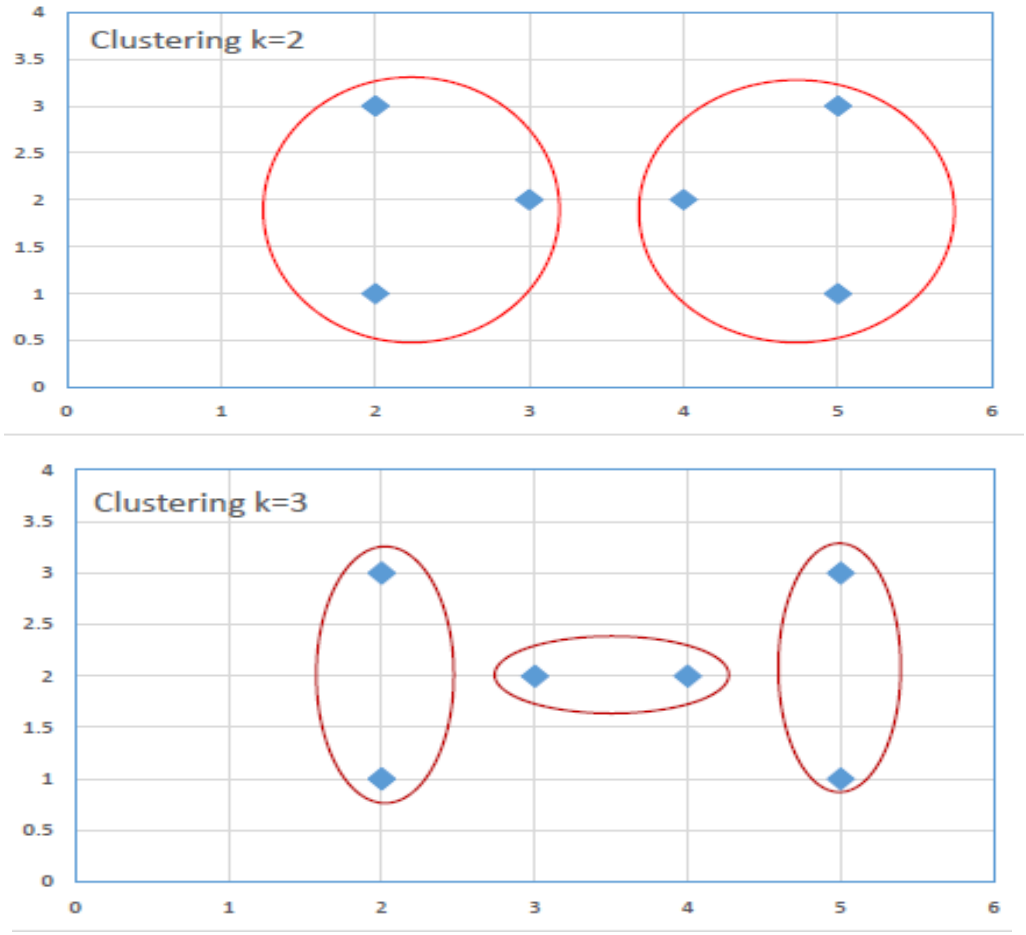
You need to address/calculate the following issues:

- ***Total parent entropy: 2 marks***
- ***Starter case: two individual entropies (2x2= 4 marks), weighted entropy (2 marks), information gain (2 mark)***
- ***Main course case: four individual entropies (4x2=8 marks), weighted entropy (2 marks), information gain (2 marks)***
- ***Final Decision (2 marks)***

Show all steps of your work (24 marks).

Question B-2

Given the results of the two clustering attempts below, obtained running the same algorithm with K=2 and K=3 clusters, calculate the related within_cluster_sums_of_squares (WSS) and between_cluster_sums_of_squares (BSS), and tell which result is better and why.



You need to address/calculate the following issues:

For k=2:

- ***Calculation of two cluster centres and centroid for all points (3x2 = 6 marks)***
- ***Calculation for WSS***
  - ***(suggestion: calculate the WSS for each cluster and then add them for the final WSS) (2x2 = 4 marks)***
- ***Calculation for BSS***
  - ***(suggestion: calculate the BSS for each cluster and then add them for the final BSS) (2x2 = 4 marks)***

For k=3:

- ***Calculation of three cluster centres (3x2=6 marks)***
- ***Calculation for WSS***
  - ***(suggestion: calculate the WSS for each cluster and then add them for the final WSS) (3x2 = 6 marks)***
- ***Calculation for BSS***
  - ***(suggestion: calculate the BSS for each cluster and then add them for the final BSS) (3x2 = 6 marks)***

Decision and justification (2 marks)

Show all steps of your work (34 marks).