Questions & Answers (ICT 2022)

Part A

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)

Indicative answer: (7+1) x 16 + (16+1) x 8 + (8+1) x 1 = 273. We have to include also the bias "weights" which is a compulsory component in MLP structure.

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).

Indicative answer: In machine learning, such "black box" models are created directly from data by an algorithm, meaning that humans, even those who design them, cannot understand how variables are being combined to make predictions. Even if one has a list of the input variables, "black box" predictive models can be such complicated functions of the variables that no human can understand how the variables are jointly related to each other to reach a final prediction (**2 marks**). Neural network (NN) models is a representative example of such models. Knowledge is accumulated into the trained, by the learning algorithm, weights and during testing phase, the input information flows via these weights and produces the outcome. A NN is a black box in the sense that while it can approximate any function (the level of accuracy depends on various factors), studying its structure won't give us any insights on the structure of the function being approximated. Or, to put it more simplistic, there is no interpretation how the input vector is applied to a "hidden" maths formula that characterises the specific function and then produces a correct output. This "black box" feature of NNs, is considered by some professionals as an obstacle for fully adoption of such learning-based models (**2 marks**).

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.

Indicative answer: For each case of linkage, we need to find the distances of X&Y, X&Z and Y&Z. Then, we need to choose the minimum of them. In single cluster linkage we need to find the minimum of individual combined components, while in the case of complete, the maximum.

Single linkage.

Dist (X,Y) = min {d(0,3), d(0,9), d(2,3), d(2,9), d(6,3), d(6,9)} = min{3,9,1,7,3,3} = 1 (1 mark)
Dist (X,Z) = min {d(0,11), d(2,11), d(6,11)} = min{11, 9, 5} = 5 (1 mark)
Dist (Y,Z) = min {d(3,11), d(9,11)} = min{8, 2} = 2 (1 mark)
For single linkage, the choice will be XY, as it has the minimum distance. (1 mark)

Complete linkage.

Dist (X,Y) = max {d(0,3), d(0,9), d(2,3), d(2,9), d(6,3), d(6,9)} = max{3,9,1,7,3,3} = 9 (1 mark)
Dist (X,Z) = max {d(0,11), d(2,11), d(6,11)} = max{11, 9, 5} = 11 (1 mark)
Dist (Y,Z) = max {d(3,11), d(9,11)} = max{8, 2} = 8 (1 mark)
For complete linkage, the choice will be YZ, as it has the minimum distance. (1 mark)

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.

Indicative answer:
The overall classification accuracy is: (6+9+10))/32 = 78.125%
As this is a multi-class problem, we need to decompose it, so to create "individual" CMs for each fruit, just to make easier the understanding of the concept. The following CM shows the details.

		Predicted Class	
		Yes	Not
Actual Class	Yes	TP	FN
	Not	FP	TN

Confusion Matrix for Fruit dataset		Predicted Class	
		Apple	Not
Actual Class	Apple	6	2
	Not	4	20

Confusion Matrix for Fruit dataset		Predicted Class	
		Pears	Not
Actual Class	Pears	9	4
	Not	0	19

Confusion Matrix for Fruit dataset		Predicted Class	
		Grapes	Not
Actual Class	Grapes	10	1
	Not	3	18

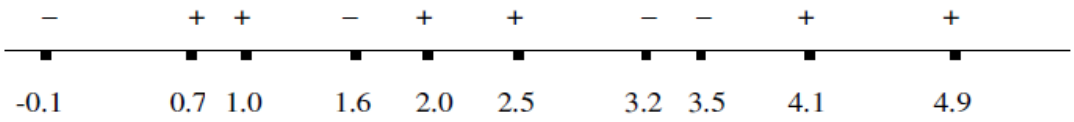
Sensitivity: TP/(TP+FN)
Specificity: TN/(TN+FP)

For apple: Sensitivity: 6/8 = 75%, specificity: 20/24 = 83.33%
For pears: Sensitivity: 9/13=69.23%, specificity: 19/19 = 100%
For grapes: Sensitivity: 10/11 = 90.91%, specificity: 18/21 = 85.71%

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).

Indicative answer: Overfitting is a problem which refers to a model that approximates the training data too well. Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new (i.e. testing) data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model. The problem is that these concepts do not apply to new data and negatively impact the models' ability to generalise. Overfitting is more likely with nonparametric and nonlinear models that have more flexibility when learning a target function (2 marks).

Consider the following dataset with one real-valued input x and one binary output y. We 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	+	

Indicative answer:

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

In the leave-one-out CV concept, each time one sample is considered as the testing sample, while all the remaining as training samples. With 1-nn, we are looking only in the first nearest neighbour. The above table shows these results, and the number of misclassifications is 4 (2 marks).

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 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). Show all steps of your work. Justify your response (10 marks).

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

Indicative answer:

Initially, from the observation table, we need to calculate the probabilities for each animal:
P(cat) = P(Turtle) = P(Parrot) = 500/1500 = 0.333
Then we need to calculate the probabilities for each attribute.
P(swim) = 1000/1500 = 0.666, P(wings) = P(green) 500/1500 = 0.333, P(sharp Teeth) = 550/1500= 0.366.
From our testing data, only the swim and green attributes are true, thus only these two variables will be considered.

In order to check if the animal is a cat:
P(Cat | swim, green) = [P(swim|cat) * P(green |cat) * P(cat)] / P(swim,Green) = [P(swim|cat) * P(green |cat) * P(cat)] / [P(swim) * PGreen)]

P(swim |cat) = 450/500 = 0.9
P(green | cat) = 0
Thus, P(Cat | Swim, Green) = (0.9 * 0 * 0.333) / (0.666 * 0.333) = 0

In order to check if the animal is a parrot:
P(parrot | swim, green) = [P(swim|parrot) * P(green |parrot) * P(parrot)] / P(swim,Green) = [P(swim|parrot) * P(green |parrot) * P(parrot)] / [P(swim) * PGreen)]

P(swim |parrot) = 50/500 = 0.1
P(green | parrot) = 400/500 = 0.8
Thus, P(parrot | Swim, Green) = (0.1 * 0.8 * 0.333) / (0.666 * 0.333) = 0.02664/0.221= 0.12

In order to check if the animal is a turtle:
P(turtle | swim, green) = [P(swim|turtle) * P(green |turtle) * P(turtle)] / P(swim,Green) = [P(swim|turtle) * P(green |turtle) * P(turtle)] / [P(swim) * PGreen)]

P(swim |turtle) = 500/500 = 1
P(green | turtle) = 100/500 = 0.2
Thus, P(turtle | Swim, Green) = (1 * 0.2 * 0.333) / (0.666 * 0.333) = 0.0666/0.221= 0.301

The value of P(Turtle| Swim, Green) is greater than P(Parrot| Swim, Green), therefore we can say that the class of the unknown animal is Turtle.

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 log2, 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).

Solution:

The total parent entropy (i.e. good menu) is calculated as:

$$Entropy(S) = -\frac{4}{7}\log_2\left(\frac{4}{7}\right) - \frac{3}{7}\log_2\left(\frac{3}{7}\right) = 0.985$$

We need to check the entropy for each one of these two attributes individually and then find which attribute maximizes the information gain.

Starter attribute:

Values	Yes	No	Entropy(S_i)
S_1 (salad)	xx	xxx	[2/5, 3/5]
S_2(soup)	xx		[1, 0]

$$Entropy(S_1) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right) = 0.971$$

$$Entropy(S_2) = -\log_2(1) = 0$$

Therefore, the entropy for this attribute is:

$$Entropy(S \mid starter) = \frac{5}{7}\cdot 0.971 + \frac{2}{7}\cdot 0 = 0.693$$

The information gain for Starter is: Entropy(S) – Entropy(S | starter) = 0.985 – 0.693 = 0.292

Main course attribute:

Values	Yes	No	Entropy(S_i)
S_1(steak)	1	0	[1, 0]
S_2(salmon)	2	0	[2/2, 0] = [1, 0]
S_3(variety-roast)	1	2	[1/3, 2/3]
S_4(surprise-bake)	0	1	[0, 1]

$$Entropy(S_1) = -\log_2(1) - 0 = 0$$
$$Entropy(S_2) = -\log_2(1) - 0 = 0$$
$$Entropy(S_3) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) = 0.918$$
$$Entropy(S_4) = 0 - \log_2(1) = 0$$

Therefore, the entropy for this attribute is:

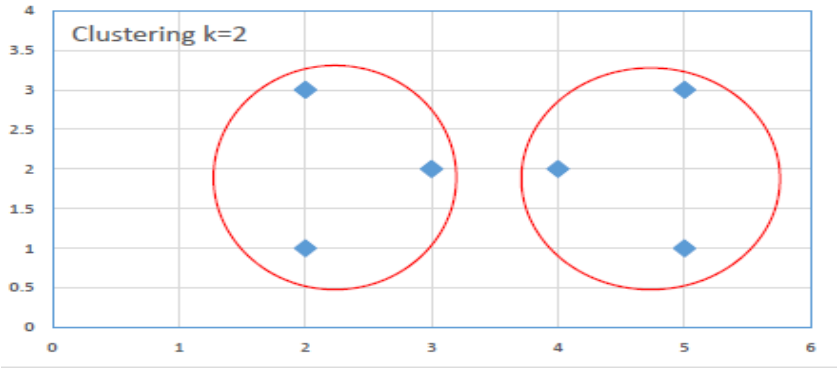
$$Entropy(S \mid main\ course) = \frac{1}{7}\cdot 0 + \frac{2}{7}\cdot 0 + \frac{3}{7}\cdot 0.918 + \frac{1}{7}\cdot 0 = 0.393$$

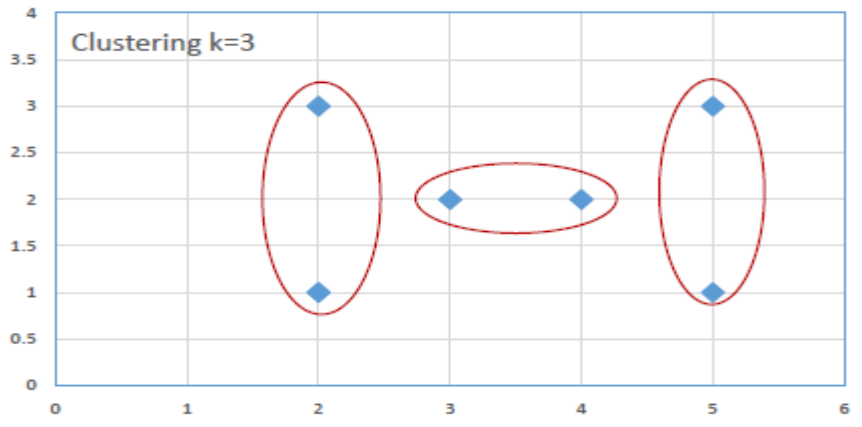
The information gain for main course is: Entropy(S) – Entropy(S |main course) = 0.985 – 0.393=0.592

Main course produces the highest gain, and thus this attribute will be the root of the tree.

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).

Solution

For $k=2$ (two clusters)

$$c_1 = \left(\frac{2+2+3}{3}, \frac{1+3+2}{3}\right) = \left(\frac{7}{3}, 2\right)$$

$$c_2 = \left(\frac{4+5+5}{3}, \frac{2+1+3}{3}\right) = \left(\frac{14}{3}, 2\right)$$

} two cluster centres

Centroid for all points

$$c = \left(\frac{2+2+3+4+5+5}{6}, \frac{1+2+3+2+1+3}{6}\right) = \left(\frac{21}{6}, 2\right) = (3.5, 2)$$

Total WSS = WSS(cluster1) + WSS(cluster2)

$$WSS_1 = \left(2-\tfrac{7}{3}\right)^2 + (1-2)^2 + \left(2-\tfrac{7}{3}\right)^2 + (3-2)^2 + \left(3-\tfrac{7}{3}\right)^2 + (2-2)^2 =$$

$$= \left(-\tfrac{1}{3}\right)^2 + (-1)^2 + \left(-\tfrac{1}{3}\right)^2 + 1^2 + \left(\tfrac{2}{3}\right)^2 =$$

$$= \tfrac{1}{9} + 1 + \tfrac{1}{9} + 1 + \tfrac{4}{9} = 2.666$$

$$WSS_2 = \left(4-\tfrac{14}{3}\right)^2 + (2-2)^2 + \left(5-\tfrac{14}{3}\right)^2 + (1-2)^2 + \left(5-\tfrac{14}{3}\right)^2 + (3-2)^2 =$$

$$= \left(-\tfrac{2}{3}\right)^2 + 0 + \left(\tfrac{1}{3}\right)^2 + (-1)^2 + \left(\tfrac{1}{3}\right)^2 + 1^2 =$$

$$\tfrac{4}{9} + \tfrac{1}{9} + 1 + \tfrac{1}{9} + 1 = 2.666$$

$$\boxed{WSS = 5.332}$$

$$BSS_1 = 3\left[\left(3.5-\tfrac{7}{3}\right)^2 + (2-2)^2\right] = 4.083$$

$$BSS_2 = 3\left[\left(3.5-\tfrac{14}{3}\right)^2 + (2-2)^2\right] = 4.083$$

$$\boxed{BSS = 8.166}$$

Total $WSS + BSS = 13.498 \simeq 13.5$

For ~~K~~ = 3 (three clusters)

Three clusters centers: $c_1 = (2,2)$, $c_2 = (3.5,2)$, $c_3 = (5,2)$

The total centroid is the same: $c = (3.5, 2)$

$$WSS_1 = (2-2)^2 + (1-2)^2 + (2-2)^2 + (3-2)^2 = 2$$

$$WSS_2 = (3-3.5)^2 + (2-2)^2 + (4-3.5)^2 + (2-2)^2 = 0.5$$

$$WSS_3 = (5-5)^2 + (1-2)^2 + (5-5)^2 + (3-2)^2 = 2$$

$$\boxed{\text{Total } WSS = 4.5}$$

$$BSS_1 = 2\left[(2-3.5)^2 + (2-2)^2\right] = 4.5$$

$$BSS_2 = 2\left[(3.5-3.5)^2 + (2-2)^2\right] = 0$$

$$BSS_3 = 2\left[(5-3.5)^2 + (2-2)^2\right] = 4.5$$

$$\boxed{BSS = 9}$$

Obviously $WSS + BSS = 13.5$ (constant)

The case $k=3$ is better, as it has both a higher cohesion (WSS is lower) and a higher separation (BSS is higher)