For each question you can either type your answers into AnswerBook, upload handwritten work, or do a combination of these.

IMPERIAL COLLEGE LONDON

TIMED REMOTE ASSESSMENTS 2020-2021

BEng Honours Degree in Computing Part III
BEng Honours Degree in Electronic and Information Engineering Part III
MEng Honours Degree in Electronic and Information Engineering Part III
MEng Honours Degree in Electronic and Information Engineering Part IV
BEng Honours Degree in Mathematics and Computer Science Part III
MEng Honours Degree in Mathematics and Computer Science Part III
MEng Honours Degrees in Computing Part III
MSc Advanced Computing
MSc Artificial Intelligence
MSc in Computing (Specialism)
for Internal Students of the Imperial College of Science, Technology and Medicine

This paper is also taken for the relevant assessments for the Associateship of the City and Guilds of London Institute

PAPER COMP70050=COMP97101=COMP97151

INTRODUCTION TO MACHINE LEARNING (TERM1)

Friday 18 December 2020, 10:00
Duration: 105 minutes
Includes 15 minutes for access and submission

Answer ALL THREE questions
Open book assessment

By completing and submitting work for this assessment, candidates confirm that the submitted work is entirely their own and they have not (i) used the services of any agency or person(s) providing specimen, model or ghostwritten work in the preparation of the work they have submitted for this assessment, (ii) given assistance in accessing this paper or in providing specimen, model or ghostwritten answers to other candidates submitting work for this assessment.

Paper contains 3 questions

1 a You would like to predict whether a piece of fruit is **sweet** or **sour** based on the *colour* and *softness* of the fruit.

You are given a dataset of 200 training samples, where 100 of the fruits are labelled as sweet, and the other 100 labelled as sour.

The colour of the fruits in the dataset are either green, red or yellow. 70% of the sour fruits are green, 20% are red and 10% are yellow. 60% of the sweet fruits are red and 40% are yellow. None of the sweet fruits are green.

Fruits can either be soft or hard. 75% of the sweet fruits are soft, while the remaining are hard. All sour fruits are hard.

i) Compute the initial entropy for the whole dataset. Show your workings (or justify your answer in one sentence). Please use $\log_2$ for your calculations.

ii) Compute the information gains for selecting the *colour* attribute and for selecting the *softness* attribute, each with respect to the initial entropy of the dataset. Please show all intermediate calculations. Please use $\log_2$ for all calculations.

iii) Based on your calculations in (ii), which attribute should be selected as the root node of a decision tree classifier?

b Assume that you have two Gaussian Mixture Models (GMMs). GMM_1 has 2 components, while GMM_2 has 3 components.

Suppose the parameters of the two GMMs are as in the table below, where μ_k is the mean of component k, σ_k the standard deviation for component k, and π_k the mixing proportion of component k.

GMM_1

k	μ_k	σ_k	π_k
1	3	1	0.6
2	8	0.5	0.4

GMM_2

k	μ_k	σ_k	π_k
1	2	0.5	0.4
2	5	1	0.3
3	8	1	0.3

i) Given a dataset X with five points $\{2,4,6,8,9\}$, which of these two GMMs fits this dataset better? Explain and justify your answer, showing any intermediate calculations. Please use the natural logarithm ln (log base e) for your calculations.

For your convenience, you may obtain the values of $\mathcal{N}(x|\mu,\sigma)$ directly from the table below.

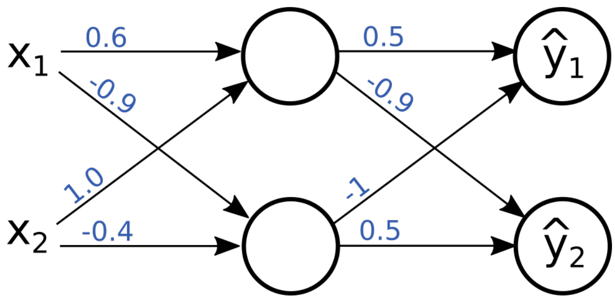
		x									
μ	σ	1	2	3	4	5	6	7	8	9	10
2	0.5	0.108	0.798	0.108	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	1.0	0.242	0.399	0.242	0.054	0.004	0.000	0.000	0.000	0.000	0.000
3	0.5	0.000	0.108	0.798	0.108	0.000	0.000	0.000	0.000	0.000	0.000
3	1.0	0.054	0.242	0.399	0.242	0.054	0.004	0.000	0.000	0.000	0.000
5	0.5	0.000	0.000	0.000	0.108	0.798	0.108	0.000	0.000	0.000	0.000
5	1.0	0.000	0.004	0.054	0.242	0.399	0.242	0.054	0.004	0.000	0.000
8	0.5	0.000	0.000	0.000	0.000	0.000	0.000	0.108	0.798	0.108	0.000
8	1.0	0.000	0.000	0.000	0.000	0.004	0.054	0.242	0.399	0.242	0.054

ii) In general, what is the effect of increasing the number of components K of a GMM? And what is the effect of increasing K to a very large value? (One to two sentences is enough for answering this question.)

The two parts carry equal marks.

2 a For this exercise, we represent words as 2-dimensional vectors. The word "exciting" is represented as a vector $[-1, 2]$.

We also have a neural network for classifying words based on their feature vector, shown in the figure. The two output neurons indicate the probability of the input word having negative ($\hat{y}_1$) or positive ($\hat{y}_2$) sentiment. The hidden layer uses tanh activation, the output layer uses softmax activation. The network is optimised with cross-entropy. Network weights are shown in the figure; there are no biases in the network.



Demonstrate the path of your calculations in this question. If you need to make any assumptions, state them clearly in your answer.

i) Using the given network, calculate the predicted output probability that the word "exciting" has positive sentiment.

ii) Representation learning is a technique that allows models to automatically discover useful features for an object or datapoint, such that they best perform on a particular task. For example, we can represent an item as a vector and then iteratively update this representation during training so that it leads to correct answers when given as input to a model. These vectors can then be used as informative features for some other task or network.

Perform one gradient descent step to update the **vector representation** for the word "exciting", using the given network. The correct target sentiment of the word is positive. Use learning rate 0.5. Give the final updated vector for the word "exciting".

Note: It is not necessary to update the weights between the layers, only the representation vector for the input word.

b Describe the neural network models and experiments you would design for the given scenarios. Specify the following details: 1) the number of input neurons, 2) the number of output neurons, 3) output activation function, 4) loss function for optimisation, 5) main evaluation metric.

i) We need to automatically determine whether two photos depict the same building. The system needs to work with any buildings, not only one specific building. You are able to use a pre-trained image encoder which provides 256-dimensional vector representations of each photo.

ii) We need to predict the number of cars and the number of bicycles that cross a particular intersection in a day. For each day, we are able to use the corresponding information from the previous 7 days to make the prediction.

The two parts carry, respectively, 60% and 40% of the marks.

3 For this question, we consider the following fictional scenario (no technical knowledge is required): A company wants to develop a new delivery service using drones. Your mission is to develop an algorithm that will automatically find the best drone design(s) for each of the delivery scenario that they may encounter. To help you in your mission, the company provides a simulator that takes a drone configuration (detailed below) and outputs the following characteristics of the drone design:

- **Maximum payload**, taking values between 1.0kg to 10.0kg.
- **Maximum parcel size**, taking values between 10.0cm to 50.0cm.
- **Average delivery time** (in minutes). If the delivery is impossible with this configuration, the value is set to -1.

The drone configuration offers the following parameters:

- **Width of the main frame**, taking values between 10.0cm to 35.0cm (float).
- **Length of the main frame**, taking values between 10.0cm to 35.0cm (float).
- **The number of motors**, taking values values between 3 to 6 (integer).
- **The type of motor** to be used (all the motors on the drone will share the same type). There are 4 possible types, numbered from 0 to 3 (integer).
- **The type of battery** to be used. There are 6 possible types, numbered from 0 to 5 (integer).

Here is an example of a drone configuration and the corresponding outputs from the simulator:
Configuration is $\{23.4, 13.9, 5, 2, 4\}$ while the output will be $\{8.8, 45.0, 24\}$.

If you need to make any assumption to answer one of the following sub-questions, please state it clearly.

We want to design an algorithm that outputs a collection of drone configurations offering different combinations of maximum payload and maximum parcel size, while for each combination providing the shorter delivery time. For this, we will use the MAP-Elites algorithm. We want the collection of configurations to contain as many configurations as possible, up to a maximum of 400 configurations.

a Explain in one or two sentences why this algorithm is appropriate in this use case.

b Define the fitness function and the behavioural descriptors (including their discretisation) that you would use in this case. Use one or two sentences for each of them.

c Give a suitable genotype, phenotype, function used to develop a genotype into a phenotype, and mutation operator, that you would use to solve the problem described above. Explain your answer in a few sentences (less than 15).

d After preliminary tests, you realise that the maximum payload and maximum parcel size values are somehow correlated and that not all the possible combinations are realistic. Indeed, it is impossible to build a drone that has a low maximum payload and that at the same time accept a large parcel size. The consequence is that the collection produced by your algorithm contains significantly fewer solutions than the targeted 400 ones.

During the preliminary tests, you have collected a large dataset of maximum payload and maximum parcel size pairs, corresponding to realistic designs. We assume that this dataset is a uniform and continuous coverage of this space (without any outliers).

We want to use this dataset to discretise the space of possible maximum payload and maximum parcel size combinations in 400 regions. These regions will then be used as a custom container for MAP-Elites, but we will not consider this part in this exam.

Which algorithm seen in this course can you use to automatically construct this discretisation into 400 regions of similar size. Explain in a few sentence how this algorithm works.

The four parts carry, respectively, 15%, 15%, 40%, and 30% of the marks.