2024 EXAMINATIONS

Lancaster University

Part II

COMPUTING AND COMMUNICATIONS – [2 hours and 30 minutes]

SCC.361 Artificial Intelligence

*Candidates are asked to answer **THREE** questions from **FOUR**; each question is worth a total of 25 marks.*

A calculator without memory is required.

Question 1

1.a Multiple Choice Questions:

i) Which of the following is a key limitation of k-means clustering? [1 mark]

a) k-means clustering cannot handle datasets with outliers effectively.

b) k-means clustering always converges to the same solution regardless of the initial cluster centroids.

c) k-means clustering is not suitable for datasets with high dimensionality.

d) k-means clustering cannot accommodate different distance functions.

ii) Consider the following statements about Generative Adversarial Networks (GANs):

A. GANs are a type of generative model that learns to generate new data samples by capturing the underlying distribution of the training data.

B. GANs consist of two neural networks: a generator and a discriminator. The generator aims to produce realistic data samples, while the discriminator aims to distinguish between real and fake samples.

C. Training a GAN involves optimizing two separate objective functions: the generator loss and the discriminator loss, which are optimized simultaneously.

Which of the statements above is/are true? [1 mark]

a) A only

b) B only

c) A and B only

d) A and B and C

iii) Consider a convolutional neural network (CNN) consisting of two convolution layers followed by a Multi Layer Perceptron and a softmax layer. The network takes an image as input and generates three outputs. Which of the following statements is correct? [2 marks]

a) If all activation functions are ReLU and all weights and biases are one, the output for the given image will be 0.3 for each of the three outputs.

b) If all activation functions are Sigmoid and all weights and biases are zero, the output for the given image will be 0.3 for each of the three outputs.

c) If all activation functions are Sigmoid and all weights and biases are one, the output for the given image will be zero for each of the three outputs.

d) If all activation functions are ReLU and all weights and biases are zero, the output for the given image will be zero for each of the three outputs.

[4 marks]

1.b You are given a dataset comprising the heights and ages of a small group of people, shown in the table below. You are asked to split the group into 2 subgroups using single-linkage hierarchical clustering and Manhattan distance.

	Height (cm)	Age (yrs)
P1	175	25
P2	164	74
P3	176	61
P4	177	78
P5	160	44
P6	166	12
P7	179	94
P8	162	23

i) Complete the distance matrix below by computing the values of $\delta_1, \delta_2, \delta_3, \delta_4$. [3 marks]

	P1	P2	P3	P4	P5	P6	P7	P8
P1	0							
P2	59	0						
P3	38	25	0					
P4	56	17	δ_1	0				
P5	33	34	33	51	0			
P6	δ_2	64	59	77	38	0		
P7	74	35	36	δ_3	69	95	0	
P8	14	53	52	70	δ_4	15	88	0

ii) Draw the dendrogram, starting with creating pairs, followed by single-linkage grouping. [3 marks]

iii) State how you can use the dendogram created in part ii to define a clustering of the data into 2. Write the membership of each group. [3 marks]

[9 marks]

1.c Using the table of data from **1.b**, k-Means clustering is performed with $k = 2$, Manhattan Distance, and initial centers (height=170, age=20) and (height=170, age=80). The distances in the below table are computed.

	P1	P2	P3	P4	P5	P6	P7	P8
Distance from cluster 1	9	60	47	65	34	12	83	11
Distance from cluster 2	59	12	25	9	46	72	23	65

i) Write down the membership of the 2 clusters at this stage and calculate the new centers of each cluster. [2 marks]

ii) Based on this, new distances are calculated (to 1 decimal place) and presented in the following table. State whether any further iterations are required and why. [1 mark]

	P1	P2	P3	P4	P5	P6	P7	P8
Distance from cluster 1	9.5	49.5	45.5	63.5	23.5	14.5	81.5	6.5
Distance from cluster 2	51.8	12.8	17.8	4.3	46.8	72.8	22.3	65.8

[3 marks]

1.d The perceptron algorithm is a fundamental building block in neural network models. It forms the basis of single-layer neural networks and plays a significant role in binary classification tasks. Consider a binary classification problem where we aim to classify data samples into two classes (Class 0 and Class 1) based on their features.

i) Explain the perceptron algorithm and its key components, including the perceptron model, activation function, weight initialization, and the update rule for adjusting the weights during training. [3 marks]

ii) Discuss the limitations of the perceptron algorithm when dealing with non-linearly separable data and explain why it fails to converge in such cases. Provide examples to illustrate your explanation. [3 marks]

iii) Given a dataset containing two classes of data points that are not linearly separable, demonstrate how you would modify the perceptron algorithm or incorporate additional components to improve its performance on the given dataset. [3 marks]

[9 marks]

Total 25 marks

Question 2

2.a Give the definition of an "agent" and state what it means for an agent to be considered "computational". Give the definition of artificial intelligence in terms of agents.

[3 marks]

2.b Multiple Choice Questions:

i) Which of the following is a characteristic of the k-nearest neighbours (k-nn) classification algorithm? [1 mark]

a) k-nn requires the assumption that the data is linearly separable.

b) k-nn is a parametric classification algorithm.

c) k-nn makes predictions based on the majority class of the nearest neighbours.

d) k-nn performs dimensionality reduction before making predictions.

ii) Which of the following statements are correct? [1 mark]

a) The memory to store facts is a positive of declarative knowledge.

b) Imperative knowledge is limited by the assumption that the past predicts the future.

c) Declarative knowledge focuses on generalisation while imperative knowledge focuses on memorisation.

d) Declarative knowledge involves the deduction of new facts from old facts.

iii) In machine learning, which of the following statements accurately describes the distinction between linearly separable and nonlinear decision boundaries? [2 marks]

a) Linearly separable boundaries can be represented by a straight line or plane, while nonlinear boundaries require complex curves or surfaces to separate the classes.

b) Linearly separable boundaries can handle only two classes, while nonlinear boundaries are capable of handling multiple classes.

c) Linearly separable boundaries are insensitive to outliers, while nonlinear boundaries are more robust in the presence of noisy data.

d) Linearly separable boundaries are less computationally efficient to train compared to nonlinear boundaries due to their inherent complexity.

[4 marks]

2.c You are given two sets of words: $S_1 = \{\text{EVEN}, \text{WELL}\}$ and $S_2 = \{\text{WEARY}, \text{CLEAR}\}$. For This question, you are asked to use k-nearest neighbours with $k = 1$ to decide to which set the word VERY should belong. Levenshtein distance will be used as the distance metric.

i) First, you computer the distances of the word VERY to each word in the sets S_1 and S_2. Most of this has been done in the tables below. Give the missing values δ_1 to δ_8. [4 marks]

		V	E	R	Y
	0	1	2	3	4
E	1	1	1	2	3
V	2	1	2	2	3
E	3	2	δ_1	2	3
N	4	3	2	δ_2	3

		V	E	R	Y
	0	1	2	3	4
W	1	1	2	3	4
E	2	2	1	2	3
L	3	3	δ_3	2	3
L	4	4	3	δ_4	3

		V	E	R	Y
	0	1	2	3	4
W	1	1	2	3	4
E	2	2	1	2	3
A	3	3	δ_5	2	3
R	4	4	3	δ_6	3
Y	5	5	4	3	2

		V	E	R	Y
	0	1	2	3	4
C	1	1	2	3	4
L	2	2	2	3	4
E	3	3	δ_7	3	4
A	4	4	3	δ_8	4
R	5	5	4	3	4

ii) Using the above tables, state the Levenshtein distances of the word VERY to each word in the sets S_1 and S_2. [2 marks]

iii) State the set to which the word VERY belongs according to k-nn with $k = 1$ and Levenshtein distance. [1 mark]

iv) Describe how you would extend this to $k = 3$ and state the set to which the word VERY belongs according to k-nn with $k = 3$ and Levenshtein distance. [2 marks]

[9 marks]

2.d Consider a deep neural network with four output neurons corresponding to four classes (Class A, Class B, Class C, and Class D). After processing an input sample through the network, the pre-softmax outputs for the four neurons are as follows:

Neuron 1: pre-softmax output= 2.5

Neuron 2: pre-softmax output= 1.8

Neuron 3: pre-softmax output= -0.9

Neuron 4: pre-softmax output= 0.7

i) Calculate the probabilities outputted by the softmax function. <u>You DO need to show your calculations.</u> [2 marks]

ii) Discuss whether applying a sigmoid function to each pre-softmax output can serve the same purpose as using the softmax function in this context. [3 marks]

[5 marks]

2.e Consider a Convolutional Neural Network (CNN) architecture designed for image classification tasks. The CNN consists of multiple convolutional layers followed by pooling layers and fully connected layers. The input images have dimensions of 32×32×3 (width, height, and channels), and the output layer consists of K neurons, where K represents the number of classes.

i) Explain the concept of parameter sharing in convolutional layers in one or two sentences. [2 marks]

ii) Discuss the role of pooling layers in CNNs in one or two sentences. [2 marks]

[4 marks]

Total 25 marks

Question 3

3.a State what overfitting means, the key issue that it causes for machine learning algorithms and describe a method of mitigating overfitting in decision trees.

[3 marks]

3.b Multiple Choice Questions.

i) Which of the following statements accurately describes a characteristic of decision trees? [1 mark]

a) Decision trees are exclusively used in regression analysis.

b) Decision trees cannot handle both categorical and numerical data.

c) Decision trees do not support handling of missing values in the dataset.

d) None of the above.

ii) Consider a neural network architecture that utilizes Transposed Convolution (also known as deconvolution) for upsampling in an image generation task. Which of the following statements accurately describes Transposed Convolution? [1 mark]

a) Transposed Convolution is a downsampling operation commonly used in convolutional neural networks (CNNs) to reduce the spatial dimensions of feature maps.

b) Transposed Convolution is a form of convolution that increases the spatial dimensions of feature maps, effectively upsampling the input data.

c) Transposed Convolution is a pooling operation used to extract important features from input images by selecting the maximum value within each pooling region.

d) Transposed Convolution is a regularization technique applied to CNNs to prevent overfitting by randomly dropping a fraction of neurons during training.

iii) In a binary classification task using a Multi-Layer Perceptron (MLP), a neural network is designed with the following specifications: The network consists of two hidden layers. The first hidden layer contains 8 neurons, and the second hidden layer consists of 6 neurons. The input to the network is flattened grayscale images with dimensions of 28×28 pixels. The output layer consists of a single neuron representing the binary classification. Each neuron in the network uses a sigmoid activation function. What is the total number of parameters in this MLP? [2 marks]

a) 6,340

b) 6,287

c) 6,341

d) 6,286

[4 marks]

3.c This question involves building a Decision Tree for the following drug trial.

	Sex (v_1)	BP (v_2)	Cholesterol (v_3)	Drug (c)
P1	M	HIGH	HIGH	A
P2	M	NORMAL	NORMAL	A
P3	F	HIGH	HIGH	A
P4	M	HIGH	NORMAL	A
P5	F	NORMAL	NORMAL	B
P6	F	HIGH	NORMAL	B
P7	M	HIGH	NORMAL	B
P8	F	NORMAL	NORMAL	B

i) Calculate the entropy of this dataset. You must show your working. [3 marks]

ii) Determine and state, with justification, the most important feature to start building a decision tree. You must show the calculations of the conditional probabilities and entropies necessary to make the decision. [6 marks]

[9 marks]

3.d Consider solving the toy problem of optimizing a simple mathematical function using a Genetic Algorithm (GA). The function to be optimized is:

$$f(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}) = {x_1}^2 + 2x_2 - 3x_3 + 4x_4 - 5x_5 + 6x_6 - 7x_7 + 8x_8 - 9x_9 + 10x_{10}$$

where $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}$ are real numbers in the range $[-10,10]$. You are tasked with implementing a Genetic Algorithm to find the global minimum of this function.

i) Define the chromosome representation and fitness function for the Genetic Algorithm in the context of solving this problem. [2 marks]

ii) Assume we have a population of 10 chromosomes, and their fitness values are as follows:

Chromosome 1: Fitness = $+8.5$
Chromosome 2: Fitness = $+7.2$
Chromosome 3: Fitness = $+6.8$
Chromosome 4: Fitness = $+9.1$
Chromosome 5: Fitness = $+5.6$
Chromosome 6: Fitness = $+8.9$
Chromosome 7: Fitness = -6.3
Chromosome 8: Fitness = $+7.8$
Chromosome 9: Fitness = $+6.7$
Chromosome 10: Fitness = $+7.0$

Now, using the provided fitness values and given the following random numbers:

1st random number = 0.2

2nd random number = 0.5

a) Identify two chromosomes which are selected as parents using the Roulette Wheel selection method and explain your reasoning for each selection. You DO need to show your calculations. [4 marks]

b) Identify two chromosomes which are selected as parents using the Rank selection method and explain your reasoning for each selection. You DO need to show your calculations. [3 marks]

[9 marks]

Total 25 marks

Question 4

4.a List the four related approaches to defining artificial intelligence and give an example of each.

[4 marks]

4.b Multiple Choice Questions.

i) Which of the following statements best describes a key assumption of the Naïve Bayes Classification approach? [1 mark]

a) Naïve Bayes requires that the dataset be linearly separable.

b) Naïve Bayes assumes that the features are independent of each other.

c) Naïve Bayes assumes that the features are dependent on each other.

d) Naïve Bayes is only applicable to binary classification problems.

ii) Which of the following statements accurately describes a key aspect of Conditional Generative Adversarial Networks (GANs)? [1 mark]

a) Conditional GANs utilize additional information, such as class labels, to guide the generation process.

b) Conditional GANs are primarily designed for unsupervised learning tasks, such as clustering and dimensionality reduction.

c) Conditional GANs do not involve a generator network and only focus on adversarial training of the discriminator.

d) Conditional GANs are exclusively used for image classification tasks and are not applicable to other types of data.

iii) Which technique, essential for spatial information reconstruction in convolutional neural networks (CNNs), is often employed as the inverse operation of max pooling? [2 marks]

a) Bilinear Interpolation

b) Random Forest Unpooling

c) Stochastic Gradient Descent Unpooling

d) Nearest-Neighbor Unpooling

[4 marks]

4.c Bayes' theorem is given as the following equation for events A and B where $P(B) \neq 0$:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

i) State which terms of the above Bayes' equation are considered as the following. [4 marks]

- "Evidence"

- "Likelihood"
- "Posterior probability"
- "Prior probability"

ii) On a pharmaceutical manufacturing line, it is found that (i) 5% of all tablets produced are defective, (ii) 95% of all tablets produced pass successfully through quality control, and (iii) 1% of tablets that pass successfully through quality control are defective. Given this information, use Bayes theorem to determine the probability that a defective tablet passes through quality control. [4 marks]

[8 marks]

4.d Suppose you are given two trained models for a 4-class classification problem. Each model employs a Softmax to generate its output probabilities as $[P_1, P_2, P_3, P_4]$, where P_1, P_2, P_3 and P_4 denote probability of the sample belonging to class A, class B, class C and class D, respectively. Model 1 and Model 2 have provided predictions for a set of 10 test samples as shown below:

Sample Number	Actual Class	Model 1 Predicted Output	Model 2 Predicted Output
1	D	[0.1, 0.2, 0.3, 0.4]	[0.3, 0.3, 0.2, 0.2]
2	A	[0.2, 0.4, 0.2, 0.2]	[0.4, 0.2, 0.3, 0.1]
3	B	[0.3, 0.2, 0.2, 0.3]	[0.2, 0.5, 0.1, 0.2]
4	C	[0.1, 0.1, 0.6, 0.2]	[0.2, 0.2, 0.2, 0.4]
5	D	[0.2, 0.2, 0.3, 0.3]	[0.3, 0.1, 0.4, 0.2]
6	A	[0.4, 0.3, 0.1, 0.2]	[0.2, 0.5, 0.2, 0.1]
7	B	[0.2, 0.2, 0.3, 0.3]	[0.3, 0.3, 0.1, 0.3]
8	C	[0.1, 0.3, 0.5, 0.1]	[0.2, 0.2, 0.3, 0.3]
9	D	[0.3, 0.2, 0.3, 0.2]	[0.2, 0.3, 0.3, 0.2]
10	B	[0.1, 0.3, 0.4, 0.2]	[0.2, 0.4, 0.1, 0.3]

i) Compute the Maximum Likelihood (ML) for both models based on the provided predictions. <u>You DO need to show your calculations.</u> [3 marks]

ii) Compute the Cross-Entropy (CE) of both models. <u>You DO need to show your calculations.</u> [3 marks]

iii) Discuss which model, Model 1 or Model 2, performs better based on the computed ML and CE values, and justify your answer in one sentence. [1 mark]

iv) Calculate the accuracy of both models. <u>You DO need to show your calculations.</u> Note that when there are two or more equal top probabilities for a sample, it means that the model is uncertain about the prediction and assigns equal probabilities to multiple classes. In such cases, we assume that the model incorrectly classifies the sample. [2 marks]

[9 marks]

Total 25 marks