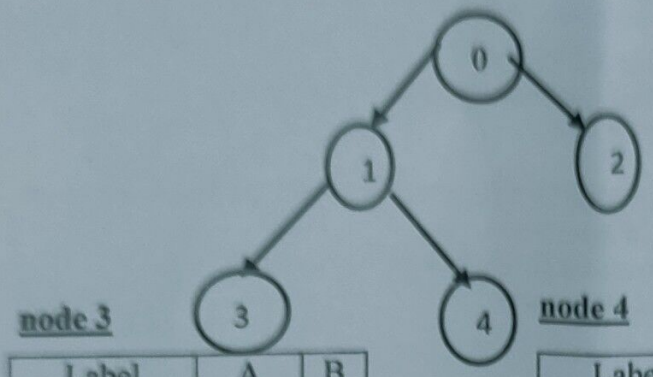
INDIAN INSTITUTE OF TECHNOLOGY ROORKEE
End-Term Examination (ETE)
Artificial Intelligence Techniques (IEC-03)

Time: 180 minutes **Total Marks: 115**

Instructions:

1. Each problem can be solved easily with a straightforward approach, and points may be lost for adding unnecessary complexity to the answers.
2. Write clearly and neatly. If your handwriting is hard to read or your answers are poorly organized, points may be deducted for lack of clarity.



node 2

Label	A	B
#Samples	12	0

node 3

Label	A	B
#Samples	4	5

node 4

Label	A	B
#Samples	2	7

1. The decision tree drawn above stores sample points from two classes, A and B. The tables indicate the number of sample points of each class stored in each leaf node. (Hint: Use entropy as the impurity function)
   a) What is entropy at the root (Treenode 0)? (2 marks)
   b) What is the information gain of the split at the root node (Treenode 0)? (3 marks)
   c) How can we increase the decision tree's training accuracy? Give a reasonable explanation for, why we might have considered doing that, but decided not to do that (2 marks)
   d) What happens if the tree is grown full? (1 marks)

2. Consider a linear regression model that is trained with mean square error (MSE) loss function and just one sample. Learning rate η is kept very less in order to ensure that weight values are not changed significantly. After performing one iteration of gradient decent weights of the model are updated. Prove that the value of loss function has decreased after one iteration. (Hint: Try to compare the values of loss functions before and after iteration) (8 marks)
3. Consider a binary classification problem with the following data points:

x_1	x_2	Class
2	2	-1
4	5	+1
7	4	+1

You are tasked with using a **Support Vector Machine (SVM)** to find the optimal hyperplane that separates the two classes. Assume that the data is linearly separable. Then answer the following question: (6+2+2= 10 marks)

   a. Find the hyperplane with maximum margin for the given data.
   b. Determine the support vectors.
   c. Calculate the margin.

4. Consider a Convolutional Neural Network (CNN) with the following architecture:

Input layer: A grayscale image of size 32×32×1 (Height H, Width W, and Channels C).
Convolutional layer (Conv1): Number of filters = 8, Filter size = 3×3, Stride = 1, Padding = 1 (same padding)
Pooling layer (Pool1): Type = Max pooling, Pool size = 2×2, Stride = 2,

Convolutional layer (Conv2): Number of filters = 16, Filter size = 3×3, Stride = 1, Padding = 0 (valid padding)
Fully Connected (FC) layer: Number of neurons = 10 (for a 10-class classification problem).
Then answer the following questions: (6 marks)

(a) Calculate the number of parameters for **Conv1**.
(b) Determine the size of the feature map after **Conv1**.
(c) Calculate the size of the feature map after **Pool1**.
(d) Calculate the number of parameters for **Conv2**.
(e) Determine the size of the feature map after **Conv2**.
(f) Calculate the total number of computations performed in **Conv1**.

5. In a GAN, the discriminator minimizes the mean squares loss rather than binary cross-entropy. For a batch of real images, the discriminator outputs: [0.9,0.85,0.95,0.92]. For a batch of generated images, the discriminator outputs: [0.2,0.15,0.25,0.3].

Then answer the following questions: (2+2+1=5 marks)

a) Calculate the mean squares loss for the real images.
b) Calculate the mean squares loss for the generated images.
c) Compute the total discriminator loss

6. Assume you are training a Generative Adversarial Network (GAN) with the following setup:
   a. The dataset contains 60,000 training samples.
   b. You use a batch size of 32 for both the generator and discriminator.
   c. You train the GAN for a total of 100 epochs.

Then answer the following questions (1+2+2=5 marks)

(a) How many iterations are performed by the generator during the entire training process?
(b) How many iterations are performed by the discriminator during the entire training process?
(c) You save the generated images from the generator and real images from the dataset to a buffer. Perform one additional update iteration of the discriminator using a new mini-batch from the saved images in the buffer. Then compute the total number of discriminator iterations performed throughout the training process (including the additional iteration).

7. (a)Explain the training method of generative adversarial networks. (8 marks)
(b)List down the problems faced during training of generative adversarial networks. Also mention ways to mitigate these problems. (6 marks)

8. (a) What is an attention mechanism?
(b) What is the difference between local and global attention mechanisms?
(c) Provide an example scenario where local attention and global attention can provide improvements in performance? (2+2+1=5 marks)

9. In Neural Networks, what is the advantage of ReLU function over sigmoid activation function? (1 mark)
(1) ReLU allows model to learn non-linear dependencies
(2) ReLU allows for faster backpropagation gradient calculations
(3) ReLU activation function can be used in output layer, while sigmoid can't be used.
(4) All of the above

10 (1.) Consider a sample of six data points $X \in R^2$, $\left\{\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix}, \begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}1\\2\end{bmatrix}, \begin{bmatrix}2\\1\end{bmatrix}, \begin{bmatrix}2\\2\end{bmatrix}\right\}$. (10 marks)

(a) [2 Marks] Compute the mean of the sample points and write the centered matrix (By subtracting the mean from each sample).
(b) [3 Marks] Find all the principal components of this sample. Write them as unit vectors.
(c) [5 Marks] [1+2+2] Find all the principal components of this sample. Write them as unit vectors.
   a. Which of those two principal components would be preferred if you use only one?
   b. What information does the PCA algorithm use to decide that one principal components is better

than another?

c. From an optimization point of view, why do we prefer that one?

(2.) Write correct answer for the following with calculation/justification (13 marks)

(a) After performing SVD on a data with 5 features, you retrieve eigenvalues 6,5,4,3,2. How many components should we include to explain atleast 75% of variance of the dataset? (2 marks)

(b) Relate the eigenvector and direction of variance in a dataset (1 marks)

(c) What is difference between squared error and absolute error, which one is better? (2 marks)

(d) Recall the objective of a soft SVM, (8 marks)

$\min(0.5 * |w|^2 + C\sum_1^m \zeta_i \text{ s.t. } y^i(w.x^i + \theta) \geq 1 - \zeta_i, \zeta_i \geq 0, \forall(x^i, y^i) \in D$, m is the number of examples, now state whether the following statements for SVM are correct, Use only one sentence in all cases to explain (no need of math derivation)

(a) When using C=0 one attains SVM hard objective

(b) Choosing higher values of C leads to overfitting the training data

(c) Slack variable ζ_i always takes value zero for a point when the point is correctly classified by hyperplane

(d) Optimal weight vector w can be calculated as a linear combination of the training data points?(you need not prove this)

11. a) Which of the following statements is **true:** (2 marks)

i) LSTM is approximately equivalent to a standard RNN when the forget gate is 0, input gate is 1 and output gate is 1 because of an additional activation function whose derivative is $f(x)*(1-f(x))$

ii) Internal state in RNN will not get updated as the sequence is processed

iii) In RNN, the state (h_t) completely forgets about the primordial (h_1) information when very long sequences are processed

iv) If the forget gate is 0, input gate is 1 and the output gate is 1 in an LSTM, the setting is almost a standard RNN

b) You are tasked with designing a variant of the LSTM model where the forget gate (f_t) and the input gate (i_t) are coupled, such that $i_t = 1 - f_t$ (8 marks)

i) Derive the equations for this coupled LSTM model, showing how the memory cell state (C_t) and the hidden state (h_t) are updated.

ii) Discuss the benefits and potential limitations of using a coupled forget and input gate in an LSTM.

12. a) An LSTM cell has an input size of 5, a hidden state size of 10, and a batch size of 3. Calculate the total number of weights required for the input, forget, cell, and output gates. Ignore the bias term in calculation. b) State the difference between LSTM and GRU c) An RNN is trained to predict the next word in a sentence. Why might the training performance degrade significantly if the sequences have varying lengths? How can this issue be mitigated? (4 + 3+ 3=10 marks)

13. a) Match the following:

1) One to many RNN architecture	i) Sentiment analysis of Netflix review
2) Many to one RNN architecture	ii) Convert French to English
3) Many to many RNN architecture	iii) Video Captioning
4) Many to many RNN architecture with unequal number of inputs and outputs	iv) Image Captioning

b) Consider a GRU cell with the following data:

$x_t = 2$

$h_{t-1} = -1$

$W_z = [1.2\ 1.4]$

$W_r = [-1.3, -1.6]$

$W = [-1, -0.7]$ Compute the following quantities (round upto 3 decimal places, bias values can be considered zero, refer formulas from sheet for computation). Update gate z_t, Reset gate r_t

New hidden state content $\tilde{h}_t$ Hidden state h_t (2 + 8 = 10 marks)

Linear Regression:

$Y = X\beta + \epsilon$; $X_{(N,P+1)}$, $\beta_{(P+1,1)}$, $\epsilon_{(N,1)}$

Least squares estimate: $\hat{\beta} = argmin_{\beta}(Y - X\beta)^T(Y - X\beta)$

$h(x) = \Sigma^n_{j=0}(\beta_j x_j)$; $J(\theta) = \frac{1}{2}\Sigma_{i=1}^{m}(h(x)^{(i)} - y^{(i)})^2$

To learn parameters $\theta(\beta_j)$, gradient descent updates weights as: $\beta_j = \beta_j - \alpha\frac{\partial J(\theta)}{\partial(\beta_j)}$

$\beta_j = \beta_j + \alpha(y^{(i)} - h(x^i))x_j^{(i)}$

Decision Tree

Information Gain IG = parent entropy (Ep) - Weighted Average entropy of children (Eavg)

SVM: Equation for hyperplane: $\overline{w}.\overline{x_i} - b = 0$

$$\phi(\alpha) = \sum_1^N \alpha_i - \frac{1}{2}\sum_1^N\sum_1^N \alpha_i\alpha_j y_i y_j(\overline{x_i}.\overline{x_j})$$

$$\overline{w} = \sum_1^N \alpha_i y_i \overline{x_i}$$

$$b = \frac{1}{2}\left(min_{i;yi=+1}(\overline{w}.\overline{x_i}) + max_{i;yi=-1}(\overline{w}.\overline{x_i})\right)$$

CNN

Feature map size after convolution: (Batchsize, #filters, H_{out}, W_{out})

Where, $H_{out} = \frac{H_{in}-K+2P}{S} + 1$, $W_{out} = \frac{W_{in}-K+2P}{S} + 1$

#parameters of convolution layer = $K * K * C * \#filters$

Computations for convolution layer = $K * K * C * \#filter * H_{out}, W_{out}$

#parameters of fully connected layer = $N_{in} * N_{out}$, where N_{in} and N_{out} denote input and output neurons.

Computations for fully connected layer = $N_{in} * N_{out}$

LSTM

$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f]$

$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i]$

$\overline{c_t} = \tanh(W_c.[h_{t-1}, x_t] + b_c]$

$C_t = f_t * C_{t-1} + i_t * \overline{c_t}$

$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o]$

$h_t = o_t * \tanh(C_t)$

GRU

$z_t = \sigma(W_z.[h_{t-1}, x_t] + b_z]$

$r_t = \sigma(W_r.[h_{t-1}, x_t] + b_r]$

$\overline{h_t} = \tanh(W.[r_t.h_{t-1}, x_t] + b]$

$h_t = (1 - z_t) * h_{t-1} + z_t * \overline{h_t}$