National University of Computer and Emerging Sciences

Islamabad Campus

Generative AI (AI4009)

Sessional-I Exam

Course Instructor(s):

Dr. Akhtar Jamil

Section(s): CS and AI (A,B)

Total Time (Hrs):	**1**
Total Marks:	**50**
Total Questions:	**3**

Date: Feb 26, 2025

_____________ _____________ _________________________

Roll No **Course Section** **Student Signature**

Do not write below this line.

Attempt all the questions.

Question No. 1. MCQ [1 x 25 = 25]

Answer the MCQs on the given answer sheet attached at the end of the question paper. Answers marked on the question paper will not be evaluated.

1. **Which of the following best describes Maximum Likelihood Estimation (MLE)?**
   a) It estimates the parameters of a probability distribution
   b) A technique used to minimize the residuals in a regression model
   c) A method to standardize a dataset for better training performance
   d) A statistical test to determine if two distributions are the same
2. **What is the effect of setting the regularization parameter λ too large?**
   a) The model becomes more likely to underfit the data.
   b) The model becomes more likely to overfit the data.
   c) The model's capacity to generalize improves.
   d) The optimization process becomes faster.
3. **Suppose we have a dataset of 100 coin flips with 60 heads and 40 tails. Using MLE, what is the estimated probability** p **of getting heads?**
   a) 0.40
   b) 0.50
   c) 0.60
   d) 0.70
4. **For a normal distribution $\mathcal{N}(\mu, \sigma^2)$, what is the MLE estimate of μ given a sample $x_1, x_2, \dots, x_n$?**
   a) The median of the sample
   b) The mode of the sample
   c) **The mean of the sample**
   d) The variance of the sample
5. **What is the primary purpose of transpose convolutions (also known as deconvolutions) in neural networks?**
   a) To increase the depth of the feature maps.
   b) To reduce the spatial dimensions of the input.
   c) To perform element-wise multiplication between feature maps.
   d) To increase the spatial dimensions of the input.

6. **Since the** learning **rate changes dynamically across epochs in scheduled learning rate strategies, can it be considered a model parameter?**
   a) True
   b) False

7. **For which of the following problems, RNN is not suitable:**
   a) Time series forecasting
   b) Sentiment analysis on text data
   c) Image classification
   d) Speech recognition
8. **Why is it crucial for latent variable models, such as VAEs, to enforce a Gaussian distribution in the latent space?**
   a) It ensures that at test time, we can sample z from a Gaussian distribution to generate meaningful outputs
   b) It prevents overfitting by limiting the complexity of the decoder network
   c) It ensures that every latent vector uniquely corresponds to a specific data sample
   d) It eliminates the need for regularization in the encoder network
9. **Why do training GANs require alternating updates between the generator and the discriminator?**
   a) To allow the generator to train without requiring the discriminator
   b) To reduce the number of parameters in the model
   c) To ensure the latent space remains structured
   d) To ensure both networks improve together rather than overpowering each other
10. **In Variational Autoencoders (VAEs), what does the KL-Divergence term in the loss function ensure?**
    a) The latent distribution closely follows the prior Gaussian distribution
    b) The generated samples are sharper than original data
    c) The VAE learns deterministic representations of input data
    d) The encoder and decoder share the same parameters
11. **What is the purpose of the entropy term in a VAE?**
    a) To encourage diversity in the latent space
    b) To sharpen the reconstructed outputs
    c) To minimize reconstruction error
    d) To enforce a strict mapping between inputs and latent vectors
12. **What does the receptive field of a neuron in a convolutional neural network (CNN) refer to?**
    a). The number of filters applied to an input.
    b). The number of parameters in a convolutional layer.
    c). The stride and padding used in convolution.
    d). **The area of the input image that influences the activation of a neuron.**
13. **Long-range dependencies can be introduced in autoregressive models using**
    a). CNN
    b). ANN
    c). RNN
    d). None of the above

14. Consider a model that learns the following joint probability distribution for three sequential inputs X_1, X_2, X_3 and the label Y:

$$P(X_1, X_2, X_3, Y) = P(Y) \cdot P(X_1|Y) \cdot P(X_2|X_1, Y) \cdot P(X_3|X_2, X_1, Y)$$

Is this model generative or discriminative?

a) Generative
b) Discriminative
c) None

15. Which of the following is not a hyperparameter

a) Learning rate
b) Batch size
c) Epochs
d) Bias

16. **Why do we often take the logarithm of the likelihood function in MLE calculations?**

a) To make the calculations easier by converting products into sums
b) To reduce the effect of small likelihood values
c) To ensure convexity of the optimization problem
d) **All of the above**

17. **If we assume a Bernoulli distribution for a coin flip experiment, what is the likelihood function for the dataset $\boldsymbol{X = \{1, 0, 1, 1, 0\}}$?**

a) $\boldsymbol{L(p) = p^3(1-p)^2}$
b) $L(p) = p^3(1-p)^3$
c) $L(p) = p(1-p)$
d) $L(p) = p^5(1-p)^5$

18. Which activation function is most appropriate if your model needs to output probabilities for multiple classes?

a) ReLU
b) Tanh
c) Softmax
d) None

19. When a model has high bias, this indicates that:

a) The model is overfitting
b) The model is underfitting
c) Converge very fast to the minimum error.
d) None of the above

20. Consider an input of size 64×64×10. To reduce the depth to 5 without changing the spatial dimensions, which of the following strategies would be the most suitable?

a) Use a convolution with a kernel size of 3×3, and 5 filters.
b) Use a convolution with a kernel size of 1×1, and 5 filters.
c) Apply max pooling with a pool size of 2×2.
d) None of the above.

21. For a convolutional layer with an input size of 5×5×3 and 3 filters of size 3×3, how many trainable parameters (excluding bias) are required?

a) 27
b) 54
c) 81
d) 243

22. Given the following probabilities, what is the joint probability $P(X = 2, Y = 1)$?

$P(X = 1) = 0.4, P(X = 2) = 0.6$, $P(Y = 1|X = 1) = 0.3$, $P(Y = 1|X = 2) = 0.5$

a) 0.20
b) 0.24
c) 0.30
d) 0.50

23. What primarily causes the vanishing gradient problem in RNNs?

a) Small gradient values exponentially decay over many time steps
b) Lack of recurrent connections in the network
c) Zero-initialized weight matrices
d) No weight sharing across time steps

24. In a Many-to-Many RNN architecture, how is the total loss computed during training?

a) By computing the loss only at the final time step
b) By comparing output with the given label at the output layer
c) By using only the first time step for loss computation
d) By summing the loss over all time steps

25. Which of the following is NOT an advantage of RNNs?

a) RNNs can process inputs of any length
b) RNNs model size remains constant regardless of input length
c) They effectively capture long-range dependencies without issues
d) RNNs utilize historical information during processing

Question No 2. Write short answers to the following questions. [3 x 5=15]

a). What is the reparameterization trick used in VAEs?

The reparameterization trick is used in Variational Autoencoders (VAEs) to enable backpropagation through stochastic nodes. Instead of sampling directly from a Gaussian distribution $z \sim \mathcal{N}(\mu, \sigma^2)$, it is rewritten as $z = \mu + \sigma \cdot \epsilon$, where $\epsilon \sim \mathcal{N}(0,1)$. This allows gradients to flow through the network during training.

b). Explain why VAEs tend to generate blurry images compared to GANs.

VAEs are trained to maximize the probability (likelihood) of the real data under a probabilistic model. They try to minimize the reconstruction error between the generated image and the original image. However, because the latent space is continuous and smooth, the model often produces an output that is the average of many probable images rather than a sharp, distinct one.

c). Write three disadvantages of RNN.

1. During backpropagation through time (BPTT), gradients can become very small (vanish) or very large (explode), making it difficult for the model to learn long-term dependencies.
2. RNNs process data sequentially, which prevents parallelization and makes training slower compared to models like CNNs or Transformers.

3. Standard RNNs struggle to retain information over long sequences, leading to poor performance on tasks requiring long-range context.

d). What is the main disadvantage of finite memory autoregressive models?

Finite memory autoregressive models struggle with long-term dependencies because they only condition on a fixed-length context window. This limits their ability to capture long-range patterns in sequences, making them less effective for tasks like long-text generation.

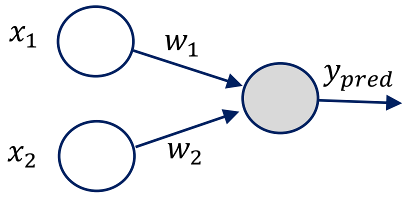
e). What is the role of regularization in training deep neural networks?

Regularization techniques prevent overfitting by constraining the model's capacity or adding noise to training. This improves generalization to unseen data by reducing reliance on specific patterns in the training set. Example of regularization include such as L1/L2 regularization, dropout, and batch normalization

Question No. 3 Consider the neural network with just one neuron as shown. [10]

Consider an Artificial Neural Network (ANN) with two inputs ($x_1 = 2$) and ($x_2 = 3$) and a single output (y_{pred}). The network has a single layer where both weights ($w_1 = w_2 = 0.5, b = 0.1$), sigmoid activation function. You are given the true target value ($y_{true} = 10$). The loss function is the Mean Squared Error (MSE), defined as:

$$L = \frac{1}{2}\left(y_{true} - y_{pred}\right)^2$$



1. Calculate the output (y_{pred}) for the given inputs and weights.
2. Perform one step of backpropagation to update the weights using gradient descent with a learning rate ($\eta = 0.01$). Calculate the new values for w_1 and w_2.

Question No. 3: Consider the neural network with just one neuron as shown. [10]

Given:

Inputs: $x_1 = 2$, $x_2 = 3$

Initial weights: $w_1 = 0.5$, $w_2 = 0.5$

Bias: $b = 0.1$

Target: $y_{\text{true}} = 10$

Activation Function: Sigmoid

Loss Function (MSE):

$$L = \frac{1}{2}(y_{\text{true}} - y_{\text{pred}})^2$$

Learning rate: $\eta = 0.01$

Step 1: Calculate the output y_{pred}

Net input:

$$z = w_1x_1 + w_2x_2 + b = (0.5)(2) + (0.5)(3) + 0.1 = 2.6$$

Sigmoid activation:

$$y_{\text{pred}} = \sigma(z) = \frac{1}{1 + e^{-2.6}} \approx 0.931$$

Step 2: Backpropagation and Weight Update

Gradient of the loss:

$$\frac{\partial L}{\partial y_{\text{pred}}} = -(y_{\text{true}} - y_{\text{pred}}) = -9.069$$

Derivative of the sigmoid:

$$\frac{\partial y_{\text{pred}}}{\partial z} = y_{\text{pred}}(1 - y_{\text{pred}}) = 0.931(1 - 0.931) \approx 0.0642$$

Partial derivatives:

$$\frac{\partial z}{\partial w_1} = x_1 = 2, \quad \frac{\partial z}{\partial w_2} = x_2 = 3$$

Gradient with respect to weights:

$$\frac{\partial L}{\partial w_1} = (-9.069)(0.0642)(2) \approx -1.164$$

$$\frac{\partial L}{\partial w_2} = (-9.069)(0.0642)(3) \approx -1.746$$

Weight updates:

$$w_1 = w_1 - \eta \cdot \frac{\partial L}{\partial w_1} = 0.5 + 0.01164 = 0.5116$$

$$w_2 = w_2 - \eta \cdot \frac{\partial L}{\partial w_2} = 0.5 + 0.01746 = 0.5175$$