Introduction to Large Language Models

Week-3 Assignment

Number of questions: 10 **Total mark: 10 X 1 = 10**

Question 1:

In backpropagation, which method is used to compute the gradients?
a. Gradient descent
b. Chain rule of derivatives
c. Matrix factorization
d. Linear regression

Correct Answer: b
Solution: Backpropagation uses the chain rule of derivatives to calculate the gradients layer by layer.

Question 2:

Which of the following functions is **not differentiable at zero**?

a. Sigmoid
b. Tanh
c. ReLU
d. Linear

Correct Answer: c

Solution: ReLU is not differentiable at zero since the left and right limits of the derivative are not equal.

Question 3:

In the context of regularization, which of the following statements is true?

a. L2 regularization tends to produce sparse weights
b. Dropout is applied during inference to improve accuracy
c. L1 regularization adds the squared weight penalties to the loss function
d. Dropout prevents overfitting by randomly disabling neurons during training

Correct Answer: d
Solution: Dropout deactivates neurons randomly during training to prevent overfitting.

Question 4:

Which activation function is least likely to suffer from vanishing gradients?

a. Tanh
b. Sigmoid
c. ReLU

Correct Answer: c

Solution: Its gradient is 1 for positive input and 0 for negative input, so it allows gradients to flow effectively

---

Question 5:

Which of the following equations correctly represents the derivative of the sigmoid function?

a. $\sigma(x) \cdot (1 + \sigma(x))$
b. $\sigma(x)^2$
c. $\sigma(x) \cdot (1 - \sigma(x))$
d. 1 / (1 + e^x)

Correct Answer: c
Solution: The derivative of sigmoid $\sigma(x)$ is $\sigma(x)(1 - \sigma(x))$.

Question 6:

What condition must be met for the Perceptron learning algorithm to converge?

a. Learning rate must be zero
b. Data must be non-linearly separable
c. Data must be linearly separable
d. Activation function must be sigmoid

Correct Answer: c

Question 7:

Which of the following logic functions requires a network with at least one hidden layer to model?

a. AND
b. OR

c. NOT
d. XOR

Correct Answer: d
Solution: XOR is the classic example of a non-linearly separable function.

Question 8:

Why is it necessary to include non-linear activation functions between layers in an MLP?

a. Without them, the network is just a linear function
c. They prevent overfitting
d. They allow backpropagation to work

Correct Answer: a
Solution: Without non-linearity, stacking linear layers results in another linear function — limiting the model's expressiveness.

Question 9:

What is typically the output activation function for an MLP solving a binary classification task?

a. Tanh
b. ReLU
c. Sigmoid
d. Softmax

Correct Answer: c
Solution: For binary classification, the output is usually a single unit with a sigmoid activation.

Question 10:

Which type of regularization encourages sparsity in the weights?
a. L1 regularization
b. L2 regularization
c. Dropout
d. Early stopping

Correct Answer: a
Solution: L1 regularization encourages sparsity in the weights.
