Tutorial Questions

Practice Problems for Quiz Preparation

1. A McCulloch-Pitts neuron is required to implement the **OR** function with three binary inputs $x_1, x_2, x_3 \in \{0, 1\}$. If all weights are set to $w_1 = w_2 = w_3 = 1$, what threshold θ should be used?

   A) $\theta = 0$

   B) $\theta = 1$

   C) $\theta = 2$

   D) $\theta = 3$

2. A perceptron with weight vector $\mathbf{w} = [2, -1, 3]^T$ and bias $b = -2$ is trained on a linearly separable dataset using the perceptron learning algorithm with learning rate $\eta = 0.5$. If a misclassified point $\mathbf{x} = [1, 2, -1]^T$ with label $y = -1$ is encountered, what is the magnitude of the updated weight vector $\|\mathbf{w}_{new}\|$?

3. In the context of the XOR problem, why can't a single perceptron solve it?

   A) The perceptron learning rate is too small

   B) XOR is not a linearly separable function

   C) The perceptron can only handle binary inputs

   D) The threshold cannot be set appropriately

4. Consider a dataset with 4 points in 2D: $\mathbf{x}_1 = [1, 2]$ (class +1), $\mathbf{x}_2 = [2, 1]$ (class +1), $\mathbf{x}_3 = [-1, -1]$ (class -1), $\mathbf{x}_4 = [-2, -2]$ (class -1). What is the maximum margin γ achievable by any linear classifier?

   A) $\gamma = \frac{1}{\sqrt{2}}$

   B) $\gamma = 1$

   C) $\gamma = \sqrt{2}$

   D) $\gamma = 2$

5. Consider a 2-layer neural network attempting to learn the XOR function. The hidden layer has 2 neurons with weights:

$$W^{(1)} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \quad \mathbf{b}^{(1)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

   Using tanh activation in the hidden layer, compute the hidden representation $\mathbf{h}$ for input $\mathbf{x} = [1, 1]^T$. What is $\|\mathbf{h}\|$?

6. In a deep network with L layers, all using sigmoid activations $\sigma(z)$, the weights are initialized such that the variance of activations is preserved across layers. If the input dimension is n_{in} and output dimension is n_{out}, the weights are sampled from $\mathcal{N}(0, \sigma_w^2)$. Derive the required variance σ_w^2 for variance preservation, assuming linear activations initially (for derivation purposes).

   A) $\sigma_w^2 = \frac{1}{n_{in}}$

   B) $\sigma_w^2 = \frac{2}{n_{in}+n_{out}}$

   C) $\sigma_w^2 = \frac{1}{n_{out}}$

   D) $\sigma_w^2 = \sqrt{\frac{2}{n_{in}}}$

7. A neural network has 2 hidden layers with ReLU activations and a sigmoid output layer. During forward propagation, which statement about the output range is correct?

A) Output is in $(-\infty, \infty)$

B) Output is in $(0, 1)$

C) Output is in $(-1, 1)$

D) Output is in $[0, \infty)$

8. A neural network uses the following activation:

$$f(z) = \begin{cases} z & \text{if } z > 0 \\ \alpha(e^z - 1) & \text{if } z \leq 0 \end{cases}$$

For $\alpha = 1.0$, compute $f(-1)$ and $f'(-1)$ (rounded to 3 decimal places).

9. A network has a bottleneck layer that compresses d-dimensional input to k-dimensional representation ($k < d$) using:

$$\mathbf{h} = \tanh(W\mathbf{x} + \mathbf{b})$$

where $W \in \mathbb{R}^{k \times d}$. If the network perfectly reconstructs the input through a decoder, what is the maximum number of linearly independent vectors in the input space that can be perfectly preserved?

A) d

B) k

C) $2k$

D) $\min(d, 2k)$

10. Given a 3-class classification problem with true label $y = [0, 1, 0]$ (one-hot) and predicted logits $z = [2.0, 1.0, 0.5]$, compute the Cross-Entropy loss after applying softmax.

11. Consider the activation: $f(z) = \max(0.01z, z)$. Compute $f(-5)$ and $f'(-5)$.

12. For a network with weight matrix $W \in \mathbb{R}^{3 \times 2}$ given by:

$$W = \begin{bmatrix} 0.5 & -0.3 \\ 0.2 & 0.4 \\ -0.1 & 0.6 \end{bmatrix}$$

and error signal $\boldsymbol{\delta} = [0.2, -0.4, 0.3]^T$, activation from previous layer $\mathbf{a}_{prev} = [1.5, 0.8]^T$. Compute the gradient $\nabla_W \mathcal{L}$.

13. In a network, the forward pass is:

$$\mathbf{a}^{(l+1)} = \mathbf{a}^{(l)} + F(\mathbf{a}^{(l)}, W^{(l)})$$

where F is a non-linear transformation. During backpropagation through L layers, what is the minimum value of $\frac{\partial \mathcal{L}}{\partial \mathbf{a}^{(1)}}$ relative to $\frac{\partial \mathcal{L}}{\partial \mathbf{a}^{(L)}}$ (assuming $\|\frac{\partial F}{\partial \mathbf{a}}\| \leq 0$)?

A) 0 (vanishes completely)

B) $(1)^L = 1$ (preserved exactly)

C) Exponentially small: $\approx 0.25^L$

D) Cannot be determined without knowing F

14. A mini-batch of size $b = 32$ has gradients with variance $\sigma_g^2 = 4.0$. To achieve the same gradient variance as a batch of size $B = 128$ using gradient accumulation, how many mini-batches must be accumulated, and what will be the effective variance?

15. For a binary classification network with sigmoid output, the loss is Binary Cross-Entropy:

$$\mathcal{L} = -[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})]$$

If the network predicts $\hat{y} = 0.95$ for a positive example ($y = 1$), and we perform gradient descent with learning rate $\eta = 0.1$, by how much does the pre-activation z change in one step? Use the fact that $\hat{y} = \sigma(z)$ and $\frac{\partial \mathcal{L}}{\partial z} = \hat{y} - y$.

16. During backpropagation in a network with skip connections:

$$a^{(l+1)} = a^{(l)} + F(a^{(l)}, W^{(l)})$$

If $\frac{\partial \mathcal{L}}{\partial a^{(l+1)}} = \delta^{(l+1)}$, what is $\frac{\partial \mathcal{L}}{\partial a^{(l)}}$?

A) $\delta^{(l+1)}$

B) $\delta^{(l+1)} \cdot \frac{\partial F}{\partial a^{(l)}}$

C) $\delta^{(l+1)} \left(1 + \frac{\partial F}{\partial a^{(l)}}\right)$

D) $\frac{\partial F}{\partial a^{(l)}}$

17. Given a mini-batch gradient descent setup with batch size $b = 64$, learning rate $\eta = 0.01$, and a dataset of size $N = 10,000$:

a) How many parameter updates occur in one epoch?

b) If the training runs for 50 epochs, what is the total number of gradient computations?

18. Derive the gradient of the sigmoid activation with respect to its input:

Given $\sigma(z) = \frac{1}{1+e^{-z}}$, show that:

$$\frac{d\sigma}{dz} = \sigma(z)(1 - \sigma(z))$$

Start from the definition and use the chain rule.