COURSE 101 HW 1

Fall 2025 Test Student

Part 1: Mathematical Foundations

Problem 1.1: Function Analysis

Consider a general function with multiple inputs and outputs for analysis.

A.

Parameter Count: State the formula for the total number of parameters (N_{params}) in a system with m inputs, n processing units, and p outputs. Calculate N_{params} if $m = 5$, $n = 100$, and $p = 3$.

Solution:

The total number of parameters is:

$$N_{\text{params}} = (m \times n + n) + (n \times p + p)$$
$$N_{\text{params}} = m \cdot n + n + n \cdot p + p$$

For $m = 5$, $n = 100$, and $p = 3$:

$$N_{\text{params}} = 5 \times 100 + 100 + 100 \times 3 + 3$$
$$N_{\text{params}} = 500 + 100 + 300 + 3$$
$$N_{\text{params}} = 903$$

B.

System Structure: For a simplified version where $m = 2$, $n = 3$, and $p = 1$, state the dimensions of the weight matrices (W) and bias vectors (b) for each layer.

Solution:

- First layer: $W_1 \in \mathbb{R}^{3\times2}$, $b_1 \in \mathbb{R}^{3\times1}$
- Second layer: $W_2 \in \mathbb{R}^{1\times3}$, $b_2 \in \mathbb{R}^{1\times1}$

C.

Theoretical Properties: Explain the mathematical properties of the composition of piecewise linear functions.

Solution:

A composition of piecewise linear functions remains piecewise linear because:

1. Each component applies a linear transformation
2. The composition preserves the piecewise structure
3. The total number of linear segments depends on the number of components
4. For one-dimensional input, the maximum segments is $n + 1$ for n components

Problem 1.2: Regional Analysis

A.

Calculate the maximum number of regions in a space based on the formula for k hyperplanes in an m-dimensional space ($k > m$).

Solution:

The formula is:

$$R(k,m) = \sum_{i=0}^{m} \binom{k}{i}$$

For $m = 2$ and $k = 4$:

$$\begin{aligned} R(4,2) &= \binom{4}{0} + \binom{4}{1} + \binom{4}{2} \\ R(4,2) &= 1 + 4 + 6 \\ R(4,2) &= 11 \end{aligned}$$

Part 2: Optimization Methods

Problem 2.1: Cost Function Analysis

Derive properties of a general cost function for classification problems.

A.

Write the general form of an entropy-based cost function for a binary problem.
Solution:

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

B.

Compute the derivative with respect to the predicted value $\hat{y}$.
Solution:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \hat{y}} &= -\left[\frac{y}{\hat{y}} - \frac{1-y}{1-\hat{y}}\right] \\ \frac{\partial \mathcal{L}}{\partial \hat{y}} &= \frac{\hat{y}-y}{\hat{y}(1-\hat{y})} \end{aligned}$$

Problem 2.2: Regularization Methods

A.

Explain two common regularization approaches and their effects on model parameters.
Solution:

- **L2 Regularization**: Adds $\lambda \sum_i w_i^2$ to the loss. Encourages small weights but rarely exactly zero. Provides smooth parameter distributions.
- **L1 Regularization**: Adds $\lambda \sum_i |w_i|$ to the loss. Encourages sparse solutions with many weights exactly zero. Useful for feature selection.

B.

When would you prefer one regularization method over another?
Solution:
Prefer L1 when:

1. You want automatic feature selection
2. You believe only a few features are truly relevant
3. Interpretability is important (sparse models are easier to interpret)
4. Storage/computation efficiency is critical

Prefer L2 when:

1. All features contribute somewhat to the prediction
2. You want smooth parameter distributions
3. Numerical stability is a concern
4. You're dealing with correlated features

Part 3: Algorithm Analysis

Problem 3.1: Gradient-Based Methods

Consider a system with layered structure. Derive the update equations.

A.

Write the forward propagation equations for a two-layer system.
Solution:

$$\begin{aligned} z^{[1]} &= W^{[1]}x + b^{[1]} \\ a^{[1]} &= \sigma(z^{[1]}) \\ z^{[2]} &= W^{[2]}a^{[1]} + b^{[2]} \\ \hat{y} &= \sigma(z^{[2]}) \end{aligned}$$

B.

Derive the gradient for the second layer weights using the chain rule.
Solution:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial W^{[2]}} &= \frac{\partial \mathcal{L}}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial W^{[2]}} \\ \frac{\partial \mathcal{L}}{\partial W^{[2]}} &= \delta^{[2]} \cdot (a^{[1]})^T \end{aligned}$$

where $\delta^{[2]} = \frac{\partial \mathcal{L}}{\partial \hat{y}} \cdot \sigma'(z^{[2]})$.

This test demonstrates all three parts with hierarchical numbering: 1.1, 1.2, 2.1, 2.2, 3.1