Assignment - 4

Problem - 1 Architectural Approaches

Approach a treat it as a regression task ~~are~~ where as approach B treat it as a classification task.

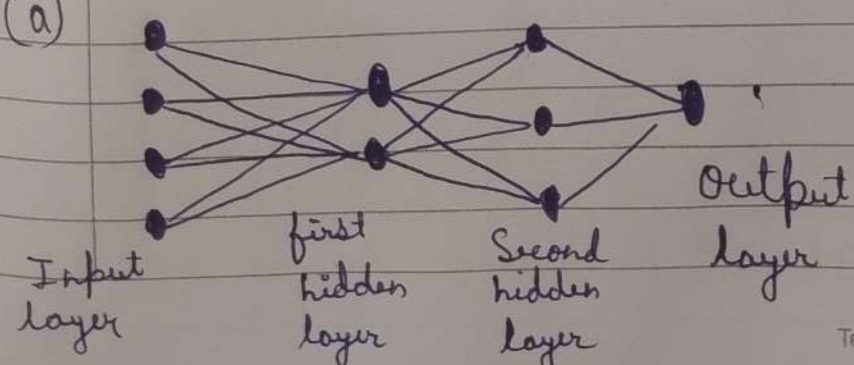
- Implied Ordinality :- This approach assumes there's an ordinal relational b/w digits which is meaning less in digit recognition

- Loss function behaviour :- if the network sees an image of a '9' but predicts '1', the error is massive if compared to predicting '8'. However misclassifying '9' as '1' is as wrong as misclassifying it to be '8'.

- Approach B is superior :- Classification using one hot encoding and softmax treats each digit as a distinct, independent class, which correctly align with the nature of the problem.

~~Problem 2~~ Problem 2 Neural Network

(a)



Teacher's Signature:

(b) let $x \in R^4$ be the input vector $x = (x_1, x_2, x_3, x_4)^T$

First ~~Layer~~ hidden layer (2 units)

$W^1 \in R^{2\times 4}$ $b^{(1)} \in R^2$.

Pre-activation

$z_1^{(1)} = \sum_{j=1}^{4} w_{1j}^{(1)} x_j + b_1^{(1)}$ $z_2^{(1)} = \sum w_{2j}^{(1)} x_j + b_2^{(1)}$

Activation $h_1^{(1)} = \sigma(z_1^{(1)})$ $h_2^{(1)} = \sigma(z_2^{(1)})$

Second hidden layer (3 units)

Pre-activations

$z_1^{(2)} = w_{11}^{(2)} h_1^{(1)} + w_{12}^{(2)} h_2^{(1)} + b_1^{(2)}$

$z_2^{(2)} = w_{21}^{(2)} h_1^{(1)} + w_{22}^{(2)} h_2^{(1)} + b_2^{(2)}$

$z_3^{(2)} = w_{31}^{(2)} h_1^{(1)} + w_{32}^{(2)} h_2^{(1)} + b_3^{(2)}$

Activation $h_k^{(2)} = \sigma(z_k^{(2)})$, $k = 1, 2, 3$

Output layer $f(x) = w_1^3 h_1^2 + w_2^3 h_2^2 + w_3^3 h_3^2 + b^3$

Fully explicity

$$f(x) = \sum_{k=1}^{3} w_k^{(3)} \max\left(0, \sum_{i=1}^{2} w_{ki}^{(2)} \max\left(0, \sum_{j=1}^{4} w_{ij}^{(1)} x_j + b_i^{(1)}\right) + b_k^{(2)}\right) + b^{(3)}$$

Teacher's Signature:

(c) Calculated value of $f(X)$

input: $x = [1, 1, 1, 1]^T$ All weights = 1
All biases : $b = 0$

1. Hidden layer 1:
$$z_1 = \sum x_i = 1+1+1+1 = 4$$
Activation $\sigma(4) = 4$ Since there are 2 units,
output vector $h^{(1)} = [4, 4]^T$

2. Hidden layer 2

$z_2 = (1 \cdot 4) + (1 \cdot 4) = 8$

Activation: $\sigma(8) = 8$ Since there are 3 units
$h^{(2)} = [8, 8, 8]^T$

3 Output layer Input is $[8, 8, 8]^T$

$$y = (1 \cdot 8) + (1 \cdot 8) + (1 \cdot 8) = 24$$

d) Total parameter

- Input → Hidden 1: (4 inputs × 2 neurons) + 2 biases = 10
Hidden 1 → Hidden 2 : (2 inputs × 3 neuron) + 3 biases = 9
Hidden 2 → Output : (3 inputs × 1 neuron) + 1 biases = 4

Total: 10 + 9 + 4 = 23 parameters

The 'Dead ReLU' Problem

$$f(z) = \max(0, z)$$

(a) $f'(z) = \begin{cases} 1 & \text{if } z > 0 \\ 0 & \text{if } z < 0 \\ \text{undefined (or 0 by convention)} & \text{if } z = 0 \end{cases}$

(b) Value of the gradient if $z < 0$ if the pre-activation sum z is negative for every example, then $f'(z) = 0$ always. By the chain rule, the gradient of the loss L w.r.t weight is:

$$\frac{\partial L}{\partial w_i} = \frac{\partial L}{\partial f} \cdot f'(z) \cdot x_i \qquad f'(z) = 0$$

(c) Will the weights change? No. Since gradient is zero the gradient Descent update term will be zero. The weight remains unchanged. This neuron is dead - it will never activate and never learn.

Problem 4: L2 Regulation and Weight Decay

Total $J_{total}(w) = J_{data}(w) + \lambda/2 \; w^2$

(a) Gradient Descent update Rule first we find the gradient of the total loss with respect to w:

$$\frac{\partial J_{total}}{\partial w} = \frac{\partial J_{data}}{\partial w} + \frac{\partial}{\partial w}\left(\frac{\lambda w^2}{2}\right)$$

EXPT. NO. NAME

The gradient descent update rule is $\cancel{w_{new}} = w$
$w_{new} = w_{old} - \eta \cdot \nabla J$. Substituting our gradient:

$$w_{new} = w_{old} - \eta \left(\frac{d J_{data}}{d w} + \lambda w_{old} \right)$$

(b) Rearranged Eqn We distribute the learning rate η and group the w_{old} terms:

$$\cancel{w_{new} = w_{old} - \eta \, d \, J_{data}}$$
$$w_{new} = w_{old} (1 - \eta \lambda) - \eta \frac{d J_{data}}{d w}$$

(c) ~~Why is~~ It is called Weight Decay because before adding the data gradient, the current weight w_{old} is multiplied by a factor of $(1 - \eta \lambda)$ which is less than 1 causing the weight to 'decay' or shrink towards zero at every step.