SC4001 Exam 2023-2024 Sem 2

1. (a) State whether each of the following statements is "TRUE" or "FALSE" with explanations. Each subquestion carries one mark.

Answer

T (i) The primary reason that gradient descent, instead of directly solving the optimization problem via matrix inversion, is used to solve neural networks is that it is too computationally expensive for large neural networks.

F (ii) You design a fully connected neural network architecture where all activations are sigmoids. The weights should be initialized with large positive numbers.

T (iii) Whether training data is shuffled or not is unimportant when you are doing gradient descent (GD) using the entire training set.

F (iv) We can use the logistic regression model for regression tasks.

T (v) Dropout is implemented differently during training and testing.

F (vi) Making your network deeper by adding more layers will always reduce the training loss.

F (vii) You are training a neural network and notice that the validation error is significantly lower than the training error. This is because you are underfitting.

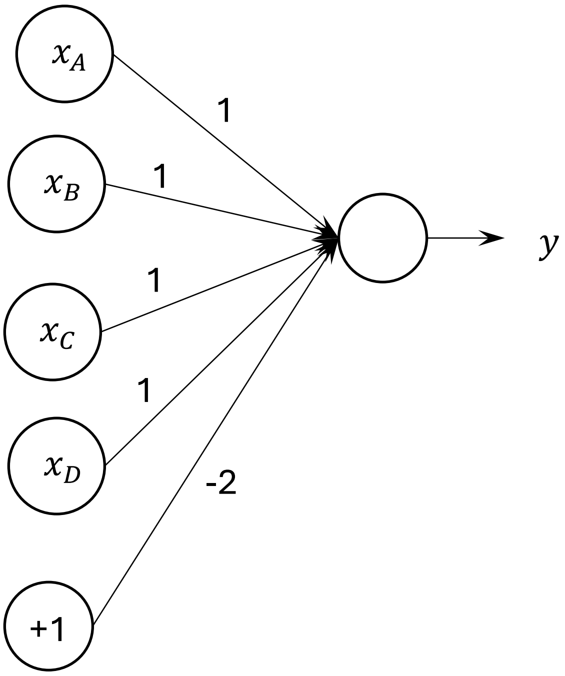
(7 marks)

1. (b) Suppose your task is to predict whether a medical image contains a TUMOR (1) or NO TUMOR (0). Now you are given predictions by n doctors, and your objective is to give a single prediction for each medical image as accurate as possible. To this end, you are expected to implement a *majority voting* algorithm - if more than half of the doctors predict TUMOR, then your final prediction should be TUMOR; otherwise, the final prediction should be NO TUMOR.

(i) Design a neural network architecture to implement this majority voting algorithm, with the assumption that we have a total of 4 doctors (named A,B,C,D). Specify the network structure and weights.

(5 marks)

Answer



$$x_i \in \{0, 1\} \qquad x = \begin{pmatrix} x_A \\ x_B \\ x_C \\ x_D \end{pmatrix} \qquad w = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \qquad b = -2.5$$

Synaptic input:

$$\begin{aligned} u &= \boldsymbol{w}^T \boldsymbol{x} + b \\ &= \begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} x_A \\ x_B \\ x_C \\ x_D \end{pmatrix} - 2.5 \\ &= x_A + x_B + x_C + x_D - 2.5 \end{aligned}$$

Output activation: logistic regression $f(\boldsymbol{u}) = \dfrac{1}{1 + e^{-u}}$

Output: $\boldsymbol{y} = 1(f(u) > 0.5)$

1. (b) (ii) Detail the learning process for at least one weight with the loss function specified and overfitting taken into consideration. At least one step of back-propagation (BP) should be run.

(5 marks)

Answer

Back-propagation:

Assume input: $x = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}$

Desired output: $d = 0$

Synaptic input:

$$\begin{aligned} u &= \boldsymbol{w}^T\boldsymbol{x} + b \\ &= (1 \quad 1 \quad 1 \quad 1)\begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} - 2.5 \\ &= 1 + 0 + 1 + 0 - 2.5 \\ &= -0.5 \end{aligned}$$

Activation Output: $f(u) = \dfrac{1}{1+e^{-u}} = \dfrac{1}{1+e^{0.5}} = 0.38$

Output:

$$\begin{aligned} y &= 1(f(u) > 0.5) \\ &= 0 \end{aligned}$$

Gradient

$$\begin{aligned} \nabla_u J &= -\big(\boldsymbol{d} - f(\boldsymbol{u})\big) \\ &= -[0 - (0.38)] \\ &= 0.38 \end{aligned}$$

Weight Update:

$$\begin{aligned} \boldsymbol{w} &= \boldsymbol{w} + \alpha x\big(\boldsymbol{d} - f(\boldsymbol{u})\big) \\ &= \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \alpha\begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}(0.38) \\ &= \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \alpha\begin{pmatrix} 0.38 \\ 0 \\ 0.38 \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} 1 + 0.38\alpha \\ 1 \\ 1 + 0.38\alpha \\ 1 \end{pmatrix} \end{aligned}$$

1. (b) (iii) Explain shortly (1-2 lines) how to generalize the network structure and weights to a more general case of n doctors.

(3 marks)

Answer

To generalize for n doctors, create an input layer with n nodes, set all weights to 1, and set the bias to $-\frac{n+1}{2}$ to ensure majority voting.

1. (b) (iv) Suppose that every doctor is based on completely different knowledge, and they all have the same accuracy level (say, 75%) on the given set of medical images. Can majority voting improve the overall accuracy in this case on this same set of medical images? Explain your answer shortly.

(5 marks)

Answer

- Yes, majority voting can improve the overall accuracy.

- When each doctor has a 75% accuracy and their predictions are independent, combining their votes reduces the likelihood of incorrect predictions.

- The collective decision is more likely to be correct than individual predictions, leveraging the wisdom of the crowd.

2. (a) Perform forward propagation on the 3-layer neural network in Figure Q2a for $x = 1$ by providing the outputs of the neurons $u_1, u_2, u_3, u_1, v_1, v_2, v_3, y$. Note that $u_1, u_2, \cdots, y$ are outputs after performing the appropriate operation as indicated in the node.

(7 marks)

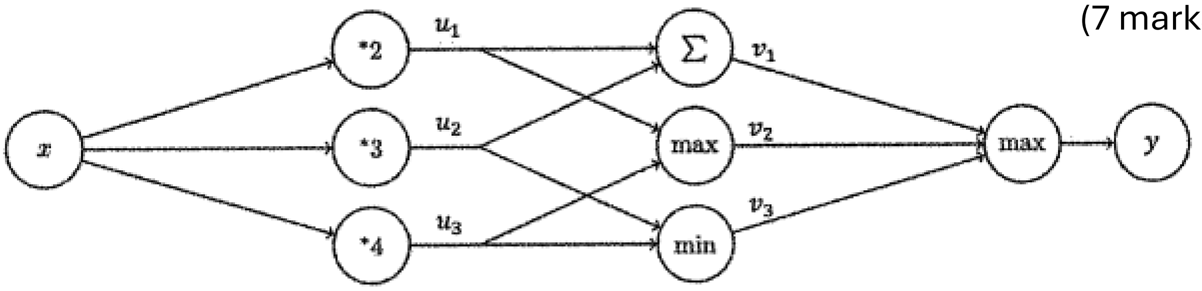


Figure Q2a

Answer

The diagram shows:

- First layer: Each path from x applies a multiplication by 2, 3, or 4
- Second layer: Then outputs are combined via sum, max, and min operations
- Output layer: Then again max operation is done

Input: $x = 1$

First layer:

- $u_1 = x * 2 = 1 * 2 = 2$
- $u_2 = x * 3 = 1 * 3 = 3$
- $u_3 = x * 4 = 1 * 4 = 4$

Second layer:

- $v_1 = u_1 + u_2 \quad = 2 + 3 \quad = 5$
- $v_2 = \max(u_1, u_3) = \max(2, 4) = 4$
- $v_3 = \min(u_2, u_3) = \min(3, 4) = 3$

Output layer:

- $y = \max(v_1, v_2, v_3) = \max(5, 4, 3) = 5$

Summary of outputs

- $u_1 = 2$
- $u_2 = 3$
- $u_3 = 4$
- $v_1 = 5$
- $v_2 = 4$
- $v_3 = 3$
- $y \ = 5$

2. (b) Below is a neural network with weights w_1, w_2, w_3, w_4, w_5, and w_6. The inputs are x_1 and x_2.

The first hidden layer computes $r_1 = max(w_1 \cdot x_1 + w_3 \cdot x_2, 0)$ and $r_2 = max(w_2 \cdot x_1 + w_4 \cdot x_2, 0)$.

The second hidden layer computes $s_1 = \frac{1}{1+\exp(-w_5 \cdot r_1)}$ and $s_2 = \frac{1}{1+\exp(-w_6 \cdot r_2)}$.

The output layer computes $y = s_1 + s_2$. Note that the weights w_1, w_2, w_3, w_4, w_5, and w_6 are indicated along the edges of the neural network here.

Suppose the network has inputs $x_1 = 1, x_2 = -1$.

The weight values are $w_1 = 4, w_2 = 1, w_3 = 2, w_4 = 2, w_5 = 1, w_6 = 1$.

Forward propagation then computes $r_1 = 2, r_2 = 0, s_1 = 0.9, s_2 = 0.5, y = 1.4$. Note: some values are rounded.

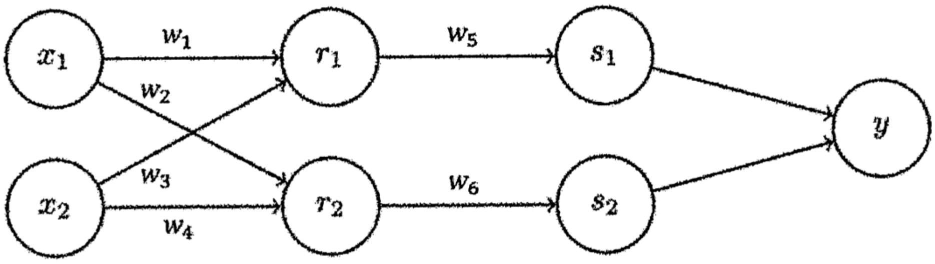


Figure Q2b

Using the values computed from forward propagation, use backpropagation to numerically calculate the partial derivatives $\frac{\partial y}{\partial w_k} (k = 1, 2, \cdots, 6)$. Write your answers as single numbers (not awk expressions). You do not need a calculator. Use scratch paper if needed.

(12 marks)

2. (b) cont

Answer

Given:

$$x_1 = 1, \ \ x_2 = -1$$

$$w_1 = 4, w_2 = 1, w_3 = 2, \ \ w_4 = 2, \ w_5 = 1, w_6 = 1$$

$$r_1 = 2, \ \ \ r_2 = 0, \ s_1 = 0.9, s_2 = 0.5, y = 1.4$$

Forward propagation results:

$$r_1 = max(w_1 \cdot x_1 + w_3 \cdot x_2, 0) = max(4 \cdot 1 + 2 \cdot (-1), 0) = max(\ \ 2, 0) = 2$$

$$r_2 = max(w_2 \cdot x_1 + w_4 \cdot x_2, 0) = max(1 \cdot 1 + 2 \cdot (-1), 0) = max(-1, 0) = 0$$

$$s_1 = \frac{1}{1 + \exp(-w_5 \cdot r_1)} = \frac{1}{1 + \exp(-1 \cdot 2)} = \frac{1}{1 + \exp(-2)} = 0.88 = 0.9(\text{rounded})$$

$$s_2 = \frac{1}{1 + \exp(-w_6 \cdot r_2)} = \frac{1}{1 + \exp(1 \cdot 0)} \ \ = \frac{1}{1 + \exp(0)} \ \ = 0.5$$

$$y = s_1 + s_2 = 0.88 + 0.5 = 1.38 = 1.4(\text{rounded})$$

Backpropagation:

Starts from output layer $y = s_1 + s_2$:

Derivatives of y:

$$\frac{\partial y}{\partial s_1} = 1, \qquad \frac{\partial y}{\partial s_2} = 1$$

2. (b) cont

Answer

Second hidden layer:

$$s_1 = \frac{1}{1 + \exp(-w_5 \cdot r_1)}) = 0.9$$

$$s_2 = \frac{1}{1 + \exp(-w_6 \cdot r_2)} = 0.5$$

Derivatives of Sigmoid

$$y = f(u) = \frac{1}{1 + e^{-u}}$$

$$\frac{\partial y}{\partial u} = f'(u) = y(1 - y)$$

Derivatives of s_1, s_2:

$$\begin{aligned} \frac{\partial s_1}{\partial (w_5 \cdot r_1)} &= s_1(1 - s_1) \\ &= 0.9 * (1 - 0.9) \\ &= 0.9 * 0.1 \\ &= 0.09 \end{aligned}$$

$$\frac{\partial (w_5 \cdot r_1)}{\partial w_5} = r_1 = 2$$

$$\begin{aligned} \frac{\partial s_2}{\partial (w_6 \cdot r_2)} &= s_2(1 - s_2) \\ &= 0.5 * (1 - 0.5) \\ &= 0.5 * 0.5 \\ &= 0.25 \end{aligned}$$

$$\frac{\partial (w_6 \cdot r_2)}{\partial (w_6)} = r_2 = 0$$

Chain rule

$$\frac{\partial J}{\partial x} = \frac{\partial J}{\partial y} \cdot \frac{\partial y}{\partial x}$$

$$\begin{aligned} \frac{\partial y}{\partial w_5} &= \frac{\partial y}{\partial s_1} \cdot \frac{\partial s_1}{\partial (w_5 \cdot r_1)} \cdot \frac{\partial (w_5 \cdot r_1)}{\partial w_5} \\ &= \frac{\partial y}{\partial s_1} * [s_1(1 - s_1)] * r_1 \\ &= 1 * 0.09 * 2 \\ &= 0.18 \end{aligned}$$

$$\begin{aligned} \frac{\partial y}{\partial w_6} &= \frac{\partial y}{\partial s_1} \cdot \frac{\partial s_1}{\partial (w_6 \cdot r_2)} \cdot \frac{\partial (w_6 \cdot r_2)}{\partial w_6} \\ &= \frac{\partial y}{\partial s_2} * [s_2(1 - s_2)] * r_2 \\ &= 1 * 0.25 * 0 \\ &= 0 \end{aligned}$$

2. (b) cont

Answer

First hidden layer:

$$r_1 = max(w_1 \cdot x_1 + w_3 \cdot x_2, 0) = max(\ \ 2, 0) = 2$$

$$r_2 = max(w_2 \cdot x_1 + w_4 \cdot x_2, 0) = max(-1, 0) = 0$$

Derivatives of r_1, r_2:

For r_1,

since $(w_1 \cdot x_1 + w_3 \cdot x_2) = 2 > 0$,

gradient is normal

$$\frac{\partial(w_5 \cdot r_1)}{\partial r_1} = w_5 = 1$$

$$\begin{aligned} \frac{\partial y}{\partial r_1} &= \frac{\partial y}{\partial s_1} \cdot \frac{\partial s_1}{\partial r_1} \\ &= \frac{\partial y}{\partial s_1} \cdot \frac{\partial s_1}{\partial(w_5 \cdot r_1)} \cdot \frac{\partial(w_5 \cdot r_1)}{\partial w_5} \\ &= \frac{\partial y}{\partial s_1} \cdot \frac{\partial s_1}{\partial(w_5 \cdot r_1)} \cdot w_5 \\ &= 1 * 0.09 * 1 \\ &= 0.09 \end{aligned}$$

For r_2,

since $(w_2 \cdot x_1 + w_4 \cdot x_2) = -1 < 0$,

So,

$$\frac{\partial y}{\partial r_2} = 0$$

Since $r_1 = max(w_1 \cdot x_1 + w_3 \cdot x_2, 0)$,

r_1 depends on w_1 and w_3

$$\frac{\partial r_1}{\partial w_1} = x_1 = 1 \qquad \frac{\partial r_1}{\partial w_3} = x_2 = -1$$

$$\begin{aligned} \frac{\partial y}{\partial w_1} &= \frac{\partial y}{\partial r_1} \cdot \frac{\partial r_1}{\partial w_1} \\ &= 0.09 * 1 \\ &= 0.09 \end{aligned}$$

$$\begin{aligned} \frac{\partial y}{\partial w_3} &= \frac{\partial y}{\partial r_1} \cdot \frac{\partial r_1}{\partial w_3} \\ &= 0.09 * (-1) \\ &= -0.09 \end{aligned}$$

Since $r_2 = max(w_2 \cdot x_1 + w_4 \cdot x_2, 0)$,

r_1 depends on w_2 and w_4

$$\frac{\partial y}{\partial w_2} = 0$$

$$\frac{\partial y}{\partial w_4} = 0$$

<u>Summary</u>

$$\frac{\partial y}{\partial w_1} = 0.09 \qquad \frac{\partial y}{\partial w_1} = 0.09$$

$$\frac{\partial y}{\partial w_2} = 0 \qquad \frac{\partial y}{\partial w_2} = 0$$

$$\frac{\partial y}{\partial w_3} = -0.09 \qquad \frac{\partial y}{\partial w_3} = -0.09$$

2. (c) Below are two plots with horizontal axis x_1 and vertical axis x_2 containing data labelled × and •. For each plot, we wish to find a function $f(x_1, x_2)$ such that $f(x_1, x_2) \geq 0$ for all data labelled × and $f(x_1, x_2) < 0$ for all data labelled •. Give at least two possible expressions such that all the data is labelled correctly for each plot.

(6 marks)

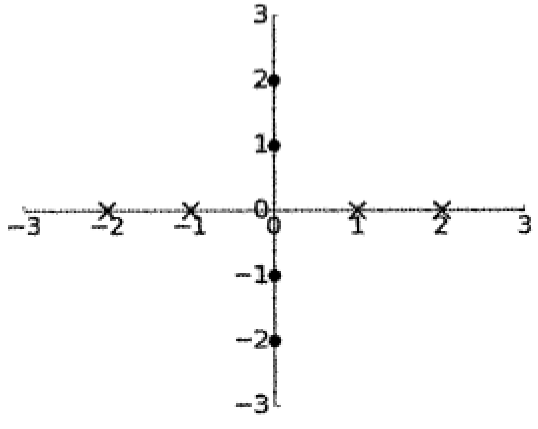


Figure Q2c

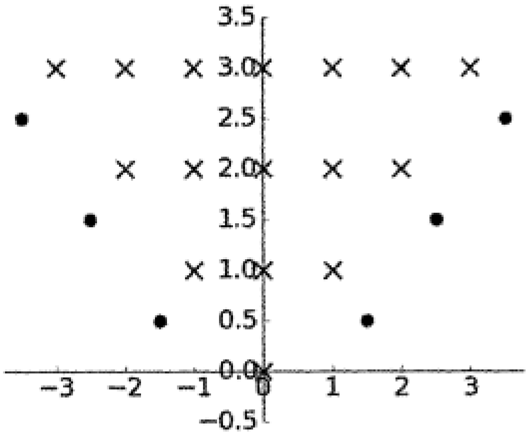


Figure Q2d

2. (c) cont

Answer

- Data Points:
  - From Figure Q2c, the data points are:
    - Labelled '×': (-2, 0), (-1, 0), (1, 0), (2, 0)
    - Labelled '•': (0, 2), (0, 1), (0, -1), (0, -2)
  - We need a function $f(x_1, x_2)$ such that:
    - $f(x_1, x_2) \geq 0$ for (-2,0),(-1,0),(1,0), (2,0)
    - $f(x_1, x_2) < 0$ for (0,2), (0,1), (0,−1),(0,-2)
  - Observations:
    - the '×' points lie on the x_1-axis $(x_2 = 0)$
    - the '•' points lie on the x_2-axis $(x_1 = 0)$.
  - Expression 1: Linear function
    - $f(x_1, x_2) = |x_1| - |x_2|$
    - Reason
      - For '×' points: $|x_1| \geq 1$ and $|x_2| = 0$, so $f(x_1, x_2) \geq 0$
      - For '•' points: $|x_1| = 0$ and $|x_2| \geq 1$ so $f(x_1, x_2) < 0$
    - Test data points
      - For '×' points
        - $f(-2, 0) = |-2| - |0| = 2 - 0 = 2$
        - $f(-1, 0) = |-1| - |0| = 1 - 0 = 1$
        - $f(\ \ 1, 0) = |1| \ \ - |0| = 1 - 0 = 1$
        - $f(\ \ 2, 0) = |2| \ \ - |0| = 2 - 0 = 2$
      - For '•' points
        - $f(0, \ \ 2) = |0| - |2| \ \ = 0 - 2 = -2$
        - $f(0, \ \ 1) = |0| - |1| \ \ = 0 - 1 = -1$
        - $f(0, -1) = |0| - |-1| = 0 - 1 = -1$
        - $f(0, -2) = |0| - |-2| = 0 - 2 = -2$
    - Tested all '×' points $f \geq 0$ and all '•' points $f < 0$

2. (c) cont

Answer

- Expression 2: Quadratic function
  - $f(x_1, x_2) = (x_1)^2 - (x_2)^2$

  - Reason
    - For '×' points: $x_2 = 0$, so $f(x_1, 0) = (x_1)^2 \geq 0$
    - For '•' points: $x_1 = 0$, so $f(0, x_2) = -(x_2)^2 < 0$

  - Test data points
    - For '×' points
      - $f(-2, 0) = (-2)^2 - (0)^2 = 4 - 0 = 4$
      - $f(-1, 0) = (-1)^2 - (0)^2 = 1 - 0 = 1$
      - $f(1, 0) = (1)^2 - (0)^2 = 1 - 0 = 1$
      - $f(2, 0) = (2)^2 - (0)^2 = 4 - 0 = 4$

    - For '•' points
      - $f(0, 2) = (0)^2 - (2)^2 = 0 - 4 = -4$
      - $f(0, 1) = (0)^2 - (1)^2 = 0 - 1 = -1$
      - $f(0, -1) = (0)^2 - (-1)^2 = 0 - 1 = -1$
      - $f(0, -2) = (0)^2 - (-2)^2 = 0 - 4 = -4$

  - Tested all '×' points $f \geq 0$ and all '•' points $f < 0$

- Data Points:
  - From Figure Q2d, the data points are:
    - Labelled '×': (-3, 3), (-2, 3), (-1, 3), (0, 3), (1, 3), (2, 3), (3, 3),
(-2, 2), (-1, 2), (0, 2), (1, 2), (2, 2),
(-1, 1), (0, 1), (1, 1),
(0,0)
    - Labelled '•': (-3.5, 2.5), (3.5, 2.5),
(-2.5, 1.5), (2.5, 1.5),
(-1.5, 0.5), (1.5, 0.5)

  - Observations:
    - For '×' points:
      - They form a kind of triangular / pyramid shape
      - A wide base at $x_2 = 3$ (from $x_1 = -3 \; to \; 3$)
    - For '•' points:
      - They are outside the '×' cluster

2. (c) cont

Answer

- Expression 1: Linear function
  - $f(x_1, x_2) = x_2 - |x_1|$

  - Reason
    - For '×' points: x_2 is large compared to $|x_1|$, so $x_2 - |x_1| \geq 0$
    - For '•' points: $|x_1|$ is large compared to x_2, so $x_2 - |x_1| < 0$

  - Test data points
    - For '×' points
      - $f(-3,3) = 3 - |-3| = 3 - 3 = 0$
      - $f(-2,3) = 3 - |-2| = 3 - 2 = 1$
      - $f(-1,3) = 3 - |-1| = 3 - 1 = 2$
      - $f(\ \ 0,3) = 3 - |\ \ 0| = 3 - 0 = 3$
      - $f(\ \ 1,3) = 3 - |\ \ 1| = 3 - 1 = 2$
      - $f(\ \ 2,3) = 3 - |\ \ 2| = 3 - 2 = 1$
      - $f(\ \ 3,3) = 3 - |\ \ 3| = 3 - 3 = 0$
      - $f(-2,2) = 2 - |-2| = 2 - 2 = 0$
      - $f(-1,2) = 2 - |-1| = 2 - 1 = 1$
      - $f(-0,2) = 2 - |\ \ 0| = 2 - 0 = 2$
      - $f(\ \ 1,2) = 2 - |\ \ 1| = 2 - 1 = 1$
      - $f(\ \ 2,2) = 2 - |\ \ 2| = 2 - 2 = 0$
      - $f(-1,1) = 1 - |-1| = 1 - 1 = 0$
      - $f(\ \ 0,1) = 1 - |\ \ 0| = 1 - 0 = 1$
      - $f(\ \ 1,1) = 1 - |\ \ 1| = 1 - 1 = 0$
      - $f(\ \ 0,0) = 0 - |\ \ 0| = 0 - 0 = 0$

    - For '•' points
      - $f(-3.5, 2.5) = 2.5 - |-3.5| = 2.5 - 3.5 = -1$
      - $f(\ \ 3.5, 2.5) = 2.5 - |\ \ 3.5| = 2.5 - 3.5 = -1$
      - $f(-2.5, 1.5) = 1.5 - |-2.5| = 1.5 - 2.5 = -1$
      - $f(\ \ 2.5, 1.5) = 1.5 - |-2.5| = 1.5 - 2.5 = -1$
      - $f(-1.5, 0.5) = 0.5 - |-1.5| = 0.5 - 1.5 = -1$
      - $f(\ \ 1.5, 0.5) = 0.5 - |-1.5| = 0.5 - 1.5 = -1$

  - Tested all '×' points $f \geq 0$ and all '•' points $f < 0$

2. (c) cont

Answer

- Expression 1: Linear function
  - $f(x_1, x_2) = x_2 + 0.5 - |x_1|$

  - Reason
    - Amplifies x_2 so that points closer to centre and higher x_2 stay positive

  - Test data points
    - For '×' points
      - $f(-3, 3) = 3 + 0.5 - |-3| = 3.5 - 3 = 0.5$
      - $f(-2, 3) = 3 + 0.5 - |-2| = 3.5 - 2 = 1.5$
      - $f(-1, 3) = 3 + 0.5 - |-1| = 3.5 - 1 = 2.5$
      - $f(\ \ 0, 3) = 3 + 0.5 - |\ \ 0| = 3.5 - 0 = 3.5$
      - $f(\ \ 1, 3) = 3 + 0.5 - |\ \ 1| = 3.5 - 1 = 2.5$
      - $f(\ \ 2, 3) = 3 + 0.5 - |\ \ 2| = 3.5 - 2 = 1.5$
      - $f(\ \ 3, 3) = 3 + 0.5 - |\ \ 3| = 3.5 - 3 = 0.5$
      - $f(-2, 2) = 2 + 0.5 - |-2| = 2.5 - 2 = 0.5$
      - $f(-1, 2) = 2 + 0.5 - |-1| = 2.5 - 1 = 1.5$
      - $f(-0, 2) = 2 + 0.5 - |\ \ 0| = 2.5 - 0 = 2.5$
      - $f(\ \ 1, 2) = 2 + 0.5 - |\ \ 1| = 2.5 - 1 = 1.5$
      - $f(\ \ 2, 2) = 2 + 0.5 - |\ \ 2| = 2.5 - 2 = 0.5$
      - $f(-1, 1) = 1 + 0.5 - |-1| = 1.5 - 1 = 0.5$
      - $f(\ \ 0, 1) = 1 + 0.5 - |\ \ 0| = 1.5 - 0 = 1.5$
      - $f(\ \ 1, 1) = 1 + 0.5 - |\ \ 1| = 1.5 - 1 = 0.5$
      - $f(\ \ 0, 0) = 0 + 0.5 - |\ \ 0| = 0.5 - 0 = 0.5$

    - For '•' points
      - $f(-3.5, 2.5) = 2.5 + 0.5 - |-3.5| = 3 - 3.5 = -0.5$
      - $f(\ \ 3.5, 2.5) = 2.5 + 0.5 - |\ \ 3.5| = 3 - 3.5 = -0.5$
      - $f(-2.5, 1.5) = 1.5 + 0.5 - |-2.5| = 2 - 2.5 = -0.5$
      - $f(\ \ 2.5, 1.5) = 1.5 + 0.5 - |-2.5| = 2 - 2.5 = -0.5$
      - $f(-1.5, 0.5) = 0.5 + 0.5 - |-1.5| = 1 - 1.5 = -0.5$
      - $f(\ \ 1.5, 0.5) = 0.5 + 0.5 - |-1.5| = 1 - 1.5 = -0.5$

  - Tested all '×' points $f \geq 0$ and all '•' points $f < 0$

3. (a) Consider a convolutional neural network (CNN) designed for image classification. The network consists of the following layers in order:

Input: Grayscale images of size 32x32 pixels.

1) Convolutional layer 1: Uses 6 filters of size 5x5 with a stride of 1 and no padding. A ReLU activation function is used.
2) Max pooling layer: Uses a 2x2 filter with a stride of 2.
3) Convolutional layer 2: Uses 16 filters of size 5x5 with a stride of 1 and no padding. A ReLU activation function is used.
4) Max pooling layer: Uses a 2x2 filter with a stride of 2.
5) Fully connected layer: Has 120 neurons with ReLU activation functions.
6) Fully connected layer: Has 10 neurons (corresponding to 10 classes) with softmax activation functions.

3. (a) (i) Calculate the dimensions of the output feature map after each layer

(6 marks)

Answer

<u>Convolution layer 1</u>

- since input size: 1x32x32, N = 32
- since filter size: 1x5x5, F = 5
- stride S = 1
- padding P = 0

- number of filters = 6

$$\text{Output} = \frac{N - F + 2P}{S} + 1$$

$$= \frac{32 - 5 + 2(0)}{1} + 1$$

$$= 27 + 1$$

$$= 28$$

Output volume size = 6x28x28

<u>Max pooling layer 1</u>

- from convolution layer 1, N = 28
- since filter size: 2x2, F = 2
- stride S = 2
- padding P = 0

- number of filters = 6

$$\text{Output} = \frac{N - F + 2P}{S} + 1$$

$$= \frac{28 - 2 + 2(0)}{2} + 1$$

$$= 13 + 1$$

$$= 14$$

Output volume size = 6x14x14

3. (a) (i) cont

Answer

Convolution layer 2

- from max pooling layer 1, N = 14
- since filter size: 1x5x5, F = 5
- stride S = 1
- padding P = 0

$$\text{Output} = \frac{N - F + 2P}{S} + 1$$
$$= \frac{14 - 5 + 2(0)}{1} + 1$$
$$= 9 + 1$$
$$= 10$$

- number of filters = 16

Output volume size = 16x10x10

Max pooling layer 2

- from convolution layer 2, N = 10
- since filter size: 2x2, F = 2
- stride S = 2
- padding P = 0

$$\text{Output} = \frac{N - F + 2P}{S} + 1$$
$$= \frac{10 - 2 + 2(0)}{2} + 1$$
$$= 4 + 1$$
$$= 5$$

- number of filters = 16

Output volume size = 16x5x5

Fully connected layer 1

- from max pooling layer, input flatten = 1 vector
- output neurons = 120

Output volume size = 120x1

Fully connected layer 2

- output neurons = 10

Output volume size = 10x1

Summary of output sizes

Convolution layer 1	= 6x28x28
Max pooling layer 1	= 6x14x14
Convolution layer 2	= 16x10x10
Max pooling layer 2	= 16x5x5
Fully connected layer 1	= 120x1
Fully connected layer 2	= 100x1

3. (a) (ii) How many parameters are there in this CNN? Show your calculations. Be reminded to account for bias terms.

(5 marks)

Answer

Convolution Layer 1

- filter size = d x F x F = 1 x 5 x 5
- d = 1
- F = 5
- convolution filters = 6, n = 6

Parameters per filer = (d * F * F) + 1 = (1 * 5 * 5) + 1 = 26
Total parameters = param * n = 26 * 6 = 156

Max pooling layer 1

- pooing layers do not have any parameters

Total parameters = 0

Convolution Layer 2

- filter size = d x F x F = 6 x 5 x 5
- d = 6
- F = 5
- convolution filters = 16, n = 16

Parameters per filer = (d * F * F) + 1 = (6 * 5 * 5) + 1 = 151
Total parameters = param * n = 151 * 16 = 2416

Max pooling layer 2

- pooing layers do not have any parameters

Total parameters = 0

Fully Connected Layer 1

- input flatten = 16 * 5 * 5 = 400
- output neurons = 120
- bias = 120

Total parameters = (input * output) + bias = (400 * 120) + 120 = 48120

3. (a) (ii) cont

Answer

Fully Connected Layer 2

- input flatten = 120 * 1 = 120
- output neurons = 10
- bias = 10

Total parameters = (input * output) + bias = (120 * 10) + 10 = 1210

Total Parameters Summary

Convolution Layer 1	=	156
Max pooling layer	=	0
Convolution Layer 2	=	2416
Max pooling layer 2	=	0
Fully Connected Layer 1	=	48120
Fully Connected Layer 2	=	1210

Total parameters in this CNN = 156 + 0 + 2416 + 0 + 48120 + 1210 = 51902

3. (a) (iii) Explain the role of the ReLU activation function in this network. What advantage does it have over other activation functions like sigmoid or tanh? (3 marks)

Answer

- The role of the ReLU activation function in this CNN is to introduce non-linearity into the network, allowing it to learn complex patterns and features from the input images.

- Without non-linear activation functions like ReLU, the entire network would behave like a single linear transformation, no matter how many layers it had, making it unable to model complex data.

- Advantage
  - Unlike sigmoid/tanh, ReLU's gradient is either 0 (for $x \leq 0$) or 1 (for $x > 0$), preventing gradient decay in deep layers during backpropagation.

  - ReLU is simple (ReLU(u) = max(0, u)), making it faster to compute compared to the more complex operations of sigmoid and tanh.

3. (b) Consider a simple convolutional neural network with a single convolutional layer. The input to this layer is a 3x4 grayscale image, and the layer uses a single 2x2 filter with weights $w = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ and no bias. The convolution operation uses a stride of 1 and no padding.

3. (b) (i) Given the input image:

$$\begin{pmatrix} -0.3 & 0.7 & -0.2 & 0.4 \\ 0.6 & -0.8 & 0.9 & -0.1 \\ -0.5 & 0.2 & -0.6 & 0.1 \end{pmatrix}$$

Compute the output feature map resulting from the convolution operation.

(4 marks)

Answer

- since input image size = 3x4,
  input height R = 3
  input width C = 4

$$\begin{pmatrix} -0.3 & 0.7 & -0.2 & 0.4 \\ 0.6 & -0.8 & 0.9 & -0.1 \\ -0.5 & 0.2 & -0.6 & 0.1 \end{pmatrix}$$

- since filter size = 2x2, F = 2

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

- stride S = 1
- padding P = 0

output size:

$$\text{row} = \frac{R - F + 2P}{S} + 1 = \frac{3 - 2 + 2(0)}{1} + 1 = 2$$

$$\text{col} = \frac{C - F + 2P}{S} + 1 = \frac{4 - 2 + 2(0)}{1} + 1 = 3$$

- output size = 2x3

3. (b) (i) cont

Answer

- stride = 1
- bias = 0

$$\left(\begin{array}{c|c|c|c} -0.3 & 0.7 & -0.2 & 0.4 \\ \hline 0.6 & -0.8 & 0.9 & -0.1 \\ -0.5 & 0.2 & -0.6 & 0.1 \end{array}\right) \qquad \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

$$\begin{aligned} u(1,1) &= (-0.3)(1) + (0.7)(0) + (0.6)(0) + (-0.8)(-1) + 0 \\ &= -0.3 + 0 + 0 + 0.8 + 0 = 0.5 \\ u(1,2) &= (0.7)(1) + (-0.2)(0) + (-0.8)(0) + (0.9)(-1) + 0 \\ &= 0.7 + 0 + 0 - 0.9 + 0 = -0.2 \\ u(1,3) &= (-0.2)(1) + (0.4)(0) + (0.9)(0) + (-0.1)(-1) + 0 \\ &= -0.2 + 0 + 0 + 0.1 + 0 = -0.1 \end{aligned}$$

$$\begin{aligned} u(2,1) &= (0.6)(1) + (-0.8)(0) + (-0.5)(0) + (0.2)(-1) + 0 \\ &= 0.6 + 0 + 0 - 0.2 + 0 = 0.4 \\ u(2,2) &= (-0.8)(1) + (0.9)(0) + (0.2)(0) + (-0.6)(-1) + 0 \\ &= -0.8 + 0 + 0 + 0.6 + 0 = -0.2 \\ u(2,3) &= (0.9)(1) + (-0.1)(0) + (-0.6)(0) + (0.1)(-1) + 0 \\ &= 0.9 + 0 + 0 - 0.1 + 0 = 0.8 \end{aligned}$$

- output feature map = $\begin{pmatrix} 0.5 & -0.2 & -0.1 \\ 0.4 & -0.2 & 0.8 \end{pmatrix}$

3. (b) (ii) If a ReLU activation function is applied to the output feature map, what would be the final output?

(2 marks)

Answer

Apply ReLU activation function:

$$\begin{pmatrix} \max(0, 0.5) = 0.5 & \max(0, -0.2) = 0 & \max(0, -0.1) = 0 \\ \max(0, 0.4) = 0.4 & \max(0, -0.2) = 0 & \max(0, 0.8) = 0.8 \end{pmatrix}$$

- output feature map = $\begin{pmatrix} 0.5 & 0 & 0 \\ 0.4 & 0 & 0.8 \end{pmatrix}$

3. (c) The statements below are all related to autoencoders. Answer "TRUE" or "FALSE" to the following statements. Each part carries 1 mark.

Answer

F (i) Autoencoders are primarily used for dimensionality reduction and cannot be used for classification tasks.

T (ii) Autoencoders can be used for anomaly detection by reconstructing input data and measuring the reconstruction error.

T (iii) Denoising autoencoders are trained to reconstruct the original input data from corrupted versions of the input.

F (iv) Overcomplete autoencoders are more likely to capture the most salient features than undercomplete autoencoders.

F (v) Sparse autoencoders encourage neuron sparsity via architecture design instead of loss function.

(5 marks)

4. (a) Consider an Elman-type recurrent neural network (RNN) that receives 2-dimensional input patterns $\boldsymbol{x}$ and has one hidden layer. The RNN has two neurons in the hidden layer (which are initialized to zeros) and one neuron in the output layer. The hidden layer neurons have *Tanh* activation functions and the output layer neurons use *Sigmoid* activation functions.

The weight matrices $\boldsymbol{U}$ connecting the input to the hidden layer, $\boldsymbol{W}$ connecting the previous hidden state to the next hidden state, and $\boldsymbol{V}$ connecting the hidden output to the output layer are given by

$$\boldsymbol{U} = \begin{pmatrix} 0.5 & -0.3 \\ 1.0 & 0.8 \end{pmatrix}, \boldsymbol{W} = \begin{pmatrix} 1.0 & -0.5 \\ 0 & 0.6 \end{pmatrix} \text{ and } \boldsymbol{V} = \begin{pmatrix} 0.2 \\ 1.0 \end{pmatrix}$$

The hidden layer bias vector $\boldsymbol{b}$ and the output layer bias c are given by

$$\boldsymbol{b} = \begin{pmatrix} 0.5 \\ 1.0 \end{pmatrix}, \text{ and } c = 0.2$$

Determine the output sequence of the RNN for an input sequence of $(\boldsymbol{x}(1), \boldsymbol{x}(2), \boldsymbol{x}(3))$ when

$$\boldsymbol{x}(1) = \begin{pmatrix} 1.0 \\ -0.5 \end{pmatrix}, x(2) = \begin{pmatrix} -2.0 \\ 1.5 \end{pmatrix} \text{ and } \boldsymbol{x}(3) = \begin{pmatrix} 0.5 \\ 0 \end{pmatrix}$$

Provide your answers rounded to three decimal places.

(12 marks)

Answer

$$\boldsymbol{U} = \begin{pmatrix} 0.5 & -0.3 \\ 1.0 & 0.8 \end{pmatrix}, \boldsymbol{W} = \begin{pmatrix} 1.0 & -0.5 \\ 0 & 0.6 \end{pmatrix} \text{ and } \boldsymbol{V} = \begin{pmatrix} 0.2 \\ 1.0 \end{pmatrix}$$

$$\boldsymbol{x}(1) = \begin{pmatrix} 1.0 \\ -0.5 \end{pmatrix}, x(2) = \begin{pmatrix} -2.0 \\ 1.5 \end{pmatrix} \text{ and } \boldsymbol{x}(3) = \begin{pmatrix} 0.5 \\ 0 \end{pmatrix}$$

$$\boldsymbol{b} = \begin{pmatrix} 0.5 \\ 1.0 \end{pmatrix}, \text{ and } c = 0.2 \qquad h = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\phi(u) = \tanh(u) \quad = \frac{e^u - e^{-u}}{e^u + e^{-u}}$$

$$\sigma(u) = \text{sigmoid}(u) = \frac{1}{1 + e^{-u}}$$

4. (a) cont

Answer

$$\phi(u) = \tanh(u) \quad = \frac{e^u - e^{-u}}{e^u + e^{-u}}$$

$$\sigma(u) = \text{sigmoid}(u) = \frac{1}{1 + e^{-u}}$$

At $t = 1, \boldsymbol{x}(1) = \begin{pmatrix} 1.0 \\ -0.5 \end{pmatrix}$,

$$\boldsymbol{h}(t) = \quad \phi(\boldsymbol{U}^T\boldsymbol{x}(t) + \boldsymbol{W}^T\boldsymbol{h}(t-1) + \boldsymbol{b})$$

$$\boldsymbol{h}(1) = tanh(\boldsymbol{U}^T\boldsymbol{x}(1) + \boldsymbol{W}^T\boldsymbol{h}(0) + \boldsymbol{b})$$

$$= \tanh\left(\begin{pmatrix} 0.5 & 1.0 \\ -0.3 & 0.8 \end{pmatrix}\begin{pmatrix} 1.0 \\ -0.5 \end{pmatrix} + \begin{pmatrix} 1.0 & 0 \\ -0.5 & 0.6 \end{pmatrix}\begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0.5 \\ 1.0 \end{pmatrix}\right)$$

$$= \tanh\left(\begin{pmatrix} (\ \ 0.5)(1) + (1.0)(-0.5) \\ (-0.3)(1) + (0.8)(-0.5) \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0.5 \\ 1.0 \end{pmatrix}\right)$$

$$= \tanh\left(\begin{pmatrix} 0 \\ -0.7 \end{pmatrix} + \begin{pmatrix} 0.5 \\ 1.0 \end{pmatrix}\right) = \tanh\begin{pmatrix} 0.5 \\ 0.3 \end{pmatrix} = \begin{pmatrix} 0.462 \\ 0.291 \end{pmatrix}$$

$$\boldsymbol{y}(t) = \qquad \sigma(\boldsymbol{V}^T\boldsymbol{h}(t) + c)$$

$$\boldsymbol{y}(1) = sigmoid(\boldsymbol{V}^T\boldsymbol{h}(1) + c)$$

$$= \text{sigmoid}\left(\begin{pmatrix} 0.2 & 1.0 \end{pmatrix}\begin{pmatrix} 0.462 \\ 0.291 \end{pmatrix} + 0.2\right) = \text{sigmoid}((0.2)(0.462) + (1)(0.291) + 0.2)$$

$$= \text{sigmoid}(0.583) = 0.642$$

At $t = 2, \boldsymbol{x}(2) = \begin{pmatrix} -2.0 \\ 1.5 \end{pmatrix}$,

$$\boldsymbol{h}(t) = \quad \phi(\boldsymbol{U}^T\boldsymbol{x}(t) + \boldsymbol{W}^T\boldsymbol{h}(t-1) + \boldsymbol{b})$$

$$\boldsymbol{h}(2) = tanh(\boldsymbol{U}^T\boldsymbol{x}(2) + \boldsymbol{W}^T\boldsymbol{h}(1) + \boldsymbol{b})$$

$$= \tanh\left(\begin{pmatrix} 0.5 & 1.0 \\ -0.3 & 0.8 \end{pmatrix}\begin{pmatrix} -2.0 \\ 1.5 \end{pmatrix} + \begin{pmatrix} 1.0 & 0 \\ -0.5 & 0.6 \end{pmatrix}\begin{pmatrix} 0.462 \\ 0.291 \end{pmatrix} + \begin{pmatrix} 0.5 \\ 1.0 \end{pmatrix}\right)$$

$$= \tanh\left(\begin{pmatrix} (\ \ 0.5)(-2) + (1.0)(1.5) \\ (-0.3)(-2) + (0.8)(1.5) \end{pmatrix} + \begin{pmatrix} (\ \ 1.0)(0.462) + (0\ \ )(0.291) \\ (-0.5)(0.462) + (0.6)(0.291) \end{pmatrix} + \begin{pmatrix} 0.5 \\ 1.0 \end{pmatrix}\right)$$

$$= \tanh\left(\begin{pmatrix} 0.5 \\ 1.8 \end{pmatrix} + \begin{pmatrix} 0.462 \\ -0.056 \end{pmatrix} + \begin{pmatrix} 0.5 \\ 1.0 \end{pmatrix}\right) = \tanh\begin{pmatrix} 1.462 \\ 2.744 \end{pmatrix} = \begin{pmatrix} 0.898 \\ 0.992 \end{pmatrix}$$

$$\boldsymbol{y}(t) = \qquad \sigma(\boldsymbol{V}^T\boldsymbol{h}(t) + c)$$

$$\boldsymbol{y}(2) = sigmoid(\boldsymbol{V}^T\boldsymbol{h}(2) + c)$$

$$= \text{sigmoid}\left(\begin{pmatrix} 0.2 & 1.0 \end{pmatrix}\begin{pmatrix} 0.898 \\ 0.992 \end{pmatrix} + 0.2\right) = \text{sigmoid}((0.2)(0.898) + (1)(0.992) + 0.2)$$

$$= \text{sigmoid}(1.372) = 0.798$$

4. (a) cont

Answer

$$\phi(u) = \tanh(u) \quad = \frac{e^u - e^{-u}}{e^u + e^{-u}}$$

$$\sigma(u) = \text{sigmoid}(u) = \frac{1}{1 + e^{-u}}$$

At $t = 3, \boldsymbol{x}(3) = \begin{pmatrix} 0.5 \\ 0 \end{pmatrix}$,

$$\boldsymbol{h}(t) = \quad \phi(\boldsymbol{U}^T\boldsymbol{x}(t) + \boldsymbol{W}^T\boldsymbol{h}(t-1) + \boldsymbol{b})$$

$$\boldsymbol{h}(3) = tanh(\boldsymbol{U}^T\boldsymbol{x}(3) + \boldsymbol{W}^T\boldsymbol{h}(2) + \boldsymbol{b})$$

$$= \tanh\left(\begin{pmatrix} 0.5 & 1.0 \\ -0.3 & 0.8 \end{pmatrix}\begin{pmatrix} 0.5 \\ 0 \end{pmatrix} + \begin{pmatrix} 1.0 & 0 \\ -0.5 & 0.6 \end{pmatrix}\begin{pmatrix} 0.898 \\ 0.992 \end{pmatrix} + \begin{pmatrix} 0.5 \\ 1.0 \end{pmatrix}\right)$$

$$= \tanh\left(\begin{pmatrix} (\ \ 0.5)(0.5) + (1.0)(0) \\ (-0.3)(0.5) + (0.8)(0) \end{pmatrix} + \begin{pmatrix} (\ \ 1.0)(0.898) + (0 \ \ )(0.992) \\ (-0.5)(0.898) + (0.6)(0.992) \end{pmatrix} + \begin{pmatrix} 0.5 \\ 1.0 \end{pmatrix}\right)$$

$$= \tanh\left(\begin{pmatrix} 0.25 \\ -0.15 \end{pmatrix} + \begin{pmatrix} 0.898 \\ 0.146 \end{pmatrix} + \begin{pmatrix} 0.5 \\ 1.0 \end{pmatrix}\right) = \tanh\begin{pmatrix} 1.648 \\ 0.996 \end{pmatrix} = \begin{pmatrix} 0.929 \\ 0.760 \end{pmatrix}$$

$$\boldsymbol{y}(t) = \quad \sigma(\boldsymbol{V}^T\boldsymbol{h}(t) + c)$$

$$\boldsymbol{y}(3) = sigmoid(\boldsymbol{V}^T\boldsymbol{h}(3) + c)$$

$$= \text{sigmoid}\left(\begin{pmatrix} 0.2 & 1.0 \end{pmatrix}\begin{pmatrix} 0.929 \\ 0.760 \end{pmatrix} + 0.2\right) = \text{sigmoid}((0.2)(0.929) + (1)(0.760) + 0.2)$$

$$= \text{sigmoid}(1.146) = 0.759$$

Summary

- Output of the network:
  - y(1) = 0.642
  - y(2) = 0.798
  - y(3) = 0.759

4. (b) The statements below are all related to Transformers. Answer "TRUE" or "FALSE" to the following statements. Each part carries 1 mark.

Answer

T (i) Dividing the dot product of the query vector with the key vector by the square root of the dimension of the key vectors helps to stabilize gradients during training.

F (ii) Multi-head attention inherently understands the sequential nature of the data, making positional encoding unnecessary.

T (iii) Self-attention is performed in both the encoder and the decoder of Transformers.

F (iv) Positional encoding only provides information about the relative position of the tokens in the sequence.

F (v) When performing language translation using a Transformer, the model processes each word one at a time in a recurrent neural network manner.

(5 marks)

4. (c) (i) Explain the roles of the generator and the discriminator in a GAN. How do they interact during the training process?

(4 marks)

Answer

- Generator:
  - The generator's role is to create synthetic data that resembles the real data.

- Discriminator:
  - The discriminator's role is to distinguish between real data and synthetic data generated by the generator.

- During training:
  - The generator updates its parameters to improve the quality of its fake data, trying to make it harder for the discriminator to tell the difference.

  - The discriminator updates its parameters to become better at distinguishing real data from fake data generated by the generator.

4. (c) (ii) In a scenario where a GAN has been trained on a dataset containing images of digits 0-9 (such as the MNIST dataset), it is observed that the trained GAN generates images representing only the digits 0-8, with the digit 9 consistently missing from the generated outputs. What is the name of this phenomenon?

(2 marks)

Answer

- The phenomenon described is called mode collapse.