Exercise 2 submission

Arnas Šniokaitis(mat. nr.: 3252523), Mika Zeitz(mat. nr.: 2905965)

November 3, 2025

1 Exercise 1

1. Explain why w_m is of dimension $\mathbb{R}^{n+1}$ and x of dimension $\mathbb{R}^n$.

   x is of dimension $\mathbb{R}^n$ because x has n different properties which interest us. w_n is of a dimension one greater than x to account for the bias.

2. Show that the operation of a fully connected layer corresponds to a matrix vector multiplication for an input vector x.

   For any node m_i in M its value is equal to the product x and w_i ($m_i = \sum_{j=1}^{M} x_j w_{i,j} + w_{i,0}$), where $w_{i,0}$ is assumed to be the bias for layer i. However this sum is equivalent to what we get if we were to take the i th feature of i and calculate its multiplication by the i th column of w. Or in other words if we were to multiply our input vector x and multiply it by our entire matrix w what we would get in each i the element of the resulting vector would be exactly equal to what each node in our layer would have after proccesing the input vector with the weights. (Note that the result needs to be transposed for us to interpret it as a vector).Mathematically this can be shown as

$$z = x' \cdot W, \quad \text{where } w_j = \begin{bmatrix} w_{1j} \\ w_{2j} \\ \vdots \\ w_{nj} \\ b_j \end{bmatrix} \in \mathbb{R}^{n+1},$$

$$W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1m} \\ w_{21} & w_{22} & \cdots & w_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nm} \\ b_1 & b_2 & \cdots & b_m \end{bmatrix} \in \mathbb{R}^{(n+1)\times m}.$$

3. Show that the operation of the fully connected layer corresponds to a matrix-vector multiplication for an input vector x.

   Following from what we showed in the previous part we can show that we can take a vector and imagine it as if it were a $1 \times n$ matrix and perform a matrix multiplication to get what would our first layer contain in each nodes(Really we are imagining that x is a $1 \times n + 1$ matrix because we also need to account for the bias, but we can chose this element to be any constant as the weight of the bias can be accordingly adjusted to yeald the same result). Since in matrix multiplication adding new rows in the input matrix doesn't change the result of the output in respect to the old inputs (that is if we have a matrix of a rows adding a new row won't change the numerical values of the first a collumns) and we already showed that the multiplication of each row with the weight matrix gives us a transposed vector of the values in our nodes, if we want to proccess a multiple input vectors we can create a $a \times n$ matrix and multiply it by our weights ant if we were to transpose the result we would get a matrix where each row will corespond with a proccessed input vector(again we will need to add a constant as the bias, but here we need all of them to be the same). Mathematically this can be shown as:

$$z = X' \cdot W \quad \text{where B - number of input vectors}$$

$$X' = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} & 1 \\ x_{2,1} & x_{2,1} & \cdots & x_{2,n} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{B,1} & x_{B,2} & \cdots & x_{B,n} & 1 \end{bmatrix} \in \mathbb{R}^{B\times(n+1)} \quad \text{and}$$

$$W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1m} \\ w_{21} & w_{21} & \cdots & w_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ w_{B1} & w_{B2} & \cdots & w_{Bm} \\ b_1 & b_2 & \cdots & b_m \end{bmatrix} \in \mathbb{R}^{(n+1)\times M}$$

4. Discuss briefly why the fully connected layer is called "fully connected".

   The layer is called fully connected because taking any two nodes in adjacent layers(or the first layer and the input vector and accordingly with the output), the is some weight w^* that connects these two nodes.

2 Exercise 2

1. State the function of the network as defined above. Use the following notation: x represents the input, $\hat{y}$ the output of the network, W0/W1 and b0 /b1 the weights and biases of the two fully connected layers l0 (first layer) and l1 (second layer). Use $\sigma(\cdot)$ to represent the piecewise sigmoid function.

   The function of the neural network can be described by the following formula $\hat{y} = \sigma((xW_0 + b_0)W_1 + b_1)$

2. Demonstrate that the network as defined above is equivalent to a network with a single fully connected layer with weights $\tilde{W}$ and $\tilde{b}$ (write down your derivation).

   The proof is equivalent to the one done in Exercise 1 part 2. We create matrices coresponding to the weights of the connections between adjacent layers and add an extra row/column to represent the biases, as well as adding an extra value to the input vector to account for the bias. After that by continuously evaluating the formula as written in Exercise 2 Task 1, we can get a single expression.

$$\begin{aligned}\hat{y} &= \sigma\big((xW_0 + b_0)W_1 + b_1\big) \\ &= \sigma\big(xW_0W_1 + b_0W_1 + b_1\big) \\ &= \sigma(xW + b), \quad \text{where } W = W_0W_1, \quad \text{and } b = b_0W_1 + b_1.\end{aligned}$$

3 3.4 Task: SoftMax Layer

We need to show that $\text{softmax}(x) = \text{softmax}(x - b)$ to prevent overflow.

Note that $\text{softmax}(x_i) = \frac{e^{x_i}}{\sum_{j \in J} e^{x_j}}$. As such, $\text{softmax}((x-b)_i) = \frac{e^{x_i - b}}{\sum_{j \in J} e^{x_j - b}} = \frac{e^{x_i} e^{-b}}{e^{-b} \sum_{j \in J} e^{x_j}}$. After simplifying e^{-b} in both the numerator and denominator, we get the original expression for $\text{softmax}(x_i)$. $\square$