Mathematical Example of Activation-Based Method in XAI (Saliency Maps)

We have a simple neural network with a single neuron that takes in 3 inputs x_1, x_2, x_3 and produces a single output.

The activation function is the sigmoid function.

Parameters:

- Weights: $w_1 = 2$, $w_2 = -1$, $w_3 = 0.5$
- Bias: $b = 0.5$
- Input values: $x_1 = 1$, $x_2 = 2$, $x_3 = -1$

The network computes the output as:

$$z = w_1 * x_1 + w_2 * x_2 + w_3 * x_3 + b$$

$$y = \sigma(z) = 1 / (1 + \exp(-z))$$

Where $\sigma(z)$ is the sigmoid activation function.

Step 1: Compute the weighted sum z

$$z = (2 * 1) + (-1 * 2) + (0.5 * -1) + 0.5$$

$$z = 2 - 2 - 0.5 + 0.5 = 0$$

Step 2: Compute the output y using the sigmoid function

$$y = \sigma(z) = 1 / (1 + e^0) = 0.5$$

Step 3: Compute the gradient of the output with respect to each input

The gradient of y with respect to each input x_i is given by:

$$dy/dx_i = (dy/dz) * (dz/dx_i)$$

Where:

- $dy/dz = y(1 - y)$ (the derivative of the sigmoid function)
- $dz/dx_i = w_i$ (the weights of the inputs)

First, calculate dy/dz:

$dy/dz = 0.5 * (1 - 0.5) = 0.25$

Now, calculate the gradient for each input:

1. For x1:

   $dy/dx1 = 0.25 * w1 = 0.25 * 2 = 0.5$

2. For x2:

   $dy/dx2 = 0.25 * w2 = 0.25 * (-1) = -0.25$

3. For x3:

   $dy/dx3 = 0.25 * w3 = 0.25 * 0.5 = 0.125$

Step 4: Interpret the saliency map

- x1 has the highest positive contribution to the output, increasing the predicted output y.
- x2 has a negative contribution, decreasing the output.
- x3 contributes positively but to a lesser extent than x1.

This provides us with an interpretation of how each input affects the model's decision.