

Computing derivatives w.r.t Hidden Layers

Part 1

The derivatives corresponding to the hidden layers

1. What we are interested in is

a.
$$\frac{\partial L(\theta)}{\partial h_{ij}} = \sum_{m=1}^k \frac{\partial L(\theta)}{\partial a_{i+1,m}} \frac{\partial a_{i+1,m}}{\partial h_{ij}}$$

b. This formula is the summation of all the paths that lead from the concerned neuron to the loss function

c. Here, i = layer number, m = neuron number for a , j = neuron number for h

d. From the previous section, we already know how to compute $\frac{\partial L(\theta)}{\partial a_{i+1,m}}$ so we need to only

focus on
$$\frac{\partial a_{i+1,m}}{\partial h_{ij}}$$

e. However, when we compute the derivative of the neuron $a_{i+1,m}$ w.r.t h_{ij} we are left with the weight component $W_{i+1,m,j}$

f. This refers to the weight component between the output neuron ($a_{i+1,m}$) and input neuron (h_{ij})

2. Thus we have
$$\frac{\partial L(\theta)}{\partial h_{ij}} = \sum_{m=1}^k \frac{\partial L(\theta)}{\partial a_{i+1,m}} W_{i+1,m,j}$$

3. Now consider these two vectors

a.

$$\nabla_{a_{i+1}} L(\theta) = \begin{bmatrix} \frac{\partial L(\theta)}{\partial a_{i+1,1}} \\ \vdots \\ \frac{\partial L(\theta)}{\partial a_{i+1,k}} \end{bmatrix}$$

$$W_{i+1, \cdot j} = \begin{bmatrix} W_{i+1,1,j} \\ \vdots \\ W_{i+1,k,j} \end{bmatrix}$$

b. Here, $\nabla_{a_{i+1}} L(\theta)$ refers to the gradient vector of the loss function w.r.t to all output neurons from $a_{i+1,1}$ to $a_{i+1,k}$

c. And $W_{i+1, \cdot j}$ refers to all rows of the j -th column of the W_{i+1} matrix, ie a vector.

4. The dot product of these two vectors is
$$(W_{i+1, \cdot j})^T \nabla_{a_{i+1}} L(\theta) = \sum_{m=1}^k \frac{\partial L(\theta)}{\partial a_{i+1,m}} W_{i+1,m,j}$$

PadhAI: Backpropagation - the full version

One Fourth Labs

5. Here, the RHS is the same as the value from step 2. Therefore, the derivative of the loss function with respect to the hidden layers is the dot-product between the gradient of loss w.r.t output layer and the corresponding weights.