PadhAI: Backpropagation - the full version

One Fourth Labs

Computing derivatives w.r.t Output Layer

Part 1

The first derivative in the chain

1. What we are actually interested in is:
   a. $\frac{\partial L(\theta)}{\partial a_{Li}} = \frac{\partial(-\log\ \hat{y}_l)}{\partial a_{Li}}$
   b. Where L = layer number, i = neuron (from 1 to k), l = index of correct output
   c. Here, we use the cross entropy loss function
   d. In the output layer L, assume we have neurons a_{L1}, a_{L2} … a_{Lk}
   e. The output layer L involves applying the softmax function the all the neurons
   f. $\hat{y}_l = \frac{e^{a_{Ll}}}{\Sigma_i\ e^{a_{Li}}}$ again, (l refers to the index of the correct output neuron)
   g. Thus, $\hat{y}_l$ depends on all the neurons' outputs as they all appear in the denominator, thereby making the derivative non-zero for all the output neurons
2. $\frac{\partial L(\theta)}{\partial a_{Li}} = \frac{\partial(-\log\ \hat{y}_l)}{\partial a_{Li}} = \frac{\partial(-\log\ \hat{y}_l)}{\partial \hat{y}_l} \frac{\partial \hat{y}_l}{\partial a_{Li}}$
3. From the previous points, we know that $\hat{y}_l$ depends on a_{Li}
4. The first part of the derivative is fairly straightforward (of the form $\frac{\partial \log x}{\partial x}$)
5. $\frac{\partial(-\log\ \hat{y}_l)}{\partial \hat{y}_l} = \frac{-1}{\hat{y}_l}$