PadhAI: Backpropagation - the full version

One Fourth Labs

Computing derivatives w.r.t Output Layer
Part 2

1. Continuing from where we left off $\frac{\partial L(\theta)}{\partial a_{Li}} = \frac{-1}{\hat{y}_l} \frac{\partial \hat{y}_l}{\partial a_{Li}}$
2. Here, we know that $\hat{y}_l = \frac{e^{a_{Ll}}}{\Sigma_i a^{Li}}$ (taking the l-th entry of the softmax function applied to vector a_L)
3. $\frac{\partial L(\theta)}{\partial a_{Li}} = \frac{-1}{\hat{y}_l} \frac{\partial\, softmax(a_L)_l}{\partial a_{Li}}$ where $a_L = [a_{L1}, a_{L2} \ldots a_{Lk}]$
   a. Where $softmax(a_L) = [\frac{e^{a_{L1}}}{\Sigma_i e^{a_{Li}}}, \frac{e^{a_{L2}}}{\Sigma_i e^{a_{Li}}} \ldots \frac{e^{a_{Lk}}}{\Sigma_i e^{a_{Li}}}]$
   b. Selecting the l-th entry would give us the value $softmax(a_L)_l = \frac{exp(a_L)_l}{\Sigma_i exp(a_L)i}$
4. $\frac{\partial L(\theta)}{\partial a_{Li}} = \frac{-1}{\hat{y}_l} \frac{\partial}{\partial a_{Li}} \frac{exp(a_L)_l}{\Sigma_{i'} exp(a_L)_{i'}}$
   a. This is of the form $\frac{g(x)}{h(x)}$ $which\ gives\ derivatives$ $\frac{\partial \frac{g(x)}{h(x)}}{\partial x} = \frac{\partial g(x)}{\partial x} \frac{1}{h(x)} - \frac{g(x)}{h(x)^2} \frac{\partial h(x)}{\partial x}$
   b. Here $g(x) = exp(a_L)_l\ and\ h(x) = \Sigma_{i'} exp(a_L)_{i'}$
   c. Substitute the values and expand the formula
5. $\frac{\partial L(\theta)}{\partial a_{Li}} = \frac{-1}{\hat{y}_l} (\frac{\frac{\partial}{\partial a_{Li}} exp(a_L)_l}{\Sigma_{i'} exp(a_L)_{i'}} - \frac{exp(a_L)_l\ (\frac{\partial}{\partial a_{Li}} \Sigma_{i'} exp(a_L)_{i'})}{(\Sigma_{i'} exp(a_L)_{i'})^2})$
   a. Here, consider $\frac{\partial}{\partial a_{Li}} exp(a_L)_l$, this value is 0 for all values of i : 0 to k except for when i = *l*
   b. Thus, we use an indicator variable $1_{(l=i)} \exp(a_L)_l$ to denote that all other values except i=*l* resolve to 0
   c. Now consider $\frac{\partial}{\partial a_{Li}} \Sigma_{i'} exp(a_L)_{i'}$, here i' ranges from 1 to k. When taking the derivative, only the index i=i' remains, which is simply a derivative of an exponent.
6. $\frac{\partial L(\theta)}{\partial a_{Li}} = \frac{-1}{\hat{y}_l} (\frac{1_{(l=i)} \exp(a_L)_l}{\Sigma_{i'} exp(a_L)_{i'}} - \frac{exp(a_L)_l}{\Sigma_{i'} exp(a_L)_{i'}} \frac{exp(a_L)_i}{\Sigma_{i'} exp(a_L)_{i'}})$
   a. This is can be rewritten in terms of the softmax function for the different variables
7. $\frac{\partial L(\theta)}{\partial a_{Li}} = \frac{-1}{\hat{y}_l} (1_{(l=i)} softmax(a_L)_l - softmax(a_L)_l softmax(a_L)_i$
   a. We know that the Softmax function is ŷ, so we rewrite it.
8. $\frac{\partial L(\theta)}{\partial a_{Li}} = \frac{-1}{\hat{y}_l} (1_{(l=i)} \hat{y}_l - \hat{y}_l \hat{y}_i)$
9. After cancellation $\frac{\partial L(\theta)}{\partial a_{Li}} = -(1_{(l=i)} - \hat{y}_i)$