

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 $\hat{y}_l = \frac{e^{a_{Ll}}}{\sum_i e^{a_{Li}}}$ (taking the l-th entry of the softmax function applied to vector a_L)
3. $\frac{\partial L(\theta)}{\partial a_{Li}} = \frac{-1}{\hat{y}_l} \frac{\partial \text{softmax}(a_L)_l}{\partial a_{Li}}$ where $a_L = [a_{L1}, a_{L2} \dots a_{Lk}]$
 - a. Where $\text{softmax}(a_L) = [\frac{e^{a_{L1}}}{\sum_i e^{a_{Li}}}, \frac{e^{a_{L2}}}{\sum_i e^{a_{Li}}} \dots \frac{e^{a_{Lk}}}{\sum_i e^{a_{Li}}}]$
 - b. Selecting the l-th entry would give us the value $\text{softmax}(a_L)_l = \frac{\exp(a_L)_l}{\sum_i \exp(a_L)_i}$
4. $\frac{\partial L(\theta)}{\partial a_{Li}} = \frac{-1}{\hat{y}_l} \frac{\partial}{\partial a_{Li}} \frac{\exp(a_L)_l}{\sum_{i'} \exp(a_L)_{i'}}$
 - 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 $h(x) = \sum_{i'} \exp(a_L)_{i'}$
 - c. Substitute the values and expand the formula
5. $\frac{\partial L(\theta)}{\partial a_{Li}} = \frac{-1}{\hat{y}_l} \left(\frac{\frac{\partial}{\partial a_{Li}} \exp(a_L)_l}{\sum_{i'} \exp(a_L)_{i'}} - \frac{\exp(a_L)_l \left(\frac{\partial}{\partial a_{Li}} \sum_{i'} \exp(a_L)_{i'} \right)}{(\sum_{i'} \exp(a_L)_{i'})^2} \right)$
 - 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 $\frac{\partial}{\partial a_{Li}} \sum_{i'} \exp(a_L)_{i'}$, 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. $\frac{\partial L(\theta)}{\partial a_{Li}} = \frac{-1}{\hat{y}_l} \left(\frac{1_{(l=i)} \exp(a_L)_l}{\sum_{i'} \exp(a_L)_{i'}} - \frac{\exp(a_L)_l}{\sum_{i'} \exp(a_L)_{i'}} \frac{\exp(a_L)_i}{\sum_{i'} \exp(a_L)_{i'}} \right)$
 - a. This can be rewritten in terms of the softmax function for the different variables
7. $\frac{\partial L(\theta)}{\partial a_{Li}} = \frac{-1}{\hat{y}_l} (1_{(l=i)} \text{softmax}(a_L)_l - \text{softmax}(a_L)_l \text{softmax}(a_L)_i)$
 - a. We know that the Softmax function is \hat{y} , 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 $\frac{\partial L(\theta)}{\partial a_{Li}} = -(1_{(l=i)} - \hat{y}_i)$