PadhAI: Backpropagation - the full version

One Fourth Labs

Computing derivatives w.r.t Output Layer
Part 3

1. So far, we have derived the partial derivative with respect to the *i*-th element of layer a_L
   a. $\frac{\partial L(\theta)}{\partial a_{Li}} = -(1_{(l=i)} - \hat{y}_i)$
2. We can now write the gradient w.r.t the vector a_L
3. As we saw earlier, $a_L = [a_{L1}, a_{L2} \dots a_{Lk}]$
4. Going by the indicator variable in step 1, it resolves to 0 for all values of *i* except for *i* = *l*
5. Let us assume a scenario where k = 4, and *l* = 2
   a. $\frac{\partial L(\theta)}{\partial a_{L1}} = -(0 - \hat{y}_i)$
   b. $\frac{\partial L(\theta)}{\partial a_{L2}} = -(1 - \hat{y}_i)$
   c. $\frac{\partial L(\theta)}{\partial a_{L3}} = -(0 - \hat{y}_i)$
   d. $\frac{\partial L(\theta)}{\partial a_{L4}} = -(0 - \hat{y}_i)$
   e. Here, the indicator variable values as a vector would be
6. The gradient w.r.t a_L is $\nabla_{a_L} =$

$$\nabla_{a_L} = \begin{bmatrix} \frac{\partial L(\theta)}{\partial a_{L1}} \\ \cdot \\ \cdot \\ \cdot \\ \frac{\partial L(\theta)}{\partial a_{Lk}} \end{bmatrix}$$

$$\nabla_{a_L} = \begin{bmatrix} -(1_{(l=1)} - \hat{y}_i) \\ \cdot \\ \cdot \\ \cdot \\ -(1_{(l=k)} - \hat{y}_i) \end{bmatrix}$$

7. The above can be seen as a difference of two vectors, $[0, 1, 0, \dots 0_k]$ and ŷ
8. The first vector is essentially the one hot representation of the true output e(*l*): $-(e(l) - \hat{y}_i)$
9. In reality, this is simply the difference between the true distribution y and the predicted distribution ŷ
10. $\nabla_{a_L} L(\theta) = -(y - \hat{y}_i)$