PadhAI: Information Theory and Cross Entropy

One Fourth Labs

Computing Partial Derivatives With Cross Entropy Loss

How do we compute Δw and Δb

1. Loss Function $L(\theta) = -[(1-y)log(1-\hat{y}) + ylog(\hat{y})]$
2. Consider Δw for 1 training example
   a. $\Delta w = \frac{\partial L(\theta)}{\partial w} = \frac{\partial L(\theta)}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial w}$
   b. The first part: $\frac{\partial L(\theta)}{\partial w} = \frac{\partial}{\partial \hat{y}}\{-(1-y)log(1-\hat{y}) - ylog(\hat{y})\}$
      i. $\frac{\partial L(\theta)}{\partial w} = (-)(-1)\frac{(1-y)}{(1-\hat{y})} - \frac{y}{\hat{y}}$
      ii. $\frac{\partial L(\theta)}{\partial w} = \frac{\hat{y}(1-y) - y(1-\hat{y})}{(1-\hat{y})\hat{y}}$
      iii. $\frac{\partial L(\theta)}{\partial w} = \frac{\hat{y}-y}{(1-\hat{y})\hat{y}}$
   c. The second part: $\frac{\partial \hat{y}}{\partial w} = \frac{\partial}{\partial w}\frac{1}{1+e^{-(w.x+b)}}$
      i. This is the exact same as from the Squared Error Loss
      ii. $\frac{\partial}{\partial w}(\frac{1}{1+e^{-(wx+b)}})$
      iii. $\frac{-1}{(1+e^{-(wx+b)})^2}\frac{\partial}{\partial w}(e^{-(wx+b)})$
      iv. $\frac{-1}{(1+e^{-(wx+b)})^2} * (e^{-(wx+b)})\frac{\partial}{\partial w}(-(wx + b))$
      v. $\frac{-1}{(1+e^{-(wx+b)})^2} * (e^{-(wx+b)}) * (-x)$
      vi. $\frac{1}{(1+e^{-(wx+b)}} * \frac{(e^{-(wx+b)})}{(1+e^{-(wx+b)})} * (x)$
      vii. $\hat{y} * (1-\hat{y}) * x$
   d. The final derivative is the first part multiplied with the second part
   e. $\Delta w = (\hat{y} - y) * x$
   f. $|||^{ly}\ \Delta b = (\hat{y} - y)$
3. We then plug the values of Δw and Δb into the learning algorithm to optimise the parameters w and b.