Assignment - 2

CS 4783/5783

27th Sept 2021

Nagasri Bhavani Santhoshi Adigoth
A20314248

① Derive the Update rule and show how to train a 3-layer (1- input layer, 1 - hidden layer, 1- Output layer) Network with Backpropogation for regression Using the mean Square Error loss. Assume that you are Using the Sigmoid activation function for the hidden Layer. Explain Breifly how this is different from the Update rule for the network trained for Binary Classification Using log loss

Backpropogation for regression Using the mean Square Error loss

$x \rightarrow$ input $\quad n \times f$

$y \rightarrow$ Labels $\quad n \times 1$

First layer weights : W_1 $\quad$ Bias : b_1

Second layer weights : W_2 $\quad$ Bias : b_2

Third layer weights : W_3 $\quad$ Bias : b_3

O/p of first layer : $Z_1 = W_1 x + b_1$

O/p of After Activation function : $a_1 = g(Z_1)$

O/p of Second layer : $Z_2 = W_2 a_1 + b_2$

After Activation function : $a_2 = g(Z_2)$

O/p of third layer : $Z_3 = W_3 a_2 + b_3$

Final prediction : $\hat{y} = a_3 = g(Z_3)$

$$L = -[(1-y)\log(1-\hat{y}) + y \cdot \log \hat{y}]$$

We update the weights using gradient descent

(1) Start with an initial guess for w_1, w_2, w_3
b_1, b_2, b_3

(2) Update weights by

$$w_i = w_i - \alpha \frac{\partial L}{\partial w_i}$$

$$b_i = b_i - \alpha \frac{\partial L}{\partial b_i}$$

Where i refers to the i^{th} layer

(3) Repeat Until Convergence

⇒ To find $\frac{\partial L}{\partial w_3}$

$$\frac{\partial L}{\partial w_3} = -\frac{\partial}{\partial w_3}\left((1-y).\log(1-\hat{y}) + y.\log\hat{y}\right)$$

$$= -(1-y).\frac{\partial}{\partial w_3}(1-\hat{y}) - y.\frac{\partial}{\partial w_3}\log\hat{y}$$

$\because \hat{y} = a_3 = g(z_3)$

$$\frac{\partial L}{\partial w_3} = -(1-y).\frac{\partial}{\partial w_3}(1-\log(g(z_3)) - y.\frac{\partial}{\partial w_3}\log(g(z$$

$\because \frac{\partial}{\partial z}\log(z) = \frac{1}{z}$

$$\frac{\partial L}{\partial w_3} = -\frac{(1-y)}{1-g(z_3)}\left(\frac{-\partial}{\partial w_3}(g(z_3))\right) - \frac{y}{g(z_3)}\frac{\partial}{\partial w_3}g(z_3)$$

$$= \frac{(1-y)}{1-g(z_3)} \cdot g'(z_3) - \frac{y}{g(z_3)} \cdot g'(z_3)$$

$\because$ for Sigmoid Activation

$$g'(z) = g(z)(1-g(z))$$

$$\frac{\partial L}{\partial w_3} = \frac{(1-y)}{1-g(z_3)} \cdot g(z_3)(1-g(z_3)) \frac{\partial z_3}{\partial w_3}$$

$$- \frac{y}{g(z_3)} \cdot g(z_3)(1-g(z_3)) \frac{\partial z_3}{\partial w_3}$$

$$= ((1-y) \cdot g(z_3) - y(1-g(z_3))] \frac{\partial z_3}{\partial w_3}$$

$$= (g(z_3) - y) \frac{\partial z_3}{\partial w_3}$$

$$= [a_3 - y] \cdot \frac{\partial}{\partial w_3}(w_3 a_2 + b_3)$$

$$\frac{\partial L}{\partial w_3} = (a_3 - y) \cdot a_2^T$$

Similarly, when we diff L w.r.t b_3 we get

$$\boxed{\frac{\partial L}{\partial b_3} = a_3 - y}$$

To derive $\frac{\partial L}{\partial w_2}$, we can use the "chain rule"

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{?}\frac{?}{\partial w_2}$$

$$= \frac{\partial L}{\partial a_3}\cdot\frac{\partial L}{?}\cdot\frac{?}{\partial w_2}$$

$\because a_3 = g(z_3)$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial a_3}\cdot\frac{\partial L}{\partial z_3}\cdot\frac{\partial z_3}{?}\cdot\frac{?}{\partial w_2}$$

$\because z_3 = w_3 a_2 + b_2$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial a_3}\cdot\frac{\partial L}{\partial z_3}\cdot\frac{\partial z_3}{\partial a_2}\cdot\frac{\partial a_2}{?}\cdot\frac{?}{\partial w_2}$$

$\because a_2 = g(z_2) = g(w_2 a_1 + b_2)$

$$\frac{\partial L}{\partial w_2} = \underbrace{\frac{\partial L}{\partial a_3}\cdot\frac{\partial L}{\partial z_3}}_{(a_2 - y)}\cdot\frac{\partial z_3}{\partial a_2}\cdot\frac{\partial a_2}{\partial z_2}\cdot\underbrace{\frac{\partial z_2}{\partial w_2}}_{a_2^T}$$

$\therefore \frac{\partial L}{\partial a_3}\cdot\frac{\partial a_3}{\partial z_3} = a_2 - y$

$z_3 = w_3 a_2 + b_3$

$\therefore \frac{\partial z_3}{\partial a_2} = w_3$

$a_2 = g(z_2)$

$\therefore \frac{\partial a_2}{\partial z_2} = g'(z_2)$

$\because Z_2 = W_2 a_1 + b_2$

$$\frac{\partial Z_2}{\partial W_2} = a_1$$

$$\therefore \frac{\partial L}{\partial W_2} = (a_3 - y) \cdot W_3 \cdot g'(Z_2) a_1$$

Consider an Example.

But Dimensions of Each Matrix dont match

$a_3 - y \longrightarrow 1 \times 1$

$W_3 \rightarrow 1 \times h_2$

$g'(Z_2) \rightarrow h_2 \times 1$

$a_1 \longrightarrow h_1 \times 1$

$$\therefore \frac{\partial L}{\partial W_2} = W_3^T \cdot g'(Z_2)(a_3 - y) \cdot a_1^T$$

Similarly,

$$\frac{\partial L}{\partial b_2} = W_3^T \cdot g'(Z_2)(a_3 - y)$$

Using chain rule,

$$\frac{\partial L}{\partial W_1} = \frac{\partial L}{\partial a_3} \cdot \frac{\partial a_3}{\partial Z_3} \cdot \frac{\partial Z_3}{\partial a_2} \cdot \frac{\partial a_2}{\partial Z_2} \cdot \frac{\partial Z_2}{\partial a_1} \cdot \frac{\partial a_1}{\partial Z_1} \cdot \frac{\partial Z_1}{\partial W_1}$$

$$\frac{\partial L}{\partial W_1} = (a_3 - y) \cdot W_3 \cdot g'(Z_2) \cdot W_2 \cdot g'(Z_1) \cdot x$$

2.1 Report the Average MSE and the accuracy.

MSE 2858.4015

Accuracy 0.6107

2.2 Plot the loss and accuracy as a function of no. of iterations

plotted in .ipynb file

2.3 What is the Effect of learning rate on the training process?

Shown in the .ipynb file

2.4 What is the Effect of number of neurons in hidden layer?

Shown in the .ipynb file

2.5 What is the Effect of Activation function in the network?

Used tanh and ReLU, Shown in .ipynb file