Matthew Maya

Homework #4

1.) $h_1 = f_1(w_1 x + b_1)$ $\quad h_3 = \hat{y} = f_3(w_3 h_1 + w_4 h_2 + b_3)$

$h_2 = f_2(w_2 x + b_2)$ $\quad f_n' = \frac{df_n(v)}{dv}$ $\quad n = 1, 2, 3$

$L = (\hat{y} - y)^2$

$\frac{dL}{d\hat{y}} = 2(\hat{y} - y)$ $\quad \frac{dL}{dw_1} = \frac{dL}{dh_3} \cdot \frac{dh_3}{dh_1} \cdot \frac{dh_1}{dw_1} = \frac{dL}{dh_3} \cdot f_3' w_3 \cdot f_1' x$

$$\frac{dL}{dw_2} = \frac{dL}{dh_3} \cdot \frac{dh_3}{dh_2} \cdot \frac{dh_2}{dw_2} = \frac{dL}{dh_3} \cdot f_3' w_4 \cdot f_2' x$$

$$\frac{dL}{dw_3} = \frac{dL}{dh_3} \cdot \frac{dh_3}{dw_3} = \frac{dL}{dh_3} \cdot f_3' h_1$$

$$\frac{dL}{dw_4} = \frac{dL}{dh_3} \cdot \frac{dh_3}{dw_4} = \frac{dL}{dh_3} \cdot f_3' h_2$$

$$\frac{dL}{db_1} = \frac{dL}{dh_3} \cdot \frac{dh_3}{dh_1} \cdot \frac{dh_1}{db_1} = \frac{dL}{dh_3} \cdot f_3' w_3 \cdot f_1'$$

$$\frac{dL}{db_2} = \frac{dL}{dh_3} \cdot \frac{dh_3}{dh_2} \cdot \frac{dh_2}{db_2} = \frac{dL}{dh_3} \cdot f_3' w_4 \cdot f_2'$$

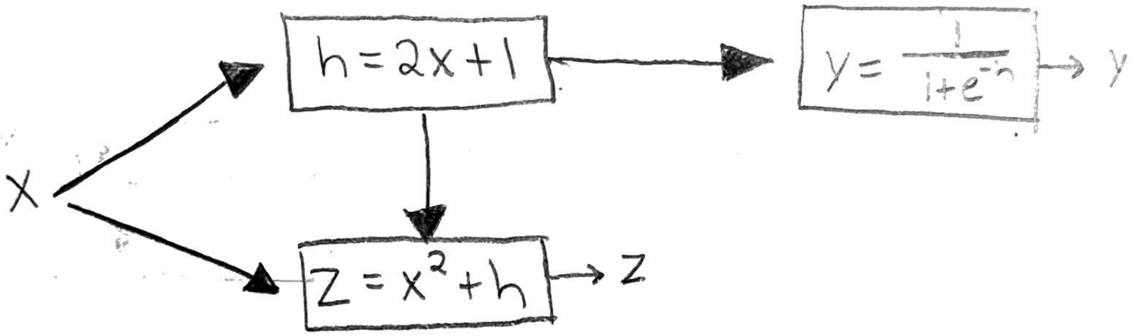
$$\frac{dL}{db_3} = \frac{dL}{dh_3} \cdot \frac{dh_3}{db_3} = \frac{dL}{dh_3} \cdot f_3'$$

$$\frac{dL}{dx} = \frac{dL}{dh_3} \cdot \frac{dh_3}{dh_1} \cdot \frac{dh_1}{dx} = \frac{dL}{dh_3} \cdot f_3' w_3 \cdot f_1' w_1$$

or $\frac{dL}{dh_3} \cdot \frac{dh_3}{dh_2} \cdot \frac{dh_2}{dx} = \frac{dL}{dh_3} \cdot f_3' w_4 \cdot f_2' w_2$

2.) $h = 2x + 1 \rightarrow z = x^2 + h \rightarrow y = \frac{1}{1+e^{-h}}$

a.)



b.) $\frac{dy}{dz} = \frac{dy}{dh} \cdot \frac{dh}{dx} \cdot \frac{dx}{dz} = \frac{e^{-h}}{(1+e^{-h})^2} \cdot 2 \cdot 0 = 0$?

$$\frac{dy}{dh} = (1+e^{-h})^{-1} \rightarrow -(1+e^{-h})^{-2} \underset{\substack{u=-h \\ du=-1}}{\frac{dy}{dh}(1+e^{-h})} \rightarrow \frac{-(-e^{-h})}{(1+e^{-h})^2} = \frac{e^{-h}}{(1+e^{-h})^2}$$

3.) $\hat{y}_{(1)}$ = monthly income (0 - 10,000)

$\hat{y}_{(2)}$ = age (0 - 100)

MSE loss : $L = (\hat{y}_{(1)} - y_{(1)})^2 + (\hat{y}_{(2)} - y_{(2)})^2$

yes, normalization is necessary because the ranges of y_1 and y_2 are pretty different. y_1 values will be dominant in the decision-making.

We can apply standardization, Min-Max normalization, or we can just divide every data point with the constraint (Max possible value. y_1 by 10,000 y_2 by 100)

4a.) ReLU function

$$f(z) = \max(0, z)$$

4b.) negative ReLU function

$$f(z) = \min(0, z)$$

4c.) combination of the sigmoid and linear activation functions

$$f(z) = a + (b+a) * \text{sigmoid}(z)$$

where sigmoid activation function is defined as:

$$\text{sigmoid}(z) = \frac{1}{(1 + \exp(-z))}$$

5a.) For batch normalization to work you need the mean and standard deviation corresponding to the mini-batch. If batch size is too small, the sample mean and sample standard deviation aren't representative of the actual distribution and the network might not learn anything meaningful

5b.) Layer Normalization normalizes each of the inputs in the batch independently across all features. It's independent of the batch size, so it can be applied to batches with smaller sizes as well.

6.) skip/residual connections are use in building deep neural networs because by using skip connections the gradients can be propagated more easily to address vanishing gradients, they enable deeper networks by allowing information to bypass some layers in the network b/c the information can be propagated directly without getting attenuated by the intermediate layers, improving convergence by learning residual mapping and the idenfity mapping, finally it can reduce overfitting.

7.) There can be many cause for randomness. Some include:

- order of training set
- initialization of the weights
- stochasity of the training set
- presence of noise in data

when writing a paper, its recommended to report the results of multiple models, rather then the best or worse as the performance of the models can vary due to those randomness factors. That's why it would be better representative to report all 3 models.