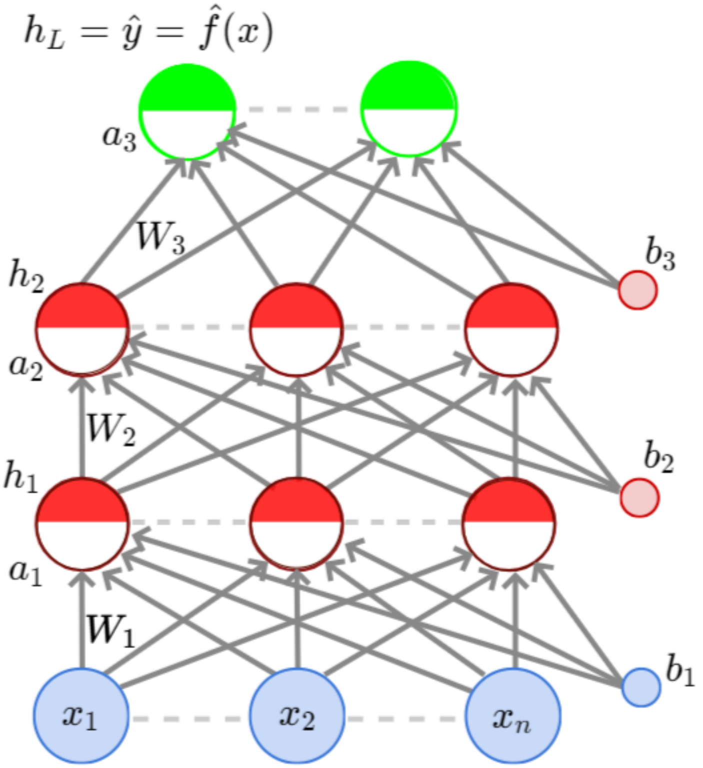
Deep Learning

Backpropagation

1 ANN

1.1 Forward Pass



Data:

$$\mathbf{x} = [x_1, x_2, \cdots, x_n]^T \in \mathbb{R}^n$$
$$\mathbf{y} \in \mathbb{R}^k \quad (k \text{ class classification problem})$$

Description of Network Structure and Convention:

Consider a network comprising an input layer, $(L-1)$ hidden layers, and an output layer. The input layer consists of n nodes (neurons), while the output layer comprises k nodes for a classification problem with k classes.

The preactivation of the ith hidden layer is denoted by a_i, and its activation is represented by h_i.

let the number of neurons in $(i-1)$th and ith layers are p and m, respectively, then

$$W_i = \begin{bmatrix} w_{i11} & w_{i12} & \cdots & w_{i1p} \\ w_{i21} & w_{i22} & \cdots & w_{i2p} \\ \vdots & \vdots & \cdots & \vdots \\ w_{im1} & w_{im2} & \cdots & w_{imp} \end{bmatrix} \in \mathbb{R}^{m \times p}$$

So, the weights in the first row are the weights coming to first node of ith layer.

$$\begin{aligned} a_i &= W_i h_{i-1} + b_i \in \mathbb{R}^m \\ h_i &= \sigma(a_i) \in \mathbb{R}^m \\ \hat{y} &= h_L = O(a_L) \in \mathbb{R}^k \end{aligned}$$

Here, σ and O are activation functions at hidden layers and output layer, respectively.

Loss function:

Let the true label for a training example $\mathbf{x}$ is l, then loss corresponding to that example is

$$L = -\ln \hat{y}_l$$

Backward Pass:

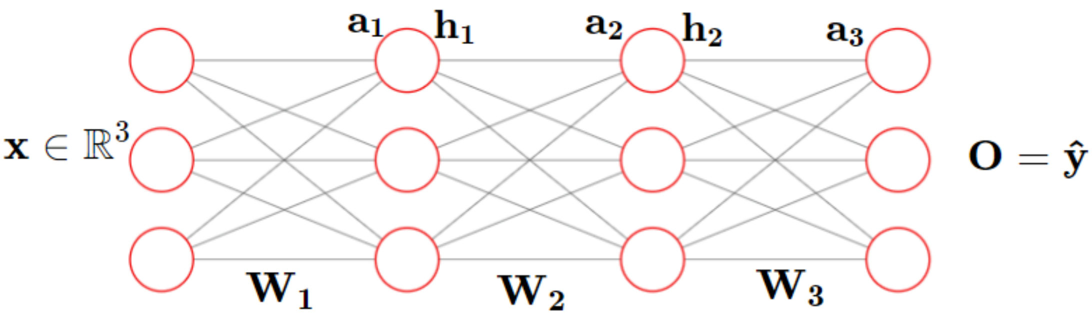
let the activation function at hidden layers be logistic function and at the output layer be softmax function.

$$\begin{aligned} \sigma(z) &= \frac{1}{1+e^{-z}} \\ \sigma'(z) &= \sigma(z)(1-\sigma(z)) \end{aligned}$$

$$\begin{aligned} \nabla_{\hat{y}}(L) &= -\frac{e_l}{\hat{y}_l} \in \mathbb{R}^k \\ \nabla_{a_L}(L) &= -(e_l - \hat{y}) \in \mathbb{R}^k \\ \nabla_{h_i}(L) &= W_{i+1}^T(\nabla_{a_{i+1}} L) \in \mathbb{R}^m \\ \nabla_{a_i}(L) &= \nabla_{h_i}(L) \odot \sigma'(a_i) \in \mathbb{R}^m \\ \nabla_{W_i}(L) &= (\nabla_{a_i} L).h_{i-1}^T \in \mathbb{R}^{m \times p} \end{aligned}$$

Example:

Consider a feed forward neural network shown below



where, $\mathbf{x}$ is an input vector. The vectors a_l, h_l correspond to pre-activation and activation at layer l. The matrices $\mathbf{W_l}$ are weights that connect neurons from layer $l-1$ to layer l. Finally, the vector $\mathbf{o}$ is an output vector $\mathbf{o} = \mathbf{h_3} = \hat{y}$. All neurons in the hidden layer use the logistic activation function, and neurons in the output layer use softmax function. Further, the network minimizes cross entropy loss. Consider no biases in the network.

Let initialization of weights are:

$$\mathbf{W_1} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

$$\mathbf{W_2} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{W_3} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

1. What will be the loss for the data point $\mathbf{x} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$, if the true label of the point is encoded as $\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$?

   Solution:

$$\mathbf{h_0} = \mathbf{x} = [1 \quad 0 \quad 1]^T$$

$$\begin{aligned}\mathbf{a_1} &= \mathbf{W_1 h_0} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}\end{aligned}$$

$$\begin{aligned}\mathbf{h_1} &= \sigma(\mathbf{a_1}) \\ &= \begin{bmatrix} \frac{1}{1+e^{-1}} \\ \frac{1}{1+e^{-2}} \\ \frac{1}{1+e^{-1}} \end{bmatrix} \\ &= \begin{bmatrix} 0.73 \\ 0.88 \\ 0.73 \end{bmatrix}\end{aligned}$$

$$\begin{aligned}\mathbf{a_2} &= \mathbf{W_2 h_1} \\ &= \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0.73 \\ 0.88 \\ 0.73 \end{bmatrix} \\ &= \begin{bmatrix} 2.34 \\ 2.34 \\ 0.73 \end{bmatrix}\end{aligned}$$

$$\begin{aligned}\mathbf{h_2} &= \sigma(\mathbf{a_2}) \\ &= \begin{bmatrix} \frac{1}{1+e^{-2.34}} \\ \frac{1}{1+e^{-2.34}} \\ \frac{1}{1+e^{-0.73}} \end{bmatrix} \\ &= \begin{bmatrix} 0.91 \\ 0.91 \\ 0.67 \end{bmatrix}\end{aligned}$$

$$\begin{aligned}\mathbf{a_3} &= \mathbf{W_3 h_2} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0.91 \\ 0.91 \\ 0.67 \end{bmatrix} \\ &= \begin{bmatrix} 0.91 \\ 0.91 \\ 2.49 \end{bmatrix}\end{aligned}$$

$$\begin{aligned}\mathbf{\hat{y}} &= \mathbf{h_3} = O(\mathbf{a_3}) \\ &= \begin{bmatrix} \dfrac{e^{0.91}}{e^{0.91} + e^{0.91} + e^{2.49}} \\ \dfrac{e^{0.91}}{e^{0.91} + e^{0.91} + e^{2.49}} \\ \dfrac{e^{2.49}}{e^{0.91} + e^{0.91} + e^{2.49}} \end{bmatrix} \\ &= \begin{bmatrix} 0.15 \\ 0.15 \\ 0.70 \end{bmatrix}\end{aligned}$$

$$\begin{aligned}L &= -\ln \hat{y}_l \\ &= -\ln 0.7 = 0.35\end{aligned}$$

2. Find the value of $\dfrac{\partial L}{\partial \mathbf{W_{311}}}$.
   Solution:

$$\frac{\partial L}{\partial \mathbf{W_{311}}} = \frac{\partial L}{\partial \mathbf{a_3}} . \frac{\partial \mathbf{a_3}}{\partial \mathbf{W_{311}}} \tag{1}$$

Now,

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{a_3}} &= -(e_l - \hat{y}) \\
&= -\left(\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} - \begin{bmatrix} 0.15 \\ 0.15 \\ 0.7 \end{bmatrix} \right) \\
&= \begin{bmatrix} 0.15 \\ 0.15 \\ -0.3 \end{bmatrix} \qquad (2)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \mathbf{a_3}}{\partial \mathbf{W_{311}}} &= \begin{bmatrix} \frac{\partial \mathbf{a_{31}}}{\partial \mathbf{W_{311}}} \\ \frac{\partial \mathbf{a_{32}}}{\partial \mathbf{W_{311}}} \\ \frac{\partial \mathbf{a_{33}}}{\partial \mathbf{W_{311}}} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{h_{21}} \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.91 \\ 0 \\ 0 \end{bmatrix} \qquad (3)
\end{aligned}$$

From eq (1), (2), and (3)

$$\frac{\partial L}{\partial \mathbf{W_{311}}} = 0.15 \times 0.91 = 0.13$$

3. Find the value of $\frac{\partial L}{\partial \mathbf{W_{123}}}$.

   Solution:

$$\frac{\partial L}{\partial \mathbf{W_{123}}} = \frac{\partial L}{\partial \mathbf{a_1}} . \frac{\partial \mathbf{a_1}}{\partial \mathbf{W_{123}}} \tag{1}$$

$$\frac{\partial L}{\partial \mathbf{a_1}} = \frac{\partial L}{\partial \mathbf{h_1}} \odot \sigma'(a_1) \tag{2}$$

$$\frac{\partial L}{\partial \mathbf{h_1}} = \mathbf{W_2}^T . \frac{\partial L}{\partial \mathbf{a_2}} \tag{3}$$

$$\frac{\partial L}{\partial \mathbf{a_2}} = \frac{\partial L}{\partial \mathbf{h_2}} \odot \sigma'(a_2) \tag{4}$$

$$\frac{\partial L}{\partial \mathbf{h_2}} = \mathbf{W_3}^T . \frac{\partial L}{\partial \mathbf{a_3}} \tag{5}$$

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{h_2}} &= \begin{bmatrix}1 & 1 & 1\\0 & 0 & 1\\0 & 0 & 1\end{bmatrix} . \begin{bmatrix}0.15\\0.15\\-0.3\end{bmatrix}\\ &= \begin{bmatrix}0\\-0.3\\-0.3\end{bmatrix}\end{aligned}$$

And

$$\begin{aligned}\sigma'(a_2) &= \begin{bmatrix}\sigma(2.34)(1-\sigma(2.34))\\\sigma(2.34)(1-\sigma(2.34))\\\sigma(0.73)(1-\sigma(0.73))\end{bmatrix}\\ &= \begin{bmatrix}0.08\\0.08\\0.21\end{bmatrix}\end{aligned}$$

Putting back in the eq (4)

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{a_2}} &= \begin{bmatrix}0\\-0.3\\-0.3\end{bmatrix} \odot \begin{bmatrix}0.08\\0.08\\0.21\end{bmatrix}\\ &= \begin{bmatrix}0\\-0.024\\-0.063\end{bmatrix}\end{aligned}$$

Putting back in eq (3)

$$\frac{\partial L}{\partial \mathbf{h_1}} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} . \begin{bmatrix} 0 \\ -0.024 \\ -0.063 \end{bmatrix}$$

$$= \begin{bmatrix} -0.024 \\ -0.024 \\ -0.087 \end{bmatrix}$$

And

$$\sigma'(a_1) = \begin{bmatrix} \sigma(1)(1-\sigma(1)) \\ \sigma(2)(1-\sigma(2)) \\ \sigma(1)(1-\sigma(1)) \end{bmatrix}$$

$$= \begin{bmatrix} 0.2 \\ 0.1 \\ 0.2 \end{bmatrix}$$

Putting back in the eq (2)

$$\frac{\partial L}{\partial \mathbf{a_1}} = \begin{bmatrix} -0.024 \\ -0.024 \\ -0.087 \end{bmatrix} \odot \begin{bmatrix} 0.2 \\ 0.1 \\ 0.2 \end{bmatrix}$$

$$= \begin{bmatrix} -0.0048 \\ -0.002 \\ -0.0174 \end{bmatrix}$$

Therefore,

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{W_{123}}} &= \frac{\partial L}{\partial \mathbf{a_{12}}} . \frac{\partial \mathbf{a_{12}}}{\partial \mathbf{W_{123}}} \\ &= \frac{\partial L}{\partial \mathbf{a_{12}}} . \mathbf{h_{03}} \\ &= -0.002 \times 1 = -0.002 \end{aligned}$$

2 CNN:

The backward pass of a convolutional layer during Backpropagation also uses Convolutions.

If The output of convolution operation on $\mathbf{X}$ by applying kernel $\mathbf{K}$ is $\mathbf{F}$, then

$$\frac{\partial L}{\partial \mathbf{K}} = \mathbf{X} * \frac{\partial L}{\partial \mathbf{F}}$$

Example:

1. Consider an input array $X = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$ and the corresponding output label $y = 1$. Suppose we use the kernel/filter $K = \begin{bmatrix} 1 & -1 \end{bmatrix}$ for the convolution operation. Convolve the kernel K over the input X with stride $s = 1$, no padding is applied.

   Pass a through the ReLU activation function and assign the output to h. Compute the output $\hat{y}$ by averaging h and calculate the squared error loss

   $$L = 0.5(\hat{y} - y)^2$$

   Enter the loss value.

   Solution:

$$X = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$
$$K = \begin{bmatrix} 1 & -1 \end{bmatrix}$$

$$\begin{aligned} a &= X * K \\ &= \begin{bmatrix} 1 & -1 & 1 & -1 & 1 & -1 & 1 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} h &= \text{Relu} \begin{bmatrix} 1 & -1 & 1 & -1 & 1 & -1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} \hat{y} &= \text{Avg}(h) \\ &= 4/7 \end{aligned}$$

Therefore,

$$\begin{aligned} L &= 0.5(1 - 4/7)^2 \\ &= 0.09 \end{aligned}$$

2. Compute the gradient of loss with respect to K (that is, $\nabla_L K$). What is the sum of the gradients?
Solution:

$$\frac{\partial L}{\partial K} = X * \frac{\partial L}{\partial a} \tag{1}$$

Now,

$$\frac{\partial L}{\partial a} = \frac{\partial L}{\partial \hat{y}} . \frac{\partial \hat{y}}{\partial h} . \frac{\partial h}{\partial a} \tag{2}$$

$$\begin{aligned} L &= 0.5(\hat{y} - y)^2 \\ \Rightarrow \frac{\partial L}{\partial \hat{y}} &= (\hat{y} - y) \\ &= (4/7 - 1) = -3/7 \end{aligned} \tag{3}$$

$$\begin{aligned} \hat{y} &= \text{Avg} \begin{bmatrix} h_1 & h_2 & \dots & h_7 \end{bmatrix} \\ \Rightarrow \frac{\partial \hat{y}}{\partial h} &= \begin{bmatrix} \frac{\partial \hat{y}}{\partial h_1} & \frac{\partial \hat{y}}{\partial h_2} & \cdots & \frac{\partial \hat{y}}{\partial h_7} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{7} & \frac{1}{7} & \cdots & \frac{1}{7} \end{bmatrix} \end{aligned} \tag{4}$$

$$\begin{aligned} h &= \text{Relu} \begin{bmatrix} a_1 & a_2 & \cdots & a_7 \end{bmatrix} \\ \Rightarrow \frac{\partial h}{\partial a} &= \begin{bmatrix} \frac{\partial h_1}{\partial a_1} & \frac{\partial h_1}{\partial a_2} & \cdots & \frac{\partial h_1}{\partial a_7} \\ \frac{\partial h_2}{\partial a_1} & \frac{\partial h_2}{\partial a_2} & \cdots & \frac{\partial h_2}{\partial a_7} \\ \vdots & \vdots & & \vdots \\ \frac{\partial h_7}{\partial a_1} & \frac{\partial h_7}{\partial a_2} & \cdots & \frac{\partial h_7}{\partial a_7} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & 0 \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \end{aligned} \tag{5}$$

Putting back in eq (2)

$$\begin{aligned}\frac{\partial L}{\partial a} &= \frac{-3}{7}\begin{bmatrix}1 & 0 & 0 & \cdots & 0\\ 0 & 0 & 0 & \cdots & 0\\ 0 & 0 & 1 & \cdots & 0\\ \vdots & \vdots & \vdots & & 0\\ 0 & 0 & 0 & \cdots & 1\end{bmatrix} \cdot \begin{bmatrix}\frac{1}{7} & \frac{1}{7} & \cdots & \frac{1}{7}\end{bmatrix}\\ &= \frac{-3}{7}\begin{bmatrix}\frac{1}{7} & 0 & \frac{1}{7} & \cdots & \frac{1}{7}\end{bmatrix}\\ &= \begin{bmatrix}\frac{-3}{49} & 0 & \frac{-3}{49} & \cdots & \frac{-3}{49}\end{bmatrix}\end{aligned}$$

Putting back in eq (1)

$$\begin{aligned}\frac{\partial L}{\partial K} &= \begin{bmatrix}1 & 0 & 1 & 0 & 1 & 0 & 1 & 0\end{bmatrix} * \begin{bmatrix}\frac{-3}{49} & 0 & \frac{-3}{49} & \cdots & \frac{-3}{49}\end{bmatrix}\\ &= \begin{bmatrix}\frac{-3}{7} & 0\end{bmatrix}\end{aligned}$$