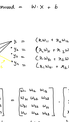
BP

Neural Net Implementation

note: the backprop math here etc. is from an implementation perspective. that why there is activation hard, activation logic handled. typically in implementation work checkboxes and lists, and other ctrls. makes it better understand. ← theoretically.

Forward = $wx + b$



activation hard
activation logic hard
activation_name = "softmax"
hidden_size = 3
output_size = 3
input_size = 3
weight_scale = 0.01
seed = 42

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

$$\underset{3\times1}{y} = \underset{3\times3}{w} \cdot \underset{3\times1}{x} + \underset{3\times1}{b}$$

$\therefore \; Y = wx + b$

backprop / loss:

– what you want: $\frac{\partial E}{\partial w} = \frac{\partial E}{\partial Y} \cdot \frac{\partial Y}{\partial w}$

for gradient descent — w.r.t. chain rule

$\frac{\partial Y}{\partial w} = \frac{\partial x^T \cdot w}{\partial w} = \frac{\partial x^T}{\partial w} \cdot w + x^T \cdot \frac{\partial w}{\partial w}$

$y = x \cdot w + b$ — $y' = x \cdot w + x \cdot w'$

– what you have: $\frac{\partial E}{\partial Y}$

$$\frac{\partial E}{\partial w} = \begin{bmatrix} \partial y_1/\partial w_{11} & \partial y_1/\partial w_{12} & \partial y_1/\partial w_{13} \\ \partial y_2/\partial w_{11} & \partial y_2/\partial w_{12} & \partial y_2/\partial w_{13} \\ \partial y_3/\partial w_{11} & \partial y_3/\partial w_{12} & \partial y_3/\partial w_{13} \end{bmatrix} \quad \therefore \quad \frac{\partial E}{\partial b} = \text{'' (same thing as b but with } \partial E/\partial b_i)$$

$$\underset{3\times3}{\frac{\partial E}{\partial w}} = \underset{3\times1}{\frac{\partial E}{\partial y}} \cdot \underset{1\times3}{x^T}$$

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial y_i} \cdot \frac{\partial y_i}{\partial w_{ij}}$$

$$\frac{\partial E}{\partial y_1} = (y-3)^2 \qquad \frac{\partial y_1}{\partial w_{11}} = \frac{\partial}{\partial w_{11}}(x_1w_{11} + x_2w_{12} + x_3w_{13} + b_1)$$

$$\frac{\partial E}{\partial y_1} = 2(y-3) \qquad = \frac{\partial}{\partial w_{11}} x_1w_{11} + \frac{\partial}{\partial w_{11}} x_2w_{12} + \frac{\partial}{\partial w_{11}} x_3w_{13} + \frac{\partial}{\partial w_{11}} b_1$$

$$= 0 + x_1 \frac{\partial}{\partial w_{11}} w_{11} + 0 + 0$$

$$= x_1$$

$$\therefore \frac{\partial E}{\partial w_{11}} = \frac{\partial E}{\partial y_1} x_1$$

or $= 2(y - y_1)x_1$

$$\therefore \frac{\partial E}{\partial w_{ji}} = \frac{\partial E}{\partial y_j} x_i \quad \text{generalized.}$$

or $= 2(y - y_j)x_i$

$$\therefore \frac{\partial E}{\partial w} = \underset{4\times3}{\begin{bmatrix} \partial E/\partial y_1 x_1 & \partial E/\partial y_1 x_2 & \partial E/\partial y_1 x_3 \\ \partial E/\partial y_2 x_1 & \partial E/\partial y_2 x_2 & \partial E/\partial y_2 x_3 \\ \partial E/\partial y_3 x_1 & \partial E/\partial y_3 x_2 & \partial E/\partial y_3 x_3 \\ \partial E/\partial y_4 x_1 & \partial E/\partial y_4 x_2 & \partial E/\partial y_4 x_3 \end{bmatrix}}$$

(from above)

$$= \underset{4\times1}{\begin{bmatrix} \partial E/\partial y_1 \\ \partial E/\partial y_2 \\ \partial E/\partial y_3 \\ \partial E/\partial y_4 \end{bmatrix}} \cdot \underset{1\times3}{[x_1 \; x_2 \; x_3]}$$

4×3, 4×1, 1×3, 4×3

unit transpose. input tensor.

remember this from before? its just $\partial E/\partial Y$

$$\therefore \boxed{\frac{\partial E}{\partial w} = \frac{\partial E}{\partial Y} \cdot x^T}$$

now, for bias:

$\therefore \frac{\partial E}{\partial b}$ (goal), $\frac{\partial E}{\partial E}$ (is calculable)

gets back up to you where the matrix form of bias. $\partial E/\partial b_i$ is just do not mess with $\partial E/\partial b_i$ instead $\longrightarrow \begin{bmatrix} \partial E/\partial b_1 \\ \partial E/\partial b_2 \end{bmatrix}$

$$\frac{\partial E}{\partial b_1} = \frac{\partial E}{\partial y_1} \cdot \frac{\partial y}{\partial b_1}$$

$$\frac{\partial E}{\partial y_1} = (y-3)^2 \qquad \frac{\partial}{\partial b_1}(w_1x_1 + w_2x_2 + w_3x_3 + b_1)$$

$$\frac{\partial E}{\partial y_1} = 2(y-3) \qquad = \frac{\partial}{\partial b_1}(w_1x_1) + \frac{\partial}{\partial b_1}(w_2x_2) + \frac{\partial}{\partial b_1}(w_3x_3) + \frac{\partial}{\partial b_1}(b_1)$$

$$= 0 + 0 + 0 + 1$$

$$= 1$$

$$\therefore \frac{\partial E}{\partial b_1} = \frac{\partial E}{\partial y_1} \cdot \frac{\partial y_1}{\partial b_1}$$

$$\therefore \frac{\partial E}{\partial b_1} = \frac{\partial E}{\partial y_1} \cdot 1$$

$$\therefore \frac{\partial E}{\partial b_1} = \frac{\partial E}{\partial y_1} \quad \text{or} \quad \frac{\partial E}{\partial b_1} = 2(y - y_1)$$

$\therefore$ generalizing this we get:

$$\frac{\partial E}{\partial b_i} = \frac{\partial E}{\partial y_i}$$

$$\therefore \boxed{\frac{\partial E}{\partial b} = \frac{\partial E}{\partial Y}}$$

now, since $\frac{\partial E}{\partial y} = \frac{\partial E}{\partial b}$

we missed that you let do for sum w/ no matrix!

$\therefore$ we were given $\frac{\partial E}{\partial Y}$ and we needed $\frac{\partial E}{\partial w}$ & $\frac{\partial E}{\partial b}$ for gradient desc. however, we also needed $\frac{\partial E}{\partial X}$ for chain rule in next layer going backwards.

$\frac{\partial E}{\partial X}$ (to calculate), $\frac{\partial E}{\partial Y}$ (given).

since we have $\frac{\partial E}{\partial X}$ a col vector X with each. item $\partial E/\partial x_i$ —

$$\frac{\partial E}{\partial X} = \begin{bmatrix} \partial E/\partial x_1 \\ \partial E/\partial x_2 \end{bmatrix}$$

$$\frac{\partial E}{\partial x_1} = \frac{\partial E}{\partial y_1} \cdot \frac{\partial y}{\partial x_1} + \frac{\partial E}{\partial y_2} \cdot \frac{\partial y_2}{\partial x_1} + \cdots + \frac{\partial E}{\partial y_3} \cdot \frac{\partial y_3}{\partial x_1} \quad \text{— (a)}$$

$\frac{\partial E}{\partial y_1} = 2(y-y_1)$ — same as each below.

this time, since all these cancel out to 0 w/ x_1 is in one of the y_i changing or $y_1 = w_{11}x_1 + w_{12}x_2 + w_{13}x_3 + b_1$, $y_2 = w_{21}x_1 + w_{22}x_2 + w_{23}x_3 + b_2$

so if we have x_1 in that element

$$\frac{\partial y_1}{\partial x_1} = w_{11}, \quad \frac{\partial y_2}{\partial x_1} = w_{21} \quad \text{etc.}$$

$$\therefore \frac{\partial y_j}{\partial x_i} = w_{ji}$$

generalizing this we get

$$\frac{\partial y_j}{\partial x_i} = w_{ji} \quad \text{— (1)}$$

$\therefore$ from previous eq (a) and (1) we get:

$$\frac{\partial E}{\partial x_i} = \frac{\partial E}{\partial y_1} w_{1i} + \frac{\partial E}{\partial y_2} w_{2i} \cdots \frac{\partial E}{\partial y_j} w_{ji}$$

recall what we wanted: $\frac{\partial E}{\partial X}$

$$\frac{\partial E}{\partial X} = \begin{bmatrix} \frac{\partial E}{\partial x_1} \\ \frac{\partial E}{\partial x_2} \\ \frac{\partial E}{\partial x_3} \\ \frac{\partial E}{\partial x_4} \end{bmatrix} = \begin{bmatrix} \frac{\partial E}{\partial y_1} w_{11} + \frac{\partial E}{\partial y_2} w_{21} + \frac{\partial E}{\partial y_3} w_{31} + \frac{\partial E}{\partial y_4} w_{41} \\ \frac{\partial E}{\partial y_1} w_{12} + \frac{\partial E}{\partial y_2} w_{22} + \frac{\partial E}{\partial y_3} w_{32} + \frac{\partial E}{\partial y_4} w_{42} \\ \vdots \\ \frac{\partial E}{\partial y_1} w_{14} + \frac{\partial E}{\partial y_2} w_{24} + \frac{\partial E}{\partial y_3} w_{34} + \frac{\partial E}{\partial y_4} w_{44} \end{bmatrix}$$

$$= \underset{3\times4}{\begin{bmatrix} w_{11} & w_{21} & w_{31} & w_{41} \\ w_{12} & w_{22} & w_{32} & w_{42} \\ w_{13} & w_{23} & w_{33} & w_{43} \end{bmatrix}} \cdot \underset{4\times1}{\begin{bmatrix} \partial E/\partial y_1 \\ \partial E/\partial y_2 \\ \partial E/\partial y_3 \\ \partial E/\partial y_4 \end{bmatrix}}$$

3×4, $r_A \times r_B$; 4×1, $r_B \times 1$

note that this is the transpose of w from using just

$$= w^T \cdot \partial E/\partial Y$$

$$\therefore \boxed{\partial E/\partial X = W \cdot \partial E/\partial Y}$$

we have all 3 steps!

$\partial E/\partial w$ ✓
$\partial E/\partial b$ ✓
$\partial E/\partial X$ ✓

Activation layer

$f(x) = Y$ (is activation fn. and $f_x(x)$)

∂ slope = Y.grad into it just replacing activation.

backprop

$$\frac{\partial E}{\partial Y} = \begin{bmatrix} \partial E/\partial y_1 \\ \partial E/\partial y_2 \\ \partial E/\partial y_3 \\ \partial E/\partial y_4 \end{bmatrix}$$

when moving back you're interesting the activation to give y you must return x. or $\partial E/\partial y \rightarrow \partial E/\partial x$

series: $y_1 = \sigma(x_1)$, $y_2 = f(x_2)$, $y_3 = f(x_3)$ & σ

$$\frac{\partial E}{\partial x_1} = \frac{\partial E}{\partial y_1} \cdot \frac{\partial y_1}{\partial x_1} + \frac{\partial E}{\partial y_2} \cdot \cancel{\frac{\partial y_2}{\partial x_1}} + \frac{\partial E}{\partial y_3} \cdot \cancel{\frac{\partial y_3}{\partial x_1}}$$

$$= \frac{\partial E}{\partial y_1} \cdot \frac{\partial y_1}{\partial x_1}$$

$$\frac{\partial}{\partial y_1}(y - y_1)^2 \qquad \frac{\partial}{x_1}(f(x))$$

$$= 2(y - y_1) \qquad = f'(x)$$

$\therefore$ generalizing above result:

$$\frac{\partial E}{\partial x_i} = \frac{\partial E}{\partial y_i} \cdot f'(x_i)$$

$$\therefore \boxed{\frac{\partial E}{\partial x} = \frac{\partial E}{\partial Y} \odot f'(x)}$$

activation — prime (self.input)

element-wise multiplication (np.multiply)

Tanh

$$\frac{\partial}{\partial x} \tanh_x = 1 - \tanh^2(x)$$

$$\frac{1 - e^{-2x}}{1 + e^{-2x}} \quad \frac{\sqrt{e^{-2x}}}{\sqrt{e^{-2x}}}$$

not force = 5

$\therefore$ tanh gets done

MSE

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \hat{Y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}$$

$$E = \frac{1}{n} \sum_i^n (y_i - \hat{y}_i)^2$$

we already assumed $\partial E/\partial y$ was given. MSE then computes it.

$$\frac{\partial E}{\partial y} = \frac{\partial}{\partial y} \frac{1}{n} ((y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots \cancel{(y_n - \hat{y}_n)^2})$$

$$= \frac{\partial}{\partial y} \frac{1}{n} (y - \hat{y})^2$$

$$\frac{\partial E}{\partial y} = \frac{2}{n}(y - \hat{y})$$

generalizing this:

$$\frac{\partial E}{\partial x} = \frac{2}{n}(x - \hat{x})$$

$$\therefore \frac{\partial E}{\partial Y} = \frac{2}{n}(Y - \hat{Y})$$