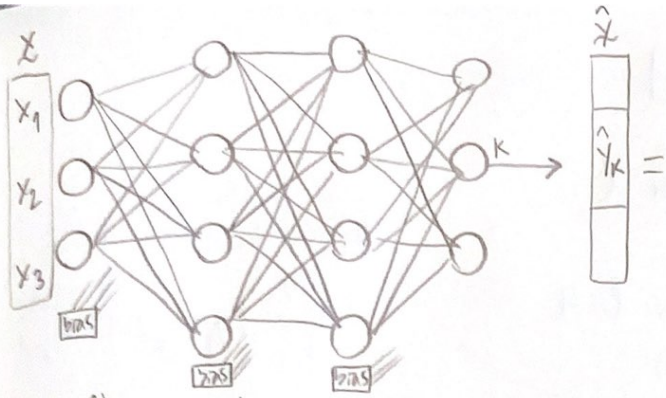

$\underset{\sim}{a}^{(1)}$ $\underset{\sim}{a}^{(2)}$ $\underset{\sim}{a}^{(3)}$ $\underset{\sim}{a}^{(4)}$

INPUT — HIDDEN — OUTPUT

layer 1 layer 2 layer 3 layer 4

X: input

a: output after activation function

z: output before activation function

f: activation function, sigmoid

$\underset{\sim}{W}^{(l)}$: weight matrix from layer l to layer $l+1$ of size $s_{l+1} \times (s_l + 1)$

(s_{l+1}: units in layer $l+1$; s_l: units in layer l; +1: bias)

— FORWARD PASS —

$\underset{\sim}{a}^{(1)} = \underset{\sim}{x}$

↳ append bias: $\underset{\sim}{a}^{(1)} \leftarrow [a_0^{(1)} = 1, \underset{\sim}{a}^{(1)T}]^T = [1 \; x_1 \; x_2 \; x_3]^T$ $(3+1)\times 1$, 4

$\underset{\sim}{z}^{(2)} = \underset{\sim}{W}^{(1)} \cdot \underset{\sim}{a}^{(1)}$ $(4\times 4)\times(4\times 1) = 4\times 1$

$\underset{\sim}{a}^{(2)} = f(\underset{\sim}{z}^{(2)})$

↳ append bias: $\underset{\sim}{a}^{(2)} \leftarrow [a_0^{(2)} = 1, \underset{\sim}{a}^{(2)T}]^T$ $(4+1)\times 1$, 5

$\underset{\sim}{z}^{(3)} = \underset{\sim}{W}^{(2)} \cdot \underset{\sim}{a}^{(2)}$ $(4\times 5)\times(5\times 1) = 4\times 1$

$\underset{\sim}{a}^{(3)} = f(\underset{\sim}{z}^{(3)})$

↳ append bias: $\underset{\sim}{a}^{(3)} \leftarrow [a_0^{(3)} = 1, \underset{\sim}{a}^{(3)T}]^T$ $(4+1)\times 1$, 5

$\underset{\sim}{z}^{(4)} = \underset{\sim}{W}^{(3)} \cdot \underset{\sim}{a}^{(3)}$ $(3\times 5)\times(5\times 1) = 3\times 1$

$\underset{\sim}{a}^{(4)} = f(\underset{\sim}{z}^{(4)}) = h_W(\underset{\sim}{x})$: 3×1

↳ model / neural network.

BACKWARD PASS

$\delta_j^{(l)}$: error of node/unit j in layer l

- For each output unit j in last layer $l=4$

$$\delta_j^{(4)} = (y_j - \hat{y}_j) = (y_j - a_j^{(4)})$$

in vectorized it's multiplied by $f'()$:

$$\delta_j^{(4)} = (y_j - \hat{y}_j) \cdot f'(z_j^{(3)})$$

↳ assemble: $\underset{\sim}{\delta}^{(4)} = [\delta_1^{(4)}, \dots, \delta_j^{(4)}, \dots]$ (num classes)

note: for the sigmoid: $f' = f \cdot (1-f)$

- Propagate errors to all units of all layers

extend because of bias, so that dims match, but then it's removed!

$$\underset{\sim}{\delta}^{(3)} = \left(\underbrace{(\underset{\sim}{W}^{(3)})^T}_{5\times 3} \cdot \underbrace{\underset{\sim}{\delta}^{(4)}}_{3\times 1} \right) .* \left[1, \underbrace{f'(\underset{\sim}{z}^{(3)})}_{1\times 4} \right]^T$$

$5\times 1 \rightarrow$ multiply 1 by 1

$$= [\underbrace{\delta_0^{(3)}}_{\text{bias}}, \delta_1^{(3)}, \dots]^T$$: each unit in layer 3 has an error.

$\delta^{(3)} \leftarrow \delta^{(3)}[1:]$ remove bias component : 4×1

$$\underset{\sim}{\delta}^{(2)} = \left(\underbrace{(\underset{\sim}{W}^{(2)})^T}_{5\times 4} \cdot \underbrace{\underset{\sim}{\delta}^{(3)}}_{4\times 1} \right) .* \left[1, \underbrace{f'(z^{(2)})^T}_{4\times 1} \right]$$

5×1

initialized as $\Delta W = 0$

$\delta^{(2)} \leftarrow \delta^{(2)}[1:]$ remove bias component: 4×1

- Weight updates: For each sample:

$\hat{y} = \text{forward}(x) \rightarrow a^{(1)}, a^{(2)}, a^{(3)}, a^{(4)} = \hat{y}$

$\text{backward}(\hat{y}) \rightarrow \delta^{(4)}, \delta^{(3)}, \delta^{(2)}$

$$\Delta \underset{\sim}{W}^{(l)} = \Delta \underset{\sim}{W}^{(l)} + \underbrace{\underset{\sim}{\delta}^{(l+1)}}_{s_{l+1}\times 1} \cdot \underbrace{a^{(l)T}}_{1\times(s_l+1)}$$

Then, after each epoch

$$\underset{\sim}{W}^{(l)} = \underbrace{\underset{\sim}{W}^{(l)}}_{\text{old}} + \underbrace{\frac{\alpha}{m} \Delta \underset{\sim}{W}^{(l)}}_{\text{step}}$$