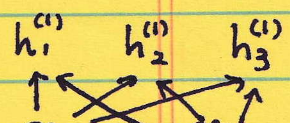
Multilayer Perceptron

$$y = \phi(W^T x + b) = \phi(w_1 x_1 + w_2 x_2 + w_3 x_3 + b)$$

as observed from the simple neural network on slides 4

Compare with logistic regression = $y = \sigma(\underbrace{W^T x + b}_{\text{logit}})$ ↑ activation

For a fully connected layer, assuming there are N input units and M output units



⇒ there are $N = 2$ inputs and $M = 3$ outputs.

⇒ W's dimensionality = $M \times N$

Activation function = $y = \phi(z)$, in which z = logit = the raw value or stimuli

$h^{(1)} = f^{(1)}(x) = \phi(W^{(1)} x + b^{(1)})$ ⇒ first layer

$h^{(2)} = f^{(2)}(h^{(1)}) = \phi(W^{(2)} h^{(1)} + b^{(2)})$ ⇒ second layer

$y = f^{(L)}(h^{(L-1)}) \Leftrightarrow y = f^{(L)} \circ \cdots \circ f^{(1)}(x)$

For regression: $y = f^L(h^{L-1}) = (W^{(L)})^T h^{(L-1)} + b^L$

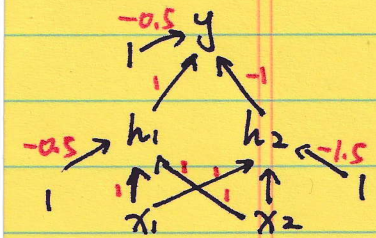
For classification = $y = f^L(h^{L-1}) = \sigma((W^L)^T h^{(L-1)} + b^L)$

↳ for example, activate the logit into a value between 0 – 1 (logistic activation)

Linear network = $y = W^3 W^2 W^1 x$

$= f^L(h^{L-1}) = f^3(h^2) = f^3(f^2(h_1)) = f^3 \circ f^2 \circ f^1(x)$

XOR



h_1 computes $-0.5 + x_1 + x_2 > 0$ (x_1 OR x_2)

h_2 computes $x_1 + x_2 - 1.5 > 0$ (x_1 AND x_2)

y computes $-0.5 + h_1 - h_2 > 0$ (h_1 AND $\neg h_2$)

Backpropagation

$$z = wx + b$$

$$y = \sigma(z)$$

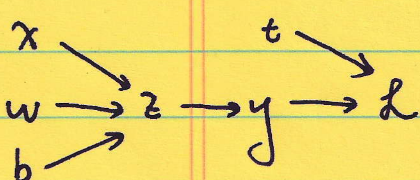
$$\mathcal{L} = \frac{1}{2}(y-t)^2$$

$$\frac{d}{dt} f(x(t)) = \frac{df}{dx} \cdot \frac{dx}{dt}$$

$$\frac{d\mathcal{L}}{dy} = y - t \qquad \frac{d\mathcal{L}}{dz} = \frac{d\mathcal{L}}{dy} \cdot \frac{dy}{dz} = \frac{d\mathcal{L}}{dy} \cdot \sigma'(z)$$

$$\frac{d\mathcal{L}}{dw} = \frac{\partial \mathcal{L}}{\partial z} \cdot \frac{\partial z}{\partial w} = \frac{d\mathcal{L}}{dz} \cdot x \qquad \frac{\partial \mathcal{L}}{\partial b} = \frac{\partial \mathcal{L}}{\partial z} \cdot \frac{\partial z}{\partial b} = \frac{\partial \mathcal{L}}{\partial z}$$

$\frac{\partial \mathcal{L}}{\partial y}$ = the error signal at y, denoted as $\bar{y}$



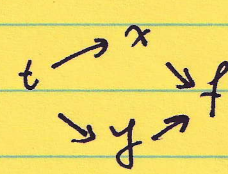
$$\bar{z} = \bar{y} \cdot \sigma'(z)$$

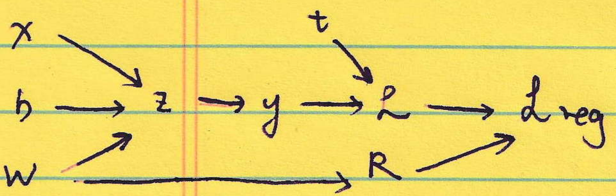
$$\bar{w} = \bar{z} \cdot x \qquad \bar{b} = \bar{z}$$

Multivariate chain rule: $\frac{d}{dt} f(x(t), y(t)) = \frac{df}{dx} \cdot \frac{dx}{dt} + \frac{df}{dy} \cdot \frac{dy}{dt}$

E.g. $f(x,y) = y + e^{xy}$, $x(t) = \cos t$, $y(t) = t^2$.

$$\frac{df}{dt} = y \cdot e^{xy} \cdot (-\sin t) + (1 + x e^{xy}) \cdot 2t$$





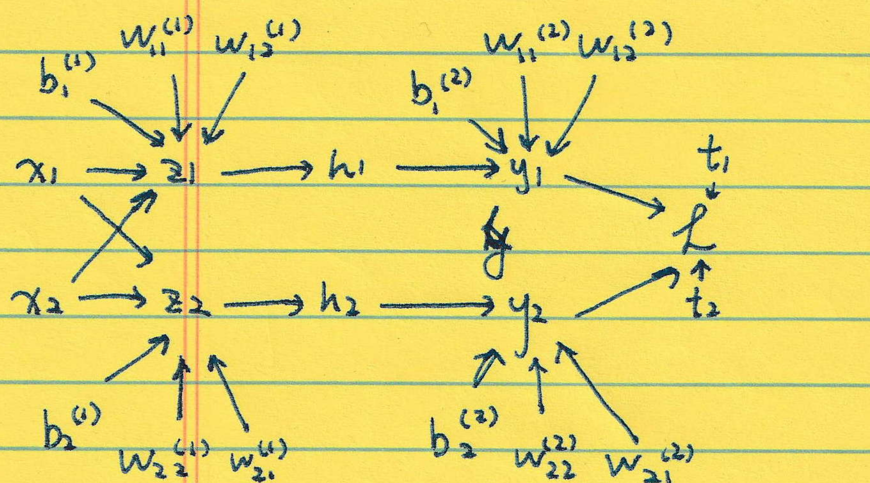
Backward pass:

$$\overline{\mathcal{L}_{reg}} = \frac{\partial \mathcal{L}_{reg}}{\partial \mathcal{L}_{reg}} = 1 \qquad \overline{R} = \overline{\mathcal{L}_{reg}} \cdot \frac{\partial \mathcal{L}_{reg}}{\partial R} = \overline{\mathcal{L}_{reg}} \cdot \lambda$$

$$\overline{\mathcal{L}} = \overline{\mathcal{L}_{reg}} \cdot \frac{\partial \mathcal{L}_{reg}}{\partial \mathcal{L}} = 1 \cdot \overline{\mathcal{L}_{reg}} = \overline{\mathcal{L}_{reg}}$$

$$\overline{y} = \overline{\mathcal{L}} \cdot \frac{\partial \mathcal{L}}{\partial y} = \overline{\mathcal{L}}(y-t) \qquad \overline{z} = \overline{y} \cdot \frac{\partial y}{\partial z} = \overline{y} \cdot \sigma'(z)$$

$$\overline{w} = \frac{\partial z}{\partial w} \cdot \overline{z} + \frac{\partial R}{\partial w} \cdot \overline{R} = \overline{z} \cdot x + \overline{R} \cdot w \qquad \overline{b} = \overline{z} \cdot \frac{\partial z}{\partial b} = \overline{z}$$



$$\overline{\mathcal{L}} = \frac{\partial \mathcal{L}}{\partial \mathcal{L}} = 1$$

$$\overline{y_k} = \overline{\mathcal{L}}(y_k - t_k)$$

$$\overline{w_{ki}^{(2)}} = \overline{y_k} \cdot h_i$$

$$\overline{b_k^{(2)}} = \overline{y_k}$$

$$\overline{h_i} = \sum_k \overline{y_k}\, w_{ki}^{(2)}$$

$$\overline{z_i} = \overline{h_i}\, \sigma'(z_i) \qquad \overline{w_{ij}^{(1)}} = \overline{z_i}\, x_j \qquad \overline{b_i^{(1)}} = \overline{z_i}$$

Remark: Start from the (regularised) loss, and propagate backwards the error signals.

You shouldn't update the values for x_i or t_i, as they are your input samples and the ground-truths.