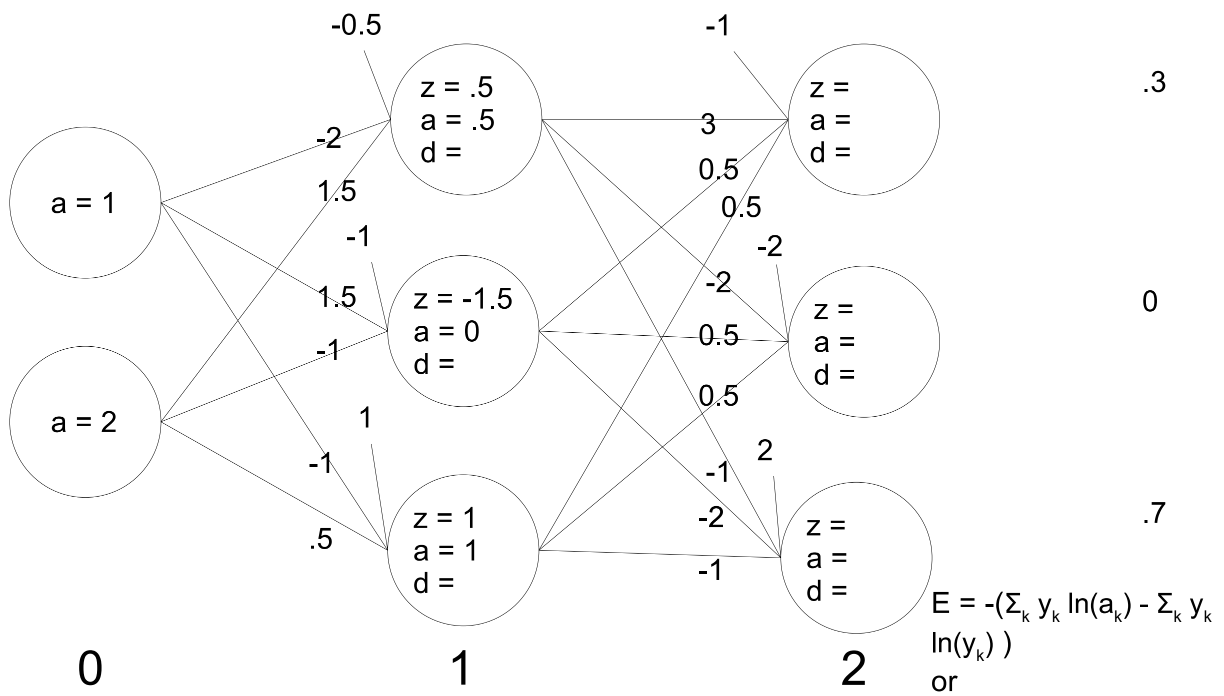
Notation:

n^L = number of neurons in layer L. n^0 is number of inputs
w_{ij}^L = weight into layer L, from neuron j to neuron i
W^L = matrix of weights from L-1 to L, dimension 2
b_i^L = bias for neuron i of layer L
B^L = bias vector for layer L
a_i^L = output i of layer L.
A^L = vector output of layer L. A^0 is input vector
z_i^L = weighted sum of neuron i of layer L incl bias.
σ = activation function. $a_i^L = \sigma(z_i^L)$, or sometimes, $a_i^L = \sigma(Z^L)$ e.g. for softmax
E = Loss function $E(A^N, Y)$, where N is final layer. (E for "error" to disambiguate from L)
d_i^L = partial of E wrt z_i^L

$\sigma = \text{relu} = \max(0, z)$

$\sigma = S(z) = 1/(1+e^{-z})$ or
$\sigma = SM(z_i) = e^{z_i}/\Sigma_k e^{z_k}$

Labels (y)



$E = -(\Sigma_k y_k \ln(a_k) - \Sigma_k y_k \ln(y_k))$
or
$E = \Sigma_k(a_k - y_k)^2 / n$

Chain Rule Partials for BackProp

For all w_{ij}:

$\partial E/\partial w_{ij}^x = a_j^{x-1}\, \partial E/\partial z_i^x$

$\partial E/\partial z_i^1 = \partial a_i^1/\partial z_i^1 * \Sigma_k(\partial z_k^2/\partial a_i^1 * \partial E/\partial z_k^2)$
$= \partial a_i^1/\partial z_i^1 * \Sigma_k(w_{ki}^2 * \partial E/\partial z_k^2)$

$\partial E/\partial z_j = \partial a_j/\partial z_j * \partial E/\partial a_j$

Partials for Loss Functions

For crossentropy: $\partial E/\partial a_j = -y_j / a_j$
For MSE: $\partial E/\partial a_j = 2(a_j - y_j)$

Partials for Activations:

For relu: 1 if a > 0, 0 otherwise

For S(z): $\partial a/\partial z = e^{-z}/(1+e^{-z})^2$

For SM(Z) it's more complex. $\partial a/\partial z$ requires a matrix of $\partial a_j/\partial z_i$ for all j, i:

for i = j: $\partial a_j/\partial z_i = a_j(1-a_i)$
for i<>j: $\partial a_j/\partial z_i = a_j(0-a_i)$
or just $\partial a_j/\partial z_i = a_j(\delta_{ij} - a_i)$

$$\begin{vmatrix} \partial E/\partial z_0 \\ \partial E/\partial z_1 \\ \partial E/\partial z_2 \end{vmatrix} = \begin{vmatrix} \partial a_0/\partial z_0 & \partial a_1/\partial z_0 & \partial a_2/\partial z_0 \\ \partial a_0/\partial z_1 & \partial a_1/\partial z_1 & \partial a_2/\partial z_1 \\ \partial a_0/\partial z_2 & \partial a_1/\partial z_2 & \partial a_2/\partial z_2 \end{vmatrix} \begin{vmatrix} \partial E/\partial a_0 \\ \partial E/\partial a_1 \\ \partial E/\partial a_2 \end{vmatrix}$$