Recitation 7: Backpropagation Examples

LOrtiz (Orig. by KKoile)

An efficient method of implementing gradient descent for neural networks

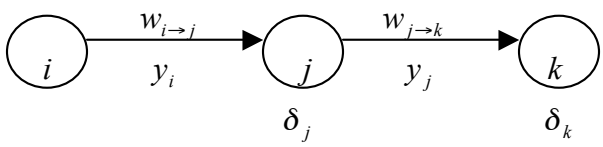
$w_{i\to j} = w_{i\to j} - r\delta_j y_i$ Descent Rule

$$\delta_j = \frac{dE_j}{dz_j} = \frac{ds(z_j)}{dz_j}$$

$$\delta_j = \frac{ds(z_j)}{dz_j}\sum_k \delta_k w_{j\to k}$$

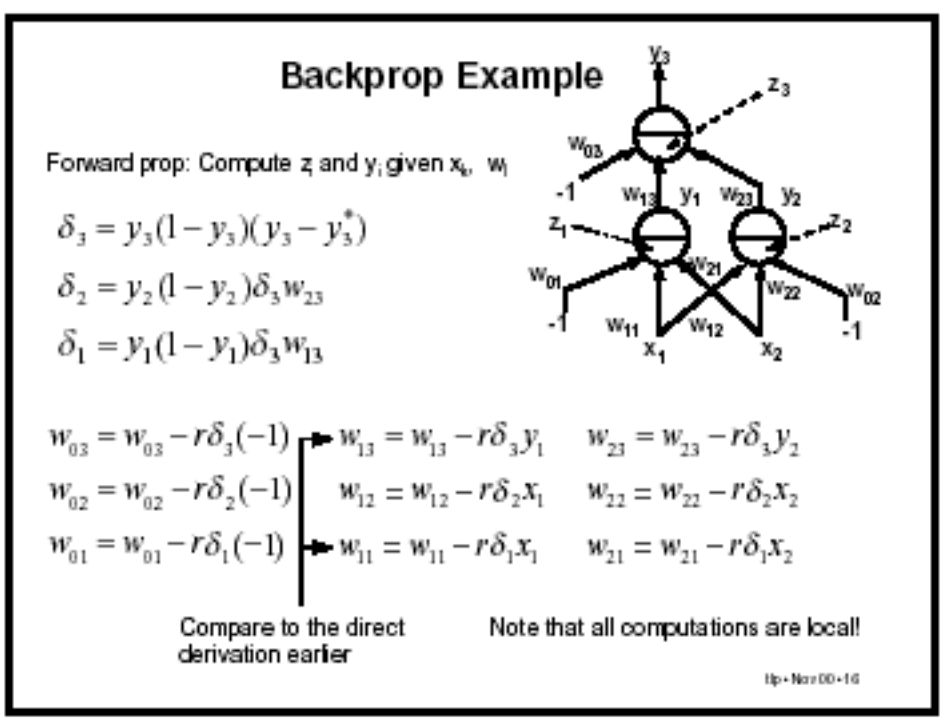
Backprop rule ($s(z_j)$ is the sigmoid function)

1. Initialize weights to small random values
2. Choose a random sample input feature vector
3. Compute total input (z_j) and output (y_j) for each unit (forward prop)
4. Compute δ_n for output layer. Assuming sigmoid neural net and standard error function: $\delta_n = \frac{ds(z_n)}{dz_n}(y_n - y_n^*) = y_n(1-y_n)(y_n - y_n^*)$
5. Compute δ_j for preceding layer by backprop rule (repeat for all layers)
6. Compute weight change by descent rule (repeat for all weights)



y_i is x_i for input layer

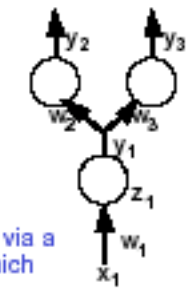
From the online tutor:



Backpropagation

$$\delta_j = \frac{ds(z_j)}{dz_j}\sum_k \delta_k w_{j\to k}$$ Backprop rule

Why the sum over k? Consider the simplified net:



$$E = \frac{1}{2}\left[(y_2^* - y_2)^2 + (y_3^* - y_3)^2\right]$$

$$y_2 = s(w_2 s(w_1 x_1))$$

$$y_3 = s(w_3 s(w_1 x_1))$$

A change in w_1 affects the error via a change in y_1, which affects y_2 and y_3

$$\frac{\partial E}{\partial w_1} = -(y_2^* - y_2)\frac{\partial y_2}{\partial w_1} - (y_3^* - y_3)\frac{\partial y_3}{\partial w_1}$$

$$\frac{\partial y_2}{\partial w_1} = \frac{\partial s(z_2)}{\partial z_2} w_2 \frac{\partial s(z_1)}{\partial z_1} x_1$$

$$\frac{\partial y_3}{\partial w_1} = \frac{\partial s(z_3)}{\partial z_3} w_3 \frac{\partial s(z_1)}{\partial z_1} x_1$$

$$\frac{\partial E}{\partial w_1} = \overbrace{\frac{\partial s(z_1)}{\partial z_1}\left[\underbrace{-(y_2^* - y_2)\frac{\partial s(z_2)}{\partial z_2} w_2}_{\delta_2} \underbrace{-(y_3^* - y_3)\frac{\partial s(z_3)}{\partial z_3} w_3}_{\delta_3}\right]}^{\delta_1} x_1$$

$$\frac{\partial E}{\partial w_1} = s(z_1)(1 - s(z_1))[(y_2 - y_2^*)(s(z_2)(1 - s(z_2))w_2 + (y_3 - y_3^*)(s(z_3)(1 - s(z_3))w_3]x_1$$

$$\frac{\partial E}{\partial w_1} = y_1(1 - y_1)[(y_2 - y_2^*)y_2(1 - y_2)w_2 + (y_3 - y_3^*)y_3(1 - y_3)w_3]x_1$$

To compute w_1:

1. Substitute above expression into: $w_1 = w_1 - r\frac{\partial E}{\partial w_1}$

or

2. Compute δ_1 and substitute into: $w_1 = w_1 - r\delta_1 x_1$

From the expression for $\frac{\partial E}{\partial w_1}$, we can derive the expression for δ_1:

$$\delta_1 = y_1(1 - y_1)[(y_2 - y_2^*)y_2(1 - y_2)w_2 + (y_3 - y_3^*)y_3(1 - y_3)w_3]$$

$$\delta_1 = y_1(1 - y_1)(\delta_2 w_2 + \delta_3 w_3)$$ (Note: This is a restatement of the backprop rule.)