An illustration for Error Backpropagation Method

Consider the case of a two-hidden layer feedforward network.

Write the output Y as

$$Y = f(V_1, \cdots, V_r, \mathbf{w})$$

with

$$V_j = g(U_1, \cdots, U_q, \theta_j)$$

and

$$U_j = h(x_1, \cdots, x_p, \psi_j)$$

where U_j's are the first hidden layer units and V_j's are second hidden layer units. The connection weights are $\mathbf{w}$; $\theta_j, j = 1, \cdots, r$; and ψ_j's. Call $\mathbf{w}$ the outer-layer weight.

Take partial derivative of Y w.r.t. $\mathbf{w}$ first, because this part is the easiest. It does not involve g and h.

Now, for each j, take partial derivative of Y w.r.t. to θ_j. The result, by chain rule, is the product of $\frac{\partial f}{\partial V_j}$ and $\frac{\partial g}{\partial \theta_j}$. It does not involve h.

So the backpropagation updates the weight $\mathbf{w}$ for the outerlayer first by adding an amount of $-\delta(Y - \hat{Y})\frac{\partial f}{\partial \mathbf{w}}$, where δ is the step size (or learning rate) to be carefully chosen.

Then move **backward** (upstream) to the next layer and update the connection weight θ_j. First, compute the *propagated error* $e_j = (y - \hat{y})\frac{\partial f}{\partial V_j}$. This propagated error should be saved because it will be needed in updating the weights for all the upstream units leading to unit V_j. Now pretend V_j as the output (Y), treat θ_j as the outer-layer weight $\mathbf{w}$, and compute $\frac{\partial V_j}{\partial \theta_j}$. Now use the propagated error computed earlier as $V_j - \hat{V}_j$. This gives an update of θ_j by increasing an amount of $-\delta e_j \frac{\partial V_j}{\partial \theta_j}$, which is exactly the same as one would have get by the chain rule.

It is now easy to generalize the procedure to obtain weight update for ψ_j. Simply treat U_j as the output node and ψ_j as the outer-layer weight. Then the error $U_j - \hat{U}_j$ should be replaced by the error propagated from all the downstream units. Thus the adjustment would be equal to $-\delta$ times $\sum_{i=1}^{r} e_i \frac{\partial V_i}{\partial U_j}$ times $\frac{\partial U_j}{\partial \psi_j}$.