The SoftMax

Forward propagation

Consider a network for selecting between k categories. It is designed such that at the last step there are k computed real values $V_1, \ldots, V_k$. The SoftMax converts these values into probabilities as follows:

$$q_i = e^{V_i}, \quad i = 1, \ldots, k \tag{1}$$

$$Z = \sum_{i=1}^{k} q_i \tag{2}$$

$$p_i = \frac{q_i}{Z} = \frac{e^{V_i}}{\sum_{j=1}^{k} e^{V_j}}, \quad i = 1, \ldots, k \tag{3}$$

Error back propagation

Suppose the true category is j. The SoftMax error is defined as follows:

$$E = E_j = -\ln p_j = \ln \frac{1}{p_j}$$

Observe that the error is always nonnegative, in the range 0 to ∞. Simple calculation shows:

$$\frac{\partial p_j}{\partial V_i} = p_j(1_{ij} - p_i), \qquad \text{where } \; 1_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

Therefore:

$$\frac{\partial E_j}{\partial V_i} = -\frac{1}{p_j}\frac{\partial p_j}{\partial V_i} = p_i - 1_{ij}$$

We now assume that the top layer is linear: $V_i = g(h_i) = h_i$. Then:

$$\frac{\partial E}{\partial h_i} = \frac{\partial E_j}{\partial h_i} = \frac{\partial E_j}{\partial V_i} = p_i - 1_{ij}$$

From the BP proof theorem:

$$\delta_i = -\frac{1}{2}\frac{\partial E}{\partial h_i} = \frac{1}{2}(1_{ij} - p_i)$$

The rest of the BP algorithm is the same as before.