backpropagation

$$y^*(t) = f[x(t)] + \epsilon$$

log likelihood:

$$\log L = \sum_t -\frac{1}{2}||y^*(t) - y(t)||^2$$

For output unit:

$$\frac{\partial \log L(n)}{\partial w_{ij}(n)} = \sum_t \frac{\partial \log L(t)}{\partial y_i(t)} \frac{\partial y_i(t)}{\partial s_i(t)} \frac{\partial s_i(t)}{\partial w_{ij}(n)}$$

For hidden unit:

$$\frac{\partial \log L(n)}{\partial w_{ij}(n)} = \sum_t \frac{\partial \log L(t)}{\partial h_i(t)} \frac{\partial h_i(t)}{\partial s_i(t)} \frac{\partial s_i(t)}{\partial w_{ij}(n)}$$

For output units:

$$\frac{\partial L(t)}{\partial y_i(t)} = y_i^*(t) - y_i(t)$$

For linear output units:

$$y_i(t) = s_i(t) = \sum_j w_{ij}(n) h_j(t)$$

$$\frac{\partial y_i(t}{\partial s_i(t)} = 1$$

$$\frac{\partial s_i(t)}{\partial w_{ij}(n)} = h_j(t)$$

$$w_{ij}(n+1) = w_{ij}(n) - \alpha \frac{\partial \log L(n)}{\partial w_{ij}(n)}$$

For hidden unit:

$$\frac{\partial \log L(t)}{\partial w_{ij}(t)} = \sum_k \frac{\partial \log L(t)}{\partial s_k(t)} w_{ki}(n)$$

sigmoid

- based on http://www.ics.uci.edu/˜pjsadows/backpropderivation.pdf

cross entropy for single sample is

$$E = -\sum_k (y_k^* \log(o_k) + (1 - y_k^*) \log(1 - o_k))$$

Activation function is logistic:

$$o_i = \frac{1}{1 + \exp(-s_i)}$$

where s_i is sum at a node:

$$s_i = \sum_j (w_{ji} o_j)$$

we want:

$$\frac{\partial E}{\partial w_{ji}}$$

$$= \frac{\partial E}{\partial o_i}\frac{\partial o_i}{\partial s_i}\frac{\partial s_i}{\partial w_{ji}}$$

$$\frac{\partial E}{o_i} = -\frac{y_i^*}{o_i} + \frac{1-y_i^*}{1-o_i}$$

$$= -\frac{y_i^* - y_i^* o_i - o_i + y_i^* o_i}{o_i(1-o_i)}$$

$$= \frac{o_i - y_i^*}{o_i(1-o_i)}$$

$$\frac{\partial o_i}{\partial s_i} = o_i(1-o_i)$$

$$\frac{\partial E}{\partial s_i} = o_i - y_i^*$$

$$\frac{\partial E}{\partial s_i^{l-1}} = \sum_k \frac{\partial E}{\partial s_k^l}\frac{\partial s_k^l}{\partial o_i^{l-1}}\frac{\partial o_i^{l-1}}{\partial s_i^{l-1}}$$

$$\frac{\partial E}{\partial s_i^{l-2}} = \sum_k \sum_j \frac{\partial E}{\partial s_k^l}\frac{\partial s_k^l}{\partial o_j^{l-1}}\frac{\partial o_j^{l-1}}{\partial s_j^{l-1}}\frac{\partial s_j^{l-1}}{\partial o_i^{l-2}}\frac{\partial o_i^{l-2}}{\partial s_i^{l-2}}$$

$$= \sum_k \sum_j \frac{\partial E}{\partial s_k^l} w_{kj}^l \frac{\partial o_j^{l-1}}{\partial s_j^{l-1}} w_{ji}^{l-1} \frac{\partial o_i^{l-2}}{\partial s_i^{l-2}}$$

classification

class has multinomial distribution:

$$p(y^*(t)|x(t)) = \prod_{k=1}^{K} y_k(t)^{y_k^*(t)}$$

log likelihood:

$$\log L = \sum_t \sum_k y_k^*(t) \log y_k(t)$$

output units use softmax activation:

$$y_i(t) = \frac{\exp[s_i(t)]}{\sum_k \exp[s_k(t)]}$$

$$\frac{\partial \log L(n)}{\partial w_{ij}(n)} = \sum_t \frac{\partial \log L(t)}{\partial s_i(t)}\frac{\partial s_i(t)}{\partial w_{ij}(n)}$$

$$\frac{\partial \log L(t)}{\partial s_i(t)} = \sum_k \frac{\partial \log L(t)}{\partial y_k(t)}\frac{\partial y_k(t)}{\partial s_i(t)}$$

$$= \sum_k \frac{y_k^*(t)}{y_k(t)} y_k(t)(\delta_{ik} - y_i(t))$$

$$\frac{\partial s_i(t)}{\partial w_{ij}(n)} = h_j(t)$$

Softmax:

$$y_i = \frac{\exp s_{n,i}}{\sum_k \exp s_{n,k}}$$

$$\frac{\exp(s_{n,i}) / \exp(\max_j s_{n,j})}{\sum_k \exp(s_{n,k}) / \exp(\max_j s_{n,j})}$$

$$= \frac{\exp(s_{n,i} - \max_j(s_{n,j}))}{\sum_k \exp(s_{n,k} - \max_j(s_{n,j}))}$$

Cross entropy for softmax (?):

$$= -\sum_n \sum_k y^*_{n,k} \log y_{n,k}$$

where $y^*_{n,k}$ should be 0, except for $y^*_{n,l_n} = 1$, where l_n is the label of instance n

sigmoid, mse (I'm trying myself, might not be correct...)

$$L = \sum_k \frac{1}{2}(y^*_k - o_k)^2$$

$$o_i = \frac{1}{1 + \exp(-s_i)}$$

$$s_i = \sum_k w_{ki} o_k$$

$$\frac{\partial L}{\partial w_{ji}} = \frac{\partial L}{\partial o_i} \frac{\partial o_i}{\partial s_i} \frac{\partial s_i}{\partial w_{ji}}$$

$$\frac{\partial L}{\partial o_i} = 2\frac{1}{2}(y^*_i - o_i)(-1)$$

$$= (o_i - y^*_i)$$

$$\frac{\partial o_i}{\partial s_i} = o_i(1 - o_i)$$

$$\frac{\partial s_i}{\partial w_{ji}} = o_j$$

$$\frac{\partial L}{\partial w_{ji}} = (o_k - y^*_k) o_i (1 - o_i) o_j$$

We want to reduce L slightly.
- therefore, we should modify w_{ji} slightly:

$$w_{ji}(n+1) = w_{ji}(n) - \alpha \frac{\partial L}{\partial w_{ji}}$$

$$= w_{ji}(n) - \alpha(o_i - y^*_i) o_i (1 - o_i) o_j$$

$$= w_{ji}(n) - \alpha \mathrm{error}(o_i, y^*_i) \mathrm{derivative}(o_i) \mathrm{output}^{l-1}(o^{l-1}_j)$$

$$\frac{\partial L}{\partial w_{kj}^{l-1}} = \sum_i \frac{\partial L}{\partial o_i^l}\frac{\partial o_i^l}{\partial s_i^l}\frac{\partial s_i^l}{o_j^{l-1}}\frac{\partial o_j^{l-1}}{\partial s_j^{l-1}}\frac{\partial s_j^{l-1}}{\partial w_{kj}^{l-1}}$$

$$s_i = \sum_j o_j^{l-1} w_{ji}^l$$

$$\frac{\partial s_i^l}{o_j^{l-1}} = w_{ji}^l$$

So,

$$\frac{\partial L}{\partial w_{kj}^{l-1}} = \sum_i (o_i^l - y_i^*) o_i^l (1 - o_i^l) w_{ji}^l o_j^{l-1}(1 - o_j^{l-1}) o_k^{l-2}$$

$$= o_j^{l-1}(1 - o_j^{l-1}) o_k^{l-2} \sum_i (o_i^l - y_i^*) o_i^l (1 - o_i^l) w_{ji}^l$$

$$= (\sum_i \text{error}^l \text{deriv}(o_i^l) w_{ji}^l) \text{deriv}(o_j^{l-1}) \text{output}(o_k^{l-2})$$

$$= \text{apparent error}_j^{l-1}(o_i^l, w_{ji}^l, y_i^*) \text{deriv}^{l-1}(o_j^{l-1}) \text{output}^{l-2}(o_k^{l-2})$$

where

$$\text{apparent error}_j^{l-1}(o_i^l, w_{ji}^l, y_i^*) = \sum_i (o_i^l - y_i^*) o_i^l (1 - o_i^l) w_{ji}^l$$

$$= \sum_i (\text{error}_i^l(o_i^l, y_i^*) \text{deriv}_i^l(o_i^l) w_{ji}^l)$$

$$= \sum_i (\frac{\partial L}{\partial o_i^l}\frac{\partial o_i^l}{\partial s_i^l} w_{ji}^l)$$

$$= \sum_i (\frac{\partial L}{\partial o_i^l}\frac{\partial o_i^l}{\partial s_i^l}\frac{\partial s_i^l}{\partial o_{ji}^{l-1}})$$

$$= \sum_i \frac{\partial L}{\partial o_{ji}^{l-1}}$$

$$= \sum_i \text{apparent error}_i^l \frac{\partial o_i^l}{\partial o_{ji}^{l-1}}$$

$$= \sum_i \text{apparent error}_i^l \frac{\partial o_i^l}{\partial s_i^l}\frac{\partial s_i^l}{\partial o_{ji}^{l-1}}$$

$$= \sum_i \text{apparent error}_i^l \frac{\partial o_i^l}{\partial s_i^l} w_{ji}^l$$

eg

$$\frac{\partial o_i^l}{\partial s_i^l}\frac{\partial s_i^l}{\partial o_{ji}^{l-1}} = o_i^l(1 - o_i^l) w_{ji}^l$$

or, for tanh

$$\frac{\partial o_i^l}{\partial s_i^l}\frac{\partial s_i^l}{\partial o_{ji}^{l-1}} = (1 - (o_i^l)^2) w_{ji}^l$$

Error functions

Cross entropy:

$$L = -\sum_k (y_k^* \log(o_k) + (1 - y_k^*) \log(1 - o_k))$$

$$\frac{\partial L}{\partial o_i} = -\frac{y_i^*}{o_i} - \frac{(1 - y_i^*)(-1)}{1 - o_i}$$

$$= -\frac{y_i^*}{o_i} + \frac{1 - y_i^*}{1 - o_i}$$

$$= \frac{-y_i^* + y_i^* o_i + o_i - y_i^* o_i}{o_i(1 - o_i}$$

$$= \frac{o_i - y_i^*}{o_i(1 - o_i)}$$

Cross-entropy, multiclass:

$$L = -\sum_i y_i^* \log(o_i)$$

$$\frac{\partial L}{\partial o_i} = -\frac{y_i^*}{o_i}$$

$$\frac{\partial o_i}{\partial s_j} = o_i(1 - o_i)$$

(when $j = i$)

$$\frac{\partial o_i}{\partial s_j} = -o_i o_j$$

(when $j != i$)

$$= \delta_{i,j} o_i - o_i o_j$$

$$\frac{\partial E}{\partial s_i} = o_i - y_i^*$$

Squared error:

$$L = \sum_k \frac{1}{2}(y_k^* - o_k)^2$$

$$\frac{\partial L}{\partial o_i} = o_i - y_i^*$$

Activation functions

Sigmoid:

$$o_i = \frac{1}{1 + \exp(-s_i)}$$

$$\frac{\partial o_i}{\partial s_i} = o_i(1 - o_i)$$

Linear:

$$o_i = s_i$$

$$\frac{\partial o_i}{\partial s_i} = 1$$

softmax:

$$o_i = \frac{\exp(s_i)}{\sum_k \exp(s_k)}$$

$$\frac{\partial o_i}{\partial s_i} =$$

errors backprop

My current method:

$$\text{error}^{l-1} = \sum_j \text{error}^l_j \frac{\partial o^l_j}{\partial o^{l-1}_i}$$

$$= \sum_j \frac{\partial L}{\partial o^l_j} \frac{\partial o^l_j}{\partial s^l_j} \frac{\partial s^l_j}{\partial o^{l-1}_i}$$

eg

$$= \sum_i \frac{\partial L}{\partial o^l_i} (1 - (o^l_i)^2)) w^l_{ji}$$

$$\frac{\partial L}{\partial w^l_{ji}} = \frac{\partial L}{\partial o^l_i} \frac{\partial o^l_i}{\partial s^l_i} \frac{\partial s^l_i}{\partial w^l_{ji}}$$

eg

$$= \frac{\partial L}{\partial o^l_i} (1 - (o^l_i)^2) o^{l-1}_j$$

Peter Sadowski ”notes on backpropagation” method:

$$\frac{\partial L}{\partial s^{l-1}_i} = \sum_j \frac{\partial L}{\partial s^l_j} \frac{\partial s^l_j}{\partial o^{l-1}_i} \frac{\partial o^{l-1}_i}{\partial s^{l-1}_i}$$

eg

$$= \sum_j \frac{\partial L}{\partial s^l_j} w^l_{ij} (1 - (o^{l-1}_i)^2)$$

$$\frac{\partial L}{\partial w^l_{ji}} = \frac{\partial L}{\partial s^l_i} \frac{\partial s^l_i}{\partial w^l_{ji}}$$

$$= \frac{\partial L}{\partial s^l_i} o^{l-1}_j$$