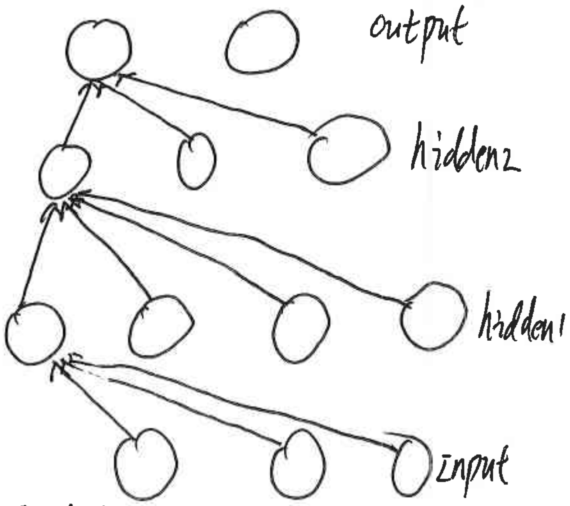
多层感知机



前向传播

$a_h = \sum_{i=1}^{I} w_{ih} x_i$ 输入→隐层

$b_h = \theta_h(a_h)$

隐层→隐层 l:隐层数量

$a_h = \sum_{h' \in H_{l-1}} w_{h'h} b_{h'}$

$b_h = \theta_h(a_h)$

隐层→输出层

$a_k = \sum_{h \in H_L} w_{hk} b_h$

对于多分类:

$p(C_k|x) = y_k = \frac{e^{a_k}}{\sum_{k'=1}^{K} e^{a_{k'}}}$

$p(z|x) = \prod_{k=1}^{K} y_k^{z_k}$

$z_1 = (1,0,0,0)$ 4类

$z_2 = (0100)$

对于一个样本 $\sum_{k=1}^{K} z_k = 1$

$\mathcal{L}(x,z) = -\sum_{k=1}^{K} z_k \ln y_k$

其实: $\mathcal{L}(S) = -\ln \prod_{(x,z)\in S} p(z|x)$

$= -\sum_{(x,z)\in S} \ln p(z|x)$

令 $\mathcal{L}(x,z) = -\ln p(z|x)$

$\mathcal{L}(S) = \sum_{(x,z)\in S} \mathcal{L}(x,z)$

$\frac{\partial \mathcal{L}(S)}{\partial w} = \sum_{(x,z)\in S} \frac{\partial \mathcal{L}(x,z)}{\partial w}$

反向传播.

输出层: $\frac{\partial \mathcal{L}(x,z)}{\partial a_k} = \sum_{k'=1}^{K} \frac{\partial \mathcal{L}(x,z)}{\partial y_{k'}} \frac{\partial y_{k'}}{\partial a_k}$

$\frac{\partial \mathcal{L}(x,z)}{\partial y_{k'}} = \frac{\partial\left(-\sum_{i=1}^{K} z_i \ln y_i\right)}{\partial y_{k'}} = -\frac{z_{k'}}{y_{k'}}$

$\frac{\partial y_{k'}}{\partial a_k} = \frac{\partial\left(\frac{e^{a_{k'}}}{\sum_{i=1}^{K} e^{a_i}}\right)}{\partial a_k}$

$= y_k \delta_{kk'} - y_k y_{k'}$ (if $k=k'$ $\delta_{kk'}=1$ else $\delta_{kk'}=0$.)

因此: $\frac{\partial \mathcal{L}(x,z)}{\partial a_k} = \cancel{-\frac{z_{k'}}{y_{k'}}} \sum_{k'=1}^{K} -\frac{z_{k'}}{y_{k'}} (y_k \delta_{kk'} - y_k y_{k'})$

$= y_k - z_k$ ($\sum_{i=1}^{K} z_i = 1$)

这一层的梯度: $\Delta w_{hk} = \frac{\partial \mathcal{L}(x,z)}{\partial a_k} \frac{\partial a_k}{\partial w_{hk}} = b_h (y_k - z_k)$

定义 $\delta_j \overset{def}{=} \frac{\partial L(x,z)}{\partial a_j}$　　H_i：第i层神经个数

h_i：第i层第h个神经元

第二层的隐层的梯度. i：第i个输入，即x_i

$$\delta_{h_2} = \frac{\partial L(x,z)}{\partial a_{h_2}} = \sum_{k=1}^{K} \frac{\partial L(x,z)}{\partial a_k}\frac{\partial a_k}{\partial b_{h_2}}\frac{\partial b_{h_2}}{\partial a_{h_2}}$$

$$= \theta'(a_{h_2}) \sum_{k=1}^{K} \delta_k w_{h_2 k}$$

$$\cancel{\Delta w_{h_1 h_2} = \frac{\partial L(x,z)}{\partial w_{h_1 h_2}} = \frac{\partial L(x,z)}{\partial a_{h_2}}\frac{\partial a_{h_2}}{\partial w_{h_1 h_2}}}$$

$$\cancel{= b_{h_1}\delta_{h_2}}$$

$$\Delta w_{h_1 h_2} = \frac{\partial L(x,z)}{\partial w_{h_1 h_2}} = \frac{\partial L(x,z)}{\partial a_{h_2}}\frac{\partial a_{h_2}}{\partial w_{h_1 h_2}}$$

$$= \delta_{h_2} b_{h_1}$$

第一层隐层的梯度.

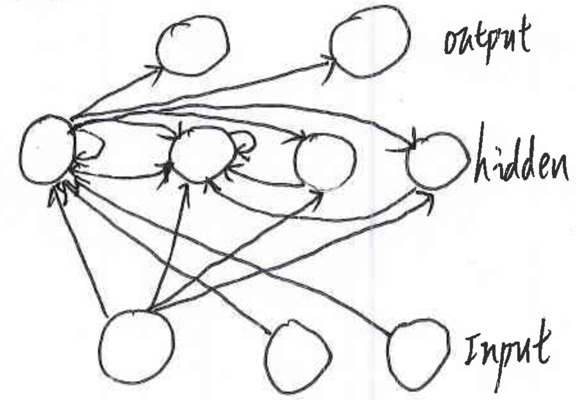
$$\delta_{h_1} = \frac{\partial L(x,z)}{\partial a_{h_1}} = \sum_{h_2=1}^{H_2} \frac{\partial L(x,z)}{\partial a_{h_2}}\frac{\partial a_{h_2}}{\partial b_{h_1}}\frac{\partial b_{h_1}}{\partial a_{h_1}}$$

$$= \theta'(a_{h_1}) \sum_{h_2=1}^{H} \delta_{h_2} w_{h_1 h_2}$$

$$w_{i h_1} = \frac{\partial L(x,z)}{\partial w_{i h_1}} = \frac{\partial L(x,z)}{\partial a_{h_1}}\frac{\partial a_{h_1}}{\partial w_{i h_1}}$$

$$= \delta_{h_1} x_i$$

RNN



前向传播 对于长度为T的序列x

H：隐层神经元个数,此处为4个

K：为类别个数

$$a_h^t = \sum_{i=1}^{I} w_{ih} x_i^t + \sum_{h'=1}^{H} w_{h'h} b_{h'}^{t-1}$$

$$b_h^t = \theta_h(a_h^t)$$

$$a_k^t = \sum_{h=1}^{H} w_{hk} b_h^t$$

$$y_k^t = \frac{e^{a_k^t}}{\sum_{k'=1}^{K} e^{a_{k'}^t}}$$

$$L(x^t, z^t) = -\sum_{k=1}^{K} z_k^t \ln y_k^t$$

损失函数 $L = \sum_{t=1}^{T} L(x^t, z^t)$

反向传播:

输出层:

$$\delta_k^t = \frac{\partial L(x^t, z^t)}{\partial a_k^t} = y_k^t - z_k^t$$

$$\Delta w_{hk} = \frac{\partial L}{\partial w_{hk}} = \sum_{t=1}^{T} \frac{\partial L(x^t, z^t)}{\partial a_k^t} \frac{\partial a_k^t}{\partial w_{hk}}$$

$$= \sum_{t=1}^{T} \delta_k^t b_h^t$$

隐层:

$$\delta_h^t = \frac{\partial L}{\partial a_h^t} = \sum_{t=1}^{T} \frac{\partial L(x^t, z^t)}{\partial a_h^t} = \frac{\partial L(x^t, z^t)}{\partial a_h^t} + \frac{\partial L(x^{t+1}, z^{t+1})}{\partial a_h^t}$$

$$= \sum_{k=1}^{K} \frac{\partial L(x^t, z^t)}{\partial a_k^t} \frac{\partial a_k^t}{\partial b_h^t} \frac{\partial b_h^t}{\partial a_h^t} + \sum_{h'=1}^{H} \frac{\partial L(x^{t+1}, z^{t+1})}{\partial a_{h'}^{t+1}} \frac{\partial a_{h'}^{t+1}}{\partial b_h^t} \frac{\partial b_h^t}{\partial a_h^t}$$

$$= \theta'(a_h^t) \sum_{k=1}^{K} \delta_k^t w_{hk} + \theta'(a_h^t) \sum_{h'=1}^{H} \delta_{h'}^{t+1} w_{hh'}$$

$$= \theta'(a_h^t) \left(\sum_{k=1}^{K} \delta_k^t w_{hk} + \sum_{h'=1}^{H} \delta_{h'}^{t+1} w_{hh'} \right)$$

$$\Delta w_{ih} = \frac{\partial L}{\partial w_{ih}} = \sum_{t=1}^{T} \frac{\partial L(x^t, z^t)}{\partial a_h^t} \frac{\partial a_h^t}{\partial w_{ih}}$$

$$= \sum_{t=1}^{T} \delta_h^t x_i^t$$

前向传播

Input gates

$$a_\iota^t = \sum_{i=1}^{I} w_{i\iota} x_i^t + \sum_{h=1}^{H} w_{h\iota} b_h^{t-1} + \sum_{c=1}^{C} w_{c\iota} s_c^{t-1}$$

$$b_\iota^t = f(a_\iota^t)$$

foget gates

$$a_\phi^t = \sum_{i=1}^{I} w_{i\phi} x_i^t + \sum_{h=1}^{H} w_{h\phi} b_h^{t-1} + \sum_{c=1}^{C} w_{c\phi} s_c^{t-1}$$

$$b_\phi^t = f(a_\phi^t)$$

Cells

$$a_c^t = \sum_{i=1}^{I} w_{ic} x_i^t + \sum_{h=1}^{H} w_{hc} b_h^{t-1}$$

$$s_c^t = b_\phi^t s_c^{t-1} + b_\iota^t g(a_c^t)$$

output gates

$$a_\omega^t = \sum_{i=1}^{I} w_{i\omega} x_i^t + \sum_{h=1}^{H} w_{h\omega} b_h^{t-1} + \sum_{c=1}^{C} w_{c\omega} s_c^t$$

$$b_\omega^t = f(a_\omega^t)$$

cell outputs

$$b_c^t = b_\omega^t h(s_c^t)$$

反向传播 $\epsilon_c^t = \frac{\partial \mathcal{L}}{\partial b_c^t}$ $\epsilon_s^t = \frac{\partial \mathcal{L}}{\partial s_c^t}$

$$y_k^t = \frac{e^{a_k^t}}{\sum_{k'=1}^{K} e^{a_{k'}^t}} \qquad \delta_j^t = \frac{\partial \mathcal{L}}{\partial a_j^t}$$

$$\mathcal{L} = \sum_{t=1}^{T} \mathcal{L}(x^t, z^t)$$

$$\mathcal{L} = -\sum_{k=1}^{K} z_k^t \ln y_k^t$$

$$a_k^t = \sum_{h=1}^{H} \sum_{c=1}^{C} b_c^t w_c^k$$

$$\frac{\partial \mathcal{L}(x^t, z^t)}{\partial a_k^t} = y_k^t - z_k^t$$

$$\Delta w_{ck} = \frac{\partial \mathcal{L}}{\partial w_{ck}} = \sum_{t=1}^{T} \frac{\partial \mathcal{L}(x^t, z^t)}{\partial w_{ck}}$$

$$= \sum_{t=1}^{T} \frac{\partial (x^t, z^t)}{\partial a_k^t} \frac{\partial a_k^t}{\partial w_{ck}}$$

$$= \sum_{t=1}^{T} \delta_k b_c^t$$

隐层参数:

$$\epsilon_c^t = \frac{\partial \mathcal{L}}{\partial b_c^t} = \sum_{k=1}^{K} \frac{\partial \mathcal{L}}{\partial a_k^t} \frac{\partial a_k^t}{\partial b_c^t} + \sum_{\omega=1}^{H} \frac{\partial \mathcal{L}}{\partial a_\omega^{t+1}} \frac{\partial a_\omega^{t+1}}{\partial b_c^t}$$

$$+ \sum_{\phi=1}^{H} \frac{\partial \mathcal{L}}{\partial a_\phi^{t+1}} \frac{\partial a_\phi^{t+1}}{\partial b_c^t} + \sum_{l=1}^{H} \frac{\partial \mathcal{L}}{\partial a_l^{t+1}} \frac{\partial a_l^{t+1}}{\partial b_c^t}$$

$$+ \sum_{m=1}^{H\times C} \frac{\partial \mathcal{L}}{\partial a_m^{t+1}} \frac{\partial a_m^{t+1}}{\partial b_c^t}$$

$$= \sum_{k=1}^{K} \delta_k^t w_{ck} + \sum_{\omega=1}^{H} \delta_\omega^{t+1} w_{c\omega} + \sum_{\phi=1}^{H} \delta_\phi^{t+1} w_{c\phi}$$

$$+ \sum_{l=1}^{H} \delta_l^{t+1} w_{cl} + \sum_{m=1}^{H\times C} \delta_m^{t+1} w_{cm}$$

$$= \sum_{k=1}^{K} \delta_k^t w_{ck} + \sum_{g}^{G} \delta_g^{t+1} w_{cg}$$

Output gates.

$$\delta_\omega^t = \sum_{c=1}^{C} \frac{\partial \mathcal{L}}{\partial b_c^t} \frac{\partial b_c^t}{\partial a_\omega^t} = f'(a_\omega^t) \sum_{c=1}^{C} \epsilon_c^t h(s_c^t)$$

$$\cancel{\Delta w_{i\omega} = \frac{\partial \mathcal{L}}{\partial w_{i\omega}} = \sum_{t=1}^{T} \frac{\partial \mathcal{L}}{\partial a_\omega^t} \frac{\partial a_\omega^t}{\partial w_{i\omega}} = \sum_{t=1}^{T}}$$

$$\Delta w_{i\omega} = \frac{\partial \mathcal{L}}{\partial w_{i\omega}} = \sum_{t=1}^{T} \frac{\partial \mathcal{L}}{\partial a_\omega^t} \frac{\partial a_\omega^t}{\partial w_{i\omega}} = \sum_{t=1}^{T} \delta_\omega^t x_i$$

$$\Delta w_{h\omega} = \frac{\partial \mathcal{L}}{\partial w_{h\omega}} = \sum_{t=1}^{T} \frac{\partial \mathcal{L}}{\partial a_\omega^t} \frac{\partial a_\omega^t}{\partial w_{h\omega}} = \sum_{t=1}^{T} \delta_\omega^t b_h^{t-1}$$

$$\Delta w_{c\omega} = \sum_{t=1}^{T} \delta_\omega^t s_c^t$$

$\epsilon_s^t = \frac{\partial \mathcal{L}}{\partial s_c^t}$ (states)

$$\epsilon_s^t = \frac{\partial \mathcal{L}}{\partial s_c^t} = \frac{\partial \mathcal{L}}{\partial b_c^t}\frac{\partial b_c^t}{\partial s_c^t} + \frac{\partial \mathcal{L}}{\partial a_\omega^t}\frac{\partial a_\omega^t}{\partial s_c^t} + \frac{\partial \mathcal{L}}{\partial a_\phi^{t+1}}\frac{\partial a_\phi^{t+1}}{\partial s_c^t} + \frac{\partial \mathcal{L}}{\partial a_\iota^{t+1}}\frac{\partial a_\iota^{t+1}}{\partial s_c^t} + \frac{\partial \mathcal{L}}{\partial s_c^{t+1}}\frac{\partial s_c^{t+1}}{\partial s_c^t}$$

$$= b_\omega^t \epsilon_c^t h'(s_c^t) + \delta_\omega^t w_{c\omega} + \delta_\phi^{t+1} w_{c\phi} + \delta_\iota^{t+1} w_{c\iota} + b_\phi^{t+1} \epsilon_s^{t+1}$$

cells:

$$\delta_c^t = \frac{\partial \mathcal{L}}{\partial a_c^t} = \frac{\partial \mathcal{L}}{\partial s_c^t}\frac{\partial s_c^t}{\partial a_c^t} = b_\iota^t \epsilon_s^t g'(a_c^t)$$

$\Delta w_{ic} = \delta_c^t x_i$ ~~$\Delta w_{h\phi} = \delta_c^t b_h^{t-1}$ $\Delta w_{c\phi} = \delta_c^t s_c^{t-1}$~~

$\Delta w_{hc} = \delta_c^t b_h^{t-1}$

Forget GagGates:

$$\delta_\phi^t = \sum_{c=1}^{C} \frac{\partial \mathcal{L}}{\partial s_c^t}\frac{\partial s_c^t}{\partial b_\phi^t}\frac{\partial b_\phi^t}{\partial a_\phi^t}$$

$$= \sum_{c=1}^{C} \epsilon_c^t s_c^{t-1} f'(a_\phi^t)$$

$\Delta w_{i\phi} = \delta_\phi^t x_i$ $\Delta w_{h\phi} = \delta_\phi^t b_h^{t-1}$ $\Delta w_{c\phi} = \delta_\phi^t s_c^{t-1}$

Input gates:

$$\delta_\iota^t = \frac{\partial \mathcal{L}}{\partial a_\iota^t} = \sum_{c=1}^{C} \frac{\partial \mathcal{L}}{\partial s_c^t}\frac{\partial s_c^t}{\partial b_\iota^t}\frac{\partial b_\iota^t}{\partial a_\iota^t}$$

$$= f'(a_\iota^t) \sum_{c=1}^{C} \epsilon_s^t g(a_c^t)$$

$\Delta w_{i\iota} = \delta_\iota^t x_i^t$

$\Delta w_{h\iota} = \delta_\iota^t b_h^{t-1}$ $\Delta w_{c\iota} = \delta_\iota^t s_c^{t-1}$