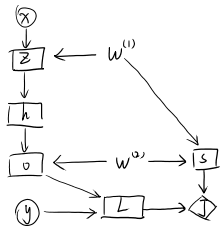
FP、BP、BPTT

2018年7月31日 10:25

- 设 $Y=f(X)$，$Z=g(Y)$

$$\frac{\partial Z}{\partial X} = \text{prod}\left(\frac{\partial Z}{\partial Y}, \frac{\partial Y}{\partial X}\right)$$

- 例：BP 化的多层感知机

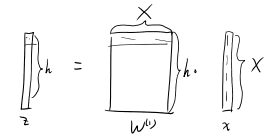


FP:

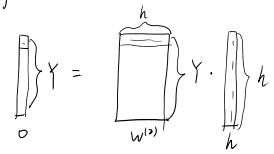
(1) $z = W^{(1)} x$

(2) $h = \varphi(z)$

(3) $o = W^{(2)} h$



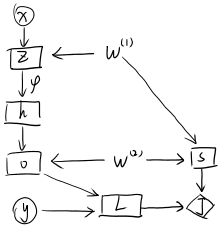
$\varphi(\cdot)$ element wise



(4) $\underset{\text{scalar}}{L} = \ell(o, y)$

(5) $S = \frac{\lambda}{2}\left(\|W^{(1)}\|_2^2 + \|W^{(2)}\|_2^2\right)$

(6) $J = L + S$



BP:

(1) $\frac{\partial J}{\partial L} = 1$，$\frac{\partial J}{\partial S} = 1$

(2) $\frac{\partial J}{\partial o} = \text{prod}\left(\frac{\partial J}{\partial L}, \frac{\partial L}{\partial o}\right)$

$= 1 \cdot \frac{\partial \ell(o,y)}{\partial o}$

$= \frac{\partial \ell}{\partial o}$

(3) $\frac{\partial S}{\partial W^{(1)}} = \lambda W^{(1)}$，$\frac{\partial S}{\partial W^{(2)}} = \lambda W^{(2)}$

(4) $\frac{\partial J}{\partial W^{(2)}}$ ($y \times h$) $= \text{prod}\left(\frac{\partial J}{\partial o}\ (y\times 1), \frac{\partial o}{\partial W^{(2)}}\ (h: h\times 1)\right) + \text{prod}\left(\frac{\partial J}{\partial S}, \frac{\partial S}{\partial W^{(2)}}\right)$

$= \frac{\partial \ell}{\partial o} \cdot h^T + \lambda W^{(2)}$

$\frac{\partial J}{\partial h}$ ($h\times 1$) $= \text{prod}\left(\frac{\partial J}{\partial o}\ (y\times 1), \frac{\partial o}{\partial h}\ (W^{(2)}: y\times h)\right)$

$= W^{(2)T} \cdot \frac{\partial \ell}{\partial o}$

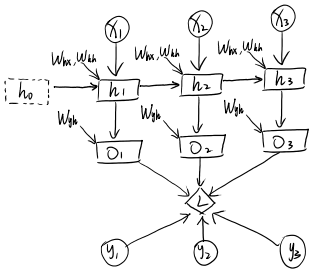
(6) $\frac{\partial J}{\partial z}$ ($h\times 1$) $= \text{prod}\left(\frac{\partial J}{\partial h}\ (h\times 1), \frac{\partial h}{\partial z}\ (h\times 1)\right)$

$= \left(W^{(2)T} \frac{\partial \ell}{\partial o}\right) \odot \varphi'(z)$

(7) $\frac{\partial J}{\partial W^{(1)}}$ ($h\times x$) $= \text{prod}\left(\frac{\partial J}{\partial z}\ (h\times 1), \frac{\partial z}{\partial W^{(1)}}\ (x: x\times 1)\right) + \text{prod}\left(\frac{\partial J}{\partial S}, \frac{\partial S}{\partial W^{(1)}}\right)$

$= \frac{\partial J}{\partial z} \cdot x^T + \lambda W^{(1)}$

- RNN



FP:

(1) h_t ($h\times 1$) $= \varphi\left(W_{hx}\ (h\times x) \cdot x_t\ (x\times 1) + W_{hh}\ (h\cdot h) \cdot h_{t-1}\ (h\times 1)\right)$ $\quad h_{T+1} = W_{hh} \cdot h_T$

(2) O_t ($y\times 1$) $= W_{yh}\ (y\times h) \cdot h_t\ (h\times 1)$

(3) $L = \frac{1}{T}\sum_{t=1}^{T} \ell(o_t, y_t)$

BP:

(1) $\frac{\partial L}{\partial O_t} = \frac{1}{T} \cdot \frac{\partial \ell(O_t, y_t)}{\partial O_t}$

(2) $\frac{\partial L}{\partial W_{yh}}$ ($y\times h$) $= \sum_{t=1}^{T} \text{prod}\left(\frac{\partial L}{\partial O_t}\ (y\times 1), \frac{\partial O_t}{\partial W_{yh}}\ (h_t: h\times 1)\right)$

$= \sum_{t=1}^{T} \frac{\partial L}{\partial O_t} h_t^T$

(3) $\frac{\partial L}{\partial h_T}$ ($h\times 1$) $= \text{prod}\left(\frac{\partial L}{\partial O_T}\ (y\times 1), \frac{\partial O_T}{\partial h_T}\ (W_{yh}: y\times h)\right)$

$= W_{yh}^T \cdot \frac{\partial L}{\partial O_T}$

(3) $\frac{\partial L}{\partial h_T} = \text{Prod}\left(\frac{\partial L}{\partial O_T}, \frac{\partial O_T}{\partial h_T}\right)$

$= W_{yh}^T \cdot \frac{\partial L}{\partial O_T}$

T 是时序的最大值，如图中 T=3

$\frac{\partial L}{\partial h_t}$ ($h\times 1$) $= \text{Prod}\left(\frac{\partial L}{\partial h_{t+1}}, \frac{\partial h_{t+1}}{\partial h_t}\right) + \text{Prod}\left(\frac{\partial L}{\partial O_t}, \frac{\partial O_t}{\partial h_t}\right)$

($\frac{\partial L}{\partial h_{t+1}}$: $h\times 1$, W_{hh}: $h\times h$; $\frac{\partial L}{\partial O_t}$: $y\times 1$, W_{yh}: $y\times h$)

error from t+1 ✲ BP through time　　error from t

$= W_{hh}^T \cdot \boxed{\frac{\partial L}{\partial h_{t+1}}} + W_{yh}^T \cdot \frac{\partial L}{\partial O_t}$　$(t<T)$

Calculate recursively from $\frac{\partial L}{\partial h_T}$　✲ gradient explosion

通项：$\frac{\partial L}{\partial h_t} = \sum_{i=t}^{T} (W_{hh}^T)^{T-i} W_{yh}^T \frac{\partial L}{\partial O_{T+t-i}}$

- when $i \to 0$, $\lim_{i\to 0} (W_{hh}^T)^{T-i} = (W_{hh}^T)^T$ is big ⇒ gradient explosion

(4) $\frac{\partial L}{\partial W_{hx}}$ ($h\times x$) $= \sum_{t=1}^{T} \text{Prod}\left(\frac{\partial L}{\partial h_t}, \frac{\partial h_t}{\partial W_{hx}}\right)$

($\frac{\partial L}{\partial h_t}$: $h\times 1$, x_t: $x\times 1$)

$= \sum_{t=1}^{T} \left[\left(\frac{\partial L}{\partial h_t} \odot \varphi'(\cdots)\right) \cdot x_t^T\right]$

($\frac{\partial L}{\partial h_t}$: $h\times 1$, h_{t-1}: $h\times 1$)

$\frac{\partial L}{\partial W_{hh}}$ ($h\times h$) $= \sum_{t=1}^{T} \text{Prod}\left(\frac{\partial L}{\partial h_t}, \frac{\partial h_t}{\partial W_{hh}}\right)$

($\frac{\partial L}{\partial h_t}$: $h\times 1$, h_{t+1}: $h\times 1$)

$= \sum_{t=1}^{T} \left[\left(\frac{\partial L}{\partial h_t} \odot \varphi'(\cdots)\right) \cdot h_{t-1}^T\right]$