CS224n : Deep Learning for NLP

Pratical Midterm #1

① $h_1 = \sigma(Wx_1 + b)$; $h_2 = \sigma(Wx_2 + b)$

$J = \frac{1}{2}\|h_1 - h_2\|_F^2 + \frac{\lambda}{2}\|W\|_F^2$

$\frac{\partial J}{\partial W} = ?$ $\frac{\partial J}{\partial b} = ?$

$W \in \mathbb{R}^{m \times n}$

$b \in \mathbb{R}^m$

$x_1, x_2 \in \mathbb{R}^n$

J could be rewritten as:

$$J = \frac{1}{2}(h_1 - h_2)^T \cdot (h_1 - h_2) + \frac{\lambda}{2}\sum_j W_j^T \cdot W_j$$

$$\Rightarrow \frac{\partial J}{\partial W} = \frac{1}{2}\frac{\partial[(h_1-h_2)^T \cdot (h_1-h_2)]}{\partial(h_1-h_2)} \cdot \frac{\partial(h_1-h_2)}{\partial W} + \frac{1}{2}\lambda \frac{\partial}{\partial W}\left(\sum_j W_j^T W_j\right)$$

$$\boxed{\frac{\partial J}{\partial W} = \frac{1}{2}(h_1 - h_2)\left[\sigma'(Wx_1 + b)x_1^T - \sigma'(Wx_2 + b)x_2^T\right] + \frac{1}{2}\lambda W}$$

Similarly; $\boxed{\frac{\partial J}{\partial b} = \frac{1}{2}(h_1 - h_2)\left[\sigma'(Wx_1 + b) - \sigma'(Wx_2 + b)\right]}$

② $W =: W - \alpha \frac{\partial J}{\partial W}$

$b =: b - \alpha \frac{\partial J}{\partial b}$

③ 60 parameters

④ This model doesn't make use of the contextual information in surrounding words

$h_1 = \sigma(W_1 x + b_1)$

$h_2 = \text{Relu}(W_2 x + b_2)$

$\hat{y} = \text{softmax}(W_3(h_1 + h_2) + b_3)$

$x \in \mathbb{R}^n$

$W_1, W_2 \in \mathbb{R}^{m \times n}$

$W_3 \in \mathbb{R}^{k \times m}$

$b_1, b_2 \in \mathbb{R}^m$

$b_3 \in \mathbb{R}^k$

$$J = -\frac{1}{N} \sum_{j=1}^{N} \sum_{i=1}^{k} y_j^i \log(\hat{y}_j^i)$$

$z_1 = W_1 x + b_1$

$z_2 = W_2 x + b_2$

let $z_3 = W_3(h_1 + h_2) + b_3$

$\frac{\partial J}{\partial h_1} = ?$ $\frac{\partial J}{\partial h_2} = ?$ $\frac{\partial J}{\partial x} = ?$

$$\frac{\partial J}{\partial h_1} = \frac{\partial J_j}{\partial z_3} \cdot \frac{\partial z_3}{\partial h_1} = -\frac{1}{N} \sum_{j=1}^{N} (y_j - \hat{y}_j)$$

$$\frac{\partial J_j}{\partial h_1} = (y_j - \hat{y}_j) \cdot W_3^T \implies \boxed{\frac{\partial J}{\partial h_1} = \frac{1}{N} \sum_j^N (y_j - \hat{y}_j) \cdot W_3^T}$$

similarly $\boxed{\frac{\partial J}{\partial h_2} = \frac{1}{N} \sum_j^N (y_j - \hat{y}_j) W_3^T}$

$$\frac{\partial J}{\partial x} = \frac{\partial J}{\partial h_1} \cdot \frac{\partial h_1}{\partial x} + \frac{\partial J}{\partial h_2} \cdot \frac{\partial h_2}{\partial x}$$

$$\frac{\partial h_1}{\partial x} = \sigma'(z_1) \cdot W_1^T$$

$$\frac{\partial h_2}{\partial x} = \begin{cases} W_2^T, & \text{if } z_2 \geq 0 \\ 0, & \text{if } z_2 < 0 \end{cases}$$

$$\implies \frac{\partial J}{\partial x} = \begin{cases} \frac{1}{N} \sum_j^N (y_j - \hat{y}_j) W_3^T \cdot (W_1^T + W_2^T) & \text{if } z_2 \geq 0 \\ \frac{1}{N} \sum_j (y_j - \hat{y}_j) W_3^T \cdot W_1^T & \text{if } z_2 < \end{cases}$$

6) W_1 will train faster because updating W_2 depends on z_2 being ≥ 0 whereas W_1 is always updated.

3) b) *) $\frac{\partial J^{(t)}}{\partial b_2} = \frac{\partial J^{(t)}}{\partial z^{(t)}} \cdot \frac{\partial z^{(t)}}{\partial b_2}$

We've show in Assignment ① that $\frac{\partial J^{(t)}}{\partial z^{(t)}} = \left(\hat{y}^{(t)} - y^{(t)}\right) = \delta^{(t)}$

$$\frac{\partial z^{(t)}}{\partial b_2} = \frac{\partial}{\partial b_2}\left(h^{(t)} U + b_2\right) = 0 + \frac{\partial b_2}{\partial b_2} = \text{Identity Matrix} = [1]$$

$$\Rightarrow \boxed{\frac{\partial J^{(t)}}{\partial b_2} = \Delta^{(t)} = \hat{y}^{(t)} - y^{(t)}}$$

*) $\frac{\partial J^{(t)}}{\partial L_{x^{(t)}}} = \frac{\partial J^{(t)}}{\partial e^{(t)}} = \frac{\partial J^{(t)}}{\partial z^{(t)}} \cdot \frac{\partial z^{(t)}}{\partial e^{(t)}} = \Delta^{(t)} \cdot \frac{\partial z^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial e^{(t)}}$

+ $\frac{\partial z^{(t)}}{\partial h^{(t)}} = \frac{\partial}{\partial h^{(t)}}\left(h^{(t)} \cdot U + b_2\right) = U^T$

+ $\frac{\partial h^{(t)}}{\partial e^{(t)}} = \text{sigmoid}'\left(h^{(t-1)} \cdot H + e^{(t)} \cdot I + b_1\right) \odot I^T = \sigma'(c^{(t)}) \odot I^T$

(with $\text{sigmoid}' = \sigma' = \sigma(1-\sigma)$) $\Rightarrow \boxed{\frac{\partial J^{(t)}}{\partial L_{x^{(t)}}} = \Delta^{(t)} \cdot U^T \odot \sigma'(c^{(t)}) \odot I}$

*) $\left.\frac{\partial J^{(t)}}{\partial I}\right|_{(t)} = \frac{\partial J^{(t)}}{\partial z^{(t)}} \cdot \frac{\partial z^{(t)}}{\partial h^{(t)}} \cdot \left.\frac{\partial h^{(t)}}{\partial I}\right|_{(t)}$

$$\boxed{\left.\frac{\partial J^{(t)}}{\partial I}\right|_{(t)} = \Delta^{(t)} \cdot U^T \cdot \left.\frac{\partial h^{(t)}}{\partial I}\right|_{(t)}}$$

*) $\boxed{\left.\frac{\partial J^{(t)}}{\partial H}\right|_{(t)} = \left.\frac{\partial J^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial H}\right|_{(t)} = \Delta^{(t)} \cdot U^T \cdot \left.\frac{\partial h^{(t)}}{\partial H}\right|_{(t)}}$

*) $\frac{\partial J^{(t)}}{\partial h^{(t-1)}} = \frac{\partial J^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial h^{(t-1)}} = \Delta^{(t)} \cdot U^T \frac{\partial h^{(t)}}{\partial h^{(t-1)}}$

$$= \Delta^{(t)} \cdot U^T \odot \sigma'(c^{(t)}) \cdot H^T$$