cs224n

(arthurg)

May 18, 2020

Question 1.

Notice that $y_w = 1$ if $w = o$ and $y_w = 0$ iff $w \neq o$

So the LHS of the equation, when $w = 0$, the terms will drop off. We end up with only the term $1 * log(\hat{y}_w)$ when $w = o$. This simplifies LHS to $-log(\hat{y}_o)$

Question 2.

$$J = -\log \frac{\exp(u_o^T v_c)}{\sum_{w \in Vocab} exp(u_w^T v_c)} = -\log \exp(u_o^T v_c) + \log \sum_{w \in Vocab} exp(u_w^T v_c)$$

$$J = -(u_o^T v_c) + \log \sum_{w \in Vocab} exp(u_w^T v_c)$$

$$\frac{\partial J}{\partial v_c} = \frac{\partial(-u_o^T v_c)}{\partial v_c} + \frac{\partial \log \sum_{w \in Vocab} exp(u_w^T v_c)}{\partial v_c}$$

$$\frac{\partial J}{\partial v_c} = -u_o + \frac{1}{\sum_{w \in Vocab} exp(u_w^T v_c)} * \frac{\partial \sum_{w \in Vocab} exp(u_w^T v_c)}{\partial v_c}$$

$$\frac{\partial J}{\partial v_c} = -u_o + \frac{1}{\sum_{w \in Vocab} exp(u_w^T v_c)} * (\sum_{w \in Vocab} \frac{\partial exp(u_w^T v_c)}{\partial v_c})$$

$$\frac{\partial J}{\partial v_c} = -u_o + \frac{1}{\sum_{w \in Vocab} exp(u_w^T v_c)} * (\sum_{w \in Vocab} exp(u_w^T v_c) \frac{\partial u_w^T v_c}{\partial v_c})$$

$$\frac{\partial J}{\partial v_c} = -u_o + \frac{1}{\sum_{w \in Vocab} exp(u_w^T v_c)} * (\sum_{w \in Vocab} exp(u_w^T v_c) u_w^T)$$

$$\frac{\partial J}{\partial v_c} = -u_o + \sum_{w \in Vocab} u_w^T \frac{(exp(u_w^T v_c))}{\sum_{x \in Vocab} exp(u_x^T v_c)}$$

$$\frac{\partial J}{\partial v_c} = -u_o + \sum_{w \in Vocab} u_w^T P(O = w|C = c)$$

$$\frac{\partial J}{\partial v_c} = -u_o + U^T \hat{y}$$

$$\frac{\partial J}{\partial v_c} = U^T(\hat{y} - y)$$

Question 3.

Find partial deriv for u_o first

$$\frac{\partial J}{\partial u_o} = \frac{\partial(-u_o^T v_c)}{\partial u_o} + \frac{\partial \log \sum_{w \in Vocab} exp(u_w^T v_c)}{\partial u_o}$$

$$\frac{\partial J}{\partial u_o} = -v_c^T + \frac{1}{\sum_{w \in Vocab} exp(u_w^T v_c)} * \frac{\partial \sum_{w \in Vocab} exp(u_w^T v_c)}{\partial u_o}$$

$$\frac{\partial J}{\partial u_o} = -v_c^T + \frac{1}{\sum_{w \in Vocab} exp(u_w^T v_c)} * \sum_{w \in Vocab} \frac{\partial exp(u_w^T v_c)}{\partial u_o}$$

$$\frac{\partial J}{\partial u_o} = -v_c^T + \frac{1}{\sum_{w \in Vocab} exp(u_w^T v_c)} * \frac{\partial exp(u_o^T v_c)}{\partial u_o}$$

$$\frac{\partial J}{\partial u_o} = -v_c^T + \frac{1}{\sum_{w \in Vocab} exp(u_w^T v_c)} * exp(u_o^T v_c) * v_c^T$$

$$\frac{\partial J}{\partial u_o} = -v_c^T + \frac{exp(u_o^T v_c)}{\sum_{w \in Vocab} exp(u_w^T v_c)} * v_c^T$$

$$\frac{\partial J}{\partial u_o} = -v_c^T + P(O = o|C = c) * v_c^T$$

Find partial deriv for $u_w, w \neq o$

$$\frac{\partial J}{\partial u_w} = \frac{\partial(-u_o^T v_c)}{\partial u_w} + \frac{\partial \log \sum_{x \in Vocab} exp(u_x^T v_c)}{\partial u_w}$$

$$\frac{\partial J}{\partial u_w} = \frac{1}{\sum_{w \in Vocab} exp(u_w^T v_c)} * \frac{\partial \sum_{x \in Vocab} exp(u_x^T v_c)}{\partial u_w}$$

$$\frac{\partial J}{\partial u_w} = \frac{1}{\sum_{w \in Vocab} exp(u_w^T v_c)} * \sum_{x \in Vocab} \frac{\partial exp(u_x^T v_c)}{\partial u_w}$$

$$\frac{\partial J}{\partial u_w} = \frac{1}{\sum_{w \in Vocab} exp(u_w^T v_c)} * \frac{\partial exp(u_w^T v_c)}{\partial u_w}$$

$$\frac{\partial J}{\partial u_w} = \frac{1}{\sum_{w \in Vocab} exp(u_w^T v_c)} * exp(u_w^T v_c) * v_c^T$$

$$\frac{\partial J}{\partial u_w} = \frac{exp(u_w^T v_c)}{\sum_{w \in Vocab} exp(u_w^T v_c)} * v_c^T$$

$$\frac{\partial J}{\partial u_w} = P(O = w|C = c) * v_c^T$$

Overall,

$$\frac{\partial J}{\partial u_w} = (\hat{y} - y) * v_c^T$$

Question 4.

$$\frac{d\sigma(x)}{dx} = \frac{d(1+e^{-x})^{-}1}{d(1+e^{-x})} \frac{d(1+e^{-x})}{d(-x)} \frac{d(-x)}{dx}$$

$$\frac{d\sigma(x)}{dx} = (1+e^{-x})^{-2} * \text{`}e^{-x} * -1$$

$$\frac{d\sigma(x)}{dx} = \frac{e^{-x}}{(1+e^{-x})^2}$$

$$\frac{d\sigma(x)}{dx} = \frac{1+e^{-x}-1}{(1+e^{-x})^2}$$

$$\frac{d\sigma(x)}{dx} = (\frac{1+e^{-x}}{(1+e^{-x})} - \frac{1}{(1+e^{-x})})\frac{1}{(1+e^{-x})}$$

$$\frac{d\sigma(x)}{dx} = (1-\sigma(x))\sigma(x)$$

Question 5.

$$J = -\log \sigma(u_o^T v_c) - \sum_{k=1}^{K} \log\left(\sigma(-u_k^T v_c)\right)$$

Find Partial WRT v_c

$$\frac{\partial J}{\partial v_c} = \frac{\partial(-\log\sigma(u_o^T v_c))}{\partial(\sigma(u_o^T v_c))}\frac{\partial(\sigma(u_o^T v_c))}{\partial(u_o^T v_c)}\frac{\partial(u_o^T v_c)}{\partial(v_c)} - \sum_{k=1}^{K}\frac{\partial\log(\sigma(-u_k^T v_c))}{\partial(\sigma(-u_k^T v_c))}\frac{\partial(\sigma(-u_k^T v_c))}{\partial(-u_k^T v_c)}\frac{\partial(-u_k^T v_c)}{\partial(v_c)}$$

$$\frac{\partial J}{\partial v_c} = \frac{-1}{(\sigma(u_o^T v_c))}(1-\sigma(u_o^T v_c))(\sigma(u_o^T v_c))u_o - \sum_{k=1}^{K}\frac{1}{(\sigma(-u_k^T v_c))}(\sigma(-u_k^T v_c))(1-\sigma(-u_k^T v_c))(-u_k)$$

$$\frac{\partial J}{\partial v_c} = (\sigma(u_o^T v_c) - 1)u_o - \sum_{k=1}^{K}(1-\sigma(-u_k^T v_c))(-u_k)$$

$$\frac{\partial J}{\partial v_c} = (\sigma(u_o^T v_c) - 1)u_o + \sum_{k=1}^{K}(1-\sigma(-u_k^T v_c))(u_k)$$

Find Partial WRT u_o

$$\frac{\partial J}{\partial u_o} = \frac{\partial(-\log\sigma(u_o^T v_c))}{\partial(\sigma(u_o^T v_c))}\frac{\partial(\sigma(u_o^T v_c))}{\partial(u_o^T v_c)}\frac{\partial(u_o^T v_c)}{\partial(u_o)} - \sum_{k=1}^{K}\frac{\partial\log(\sigma(-u_k^T v_c))}{\partial(u_o)}$$

$$\frac{\partial J}{\partial u_o} = \frac{-1}{(\sigma(u_o^T v_c))}(1-\sigma(u_o^T v_c))(\sigma(u_o^T v_c))v_c$$

$$\frac{\partial J}{\partial v_c} = (\sigma(u_o^T v_c) - 1)v_c$$

Find Partial WRT $u_x, x \neq o$

$$\frac{\partial J}{\partial u_x} = \frac{\partial(-\log\sigma(u_o^T v_c))}{\partial(u_x)} - \sum_{k=1}^{K}\frac{\partial\log(\sigma(-u_k^T v_c))}{\partial(\sigma(-u_k^T v_c))}\frac{\partial(\sigma(-u_k^T v_c))}{\partial(-u_k^T v_c)}\frac{\partial(-u_k^T v_c)}{\partial(u_x)}$$

$$\frac{\partial J}{\partial u_x} = -\frac{\partial\log(\sigma(-u_x^T v_c))}{\partial(\sigma(-u_x^T v_c))}\frac{\partial(\sigma(-u_x^T v_c))}{\partial(-u_x^T v_c)}\frac{\partial(-u_x^T v_c)}{\partial(v_c)}$$

$$\frac{\partial J}{\partial u_x} = \frac{-1}{(\sigma(-u_x^T v_c))}(\sigma(-u_x^T v_c))(1-\sigma(-u_x^T v_c))(-v_c)$$

$$\frac{\partial J}{\partial u_x} = (1-\sigma(-u_x^T v_c))(v_c)$$

This is a lot more efficient than the naive-softmax implementation because naive-softmax uses U, V matricies, which are $O(|vocab||embedding|)$. Meanwhile, this new implementation only uses certain rows of U, V, making the runtime $O(|K||embedding|)$ which is much smaller

Question 6.

$$\frac{\partial J_{skip\_gram}}{\partial U} = \sum_{-m \leq j \leq m, j \neq 0} \frac{\partial (J)}{\partial U}$$

$$\frac{\partial J_{skip\_gram}}{\partial v_c} = \sum_{-m \leq j \leq m, j \neq 0} \frac{\partial (J)}{\partial v_c}$$

$$\frac{\partial J_{skip\_gram}}{\partial v_w} = 0, w \neq c$$