Assignment 2: Word Vectors
March 15, 2021

1. Written

a) Given: $\text{softmax}(x)_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$

Similarly, $\text{softmax}(x+c)_i = \frac{e^{(x+c)_i}}{\sum_j e^{(x+c)_j}} = \frac{e^{x_i} e^{c}}{\sum_j e^{x_j} e^{c}} = \frac{e^{c} e^{x_i}}{e^{c} \sum_j e^{x_j}} = \frac{e^{x_i}}{\sum_j e^{x_j}}$

$= \text{softmax}(x)_i$.

Therefore, $\text{softmax}(x) = \text{softmax}(x+c)$.

b) Given: $\sigma(x) = \frac{1}{1+e^{-x}}$

$$\frac{\partial \sigma(x)}{\partial x} = \frac{\partial}{\partial x}\left(\frac{1}{1+e^{-x}}\right) = \frac{\partial}{\partial x}(1+e^{-x})^{-1} = -(1+e^{-x})^{-2}\frac{\partial}{\partial x}(1+e^{-x})$$

$$= -(1+e^{-x})^{-2}(-e^{-x}) = -(-e^{-x})\left(\frac{1}{(1+e^{-x})^2}\right)$$

$$= \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}}\left(\frac{e^{-x}}{1+e^{-x}}\right) = \frac{1}{1+e^{-x}}\left(\frac{(1+e^{-x})-1}{1+e^{-x}}\right)$$

$$\frac{\partial \sigma(x)}{\partial x} = \frac{1}{1+e^{-x}}\left(1 - \frac{1}{1+e^{-x}}\right)$$

Using the given definition of $\sigma(x)$, this can be expressed as:

$$\frac{\partial \sigma(x)}{\partial x} = \sigma(x)\left(1-\sigma(x)\right).$$

c) i. $J = -\sum_i y_i \log(\hat{y}_i) = -\sum_i y_i \log(p(x_i|c)) = -\log\left(\frac{\exp(u_K^T v_c)}{\sum_{w=1}^{W}\exp(u_w^T v_c)}\right)$ (Because $y_i = 1$).

$$= -\left(\log(\exp(u_K^T v_c)) - \log\left(\sum_{w=1}^{W}\exp(u_w^T v_c)\right)\right) = -(u_K^T v_c) - \log\left(\sum_{w=1}^{W}\exp(u_w^T v_c)\right)$$

$$\frac{\partial J}{\partial v_c} = -\left(\frac{\partial}{\partial v_c}(u_K^T v_c) - \frac{\partial}{\partial v_c}\log\left(\sum_{w=1}^{W}\exp(u_w^T v_c)\right)\right)$$

$$\frac{\partial}{\partial v_c}(u_K^T v_c) = u_K$$

Continued on next page

$$\frac{\partial}{\partial v_c} \log\left(\sum_{w=1}^{W} \exp(u_w^T v_c)\right) = \frac{1}{\sum_{w=1}^{W} \exp(u_w^T v_c)} \cdot \sum_{x=1}^{W} \frac{\partial}{\partial v_c} \exp(u_x^T v_c)$$

$$= \frac{1}{\sum_{w=1}^{W} \exp(u_w^T v_c)} \cdot \sum_{x=1}^{W} \exp(u_x^T v_c) \cdot \frac{\partial}{\partial v_c}(u_x^T v_c)$$

$$= \frac{1}{\sum_{w=1}^{W} \exp(u_w^T v_c)} \cdot \sum_{x=1}^{W} \exp(u_x^T v_c) \cdot u_x$$

$$= \sum_{x=1}^{W} \frac{\exp(u_x^T v_c)}{\sum_{w=1}^{W} \exp(u_w^T v_c)} \cdot u_x = \sum_{x=1}^{W} p(u_x | v_c) \cdot u_x = \sum_{x=1}^{W} \hat{y}_x u_x$$

therefore,

$$\frac{\partial J}{\partial v_c} = -\left(u_k - \sum_{x=1}^{W} \hat{y}_x u_x\right) = -u_k + \sum_{x=1}^{W} \hat{y}_x u_x$$

$$= -u \cdot y + u \cdot \hat{y} = u(-y + \hat{y}) = u(\hat{y} - y).$$

c) ii. From part i,

$$J = -\left(\log(\exp(u_o^T v_c)) - \log\left(\sum_{w=1}^{W} \exp(u_w^T v_c)\right)\right)$$

$$= -(u_o^T v_c) + \log\left(\sum_{w=1}^{W} \exp(u_w^T v_c)\right)$$

consider two cases for the value of u_w:

① $w = o$

$$\frac{\partial}{\partial u_w} -(u_o^T v_c) = -v_c$$

$$\frac{\partial}{\partial u_w} \log\left(\sum_{w=1}^{W} \exp(u_w^T v_c)\right) = \frac{1}{\sum_{w=1}^{W} \exp(u_w^T v_c)} \cdot \sum_{x=1}^{W} \frac{\partial}{\partial u_w} \exp(u_x^T v_c)$$

$$= \frac{1}{\sum_{w=1}^{W} \exp(u_w^T v_c)} \cdot \sum_{x=1}^{W} \exp(u_x^T v_c) \cdot v_c$$

$$= \sum_{x=1}^{W} \frac{\exp(u_x^T v_c) \cdot v_c}{\sum_{w=1}^{W} \exp(u_w^T v_c)} = \sum_{x=1}^{W} \hat{y}_x v_c$$

$$\frac{\partial J}{\partial u_w}\Big|_{w=o} = -v_c + \sum_{x=1}^{W} \hat{y}_x v_c = v_c(1 - \hat{y}_w)$$

continued on next page

② $w \neq 0$

$$\frac{\partial}{\partial u_w} - (u_0^T v_c) = 0 \qquad \frac{\partial}{\partial u_w} \log\left(\sum_{w=1}^{W} \exp(u_w^T v_c)\right) = \frac{1}{\sum_{w=1}^{W} \exp(u_w^T v_c)} \cdot \sum_{x=1}^{W} \frac{\partial}{\partial u_w} \exp(u_x^T v_c)$$

$$= \frac{1}{\sum_{w=1}^{W} \exp(u_w^T v_c)} \cdot \sum_{x=1}^{W} \exp(u_x^T v_c) \cdot v_c$$

$$= \sum_{x=1}^{W} \frac{\exp(u_x^T v_c)}{\sum_{w=1}^{W} \exp(u_w^T v_c)} \cdot v_c = \sum_{x=1}^{W} \hat{y}_x v_c$$

$$\frac{\partial J}{\partial u_w}\Big|_{w \neq 0} = 0 + \sum_{x=1}^{W} \hat{y}_x v_c = \hat{y}_w v_c$$

Therefore,

$$\frac{\partial J}{\partial u_w} = \begin{cases} (1-\hat{y}_w) v_c, & w = 1 \\ \hat{y}_w v_c, & w \neq 0 \end{cases}.$$

c) iii. $J = -\log(\sigma(u_0^T v_c)) - \sum_{k=1}^{K} \log(\sigma(-u_k^T v_c))$

i) $\frac{\partial J}{\partial v_c} = \frac{\partial}{\partial v_c}\left(-\log(\sigma(u_0^T v_c))\right) - \frac{\partial}{\partial v_c}\left(\sum_{k=1}^{K} \log(\sigma(-u_k^T v_c))\right)$

$$\frac{\partial}{\partial v_c}\left(-\log(\sigma(u_0^T v_c))\right) = -\frac{1}{\sigma(u_0^T v_c)} \cdot \frac{\partial}{\partial v_c}(\sigma(u_0^T v_c)) = -\frac{1}{\sigma(u_0^T v_c)} \cdot \sigma(u_0^T v_c)(1-\sigma(u_0^T v_c)) \cdot \frac{\partial}{\partial v_c}(u_0^T v_c)$$

$$= -(1-\sigma(u_0^T v_c)) u_0 = (\sigma(u_0^T v_c) - 1) u_0$$

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

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

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

Therefore,

$$\frac{\partial J}{v_c} = (\sigma(u_0^T v_c) - 1) u_0 + \sum_{k=1}^{K} (1-\sigma(-u_k^T v_c)) u_k.$$

↓ continued on next page.

ii) Consider two cases for the value of u_w:

① $w=0$

$$\frac{\partial}{\partial u_w}\left(-\log\left(\sigma(u_0^T v_c)\right)\right) = -\frac{1}{\sigma(u_0^T v_c)} \cdot \sigma(u_0^T v_c)\left(1-\sigma(u_0^T v_c)\right) \cdot \frac{\partial}{\partial u_w}(u_0^T v_c)$$

$$= -\left(1-\sigma(u_0^T v_c)\right) v_c = \left(\sigma(u_0^T v_c)-1\right) v_c$$

$$\frac{\partial}{\partial u_w}\left(\sum_{k=1}^{K} \log\left(\sigma(-u_k^T v_c)\right)\right) = \sum_{k=1}^{K} \frac{1}{\sigma(-u_k^T v_c)} \cdot \sigma(-u_k^T v_c)\left(1-\sigma(-u_k^T v_c)\right) \cdot \underbrace{\frac{\partial}{\partial u_w}(-u_k^T v_c)}_{=0}$$

$$\frac{\partial J}{\partial u_w}\Big|_{w=0} = \left(\sigma(u_0^T v_c)-1\right) v_c$$

② $w \neq 0$ (or, $w=k$)

$$\frac{\partial}{\partial u_w}\left(-\log\left(\sigma(u_0^T v_c)\right)\right) = -\frac{1}{\sigma(u_0^T v_c)} \cdot \sigma(u_0^T v_c)\left(1-\sigma(u_0^T v_c)\right) \cdot \underbrace{\frac{\partial}{\partial u_w}(u_0^T v_c)}_{=0}$$

$$\frac{\partial}{\partial u_w}\left(-\sum_{k=1}^{K} \log\left(\sigma(-u_k^T v_c)\right)\right) = -\sum_{k=1}^{K} \frac{1}{\sigma(-u_k^T v_c)} \cdot \sigma(-u_k^T v_c)\left(1-\sigma(-u_k^T v_c)\right) \cdot \frac{\partial}{\partial u_w}(-u_k^T v_c)$$

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

$$\frac{\partial J}{\partial u_w}\Big|_{w\neq 0} = \sum_{k=1}^{K} \left(1-\sigma(-u_k^T v_c)\right) v_c$$

Therefore,

$$\frac{\partial J}{\partial u_w} = \begin{cases} \left(\sigma(u_0^T v_c)-1\right) v_c, & w=0 \\ \sum_{k=1}^{K} \left(1-\sigma(-u_k^T v_c)\right) v_c, & w \neq 0. \end{cases}$$

v. $J_{\text{skip-gram}}(\text{word}_{c-m}, \ldots, \text{word}_c, \ldots, \text{word}_{c+m}) = \sum_{-m \leq j \leq m, j \neq 0} F(w_{c+j}, v_c)$

Use $F(o, v_c)$ as placeholder for JCE or $J_{\text{neg-sample}}$.

$$\frac{\partial J_{\text{skip-gram}}}{\partial v_c} = \sum_{-m \leq j \leq m, j \neq 0} \frac{\partial F(o, v_c)}{\partial v_c} \qquad \frac{\partial J_{\text{skip-gram}}}{\partial u_k} = \sum_{-m \leq j \leq m, j \neq 0} \frac{\partial F(o, v_c)}{\partial u_k}$$