Assignment 2

1. Written: Understanding word2vec

(a) $-\sum_{w \in Vocab} y_w \log(\hat{y}_w) = -\underset{\overset{\shortparallel}{0}}{y_{w_0}} \log(\hat{y}_{w_0}) - \dots - \underset{\overset{\shortparallel}{1}}{y_o} \log(\hat{y}_o) - \dots - \underset{\overset{\shortparallel}{0}}{y_{w_n}} \log(\hat{y}_{w_n}) = -\log(\hat{y}_o)$

(b) $J_{naive\text{-}softmax}(v_c, o, U)$

$= -\log P(O=o \mid C=c)$

$= -\log\left(\dfrac{\exp(u_o^T v_c)}{\sum_{w \in Vocab} \exp(u_w^T v_c)}\right)$

$= -\log(\exp(u_o^T v_c)) + \log \sum_w \exp(u_w^T v_c)$

$= -u_o^T v_c + \log \sum_w \exp(u_w^T v_c)$

$\therefore \dfrac{\partial J}{\partial v_c} = -u_o + \dfrac{1}{\sum_w \exp(u_w^T v_c)} \dfrac{\partial}{\partial v_c} \sum_x \exp(u_x^T v_c)$

$= -u_o + \dfrac{1}{\sum_w \exp(u_w^T v_c)} \sum_x u_x \cdot \exp(u_x^T v_c)$

$= -u_o + \sum_x u_x \cdot \dfrac{\exp(u_x^T v_c)}{\sum_w \exp(u_w^T v_c)}$

$= -u_o + \sum_x u_x \cdot \hat{y}_x$

$= -U^T y + U^T \hat{y}$

$= U^T(\hat{y} - y)$

(c) when $w = o$,

$\dfrac{\partial J}{\partial u_w} = -v_c + \dfrac{1}{\sum_w \exp(u_w^T v_c)} \exp(u_o^T v_c)\, v_c$

$= -v_c + \hat{y}_o \cdot v_c$

$= (\hat{y}_o - 1) \cdot v_c$

when $w \neq o$,

$\dfrac{\partial J}{\partial u_w} = \dfrac{\partial}{\partial u_w} \log \sum_w \exp(u_w^T v_c)$

$= \dfrac{1}{\sum_w \exp(u_w^T v_c)} \cdot \sum_{x \neq o} \exp(u_x^T v_c) \cdot v_c^T$

$$= \sum_{x \neq o} \frac{\exp(u_x^T v_c)}{\sum_w \exp(u_w^T v_c)} \cdot v_c^T$$

$$= \left(\sum_{x \neq o} \hat{y}_x\right) \cdot v_c^T$$

$$\therefore \frac{\partial J}{\partial u_w} = \begin{cases} (\hat{y}_o - 1) \cdot v_c^T, & \text{if } w = o \\ \left(\sum_{w \neq o} \hat{y}_w\right) \cdot v_c^T, & \text{otherwise} \end{cases}$$

$$= (\hat{y} - y) \cdot v_c^T$$

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

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

$$= -(1+e^{-x})^{-2} \cdot (-e^{-x})$$

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

$$= \frac{e^{-x}}{1+e^{-x}} \cdot \frac{1}{1+e^{-x}}$$

$$= (1 - \sigma(x))\sigma(x)$$

(e) $J_{\text{neg-sample}}(v_c, o, U) = -\log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} \log(\sigma(-u_k^T v_c))$

① Derive $\partial J / \partial v_c$

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

$$= -\cancel{\frac{1}{\sigma(u_o^T v_c)}} \cdot \cancel{\sigma(u_o^T v_c)}(1 - \sigma(u_o^T v_c)) \cdot u_o$$

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

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

② Derive $\partial J / \partial u_o$

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

$$= (\sigma(u_o^T v_c) - 1) \cdot v_c - 0 \quad (\text{since } o \notin K)$$

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

③ Derive $\partial J / \partial u_k$

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

$$= \sigma(u_k^T v_c) \cdot v_c$$

(f) $J_{skip\text{-}gram}(v_c, w_{t-m}, \ldots, w_{t+m}, U) = \sum_{\substack{-m \le j \le m \\ j \ne 0}} J(v_c, w_{t+j}, U)$

(i) $\partial J_{skip\text{-}gram}(v_c, w_{t-m}, \ldots, w_{t+m}, U) / \partial U = \sum_{\substack{-m \le j \le m \\ j \ne 0}} J(v_c, w_{t+j}, U) / \partial U$

(ii) $\partial J_{skip\text{-}gram}(v_c, w_{t-m}, \ldots, w_{t+m}, U) / \partial v_c = \sum_{\substack{-m \le j \le m \\ j \ne 0}} J(v_c, w_{t+j}, U) / \partial v_c$

(iii) $\partial J_{skip\text{-}gram}(v_c, w_{t-m}, \ldots, w_{t+m}, U) / \partial v_w = 0$, if $w \ne c$