CS224N HW2

1. Understanding word2vec.

(a) For all $w \neq 0$, $y_w = 0$; For $w=0$ $y_w = y_o = 1$; thus:

$$-\sum_{w \in V} y_w \log(\hat{y}_w) = 0 - 1 \cdot \log(\hat{y}_o) = -\log(\hat{y}_o)$$

(b) $$J_{naive\text{-}softmax}(v_c, o, U) = -\log\left(\frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}\right)$$

$$\frac{\partial}{\partial v_c} J_{naive\text{-}softmax}(v_c, o, U) = -\frac{\exp(u_o^T v_c)\cdot u_o}{\exp(u_o^T v_c)} + \frac{\sum_{w \in V} \exp(u_w^T v_c) u_w}{\sum_{w \in V} \exp(u_w^T v_c)}$$

$$= -u_o + \sum_{w \in V} (\hat{y}_w \cdot u_w)$$

(1) Note that $u_o = U \cdot y$, where $U = [u_1, u_2, \cdots, u_V]$ and y is true label for word v_c and o.

(2) Note that $\sum_{w \in V} (\hat{y}_w \cdot u_w) = U \cdot \hat{y}$, where $\hat{y}$ is predicted values for each word

$$\therefore \frac{\partial}{\partial v_c} J_{naive\text{-}softmax}(v_c, o, U) = -U \cdot y + U\hat{y}$$
$$= U(\hat{y} - y)$$

(c) $$J_{naive\text{-}softmax}(v_c, o, U) = -\log(\exp(u_o^T v_c)) + \log\left(\sum_{w \in V} \exp(u_w^T v_c)\right)$$

- Case 1. when $u_w \neq u_o$:

$$\frac{\partial}{\partial u_w} J_{naive\text{-}softmax}(v_c, o, U) = \frac{\exp(u_w^T v_c)\cdot v_c}{\sum_{x \in V} \exp(u_x^T v_c)}$$

$$= \hat{y}_w \cdot v_c \quad (d,)$$

- Case 2. when $u_w = u_o$:

$$\frac{\partial}{\partial u_o} J_{naive\text{-}softmax}(v_c, o, U) = -\frac{\exp(u_o^T v_c)\cdot v_c}{\exp(u_o^T v_c)} + \frac{\exp(u_o^T v_c)\, v_c}{\sum_{w \in V} \exp(u_w^T v_c)}$$

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

(d) $\frac{\partial}{\partial U} J_{naive\text{-}softmax}(v_c, o, U) = \left[\frac{\partial J(v_c,o,U)}{\partial u_1}, \frac{\partial J(v_c,o,U)}{\partial u_2}, \cdots, \frac{\partial J(v_c,o,U)}{\partial u_{|vocab|}} \right]$

(e) Let $z = e^{-x}$, then:

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

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

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

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

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

② $$\frac{\partial}{\partial u_o} J(v_c, o, U) = -\frac{\sigma(u_o^T v_c)}{\sigma(u_o^T v_c)} (1-\sigma(u_o^T v_c)) \cdot v_c^T = -(1-\sigma(u_o^T v_c)) \cdot v_c^T$$

③ $$\frac{\partial}{\partial u_k} J(v_c, o, U) = -\frac{\sigma(-u_k^T v_c)}{\sigma(-u_k^T v_c)} (1-\sigma(-u_k^T v_c)) \cdot (-v_c^T) = (1-\sigma(-u_k^T v_c)) \cdot v_c^T$$

(g) $$\frac{\partial}{\partial u_k} J(v_c, o, U) = \sum_{u_w = u_k} \frac{\sigma(-u_w^T v_c)}{\sigma(-u_w^T v_c)} (1-\sigma(-u_w^T v_c)) \cdot (-v_c^T)$$

Assume there are n samples that equals to u_k.

$$= n(1-\sigma(-u_k^T v_c)) \cdot (-v_c^T)$$

(h)

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

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

(iii) $\frac{\partial}{\partial v_w} J_{skip\text{-}gram} = \sum_{\substack{-m \le j \le m \\ j \ne 0 \\ \text{if } t+j = w}} \frac{\partial}{\partial v_w} J(v_c, w_{t+j}, U)$

(C) Some analogies can be seen like "man" - "king" v.s "woman" - "queen". Some clusters of adj., such as "great", "wonderful" and "amazing" are clustered together. Positive adjectives and negative ones like "boring" are not divided in low dimension visualization.