CS 224N Assignment #2

1. Written: Understanding word2vec

(a) Note $\vec{y}$ is a one-hot vector with one at index $w = o$

$$-\sum_{w \in V} y_w \log \hat{y}_w = -(y_1 \log \hat{y}_1 + y_2 \log \hat{y}_2 + \cdots + y_{|V|} \log \hat{y}_{|V|})$$

$$= -\log \hat{y}_o$$

(b) $J_{native\text{-}softmax}(v_c, o, U) = -\log P(O = o \mid C = c)$

$$= -\log \frac{e^{u_o^T v_c}}{\sum_{w \in Vocab} e^{u_w^T v_c}}$$

$$= -u_o^T v_c + \log \sum_{w \in Vocab} e^{u_w^T v_c}$$

$$\frac{\partial J_{native}}{\partial v_c} = -\frac{\partial}{\partial v_c} u_o^T v_c + \frac{\partial}{\partial v_c} \log \sum_{w \in Vocab} e^{u_w^T v_c}$$

$$= -u_o + \frac{1}{\sum_{w \in Vocab} e^{u_w^T v_c}} \cdot \sum_{w \in Vocab} u_w \cdot e^{u_w^T v_c}$$

$$= -u_o + \sum_{w \in Vocab} \frac{e^{u_w^T v_c}}{\sum_{w \in Vocab} e^{u_w^T v_c}} \cdot u_w$$

$$= -u_o + \sum_{w \in Vocab} P(u_w \mid v_c) \cdot u_w$$

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

(c) If $w = o$

$$\frac{\partial J_{native}}{\partial u_o} = -\frac{\partial}{\partial u_o} u_o^T v_c + \frac{\partial}{\partial u_o} \log \sum_{w \in Vocab} e^{u_w^T v_c}$$

$$= -v_c + \frac{1}{\sum_{w \in Vocab} e^{u_w^T v_c}} \cdot \frac{\partial}{\partial u_o} \sum_{w \in Vocab} e^{u_w^T v_c}$$

$$= -v_c + \frac{1}{\sum_{w \in Vocab} e^{u_w^T v_c}} \cdot e^{u_o^T v_c} \cdot v_c$$

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

else (i.e., $w \neq o$)

$$\frac{\partial J_{native}}{\partial u_w} = -\frac{\partial}{\partial u_w} u_o^T v_c + \frac{\partial}{\partial u_w} \log \sum_{x \in Vocab} e^{u_x^T v_c}$$

$$= 0 + \frac{1}{\sum_{x \in Vocab} e^{u_x^T v_c}} \frac{\partial}{\partial u_w} \sum_{x \in Vocab} e^{u_x^T v_c}$$

$$= \frac{1}{\sum_{x \in Vocab} e^{u_x^T v_c}} e^{u_w^T v_c} \cdot v_c$$

$$= v_c \hat{y}_w$$

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

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

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

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

$$= \sigma(x)\left[1-\sigma(x)\right)$$

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

$$= -\log \frac{1}{1+e^{-u_o^T v_c}} - \sum_{k=1}^{K} \log \frac{1}{1+e^{u_k^T v_c}}$$

compute partial derivative for v_c

$$\frac{\partial J_{neg}}{\partial v_c} = -\frac{\partial}{\partial v_c} \log \frac{1}{1+e^{-u_o^T v_c}} - \sum_{k=1}^{K} \frac{\partial}{\partial v_c} \log \frac{1}{1+e^{u_k^T v_c}}$$

$$= -\frac{1}{\sigma(u_o^T v_c)} \cdot \sigma(u_o^T v_c) \cdot \left[1-\sigma(u_o^T v_c)\right] \cdot u_o$$

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

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

Compute partial derivative for u_o

$$\frac{\partial J_{neg}}{\partial u_o} = -\frac{1}{\sigma(u_o^T v_c)} \cdot \sigma(u_o^T v_c) \cdot [1 - \sigma(u_o^T v_c)] \cdot v_c$$

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

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

Compute partial derivative for u_k

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

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

$$= -\frac{1}{\sigma(-u_k^T v_c)} \cdot \sigma(-u_k^T v_c) \cdot [1 - \sigma(-u_k^T v_c)] \cdot (-v_c)$$

$$= v_c [1 - \sigma(-u_k^T v_c)]$$

For the naive softmax loss function, we will have to use all the word vectors in the vocabulary to normalize the probabilities. In contrast, with negative sampling, we just need to take K word vectors into account.

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

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

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

(iii) $$\frac{\partial J_{\text{skip-gram}}}{\partial v_w} = 0$$