CS224N Assignment #2 : Word2Vec

1. Understanding Word2Vec

(a) (3 points) Show that the naive-softmax loss given in Equation (2) is the same as the cross-entropy loss between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$; i.e., show that

$$-\sum_{w \in Vocab} y_w \log(\hat{y}_w) = -\log(\hat{y}_o). \tag{3}$$

$$-\sum_{w \in Vocab} y_w \log(\hat{y}_w) = -\{ y_1 \log(\hat{y}_1) + \cdots + y_o \log(\hat{y}_o) + \cdots + y_k \log(\hat{y}_k) \} = -y_o \log(\hat{y}_o)$$

$y=0$ 이 되는 주변 단어 값들은 다 지워지고, $y=1$ 인 값을 갖는 중심 단어만 남게 되므로 식은 성립한다.

(b) (5 points) Compute the partial derivative of $\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})$ with respect to $\boldsymbol{v}_c$. Please write your answer in terms of $\boldsymbol{y}$, $\hat{\boldsymbol{y}}$, and $\boldsymbol{U}$.

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

center word 가 주어졌을 때, outside word 가 나올 확률

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

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

$$\frac{\partial J}{\partial v_c} = -u_o^T + \sum_w \frac{\exp(u_w^T v_c) \cdot u_w^T}{\sum_w \exp(u_w^T v_c)}$$

$$= -u_o + \sum_w P(O=w \mid C=c) \cdot u_w$$

$$= -u_o + \sum_w \hat{y}_w u_w$$

$\because$ answer (a) : 답에 대해서만 1이고, 나머지 값은 모두 0

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

(c) (5 points) Compute the partial derivatives of $\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})$ with respect to each of the 'outside' word vectors, $\boldsymbol{u}_w$'s. There will be two cases: when $w = o$, the true 'outside' word vector, and $w \neq o$, for all other words. Please write you answer in terms of $\boldsymbol{y}$, $\hat{\boldsymbol{y}}$, and $\boldsymbol{v}_c$.

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

i) $w=o$ outside word 가 실제 정답인 경우 (correct class)

$$\frac{\partial J}{\partial u_w} = -v_c + \frac{\exp(u_w^T v_c)}{\sum_w \exp(u_w^T v_c)} \cdot v_c$$

$$= -v_c + P(O=w \mid C=c) \cdot v_c$$

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

ii) $w \neq o$ outside word 가 실제 정답이 아닌 경우 (incorrect class)

$$\frac{\partial J}{\partial u_w} = 0 + \frac{\exp(u_w^T v_c)}{\sum_w \exp(u_w^T v_c)} \cdot v_c$$

$$= P(O=w \mid C=c) \cdot v_c$$

$$= \hat{y} v_c$$

(d) (3 Points) The sigmoid function is given by Equation 4:

$$\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1} \tag{4}$$

Please compute the derivative of $\sigma(x)$ with respect to x, where x is a scalar. Hint: you may want to write your answer in terms of $\sigma(x)$.

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

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

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

(e) (4 points) Now we shall consider the Negative Sampling loss, which is an alternative to the Naive Softmax loss. Assume that K negative samples (words) are drawn from the vocabulary. For simplicity of notation we shall refer to them as $w_1, w_2, \dots, w_K$ and their outside vectors as $\boldsymbol{u}_1, \dots, \boldsymbol{u}_K$. Note that $o \notin \{w_1, \dots, w_K\}$. For a center word c and an outside word o, the negative sampling loss function is given by:

$$\boldsymbol{J}_{\text{neg-sample}}(\boldsymbol{v}_c, o, \boldsymbol{U}) = -\log(\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)) - \sum_{k=1}^{K} \log(\sigma(-\boldsymbol{u}_k^\top \boldsymbol{v}_c)) \tag{5}$$

for a sample $w_1, \dots w_K$, where $\sigma(\cdot)$ is the sigmoid function.[3]

Please repeat parts (b) and (c), computing the partial derivatives of $\boldsymbol{J}_{\text{neg-sample}}$ with respect to $\boldsymbol{v}_c$, with respect to $\boldsymbol{u}_o$, and with respect to a negative sample $\boldsymbol{u}_k$. Please write your answers in terms of the vectors $\boldsymbol{u}_o$, $\boldsymbol{v}_c$, and $\boldsymbol{u}_k$, where $k \in [1, K]$. After you've done this, describe with one sentence why this loss function is much more efficient to compute than the naive-softmax loss. Note, you should be able to use your solution to part (d) to help compute the necessary gradients here.

i) $\dfrac{\partial J}{\partial v_c} = -\dfrac{1}{\cancel{\sigma(u_o^T v_c)}} \cdot \cancel{\sigma(u_o^T v_c)} \cdot \{1-\sigma(u_o^T v_c)\} \cdot u_o^T$

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

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

ii) 모든 k에 대해 $\dfrac{\partial}{\partial u_o} \log(\sigma(-u_k^T u_c)) = 0 \quad (\because o \notin \{w_1, \cdots, w_k\})$

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

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

iii) $\dfrac{\partial}{\partial u_k} \log(\sigma(u_o^T u_c)) = 0 \quad (\because o \notin \{w_1, \cdots, w_k\})$

$\dfrac{\partial J}{\partial u_k} = -\displaystyle\sum_{k=1}^{K} \dfrac{1}{\cancel{\sigma(-u_k^T v_c)}} \cdot \cancel{\sigma(-u_k^T v_c)} \cdot \{1-\sigma(-u_k^T \cdot v_c)\} \cdot (-v_c)$

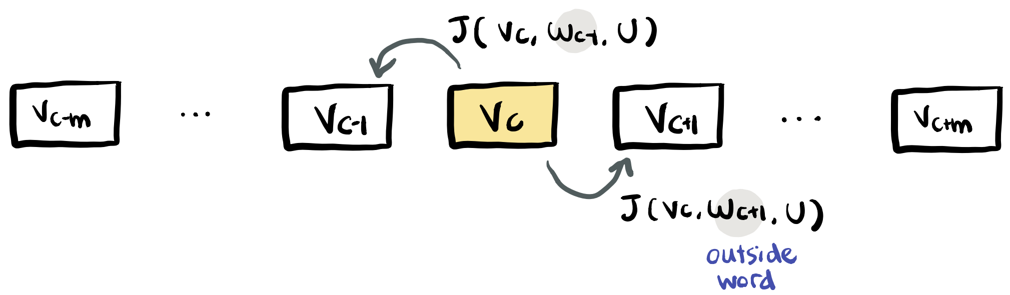
$= \displaystyle\sum_{k=1}^{K} \{1-\sigma(-u_k^T \cdot v_c)\} \cdot v_c$

(f) (3 points) Suppose the center word is $c = w_t$ and the context window is $[w_{t-m}, \ldots, w_{t-1}, w_t, w_{t+1}, \ldots, w_{t+m}]$, where m is the context window size. Recall that for the skip-gram version of `word2vec`, the total loss for the context window is:

$$\boldsymbol{J}_{\text{skip-gram}}(\boldsymbol{v}_c, w_{t-m}, \ldots w_{t+m}, \boldsymbol{U}) = \sum_{\substack{-m \le j \le m \\ j \ne 0}} \boldsymbol{J}(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U}) \tag{6}$$

Here, $\boldsymbol{J}(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U})$ represents an arbitrary loss term for the center word $c = w_t$ and outside word w_{t+j}. $\boldsymbol{J}(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U})$ could be $\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U})$ or $\boldsymbol{J}_{\text{neg-sample}}(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U})$, depending on your implementation.

Write down three partial derivatives:



(i) $\frac{\partial J}{\partial U} = \sum \frac{\partial J(V_c, W_{t+j}, U)}{\partial U}$

(ii) $\frac{\partial J}{\partial V_c} = \sum \frac{\partial J(V_c, W_{t+j}, U)}{\partial V_c}$

(iii) $\frac{\partial J}{\partial W_c} = 0$ center word 가 아닌 경우 update 되지 않기 때문