Assignment2

April 15, 2020

1 Question(a)

We know that:

$$y_w = \begin{cases} 1 & w \text{ shown in context of } o \\ 0 & w \text{ not shown in context of } o \end{cases} \tag{1}$$

And since y is a one-hot vector, so only for y_o, it is 1.

$$\text{cross\_entropy}(y, \hat{y}) = \sum_{w \in V} y_w log(\hat{y_w}) = y_o log(\hat{y_o}) \tag{2}$$

2 Question(b)

U is a matrix where each column is u_i for context word i. c is the center word's index, o is the outside word's index.

$$\begin{aligned} J(o, c, U) &= -log \frac{exp(u_o^T v_c)}{\sum_{w \in V} exp(u_w^T v_c)} \\ &= -u_o^T v_c + log(\sum_{w \in V} exp(u_w^T v_c)) \end{aligned} \tag{3}$$

$$\begin{aligned} \frac{\partial J(o, c, U)}{\partial v_c} &= -u_o + \frac{\partial log \sum_{w \in V} exp(u_w^T v_c)}{\partial v_c} \\ &= -u_o + \frac{1}{\sum_{w \in V} exp(u_w^T v_c)} \frac{\partial \sum_{w \in V} exp(u_w^T v_c)}{\partial v_c} \\ &= -u_o + \frac{\sum_{w \in V} u_w exp(u_w^T v_c)}{\sum_{w \in V} exp(u_w^T v_c)} \\ &= -Uy + \sum_{w \in V} u_w \hat{y} \\ &= U(\hat{y} - y) \end{aligned} \tag{4}$$

3 Question(c)

If $w = o$(this word is the outside word), then:

$$\begin{aligned}
\frac{\partial J(o,c,U)}{\partial u_w} &= -v_c + \frac{\partial log \sum_{w \in V} exp(u_w^T v_c)}{\partial u_o} \\
&= -v_c + \frac{1}{\sum_{w \in V} exp(u_w^T v_c)} \frac{\partial \sum_{w \in V} exp(u_w^T v_c)}{\partial u_o} \\
&= -v_c + \frac{v_c exp(u_w^T v_c)}{\sum_{w \in V} exp(u_w^T v_c)} \\
&= -v_c + v_c \hat{y} \\
&= v_c(\hat{y} - y)
\end{aligned} \tag{5}$$

Otheriwise $v_c \hat{y}$

4 Question(d)

$$\begin{aligned}
\frac{d\sigma(x)}{dx} &= \frac{d(1+e^{-x})^{-1}}{dx} \\
&= -(1+e^{-x})^{-2} e^{-x} \\
&= -\frac{1}{(1+e^{-x})(1+e^{x})} \\
&= \sigma(x)\sigma(-x)
\end{aligned} \tag{6}$$

5 Question(e)

$$\begin{aligned}
J_{\text{neg-sample}}(v_o, c, U) &= -log(\sigma(u_o^T v_c)) - \sum_{u_k \in \text{neg}} log(\sigma(-u_k^T v_c)) \\
\frac{\partial J_{\text{neg-sample}}(v_o, c, U)}{\partial v_c} &= -\frac{\partial (log(\sigma(u_o^T v_c))}{\partial v_c} - \sum_{v_k \in \text{neg}} \frac{\partial log(\sigma(-u_k^T v_c))}{\partial v_c} \\
&= -u_o \sigma(-u_o^T v_c) + \sum_{v_k \in \text{neg}} u_k \sigma(u_k^T v_c) \\
&= u_o[\sigma(u_o^T v_c) - 1] + \sum_{v_k \in \text{neg}} u_k \sigma(u_k^T v_c)
\end{aligned} \tag{7}$$

$$\begin{aligned}
\frac{\partial J_{\text{neg-sample}}(v_o, c, U)}{\partial u_k} &= -\frac{\partial (log(\sigma(u_o^T v_c))}{\partial u_k} - \sum_{v_k \in \text{neg}} \frac{\partial log(\sigma(-u_k^T v_c))}{\partial u_k} \\
&= v_c \sigma(u_k^T v_c)
\end{aligned} \tag{8}$$

$$\frac{\partial J_{\text{neg-sample}}(v_o, c, U)}{\partial u_o} = -v_c \sigma(-u_o^T v_c) \tag{9}$$

This negative-sampling loss function is more efficient to compute than naive softmax loss function because here only one vector u_k or v_c or u_o involved, but in naive softmax, we need the whole matrix U to be involved.

6 Question(f)

(i)

$$\begin{aligned}\frac{\partial J_{\text{skip-gram}}(v_o, w_{t-m}, ..., w_{t+m}, U)}{\partial U} &= \sum_{-m\leq j\leq m, j\neq 0} \frac{\partial J(v_c, w_{t+j}, U)}{\partial U} \\ &= \sum_{-m\leq j\leq m, j\neq 0} v_c(\hat{y}_j - y_j)\end{aligned} \tag{10}$$

(ii)

$$\begin{aligned}\frac{\partial J_{\text{skip-gram}}(v_o, w_{t-m}, ..., w_{t+m}, U)}{\partial v_c} &= \sum_{-m\leq j\leq m, j\neq 0} \frac{\partial J(v_c, w_j, U)}{\partial v_c} \\ &= \sum_{-m\leq j\leq m, j\neq 0} U(\hat{y}_j - y_j)\end{aligned} \tag{11}$$

(iii)

$$\frac{\partial J_{\text{skip-gram}}(v_o, w_{t-m}, ..., w_{t+m}, U)}{\partial v_w}(\text{ when } w \neq c) = 0 \tag{12}$$