Name: Bhavnick Minhas

CS 224N

Assignment 2 [Written]

Understanding Word2Vec

Given:

$$P(O=o \mid C=c) = \frac{\exp(u_o^T v_c)}{\sum_{w \in Vocab} \exp(u_w^T v_c)}$$

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

a) To prove:

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

Since, $\bar{y}_o = [0 \cdots \underbrace{1}_{o^{th}\text{ index}} \cdots 0]$

as y_w is just one-hot pointing to the true outer word, hence

$$y_w = \begin{cases} 0 & \text{for } w \neq o \\ 1 & \text{for } w = o \end{cases}$$

$$\Rightarrow \quad -\sum y_w \log(\hat{y}_w) = 0 + 0 \cdots - \log(\hat{y}_o) + 0 \cdots$$

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

b) To get: $\dfrac{\partial J_{naive\text{-}softmax}}{\partial v_c}$

First, lets simplify J_{ns} a bit.

$$J_{ns} = -\log\left(\frac{\exp(u_o^T v_c)}{\sum_w \exp(u_w^T v_c)}\right)$$

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

Then, we have,

$$\frac{\partial J_{ns}}{\partial v_c} = -u_o + \frac{1}{\sum_w \exp(u_w^T v_c)} \cdot \sum_w \exp(u_w^T v_c) \cdot u_w$$

$$= -u_o + \frac{\overset{d\times m}{U} \cdot \exp(\overset{m\times d}{U^T} \overset{d\times 1}{v_c})}{\underset{1\times m}{\vec{1}} \cdot \exp(U^T v_c)}$$

where $U \in \mathbb{R}^{d\times m}$

$v_c \in \mathbb{R}^{d\times 1}$

$\vec{1} = [1, 1, \dots]$

$$= -U\cdot \vec{y} + U\cdot \hat{y}$$

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

c) To compute: $\frac{\partial J_{ns}}{\partial \vec{u}_w}$

for $w \in \{1, \dots m\}$

case I: $w \neq 0$, $u_{w'}$

$$\frac{\partial}{\partial u_{w'}}\left(-u_o^T v_c + \log\left(\sum \exp(u_w^T v_c)\right)\right)$$

$$= 0 + \frac{1}{\sum_w \exp(u_w^T v_c)} \cdot \exp(u_{w'}^T v_c) \cdot v_c$$

$$= \hat{y}_{w'} \cdot \bar{v}_c$$

case II: $w = 0$

$$\frac{\partial}{\partial u_o}\left(-u_o^T v_c + \log\left(\sum \exp(u_w^T v_c)\right)\right)$$

$$= -v_c + \frac{1}{\sum \exp(u_w^T v_c)} \cdot \exp(u_o^T v_c) \cdot v_c$$

$$= -\bar{v}_c + \hat{y}_o \cdot \bar{v}_c$$

$$= \bar{v}_c(\hat{y}_o - y_o)$$

d)
$$\frac{\partial J_{ns}}{\partial U} = \left[\frac{\partial J}{\partial u_0} \; \frac{\partial J}{\partial u_1} \; \cdots \; \frac{\partial J}{\partial u_m} \right]$$

$$= \left[\hat{y}_0 \bar{v}_c , \hat{y}_1 \bar{v}_c \cdots \bar{v}_c(\hat{y}_o - y_o) \cdots \hat{y}_m v_c \right]$$

$$= \underset{d\times 1}{\bar{v}_c} \cdot (\underset{m\times 1}{\vec{\hat{y}} - \vec{y}})^T$$

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

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

$$= \frac{e^{-x}}{(1+e^{-x})^2} = \sigma(x) \left[\frac{e^{-x}}{1+e^{-x}} \right]$$

$$= \sigma(x) \left[\frac{1+e^{-x} - 1}{1+e^{-x}} \right]$$

$$= \sigma(x) \left[1 - \frac{1}{1+e^{-x}} \right]$$

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

f)

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

i) To get: $\dfrac{\partial J_{neg\text{-}sampling}}{\partial v_c}$

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

$$+ \sum_k \frac{1}{\cancel{\sigma(-u_k^T v_c)}} \cancel{\sigma(-u_k^T v_c)} \cdot (1 - \sigma(-u_k^T v_c)) \cancel{+} u_k$$

$$= -1(1 - \sigma(u_o^T v_c))\bar{u}_o + \sum \bar{u}_k (1 - \sigma(-u_k^T v_c))$$

ii) To get: $\dfrac{\partial J_{ns}}{\partial u_o}$

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

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

iii) To get : $\frac{\partial J_{ns}}{\partial u_k}$

$$\frac{\partial J_{ns}}{\partial u_k} = 0 - \frac{1}{\sigma(-u_k^T v_c)} \cdot \sigma(-u_k^T v_c)\left(1-\sigma(-u_k^T v_c)\right) \cdot -v_c$$

$$= \left(1-\sigma(-u_k^T v_c)\right) v_c$$

g) In the case that the indices are distinct, the indices can be devided into two parts:

i) One containing same index k
ii) Other not being k

$$\Rightarrow \frac{\partial J_{ns}}{\partial u_k} = -\sum_{d=1}^{d_k} \frac{1}{\sigma(-u_k^T v_c)} \cdot \sigma(-u_k^T v_c)\left(1-\sigma(-u_k^T v_c)\right) \cdot -v_c$$

$$= d_k\left(1-\sigma(-u_k^T v_c)\right) v_c$$

where d_k are the no. of k^{th} indices in samples.

b) For skip gram, since,

$$J_{sg}(v_c, w_{t-m}, \ldots, w_{t+m}, U)$$

$$= \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} J(v_c, w_{t+j}, U)$$

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

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

iii) $$\frac{\partial J_{sg}}{\partial v_w} = 0$$ [since the term v_w doesn't occur at all in the expression.]