CS224n Assignment #2 : word2vec

1. (a) $$-\sum_{w \in Vocab} y_w \log(\hat{y}_w)$$

$$= -\Big(\sum_{\substack{w \in vocab \\ w \neq o}} 0 \times \log(\hat{y}_w) + 1 \times \log(\hat{y}_o)\Big)$$

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

(b) $$J_{naive\text{-}softmax}(V_c, o, U) = -\log P(O=o \mid C=c) = -\log \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}$$

$$-\frac{\partial}{\partial v_c} \log \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)} = -\frac{\partial}{\partial v_c}\Big(\log \cancel{\exp}(u_o^T v_c) - \log \sum_{w \in V} \exp(u_w^T v_c)\Big)$$

$$= -\frac{\partial}{\partial v_c} u_o^T v_c + \frac{\partial}{\partial v_c} \log \sum_{w \in V} \exp(u_w^T v_c)$$

$$= -u_o + \frac{\sum_{w \in V} \frac{\partial}{\partial v_c} \exp(u_w^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}$$

$$= -u_o + \frac{\sum_{w \in V} \exp(u_w^T v_c)\, u_w}{\sum_{w \in V} \exp(u_w^T v_c)}$$

$$= -u_o + \sum_{w \in V} \frac{\exp(u_w^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)} u_w$$

$$= -u_o + \sum_{w \in V} u_w P(w \mid c)$$

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

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

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

$$-\frac{\partial}{\partial u_w} \log \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}$$

case 1: $w = o$

$$= -\frac{\partial}{\partial u_w} \log \frac{\exp(u_w^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}$$

$$= -\frac{\partial}{\partial u_w} \cancel{\log \exp}(u_w^T v_c) + \frac{\partial}{\partial u_w} \log \sum_{w \in V} \exp(u_w^T v_c)$$

$$= -v_c + \frac{\frac{\partial}{\partial u_w} \sum_{w \in V} \exp(u_w^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}$$

$$= -v_c + \frac{\sum_{w \in V} \exp(u_w^T v_c)\, v_c}{\sum_{w \in V} \exp(u_w^T v_c)}$$

$$= -v_c + \sum_{w \in V} \frac{\exp(u_w^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)} v_c$$

$$= -v_c + \sum_{w \in V} v_c\, p(w|c)$$

$$= -v_c + v_c \cdot \vec{c} \cdot \hat{y}$$

case 2: $w \neq o$

$$= -\frac{\partial}{\partial u_w} \log \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}$$

$$= -\frac{\partial}{\partial u_w} \cancel{\log \exp}(u_o^T v_c) + \frac{\partial}{\partial u_w} \log \sum_{w \in V} \exp(u_w^T v_c)$$

$$= -0 + \frac{\sum_{w \in V} \exp(u_w^T v_c)\, v_c}{\sum_{w \in V} \exp(u_w^T v_c)}$$

$$= -\sum_{w \in V} \frac{\exp(u_w^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)} v_c$$

$$= -\sum_{w \in V} v_c\, p(w|c)$$

$$= -v_c \sum_{w \in V} p(w|c)$$

$$= -v_c \cdot \vec{c} \cdot \hat{y}$$

$\vec{c}$ is a $1 \times V$ vector with a 1 at each position.

(d) $\frac{d}{dx}\left(\frac{e^x}{e^x+1}\right) = \frac{d}{dx} e^x (e^x+1)^{-1} = -e^x(e^x+1)^{-2} e^x + e^x (e^x+1)^{-1}$

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

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

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

$$= \sigma(x)\,(1 - \sigma(x))$$

(e)

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

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

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

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

$$= -\frac{\cancel{\sigma(u_o^T v_c)}(1-\sigma(u_o^T v_c))\,v_c}{\cancel{\sigma(u_o^T v_c)}} - \sum_{k=1}^{k}\frac{0}{\sigma(-u_k^T v_c)}$$

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

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

$$= -\sum_{k=1}^{k}\frac{\cancel{\sigma(-u_k^T v_c)}(1-\sigma(-u_k^T v_c))(-v_c)}{\cancel{\sigma(-u_k^T v_c)}}$$

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

The calculation does not involve any matrix multiplication. In addition, since sigmoid values are calculated during forward propagation, they can be reused here.

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

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

(iii) $$\partial J_{skip\text{-}gram}(v_c, w_{t-m}, \cdots w_{t+m}, U)/\partial v_w = 0$$