CS224n: HW2

2021/3/27

Written part

1.

(a). since the true empirical distribution $\vec{y}$ is by one-hot encoding in the vector $\vec{y}$, only $y_o = 1$, all other $y_w = 0$, $w \in$ Vocabulary

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

(b) $$\frac{\partial J(v_c, o, U)}{\partial v_c} = \frac{\partial \; -u_o^T v_c + \log\left(\sum_{w \in Vocab} \exp(u_w^T v_c)\right)}{\partial v_c}$$

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

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

$$= -u_o + \sum_{w' \in Vocab} \frac{\exp(u_{w'}^T v_c)}{\sum_{w \in Vocab} \exp(u_w^T v_c)} u_{w'}$$

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

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

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

(c) 2 cases can be merged into 1

1) When $w = o$

$$\frac{\partial J(v_c, o, U)}{\partial u_w} = \frac{\partial \; -u_o^T v_c + \log\left(\sum_{w \in Vocab} \exp(u_w^T v_c)\right)}{\partial u_w}$$

$$= -v_c \cdot \mathbb{I}\{w = o\} + P(w \mid c) \cdot v_c$$

$$= v_c\left(P(w|c) - \mathbb{I}\{w=o\}\right)$$

$$= v_c\left(\hat{y}_w - y_w\right)$$

(d) for $\vec{x}^T = [x_1, x_2, \cdots, x_n]$

the sigmoid function σ acts as

$$\sigma(\vec{x})^T = \left[\frac{e^{x_1}}{e^{x_1}+1}, \frac{e^{x_2}}{e^{x_2}+1}, \cdots, \frac{e^{x_n}}{e^{x_n}+1}\right]$$

So $\frac{\partial\, \sigma(\vec{x})_i}{\partial\, x_j} = \mathbb{I}\{i=j\}\, \sigma(x_i)\cdot(1-\sigma(x_i))$

$$= \text{diag}\left[\sigma(x)\cdot(1-\sigma(x))^T\right]$$

notice the dimension here, by taking transpose to get a matrix.

(e) $J_{\text{neg-sample}}(v_c, o, U) = -\log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} \log(\sigma(-u_k^T v_c))$

① $\frac{\partial J_{\text{neg-sample}}(v_c, o, U)}{\partial\, v_c} = \frac{\partial\, -\log(\sigma(u_o^T v_c))}{\partial\, v_c} - \sum_{k=1}^{K} \frac{\partial \log(\sigma(-u_k^T v_c))}{\partial v_c}$

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

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

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

② $\frac{\partial J_{\text{neg-sample}}(v_c, o, U)}{\partial\, u_o} = \frac{\partial\, -\log(\sigma(u_o^T v_c))}{\partial\, u_o}$

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

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

③ $\forall k, \quad \dfrac{\partial J_{neg\text{-}sample}(v_c, o, U)}{\partial u_k} = (1 - \sigma(u_k^T v_c)) \cdot v_c$

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

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

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

(iii) when $w \neq c$

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