center word & window

$$\begin{cases} P(O|C) = P(O=o|C=c) = \dfrac{\exp(u_o^T v_c)}{\sum_{w \in Vocab} \exp(u_w^T v_c)} \\ u_o \in U = \begin{bmatrix} | & & | \\ u_1 & \cdots & u_W \\ | & & | \end{bmatrix} \qquad v_c \in V = \begin{bmatrix} | & & | \\ v_1 & \cdots & v_W \\ | & & | \end{bmatrix} \end{cases}$$

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

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

The mechanism of y_w (one-hot) is to choose the element of the highest probability in $\hat{y}_w$.

(b)

(i) $\nabla_{v_c} J(v_c, o, U) = -\nabla_{v_c} \log \dfrac{\exp(u_o^T v_c)}{\sum_w \exp(u_w^T v_c)}$

$= -\nabla_{v_c} \log \exp(u_o^T v_c) + \nabla_{v_c} \log \sum_w \exp(u_w^T v_c)$

$= -\dfrac{\exp(u_o^T v_c)}{\exp(u_o^T v_c)} \cdot u_o + \dfrac{1}{\sum_w \exp(u_w^T v_c)} \cdot \sum_w \nabla_{v_c} \exp(u_w^T v_c)$

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

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

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

$= -u_o + U^T \hat{y} = U^T(\hat{y} - y)$

(ii) $-u_o + U^T \hat{y} = 0 \Rightarrow U^T \hat{y} = u_o$ (right prediction)

When $\hat{y} = (U^T)^{-1} u_o$, the gradient $= 0$ hold True

(iii) $v_c := v_c + \alpha u_o - \alpha U^T \hat{y}$

$+\alpha u_o$: make the v_c more similar to u_o with LR "α"

$-\alpha U^T \hat{y}$: similar to expected class u_w (soft)

(iv) $u(good) \approx \alpha\, u(best)$

when L_2 - normalization. $\dfrac{u(good)}{\|u\|_2} \approx \dfrac{\alpha \cdot u(best)}{\|\alpha u\|_2} = \dfrac{u(best)}{\|u_2\|_2}$

the semantic change happens.

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

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

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

$= -\nabla_U u_o^T v_c + \sum_w \hat{y}_w \nabla_U \cdot u_w^T v_c$

$= (\sum_w \hat{y}_w \nabla_U u_w^T - \nabla_U \cdot u_o^T) \cdot v_c$

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

(d) $\nabla_U J(v_c, o, U) = \left[\frac{\partial J(v_c, o, U)}{\partial u_1}, \frac{\partial J(v_c, o, U)}{\partial u_2}, \dots, \frac{\partial J(v_c, o, U)}{\partial u_{|vocab|}} \right]$

$= \left[\hat{y}_1^T v_c, \hat{y}_2^T v_c, \dots, (\hat{y}_o - 1) v_c, \dots, \hat{y}_{|vocab|} v_c \right]$ (1: one-hot)

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

(e) Leaky ReLU: $f(x) = \max(\alpha x, x)$

$\frac{\partial}{\partial x} f(x) = \begin{cases} \alpha & x < 0 \\ 1 & x > 0 \end{cases}$

(f) Sigmoid:

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

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

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

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

(g) i) $J_{neg\text{-}sample}(v_c, o, U) = -\log \sigma(u_o^T v_c) - \sum_{s=1}^{k} \log(\sigma(-u_{w_s}^T v_c))$

$\nabla_{v_c} J(v_c, o, U) = -\nabla_{v_c} \log \sigma(u_o^T v_c) - \sum_{s=1}^{k} \nabla_{v_c} \log(\sigma(-u_{w_s}^T v_c))$

$= -\frac{1}{\sigma(u_o^T v_c)} \cdot \sigma(u_o^T v_c)(1-\sigma(u_o^T v_c)) u_o - \sum_{s=1}^{k} \frac{\sigma(-u_{w_s}^T v_c)}{\sigma(-u_{w_s}^T v_c)}(1-\sigma(-u_{w_s}^T v_c)) \cdot -u_{w_s}$

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

$\nabla_{u_o} J(v_c, o, U) = -(1-\sigma(u_o^T v_c)) v_c$

$\nabla_{u_{w_s}} J(v_c, o, U) = \sum_{s=1}^{k} (1-\sigma(-u_{w_s}^T v_c)) v_c$

(ii) $U_{o,\{w_1,\dots,w_k\}} = \begin{bmatrix} u_o \\ -u_{w_1} \\ \vdots \\ -u_{w_k} \end{bmatrix} \in \mathbb{R}^{(k+1)\times d}$ $\quad v_c \in \mathbb{R}^{d\times 1}$

$\Rightarrow \sigma(1 - U_{o,\{w_1,\dots,w_k\}} v_c) \in \mathbb{R}^{(k+1)\times 1}$

$\Rightarrow \nabla_{v_c} J(v_c, o, U) = -U^T_{o,\{w_1,\dots,w_k\}} \sigma(1 - U_{o,\{w_1,\dots,w_k\}} v_c)$ ①

$\nabla_{U_{o,\{w_1,\dots,w_k\}}} J(v_c, o, U) = \frac{U_{o,\{w_1,\dots,w_k\}}}{|U_{o,\{w_1,\dots,w_k\}}|} \odot \underbrace{\sigma(1 - U_{o,\{w_1,\dots,w_k\}} v_c)}_{(k+1)\times 1} \cdot \underbrace{v_c^T}_{1\times d}$

(iii) This function only sample k negative samples, but the naive softmax loss need to calculate hole dataset.

(h) $J_{neg\text{-}sample}(v_c, o, U) = -\log(\sigma(u_o^T v_c)) - \sum_{s=1}^{k} \log(\sigma(-u_{w_s}^T v_c))$

from g(i), we have $\nabla_{u_{w_s}} J_{neg\text{-}sample}(v_c, o, U) = (1 - \sigma(-u_{w_s}^T v_c)) v_c$

for not distinct case: ∇ = sum of all $w = w_s$

$\hookrightarrow \nabla_{u_{w_s}} J_{neg\text{-}sample}(v_c, o, U) = \sum_{\substack{w\in V \\ w=w_s}} (1 - \sigma(u_{w_s}^T v_c)) v_c$

$= \text{num\_non-distinct} \times (1 - \sigma(u_{w_s}^T v_c)) v_c$

$= |\{w \in \text{Vocab} \mid w = w_s\}| (1 - \sigma(u_{w_s}^T v_c)) v_c$

(i)

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

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

$$\frac{\partial J_{skip\text{-}gram}(v_c, w_{t-m}, \dots w_{t+m}, U)}{\partial u_w} = 0$$