CS224n Assignment #2: word2vec

1 Written: Understanding Word2vec

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

y_w is a scalar, y is a vector *

y_w 是 y 中第 w 个词对应的标量

$\hat{y}_w$ 是 y 中第 w 个词计算出的 $P(w|o)$.

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

(b). $\frac{\partial}{\partial v_c} J_{naive\text{-}softmax}(v_c, o, U)$ answers in terms of $y, \hat{y}, U$

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

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

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

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

$= -u_o + \sum_{x \in V} P(x|c)\, u_x$

$= -y_o^T \cdot U^T + \sum \hat{y}_x^T \cdot U^T$

$= \sum_{x \in V} (\hat{y}_x - y_o)^T \cdot U^T$

rows/columns of U: outside vectors u_w

$y = U v_c$

$u_o = U^T \cdot y$

$P(x|c) = \hat{y}_x$

$\sum_{x \in V} P(x|c) \cdot u_x = U^T \hat{y}$

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

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

$(y_1, y_2, \ldots v)$

(c). $\frac{\partial}{\partial u_w} J_{naive\text{-}softmax}(v_c, o, U)$ answers: in terms of $y, \hat{y}, v_c$.

① $w = o$

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

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

$= -v_c + \frac{1}{\sum_{w \in V} \exp(u_w^T v_c)} \cdot \sum_{\substack{w \in V \\ w \neq o}} \exp(u_w^T v_c) \cdot \frac{\partial e^{u_o^T v_c}}{\partial u_o}$

daoben®

$$= -v_c + \frac{1}{\sum_{w\in V}\exp(u_w^T v_c)} \cdot \sum_{\substack{w\in V \\ w\neq o}} 0$$

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

$$= -v_c + \sum_{w\in V} P(w|c)\cdot 1$$

$$= -v_c + \sum_{w\in Vocab} \hat{y}_w v_c$$

$P(o|c)$

(d). $\frac{\partial J(v_c, o, U)}{\partial U}$ answer

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

$$= \left(\frac{\partial J}{\partial u_1}, \frac{\partial J}{\partial u_2}, \cdots \frac{\partial J}{\partial u_{|Vocab|}}\right)^T$$

$$= \left(\hat{y}_1 \cdot v_c, \hat{y}_2 \cdot v_c, \cdots \hat{y}_{|Vocab|} \cdot v_c\right)^T = \cancel{(\hat{y}\cdot v_c)}$$

$(\hat{y}) \cdot v_c$

(e). $\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}$ x is a scalar. answer : $\sigma(x)$ in terms of

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

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

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

answer in terms of u_o, v_c, u_k

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

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

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

neg-sample 's loss function is more efficient than naive-softmax cuz no $\hat{y}$ ($\sum$)

② $w \neq o$. 又 $u_x = u_w$. $x \neq o$.

$$\frac{\partial}{\partial u_x} J_{naive\text{-}softmax}(v_c, o, U)$$

$$J_{naive\text{-}softmax} = -u_o^T v_c + \log\sum_{w\in V}\exp(u_w^T v_c)$$

$$= -u_o^T v_c + \log\left[\exp(u_x^T v_c) + \sum_{\substack{w\in V \\ w\neq x \\ w\neq o}}\exp(u_w^T v_c)\right]$$

$$\frac{\partial J}{\partial u_x} = \frac{\partial \log\left[\exp(u_x^T v_c) + \sum_{\substack{w\in V \\ w\neq x \\ w\neq o}}\exp(u_w^T v_c)\right]}{\partial u_x}$$

$$= \frac{1}{\log\sum_{w\in V}\exp(u_w^T v_c)}\cdot\exp(u_x^T v_c)\cdot v_c$$

$$= P(x|c)\cdot v_c$$

$$= \hat{y}_x \cdot v_c$$

又 $\hat{y}_w \cdot v_c$.

(g). $\frac{\partial J_{neg\text{-}sample}}{\partial u_k}$ answers in terms of v_c, u_k, $k \in [1, K]$

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

$$\frac{\partial J_{neg\text{-}sample}}{\partial u_k} = [\sigma(-u_k^T v_c) - 1] \cdot v_c$$

skip-gram of word2vec

(h). suppose $c = w_t$. context window size $= m$.

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

$J(v_c, w_{t+j}, U)$: $J_{naivesoftmax}(v_c, w_{t+j}, U)$ or $J_{neg\text{-}sample}(v_c, w_{t+j}, U)$

answers in terms of $\frac{\partial J(v_c, w_{t+j}, U)}{\partial U}$ 、 $\frac{\partial J(v_c, w_{t+j}, U)}{\partial v_c}$.

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

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

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