MAC5725 – Linguística Computacional
EP 1: Word2Vec
Wesley Seidel Carvalho
NUSP: 6544342
2020-02

Parte 1: Fundamentos matemáticos do word2vec

Questão (a)

Sendo y o vetor **one-hot** de w, então a posição y_w será 1 apenas quando $w = o$, sendo assim:

$$-\sum_{w \in Vocab} y_w \log(\hat{y}_w) = -[0\log(\hat{y}_1) + .. + 1\log(\hat{y}_o) + .. + 0\log(\hat{y}_{|V|})] = -\log(\hat{y}_o) \tag{1}$$

Questão (b)

Para simplificar escrita, fazemos: $J = J_{naive-softmax}(v_c, o, U)$

$$J = -\log(\frac{\exp(u_o^T.v_c)}{\sum_{w \in Vocab} \exp(u_w^T.v_c)}) \tag{2}$$

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

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

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

$$\frac{\partial J}{\partial v_c} = -u_o^T + \frac{1}{\sum_{w \in Vocab} \exp(u_w^T.v_c)}.\sum_{k \in Vocab} \exp(u_k^T.v_c)u_k^T$$

$$\frac{\partial J}{\partial v_c} = -u_o^T + \sum_{k \in Vocab} \frac{\exp(u_k^T.v_c)u_k^T}{\sum_{w \in Vocab} \exp(u_w^T.v_c)} \tag{3}$$

Até aqui a derivada já está calculada, iremos agora escrever em termos de y, $\hat{y}$ e U. Repare que a expresão $\frac{\exp(u_k^T.v_c)}{\sum_{w \in Vocab} \exp(u_w^T.v_c)}$ é y_k, logo:

$$\frac{\partial J}{\partial v_c} = -u_o^T + \sum_{k \in Vocab} y_k u_k^T$$

$$\frac{\partial J}{\partial v_c} = -u_o^T + U^T\hat{y} = U^T\hat{y} - u_o^T$$

Aqui, temo que $u_o^T = y$, pois y é one-hot de o sobre U, logo:

$$\frac{\partial J}{\partial v_c} = U^T\hat{y} - U^T.y \tag{4}$$

E finalmente temos:

$$\frac{\partial J_{naive-softmax}(v_c, o, U)}{\partial v_c} = U^T[\hat{y} - y] \tag{5}$$

Questão (c)

Para simplificar escrita, fazemos: $J = J_{naive-softmax}(v_c, o, U)$

$$J = -\log\left(\frac{\exp(u_o^T.v_c)}{\sum_{w\in Vocab}\exp(u_w^T.v_c)}\right) \tag{6}$$

$$J = -[\log(\exp(u_o^T.v_c)) - \log(\sum_{w\in Vocab}\exp(u_w^T.v_c))] \tag{7}$$

(c) Caso 1: $w = o$

$$\frac{\partial J(v_c, o, U)}{\partial u_{w=0}} = -\frac{\partial \log(\exp(u_o^T.v_c))}{\partial u_{w=0}} + \frac{\partial \log(\sum_{w\in Vocab}\exp(u_w^T.v_c))]}{\partial u_{w=0}}$$

$$\frac{\partial J(v_c, o, U)}{\partial u_{w=0}} = -v_c + \frac{\partial \log(\sum_{w\in Vocab}\exp(u_w^T.v_c))]}{\partial u_{w=0}} \tag{8}$$

Trabalhando o segundo termo de forma análoga à (b), temos:

$$\frac{\partial J(v_c, o, U)}{\partial u_{w=0}} = -v_c + \sum_{w\in Vocab}\hat{y}_w.v_c = -v_c + \hat{y}_{w=o}.v_c$$

$$\frac{\partial J(v_c, o, U)}{\partial u_{w=0}} = v_c[\hat{y}_{w=0} - 1] \tag{9}$$

(c) Caso 2: $w \neq o$

$$\frac{\partial J(v_c, o, U)}{\partial u_{w\neq o}} = -\frac{\partial \log(\exp(u_o^T.v_c))}{\partial u_{w\neq o}} + \frac{\partial \log(\sum_{w\in Vocab}\exp(u_w^T.v_c))]}{\partial u_{w\neq o}} \tag{10}$$

$$\frac{\partial J(v_c, o, U)}{\partial u_{w\neq o}} = 0 + \sum_{w \in Vocab} \hat{y}_w . v_c = \hat{y}_{w\neq o} . v_c$$

$$\frac{\partial J(v_c, o, U)}{\partial u_{w\neq o}} = \hat{y}_{w\neq o} v_c \tag{11}$$

Para termos os dois resultado em termos de y, $\hat{y}$ e v_c, fazemos:

Como no vetor y só teremos o valor 1 na posição $w = o$, o que recai no caso 1; E todas as posições que dizem respeito à $w \neq o$ são igual a 0, então podemos afirmar que:

$$\frac{\partial J(v_c, o, U)}{\partial u_w} = v_c[\hat{y}_w - y_w] \tag{12}$$

Questão (d)

$$\sigma(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1} \tag{13}$$

Fazemos $u = e^x$ e $v = e^x + 1$ e suas derivadas $u' = e^x$ e $v' = e^x$. Sendo assim, $\sigma(x) = \frac{u}{v}$, e sua derivada:

$$\sigma'(x) = \frac{e^x}{(e^x + 1)^2} \tag{14}$$

Escrevendo em termos de $\sigma(x)$:

$$\sigma'(x) = \frac{e^x}{(e^x + 1)^2} = \frac{e^x}{(e^x + 1)} \frac{1}{(e^x + 1)} = \sigma(x).(\frac{1}{(e^x + 1)}) \tag{15}$$

Desenvolvendo o termo $\frac{1}{(e^x+1)}$ da expressão (18) temos:

$$\frac{1}{e^x + 1} = \frac{1}{e^x + 1} + e^x - e^x = \frac{1 + e^x - e^x}{e^x + 1} = \frac{e^x + 1}{e^x + 1} - \frac{e^x}{e^x + 1} = 1 - \frac{e^x}{e^x + 1} \tag{16}$$

$$\frac{1}{e^x + 1} = 1 - \sigma(x) \tag{17}$$

Substituindo o resultado de (17) em (18), temos finalmente:

$$\sigma'(x) = \sigma(x).(1 - \sigma(x)) \tag{18}$$

Amostragem Negativa

Questão (e)

$$J_{amostra-negativa}(v_c, o, U) = -log(\sigma(u_o^T.v_c)) - \sum_{k=1}^{K} log(\sigma(-u_k^T.v_c)) \tag{19}$$

Para simplificação fazemos $J = J_{amostra-negativa}(v_c, o, U)$.
e.1: Derivada de J em relação v_c:

$$\frac{\partial J}{\partial v_c} = \frac{\partial[-log(\sigma(u_o^T.v_c)) - \sum_{k=1}^{K} log(\sigma(-u_k^T.v_c))]}{\partial v_c} \tag{20}$$

$$\frac{\partial J}{\partial v_c} = \frac{\partial}{\partial v_c}[-log(\sigma(u_o^T.v_c))] - \frac{\partial}{\partial v_c}[\sum_{k=1}^{K} log(\sigma(-u_k^T.v_c))] \tag{21}$$

$$\frac{dJ}{dv_c} = -[1 - \sigma(u_o^T.v_c)]u_o + \sum_{k=1}^{K} u_k^T[1 - \sigma(-u_k^T.v_c)] \tag{22}$$

e.2: Derivada de J em relação u_o:

$$\frac{\partial J}{\partial u_o} = \frac{\partial[-log(\sigma(u_o^T.v_c)) - \sum_{k=1}^{K} log(\sigma(-u_k^T.v_c))]}{\partial u_o} \tag{23}$$

$$\frac{\partial J}{\partial u_o} = \frac{\partial}{\partial u_o}[-log(\sigma(u_o^T.v_c))] - \frac{\partial}{\partial u_o}[\sum_{k=1}^{K} log(\sigma(-u_k^T.v_c))] \tag{24}$$

$$\frac{\partial J}{\partial u_o} = -[1 - \sigma(u_o^T.v_c)]v_c - \sum_{k=1}^{K} 1 - \sigma(-u_k^T.v_c)\frac{\partial u_k^T.v_c}{\partial u_o} \tag{25}$$

Como o não pertence ao conjunto K, logo o resultado do segundo termo é 0, então:

$$\frac{dJ}{du_o} = -[1 - \sigma(u_o^T.v_c)]v_c + 0 \tag{26}$$

e.3: Derivada de J em relação u_k:

$$\frac{\partial J}{\partial u_k} = \frac{\partial[-log(\sigma(u_o^T.v_c)) - \sum_{k=1}^{K} log(\sigma(-u_k^T.v_c))]}{\partial u_k} \tag{27}$$

$$\frac{\partial J}{\partial u_k} = \frac{\partial}{\partial u_k}[-log(\sigma(u_o^T.v_c))] - \frac{\partial}{\partial u_k}[\sum_{k=1}^{K} log(\sigma(-u_k^T.v_c))] \tag{28}$$

$$\frac{\partial J}{\partial u_k} = \frac{\partial}{\partial u_k}[-log(\sigma(u_o^T.v_c))] - [\sum_{k=1}^{K} \frac{1}{log(\sigma(-u_k^T.v_c))}.\sigma'(-u_k^T.v_c).\frac{\partial - u_k^T.v_c}{\partial u_k}] \tag{29}$$

$$\frac{\partial J}{\partial u_k} = -\frac{1}{\sigma(u_o^T.v_c)}.\sigma'(u_o^T.v_c)\frac{\partial u_o^T.v_c}{\partial u_k} - \sum_{k=1}^{K}[-v_c(1 - \sigma(-u_k^T.v_c)] \tag{30}$$

Como a derivada de $u_o^T.v_c$ é 0 em relação a u_k, temos:

$$\frac{\partial J}{\partial u_k} = 0 + \sum_{k=1}^{K}[v_c(1 - \sigma(-u_k^T.v_c)] \tag{31}$$

E finalmente:

$$\frac{dJ}{du_k} = \sum_{k=1}^{K} v_c[1 - \sigma(-u_k^T.v_c)] \tag{32}$$

Comentário: Na **função de custo naive softmax** é necessário utilizar todo o vocabulário para o cálculo do custo, enquanto na **função de custo das amostragem negativa** é utilizado apenas um subconjunto de palavras para a realização do cálculo.

Questão (f)

Derivar J em relação à U, v_c e v_w.

$$J_{skip-gram}(v_c, w_{tm}, ..., w_{t+m}, U) = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} J(v_c, w_{t+1}, U) \tag{33}$$

f.1: Derivada de J em relação U:

$$\frac{\partial J}{\partial U} = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \frac{dJ(v_c, w_{t+j}, U)}{\partial U} \tag{34}$$

f.2: Derivada de J em relação v_c:

$$\frac{\partial J}{\partial v_c} = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \frac{dJ(v_c, w_{t+j}, U)}{\partial v_c} \tag{35}$$

f.3: Derivada de J em relação $v_{w \neq c}$:

$$\frac{\partial J}{\partial v_{w \neq c}} = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \frac{dJ(v_c, w_{t+j}, U)}{\partial v_{w \neq c}} = 0 \tag{36}$$