MAC5725- Linguística Computacional EP 1

Andre Barbosa - NUSP7971751

andre.barbosa@ime.usp.br

25 de Setembro de 2020

1. Mostre que a perda/custo naive-softmax dada na Equação (2) é a mesma que a perda de entropia cruzada entre y e $\hat{y}$;ou seja, mostre que:

$$- \sum_{w \in Vocab} y_w log(\hat{y}_w) = -log(\hat{y}_o) \tag{1}$$

Considerando que a representação de palavras é um *one hot encoding/bag of words*, para uma determinada palavra o, ela será 1 quando $w = o$ e 0 para as demais palavras. Logo, teremos o seguinte:

Demonstração.

$$\begin{aligned} - \sum_{w \in Vocab} y_w log(\hat{y}_w) = \\ = -(y_o * log(\hat{y}_o) + \sum_{w \in Vocab - \{o\}} y_w log(\hat{y}_w)) = \\ = -(1 * log(\hat{y}_o) + \sum_{w \in Vocab - \{o\}} 0 * log(\hat{y}_w)) = -log(\hat{y}_o) \end{aligned} \tag{2}$$

□

2. Calcule a derivada parcial de $J_{naive-softmax}(v_c, o, U)$ em relação a v_c. Por favor escreva a resposta em termos de y, $\hat{y}$ e U.

Por questões de simplificação, usaremos J para denotar $J_{naive-softmax}(v_c, o, U)$. Além disso, da equação (2) temos que:

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

Disso, temos que:

$$
\begin{aligned}
(\frac{\partial}{\partial v_c} - log\left[\frac{exp(u_o^T \cdot v_c)}{\sum_{w \in \text{Vocab}} exp(u_w^T \cdot v_c)}\right]) = \\
= -(\frac{\partial}{\partial v_c}[log(exp(u_o^T \cdot v_c)) - log(\sum_{w \in \text{Vocab}} exp(u_w^T \cdot v_c))]) \\
= -(\frac{\partial}{\partial v_c} log(exp(u_o^T \cdot v_c)) - \frac{\partial}{\partial v_c} log(\sum_{w \in \text{Vocab}} exp(u_w^T \cdot v_c))) \\
= -(u_o - \frac{\partial}{\partial v_c} log(\sum_{w \in \text{Vocab}} exp(u_w^T \cdot v_c))) \\
= -(u_o - \frac{1}{\sum_{w \in \text{Vocab}} exp(u_w^T \cdot v_c)} \frac{\partial}{\partial v_c} \sum_{k \in \text{Vocab}} exp(u_k^T \cdot v_c)) \\
= -(u_o - \frac{1}{\sum_{w \in \text{Vocab}} exp(u_w^T \cdot v_c)} \sum_{k \in \text{Vocab}} \frac{\partial}{\partial v_c} exp(u_k^T \cdot v_c)) \\
= -(u_o - \frac{1}{\sum_{w \in \text{Vocab}} exp(u_w^T \cdot v_c)} \sum_{k \in \text{Vocab}} exp(u_k^T \cdot v_c) \frac{\partial}{\partial v_c} u_k^T \cdot v_c) \\
= -(u_o - \sum_{k \in \text{Vocab}} \frac{exp(u_k^T \cdot v_c)}{\sum_{w \in \text{Vocab}} exp(u_w^T \cdot v_c)} u_k)
\end{aligned}
$$

E, segundo a equação 1, podemos reescrever da seguinte forma:

$$= -(u_o - \sum_{k \in \text{Vocab}} p(O = k | C = c) u_k)$$

Em que u_k é a palavra de índice k em U. Aplicando o somatório, então, temos que:

$$\frac{\partial J}{\partial v_c} = U^T(\hat{y} - y)$$

3. Calcule as derivadas parciais de $J_{naive-softmax}(v_c, o, U)$ em relação a cada um dos vetores de palavras "externas", $u_w's$. Há dois casos: quando $w = o$, o verdadeiro vetor de palavras "externas" e $w \neq o$, para todas as outras palavras. Escreva a sua resposta em termos de y, $\hat{y}$ e v_c.

$$
\begin{aligned}
(\frac{\partial}{\partial u_w} - log\left[\frac{exp(u_o^T \cdot v_c)}{\sum_{w \in \text{Vocab}} exp(u_w^T \cdot v_c)}\right]) = \\
= -(\frac{\partial}{\partial u_w}[log(exp(u_o^T \cdot v_c)) - log(\sum_{w \in \text{Vocab}} exp(u_w^T \cdot v_c))]) \\
= -(\frac{\partial}{\partial u_w} log(exp(u_o^T \cdot v_c)) - \frac{\partial}{\partial u_w} log(\sum_{w \in \text{Vocab}} exp(u_w^T \cdot v_c)))
\end{aligned}
$$

Disso temos os dois casos:

1. $w = o$

$$
\begin{aligned}
&= -(\frac{\partial}{\partial u_o} log(exp(u_o^T \cdot v_c)) - \frac{\partial}{\partial u_o} log(\sum_{w \in \text{Vocab}} exp(u_w^T \cdot v_c))) \\
&= -(v_c - \frac{\partial}{\partial u_o} log(\sum_{w \in \text{Vocab}} exp(u_w^T \cdot v_c))) \\
&= -(v_c - \frac{1}{\sum_{w \in \text{Vocab}} exp(u_w^T \cdot v_c)} (\frac{\partial}{\partial u_o} \sum_{w \in \text{Vocab}} exp(u_w^T \cdot v_c))) \\
&= -(v_c - \frac{1}{\sum_{w \in \text{Vocab}} exp(u_w^T \cdot v_c)} (\frac{\partial}{\partial u_o} exp(u_o^T \cdot v_c))) \\
&= -(v_c - \frac{1}{\sum_{w \in \text{Vocab}} exp(u_w^T \cdot v_c)} (exp(u_o^T \cdot v_c) v_c)) \\
&= -(v_c - \frac{(exp(u_o^T \cdot v_c)}{\sum_{w \in \text{Vocab}} exp(u_w^T \cdot v_c)} v_c) \\
&= -(v_c - p(O = o|C = c) v_c) \\
&= (p(O = o|C = c) - 1) v_c
\end{aligned}
$$

2. $w \neq o$

$$
\begin{aligned}
&= -(\frac{\partial}{\partial u_w} log(exp(u_o^T \cdot v_c)) - \frac{\partial}{\partial u_w} log(\sum_{w \in \text{Vocab}} exp(u_w^T \cdot v_c))) \\
&= -(0 - \frac{\partial}{\partial u_w} log(\sum_{w \in \text{Vocab}} exp(u_w^T \cdot v_c))) \\
&= -(0 - \frac{1}{\sum_{w \in \text{Vocab}} exp(u_w^T \cdot v_c)} (\frac{\partial}{\partial u_w} \sum_{w \in \text{Vocab}} exp(u_w^T \cdot v_c))) \\
&= -(0 - \frac{1}{\sum_{w \in \text{Vocab}} exp(u_w^T \cdot v_c)} (\frac{\partial}{\partial u_w} exp(u_w^T \cdot v_c))) \\
&= -(0 - \frac{1}{\sum_{w \in \text{Vocab}} exp(u_w^T \cdot v_c)} (exp(u_o^T \cdot v_c) v_c)) \\
&= -(0 - \frac{(exp(u_w^T \cdot v_c)}{\sum_{w \in \text{Vocab}} exp(u_w^T \cdot v_c)} v_c) \\
&= -(-p(O = w|C = c) v_c) \\
&= p(O = w|C = c) v_c
\end{aligned}
$$

De (1) e (2), então, temos que

$$\frac{\partial J}{\partial u_w} = (\hat{y} - y) v_c$$

4. A função sigmóide é dada pela Equação :

$$\sigma(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1} \tag{3}$$

Calcule a derivada de $\sigma(x)$ em relação a x, onde x é um escalar

Podemos dar o resultado por:

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

5. Repita as partes (b) e (c), calculando as derivadas parciais de $J_{neg-sample}$ em relação a v_c, em relação a u_o, e em relação a uma amostra negativa u_k:

Derivada Parcial de $J_{neg-sample}$ em relação a v_c

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

Derivada Parcial de $J_{neg-sample}$ em relação a u_o

$$\frac{\partial}{\partial u_o}[-log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} log(\sigma(-u_k^T v_c))] =$$
$$= -\frac{\partial}{\partial u_o} log(\sigma(u_o^T v_c)) - \frac{\partial}{\partial u_o} \sum_{k=1}^{K} log(\sigma(-u_k^T v_c))$$
$$= -\frac{\partial}{\partial u_o} log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} \frac{\partial}{\partial u_o} log(\sigma(-u_k^T v_c))$$
$$= -\frac{\sigma(u_o^T v_c)(1-\sigma(u_o^T v_c))}{\sigma(u_o^T v_c)} \frac{\partial}{\partial u_o} u_o^T v_c - \sum_{k=1}^{K} \frac{\sigma(-u_k^T v_c)(1-\sigma(-u_k^T v_c))}{\sigma(-u_k^T v_c)} \frac{\partial}{\partial u_o} - u_k^T v_c$$
$$= -(1-\sigma(u_o^T v_c))v_c - \sum_{k=1}^{K}(1-\sigma(-u_k^T v_c))0$$
$$= -(1-\sigma(u_o^T v_c))v_c$$

Derivada Parcial de $J_{neg-sample}$ em relação a uma amostra negativa u_k

$$\frac{\partial}{\partial u_k}[-log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} log(\sigma(-u_k^T v_c))] =$$
$$= -\frac{\partial}{\partial u_k} log(\sigma(u_o^T v_c)) - \frac{\partial}{\partial u_k} \sum_{k=1}^{K} log(\sigma(-u_k^T v_c))$$
$$= -\frac{\partial}{\partial u_k} log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} \frac{\partial}{\partial u_k} log(\sigma(-u_k^T v_c))$$
$$= -\frac{\sigma(u_o^T v_c)(1-\sigma(u_o^T v_c))}{\sigma(u_o^T v_c)} \frac{\partial}{\partial u_k} u_o^T v_c - \sum_{k=1}^{K} \frac{\sigma(-u_k^T v_c)(1-\sigma(-u_k^T v_c))}{\sigma(-u_k^T v_c)} \frac{\partial}{\partial u_k} - u_k^T v_c$$
$$= -(1-\sigma(u_o^T v_c)0 - \sum_{k=1}^{K}(1-\sigma(-u_k^T v_c))(-v_c)$$
$$= +\sum_{k=1}^{K}(1-\sigma(-u_k^T v_c))v_c$$

Então, no caso de uma amostra negativa u_k, então: $\frac{J_{\text{neg-sample}}}{\partial u_k} = (1-\sigma(-u_k^T v_c))v_c$

Esta função de custo é mais eficente proquê ela não precisa carregar todo o vocabulário de palavras em memória, levando em conta apenas a palavra observada, a central e as K amostras selecionadas

6. Escreva três derivadas parciais, em que:

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

---

i $\frac{\partial J_{\text{skip-gram}}(v_c, w_{t-m}, \cdots, w_{t+m}, U)}{\partial U}$

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

ii $\frac{\partial J_{\text{skip-gram}}(v_c, w_{t-m}, \cdots, w_{t+m}, U)}{\partial v_c}$

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

iii $\frac{\partial J_{\text{skip-gram}}(v_c, w_{t-m}, \cdots, w_{t+m}, U)}{\partial v_w}$, com $w \neq c$

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