MAC5725
Linguística Computacional
Soluções dos exercícios da 1° parte do EP1

Aluno: Felipe de Lima Peressim
NUSP: 11823558

17 de setembro de 2020

Parte 1: Fundamentos matemáticos do word2vec

(a) 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

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

Demonstração. Suponha que w seja a k-éssima palavra do vocabulário de tal modo que $w = o$. E, sabendo-se que y é um vetor one-hot com 1 para a verdadeira palavra externa o e 0 para as demais, então, podemos expandir a somatória da equação (1) para concluir que

$$\begin{aligned}
-\sum_{w \in Vocab} y_w \log(\hat{y}_w) &= -(y_0 \log(\hat{y}_0) + y_1 \log(\hat{y}_1) + \cdots + y_w \log(\hat{y}_w) + \cdots + y_{n-1} \log(\hat{y}_{n-1})) \\
&= -(0 \cdot \log(\hat{y}_0) + 0 \cdot \log(\hat{y}_1) + \cdots + 1 \cdot \log(\hat{y}_w) + \cdots + 0 \cdot \log(\hat{y}_{n-1})) \\
&= -\log(\hat{y}_w) \\
&= -\log(\hat{y}_o)
\end{aligned}$$

com $n = |Vocab|$.

□

(b) 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.

$$
\begin{aligned}
\frac{\partial J_{naive-softmax}(v_c, o, U)}{\partial v_c} &= \frac{\partial - \log(P(O = o|C = c))}{\partial v_c} \\
&= -\frac{\partial \log(\frac{\exp(u_o^\top \cdot v_c)}{\sum_{w \in Vocab} \exp(u_w^\top \cdot v_c)})}{\partial v_c} \\
&= \frac{\partial \log(\sum_{w \in Vocab} \exp(u_w^\top \cdot v_c))}{\partial v_c} - \frac{\partial \log(\exp(u_o^\top \cdot v_c))}{\partial v_c} \\
&= \frac{\partial \log(\sum_{w \in Vocab} \exp(u_w^\top \cdot v_c))}{\partial v_c} - \frac{\partial u_o^\top \cdot v_c}{\partial v_c} \\
&= \frac{1}{\sum_{w \in Vocab} \exp(u_w^\top \cdot v_c)} \cdot \frac{\partial \sum_{w \in Vocab} \exp(u_w^\top \cdot v_c)}{\partial v_c} - u_o \\
&= \sum_{w \in Vocab} \frac{\exp(u_w^\top \cdot v_c)}{\sum_{z \in Vocab} \exp(u_z^\top \cdot v_c)} \cdot \frac{\partial u_w^\top \cdot v_c}{\partial v_c} - u_o \\
&= \sum_{w \in Vocab} \frac{\exp(u_w^\top \cdot v_c)}{\sum_{z \in Vocab} \exp(u_z^\top \cdot v_c)} u_w - u_o \\
&= \sum_{w \in Vocab} (\hat{y}_w \cdot u_w) - u_o \\
&= U^\top \hat{y} - u_o
\end{aligned}
$$

como y é vetor one-hot, podemos reescrever u_o como $U^\top y$ e portanto

$$
\begin{aligned}
U^\top \hat{y} - u_o &= U^\top \hat{y} - U^\top y \\
&= U^\top (\hat{y} - y)
\end{aligned}
$$

(c) 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.

$$
\begin{aligned}
\frac{\partial J_{naive-softmax}(v_c, o, U)}{\partial u_w} &= \frac{\partial - \log(P(O = o|C = c))}{\partial u_w} \\
&= -\frac{\partial \log(\frac{\exp(u_o^\top \cdot v_c)}{\sum_{w \in Vocab} \exp(u_w^\top \cdot v_c)})}{\partial u_w} \\
&= \frac{\partial \log(\sum_{w \in Vocab} \exp(u_w^\top \cdot v_c))}{\partial u_w} - \frac{\partial \log(\exp(u_o^\top \cdot v_c))}{\partial u_w} \\
&= \frac{\partial \log(\sum_{w \in Vocab} \exp(u_w^\top \cdot v_c))}{\partial u_w} - \frac{\partial u_o^\top \cdot v_c}{\partial u_w} \\
&= \frac{1}{\sum_{w \in Vocab} \exp(u_w^\top \cdot v_c)} \cdot \frac{\partial \sum_{w \in Vocab} \exp(u_w^\top \cdot v_c)}{\partial u_w} - \frac{\partial u_o^\top \cdot v_c}{\partial u_w}
\end{aligned}
$$

Tomando um w_k qualquer,

$$
\begin{aligned}
\frac{\partial \sum_{w \in Vocab} \exp(u_w^\top \cdot v_c)}{\partial u_{w_k}} &= \frac{\partial \exp(u_{w_0}^\top \cdot v_c)}{\partial u_{w_k}} + \frac{\partial \exp(u_{w_1}^\top \cdot v_c)}{\partial u_{w_k}} \\
&+ \cdots + \frac{\partial \exp(u_{w_k}^\top \cdot v_c)}{\partial u_{w_k}} + \cdots + \frac{\partial \exp(u_{w_n}^\top \cdot v_c)}{\partial u_{w_k}} \\
&= \exp(u_{w_k}^\top \cdot v_c) \cdot \frac{\partial u_{w_k}^\top \cdot v_c}{\partial u_{w_k}} \\
&= \exp(u_{w_k}^\top \cdot v_c) \cdot v_c
\end{aligned}
$$

logo,

$$
\begin{aligned}
\frac{1}{\sum_{w \in Vocab} \exp(u_w^\top \cdot v_c)} \cdot \frac{\partial \sum_{w \in Vocab} \exp(u_w^\top \cdot v_c)}{\partial u_w} - \frac{\partial u_o^\top \cdot v_c}{\partial u_w} &= \frac{\exp(u_w^\top \cdot v_c)}{\sum_{w \in Vocab} \exp(u_w^\top \cdot v_c)} \cdot v_c \\
-\frac{\partial u_o^\top \cdot v_c}{\partial u_w}
\end{aligned}
$$

se $w = o$

$$
\begin{aligned}
\frac{\exp(u_w^\top \cdot v_c)}{\sum_{w \in Vocab} \exp(u_w^\top \cdot v_c)} \cdot v_c - \frac{\partial u_o^\top \cdot v_c}{\partial u_w} &= \frac{\exp(u_w^\top \cdot v_c)}{\sum_{w \in Vocab} \exp(u_w^\top \cdot v_c)} \cdot v_c - v_c \\
&= \hat{y}_w \cdot v_c - v_c \\
&= (\hat{y}_w - 1) \cdot v_c
\end{aligned}
$$

se $w \neq o$

$$
\begin{aligned}
\frac{\exp(u_w^\top \cdot v_c)}{\sum_{w \in Vocab} \exp(u_w^\top \cdot v_c)} \cdot v_c - \frac{\partial u_o^\top \cdot v_c}{\partial u_w} &= \frac{\exp(u_w^\top \cdot v_c)}{\sum_{w \in Vocab} \exp(u_w^\top \cdot v_c)} \cdot v_c - 0 \\
&= \hat{y}_w \cdot v_c
\end{aligned}
$$

para combinar ambos os casos basta notar que y é 1 na posição da palavra w e 0 em todas as outras posições, logo podemos reescrever a derivada parcial em termos de y, $\hat{y}$, e v_c, portanto

$$
\frac{\partial J_{naive-softmax}(v_c, o, U)}{\partial u_w} = (\hat{y} - y)^\top \cdot v_c
$$

(d) A função sigmóide é dada pela Equação (2):

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

Calcule a derivada de $\sigma(x)$ em relação a x, onde x é um escalar. Dica: dê sua resposta em termos $\sigma(x)$.

$$
\begin{aligned}
\frac{\mathrm{d}\sigma(x)}{\mathrm{d}x} &= \frac{\mathrm{d}\frac{e^x}{e^x+1}}{\mathrm{d}x} \\
&= \frac{\frac{\mathrm{d}e^x}{\mathrm{d}x}\cdot(e^x+1) - e^x \cdot \frac{\mathrm{d}(e^x+1)}{\mathrm{d}x}}{(e^x+1)^2} \\
&= \frac{e^x\cdot(e^x+1) - e^x\cdot e^x}{(e^x+1)^2} \\
&= \frac{e^x}{(e^x+1)^2} \\
&= \frac{e^x}{e^x+1}\cdot\frac{1}{e^x+1} \\
&= \frac{e^x}{e^x+1}\cdot\frac{1+e^x-e^x}{e^x+1} \\
&= \frac{e^x}{e^x+1}\cdot(\frac{e^x+1}{e^x+1} - \frac{e^x}{e^x+1}) \\
&= \sigma(x)\cdot(1-\sigma(x))
\end{aligned}
$$

(e) Vamos considerar a perda de Amostragem Negativa, que é uma alternativa para a Perda Naive-Softmax. Suponha que K amostras negativas (palavras) sejam retiradas do vocabulário. Por simplicidade de notação, devemos nos referir a eles como $w_1, w_2, \cdots, w_K$ e seus vetores externos como $u_1, \cdots, u_K$. Observe que o $o \notin w_1, \cdots, w_K$. Para uma palavra central c e uma palavra externa o, a função de custo de amostragem negativa é dada pela soma da entropia cruzada do exemplo positivo e das entropias cruzadas das amostras negativas:

$$
J_{amostra\ negativa}(v_c, o, U) = -\log(\sigma(u_o^\top \cdot v_c)) - \sum_{k=1}^{K} \log(\sigma(-u_k^\top \cdot v_c)) \qquad (3)
$$

para uma amostra $w_1, \cdots, w_K$, onde $\sigma(\cdot)$ é a função sigmóide. Repita as partes (b) e (c), calculando as derivadas parciais de $J_{amostra\ negativa}$ em relação a v_c, em relação a u_o, e em relação a uma amostra negativa u_k. Dê suas respostas em termos dos vetores u_o , v_c e u_k , onde $k \in [1, K]$. Descreva com uma frase por que esta função de custo é muito mais eficiente de calcular do que o custo do naive *Softmax*. Você deve ser capaz de usar sua solução da parte (d) para ajudar a calcular os gradientes necessários aqui.

- Derivada parcial com relação a v_c.

$$
\begin{aligned}
\frac{\partial J_{amostra\ negativa}(v_c, o, U)}{\partial v_c} &= \frac{\partial - \log(\sigma(u_o^\top \cdot v_c)) - \sum\limits_{k=1}^{K} \log(\sigma(-u_k^\top \cdot v_c))}{\partial v_c} \\
&= -\frac{\partial \log(\sigma(u_o^\top \cdot v_c))}{\partial v_c} - \sum_{k=1}^{K} \frac{\partial \log(\sigma(-u_k^\top \cdot v_c))}{\partial v_c} \\
&= -\frac{1}{\sigma(u_o^\top \cdot v_c)} \cdot \frac{\partial \sigma(u_o^\top \cdot v_c)}{\partial v_c} - \sum_{k=1}^{K} \Big(\frac{1}{\sigma(-u_k^\top \cdot v_c)} \cdot \frac{\partial \sigma(u_k^\top \cdot v_c)}{\partial v_c} \Big) \\
&= -\frac{\sigma(u_o^\top \cdot v_c) \cdot (1 - \sigma(u_o^\top \cdot v_c))}{\sigma(u_o^\top \cdot v_c)} \cdot \frac{\partial \sigma(u_o^\top \cdot v_c)}{\partial v_c} \\
&- \sum_{k=1}^{K} \Big(\frac{\sigma(-u_k^\top \cdot v_c) \cdot (1 - \sigma(-u_k^\top \cdot v_c))}{\sigma(-u_k^\top \cdot v_c)} \cdot \frac{\partial(-u_k^\top \cdot v_c)}{\partial v_c} \Big) \\
&= -(1 - \sigma(u_o^\top \cdot v_c)) \cdot u_o^\top - \sum_{k=1}^{K} (1 - \sigma(-u_k^\top \cdot v_c)) \cdot (-u_k) \\
&= \sum_{k=1}^{K} (1 - \sigma(-u_k^\top \cdot v_c)) \cdot u_k - (1 - \sigma(u_o^\top \cdot v_c)) \cdot u_o^\top
\end{aligned}
$$

- Derivada parcial com relação a u_o.

$$
\begin{aligned}
\frac{\partial J_{amostra\ negativa}(v_c, o, U)}{\partial u_o} &= \frac{\partial - \log(\sigma(u_o^\top \cdot v_c)) - \sum\limits_{k=1}^{K} \log(\sigma(-u_k^\top \cdot v_c))}{\partial u_o} \\
&= -\frac{\partial \log(\sigma(u_o^\top \cdot v_c))}{\partial u_o} - \sum_{k=1}^{K} \frac{\partial \log(\sigma(-u_k^\top \cdot v_c))}{\partial u_o} \\
&= -\frac{1}{\sigma(u_o^\top \cdot v_c)} \cdot \frac{\partial \sigma(u_o^\top \cdot v_c)}{\partial u_o} - 0 \\
&= -\frac{\sigma(u_o^\top \cdot v_c) \cdot (1 - \sigma(u_o^\top \cdot v_c))}{\sigma(u_o^\top \cdot v_c)} \cdot \frac{\partial(u_o^\top \cdot v_c)}{\partial u_o} \\
&= -(1 - \sigma(u_o^\top \cdot v_c)) \cdot v_c
\end{aligned}
$$

- Derivada parcial com relação a u_k.

$$
\begin{aligned}
\frac{\partial J_{amostra\ negativa}(v_c, o, U)}{\partial u_k} &= \frac{\partial - \log(\sigma(u_o^\top \cdot v_c)) - \sum_{k=1}^{K} \log(\sigma(-u_k^\top \cdot v_c))}{\partial u_k} \\
&= -\frac{\partial \log(\sigma(u_o^\top \cdot v_c))}{\partial u_k} - \sum_{k=1}^{K} \frac{\partial \log(\sigma(-u_k^\top \cdot v_c))}{\partial u_k} \\
&= -0 - \frac{1}{\sigma(-u_k^\top \cdot v_c)} \cdot \frac{\partial \sigma(-u_k^\top \cdot v_c)}{\partial u_k} \\
&= -\frac{\sigma(-u_k^\top \cdot v_c) \cdot (1 - \sigma(-u_k^\top \cdot v_c))}{\sigma(-u_k^\top \cdot v_c)} \cdot \frac{\partial(-u_k^\top \cdot v_c)}{\partial u_k} \\
&= (1 - \sigma(-u_k^\top \cdot v_c)) \cdot v_c
\end{aligned}
$$

- A função de custo de amostragem negativa é mais eficiente porque apenas um subconjunto de palavras é utilizado para a realização dos cálculos, em contrapartida, para computar a função naive Softmax é necessário passar por todo o vocabulário.

(f) Suponha que a palavra central seja $c = w_t$ e a janela de contexto seja $[w_{t-m}, \cdots, w_{t-1}, w_t, w_{t+1}, \cdots, w_{t+m}]$, onde m é o tamanho da janela de contexto. Lembre-se de que, para a versão skip-gram do word2vec, o custo total da janela de contexto é:

$$
J_{skip-gram}(v_c, w_{t-m}, \cdots, w_{t+m}, U) = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} J(v_c, w_{t+j}, U) \tag{4}
$$

onde $J(v_c, w_{t+j}, U)$ representa um termo de custo arbitrário para a palavra central $c = w_t$ e palavra externa w_{t+j} . Este custo pode ser $J_{naive-softmax}(v_c, w_{t+j}, U)$ ou $J_{neg-sample}(v_c, w_{t+j}, U)$, dependendo da sua implementação. Escreva três derivadas parciais:

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

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

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

Dê suas respostas em termos de $\frac{\partial J_{skip-gram}(v_c, w_{t+j}, U)}{\partial U}$ e $\frac{\partial J_{skip-gram}(v_c, w_{t+j}, U)}{\partial v_c}$.
Isso deve ser muito simples – cada solução deve ser uma linha.

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

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

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

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

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

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