

# Stanford CS 224n Assignment 2

Yufan Liu

January 7, 2022

## 1 Written: Understanding word2Vec (23 points)

(a) (3 points) Show that the naive-softmax loss given in Equation (2) is the same as the cross-entropy loss between  $y$  and  $\hat{y}$ ; i.e., show that

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

Your answer should be one line.

**Answer:** Given outside word  $o$  and context word  $c$ ,

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

(b) (5 points) Compute the partial derivative of  $J_{naive-softmax}(v_c, o, U)$  with respect to  $v_c$ . Please write your answer in terms of  $y$ ,  $\hat{y}$ , and  $U$ .

**Answer:**

$$\begin{aligned} \frac{\partial J}{\partial v_c} &= \frac{\partial \log(P(O = o \mid C = c))}{\partial v_c} \\ &= \frac{\partial \log(\exp(u_o^T v_c))}{\partial v_c} + \frac{\partial \log(\sum_{w=1}^V \exp(u_w^T v_c))}{\partial v_c} \\ &= -u_o + \sum_{w=1}^V P(O = w \mid C = c) u_w \\ &= U^T(\hat{y} - y) \end{aligned}$$

(c) (5 points) Compute the partial derivatives of  $J_{naive-softmax}(\mathbf{v}_c, o, \mathbf{U})$  with respect to each of the ‘outside’ word vectors,  $u_w$ ’s. There will be two cases: when  $w = o$ , the true ‘outside’ word vector, and  $w \neq o$ , for all other words. Please write your answer in terms of  $y$ ,  $\hat{y}$ , and  $v_c$ .

**Answer:** Similar to answer (b) above.

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

(d) (3 Points) The sigmoid function is given by Equation 4:

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

Please compute the derivative of  $\sigma(x)$  with respect to  $x$ , where  $x$  is a scalar. Hint: you may want to write your answer in terms of  $\sigma(x)$ .

**Answer:**

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

(e) (4 points) Now we shall consider the Negative Sampling loss, which is an alternative to the Naive Softmax loss. Assume that  $K$  negative samples (words) are drawn from the vocabulary. For simplicity of notation we shall refer to them as  $w_1, w_2, \dots, w_K$  and their outside vectors as  $\mathbf{u}_1, \dots, \mathbf{u}_K$ . Note that  $o \notin w_1, \dots, w_K$ . For a center word  $c$  and an outside word  $o$ , the negative sampling loss function is given by:

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

for a sample  $w_1, \dots, w_K$ , where  $\sigma(\cdot)$  is the sigmoid function.

Please repeat parts (b) and (c), computing the partial derivatives of  $J_{neg-sample}$  with respect to  $v_c$ , with respect to  $u_o$ , and with respect to a negative sample  $u_k$ . Please write your answers in terms of the vectors  $u_o, v_c$ , and  $u_k$ , where  $k \in [1, K]$ . After you've done this, describe with one sentence why this loss function is much more efficient to compute than the naive-softmax loss. Note, you should be able to use your solution to part (d) to help compute the necessary gradients here.

**Answer:**

$$\begin{aligned}
\frac{\partial J}{\partial u_o} &= -(1 - \sigma(u_o^T v_c))v_c \\
\frac{\partial J}{\partial u_k} &= (1 - \sigma(-u_k^T v_c))v_c \\
\frac{\partial J}{\partial v_c} &= -(1 - \sigma(u_o^T v_c))u_o - \sum_{k=1}^K (1 - \sigma(-u_k^T v_c))u_k
\end{aligned}$$

**Answer:** The loss function changes from  $V$  multi-classifier to  $\{0,1\}$  binary classifiers, and the probability need to output decreases from  $V$  to  $2 \cdot K$ .

(f) (3 points) Suppose the center word  $w_t$  and the context window is  $[w_{t-m}, \dots, w_{t-1}, w_t, w_{t+1}, \dots, w_{t+m}]$ , where  $m$  is the context window size. Recall that for the skip-gram version of word2Vec, the total loss for the context window is:

$$J_{skip-gram}(v_c, w_{t-m}, \dots, w_{t+m}, U) = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} J(v_c, w_{t+j}, U) \quad (2)$$

Here,  $J(v_c, w_{t+j}, U)$  represents an arbitrary loss term for the center word  $c = w_t$  and outside word  $w_{t+j}$ .  $J(v_c, w_{t+j}, U)$  could be  $J_{naive-softmax}(v_c, w_{t+j}, U)$  or  $J_{neg-sample}(v_c, w_{t+j}, U)$ , depending on your implementation.

Write down three partial derivatives:

- (i)  $\partial J_{skip-gram}(v_c, w_{t-m}, \dots, w_{t+m}) / \partial U$
- (ii)  $\partial J_{skip-gram}(v_c, w_{t-m}, \dots, w_{t+m}) / \partial v_c$
- (iii)  $\partial J_{skip-gram}(v_c, w_{t-m}, \dots, w_{t+m}) / \partial w_c$  when  $w \neq c$

Write your answers in terms of  $\partial J(v_c, w_{t+j}, U) / \partial U$  and  $\partial J(v_c, w_{t+j}, U) / \partial v_c$ . This is very simple – each solution should be one line.

**Answer:**

$$\begin{aligned}\frac{\partial J_{skip-gram}(v_c, w_{t-m}, \dots, w_{t+m})}{\partial U} &= \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \frac{J(v_c, w_{t+j}, U)}{\partial U} \\ \frac{\partial J_{skip-gram}(v_c, w_{t-m}, \dots, w_{t+m})}{\partial v_c} &= \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \frac{J(v_c, w_{t+j}, U)}{\partial v_c} \\ \frac{\partial J_{skip-gram}(v_c, w_{t-m}, \dots, w_{t+m})}{\partial w_c(w \neq c)} &= 0\end{aligned}$$