

# Stanford CS 224N Assignment #2 - Written Solution

December 13, 2020

## 1. Understanding word2vec

In word2vec

- conditional probability distribution is given by

$$P(O = o|C = c) = \frac{\exp(\mathbf{u}_o^T \mathbf{v}_c)}{\sum_{w \in Vocab} \exp(\mathbf{u}_w^T \mathbf{v}_c)} \quad (1)$$

- naive-softmax loss function is given by

$$\mathbf{J}_{naive-softmax}(\mathbf{v}_c, o, \mathbf{U}) = -\log P(O = o|C = c) \quad (2)$$

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

$$-\sum_{w \in Vocab} y_w \log(\hat{y}_w) = -\log(\hat{y}_o) \quad (3)$$

Solution:

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

As  $\mathbf{y}$  is a one-hot encoded vector with a 1 for true outside word  $o$ , and 0 everywhere else,  $y_w = 0$  if  $w \neq o$ , resulting in the below equality

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

(b) Compute the partial derivate of  $\mathbf{J}_{naive-softmax}(\mathbf{v}_c, o, \mathbf{U})$  with respect to  $\mathbf{v}_c$ . Please write your answer in terms of  $\mathbf{y}, \hat{\mathbf{y}}$  and  $\mathbf{U}$ .

Solution:

$$\frac{\partial \mathbf{J}_{naive-softmax}(\mathbf{v}_c, o, \mathbf{U})}{\partial \mathbf{v}_c} = \frac{\partial -\log P(O = o|C = c)}{\partial \mathbf{v}_c} = \frac{\partial -\log\left(\frac{\exp(\mathbf{u}_o^T \mathbf{v}_c)}{\sum_{w \in Vocab} \exp(\mathbf{u}_w^T \mathbf{v}_c)}\right)}{\partial \mathbf{v}_c}$$

$$\begin{aligned}
&= \frac{\partial - \log(\exp(\mathbf{u}_o^T \mathbf{v}_c))}{\partial \mathbf{v}_c} - \frac{\partial - \log(\sum_{w \in V_{ocab}} \exp(\mathbf{u}_w^T \mathbf{v}_c))}{\partial \mathbf{v}_c} \\
&= -\mathbf{u}_o + \frac{1}{\sum_{w \in V_{ocab}} \exp(\mathbf{u}_w^T \mathbf{v}_c)} \frac{\partial \sum_{x \in V_{ocab}} \exp(\mathbf{u}_x^T \mathbf{v}_c)}{\partial \mathbf{v}_c} \\
&= -\mathbf{u}_o + \frac{1}{\sum_{w \in V_{ocab}} \exp(\mathbf{u}_w^T \mathbf{v}_c)} \sum_{x \in V_{ocab}} \frac{\partial \exp(\mathbf{u}_x^T \mathbf{v}_c)}{\partial \mathbf{v}_c} \\
&= -\mathbf{u}_o + \frac{1}{\sum_{w \in V_{ocab}} \exp(\mathbf{u}_w^T \mathbf{v}_c)} \sum_{x \in V_{ocab}} \exp(\mathbf{u}_x^T \mathbf{v}_c) \mathbf{u}_x \\
&= -\mathbf{u}_o + \sum_{x \in V_{ocab}} \frac{\exp(\mathbf{u}_x^T \mathbf{v}_c)}{\sum_{w \in V_{ocab}} \exp(\mathbf{u}_w^T \mathbf{v}_c)} \mathbf{u}_x \\
&= -\mathbf{u}_o + \sum_{x \in V_{ocab}} P(O = x | C = c) \mathbf{u}_x
\end{aligned}$$

In terms of  $\mathbf{y}$ ,  $\hat{\mathbf{y}}$  and  $\mathbf{U}$ :

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

(c) Compute the partial derivate of  $\mathbf{J}_{naive-softmax}(\mathbf{v}_c, o, \mathbf{U})$  with respect to each of the 'outside' word vectors,  $\mathbf{u}_w$ 's.

Solution:

$$\begin{aligned}
\frac{\partial \mathbf{J}_{naive-softmax}(\mathbf{v}_c, o, \mathbf{U})}{\partial \mathbf{u}_w} &= \frac{\partial - \log P(O = o | C = c)}{\partial \mathbf{u}_w} = \frac{\partial - \log(\frac{\exp(\mathbf{u}_o^T \mathbf{v}_c)}{\sum_{w \in V_{ocab}} \exp(\mathbf{u}_w^T \mathbf{v}_c)})}{\partial \mathbf{u}_w} \\
&= \frac{\partial - \log(\exp(\mathbf{u}_o^T \mathbf{v}_c))}{\partial \mathbf{u}_w} - \frac{\partial - \log(\sum_{w \in V_{ocab}} \exp(\mathbf{u}_w^T \mathbf{v}_c))}{\partial \mathbf{u}_w}
\end{aligned}$$

If  $w = o$ , then

$$\begin{aligned}
&= \frac{\partial - \log(\exp(\mathbf{u}_o^T \mathbf{v}_c))}{\partial \mathbf{u}_o} - \frac{\partial - \log(\sum_{w \in V_{ocab}} \exp(\mathbf{u}_w^T \mathbf{v}_c))}{\partial \mathbf{u}_o} \\
&= -\mathbf{v}_c + \frac{1}{\sum_{w \in V_{ocab}} \exp(\mathbf{u}_w^T \mathbf{v}_c)} \frac{\partial \sum_{w \in V_{ocab}} \exp(\mathbf{u}_w^T \mathbf{v}_c)}{\partial \mathbf{u}_o}
\end{aligned}$$

$$\begin{aligned}
&= -\mathbf{v}_c + \frac{\exp(\mathbf{u}_o^T \mathbf{v}_c)}{\sum_{w \in V_{ocab}} \exp(\mathbf{u}_w^T \mathbf{v}_c)} \mathbf{v}_c \\
&= (P(O = o | C = c) - 1) \mathbf{v}_c
\end{aligned}$$

If  $w \neq o$ , then

$$\begin{aligned}
&= \frac{\partial -\log(\exp(\mathbf{u}_o^T \mathbf{v}_c))}{\partial \mathbf{u}_w} - \frac{\partial -\log(\sum_{w \in V_{ocab}} \exp(\mathbf{u}_w^T \mathbf{v}_c))}{\partial \mathbf{u}_w} \\
&= 0 + \frac{1}{\sum_{w \in V_{ocab}} \exp(\mathbf{u}_w^T \mathbf{v}_c)} \frac{\partial \sum_{w \in V_{ocab}} \exp(\mathbf{u}_w^T \mathbf{v}_c)}{\partial \mathbf{u}_w} \\
&= 0 + \frac{\exp(\mathbf{u}_w^T \mathbf{v}_c)}{\sum_{w \in V_{ocab}} \exp(\mathbf{u}_w^T \mathbf{v}_c)} \mathbf{v}_c \\
&= P(O = w | C = c) \mathbf{v}_c
\end{aligned}$$

In terms of terms of  $\mathbf{y}$ ,  $\hat{\mathbf{y}}$  and  $\mathbf{v}_c$ :

$$= (\hat{\mathbf{y}} - \mathbf{y})^T \mathbf{v}_c$$

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

$$\sigma(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1} \quad (4)$$

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  $\sigma(x)$ .

Solution:

$$\begin{aligned}
\frac{d\sigma(x)}{dx} &= \frac{d \frac{e^x}{e^x + 1}}{dx} = \frac{\frac{de^x}{dx}(e^x + 1) - e^x \frac{d(e^x + 1)}{dx}}{(e^x + 1)^2} = \frac{e^x(e^x + 1) - e^x(e^x + 0)}{(e^x + 1)^2} \\
&= \frac{e^x}{(e^x + 1)^2} = \frac{e^x}{e^x + 1} \left(1 - \frac{e^x}{e^x + 1}\right) = \sigma(x)(1 - \sigma(x))
\end{aligned}$$

(e) 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  $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_K$ . Note that  $o \notin \{w_1, w_2, \dots, w_K\}$ . For a center word  $c$  and an outside word  $o$ , the negative sampling loss function is given by:

$$\mathbf{J}_{neg-sample}(\mathbf{v}_c, o, \mathbf{U}) = -\log(\sigma(\mathbf{u}_o^T \mathbf{v}_c)) - \sum_{k=1}^K \log(\sigma(-\mathbf{u}_k^T \mathbf{v}_c)) \quad (5)$$

for a sample  $w_1, w_2, \dots, w_K$ , where  $\sigma(\cdot)$  is a sigmoid function. Please report parts (b) and (c), computing the partial derivatives of  $\mathbf{J}_{neg-sample}$  with respect to  $\mathbf{v}_c$ , with respect to  $\mathbf{u}_o$ , and with respect to a negative sample  $\mathbf{u}_k$ . Please write your answers in terms of the vectors  $\mathbf{u}_o$ ,  $\mathbf{v}_c$ , and  $\mathbf{u}_k$ , where  $k \in [1, K]$ . After you have done this, describe with one sentence why this loss function is much more efficient to compute than the naive-softmax loss.

Solution:

Partial derivatives of  $\mathbf{J}_{neg-sample}$  with respect to  $\mathbf{v}_c$ :

$$\begin{aligned} \frac{\partial \mathbf{J}_{neg-sample}(\mathbf{v}_c, o, \mathbf{U})}{\partial \mathbf{v}_c} &= -\frac{\partial \log(\sigma(\mathbf{u}_o^T \mathbf{v}_c))}{\partial \mathbf{v}_c} - \sum_{k=1}^K \frac{\partial \log(\sigma(-\mathbf{u}_k^T \mathbf{v}_c))}{\partial \mathbf{v}_c} \\ &= -\frac{1}{\sigma(\mathbf{u}_o^T \mathbf{v}_c)} \frac{\partial \sigma(\mathbf{u}_o^T \mathbf{v}_c)}{\partial \mathbf{v}_c} - \sum_{k=1}^K \frac{1}{\sigma(-\mathbf{u}_k^T \mathbf{v}_c)} \frac{\partial \sigma(-\mathbf{u}_k^T \mathbf{v}_c)}{\partial \mathbf{v}_c} \\ &= -\frac{\sigma(\mathbf{u}_o^T \mathbf{v}_c)(1 - \sigma(\mathbf{u}_o^T \mathbf{v}_c))}{\sigma(\mathbf{u}_o^T \mathbf{v}_c)} \mathbf{u}_o + \sum_{k=1}^K \frac{\sigma(-\mathbf{u}_k^T \mathbf{v}_c)(1 - \sigma(-\mathbf{u}_k^T \mathbf{v}_c))}{\sigma(-\mathbf{u}_k^T \mathbf{v}_c)} \mathbf{u}_k \\ &= -(1 - \sigma(\mathbf{u}_o^T \mathbf{v}_c)) \mathbf{u}_o + \sum_{k=1}^K (1 - \sigma(-\mathbf{u}_k^T \mathbf{v}_c)) \mathbf{u}_k \end{aligned}$$

Partial derivatives of  $\mathbf{J}_{neg-sample}$  with respect to  $\mathbf{u}_o$ :

$$\begin{aligned} \frac{\partial \mathbf{J}_{neg-sample}(\mathbf{v}_c, o, \mathbf{U})}{\partial \mathbf{u}_o} &= -\frac{\partial \log(\sigma(\mathbf{u}_o^T \mathbf{v}_c))}{\partial \mathbf{u}_o} - \sum_{k=1}^K \frac{\partial \log(\sigma(-\mathbf{u}_k^T \mathbf{v}_c))}{\partial \mathbf{u}_o} \\ &= -\frac{1}{\sigma(\mathbf{u}_o^T \mathbf{v}_c)} \frac{\partial \sigma(\mathbf{u}_o^T \mathbf{v}_c)}{\partial \mathbf{u}_o} - 0 = -\frac{\sigma(\mathbf{u}_o^T \mathbf{v}_c)(1 - \sigma(\mathbf{u}_o^T \mathbf{v}_c))}{\sigma(\mathbf{u}_o^T \mathbf{v}_c)} \mathbf{v}_c = (1 - \sigma(\mathbf{u}_o^T \mathbf{v}_c)) \mathbf{v}_c \end{aligned}$$

Partial derivatives of  $\mathbf{J}_{neg-sample}$  with respect to  $\mathbf{u}_k$ :

$$\begin{aligned} \frac{\partial \mathbf{J}_{neg-sample}(\mathbf{v}_c, o, \mathbf{U})}{\partial \mathbf{u}_k} &= -\frac{\partial \log(\sigma(\mathbf{u}_o^T \mathbf{v}_c))}{\partial \mathbf{u}_k} - \sum_{k=1}^K \frac{\partial \log(\sigma(-\mathbf{u}_k^T \mathbf{v}_c))}{\partial \mathbf{u}_k} \\ &= 0 - \frac{1}{\sigma(-\mathbf{u}_k^T \mathbf{v}_c)} \frac{\partial \sigma(-\mathbf{u}_k^T \mathbf{v}_c)}{\partial \mathbf{u}_k} = +\frac{\sigma(-\mathbf{u}_k^T \mathbf{v}_c)(1 - \sigma(-\mathbf{u}_k^T \mathbf{v}_c))}{\sigma(-\mathbf{u}_k^T \mathbf{v}_c)} \mathbf{v}_c = (1 - \sigma(-\mathbf{u}_k^T \mathbf{v}_c)) \mathbf{v}_c \end{aligned}$$

(f) Suppose the center word is  $c = w_t$  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 skip-gram version of word2vec, the total loss for the context window is:

$$\mathbf{J}_{\text{skip-gram}}(\mathbf{v}_c, w_{t-m}, \dots, w_{t+m}, \mathbf{U}) = \sum_{-m \leq j \leq m, j \neq 0} \mathbf{J}(\mathbf{v}_c, w_j, \mathbf{U})$$

Write down three partial derivatives:

- (i)  $\frac{\partial \mathbf{J}_{\text{skip-gram}}(\mathbf{v}_c, w_{t-m}, \dots, w_{t+m}, \mathbf{U})}{\partial \mathbf{U}}$
- (ii)  $\frac{\partial \mathbf{J}_{\text{skip-gram}}(\mathbf{v}_c, w_{t-m}, \dots, w_{t+m}, \mathbf{U})}{\partial \mathbf{v}_c}$
- (iii)  $\frac{\partial \mathbf{J}_{\text{skip-gram}}(\mathbf{v}_c, w_{t-m}, \dots, w_{t+m}, \mathbf{U})}{\partial \mathbf{v}_w}$  when  $w \neq c$

Solution:

$$(i) \frac{\partial \mathbf{J}_{\text{skip-gram}}(\mathbf{v}_c, w_{t-m}, \dots, w_{t+m}, \mathbf{U})}{\partial \mathbf{U}} = \sum_{-m \leq j \leq m, j \neq 0} \frac{\partial \mathbf{J}(\mathbf{v}_c, w_j, \mathbf{U})}{\partial \mathbf{U}}$$

$$(ii) \frac{\partial \mathbf{J}_{\text{skip-gram}}(\mathbf{v}_c, w_{t-m}, \dots, w_{t+m}, \mathbf{U})}{\partial \mathbf{v}_c} = \sum_{-m \leq j \leq m, j \neq 0} \frac{\partial \mathbf{J}(\mathbf{v}_c, w_j, \mathbf{U})}{\partial \mathbf{v}_c}$$

$$(iii) \frac{\partial \mathbf{J}_{\text{skip-gram}}(\mathbf{v}_c, w_{t-m}, \dots, w_{t+m}, \mathbf{U})}{\partial \mathbf{u}_w} = \sum_{-m \leq j \leq m, j \neq 0} \frac{\partial \mathbf{J}(\mathbf{v}_c, w_j, \mathbf{U})}{\partial \mathbf{u}_w} = 0$$