Stanford CS 224N Assignment #2 - Written Solution

December 13, 2020

1. Understanding word2vec

In word2vec

- conditional probability distribution is given by

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

- naive-softmax loss function is given by

$$\boldsymbol{J}_{naive-softmax}(\boldsymbol{v}_c, o, \boldsymbol{U}) = -logP(O = o|C = c) \tag{2}$$

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

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

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 $\boldsymbol{y}$is a one-hot encoded vector with a 1 for true outside word o, and 0 everwhere 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 $\boldsymbol{J}_{naive-softmax}(\boldsymbol{v}_c, o, \boldsymbol{U})$ with respect to $\boldsymbol{v}_c$. Please write your answer in terms of $\boldsymbol{y}$, $\hat{\boldsymbol{y}}$ and $\boldsymbol{U}$.

Solution:

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

$$= \frac{\partial - log(exp(\boldsymbol{u}_o^T \boldsymbol{v}_c))}{\partial \boldsymbol{v}_c} - \frac{\partial - log(\sum_{w \in Vocab} exp(\boldsymbol{u}_w^T \boldsymbol{v}_c))}{\partial \boldsymbol{v}_c}$$

$$= -\boldsymbol{u}_o + \frac{1}{\sum_{w \in Vocab} exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)} \frac{\partial \sum_{x \in Vocab} exp(\boldsymbol{u}_x^T \boldsymbol{v}_c)}{\partial \boldsymbol{v}_c}$$

$$= -\boldsymbol{u}_o + \frac{1}{\sum_{w \in Vocab} exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)} \sum_{x \in Vocab} \frac{\partial exp(\boldsymbol{u}_x^T \boldsymbol{v}_c)}{\partial \boldsymbol{v}_c}$$

$$= -\boldsymbol{u}_o + \frac{1}{\sum_{w \in Vocab} exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)} \sum_{x \in Vocab} exp(\boldsymbol{u}_x^T \boldsymbol{v}_c) \boldsymbol{u}_x$$

$$= -\boldsymbol{u}_o + \sum_{x \in Vocab} \frac{exp(\boldsymbol{u}_x^T \boldsymbol{v}_c)}{\sum_{w \in Vocab} exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)} \boldsymbol{u}_x$$

$$= -\boldsymbol{u}_o + \sum_{x \in Vocab} P(O = x | C = c) \boldsymbol{u}_x$$

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

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

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

Solution:

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

$$= \frac{\partial - log(exp(\boldsymbol{u}_o^T \boldsymbol{v}_c))}{\partial \boldsymbol{u}_w} - \frac{\partial - log(\sum_{w \in Vocab} exp(\boldsymbol{u}_w^T \boldsymbol{v}_c))}{\partial u_w}$$

If $w = o$, then

$$= \frac{\partial - log(exp(\boldsymbol{u}_o^T \boldsymbol{v}_c))}{\partial \boldsymbol{u}_o} - \frac{\partial - log(\sum_{w \in Vocab} exp(\boldsymbol{u}_w^T \boldsymbol{v}_c))}{\partial \boldsymbol{u}_o}$$

$$= -\boldsymbol{v}_c + \frac{1}{\sum_{w \in Vocab} exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)} \frac{\partial \sum_{w \in Vocab} exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)}{\partial \boldsymbol{u}_o}$$

$$= -\boldsymbol{v}_c + \frac{exp(\boldsymbol{u}_o^T \boldsymbol{v}_c)}{\sum_{w \in Vocab} exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)} \boldsymbol{v}_c$$

$$= (P(O = o|C = c) - 1)\boldsymbol{v}_c$$

If $w \neq o$, then

$$= \frac{\partial - log(exp(\boldsymbol{u}_o^T \boldsymbol{v}_c))}{\partial \boldsymbol{u}_w} - \frac{\partial - log(\sum_{w \in Vocab} exp(\boldsymbol{u}_w^T \boldsymbol{v}_c))}{\partial \boldsymbol{u}_w}$$

$$= 0 + \frac{1}{\sum_{w \in Vocab} exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)} \frac{\partial \sum_{w \in Vocab} exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)}{\partial \boldsymbol{u}_w}$$

$$= 0 + \frac{exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)}{\sum_{w \in Vocab} exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)} \boldsymbol{v}_c$$

$$= P(O = w|C = c)\boldsymbol{v}_c$$

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

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

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

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

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:

$$\frac{d\sigma(x)}{dx} = \frac{d\frac{e^x}{e^x+1}}{dx} = \frac{\frac{de^x}{dx}(e^x + 1) - e^x \frac{d(e^x+1)}{dx}}{(e^x + 1)^2} = \frac{e^x(e^x + 1) - e^x(e^x + 0)}{(e^x + 1)^2}$$

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

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

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

for a sample $w_1, w_2, ..., w_K$, where $\sigma(.)$ is a sigmoid function. Please report parts (b) and (c), computing the partial derivatives of $\boldsymbol{J}_{neg-sampe}$with respect to $\boldsymbol{v}_c$, with respect to $\boldsymbol{u}_o$, and with respect to a negative sample $\boldsymbol{u}_k$. Please write your answers in terms of the vectors $\boldsymbol{u}_o$, $\boldsymbol{v}_c$, and $\boldsymbol{u}_k$, 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 $\boldsymbol{J}_{neg-sampe}$with respect to $\boldsymbol{v}_c$:

$$\frac{\partial \boldsymbol{J}_{neg-sampe}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{v}_c} = -\frac{\partial log(\sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c))}{\partial \boldsymbol{v}_c} - \sum_{k=1}^{K} \frac{\partial log(\sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c))}{\partial \boldsymbol{v}_c}$$

$$= -\frac{1}{\sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c)} \frac{\partial \sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c)}{\partial \boldsymbol{v}_c} - \sum_{k=1}^{K} \frac{1}{\sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c)} \frac{\partial \sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c)}{\partial \boldsymbol{v}_c}$$

$$= -\frac{\sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c)(1-\sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c))}{\sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c)} \boldsymbol{u}_o + \sum_{k=1}^{K} \frac{\sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c)(1-\sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c))}{\sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c)} \boldsymbol{u}_k$$

$$= -(1-\sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c) \boldsymbol{u}_o + \sum_{k=1}^{K} (1-\sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c)) \boldsymbol{u}_k$$

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

$$\frac{\partial \boldsymbol{J}_{neg-sampe}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_o} = -\frac{\partial log(\sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c))}{\partial \boldsymbol{u}_o} - \sum_{k=1}^{K} \frac{\partial log(\sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c))}{\partial \boldsymbol{u}_o}$$

$$= -\frac{1}{\sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c)} \frac{\partial \sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c)}{\partial \boldsymbol{u}_o} - 0 = -\frac{\sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c)(1-\sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c))}{\sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c)} \boldsymbol{v}_c = (1-\sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c)) \boldsymbol{v}_c$$

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

$$\frac{\partial \boldsymbol{J}_{neg-sampe}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_k} = -\frac{\partial log(\sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c))}{\partial \boldsymbol{u}_k} - \sum_{k=1}^{K} \frac{\partial log(\sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c))}{\partial \boldsymbol{u}_k}$$

$$= -0 - \frac{1}{\sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c)} \frac{\partial \sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c)}{\partial \boldsymbol{u}_k} = +\frac{\sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c)(1-\sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c))}{\sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c)} \boldsymbol{v}_c = (1-\sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c)) \boldsymbol{v}_c$$

(f) Suppose the center word is $c = w_t$ and the context window is $[w_{t-m}, ..., w_{t-1}, w_t, w_{t+1}, ..., w_{t+m}]$, where m is the context window size. Recall that for skip-gram version of word2vec, the total loss for the context window is:

$$\boldsymbol{J}_{skip-gram}(\boldsymbol{v}_c, w_{t-m}, ..., w_{t+m}, \boldsymbol{U}) = \sum_{-m<=j<=m\ j\neq 0} \boldsymbol{J}(\boldsymbol{v}_c, w_j, \boldsymbol{U})$$

Write down three partial derivatives:
(i) $\frac{\partial \boldsymbol{J}_{skip-gram}(\boldsymbol{v}_c, w_{t-m}, ..., w_{t+m}, \boldsymbol{U})}{\partial \boldsymbol{U}}$
(ii) $\frac{\partial \boldsymbol{J}_{skip-gram}(\boldsymbol{v}_c, w_{t-m}, ..., w_{t+m}, \boldsymbol{U})}{\partial \boldsymbol{v}_c}$
(iii) $\frac{\partial \boldsymbol{J}_{skip-gram}(\boldsymbol{v}_c, w_{t-m}, ..., w_{t+m}, \boldsymbol{U})}{\partial \boldsymbol{v}_w}$ when $w \neq c$
Solution:

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

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

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