UNDERSTANDING WORD2VEC

YECHAN LEE

(1) For some given o, c

$$\begin{aligned}
&\mathbf{y} = [0, \cdots, 1 \text{ (c-th position)}, \cdots, 0] \\
&\hat{\mathbf{y}} = [P(O = 1|C = c), P(O = 2|C = c), \cdots] \\
&J_{\text{naitve-softmax}}(v_c, o, U) = -\log P(O = o|C = c) \\
&= -\log \hat{y}_0 = -\sum_{w \in \text{Vocab}} y_w \log \hat{y}_w
\end{aligned}$$

(2) Let D be the dimension of word vector, W be the number of words in vocabulary. Then, shape of $\mathbf{v}_c$ is $[D, 1]$, $\mathbf{U}$ is $[W, D]$, $\mathbf{y}, \hat{\mathbf{y}}$ is $[W, 1]$

$$\begin{aligned}
\frac{\partial J}{\partial \mathbf{v}_c} &= \frac{\partial}{\partial \mathbf{v}_c} - \log P(O = o|C = c) \\
&= \frac{\partial}{\partial \mathbf{v}_c} \left(-\mathbf{u}_0^T \mathbf{v}_c + \log \sum_{w \in \text{Vocab}} \exp \mathbf{u}_w^T \mathbf{v}_c \right) \\
&= -\mathbf{u}_0 + \frac{\sum_{w \in \text{Vocab}} \exp \left(\mathbf{u}_w^T \mathbf{v}_c \right) \cdot \mathbf{u}_w}{\sum_{w \in \text{Vocab}} \exp \mathbf{u}_w^T \mathbf{v}_c} \\
&= -\mathbf{u}_0 + \sum_{w \in \text{Vocab}} P(O = w|C = c) \cdot \mathbf{u}_w \\
&= -U^T \mathbf{y} + U^T \hat{\mathbf{y}} \\
&= U^T (\hat{\mathbf{y}} - \mathbf{y})
\end{aligned}$$

$U^T \hat{\mathbf{y}}$ has $[D, 1]$ shape because U^T has $[D, W]$ shape, and $\hat{\mathbf{y}}$ has $[W, 1]$ shape.

(3)

(a) If $w = o$ then,

$$\begin{aligned}
\frac{\partial J}{\partial \mathbf{u}_o} &= \frac{\partial}{\partial \mathbf{u}_o} - \log P(O = o|C = c) \\
&= \frac{\partial}{\partial \mathbf{u}_o} \left(-\mathbf{u}_0^T \mathbf{v}_c + \log \sum_{w \in \text{Vocab}} \exp \mathbf{u}_w^T \mathbf{v}_c \right) \\
&= -\mathbf{v}_c + \frac{\exp \left(\mathbf{u}_o^T \mathbf{v}_c \right) \mathbf{v}_c}{\sum_{w \in \text{Vocab}} \exp \left(\mathbf{u}_w^T \mathbf{v}_c \right) \cdot \mathbf{v}_c} \\
&= (P(O = o|C = c) - 1) \mathbf{v}_c \\
&= (\hat{y}_o - 1) \mathbf{v}_c
\end{aligned}$$

Date: July 4, 2021.

(b) If $w \neq o$ then,

$$\begin{aligned}\frac{\partial J}{\partial \mathbf{u}_w} &= \frac{\partial}{\partial \mathbf{u}_w} - \log P(O = o|C = c) \\ &= \frac{\partial}{\partial \mathbf{u}_w}\left(-\mathbf{u}_0^T\mathbf{v}_c + \log \sum_{w' \in \text{Vocab}} \exp \mathbf{u}_{w'}^T \mathbf{v}_c\right) \\ &= \frac{\exp\left(\mathbf{u}_w^T\mathbf{v}_c\right)\mathbf{v}_c}{\sum_{w' \in \text{Vocab}} \exp\left(\mathbf{u}_{w'}^T\mathbf{v}_c\right) \cdot \mathbf{v}_c} \\ &= P(O = w|C = c)\mathbf{v}_c \\ &= \hat{y}_w \mathbf{v}_c\end{aligned}$$

(4) Noted: We haved defined the length of Vocabulary as W.

$$\frac{\partial J}{\partial \mathbf{U}} = \begin{bmatrix} \frac{\partial J}{\partial \mathbf{u}_1} \\ \vdots \\ \frac{\partial J}{\partial \mathbf{u}_o} \\ \vdots \\ \frac{\partial J}{\partial \mathbf{u}_W} \end{bmatrix} = \begin{bmatrix} \hat{y}_1 \mathbf{v}_c^T \\ \vdots \\ (\hat{y}_0 - 1)\mathbf{v}_c^T \\ \vdots \\ \hat{y}_W \mathbf{v}_c^T \end{bmatrix} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_o - 1 \\ \vdots \\ \hat{y}_W \end{bmatrix} \mathbf{v}_c^T$$

(5) Differentiate the sigmoid function.

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

(6)

(a) Repeat (2). Differentiate J with respect to $\mathbf{v}_c$

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

(b) Repeat (3). Differentiate J with respect to $\mathbf{u}_o$

$$\begin{aligned}\frac{\partial}{\partial \mathbf{u}_0} J &= -\frac{\partial}{\partial \mathbf{u}_0}\log(\sigma(\mathbf{u}_o^T\mathbf{v}_c)) \\ &= -\frac{1}{\sigma(\mathbf{u}_o^T\mathbf{v}_c)}\frac{\partial}{\partial \mathbf{u}_0}\sigma(\mathbf{u}_o^T\mathbf{v}_c) \\ &= -\frac{1}{\sigma(\mathbf{u}_o^T\mathbf{v}_c)}\sigma(\mathbf{u}_o^T\mathbf{v}_c)(1-\sigma(\mathbf{u}_o^T\mathbf{v}_c))\frac{\partial}{\partial \mathbf{u}_0}\mathbf{u}_o^T\mathbf{v}_c \\ &= (\sigma(\mathbf{u}_o^T\mathbf{v}_c) - 1)\mathbf{v}_c\end{aligned}$$

(c) Repeat (3). Differentiate J with respect to $\mathbf{u}_k$

$$\begin{aligned}\frac{\partial}{\partial \mathbf{u}_k} J &= -\frac{\partial}{\partial \mathbf{u}_k}\log(\sigma(-\mathbf{u}_k^T\mathbf{v}_c)) \\ &= (1-\sigma(-\mathbf{u}_k^T\mathbf{v}_c))\mathbf{v}_c\end{aligned}$$

The reason is that it takes $O(W^2)$ times to calculate $\hat{\mathbf{y}}, \mathbf{y}$. Therefore, it takes quadratic time to compute the native-softmax loss. On the other hand, it takes $O(k)$ times to compute the Negative Sampling loss.

(7) Repeat the previous exercise without the distinct sampling assumption. As you can see, calculating the derivative with respect to $\mathbf{v}_c, \mathbf{u}_o$ does not use the assumption. Therefore, these derivatives are the same as the previous ones.

$$\frac{\partial}{\partial \mathbf{u}_k} J = -\sum_{w_k = w_{k'}} \log(\sigma(-\mathbf{u}_k^T \mathbf{v}_c)) = (1 - \sigma(-\mathbf{u}_k^T \mathbf{v}_c))\mathbf{v}_c \times \sum_{k'=1}^{K} [w_k = w_{k'}]$$

, where $[\text{true}] = 1, [\text{false}] = 0$.

(8)

$$\frac{\partial J_{\text{skip-gram}}(\mathbf{v}_c, w_{t-m}, \dots, w_{t+m}, \mathbf{U})}{\partial \mathbf{U}} = \sum_{-m \le j \le m, j \ne 0} \frac{\partial J(\mathbf{v}_c, w_{t+j}, \mathbf{U})}{\partial \mathbf{U}}$$

$$\frac{\partial J_{\text{skip-gram}}(\mathbf{v}_c, w_{t-m}, \dots, w_{t+m}, \mathbf{U})}{\partial \mathbf{v}_c} = \sum_{-m \le j \le m, j \ne 0} \frac{\partial J(\mathbf{v}_c, w_{t+j}, \mathbf{U})}{\partial \mathbf{v}_c}$$

$$\frac{\partial J_{\text{skip-gram}}(\mathbf{v}_c, w_{t-m}, \dots, w_{t+m}, \mathbf{U})}{\partial \mathbf{v}_w} = 0$$