CS224N A2 Written

Yannik Kumar

July 2020

1 Understanding word2vec

a)

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

Since $\mathbf{y}$ is a one-hot vector and y_w is 0 for every term in the summation on the LHS except when $w = o$, in which case $y_w = 1$ and the LHS equals the RHS.

b)

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{v}_c} - \log\left(\frac{\exp(\mathbf{u}_o^\intercal \mathbf{v}_c)}{\sum\limits_{w \in V} \exp(\mathbf{u}_w^\intercal \mathbf{v}_c)}\right) &= \frac{\partial}{\partial \mathbf{v}_c} - \log\left(\exp(\mathbf{u}_o^\intercal \mathbf{v}_c)\right) + \frac{\partial}{\partial \mathbf{v}_c} \log\left(\sum_{w \in V} \exp(\mathbf{u}_w^\intercal \mathbf{v}_c)\right) \\
&= -\frac{\partial}{\partial \mathbf{v}_c} \mathbf{u}_o^\intercal \mathbf{v}_c + \frac{\partial}{\partial \mathbf{v}_c} \log\left(\sum_{w \in V} \exp(\mathbf{u}_w^\intercal \mathbf{v}_c)\right) \\
&= -\mathbf{u}_o + \frac{\partial}{\partial \mathbf{v}_c} \log\left(\sum_{w \in V} \exp(\mathbf{u}_w^\intercal \mathbf{v}_c)\right) \\
&= -\mathbf{u}_o + \frac{1}{\sum\limits_{w \in V} \exp(\mathbf{u}_w^\intercal \mathbf{v}_c)} \cdot \frac{\partial}{\partial \mathbf{v}_c} \sum_{x \in V} \exp(\mathbf{u}_x^\intercal \mathbf{v}_c) \\
&= -\mathbf{u}_o + \frac{1}{\sum\limits_{w \in V} \exp(\mathbf{u}_w^\intercal \mathbf{v}_c)} \cdot \sum_{x \in V} \exp(\mathbf{u}_x^\intercal \mathbf{v}_c) \mathbf{u}_x \\
&= -\mathbf{u}_o + \sum_{x \in V} \frac{\exp(\mathbf{u}_x^\intercal \mathbf{v}_c)}{\sum\limits_{w \in V} \exp(\mathbf{u}_w^\intercal \mathbf{v}_c)} \cdot \mathbf{u}_x
\end{aligned}$$

$-\mathbf{u_o} = -\mathbf{U}\mathbf{y}$ since $\mathbf{y}$ is a one-hot vector with a 1 at index o. The fraction represents $p(x|c)$ (probability of the context word given the center word according to our model). The summation over the vocabulary with each $p(x|c)$ multiplied with $\mathbf{u_x}$ gives you a vector that is a weighted average of the vectors of $\mathbf{U}$ – a vector the models believes is the context word. This is equivalent to $\mathbf{U}\hat{\mathbf{y}}$.

$$\therefore \frac{\partial \mathbf{J}}{\partial \mathbf{v_c}} = \mathbf{U}\hat{\mathbf{y}} - \mathbf{U}\mathbf{y}$$

c) Case one, when $w = o$ (outside word vector is the true context vector):

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{u}_w} - \log\left(\frac{\exp(\mathbf{u}_o^\intercal \mathbf{v}_c)}{\sum\limits_{w \in V} \exp(\mathbf{u}_w^\intercal \mathbf{v}_c)}\right) &= \frac{\partial}{\partial \mathbf{u}_w} - \log\left(\exp(\mathbf{u}_o^\intercal \mathbf{v}_c)\right) + \frac{\partial}{\partial \mathbf{u}_w} \log\left(\sum_{w \in V} \exp(\mathbf{u}_w^\intercal \mathbf{v}_c)\right) \\
&= -\mathbf{v}_c + \frac{\partial}{\partial \mathbf{u}_w} \log\left(\sum_{w \in V} \exp(\mathbf{u}_w^\intercal \mathbf{v}_c)\right) \\
&= -\mathbf{v}_c + \frac{1}{\sum\limits_{w \in V} \exp(\mathbf{u}_w^\intercal \mathbf{v}_c)} \cdot \frac{\partial}{\partial \mathbf{u}_w} \sum_{x \in V} \exp(\mathbf{u}_x^\intercal \mathbf{v}_c) \\
&= -\mathbf{v}_c + \frac{1}{\sum\limits_{w \in V} \exp(\mathbf{u}_w^\intercal \mathbf{v}_c)} \cdot \exp(\mathbf{u}_x^\intercal \mathbf{v}_c)\mathbf{v}_c \\
&= -\mathbf{v}_c + \frac{\exp(\mathbf{u}_x^\intercal \mathbf{v}_c)}{\sum\limits_{w \in V} \exp(\mathbf{u}_w^\intercal \mathbf{v}_c)} \mathbf{v}_c \\
&= \hat{y}_w \mathbf{v}_c - \mathbf{v}_c \\
&= \mathbf{v}_c(\hat{y}_o - y_o)
\end{aligned}$$

Case two, when $w \neq o$ (outside vector is not the true context vector):

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{u}_w} - \log\left(\frac{\exp(\mathbf{u}_o^\intercal \mathbf{v}_c)}{\sum\limits_{w \in V} \exp(\mathbf{u}_w^\intercal \mathbf{v}_c)}\right) &= \frac{\partial}{\partial \mathbf{u}_w} - \log\left(\exp(\mathbf{u}_o^\intercal \mathbf{v}_c)\right) + \frac{\partial}{\partial \mathbf{u}_w} \log\left(\sum_{w \in V} \exp(\mathbf{u}_w^\intercal \mathbf{v}_c)\right) \\
&= 0 + \frac{\partial}{\partial \mathbf{u}_w} \log\left(\sum_{w \in V} \exp(\mathbf{u}_w^\intercal \mathbf{v}_c)\right) \\
&= \frac{1}{\sum\limits_{w \in V} \exp(\mathbf{u}_w^\intercal \mathbf{v}_c)} \cdot \frac{\partial}{\partial \mathbf{u}_w} \sum_{x \in V} \exp(\mathbf{u}_x^\intercal \mathbf{v}_c) \\
&= \frac{\exp(\mathbf{u}_x^\intercal \mathbf{v}_c}{\sum\limits_{w \in V} \exp(\mathbf{u}_w^\intercal \mathbf{v}_c)} \cdot \mathbf{v}_c \\
&= \hat{y}_{x \neq o} \mathbf{v}_c
\end{aligned}$$

d) With $\mathbf{x} \in \mathbb{R}^n$, $\sigma(\mathbf{x})$ has n inputs and n outputs, giving it an $n \times n$ jacobian.

$$\frac{d}{d\mathbf{x}} \frac{1}{1 + e^{-\mathbf{x}}} = \frac{d\sigma(\mathbf{x})}{d\mathbf{x}} \in \mathbb{R}^{n \times n}$$

.

When $i = j$:

$$\begin{aligned}
\left(\frac{d\sigma(\mathbf{x})}{d\mathbf{x}}\right)_{i,j} &= \frac{d}{dx_j} \frac{1}{1 + e^{-x_i}} \\
&= \frac{d}{dx_j} (1 + e^{-x_i})^{-1} \\
&= -1 \cdot (1 + e^{-x_i})^{-2} \cdot e^{-x_i} \cdot -1 \\
&= \frac{e^{-x_i}}{(1 + e^{-x_i})^2}
\end{aligned}$$

When $i \neq j$:

$$\left(\frac{d\sigma(\mathbf{x})}{d\mathbf{x}}\right)_{i,j} = 0$$

$$
\begin{aligned}
\therefore \frac{d\sigma(\mathbf{x})}{d\mathbf{x}} &= \operatorname{diag}\left(\frac{1}{1+e^{-\mathbf{x}}}\right) \\
&= \operatorname{diag}\left(\left(\frac{1+e^{-\mathbf{x}}-1}{1+e^{-\mathbf{x}}}\right)\left(\frac{1}{1+e^{-\mathbf{x}}}\right)\right) \\
&= \operatorname{diag}\left(\left(-\frac{1}{1+e^{-\mathbf{x}}}\right)\left(\frac{1}{1+e^{-\mathbf{x}}}\right)\right) \\
&= \operatorname{diag}\left((1-\sigma(\mathbf{x}))\sigma(\mathbf{x})\right) \\
&= \operatorname{diag}(\sigma'(\mathbf{x}))
\end{aligned}
$$

e) w.r.t. $\mathbf{v_c}$:

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

w.r.t. $\mathbf{u}_o$:

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

w.r.t. $\mathbf{u}_k$:

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

The negative sampling loss is more efficient than the naive soft-max loss because it involves no summations over the vocab, i.e. $O(K)$ vs. $O(|V|)$.

f) i)

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

ii)

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

iii)

$$\frac{\partial \mathbf{J}_{skip-gram}(\mathbf{v}_c, w_{t-m}, ..., w_{t+m}, \mathbf{U})}{\partial \mathbf{v}_w} \text{ when } w \neq c = 0$$