CS 224n Assignment #2: Written Assignment

Yoshihiro Kumazawa

May 28, 2020

(a)

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

(b)

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{v}_c} \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) &= \frac{\partial}{\partial \boldsymbol{v}_c} \left(-\log \left(\frac{\exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}{\sum_{w \in Vocab} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)} \right) \right) \\
&= \frac{\partial}{\partial \boldsymbol{v}_c} \left(\log \left(\sum_{w \in Vocab} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c) \right) - \log \left(\exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c) \right) \right) \\
&= \frac{\partial}{\partial \boldsymbol{v}_c} \log \left(\sum_{w \in Vocab} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c) \right) - \frac{\partial}{\partial \boldsymbol{v}_c} \boldsymbol{u}_o^\top \boldsymbol{v}_c \\
&= \frac{\sum_{x \in Vocab} \exp(\boldsymbol{u}_x^\top \boldsymbol{v}_c) \boldsymbol{u}_x}{\sum_{w \in Vocab} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)} - \boldsymbol{u}_o \\
&= \sum_{x \in Vocab} \frac{\exp(\boldsymbol{u}_x^\top \boldsymbol{v}_c)}{\sum_{w \in Vocab} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)} \boldsymbol{u}_x - \boldsymbol{u}_o \\
&= \sum_{x \in Vocab} \hat{y}_x \boldsymbol{u}_x - \sum_{x \in Vocab} y_x \boldsymbol{u}_x \\
&= \sum_{x \in Vocab} \boldsymbol{u}_x (\hat{y}_x - y_x) \\
&= \boldsymbol{U}(\hat{\boldsymbol{y}} - \boldsymbol{y}).
\end{aligned}
$$

(c)

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{u}_w} \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) &= \frac{\partial}{\partial \boldsymbol{u}_w} \log \left(\sum_{w \in Vocab} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c) \right) - \frac{\partial}{\partial \boldsymbol{u}_w} \boldsymbol{u}_o^\top \boldsymbol{v}_c \\
&= \frac{\exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}{\sum_{w \in Vocab} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)} \boldsymbol{v}_c - y_w \boldsymbol{v}_c \\
&= \hat{y}_w \boldsymbol{v}_c - y_w \boldsymbol{v}_c \\
&= \boldsymbol{v}_c (\hat{y}_w - y_w).
\end{aligned}
$$

(d)

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

(e)

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

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

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

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

$$\frac{\partial}{\partial \boldsymbol{u}_o} \boldsymbol{J}_{\text{neg-sample}}(\boldsymbol{v}_c, o, \boldsymbol{U}) = -\frac{\partial}{\partial \boldsymbol{u}_o} \log(\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)) = -(1 - \sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c))\boldsymbol{v}_c.$$

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

These are computationally less expensive than the naive-softmax loss because its summation ranges over only K numbers, which is usually much smaller than $|Vocab|$.

(f) (i) $\partial \boldsymbol{J}_{\text{skip-gram}}(\boldsymbol{v}_c, w_{t-m}, \ldots, w_{t+m}, \boldsymbol{U})/\partial \boldsymbol{U} = \sum\limits_{\substack{-m \leq j \leq m \\ j \neq 0}} \partial \boldsymbol{J}(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U})/\partial \boldsymbol{U}.$

(ii) $\partial \boldsymbol{J}_{\text{skip-gram}}(\boldsymbol{v}_c, w_{t-m}, \ldots, w_{t+m}, \boldsymbol{U})/\partial \boldsymbol{v}_c = \sum\limits_{\substack{-m \leq j \leq m \\ j \neq 0}} \partial \boldsymbol{J}(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U})/\partial \boldsymbol{v}_c.$

(iii) $\partial \boldsymbol{J}_{\text{skip-gram}}(\boldsymbol{v}_c, w_{t-m}, \ldots, w_{t+m}, \boldsymbol{U})/\partial \boldsymbol{v}_w = 0.$