Natural Language Processing: Homework 1

Taewook Kang

September 24, 2024

Problem (a)

Prove that the loss $J_{\text{naive-softmax}}$ is the same as the cross-entropy loss between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$.

Solution. The elements of one-hot vector $\boldsymbol{y}_w$ (scalar) can be defined as:

$$\boldsymbol{y}_w = \begin{cases} 1 & \text{if } w = o \\ 0 & \text{else.} \end{cases} \tag{1}$$

Thus, the cross-entropy loss can be reformulated by utilizing the elements of $\boldsymbol{y}_w$ as follows:

$$\begin{aligned} -\sum_{w \in \text{Vocab}} \boldsymbol{y}_w \log(\hat{\boldsymbol{y}}_w) &= -\boldsymbol{y}_o \log(\hat{\boldsymbol{y}}_o) - \sum_{\substack{w \in \text{Vocab} \\ w \neq o}} \boldsymbol{y}_w \log(\hat{\boldsymbol{y}}_w) & (2) \\ &= -1 \times \log(\hat{\boldsymbol{y}}_o) - \sum_{\substack{w \in \text{Vocab} \\ w \neq o}} 0 \times \log(\hat{\boldsymbol{y}}_w) & (3) \\ &= -\log(\hat{\boldsymbol{y}}_o). & (4) \end{aligned}$$

$$\therefore J_{\text{naive-softmax}} = -\log(\hat{\boldsymbol{y}}_o) = -\sum_{w \in \text{Vocab}} \boldsymbol{y}_w \log(\hat{\boldsymbol{y}}_w) \tag{5}$$

Problem (b)

Compute the partial derivative of $J_{\text{naive-softmax}}(\boldsymbol{v}_c, o, U)$ with respect to $\boldsymbol{v}_c$.

Solution.

$$J_{\text{naive-softmax}} = -\log \frac{\exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}{\sum_{w\in\text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)} \tag{6}$$

$$= -\log \exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c) + \log \sum_{w\in\text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c) \tag{7}$$

$$= -\boldsymbol{u}_o^\top \boldsymbol{v}_c + \log \sum_{w\in\text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c). \tag{8}$$

Using equation (8), we can express the partial derivative $\frac{\partial J_{\text{naive-softmax}}}{\partial \boldsymbol{v}_c}$ as follows:

$$\frac{\partial J_{\text{naive-softmax}}}{\partial \boldsymbol{v}_c} = -\frac{\partial}{\partial \boldsymbol{v}_c}\boldsymbol{u}_o^\top \boldsymbol{v}_c + \frac{\partial}{\partial \boldsymbol{v}_c} \log \sum_{w\in\text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c). \tag{9}$$

The first term of (9), $\frac{\partial}{\partial \boldsymbol{v}_c}\boldsymbol{u}_o^\top \boldsymbol{v}_c$, can be computed as follows:

$$-\frac{\partial}{\partial \boldsymbol{v}_c}\boldsymbol{u}_o^\top \boldsymbol{v}_c = -\boldsymbol{u}_o. \tag{10}$$

The second term of (9) can be computed as follows:

$$\frac{\partial}{\partial \boldsymbol{v}_c} \log \sum_{w\in\text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c) = \frac{1}{\sum_{w\in\text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)} \left[\frac{\partial}{\partial \boldsymbol{v}_c} \sum_{x\in\text{Vocab}} \exp \boldsymbol{u}_x^\top \boldsymbol{v}_c\right] \tag{11}$$

$$= \frac{1}{\sum_{w\in\text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)} \left[\sum_{x\in\text{Vocab}} \boldsymbol{u}_x \exp \boldsymbol{u}_x^\top \boldsymbol{v}_c\right] \tag{12}$$

$$= \sum_{x\in\text{Vocab}} \frac{\boldsymbol{u}_x \exp \boldsymbol{u}_x^\top \boldsymbol{v}_c}{\sum_{w\in\text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)} \tag{13}$$

$$= \sum_{x\in\text{Vocab}} \boldsymbol{u}_x p(x|c). \tag{14}$$

Using (10) and (14), (9) can be expressed as the following equation:

$$\frac{\partial J_{\text{naive-softmax}}(\boldsymbol{v}_c, o, U)}{\partial \boldsymbol{v}_c} = -\boldsymbol{u}_o + \sum_{x\in\text{Vocab}} \boldsymbol{u}_x p(x|c). \tag{15}$$

Since $\boldsymbol{y}$ is a one-hot vector with a 1 at the position corresponding to the true outside word o and 0 everywhere else, the following relationship holds::

$$\boldsymbol{u}_o = U\boldsymbol{y}. \tag{16}$$

Also, given that $p(x|c) = \hat{\boldsymbol{y}}_x$, the relationship can be expressed as:

$$\sum_{x\in\text{Vocab}} \mathbf{u}_x p(x|c) = U\hat{\boldsymbol{y}}. \tag{17}$$

Therefore, using (16) and (17), equation (15) can be reformulated in terms of $\boldsymbol{y}, \hat{\boldsymbol{y}}$, and U.

$$\frac{\partial J_{\text{naive-softmax}}(\boldsymbol{v}_c, o, U)}{\partial \boldsymbol{v}_c} = U(\hat{\boldsymbol{y}} - \boldsymbol{y}). \tag{18}$$

Problem (c)

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

Solution. By equation (8), $\frac{\partial J_{\text{naive-softmax}}(\boldsymbol{v}_c, o, U)}{\partial \boldsymbol{u}_w}$ can be calculated as follows:

$$\frac{\partial J_{\text{naive-softmax}}(\boldsymbol{v}_c, o, U)}{\partial \boldsymbol{u}_w} = \frac{\partial}{\partial \boldsymbol{u}_w} \left\{ -\boldsymbol{u}_o^\top \boldsymbol{v}_c + \log \sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c) \right\}. \tag{19}$$

There are two cases for w: either $w = o$ or $w \neq o$.
i) $w = o$
The first term of (19) can be calculated as follows, since $w = o$:

$$\frac{\partial}{\partial \boldsymbol{u}_w} - \boldsymbol{u}_o^\top \boldsymbol{v}_c = \frac{\partial}{\partial \boldsymbol{u}_o} - \boldsymbol{u}_o^\top \boldsymbol{v}_c = -\boldsymbol{v}_c. \tag{20}$$

The second term of (19) can be calculated as follows:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{u}_w} \log \sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c) &= \frac{1}{\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)} \left\{ \frac{\partial}{\partial \boldsymbol{u}_w} \sum_{x \in \text{Vocab}} \exp(\boldsymbol{u}_x^\top \boldsymbol{v}_c) \right\} && (21) \\
&= \frac{1}{\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)} \left\{ \frac{\partial}{\partial \boldsymbol{u}_o} \sum_{x \in \text{Vocab}} \exp(\boldsymbol{u}_x^\top \boldsymbol{v}_c) \right\} && (22) \\
&= \frac{\exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c) \boldsymbol{v}_c}{\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)} && (23) \\
&= p(o|c) \boldsymbol{v}_c && (24) \\
&= \hat{\boldsymbol{y}}_o \boldsymbol{v}_c. && (25)
\end{aligned}$$

Therefore, (19) can be expressed as the following equation by combining the results of (20) and (25):

$$\frac{\partial J_{\text{naive-softmax}}(\boldsymbol{v}_c, o, U)}{\partial \boldsymbol{u}_w} = (\hat{\boldsymbol{y}}_w - 1)\boldsymbol{v}_c = (\hat{\boldsymbol{y}}_o - 1)\boldsymbol{v}_c. \tag{26}$$

ii) $w \neq o$
The first term of (19) can be calculated as follows, since $w \neq o$:

$$\frac{\partial}{\partial \boldsymbol{u}_w} - \boldsymbol{u}_o^\top \boldsymbol{v}_c = 0 \tag{27}$$

The second term of (19) can be calculated as follows:

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{u}_w} \log \sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c) &= \frac{1}{\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)} \left\{ \frac{\partial}{\partial \boldsymbol{u}_w} \sum_{x \in \text{Vocab}} \exp(\boldsymbol{u}_x^\top \boldsymbol{v}_c) \right\} && (28) \\
&= \frac{1}{\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)} \frac{\partial \{ \cdots + \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c) + \cdots \}}{\partial \boldsymbol{u}_w} && (29) \\
&= \frac{\exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c) \boldsymbol{v}_c}{\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)} && (30) \\
&= p(w|c) \boldsymbol{v}_c && (31) \\
&= \hat{\boldsymbol{y}}_w \boldsymbol{v}_c. && (32)
\end{aligned}
$$

By combining the results from equations (27) and (32), we can express equation (19) as follows:

$$
\frac{\partial J_{\text{naive-softmax}}(\boldsymbol{v}_c, o, U)}{\partial \boldsymbol{u}_w} = \hat{\boldsymbol{y}}_w \boldsymbol{v}_c. \tag{33}
$$

Problem (d)

Write down the partial derivative of $J_{\text{naive-softmax}}(\boldsymbol{v}_c, o, U)$ with respect to U.

Solution. By equation 8, $\frac{\partial J_{\text{naive-softmax}}(\boldsymbol{v}_c, o, U)}{\partial U}$ can be calculated as follows:

$$
\begin{aligned}
\frac{\partial J_{\text{naive-softmax}}(\boldsymbol{v}_c, o, U)}{\partial U} &= \frac{\partial}{\partial U} \left[-\boldsymbol{u}_o^\top \boldsymbol{v}_c + \log \sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c) \right] && (34) \\
&= -\frac{\partial \boldsymbol{u}_o^\top \boldsymbol{v}_c}{\partial U} + \frac{\partial}{\partial U} \left[\log \sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c) \right]. && (35)
\end{aligned}
$$

The first term of (35) can be calculated as follows, since $\boldsymbol{y}$ is an one-hot vector:

$$
-\frac{\partial \boldsymbol{u}_o^\top \boldsymbol{v}_c}{\partial U} = -\boldsymbol{v}_c \boldsymbol{y}^\top. \tag{36}
$$

The second term of (35) can be expressed as follows:

$$
\begin{aligned}
\frac{\partial}{\partial U} \left[\log \sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c) \right] &= \frac{1}{\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)} \left\{ \frac{\partial}{\partial U} \sum_{x \in \text{Vocab}} \exp(\boldsymbol{u}_x^\top \boldsymbol{v}_c) \right\} && (37) \\
&= \sum_{x \in \text{Vocab}} \frac{\exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}{\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)} \frac{\partial \boldsymbol{u}_w^\top \boldsymbol{v}_c}{\partial U} && (38) \\
&= \sum_{x \in \text{Vocab}} \hat{\boldsymbol{y}}_x \frac{\partial \boldsymbol{u}_w^\top \boldsymbol{v}_c}{\partial U} && (39) \\
&= \boldsymbol{v}_c \hat{\boldsymbol{y}}^\top. && (40)
\end{aligned}
$$

Therefore, using (36) and (40), equation (35) can be reformulated as a following equation:

$$\begin{aligned}
\frac{\partial J_{\text{naive-softmax}}(\boldsymbol{v}_c, o, U)}{\partial U} &= -\boldsymbol{v}_c \boldsymbol{y}^\top + \boldsymbol{v}_c \hat{\boldsymbol{y}}^\top && (41)\\
&= \boldsymbol{v_c}(\hat{\boldsymbol{y}} - \boldsymbol{y})^\top && (42)\\
&= \begin{bmatrix} \hat{y}_1 \boldsymbol{v}_c & \hat{y}_2 \boldsymbol{v}_c & \cdots & (\hat{y}_o - 1)\boldsymbol{v}_c & \cdots & \hat{y}_{|\text{Vocab}|}\boldsymbol{v}_c \end{bmatrix}. && (43)
\end{aligned}$$

Problem (e.i)

Please repeat parts (b) and (c), computing the partial derivatives of $J_{\text{neg-sample}}$ with respect to $\mathbf{v}_c$, with respect to $\mathbf{u}_o$, and with respect to the sth negative sample $\mathbf{u}_{w_s}$.

Solution. The derivative of a sigmoide function is computed as follows:

$$\frac{\partial \sigma(\mathbf{x})}{\partial \mathbf{x}} = \sigma(\mathbf{x})(1 - \sigma(\mathbf{x})). \tag{44}$$

Based on (44), all derivatives of $J_{\text{neg-sample}}$ can be expressed as follows:

$$\begin{aligned}
\frac{\partial J_{\text{neg-sample}}(\mathbf{v}_c, o, U)}{\partial \mathbf{v}_c} &= \frac{\partial}{\partial \mathbf{v}_c}\left[-\log(\sigma(\mathbf{u}_o^\top \mathbf{v}_c)) - \sum_{s=1}^{K} \log(\sigma(-\mathbf{u}_{w_s}^\top \mathbf{v}_c))\right] && (45)\\
&= -\frac{\sigma(\mathbf{u}_o^\top \mathbf{v}_c)(1 - \sigma(\mathbf{u}_o^\top \mathbf{v}_c))}{\sigma(\mathbf{u}_o^\top \mathbf{v}_c)} \frac{\partial \sigma(\mathbf{u}_o^\top \mathbf{v}_c)}{\partial \mathbf{v}_c} && (46)\\
&\quad - \sum_{s=1}^{K} \frac{\sigma(-\mathbf{u}_{w_s}^\top \mathbf{v}_c)(1 - \sigma(\mathbf{u}_{w_s}^\top \mathbf{v}_c))}{\sigma(-\mathbf{u}_{w_s}^\top \mathbf{v}_c)} \frac{\partial \sigma(-\mathbf{u}_{w_s}^\top \mathbf{v}_c)}{\partial \mathbf{v}_c} && (47)\\
&= -(1 - \sigma(\mathbf{u}_o^\top \mathbf{v}_c)) \frac{\partial \sigma(\mathbf{u}_o^\top \mathbf{v}_c)}{\partial \mathbf{v}_c} - \sum_{s=1}^{K} (1 - \sigma(-\mathbf{u}_{w_s}^\top \mathbf{v}_c)) \frac{\partial \sigma(-\mathbf{u}_{w_s}^\top \mathbf{v}_c)}{\partial \mathbf{v}_c} && (48)\\
&= -(1 - \sigma(\mathbf{u}_o^\top \mathbf{v}_c))\mathbf{u}_o + \sum_{s=1}^{K} (1 - \sigma(-\mathbf{u}_{w_s}^\top \mathbf{v}_c))\mathbf{u}_{w_s}. && (49)
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial J_{\text{neg-sample}}(\mathbf{v}_c, o, U)}{\partial \mathbf{u}_o} &= \frac{\partial}{\partial \mathbf{u}_o}\left[-\log(\sigma(\mathbf{u}_o^\top \mathbf{v}_c)) - \sum_{s=1}^{K}\log(\sigma(-\mathbf{u}_{w_s}^\top \mathbf{v}_c))\right] && (50)\\
&= -\frac{\partial \log(\sigma(\mathbf{u}_o^\top \mathbf{v}_c))}{\partial \mathbf{u}_o} - \frac{\partial}{\partial \mathbf{u}_o}\sum_{s=1}^{K}\log(\sigma(-\mathbf{u}_{w_s}^\top \mathbf{v}_c)) && (51)\\
&= -\frac{\partial \log(\sigma(\mathbf{u}_o^\top \mathbf{v}_c))}{\partial \mathbf{u}_o} - 0 && (52)\\
&= -\frac{\sigma(\mathbf{u}_o^\top \mathbf{v}_c)(1-\sigma(\mathbf{u}_o^\top \mathbf{v}_c))}{\sigma(\mathbf{u}_o^\top \mathbf{v}_c)}\frac{\partial \sigma(\mathbf{u}_o^\top \mathbf{v}_c)}{\partial \mathbf{u}_o} && (53)\\
&= -(1-\sigma(\mathbf{u}_o^\top \mathbf{v}_c))\frac{\partial \sigma(\mathbf{u}_o^\top \mathbf{v}_c)}{\partial \mathbf{u}_o} && (54)\\
&= -(1-\sigma(\mathbf{u}_o^\top \mathbf{v}_c))\mathbf{v}_c. && (55)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial J_{\text{neg-sample}}(\mathbf{v}_c, o, U)}{\partial \mathbf{u}_{w_s}} &= \frac{\partial}{\partial \mathbf{u}_{w_s}}\left[-\log(\sigma(\mathbf{u}_o^\top \mathbf{v}_c)) - \sum_{s=1}^{K}\log(\sigma(-\mathbf{u}_{w_s}^\top \mathbf{v}_c))\right] && (56)\\
&= -\frac{\partial \log(\sigma(\mathbf{u}_o^\top \mathbf{v}_c))}{\partial \mathbf{u}_{w_s}} - \frac{\partial}{\partial \mathbf{u}_{w_s}}\sum_{s=1}^{K}\log(\sigma(-\mathbf{u}_{w_s}^\top \mathbf{v}_c)) && (57)\\
&= -0 - \frac{\partial}{\partial \mathbf{u}_{w_s}}\sum_{s=1}^{K}\log(\sigma(-\mathbf{u}_{w_s}^\top \mathbf{v}_c)) && (58)\\
&= -\frac{\sigma(-\mathbf{u}_{w_s}^\top \mathbf{v}_c)(1-\sigma(-\mathbf{u}_{w_s}^\top \mathbf{v}_c))}{\sigma(-\mathbf{u}_{w_s}^\top \mathbf{v}_c)}\frac{\partial \sigma(-\mathbf{u}_{w_s}^\top \mathbf{v}_c)}{\partial \mathbf{u}_{w_s}} && (59)\\
&= -(1-\sigma(-\mathbf{u}_{w_s}^\top \mathbf{v}_c))\frac{\partial \sigma(-\mathbf{u}_{w_s}^\top \mathbf{v}_c)}{\partial \mathbf{u}_{w_s}} && (60)\\
&= (1-\sigma(-\mathbf{u}_{w_s}^\top \mathbf{v}_c))\mathbf{v}_c. && (61)
\end{aligned}
$$

Problem (e.ii)

Describe with one sentence why this loss function is much more efficient to compute than the naive-softmax loss.

Solution. Negative sampling is more efficient than naive-softmax because it reduces computational complexity by sampling K negative words instead of traversing the entire vocabulary, while still approximating the naive-softmax result effectively through Monte-Carlo sampling ($K << |\text{Vocab}|$).

Problem (f.1)

Write down partial derivative: $\frac{\partial J_{\text{skip-gram}}(\boldsymbol{v}_c, w_{t-m}, \cdots, w_{t+m}, U)}{\partial U}$.

Solution.

$$\frac{\partial J_{\text{skip-gram}}(\boldsymbol{v}_c, w_{t-m}, \cdots, w_{t+m}, U)}{\partial U} = \sum_{-m \leq j \leq m, j \neq 0} \frac{\partial J(\boldsymbol{v}_c, w_{t+j}, U)}{\partial U}. \tag{62}$$

Problem (f.ii)

Write down partial derivative: $\frac{\partial J_{\text{skip-gram}}(\boldsymbol{v}_c, w_{t-m}, \cdots, w_{t+m}, U)}{\partial \boldsymbol{v}_c}$.

Solution.

$$\frac{\partial J_{\text{skip-gram}}(\boldsymbol{v}_c, w_{t-m}, \cdots, w_{t+m}, U)}{\partial \boldsymbol{v}_c} = \sum_{-m \leq j \leq m, j \neq 0} \frac{\partial J(\boldsymbol{v}_c, w_{t+j}, U)}{\partial \boldsymbol{v}_c}. \tag{63}$$

Problem (f.iii)

Write down partial derivative: $\frac{\partial J_{\text{skip-gram}}(\boldsymbol{v}_c, w_{t-m}, \cdots, w_{t+m}, U)}{\partial \boldsymbol{v}_w}$ when $w \neq c$.

Solution.

$$\frac{\partial J_{\text{skip-gram}}(\boldsymbol{v}_c, w_{t-m}, \cdots, w_{t+m}, U)}{\partial \boldsymbol{v}_w} = 0. \tag{64}$$