Questions

Question a:

Show that the naive-softmax loss given in Equation (2) is the same as the cross-entropy loss
between y and $\hat{y}$;

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

Answer:

Since y is one_hot vector

The true empirical distribution y is a one-hot vector with a 1 for the true outside word o, and 0 everywhere else. The predicted distribution ˆ y is the probability distribution P(O|C = c) given by our model.

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

Question b:

Compute the partial derivative with respect to v_c

$$J_{naive-softmax}(v_c, o, U) = -\log P(O = o | C = c)$$

Answer:

$$
\begin{aligned}
\frac{\partial J}{\partial v_c} &= \frac{\partial}{\partial v_c}(-\log(\frac{exp(u_o^T v_c)}{\sum_{w\in Vocab} exp(u_w^T v_c)})) \\
&= \frac{\partial}{\partial v_c} - (u_o^T v_c - \log \sum_{w\in Vocab} exp(u_w^T v_c)) \\
&= \frac{\partial}{\partial v_c}(-u_o^T v_c) + \frac{\partial}{\partial v_c} \log \sum_{w\in Vocab} exp(u_w^T v_c) \\
&= -u_o + \frac{\sum_w exp(u_w^T v_c) u_w}{\sum_w exp(u_w^T v_c)} \\
&= -u_o + \sum_w \frac{exp(u_w^T v_c) u_w}{\sum_w exp(u_w^T v_c)} \\
&= -u_o + \sum_w P(O=w|C=c) u_w \\
&= -u_o + \sum_w \hat{y}_w u_w \\
&= U(\hat{y} - y)^T
\end{aligned}
$$

note: $y, \hat{y}$ are vectors (in) containing $\hat{y}_w (number)$ in row

u,v in column

Question c:

Compute the partial derivatives of $J_{naive-softmax}(v_c, o, U)$ with respect to each of the 'outside'
word vectors, u_w's. There will be two cases: when w = o, the true 'outside' word vector, and w != o, for all other words. Please write you answer in terms of y, ˆ y, and v_c

Answer:

$$
\begin{aligned}
\frac{\partial J}{\partial u_w} &= \frac{\partial}{\partial u_w}(-\log(\frac{exp(u_o^T v_c)}{\sum_{w\in Vocab} exp(u_w^T v_c)})) \\
&= \frac{\partial}{\partial u_w} - (u_o^T v_c - \log \sum_{w\in Vocab} exp(u_w^T v_c)) \\
&= \frac{\partial}{\partial u_w}(-u_o^T v_c) + \frac{\partial}{\partial u_w} \log \sum_{w\in Vocab} exp(u_w^T v_c)
\end{aligned}
$$

when $w \neq o$

$$
\begin{aligned}
\frac{\partial J}{\partial u_w} &= 0 + \frac{exp(u_w^T v_c) v_c}{\sum_w exp(u_w^T v_c)} \\
&= \frac{exp(u_w^T v_c) v_c}{\sum_w exp(u_w^T v_c)} \\
&= P(O=w|C=c) v_c \\
&= \hat{y}_w v_c
\end{aligned}
$$

when $w = o$

$$
\begin{aligned}
\frac{\partial J}{\partial u_w} &= -v_c + \frac{exp(u_w^T v_c) v_c}{\sum_w exp(u_w^T v_c)} \\
&= -v_c + \frac{exp(u_w^T v_c v_c}{\sum_w exp(u_w^T v_c)} \\
&= -v_c + P(O=w|C=c) v_c \\
&= -v_c + \hat{y}_w v_c
\end{aligned}
$$

In conclusion:

$$\frac{\partial J}{\partial u_w} = (\hat{y}_w - y_w)v_c$$

and

$$\frac{\partial J}{\partial U} = v_c(\hat{y} - y)$$

Question d:

The sigmoid function is given by Equation 4:
$\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x}$
Please compute the derivative of σ(x) with respect to x, where x is a scalar.
Hint: you may want to
write your answer in terms of σ(x).

$$\begin{aligned}\frac{\partial \sigma(x_i)}{\partial x_i} &= \frac{e^{x_i} \times (1+e^{x_i}) - e^{x_i} \times e^{x_i}}{(1+e^{x_i})^2} \\ &= \frac{e^{x_i}}{(1+e^{x_i})^2} \\ &= \frac{e^{x_i}}{1+e^{x_i}} \times \frac{1}{1+e^{x_i}} \\ &= \sigma(x_i)(1-\sigma(x_i))\end{aligned}$$

$$\begin{aligned}\frac{\partial \sigma(x_i)}{\partial x} &= \left[\frac{\partial \sigma(x_i)}{\partial x_i}\right]_{n\times n} \\ &= \begin{bmatrix} \sigma'(x_1) & 0 & \cdots & 0 \\ 0 & \sigma'(x_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma'(x_n) \end{bmatrix}\end{aligned}$$

Question 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 $w_1, w_2, \ldots, w_K$ and their outside vectors as$u_1, \ldots, u_K$. Note that $o / \in w_1, \ldots, w_K$. For a center word c and an outside word o, the negative sampling loss function is
given by:

$$J_{neg-sample}(v_c, o, U) = -log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} log(\sigma(-u_k^T v_c))$$

Answer:

For v_c

$$
\begin{aligned}
\frac{\partial J}{\partial v_c} &= \frac{\partial}{\partial v_c}(-log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} log(\sigma(-u_k^T v_c))) \\
&= \frac{\partial}{\partial v_c}(-log(\sigma(u_o^T v_c))) - \sum_{k=1}^{K} \frac{\partial}{\partial v_c}(log(\sigma(-u_k^T v_c))) \\
&= -\frac{\sigma(u_o^T v_c)(1-\sigma(u_o^T v_c))u_o}{\sigma(u_o^T v_c)} + \sum_{k=1}^{K} \frac{\sigma(-u_k^T v_c)(1-\sigma(-u_k^T v_c))u_k}{\sigma(-u_k^T v_c)} \\
&= -(1-\sigma(u_o^T v_c))u_o + \sum_{k=1}^{K}(1-\sigma(-u_k^T v_c))u_k \\
&= (\sigma(u_o^T v_c) - 1)u_o + \sum_{k=1}^{K} \sigma(u_k^T v_c)u_k
\end{aligned}
$$

For $u_o \quad o \notin w_1, w_2, \ldots w_K$

$$
\begin{aligned}
\frac{\partial J}{\partial u_o} &= \frac{\partial}{\partial u_o}(-log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} log(\sigma(-u_k^T v_c))) \\
&= \frac{\partial}{\partial u_o}(-log(\sigma(u_o^T v_c))) - \sum_{k=1}^{K} \frac{\partial}{\partial u_o}(log(\sigma(-u_k^T v_c))) \\
&= (\sigma(u_o^T v_c) - 1)v_c
\end{aligned}
$$

For $u_k \quad k \in w_1, w_2, \ldots w_K$

$$
\begin{aligned}
\frac{\partial J}{\partial u_k} &= \frac{\partial}{\partial u_k}(-log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} log(\sigma(-u_k^T v_c))) \\
&= \frac{\partial}{\partial u_k}(-log(\sigma(u_o^T v_c))) - \sum_{k=1}^{K} \frac{\partial}{\partial u_k}(log(\sigma(-u_k^T v_c))) \\
&= \sigma(u_k^T v_c)v_c
\end{aligned}
$$

Compare

When we use naive loss function, we have to compute a large matrix U multiplication, and derivative, which costs much time.

But in negative sampling, we only need to calculate K numbers and their derivation which saves us a lot time.

Question f:

$$
J_{skip-gram}(v_c, w_{t-m}, \ldots w_{t+m}, U) = \sum_{-m \le j \le m \& j \ne 0} J(v_c, w_{t+j}, U)
$$

Answer:

$$
\begin{aligned}
\frac{\partial J_s}{\partial U} &= \sum_{-m \le j \le m \& j \ne 0} \frac{\partial J(v_c, w_{t+j}, U)}{\partial U} \\
\frac{\partial J_s}{\partial v_c} &= \sum_{-m \le j \le m \& j \ne 0} \frac{\partial J(v_c, w_{t+j}, U)}{\partial v_c} \\
\frac{\partial J_s}{\partial v_w} &= 0(when\ w \ne c)
\end{aligned}
$$