CS224n： Assignment2 ， written part

qusetion website

Variables notation

Attention: All the variables' dimensions here are consistent with the code part in Assignment 2 for easy understanding.

U , matrix of shape (vocab_size,embedding_dim) ,all the 'outside' vectors .

V, matrix of shape (vocab_size,embedding_dim) ,all the 'center' vectors .

y, vector of shape (vocab_size,1), the true empirical distribution **y** is a one-hot vector with a 1 for the true outside word o, and 0 everywhere else .

$\hat{\boldsymbol{y}}$, vector of shape (vocab_size,1), the predicted distribution $\hat{\boldsymbol{y}}$ is the probability distribution $P(O|C=c)$ given by our model .

question a

Given outside word o and context word c.

The distribution of **y** is as follows:

$$y_w = \begin{cases} 1 & \text{w=o} \\ 0 & \text{w!=o} \end{cases}$$

$$-\sum_{w=1}^{V} y_w log(\hat{y_w}) = -y_o log(\hat{y_o}) = -log(\hat{y_o})$$

Here , V represents the vocab_size.

question b

$$
\begin{aligned}
&\frac{\partial J_{naive-softmax}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{v}_c} \\
&= -\frac{\partial log(P(O=o|C=c))}{\partial \boldsymbol{v}_c} \\
&= -\frac{\partial log(exp(\boldsymbol{u}_o^T \boldsymbol{v}_c))}{\partial \boldsymbol{v}_c} + \frac{\partial log(\sum_{w=1}^{V} exp(\boldsymbol{u}_w^T \boldsymbol{v}_c))}{\partial \boldsymbol{v}_c} \\
&= -\boldsymbol{u}_o + \sum_{w=1}^{V} \frac{exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)}{\sum_{w=1}^{V} exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)} \boldsymbol{u}_w \\
&= -\boldsymbol{u}_o + \sum_{w=1}^{V} P(O=w|C=c) \boldsymbol{u}_w \\
&= \boldsymbol{U}^T(\hat{\boldsymbol{y}} - \boldsymbol{y})
\end{aligned}
$$

question c

$$
\begin{aligned}
&\frac{\partial J_{naive-softmax}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_w} \\
&= -\frac{\partial log(exp(\boldsymbol{u}_o^T \boldsymbol{v}_c))}{\partial \boldsymbol{u}_w} + \frac{\partial log(\sum_{w=1}^{V} exp(\boldsymbol{u}_w^T \boldsymbol{v}_c))}{\partial \boldsymbol{u}_w}
\end{aligned}
$$

when w = o,

$$
\begin{aligned}
&\frac{\partial J_{naive-softmax}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_w} \\
&= -\boldsymbol{v}_c + \frac{1}{\sum_{w=1}^{V} exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)} \frac{\partial \sum_{w=1}^{V} exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)}{\partial \boldsymbol{u}_o} \\
&= -\boldsymbol{v}_c + \frac{1}{\sum_{w=1}^{V} exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)} \frac{\partial exp(\boldsymbol{u}_o^T \boldsymbol{v}_c)}{\partial \boldsymbol{u}_o} \\
&= -\boldsymbol{v}_c + \frac{exp(\boldsymbol{u}_o^T \boldsymbol{v}_c)}{\sum_{w=1}^{V} exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)} \boldsymbol{v}_c \\
&= (P(O=o|C=c) - 1))\boldsymbol{v}_c
\end{aligned}
$$

when w != o,

$$
\begin{aligned}
&\frac{\partial J_{naive-softmax}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_w} \\
&= \frac{exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)}{\sum_{w=1}^{V} exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)} \boldsymbol{v}_c \\
&= P(O = w | C = c) \boldsymbol{v}_c
\end{aligned}
$$

In summary,

$$
\begin{aligned}
&\frac{\partial J_{naive-softmax}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{U}} \\
&= (\hat{\boldsymbol{y}} - \boldsymbol{y})^T \boldsymbol{v}_c
\end{aligned}
$$

question d

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

question e

$$
\begin{aligned}
&\frac{\partial J_{neg-sample}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{v}_c} \\
&= \frac{\partial(-log(\sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c)) - \sum_{k=1}^{K} log(\sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c)))}{\partial \boldsymbol{v}_c} \\
&= -\frac{\sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c)(1 - \sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c))}{\sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c)} \frac{\partial \boldsymbol{u}_o^T \boldsymbol{v}_c}{\partial \boldsymbol{v}_c} - \sum_{k=1}^{K} \frac{\partial log(\sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c))}{\partial \boldsymbol{v}_c} \\
&= -(1 - \sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c)) \boldsymbol{u}_o + \sum_{k=1}^{K} (1 - \sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c)) \boldsymbol{u}_k
\end{aligned}
$$

$$\frac{\partial J_{neg-sample}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_o}$$

$$= \frac{\partial(-log(\sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c))}{\partial \boldsymbol{u}_o} = -(1 - \sigma(\boldsymbol{u}_o^T \boldsymbol{v}_c))\boldsymbol{v}_c$$

$$\frac{\partial J_{neg-sample}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_k}$$

$$= \frac{\partial(-log(\sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c))}{\partial \boldsymbol{u}_k} = (1 - \sigma(-\boldsymbol{u}_k^T \boldsymbol{v}_c))\boldsymbol{v}_c$$

qustion f

i)

$$\frac{\partial J_{skip-gram}(\boldsymbol{v}_c, w_{t-m}, \ldots, w_{t+m}, \boldsymbol{U})}{\partial \boldsymbol{U}}$$

$$= \sum_{-m<=j<=m, j!=0} \frac{\partial J(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U})}{\partial \boldsymbol{U}}$$

ii)

when w=c,

$$\frac{\partial J_{skip-gram}(\boldsymbol{v}_c, w_{t-m}, \ldots, w_{t+m}, \boldsymbol{U})}{\partial \boldsymbol{v}_c}$$

$$= \sum_{-m<=j<=m, j!=0} \frac{\partial J(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U})}{\partial \boldsymbol{v}_c}$$

iii)

when w!=c,

$$\frac{\partial J_{skip-gram}(\boldsymbol{v}_c, w_{t-m}, \ldots, w_{t+m}, \boldsymbol{U})}{\partial \boldsymbol{v}_w}$$

$$= \boldsymbol{0}$$