CS224N-2019 Assignment 2 Written Part

Understanding word2vec

Variable Notation

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

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

y , matrix of shape(vocab_size, 1), means one-hot vector with 1 for outside word and 0 for anything else.

$\hat{y}$, matrix of shape(vocab_size, 1), means distributed prediction vector for all words.

$$J_{naive-softmax}(v_c, o, U) = -logP(o|c) = -log\frac{exp(u_o^{\mathrm{T}}v_c)}{\sum_w exp(u_w^{\mathrm{T}}v_c)} \tag{1}$$

Question(a) Ans: Given one outside word, we know the distribution of y as following:

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

It's obvious that $-\sum_w y_w log(\hat{y_w}) = -y_o log(\hat{y_o}) = -log(\hat{y_o})$

Question(b) Ans: Firstly we simplify $J_{naive-softmax}(v_c, o, U)$ as following:

$$J_{naive-softmax}(v_c, o, U) = -u_o^{\mathrm{T}}v_c + log\sum_w exp(u_w^{\mathrm{T}}v_c)$$

Then compute the partial derivative of $J_{naive-softmax}(v_c, o, U)$:

$$\begin{aligned}
\frac{\partial J_{naive-softmax}(v_c, o, U)}{\partial v_c} &= -u_o + \frac{\partial log\sum_w exp(u_w^{\mathrm{T}}v_c)}{v_c} \\
&= -u_o + \frac{1}{\sum_w exp(u_w^{\mathrm{T}}v_c)} \times \frac{\sum_x \partial exp(u_x^{\mathrm{T}}v_c)}{\partial v_c} \\
&= -u_o + \frac{1}{\sum_w exp(u_w^{\mathrm{T}}v_c)} \times \sum_x exp(u_x^{\mathrm{T}}v_c)u_x \\
&= -u_o + \sum_x \frac{exp(u_x^{\mathrm{T}}v_c)}{\sum_w exp(u_w^{\mathrm{T}}v_c)}u_x \\
&= -u_o + \sum_x P(x|c)u_x
\end{aligned}$$

Also according to notation, we have $-u_o^{\mathrm{T}} = -Uy,\ \sum_x P(x|c)u_x^{\mathrm{T}} = U\hat{y}$ and get the partial derivative respect of v_c in terms of $y, \hat{y}$ and U as $\frac{\partial J_{naive-softmax}(v_c, o, U)}{\partial v_c} = U(\hat{y} - y)$.

Question(c) Ans: There are two cases for the condition, $w = o$ and $w! = o$.

First consider the case $w = o$ which is very similar to that of Question(b):

$$\begin{aligned}
\frac{\partial J_{naive-softmax}(v_c, o, U)}{\partial u_w} &= -v_c + \frac{1}{\sum_i exp(u_i^{\mathrm{T}}v_c)} \times \frac{\sum_x \partial exp(u_x^{\mathrm{T}}v_c)}{\partial u_w} \\
&= -v_c + \frac{exp(u_w^{\mathrm{T}}v_c)}{\sum_i exp(u_i^{\mathrm{T}}v_c)}v_c \\
&= -v_c + P(w|c)v_c
\end{aligned}$$

Now it's turn for the case $w! = o$:

$$
\begin{aligned}
\frac{\partial J_{naive-softmax}(v_c, o, U)}{\partial u_w} &= \frac{1}{\sum_i exp(u_i^{\mathrm{T}} v_c)} \times \frac{\sum_x \partial exp(u_x^{\mathrm{T}} v_c)}{\partial u_w} \\
&= \frac{exp(u_w^{\mathrm{T}} v_c)}{\sum_i exp(u_i^{\mathrm{T}} v_c)} v_c \\
&= P(w|c) v_c
\end{aligned}
$$

The partial derivatives of $J_{naive-softmax}(v_c, o, U)$ with respect to u_w's makes a matrix of shape(embedding_dim, vocab_size) $[P(w_1|c)v_c, P(w_2|c)v_c, \ldots, (P(o|c)-1)v_c, \ldots, P(w_n|c)v_c]$, which actually equals to $v_c(\hat{y}-y)^{\mathrm{T}}$

Question(d) Ans: Despite x as a vector, the calculation of derivative of $\sigma(x)$ is the same as a real number.

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

Question(e) Ans: Note that use of the conclusion in part(d) makes great convenience.

First repeat part(b).

$$
\begin{aligned}
\frac{\partial J_{neg-sample}(v_c, o, U)}{\partial v_c} &= -\frac{1}{\sigma(u_o^{\mathrm{T}} v_c)} \times \frac{\partial \sigma(u_o^{\mathrm{T}} v_c)}{\partial v_c} - \sum_{k=1}^{K} \frac{1}{\sigma(-u_k^{\mathrm{T}} v_c)} \times \frac{\partial \sigma(-u_k^{\mathrm{T}} v_c)}{\partial v_c} \\
&= (\sigma(u_o^{\mathrm{T}} v_c) - 1) u_o + \sum_{k=1}^{K} (1 - \sigma(-u_k^{\mathrm{T}} v_c)) u_k
\end{aligned}
$$

Next repeat part(c).

case $w = o$

$$
\begin{aligned}
\frac{\partial J_{neg-sample}(v_c, o, U)}{\partial u_w} &= -\frac{1}{\sigma(u_o^{\mathrm{T}} v_c)} \times \frac{\partial \sigma(u_o^{\mathrm{T}} v_c)}{\partial u_w} + 0 \\
&= (\sigma(u_o^{\mathrm{T}} v_c) - 1) v_c
\end{aligned}
$$

case $w \in [1, K]$

$$
\begin{aligned}
\frac{\partial J_{neg-sample}(v_c, o, U)}{\partial u_w} &= 0 - \sum_{k=1}^{K} \frac{1}{\sigma(-u_k^{\mathrm{T}} v_c)} \times \frac{\partial \sigma(-u_k^{\mathrm{T}} v_c)}{\partial u_w} \\
&= (1 - \sigma(-u_w^{\mathrm{T}} v_c)) v_c
\end{aligned}
$$

Question(f) Ans:

(i) According to attribute of derivatives , we have:

$$
\frac{\partial J_{skig-gram}(v_c, W_{t-m}, \ldots, W_{t+m}, U)}{\partial U} = \sum_{-m \le j \le m} \frac{J_{skig-gram}(v_c, W_{t+j}, U)}{\partial U}
$$

(ii) Similarly, we also have:

$$
\frac{\partial J_{skig-gram}(v_c, W_{t-m}, \ldots, W_{t+m}, U)}{\partial v_c} = \sum_{-m \le j \le m} \frac{J_{skig-gram}(v_c, W_{t+j}, U)}{\partial v_c}
$$

(iii) Obviously there is no such variable $v_w (w \neq c)$ in $J_{skig-gram}(v_c, W_{t-m}, \ldots, W_{t+m}, U)$ and the answer is zero.