CS224n Assignment2

September 26, 2022

1.Written: Understanding word2vec

(a) **y** is the true empirical distribution, one-hot vector with a 1 for true outside word o, and 0 everywhere else. Hence the left side of the function can be rewritten as:

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

(b) Explanation by line:

- L1-L2 take $exp(u_o^T v_c)$ as $f(v_c)$, $exp(u_w^T v_c)$ as $g(v_c)$ and use the quotient rule.
- L2-L3 Suppose we have a vocabulary of size k, and an embedding size of 200. Then U is of size (200,k). $u_o = Uy$ (y is a one-hot vector of size (k,1)). For the summation part,it should be clear if you just open it up.
- L3-L4 $\hat{y}_o = \frac{exp(u_o^T v_c)}{\sum_w exp(u_w^T v_c)}$, note $\hat{y}_o$ is a scalar.
- L4-L5 $\hat{y}$ is of size (k,1), convert the summation into matrix multiplication.

$$\begin{aligned}
\frac{\partial J}{\partial v_c} &= \frac{\partial}{\partial v_c} - log \frac{exp(u_o^T v_c)}{\sum_w exp(u_w^T v_c)} \\
&= -u_o + \frac{\sum_w u_w exp(u_w^T v_c)}{\sum_w exp(u_w^T v_c)} \\
&= -Uy + \sum_w \frac{u_w exp(u_w^T v_c)}{\sum_m exp(u_m^T v_c)} \\
&= -Uy + \sum_w \hat{y}_w u_w \\
&= -Uy + U\hat{y} = U(\hat{y} - y)
\end{aligned}$$

(c) Consider two cases:
when $w \neq o$:

- L1-L2 $log\frac{a}{b} = loga - logb$, and the first term does not contain u_w.
- L2-L3 Take the derivative.
- L3-L4 $\hat{y}_w = \frac{exp(u_w^T v_c)}{\sum_w exp(u_w^T v_c)}$

$$\begin{aligned}
\frac{\partial J}{\partial u_w} &= \frac{\partial}{\partial u_w} - log\frac{exp(u_o^T v_c)}{\sum_w exp(u_w^T v_c)} \\
&= \frac{\partial}{\partial u_w} log\sum_w exp(u_w^T v_c) \\
&= \frac{v_c exp(u_w^T v_c)}{\sum_m exp(u_m^T v_c)} \\
&= v_c \hat{y}_w
\end{aligned}$$

when $w = o$:

$$\begin{aligned}
\frac{\partial J}{\partial u_w} &= \frac{\partial}{\partial u_w} - log\frac{exp(u_o^T v_c)}{\sum_w exp(u_w^T v_c)} \\
&= \frac{\partial}{\partial u_w} log\sum_w exp(u_w^T v_c) - logexp(u_o^T v_c) \\
&= \frac{v_c exp(u_w^T v_c)}{\sum_m exp(u_m^T v_c)} - v_c \\
&= v_c(\hat{y}_o - 1)
\end{aligned}$$

(d) The following term is calculated as above.

$$\frac{\partial J}{\partial U} = [\frac{\partial J}{\partial u_1}, \frac{\partial J}{\partial u_2}, ..., \frac{\partial J}{\partial u_{|Vocab|}}]$$

(e) When $x > 0$, $f'(x) = 1$, when $x < 0$, $f'(x) = 0$

(f)

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

(g)
(i) Here J refers to the negative sampling loss

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

Note that $o \notin w_1, ..., w_k$, the second term does not contain u_o

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

The first term does not contain u_{w_s}

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

(ii) $Uv_c - 1$

(iii) The calculation only depends on the K negative samples instead of the whole vocabulary.

(h)
We now suppose that the K sampled words are not distinct. The basic intuition here is that we multiply the gradient by the number of a sample word's appearance.

$$\begin{aligned}\frac{\partial J}{\partial u_{w_s}} &= \frac{\partial}{\partial u_{w_s}} - \sum_{s=1}^{K} log(\sigma(-u_{w_s}^T v_c)) \\ &= \sum_{w=w_s} v_c(1 - \sigma(-u_{w_s}^T v_c))\end{aligned}$$

(i) Here J refers to the skip-gram loss

$$\frac{\partial J}{\partial U} = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \frac{\partial J_{neg-sample}(v_c, w_{t+j}, U)}{\partial U}$$

$$\frac{\partial J}{\partial v_c} = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \frac{\partial J_{neg-sample}(v_c, w_{t+j}, U)}{\partial v_c}$$

when $w \neq c$,

$$\frac{\partial J}{\partial v_w} = 0$$