CS 224n: Assignment 2 Written Part

April 20, 2022

Question 1: Understanding Word2Vec (26 points)

a) Show that the naive-softmax loss given in Equation (2) is the same as the cross-entropy loss between y and yˆ.

Solution: Because the true empirical distribution y is a one-hot vector with a 1 for the true outside word o, and 0 everywhere else. Therefore, the LHS equals to

$$-(y_0 log(\hat{y_0}) + y_1 log(\hat{y_1}) + ... + y_o log(\hat{y_o}) + ...) = -y_o log(\hat{y_o}) = -log(\hat{y_o})$$

b) Compute the partial derivative of $J_{naive-softmax}(v_c, o, U)$ with respect to v_c.

Solution:

$$\begin{aligned}
\frac{\partial}{\partial v_c} J_{naive-softmax}(v_c, o, U) &= \frac{\partial}{\partial v_c} - \log P(O = o|C = c) \\
&= \frac{\partial}{\partial v_c} - \log \frac{\exp(u_o^T v_c)}{\sum_w \exp(u_w^T v_c)} \\
&= \frac{\partial}{\partial v_c} - [\log \exp(u_o^T v_c) - \log \sum_w \exp(u_w^T v_c)] \\
&= \frac{\partial}{\partial v_c} [-u_o^T v_c + \log \sum_w \exp(u_w^T v_c)] \\
&= -u_o + \frac{\sum_x \exp(u_x^T v_c) \cdot u_x}{\sum_w \exp(u_w^T v_c)} \\
&= -u_o + \sum_x \frac{\exp(u_x^T v_c) \cdot u_x}{\sum_w \exp(u_w^T v_c)} \\
&= -u_o + \sum_x P(u_x|v_c) \cdot u_x \\
&= -u_o + \sum_x \hat{y_x} \cdot u_x \\
&= U \cdot (\hat{y} - y)
\end{aligned}$$

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.

Solution:

①When w = o:

$$\begin{aligned}
\frac{\partial}{\partial u_w} J_{naive-softmax}(v_c, o, U) &= \frac{\partial}{\partial u_o} - \log \frac{\exp(u_o^T v_c)}{\sum_w \exp(u_w^T v_c)} \\
&= \frac{\partial}{\partial u_o} - [\log \exp(u_o^T v_c) - \log \sum_w \exp(u_w^T v_c)] \\
&= \frac{\partial}{\partial u_o} (-u_o^T v_c + \log \sum_w \exp(u_w^T v_c)) \\
&= -v_c + \frac{\exp(u_o^T v_c) \cdot v_c}{\sum_w \exp(u_w^T v_c)} \\
&= -v_c + \hat{y_o} \cdot v_c
\end{aligned}$$

②When $w \neq o$:

$$\begin{aligned}
\frac{\partial}{\partial u_w} J_{naive-softmax}(v_c, o, U) &= \frac{\partial}{\partial u_w} - \log \frac{\exp(u_o^T v_c)}{\sum_w \exp(u_w^T v_c)} \\
&= \frac{\partial}{\partial u_w} - (u_o^T v_c + \log \sum_w \exp(u_w^T v_c)) \\
&= 0 + \frac{\exp(u_w^T v_c) \cdot v_c}{\sum_x \exp(u_x^T v_c)} \\
&= \hat{y_{w \neq o}} \cdot v_c
\end{aligned}$$

In summary, $\frac{\partial}{\partial u_w} J_{naive-softmax}(v_c, o, U) = (\hat{y} - y)^T v_c$.

d) Compute the partial derivative of $J_{naive-softmax}(v_c, o, U)$ with respect to U.

Solution:

$$\begin{aligned}
\frac{\partial}{\partial U} J_{naive-softmax}(v_c, o, U) &= \frac{\partial}{\partial U} - \log \frac{\exp(u_o^T v_c)}{\sum_w \exp(u_w^T v_c)} \\
&= \frac{\partial}{\partial U} (-u_o^T v_c + \log \sum_w \exp(u_w^T v_c)) \\
&= [0, 0, ..., v_c, 0, 0, ..., 0] + \frac{[\exp(u_1^T v_c) \cdot v_c, \exp(u_2^T v_c) \cdot v_c, ...]}{\sum_w \exp(u_w^T v_c)} \\
&= y + \hat{y} \cdot v_c
\end{aligned}$$

e) Please compute the derivative of $\sigma(x)$ with respect to x, where x is a scalar.

Solution:

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

f) Please repeat parts (b) and (c), computing the partial derivatives of $J_{neg-sample}$ with respect to v_c, with respect to u_o, and with respect to a negative sample u_k. After you've done this, describe with one sentence why this loss function is much more efficient to compute than the naive-softmax loss.

Solution:

①With respect to v_c

$$\begin{aligned}
\frac{\partial}{\partial v_c} J_{neg-sample}(v_c, o, U) &= \frac{\partial}{\partial u_o}[-\log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} l(\sigma(-u_k^T v_c))] \\
&= -\frac{(\sigma(u_o^T v_c) - \sigma^2(u_o^T v_c))u_o}{\sigma(u_o^T v_c)} - \sum_{k=1}^{K} \frac{(\sigma(-u_k^T v_c) - \sigma^2(-u_k^T v_c)) \cdot (-u_k)}{\sigma(-u_k^T v_c)} \\
&= -u_o + u_o \cdot \sigma(u_o^T v_c) - \sum_{k=1}^{K} (1 - \sigma(-u_k^T v_c)) \cdot (-u_k) \\
&= u_o(\sigma(u_o^T v_c) - 1) + \sum_{k=1}^{K} u_k(1 - \sigma(-u_k^T v_c))
\end{aligned}$$

②With respect to u_o

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

③With respect to u_k

$$\begin{aligned}
\frac{\partial}{\partial u_k} J_{neg-sample}(v_c, o, U) &= \frac{\partial}{\partial u_k}[-\log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} \log(\sigma(-u_k^T v_c))] \\
&= \frac{(\sigma(-u_k^T v_c) - \sigma^2(-u_k^T v_c)) \cdot (v_c)}{\sigma(-u_k^T v_c)} \\
&= v_c \cdot (1 - \sigma(-u_k^T v_c))
\end{aligned}$$

The naive-softmax loss requires to run through the whole word vectors, which is O(—V—). While negative sample loss only requires us to look at K words, which is O(—K—). Thus, the negative sample loss function is more efficient.

g) Now we will repeat the previous exercise, but without the assumption that the K sampled words are distinct. Compute the partial derivative of $J_{neg-sample}$ with respect to a negative sample u_k.

Solution:

$$\begin{aligned}
&\frac{\partial}{\partial u_k} J_{neg-sample}(v_c, o, U) \\
&= \frac{\partial}{\partial u_k}[-\log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} \log(\sigma(-u_k^T v_c))] \\
&= \frac{\partial}{\partial u_k}(-\log(\sigma(u_o^T v_c)) - \sum_{x \in k \ \& \ u_x = u_o} \log(\sigma(-u_o^T v_c)) - \sum_{y \in k \ \& \ u_y \neq u_o} \log(\sigma(-u_y^T v_c))
\end{aligned}$$

if $u_k = u_o$:

$$\begin{aligned}\frac{\partial}{\partial u_k} J_{neg-sample}(v_c, o, U) &= -\frac{(\sigma(u_o^T v_c) - \sigma^2(u_o^T v_c)) \cdot v_c}{\sigma(u_o^T v_c)} + \sum_{x \in K \ \& \ u_x = u_o} \frac{(\sigma(-u_o^T v_c) - \sigma^2(-u_o^T v_c)) \cdot v_c}{\sigma(-u_o^T v_c)} \\ &= (\sigma(u_o^T v_c) - 1) \cdot v_c + \sum_{x \in K \ \& \ u_x = u_o} (1 - \sigma(-u_o^T v_c)) \cdot v_c) \\ &= (\sigma(u_k^T v_c) - 1) \cdot v_c + \sum_{u_k = u_o}^{K} (1 - \sigma(-u_k^T v_c)) \cdot v_c\end{aligned}$$

if $u_k \neq u_o$:

$$\begin{aligned}\frac{\partial}{\partial u_k} J_{neg-sample}(v_c, o, U) &= -\sum_{u_k \neq u_o}^{K} \frac{(\sigma(-u_k^T v_c) - \sigma^2(-u_k^T v_c)) \cdot (-v_c)}{\sigma(-u_k^T v_c)} \\ &= \sum_{u_k \neq u_o}^{K} (1 - \sigma(-u_k^T v_c))\end{aligned}$$

h) Write down three partial derivatives of $J_{skip-gram}(v_c, w_{t-m}, ..., w_{t+m}, U)$

Solution:

$$(i) \partial J_{skip-gram}(v_c, w_{t-m}, ..., w_{t+m}, U)/\partial U = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \partial J(v_c, w_{t+j}, U)/\partial U$$

$$(ii) \partial J_{skip-gram}(v_c, w_{t-m}, ..., w_{t+m}, U)/\partial v_c = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \partial J(v_c, w_{t+j}, U)/\partial v_c$$

$$(iii) \partial J_{skip-gram}(v_c, w_{t-m}, ..., w_{t+m}, U)/\partial v_c = 0$$