CS224n: Aassignment 2

Amirali Abdullah

Winter 2020

Problem 1

Consider a single pair of words c and o co-occurring, where c is the "center" word of context and o is the "outside" word in the window. We lay out the following definitions:
Define the naive softmax loss as:

$$J_{nsoftmax}(v_c, o, U) = -\log P(O = o|C = c). \tag{1}$$

Define y and $\hat{y}$ as the true empirical distribution and predicted distribution for a given c respectively. Namely the k^{th} entry of $\hat{y}$ indicates the probability that o is an outside word when c occurs. Whereas the k^{th} entry of y is a 1-hot vector with a 1 for the true outside word o, and 0 everywhere else.

Part (a)

Assume we are given a single pair of words c and o. Show that the naive-softmax loss is the same as the cross entropy loss between y and $\hat{y}$. Namely show that:

$$-\sum_{w \in Vocab} y_w \log(\hat{y}_w) = -\log(\hat{y}_o) \tag{2}$$

Ans

$$\begin{aligned}
-\sum_{w \in Vocab} y_w \log(\hat{y}_w) &= -y_o \log(\hat{y}_0) - \sum_{w \neq o} y_w \log(\hat{y}_w) \\
&= -y_o \log(\hat{y}_0) - \sum_{w \neq o} 0. \log(\hat{y}_w) \\
&= -y_0 \log(\hat{y}_0)
\end{aligned}$$

Part (b)

Find derivative of $J_{nsoftmax}(v_c, o, U)$ w.r.t v_c. Please write your answer in terms of y, y' and U.

Ans

From part (a), we have that:

$$J_{nsoftmax}(v_c, o, U) = CE(y, \hat{y})$$

Recall also that $\hat{y} = softmax(\theta)$, where we define θ as a vector with the elements $u_w^T v_c$ (or specifically, $\theta = U^T v_c$.) We now observe the two straightforward formulae:

1. Applying the formula for Jacobian of cross entropy, we immediately have that $\frac{\partial J}{\partial \theta} = (\hat{y} - y)^T$.
2. Next we see that $\frac{\partial \theta}{\partial v_c} = U$.

Finally, the chain rule yields us:

$$\frac{\partial J}{\partial v_c} = \frac{\partial J}{\partial \theta} \frac{\partial \theta}{\partial v_c} = (\hat{y} - y)^T U \tag{3}$$

Alternate Ans

Proceeding directly:

$$
\begin{aligned}
J_{nsoftmax}(v_c, o, U) &= -\log\left(\frac{exp(u_0^T v_c)}{\sum_{w\in vocab} exp(u_w^T v_c)}\right) \\
&= -u_0^T v_c + \log\left(\sum_{w\in vocab} exp(u_w^T v_c)\right)
\end{aligned}
$$

So the derivative becomes:

$$
\begin{aligned}
\frac{\partial}{\partial v_c} J_{nsoftmax}(v_c, o, U) =& - u_o + \frac{1}{\sum_{x\in vocab} exp(u_x^T v_c)}\left(\sum_{w\in vocab} \frac{\partial}{\partial v_c} exp(u_w^T v_c)\right) \\
=& - u_0 + \sum_{w\in vocab} u_w \frac{exp(u_w^T v_c)}{\sum_{x\in vocab} exp(u_x^T v_c)} \\
=& - u_0 + \sum_{w\in vocab} u_w Pr(O = w|C = c) \\
=& (\hat{y} - y)^T U
\end{aligned}
$$

Part (c)

Find derivative of $J_{nsoftmax}(v_c, o, U)$ w.r.t u_w. There will be two cases, one where u_w corresponds to the "true" outside word o and one where it does not. Please write your answer in terms of y, y' and U.

Ans

Proceeding directly:

$$
\begin{aligned}
J_{nsoftmax}(v_c, o, U) &= -\log\left(\frac{exp(u_0^T v_c)}{\sum_{w\in vocab} exp(u_w^T v_c)}\right) \\
&= -u_0^T v_c + \log\left(\sum_{w\in vocab} exp(u_w^T v_c)\right)
\end{aligned}
$$

So the derivative becomes.
Case 1 (u_w for $w = o$).

$$
\begin{aligned}
\frac{\partial}{\partial u_o} J_{nsoftmax}(v_c, o, U) =& - v_c + \frac{1}{\sum_{x\in vocab} exp(u_x^T v_c)}\left(\sum_{w\in vocab} \frac{\partial}{\partial u_o} exp(u_w^T v_c)\right) \\
=& - v_c + v_c \frac{exp(u_o^T v_c)}{\sum_{x\in vocab} exp(u_x^T v_c)} \\
=& - v_c + v_c Pr(O = o|C = c) \\
=& v_c Pr(O = o|C = c) - v_c \\
=& (\hat{y} - y) v_c
\end{aligned}
$$

Case 2 (u_s for $s \neq o$).

$$
\frac{\partial}{\partial u_s} J_{nsoftmax}(v_c, o, U) = \frac{1}{\sum_{x\in vocab} exp(u_x^T v_c)}\left(\sum_{w\in vocab} \frac{\partial}{\partial u_s} exp(u_w^T v_c)\right)
$$

$$
\begin{aligned}
=&v_c\frac{exp(u_s^T v_c)}{\sum_{x\in vocab} exp(u_x^T v_c)}\\
=&v_c Pr(O=s|C=c)\\
=&y v_c = (\hat{y}-y)v_c
\end{aligned}
$$

Part (d)

Find the derivative of the sigmoid function, namely:

$$\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}$$

Ans

Recall first this formula for derivatives: $\frac{d}{dx}\frac{1}{f(x)} = \frac{-f'(x)}{f(x)^2}$.
Applying this here to the sigmoid function, we get:

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

Part (e)

Consider negative sampling gradient as follows:

$$J_{neg-sample}(v_c, o, U) = -\log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K}\log(\sigma(-u_k^T v_c)) \tag{4}$$

Ans

Then we compute

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

And for u_o:

$$
\begin{aligned}
\frac{\partial}{\partial u_0} J_{negsample}(v_c, o, U) &= -\frac{1}{\sigma(u_o^T v_c)}(\sigma(u_o^T v_c))(1-\sigma(u_o^T v_c))v_c\\
&= -(1-\sigma(u_o^T v_c))v_c
\end{aligned}
$$

And for another u_k, where $k \neq o$, we have:

$$\begin{aligned}\frac{\partial}{\partial u_k} J_{negsample}(v_c, o, U) &= \frac{1}{\sigma(u_k^T v_c)}(\sigma(-u_k^T v_c))(1 - \sigma(u_k^T v_c))v_c \\ &= (1 - \sigma(-u_k^T v_c))v_c\end{aligned}$$

Part (f)

Consider skipgram context window version of word2vec

$$J_{skipgram}(v_c, w_{t-m}, ..., w_{t+m}, U) = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} J(v_c, w_{t+j}, U) \tag{5}$$

Use skipgram derivatives from (e) to find gradients w.r.t U, v_c and v_w for when $w \neq c$.

Ans

We derive the following results:

$$\frac{\partial}{\partial U} J_{skipgram}(v_c, w_{t-m}, ..., w_{t+m}, U) = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \frac{\partial}{\partial U} J(v_c, w_{t+j}, U) \tag{6}$$

$$\frac{\partial}{\partial v_c} J_{skipgram}(v_c, w_{t-m}, ..., w_{t+m}, U) = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \frac{\partial}{\partial v_c} J(v_c, w_{t+j}, U) \tag{7}$$

$$\frac{\partial}{\partial v_{w, w \neq c}} J_{skipgram}(v_c, w_{t-m}, ..., w_{t+m}, U) = 0 \tag{8}$$