CS 224n Assignment #2: word2vec

(a) (3 points)

Show that the naive-softmax loss given in Equation (2) is the same as the cross-entropy loss between y and $\hat{y}$; i.e., show that

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

Your answer should be one line.

Note:

$$J_{naive-softmax}(v_c, o, U) = -logP(O = o|C = c)$$
$$P(O = o|C = c) = \frac{\exp(u_o^T v_c)}{\sum_{w \in \text{Vocab}} \exp(u_w^T v_c)}$$

Answer:

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

y_w is a one hot vector where the index of the correct word o is 1 and all others are 0.

Therefore

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

(b) (5 points)

Compute the partial derivative of $J_{naive-softmax}(v_c, o, U)$ with respect to v_c. Please write your answer in terms of $y, \hat{y}$ and U.

Answer:

Note that

- $J = CE(y, \hat{y}) = -\sum_w y_w log(\hat{y}_w)$
- $\hat{y} = \text{softmax}(\theta) = \frac{\exp(\theta_o)}{\sum_{w \in \text{Vocab}} \exp(\theta_w)} = \frac{\exp(u_o^T v_c)}{\sum_{w \in \text{Vocab}} \exp(u_w^T v_c)}$

For $\frac{\partial J}{\partial \theta}$:

$$J = -\sum_w y_w log(\hat{y}_w)$$
$$\frac{\partial J}{\partial \theta_o} = -\sum_w y_w \frac{\partial log(\hat{y}_w)}{\partial \theta_o}$$
$$= -\sum_w y_w \frac{\partial log(\hat{y}_w)}{\partial \hat{y}_w} \frac{\partial \hat{y}_w}{\theta_o}$$
$$= -\sum_w y_w \frac{1}{\hat{y}_w} \frac{\partial \hat{y}_w}{\theta_o}$$

where $\frac{\partial \hat{y}_w}{\theta_o}$ is the derivative of a softmax regression function.

It can be proven that the derivative of a softmax regression $\hat{y} = \text{softmax}(\theta)$ satisfies

$$\frac{\partial \hat{y}_w}{\partial \theta_o} = \begin{cases} \hat{y}_w(1 - \hat{y}_o) & \text{if } w = o \\ -\hat{y}_o \cdot \hat{y}_w & \text{if otherwise} \end{cases}$$

Therefore, we can rewrite $\frac{\partial J}{\partial \theta}$ as

$$\begin{aligned} \frac{\partial J}{\partial \theta} &= -\sum_w y_w \frac{1}{\hat{y}_w} \frac{\partial \hat{y}_w}{\partial \theta_o} \\ &= -y_o(1 - \hat{y}_o) - \sum_{w \neq o} y_w \frac{1}{\hat{y}_w}(-\hat{y}_w \hat{y}_o) \\ &= -y_o(1 - \hat{y}_o) + \sum_{w \neq o} y_w \hat{y}_o \\ &= -y_o + y_o \hat{y}_o + \sum_{w \neq o} y_w \hat{y}_o \\ &= \hat{y}_o(y_o + \sum_{w \neq o} y_w) - y_o \end{aligned}$$

y is a one hot encoded vector for the labels, so $\sum_w y_w = 1$ and $y_o + \sum_{w \neq o} y_w = 1$. So we have

$$\begin{aligned} \frac{\partial J}{\partial \theta} &= \hat{y}_o - y_o \\ &\text{or} \\ \frac{\partial J}{\partial \theta} &= (\hat{y}_o - y_o)^T \quad \text{if y is a column vector} \end{aligned}$$

For $\frac{\partial \theta}{\partial v_c}$:

$$\frac{\partial \theta}{\partial v_c} = U$$

Therefore

$$\begin{aligned} \frac{\partial J}{\partial v_c} &= \frac{\partial J}{\partial \theta} \frac{\partial \theta}{\partial v_c} \\ &= (\hat{y} - y)^T U \end{aligned}$$

(c) (5 points)

Compute the partial derivatives of $J_{naive-softmax}(v_c, o, U)$ with respect to each of the 'outside' word vectors, u_w's. There will be two cases: when $w = o$, the true 'outside' word vector, and $w \neq o$, for all other words. Please write you answer in terms of $y, \hat{y}, \text{ and } v_c$.

Answer:

Note that

- $J = CE(y, \hat{y}) = -\sum_w y_w log(\hat{y}_w)$
- $\hat{y} = \text{softmax}(\theta) = \frac{\exp(\theta_o)}{\sum_{w \in \text{Vocab}} \exp(\theta_w)} = \frac{\exp(u_o^T v_c)}{\sum_{w \in \text{Vocab}} \exp(u_w^T v_c)}$

We have model prediction $\hat{y} = \text{softmax}(\theta)$ and cross-entropy loss $J = CE(y, \hat{y})$. We want to compute $\frac{\partial J}{\partial \theta}\frac{\partial \theta}{\partial u}$.

For $\frac{\partial J}{\partial \theta}$ we know from question (b) that:

$$\frac{\partial J}{\partial \theta} = (\hat{y}_o - y_o)^T$$

For $\frac{\partial \theta}{\partial u}$:

$$\frac{\partial \theta}{\partial u} = v_c$$

Therefore

$$\begin{aligned}\frac{\partial J}{\partial u} &= \frac{\partial J}{\partial \theta}\frac{\partial \theta}{\partial u} \\ &= (\hat{y} - y)^T v_c\end{aligned}$$

(d) (3 Points)

The sigmoid function is given by:

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

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

Answer:

Let $z = e^x \text{ and } \sigma(z) = \frac{z}{z+1}$

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

(e) (4 points)

Now we shall consider the Negative Sampling loss, which is an alternative to the Naive Softmax loss. Assume that K negative samples (words) are drawn from the vocabulary. For simplicity of notation we shall refer to them as $w_1; w_2; \cdots; w_K$ and their outside word vectors as $u_1; u_2; \cdots; u_K$. Note that $o \notin \{w_1; w_2; \cdots; w_K\}$. For a center word c and an outside word o, the negative sampling loss function is given by:

$$\boldsymbol{J}_{\text{neg-sample}}\ (\boldsymbol{v}_c, o, \boldsymbol{U}) = -\log\left(\sigma\left(\boldsymbol{u}_o^\top \boldsymbol{v}_c\right)\right) - \sum_{k=1}^{K} \log\left(\sigma\left(-\boldsymbol{u}_k^\top \boldsymbol{v}_c\right)\right)$$

for a sample $w_1; w_2; \cdots; w_K$, where $\sigma(\cdot)$ is the sigmoid function.

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. Please write your answers in terms of the vectors $u_o, v_c,$ and $u_k,$ where $k \in [1, K]$. After you've done this, describe with one sentence why this loss function is much more efficient to compute than the naive-softmax loss. Note, you should be able to use your solution to part (d) to help compute the necessary gradients here.

Answer:

1. with respect to center word vector v_c

$$\begin{aligned}\frac{\partial J_{neg-sample}}{\partial v_c} &= \frac{\partial(-\log\left(\sigma\left(\boldsymbol{u}_o^\top \boldsymbol{v}_c\right)\right) - \sum_{k=1}^{K}\log\left(\sigma\left(-\boldsymbol{u}_k^\top \boldsymbol{v}_c\right)\right))}{\partial v_c} \\ &= -\frac{\sigma(u_o^T v_c)(1-\sigma(u_o^T v_c))}{\sigma(u_o^T v_c)}\frac{\partial u_o^T v_c}{\partial v_c} - \sum_{k=1}^{K}\frac{\partial log(\sigma(-u_k^T v_c))}{\partial v_c} \\ &= -(1-\sigma(u_o^T v_c))u_o + \sum_{k=1}^{K}(1-\sigma(-u_k^T v_c))u_k\end{aligned}$$

2. with respect to positive sample's outside word vector u_o

We only need to consider the case whereby the word is not a negative sample. Therefore the second part of the loss function can be ignored.

$$\begin{aligned}\frac{\partial J_{neg-sample}}{\partial u_o} &= \frac{\partial(-\log\left(\sigma\left(\boldsymbol{u}_o^\top \boldsymbol{v}_c\right)\right) - \sum_{k=1}^{K}\log\left(\sigma\left(-\boldsymbol{u}_k^\top \boldsymbol{v}_c\right)\right))}{\partial u_o} \\ &= \frac{\partial(-\log\left(\sigma\left(\boldsymbol{u}_o^\top \boldsymbol{v}_c\right)\right)}{\partial u_o} \\ &= -(1-\sigma(u_o^T v_c))v_c\end{aligned}$$

3. with respect to a negative sample's outside word vector u_k

We only need to consider the case whereby the word is a negative sample. Therefore the first part of the loss function can be ignored. Note that here is only asking for one negative sample and therefore the summation can be removed.

$$\begin{aligned}\frac{\partial J_{neg-sample}}{\partial u_k} &= \frac{\partial(-\log\left(\sigma\left(\boldsymbol{u}_o^\top \boldsymbol{v}_c\right)\right) - \sum_{k=1}^{K}\log\left(\sigma\left(-\boldsymbol{u}_k^\top \boldsymbol{v}_c\right)\right))}{\partial u_k} \\ &= \frac{\partial log(\sigma(-u_k^T v_c))}{\partial u_k} \\ &= (1-\sigma(-u_k^T v_c))v_c\end{aligned}$$

(f) (3 points)

Suppose the center word is $c = w_t$ and the context window is $[w_{t-m}, \cdots, w_{t-1}, w_t, w_{t+1}, \cdots, w_{t+m}]$, where m is the context window size. Recall that for the skip-gram version of word2vec (using center word to predict outside words), the total loss for the context window is:

$$J_{skip-gram}(v_c, w_{t-m}, \cdots, w_{t+m}, U) = \sum_{-m \le j \le m, j \neq 0} J(v_c, w_{t+j}, U)$$

Here, $J(v_c, w_{t+j}, U)$ represents an arbitrary loss term for the center word $c = w_t$ and outside word w_{t+j}. $J(v_c, w_{t+j}, U)$ could be $J_{naive-softmax}(v_c, w_{t+j}, U)$ or $J_{neg-sample}(v_c, w_{t+j}, U)$, depending on your implementation.

Write down three partial derivatives:

(i) $\frac{\partial J_{skip-gram}(v_c, w_{t-m}, \cdots, w_{t+m}, U)}{\partial U}$

(ii) $\frac{\partial J_{skip-gram}(v_c, w_{t-m}, \cdots, w_{t+m}, U)}{\partial v_c}$

(iii) $\frac{\partial J_{skip-gram}(v_c, w_{t-m}, \cdots, w_{t+m}, U)}{\partial v_c}$ when $w \neq c$

Write your answers in terms of $\frac{\partial J(v_c, w_{t+j}, U)}{\partial U}$ and $\frac{\partial J(v_c, w_{t+j}, U)}{\partial v_c}$. This is very simple – each solution should be one line.

Answer:

1.

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

2.

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

3.

$$\frac{\partial J_{skip-gram}(v_c, w_{t-m}, \cdots, w_{t+m}, U)}{\partial v_w}_{\text{where } w \neq c} = 0$$