Домашнее задание
Natural Language Processing - Word2Vec

Газимова Айгуль, 796

декабрь 2020 г.

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

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

Решение. Заметим, что среди компонент $(y_w)_{w \in V}$ одна 1 на позиции $w = o$, а все остальные - нули. Поэтому получаем:

$$-\sum_{w \in V} y_w \log(\hat{y}_w) = -1 \cdot \log(\hat{y}_o) - \sum_{\substack{w \in V \\ w \neq o}} 0 \cdot \log(\hat{y}_w) = -\log(\hat{y}_o)$$

■

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

Решение. Известно:

- $\hat{y}_w = P(W = w | C = c) = \frac{\exp u_w^T v_c}{\sum\limits_{x \in V} \exp(u_x^T v_c)}$
- $J_{\text{naive-softmax}}(v_c, o, U) = -\log\big(\hat{y}_o\big)$

$$\begin{aligned} J_{\text{naive-softmax}}(v_c, o, U) = \log \frac{\exp(u_o^T v_c)}{\sum\limits_{x \in V} \exp(u_x^T v_c)} &= \\ &= -\log\big(\exp(u_o^T v_c)\big) + \log\big(\sum_{x \in V} \exp(u_x^T v_c)\big) = \\ &= -u_o^T v_c + \log\big(\sum_{x \in V} \exp(u_x^T v_c)\big) \end{aligned}$$

Получаем, что:

$$
\begin{aligned}
\frac{\partial J_{\text{naive-softmax}}(v_c, o, U)}{\partial v_c} &= -u_o + \frac{1}{\sum\limits_{x \in V} \exp(u_x^T\, v_c)} \sum_{w \in V} \big(\exp(u_w^T\, v_c)\, u_w\big) = \\
&= -u_o + \sum_{w \in V} \left[\frac{\exp(u_w^T\, v_c)}{\sum\limits_{x \in V} \exp(u_x^T\, v_c)}\, u_w \right] = \\
&= -u_o + \sum_{w \in V} \hat{y}_w\, u_w = \\
&= -u_o + U\, \hat{y}
\end{aligned}
$$

■

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

Решение. Из предыдущей задачи воспользуемся тем, что:

$$
J_{\text{naive-softmax}}(v_c, o, U) = -u_o^T\, v_c + \log\big(\sum_{x \in V} \exp(u_x^T\, v_c)\big)
$$

Рассмотрим 2 случая:

1. $w = o$:

$$
\frac{\partial J_{\text{naive-softmax}}(v_c, o, U)}{\partial u_o} = -v_c + \frac{\exp(u_o^T\, v_c)}{\sum\limits_{w \in V} \exp(u_w^T\, v_c)}\, v_c = -v_c + \hat{y}_o\, v_c = v_c\,(\hat{y}_o - 1)
$$

2. $w \neq o$:

$$
\frac{\partial J_{\text{naive-softmax}}(v_c, o, U)}{\partial u_w} = \frac{\exp(u_w^T\, v_c)}{\sum\limits_{x \in V} \exp(u_x^T\, v_c)}\, v_c = \hat{y}_w\, v_c
$$

■

Задача d. The sigmoid function is given by Equation 4:

$$
\sigma(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1} \tag{4}
$$

Please compute the derivative of $\sigma(x)$ with respect to x, where x is scalar. Hint: you may want to write your answer in terms of $\sigma(x)$.

Решение.
$$\begin{aligned}\frac{\partial \sigma(x)}{\partial x} &= \left(\frac{\exp(x)}{\exp(x)+1}\right)'_x = \\ &= \frac{\exp(x)\left(\exp(x)+1\right) - \exp(2x)}{(\exp(x)+1)^2} = \\ &= \frac{\exp(x)}{\exp(x)+1}\cdot\frac{1}{\exp(x)+1} = \\ &= \frac{\exp(x)}{\exp(x)+1}\cdot\left(1-\frac{\exp(x)}{\exp(x)+1}\right) = \\ &= \sigma(x)\left(1-\sigma(x)\right)\end{aligned}$$

■

Задача e. Now we shall consider the Negative Sampling loss, which is an alternative of 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 $w_1, w_2, \dots, w_K$ and their outside vectors as $\mathbf{u_1}, \dots, \mathbf{u_K}$. Note that $o \notin \{w_1, \dots, w_K\}$. For a center word c and an outside word o, the negative sampling loss function is given by:

$$\mathbf{J}_{\text{neg-sample}}(\mathbf{v_c}, o, \mathbf{U}) = -\log(\sigma(u_o^T\, v_c)) - \sum_{k=1}^{K} \log(\sigma(-u_k^T\, v_c)) \tag{5}$$

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

Please repeat parts (b) and (c), computing the partial derivatives of $\mathbf{J}_{\text{neg-sample}}$ with respect to $\mathbf{v_c}$, with respect to $\mathbf{u_o}$, and with respect to a negative sample $\mathbf{u_k}$. Please write your answers in terms of the vectors $\mathbf{u_o}$, $\mathbf{v_c}$ and $\mathbf{u_k}$, 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, that you should be able to use your solution to part (d) to help compute the necessary gradients here.

Решение. Посчитаем частные производные $\log(\sigma(x))$ и $\log(\sigma(-x))$ по x:

$$\begin{aligned}\frac{\partial \log(\sigma(x))}{\partial x} &= \frac{\partial\sigma(x)/\partial x}{\sigma(x)} = \frac{\cancel{\sigma(x)}\,(1-\sigma(x))}{\cancel{\sigma(x)}} = 1-\sigma(x) \\ \frac{\partial \log(\sigma(-x))}{\partial x} &= \frac{\partial\sigma(-x)/\partial x}{\sigma(-x)} = \frac{-\cancel{\sigma(-x)}\,(1-\sigma(-x))}{\cancel{\sigma(-x)}} = \\ &= \sigma(-x) - 1 = 1-\sigma(x) - 1 = -\sigma(x)\end{aligned}$$

1. Частная производная по v_c:

$$\frac{\partial J_{\text{neg-sample}}(v_c, o, U)}{\partial v_c} = -\frac{\partial \log(\sigma(u_o^T\, v_c))}{\partial v_c} - \sum_{k=1}^{K} \frac{\partial \log(\sigma(-u_k^T\, v_c))}{\partial v_c} =$$
$$= -\big(1 - \sigma(u_o^T\, v_c)\big)\, u_o + \sum_{k=1}^{K} \sigma(u_k^T\, v_c)\, u_k$$

2. Частная производная по u_o:

$$\begin{aligned}\frac{\partial J_{\text{neg-sample}}(v_c, o, U)}{\partial u_o} &= -\frac{\partial \log(\sigma(u_o^T\, v_c))}{\partial u_o} - \sum_{k=1}^{K} \frac{\partial \log(\sigma(-u_k^T\, v_c))}{\partial u_o} = \\ &= |\text{поскольку } u_o \text{ не входит в } u_1, \ldots, u_K| = \\ &= -\big(1 - \sigma(u_o^T\, v_c)\big)\, v_c = \\ &= \big(\sigma(u_o^T\, v_c) - 1\big)\, v_c\end{aligned}$$

3. Частная производная по u_k:

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

При обучении Word2Vec для каждого из обучаемых векторов необходимо считать огромную сумму в знаменателе ($\sum\limits_{w \in V} \exp(u_w^T\, v_c)$) по пространству большой размерности V (т. к. слов обычно бывает очень много). Negative-sample эффективнее тем, что помогает сильно уменьшить вычислительные затраты.

■

Задача f. Suppose the center word is $c = w_t$ and the context window is $[w_{t-m}, \ldots, w_{t-1}, w_t, w_{t+1}, \ldots, w_{t+m}]$, where m is the context window size. Recall that for the skip-gram version of ‘word2vec‘, the total loss for the context window is:

$$\mathbf{J}_{\text{skip-gram}}(\mathbf{v_c}, w_{t-m}, \ldots, w_{t+m}, \mathbf{U}) = \sum_{\substack{-m \le j \le m \\ j \ne 0}} \mathbf{J}(\mathbf{v_c}, w_{t+j}, \mathbf{U})$$

Here, $\mathbf{J}(\mathbf{v_c}, w_{t+j}, \mathbf{U})$ represents an arbitrary loss term for the center word $c = w_t$ and outside word w_{t+j}. $\mathbf{J}(\mathbf{v_c}, w_{t+j}, \mathbf{U})$ could be $\mathbf{J}_{\text{naive-softmax}}(\mathbf{v_c}, w_{t+j}, \mathbf{U})$ or $\mathbf{J}_{\text{neg-sample}}(\mathbf{v_c}, w_{t+j}, \mathbf{U})$, depending on your implementation.
Write down three partial derivatives:

1. $\partial \mathbf{J}_{\text{skip-gram}}(\mathbf{v_c}, w_{t+j}, \mathbf{U})/\partial \mathbf{U}$
2. $\partial \mathbf{J}_{\text{skip-gram}}(\mathbf{v_c}, w_{t+j}, \mathbf{U})/\partial \mathbf{v_c}$
3. $\partial \mathbf{J}_{\text{skip-gram}}(\mathbf{v_c}, w_{t+j}, \mathbf{U})/\partial \mathbf{v_w}$ when $w \neq c$

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

Решение.

$$\partial \mathbf{J}_{\text{skip-gram}}(\mathbf{v_c}, w_{t+j}, \mathbf{U})/\partial \mathbf{U} = \sum_{\substack{-m \le j \le m \\ j \neq 0}} \partial \mathbf{J}(\mathbf{v_c}, w_{t+j}, \mathbf{U})/\partial \mathbf{U}$$

$$\partial \mathbf{J}_{\text{skip-gram}}(\mathbf{v_c}, w_{t+j}, \mathbf{U})/\partial \mathbf{v_c} = \sum_{\substack{-m \le j \le m \\ j \neq 0}} \partial \mathbf{J}(\mathbf{v_c}, w_{t+j}, \mathbf{U})/\partial \mathbf{v_c}$$

$$\partial \mathbf{J}_{\text{skip-gram}}(\mathbf{v_c}, w_{t+j}, \mathbf{U})/\partial \mathbf{v_w} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

■