به نام خدا

درس مبانی پردازش زبان و گفتار

تمرین سری دوم

مدرس درس:
جناب آقای دکتر اعتمادی

تهیه شده توسط:
الناز رضایی ۹۸۴۱۱۳۸۷

تاریخ ارسال: ۱۴۰۱/۱۲/۲۶

سوال ۱:

پاسخ (a)

y is a one-hot vector, so we have:

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

Therefore, we rewrite the equation as follows:

$$-\sum_{w \in Vocab} y_w \log(\hat{y_w}) = -(y_1 \log(\hat{y_1}) + ... + y_o \log(\hat{y_o}) + ... + y_n \log(\hat{y_n}))$$

$$= -(0 + ... + 1 * \log(\hat{y_o}) + ... + 0) = -\log(\hat{y_o})$$

پاسخ (b)

i) Consider input vector as i $= U^T v_c$. Also we know that $\hat{y} = S(i)$ (S represents softmax function).

$$\frac{\partial J}{\partial v_c} = \frac{\partial J}{\partial i} * \frac{\partial i}{\partial v_c} = (\hat{y} - y)^T * U^T = (\hat{y} - y)U$$

ii) If prediction function predict the correct output, then $\hat{y}$ is equal to y. Therefore, $(\hat{y} - y) = 0$ and according to part (i), the result of gradient equals to zero.

iii) We want to find a direction in which we should move from a random point to increase the altitude the fastest. That direction is given by the gradient vector at that particular point. Thus, the gradient gives us a direction of steepest ascent. In our gradient descent algorithm, we aim to minimize the value of loss function, therefore at any given point, we need to move in direction where the value of the function decreases the most, i.e. in the opposite direction of gradient which is direction of steepest descent. Therefore we subtract the gradient in the algorithm.

Auxiliary link

iv) When the downstream applications only care about the direction of the word vectors (e.g. they only pay attention to the cosine similarity of two words), then normalize, and forget about length.

If the downstream applications are able to (or need to) consider more sensible aspects, such as word significance, or consistency in word usage (see below), then normalization might not be such a good idea.

Auxiliary link

پاسخ (c)

$$\frac{\partial J}{\partial u_w} = \begin{cases} -v_c + (P(O=o|C=c))v_c & \text{if w = o} \\ 0 + (P(O=o|C=c))v_c & \text{otherwise} \end{cases}$$

Thus:

$$\frac{\partial J}{\partial u_w} = (\hat{y} - y)v_c$$

پاسخ (d)

$$\begin{bmatrix} (\hat{y_1} - y_1)v_c & (\hat{y_2} - y_2)v_c & .... & (y_{|V\hat{o}cab|} - y_{|Vocab|})v_c \end{bmatrix}$$

پاسخ (e)

$$\frac{\partial f(x)}{\partial x} = \begin{cases} \frac{\partial \alpha x}{\partial x} = \alpha & \text{if x} < 0 \\ \frac{\partial x}{\partial x} = 1 & \text{if x} > 0 \end{cases}$$

پاسخ (f)

$$\frac{\partial e^x}{\partial x} = e^x$$

$$\frac{\partial \sigma(x)}{\partial x} = \frac{\partial \frac{e^x}{1+e^x}}{\partial x} = \frac{(e^x * (1+e^x)) - (e^x * e^x)}{(1+e^x)^2} = \frac{e^x * (1+e^x - e^x)}{(1+e^x)^2}$$

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

پاسخ (g)

i)

$$\frac{\partial J}{\partial v_C} = \frac{\partial j}{\partial i}.\frac{\partial i}{\partial v_c} = \frac{-\sigma(u_o^T v_c)(1-\sigma(u_o^T v_c))}{u_o^T v_c}.u_o + \sum_{k=1}^{k}(1-\sigma(u_o^T v_c)).u_k$$

$$= -u_o(1-\sigma(u_o^T v_c)) + \sum_{k=1}^{k}(1-\sigma(u_o^T v_c)).u_k$$

$$\frac{\partial J}{\partial u_o} = \frac{\partial j}{\partial i}.\frac{\partial i}{\partial u_o} = (1-\sigma(u_o^T v_c)).v_c$$

$$\frac{\partial J}{\partial u_k} = \frac{\partial j}{\partial i}.\frac{\partial i}{\partial u_k} = (1-\sigma(u_o^T v_c)).v_c$$

ii)

$$\begin{bmatrix} (1-\sigma(u_o^T v_c)).v_c & -(1-\sigma(u_{w_1}^T v_c)).v_c & \ldots & -(1-\sigma(u_{w_k}^T v_c)).v_c \end{bmatrix}$$

We reuse $-(1-\sigma(u_{w_i}^T v_c))$, $i \in [1,k]$

iii) Negative sampling reduced the computations in the most cumbersome last layer, thereby making the gradient update procedure efficient.

پاسخ (h)

$$\frac{\partial J}{\partial u_{w_s}} = v_c(1 - \sigma(u_{w_s}^T v_c))$$

پاسخ (i)

i)

$$\frac{\partial j}{\partial U} = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \frac{J(v_c, wt + j, U)}{\partial U}$$

ii)

$$\frac{\partial j}{\partial v_c} = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \frac{J(v_c, wt + j, U)}{\partial v_c}$$

iii)

$$\frac{\partial j}{\partial w}(w \neq c) = 0$$