Date: تاریخ:

Subject: موضوع:

a) As mentioned before, y is a one-hot vector with a 1 for the true outside word o, so the answer of Σ is just $-\log(\hat{y}_o)$.

b) i) $$\frac{\partial J_{naive\text{-}softmax}(v_c, o, U)}{\partial v_c} = \frac{\partial}{\partial v_c}\left(-\log\left(\frac{\exp(u_o^T v_c)}{\sum_{w \in Vocab} \exp(u_w^T v_c)}\right)\right)$$

$$= \frac{\partial}{\partial v_c}\left(-\log(\exp(u_o^T v_c)) + \log\left(\sum_{w \in Vocab} \exp(u_w^T v_c)\right)\right)$$

$$= \frac{\partial}{\partial v_c}\left(-\exp(u_o^T v_c) + \log\left(\sum_{w \in Vocab} \exp(u_w^T v_c)\right)\right)$$

$$= -u_o + \underbrace{\frac{\sum_{w \in Vocab} u_w \exp(u_w^T v_c)}{\sum_{w \in Vocab} \exp(u_w^T v_c)}}_{\sum_{w \in Vocab} \hat{y}_w u_w}$$

$$= U\hat{y} - u_o$$

Date: تاریخ:

Subject: موضوع:

b) ii) the softmax prediction is perfect and equal to true distribution.

iii) in this gradiant ,each time we update the center word and the loss function get minimized which means the probability get closer.

iv) Sometimes word vectors are really similar to eachother according to similarity of that words in meaning. when our data is too big and we need to make it small considering that words equal is good for us but sometimes it takes useful information from us and may effect badly in our downstream task for example word vector of "very bad" and "not bad" maybe very similar to eachother and by doing this normalization they will consider the same but we know that they aren't.

تاریخ:
Date:
موضوع:
Subject:

c) $\frac{\partial J_{naive\text{-}softmax}}{\partial U} = \frac{\partial}{\partial U}\left(-\log\left(\frac{\exp(u_o^T v_c)}{\sum_{w \in Vocab} \exp(u_w^T v_c)}\right)\right)$

$U y^T$

$= \frac{-\partial u_o^T v_c}{\partial U} + \frac{1}{\sum_{w \in Vocab} \exp(u_w^T v_c)} \left(\sum_{w \in Vocab} \exp(u_w^T v_c) \times \frac{\partial u_w^T v_c}{\partial U}\right)$

$= -v_c y^T + \underbrace{\sum_{w \in Vocab} \hat{y}_w \times \frac{\partial u_w^T v_c}{\partial U}}_{v_c \hat{y}^T} = v_c(\hat{y}^T - y^T)$

$= v_c(\hat{y} - y)^T$

$\rightarrow \begin{cases} \text{if } w = 0 \rightarrow v_c \hat{y}^T - v_c \\ \text{if } w \neq 0 \rightarrow v_c \hat{y}^T \end{cases}$

d) $\frac{\partial J_{naive\text{-}softmax}}{\partial U} = \left[\frac{\partial J(v_c, o, U)}{\partial u_1} \; \frac{\partial J(v_c, o, U)}{\partial u_2} \; \cdots \; \frac{\partial J(v_c, o, U)}{\partial u_{|Vocab|}}\right]$

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

Date: تاریخ:

Subject: موضوع:

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

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

$$= \sigma(x) - \sigma^2(x)$$

g) i) $$\frac{\partial J_{neg\_sample}}{\partial v_c} = \frac{\partial}{\partial v_c}\left(-\log(\sigma(u_o^T v_c)) - \sum_{s=1}^{K} \log(\sigma(-u_{ws}^T v_c))\right)$$

$$= \frac{-1}{\sigma(u_o^T v_c)} \times \left(u_o \times \left(\sigma(u_o^T v_c) - \sigma(u_o^T v_c)^2\right)\right)$$

$$- \sum_{s=1}^{K}\left(\frac{1}{\sigma(-u_{ws}^T v_c)} \times \left(-u_{ws} \times \left(\sigma(-u_{ws}^T v_c) - \sigma(-u_{ws}^T v_c)^2\right)\right)\right)$$

$$= -u_o(1-\sigma(u_o^T v_c)) + \sum_{s=1}^{K} u_{ws}(1-\sigma(-u_{ws}^T v_c))$$

$$\frac{\partial J_{neg\_sample}}{\partial u_o} = \frac{\partial}{\partial u_o}\left(-\log(\sigma(u_o^T v_c)) - \sum_{s=1}^{K} \log(\sigma(-u_{ws}^T v_c))\right)$$

$$= \frac{-1}{\sigma(u_o^T v_c)} \times \left(v_c \times \left(\sigma(u_o^T v_c) - \sigma(u_o^T v_c)^2\right)\right) - 0$$

$$= -v_c(1-\sigma(u_o^T v_c))$$

Date: تاریخ:

Subject: موضوع:

g) i) $\frac{\partial J_{neg\text{-}sample}}{\partial u_{ws}} = \frac{\partial}{\partial u_{ws}}\left(-\log(\sigma(u_o^T v_c) - \sum_{s=1}^{K} \log(\sigma(-u_{ws}^T v_c))\right)$

$= 0 - \frac{1}{\sigma(-u_{ws}^T v_c)} \times (-v_c \times (\sigma(-u_{ws}^T v_c) - \sigma(-u_{ws}^T v_c)^2)$

$= v_c (1 - \sigma(-u_{ws}^T v_c))$

ii) the first one because it has both terms and we can use it in second and third one.

$\frac{\partial J_{neg\text{-}sample}}{\partial v_c} = -u_o (1 - \sigma(u_o^T v_c))$

iii) because in negative sampling we compute on just K words but in softmax the computation is on the whole vocabulary words, so using negative sampling is more efficient.

Date: تاریخ:

Subject: موضوع:

h) because this time our words are not distinct so the answer would be the Σ of the words which are equal to w_s so:

$$\frac{\partial J_{neg-sampling}}{\partial u_{w_s}} = \sum_{w=w_s} v_c \left(1 - \sigma(-u_{w_s}^T v_c)\right)$$

i) $$\sum_{\substack{-m \le j \le m \\ j \neq 0}} \frac{\partial J(v_c, w_{t+j}, U)}{\partial U}$$

ii) $$\sum_{\substack{-m \le j \le m \\ j \neq 0}} \frac{\partial J(v_c, w_{t+j}, U)}{\partial v_c}$$

iii) $= 0$