

Assignment #2 writting answers

(a) Recall that the cross entropy loss between the true probability distribution p and another distribution q is $-\sum_i p_i \log(q_i)$. With given center word c , y is the true emprical distribution(a one-hot vector with a 1 for the true outside word o , and 0 everywhere), and \hat{y} is the predicted distribution, k^{th} entry in these vectors indicates the conditional probability of the k^{th} word being an outside word for the given c . So the cross entropy loss between y and \hat{y} is:

$$-\sum_{w \in Vocab} y_w \log(\hat{y}_w) = -\sum_{w \neq o, w \in Vocab} 0 * \log(\hat{y}_w) - 1 * \log(\hat{y}_o) = -\log(\hat{y}_o)$$

just the same as the naive softmax loss for single pair words of o and c :

$$J_{naive-softmax}(v_c, o, U) = -\log P(O = o | C = c)$$

(b)

$$\begin{aligned} & \frac{\partial}{\partial v_c} J_{naive-softmax}(v_c, o, U) \\ &= \frac{\partial}{\partial v_c} -\log(\hat{y}_o) \\ &= -\frac{\partial}{\partial v_c} \log\left(\frac{\exp(u_o^T v_c)}{\sum_{w \in Vocab} \exp(u_w^T v_c)}\right) \\ &= -\left(\frac{\partial}{\partial v_c} \log \exp(u_o^T v_c) - \frac{\partial}{\partial v_c} \log \sum_{w \in Vocab} \exp(u_w^T v_c)\right) \\ &= -\left(u_o - \frac{1}{\sum_{w \in Vocab} \exp(u_w^T v_c)} \sum_{w \in Vocab} \frac{\partial}{\partial v_c} \exp(u_w^T v_c)\right) \\ &= -\left(u_o - \frac{1}{\sum_{w \in Vocab} \exp(u_w^T v_c)} \sum_{w \in Vocab} \exp(u_w^T v_c) u_w\right) \\ &= -\left(u_o - \frac{\sum_{w \in Vocab} \exp(u_w^T v_c) u_w}{\sum_{w \in Vocab} \exp(u_w^T v_c)}\right) \\ &= -\left(u_o - \sum_{x \in Vocab} \frac{\exp(u_x^T v_c)}{\sum_{w \in Vocab} \exp(u_w^T v_c)} u_x\right) \\ &= -\left(u_o - \sum_{x \in Vocab} \hat{y}_x u_x\right) \\ &= \sum_{x \in Vocab} \hat{y}_x u_x - u_o \\ &= \sum_{x \in Vocab} \hat{y}_x u_x - \sum_{x \in Vocab} y_x u_x \\ &= U(\hat{y} - y)^T \end{aligned}$$

(c) case 1: $w \neq o$

$$\begin{aligned} & \frac{\partial}{\partial u_w} J_{naive-softmax}(v_c, o, U) \\ &= \frac{\partial}{\partial u_w} -\log(\hat{y}_o) \\ &= -\frac{\partial}{\partial u_w} \log\left(\frac{\exp(u_o^T v_c)}{\sum_{w \in Vocab} \exp(u_w^T v_c)}\right) \\ &= -\left(\frac{\partial}{\partial u_w} \log \exp(u_o^T v_c) - \frac{\partial}{\partial u_w} \log \sum_{w \in Vocab} \exp(u_w^T v_c)\right) \\ &= \frac{\partial}{\partial u_w} \log \sum_{w \in Vocab} \exp(u_w^T v_c) \\ &= \frac{1}{\sum_{w \in Vocab} \exp(u_w^T v_c)} \sum_{w \neq o} \frac{\partial}{\partial u_w} \exp(u_w^T v_c) \\ &= \frac{1}{\sum_{w \in Vocab} \exp(u_w^T v_c)} \sum_{w \neq o} \exp(u_w^T v_c) v_c \\ &= \sum_{x \neq o} \frac{u_x^T v_c}{\sum_{w \in Vocab} \exp(u_w^T v_c)} v_c \\ &= \sum_{x \neq o} \hat{y}_x v_c \end{aligned}$$

case 2: $w = o$

$$\begin{aligned} & \frac{\partial}{\partial u_o} J_{naive-softmax}(v_c, o, U) \\ &= \frac{\partial}{\partial u_o} -\log(\hat{y}_o) \\ &= -\frac{\partial}{\partial u_o} \log\left(\frac{\exp(u_o^T v_c)}{\sum_{w \in Vocab} \exp(u_w^T v_c)}\right) \\ &= -\left(\frac{\partial}{\partial u_o} \log \exp(u_o^T v_c) - \frac{\partial}{\partial u_o} \log \sum_{w \in Vocab} \exp(u_w^T v_c)\right) \\ &= -\left(v_c - \frac{1}{\sum_{w \in Vocab} \exp(u_w^T v_c)} \exp(u_o^T v_c) v_c\right) \\ &= \hat{y}_o v_c - v_c \end{aligned}$$

综合上述两种情况：

$$\begin{aligned} & \frac{\partial}{\partial u_x} J_{naive-softmax}(v_c, o, U) \\ &= \sum_{x \in Vocab} v_c (\hat{y}_x (1 - y_x) + (\hat{y}_x - y_x) y_x) \\ &= \sum_{x \in Vocab} v_c (\hat{y}_x - y_x) \\ &= (\hat{y} - y) v_c \end{aligned}$$

(d)

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

(e)

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

Negative sampling only updates u_o and K u_k , ($K + 1$) total, while naive softmax needs to update all the u_k .

(f)

$$\begin{aligned} \frac{\partial}{\partial U} J_{skip-gram}(v_c, w_{t-m}, \dots, w_{t+m}, U) &= \sum_{-m \leq j \leq m, j \neq 0} \frac{\partial J(v_c, w_{t+j}, U)}{\partial U} \\ \frac{\partial}{\partial v_c} J_{skip-gram}(v_c, w_{t-m}, \dots, w_{t+m}, U) &= \sum_{-m \leq j \leq m, j \neq 0} \frac{\partial J(v_c, w_{t+j}, U)}{\partial v_c} \\ \frac{\partial}{\partial v_w} J_{skip-gram}(v_c, w_{t-m}, \dots, w_{t+m}, U) &= 0, \forall w \neq c \end{aligned}$$