Zeyad Ezzat
4492

assignment 2

ⓐ $-\sum y_w \log(\hat{y}_w) = -\log(\hat{y}_o)$

y_w is one hot incoded so it's equal to zero if it's not the correct label

$y_w = 1$ only at the correct context

$-\sum(0, 0, 0, + y_o \log y_o + \dots) = -\log(\hat{y}_o)$

ⓑ $J_{naive\text{-}softmax} = -\log \dfrac{e^{u_o^T v_c}}{\sum_{w \in V} e^{u_w^T v_c}}$

$= -u_o^T v_c + \log \sum e^{u_w^T v_c}$

$J(\ ) = \log \sum e^{u_w^T v_c} - u_o^T v_c$

$\dfrac{\partial J}{\partial v_c} = \sum_w \dfrac{u_w e^{u_w^T v_c}}{\sum_w e^{u_w^T v_c}} - u_o$

$= \sum u_w \hat{y}_w - u_o$ $\quad u_o = y_{w=o} = y_w$

$= \sum u_w \hat{y}_w - y_w$

ⓒ $J = -u_{(o)}^T v_c + \log \sum e^{u_w^T v_c}$

$\dfrac{J}{\partial w} = -y_w v_c + \dfrac{v_c\ e^{u_w^T v_c}}{\sum e^{u_w^T v_c}}$

$= v_c(\hat{y}_w - y_w) \rightarrow \boxed{w = o}$

$\boxed{= v_c \hat{y}_w \quad w \neq o}$

(d)

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

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

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

$$= \frac{e^x}{e^x+1} - \frac{e^x}{e^x+1}\,\frac{e^x}{e^x+1}$$

$$= \sigma(x) - \sigma(x)\,\sigma(x)$$

$$= \boxed{\sigma(x)\,(1-\sigma(x))}$$

(e)

$$x_o = u_o^T v_c$$

$$\frac{\partial x_c}{\partial v_c} = u_o$$

$$J_{neg\text{-}sample} = -\log(\sigma(x_o)) - \sum_{k=1}^{K} \log(\sigma(x_k))$$

$$\frac{\partial J}{\partial v_c} = -\frac{\frac{\partial\sigma(x_o)}{\partial v_c}}{\sigma(x_o)} - \sum \frac{\frac{\partial\sigma(x_k)}{\partial v_c}}{\sigma(x_k)}$$

$$\frac{\partial J}{\partial v_c} = -\frac{\cancel{\sigma(x)}\,(1-\sigma(x))\,u_o}{\cancel{\sigma(x_o)}} - \sum \frac{\cancel{\sigma(x_k)}\,(1-\sigma(x_k))(-u_k)}{\cancel{\sigma(x_k)}}$$

$$\frac{\partial J}{\partial v_c} = -(1-\sigma(x_o))\,u_o + \sum_k (1-\sigma(x_k))\,u_k$$

$$= -(1-\sigma(u_o^T v_c))\,u_o + \sum (1-\sigma(-u_k^T v_c))\,u_k$$

Scanned by CamScanner

(e)

$$J_{neg\text{-}sample} = -\log(\sigma(u_o^T v_c)) - \sum_k \log \sigma(-u_k^T v_c)$$ $k \neq 0$

$$\frac{\partial J}{\partial u_k} = \text{zero} - \frac{(1-\sigma(-u_w^T v_c))\,\cancel{\sigma(-u_w^T v_c)}(-v_c)}{\cancel{\sigma}(-u_w^T v_c)}$$ $\boxed{w \neq 0}$

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

$$\frac{\partial J}{\partial u_o} = -\frac{\cancel{\sigma(u_o^T v_c)}(1-\sigma(u_o^T v_c))v_c}{\cancel{\sigma(u_o^T v_c)}} - \text{zero} \longrightarrow k \neq 0$$ $w = 0$

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

This a lot better because we don't need to calculate the normalization factor

(f)

(f)

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

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

(iii) $\displaystyle -\sum \frac{\partial J(v_c, w_{t-m}, U)}{\partial v_c}$