श्री

classmate
Date 10-06-21
Page

CS-224 N

(1)

Assignment - 2

(a) Ans:- To prove: $- \sum_{w \in Vocab} y_w \log(\hat{y}_w) = -\log(\hat{y}_o)$

$y \rightarrow$ one-hot vector with 1 at true outside word -o and 0 at other word's places.

$\therefore y = [0 \ldots \underset{\substack{\downarrow \\ \text{o-word's position}}}{1} \ldots 0] \qquad \hat{y} = [\ldots \underset{\substack{\downarrow \\ \text{Prediction for o-word}}}{\hat{y}_o} \ldots]$

$\therefore - \sum_{w \in Vocab} y * \log(\hat{y}_w) = -(0 \times \log(\hat{y}_0) + 0 \times \log(\hat{y}_1) \ldots + 1 \times \log(\hat{y}_o) \ldots)$

$= -\log(\hat{y}_o)$

hence proved

(b) Ans.

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

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

To find: $\dfrac{\partial J}{\partial v_c}$.

Ans: We know that $-\log\left(\dfrac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}\right)$

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

(2) We know that, $P(O=o \mid C=c) = \dfrac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}$ $\Rightarrow \hat{y}_w$

$$\therefore \frac{\partial J}{\partial v_c} = -u_o + \frac{\sum_w \exp(u_w^T v_c) \cdot u_w}{\sum_w \exp(u_w^T v_c)}$$

$$= -u_o + \sum_{w \in V} \underset{1\times 1}{\hat{y}_w} \cdot \underset{d\times 1}{u_w} \qquad [U \to d \times V]$$

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

$$\boxed{= -y \cdot U + \hat{y} \cdot U} \quad \underline{\underline{\text{Ans}}}$$

$\left(y \to 0 - - 0 - - 1 \right.$ for true label o $\left. \right)$

$\therefore y \cdot U = u_o$

(c) <u>Case I</u> : $w = o$,

$$\therefore J = -u_o^T v_c + \log\left(\sum_{w \in V} \exp(u_w^T v_c)\right)$$

$$\Rightarrow \frac{\partial J}{\partial u_o} = -v_c + \frac{1}{\sum_{w \in V} (u_w^T v_c)} \cdot v_c \cdot \exp(u_o^T v_c)$$

$$= -v_c + \hat{y}_o \cdot v_c$$

$$= v_c(\hat{y}_o - y)$$

(3)

Case II: $w \neq 0$, Let $w = k$

$$\therefore \frac{\partial J}{\partial u_k} = 0 + \frac{1}{\sum_{w \in V} \exp(u_w^T v_c)} \cdot \exp(u_k^T v_c) \cdot v_c$$

$$= \boxed{\hat{y}_k \cdot v_c}$$ Ans

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

$$\therefore \frac{dk}{dx} = -(1+e^{-x})^{-2} \times (0 + e^{-x} \times (-1))$$

$$\Rightarrow \frac{dk}{dx} = \frac{e^{-x}}{(1+e^{-x})^2} = k^2 \cdot e^{-x}$$

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

$$\therefore \frac{dk}{dx} = k^{\cancel{2}} \cdot \left(\frac{1-k}{\cancel{k}}\right)$$

$$\therefore \frac{d(\sigma(x))}{dx} = \sigma(x)(1-\sigma(x))$$

Ans

(4)

(e) $J = -\log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} \log(\sigma(-u_k^T v_c))$

$\therefore$ For any x, $$\frac{\partial J}{\partial x} = -\frac{1}{\cancel{\sigma(u_o^T v_c)}} \cdot \cancel{\sigma(u_o^T v_c)} \cdot (1-\sigma(u_o^T v_c)) \cdot \frac{\partial}{\partial x}(u_o^T v_c)$$

$$+ \sum_{k=1}^{K} \frac{1}{\cancel{\sigma(-u_k^T v_c)}} \cdot \cancel{\sigma(-u_k^T v_c)} (1-\sigma(-u_k^T v_c)) \cdot \frac{\partial}{\partial x}(u_k^T v_c)$$

$$\Rightarrow \frac{\partial J}{\partial x} = -(1-\sigma(u_o^T v_c)) \frac{\partial}{\partial x}(u_o^T v_c) + \sum_{k=1}^{K} (1-\sigma(-u_k^T v_c)) \frac{\partial}{\partial x}(u_k^T v_c)$$

$\therefore$ (1) $$\boxed{\frac{\partial J}{\partial v_c} = -(1-\sigma(u_o^T v_c)) \cdot u_o + \sum_{k=1}^{K} (1-\sigma(-u_k^T v_c)) \cdot u_k}$$

(2) $$\boxed{\frac{\partial J}{\partial u_o} = -(1-\sigma(u_o^T v_c)) \cdot v_c}$$

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

Ans

This loss is more efficient because we got rid of the Normalization term which made the naive - softmax computationally expensive.

(5)

(f) Given: $J_{sg} = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} J(v_c, w_{t+j}, U)$

$\therefore$ (i) $\frac{\partial J_{sg}}{\partial U} = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \frac{\partial J}{\partial U}$

(ii) $\frac{\partial J_{sg}}{\partial v_c} = \sum \frac{\partial J}{\partial v_c}$

(iii) $\frac{\partial J_{sg}}{\partial v_w} = 0$