CS 224n Assignment 2 Written Part.

1/a) Prove negative-sampling loss is the same as cross-entropy loss between y and $\hat{y}$, i.e. show that

$$-\sum_{w \in Vocab} y_w \log(\hat{y}_w) = -\log(\hat{y}_o)$$

The probability for an outside word o given the center word is defined as

$$\hat{y}_o = p(u_o | v_c) = \frac{\exp(u_o^T v_c)}{\sum_{i \in Vocab} \exp(u_i^T v_c)}$$

Therefore the following equation is ...

$$-\sum_{w \in Vocab} y_w \log(\hat{y}_w) = -y_1 \log(\hat{y}_1) - y_2 \log(\hat{y}_2) - \cdots - y_{|V|} \log(\hat{y}_{|V|})$$

Since for true label y, it's a one-hot encoded vector which only true outside context with value 1 and the rest as 0. So the previous equation could be simplified as.

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

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

1(b). partial derivative of $J_{naive-softmax}$ with respect to v_c.

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

$$= -\log \frac{\exp(u_o^T v_c)}{\sum_{w \in vocab} \exp(u_w^T v_c)}$$

$$= \underbrace{-u_o^T v_c}_{A} + \underbrace{\log\left[\sum_{w \in vocab} \exp(u_w^T v_c)\right]}_{B}$$

$$\frac{\partial J}{\partial v_c} = \frac{\partial A}{\partial v_c} + \frac{\partial B}{\partial v_c}$$

$$\frac{\partial A}{\partial v_c} = \frac{-\partial u_o^T v_c}{\partial v_c} = -u_o^T$$

$$\frac{\partial B}{\partial v_c} = \frac{\sum u_w^T \exp(u_w^T v_c)}{\sum \exp(u_w^T v_c)}$$

$$= \sum_{x \in vocab} \frac{\exp(u_x^T v_c)}{\sum_{w \in vocab} \exp(u_w^T v_c)} \cdot u_x^T$$

$$= \sum_{x \in vocab} P(u_x|v_c) \cdot u_x^T$$

$$= \sum_{x \in vocab} \hat{y}_x u_x$$

$$\therefore \frac{\partial J}{\partial v_c} = -u_o + \sum_{x \in vocab} \hat{y}_x u_x$$

The slope of loss function w.r.t center words is the difference between of the true context words and expected context words in the model.

v(c) Part 1 : if $u_w = u_o$.

$$\frac{\partial J}{\partial u_o} = \frac{\partial A}{\partial u_o} + \frac{\partial B}{\partial u_o}.$$

- $\frac{\partial A}{\partial u_o} = -v_c.$
- $\frac{\partial B}{\partial u_o} = \frac{v_c \exp(u_o^T v_c)}{\sum \exp(u_w^T v_c)}.$

$$\frac{\partial J}{\partial u_o} = -v_c + v_c \cdot \hat{y}_c = (\hat{y}_c - 1) \cdot v_c.$$

Part 2 if $u_w \neq u_o$.

- $\frac{\partial A}{\partial u_i} = 0.$
- $\frac{\partial B}{\partial u_i} = \frac{v_c \exp(u_i^T v_c)}{\sum_{w \neq o} \exp(u_w^T v_c)} = v_c \cdot \hat{y}_i$

$$\frac{\partial J}{\partial u_i} = 0 + v_c \cdot \hat{y}_i$$

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

$$\frac{d\sigma}{dx} = -\frac{1}{(e^{-x}+1)^2} \cdot e^{-x} \cdot -1$$

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

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

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

$$= \sigma(x)(1-\sigma(x))$$

1/e). $f(x) = \max(0, x)$ is the ReLu

if $x > 0$, $\frac{df(x)}{dx} = 1$.

$x < 0$. $\frac{df(x)}{dx} = 0$.

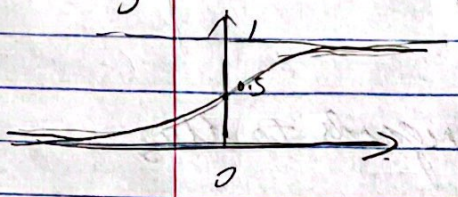
1/f). $\frac{dJ}{dw}$ is just the combination of partial derivatives of all w_i which $w_i \in$ locals.

$$\frac{dJ}{dw} = \begin{cases} (\hat{y}_i - 1) \cdot v_i & \text{if } i = 0. \\ \hat{y}_i \, v_i & \text{if } i \neq 0. \end{cases}$$

1/3). Negative Sample.

$$J_{neg\text{-}sample}(v_c, o, U) = -\log(\sigma(u_o^T v_c)) - \sum \log(\sigma(-u_w^T v_c))$$

As for Sigmoid function



i) $\frac{\partial J}{\partial v_c} = \frac{\partial J}{\partial \sigma} \cdot \frac{\partial \sigma}{\partial v_c}$.

$$= -\frac{u_o \sigma(u_o^T v_c)[1-\sigma(u_o^T v_c)]}{\sigma(u_o^T v_c)}$$

$$-\sum \frac{-u_w \sigma(-u_w^T v_c)[1-\sigma(-u_w^T v_c)]}{\sigma(-u_w^T v_c)}$$

$$= -u_o[1-\sigma(u_o^T v_c)] + \sum u_w[1-\sigma(-u_w^T v_c)].$$

ii) $\frac{\partial J}{\partial u_o} = -v_c[1-\sigma(u_o^T v_c)]$.

$\frac{\partial J}{\partial u_i} = v_c[1-\sigma(u_i^T v_c)]$ i≠o.

Negative Sample is more efficient because it takes only apart of the original sample while loss function considers all the sample.

1/(h) Now without the assumption that K sampled are distinct. which means we still have $O \notin \{w_1, w_2, \cdots w_K\}$ but now $i \neq j$ no longer holds.

$\frac{\partial J_{neg-sample}}{\partial v_c}$ is still the same

while for partial derivative with regards to u_{w_s}

$$\frac{\partial J_{neg-sample}}{\partial u_{w_s}} = n_{w_s} v_c [1 - \sigma(u_{w_s}^T v_c)]$$

n_{w_s} is the number of negative sample equals to w_s.

1/(i). $$\frac{\partial J_{skip-gram}}{\partial u} = \sum_{-m \le j \le m, j \neq 0} \frac{\partial J(v_c, w_j, u)}{\partial u}$$

$J(v_c, w_j, u)$ could be the $J_{softmax}$ or J_{naive} based on the setting