1 Understanding word2vec

(a)

$-\sum_{w \in V} y_w \log(\hat{y}_w) = \begin{bmatrix} y_w = 1 & \text{if } w = o \\ y_w = 0 & otherwise \end{bmatrix} = -\log(\hat{y}_o)$

(b) $J_{naive\_softmax}(v_c, o, U) = -\log P(O = o|C = c) = -\log \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}$

$$\frac{\partial J_{naive\_softmax}(v_c, o, U)}{\partial v_c} = \frac{\partial}{\partial v_c} - \log \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)} = \frac{\partial}{\partial v_c} - u_o^T v_c + \log \sum_{w \in V} \exp(u_w^T v_c) =$$

$$= u_o + \frac{\sum_{w \in V} \exp(u_w^T v_c) u_w}{\sum_{w \in V} \exp(u_w^T v_c)} = -u_o + \sum_{x \in V} \frac{exp(u_x^T)}{\sum_{w \in V} (u_w^T v_c)} u_x = -u_o + \sum_{x \in V} \hat{y}_x u_x = U\hat{y} - y$$

(c)
case $w \neq o$:

$$\frac{\partial J_{naive\_softmax}(v_c, o, U)}{\partial u_w} = \frac{\partial}{\partial u_w} - u_o^T v_c + \log \sum_{t \in V} \exp(u_t^T v_c) = \frac{v_c \exp(u_w^T v_c)}{\sum_{t \in V} \exp(u_t^T v_c)} = \hat{y}_w v_c$$

case $w = o$:

$$\frac{\partial J_{naive\_softmax}(v_c, o, U)}{\partial u_o} = -v_c + \frac{v_c \exp(U_o^T v_c)}{\sum_{t \in V} \exp(u_t^T v_c)} = (\hat{y}_o - 1) v_c$$

(d) $\sigma(x) = \frac{1}{1+\exp(-x)}$

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

$$= \frac{1}{1+\exp(-x)} * \frac{1+\exp(-x)-1}{1+\exp(-x)} = \frac{1}{1+\exp(-x)} * (1 - \frac{1}{1+\exp(-x)}) = \sigma(1-\sigma)$$

(e) $J_{neg\_sample}(v_c, o, U) = -log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} \log(\sigma(-u_k^T v_c))$

$$\frac{\partial J_{neg\_sample}(v_c, o, U)}{\partial v_c} = \frac{-\sigma(u_o^T v_c)(1-\sigma(u_o^T v_c))}{\sigma(u_o^T v_c)} - \sum_{k=1}^{K} \frac{\sigma(-u_k^T v_c)(1-\sigma(-u_k^T v_c))(-u_k)}{\sigma(-u_k^T v_c)} =$$

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

This objective is much more efficient because K is much smaller than vocabulary size, which might be several million.

(f) $J_{skip\_gram}(v_c, w_{t-m}, ..., w_{t+m}, U) = \sum_{-m \leq j \leq m; j \neq 0} J(v_c, w_{t+j}, U)$

$$\frac{\partial J_{skip\_gram}}{\partial U} = \frac{\partial \sum_{-m \leq j \leq m; j \neq 0} J(v_c, w_{t+j}, U)}{\partial U} = \sum_{-m \leq j \leq m; j \neq 0} \frac{\partial J(v_c, w_{t+j}, U)}{\partial U}$$

$$\frac{\partial J_{skip\_gram}}{\partial v_c} = \sum_{-m \leq j \leq m; j \neq 0} \frac{\partial J(v_c, w_{t+j}, U)}{\partial v_c}$$

$$\frac{\partial J_{skip\_gram}}{\partial v_{w \neq c}} = 0$$