

a. naive - softmax loss가 cross entropy loss와 같아진 이유

$$-\sum_{w \in \text{vocab}} y_w \log \hat{y}_w = -\log \hat{y}_0$$

$y$ 는 실제 클래스 분포,  $\hat{y}$ 는 모델에서 주한 클래스 분포

$y$ 는 context word  $o$ 에 해당하는 element 한 1인 one-hot vector  $\therefore$  참

b, c b = naive softmax loss 식을  $v_c$ 에 대해 편미분

$$\frac{\partial J}{\partial v_c} = -u_0 + \sum_{z \in \text{vocab}} p(z|c) u_z$$

실제 분포와 가중치가 있는 클래스 분포의 차이값 계산

$$\frac{\partial J}{\partial u_c} = u(\hat{y} - y)$$

c = naive softmax loss 식을  $u$ 에 대해 편미분 ( $u=0$ 인 경우)

$$\frac{\partial J}{\partial u_0} = (\hat{y}_0 - y_0) v_c$$

클래스 분포의 element 가 1이냐 0이냐 (scalar 값)

( $u \neq 0$ 인 경우)

$$\frac{\partial J}{\partial u_w} = \hat{y}_w v_c$$

$$\frac{\partial J}{\partial u} = (\hat{y} - y) v_c^T \quad / \quad \text{전체 } u \text{에 대해 편미분}$$

d sigmoid 편미분

$$\frac{\partial b}{\partial x} = b(1-b)$$

e negative sample에 대한 loss의 편미분 식, neg sample의 loss

$$J_{\text{neg-sample}}(v_c, o, u) = -\log(b(u_o^T v_c)) - \sum_{k=1}^K \log(b(-u_k^T v_c))$$

$K$  : negative samples

$o$  : neg sample  $\times$

<각각 미분한 결과>

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

$$\frac{\partial J}{\partial u_0} = -(1-b(u_o^T v_c)) v_c$$

$$\frac{\partial J}{\partial u_k} = \sum_{k=1}^K (1-b(-u_k^T v_c)) \frac{\partial u_k^T v_c}{\partial u_k}$$