CS224n Assignment 2 Solutions

Tianjian Li

February 2022

(a) Because

$$-\sum_{w\in vocab} y_w \log(\hat{y}_w) = -\Big(\sum_{w\in vocab\setminus\{o\}} y_w \log(\hat{y}_w)\Big) - y_o \log(\hat{y}_o)$$
$$= 0 - \log(\hat{y}_0)$$

Therefore

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

(b) The Cross Entropy Loss is given by

$$J = -\sum_{i=1}^{W} y_i \log(\hat{y}_i)$$

$$= -\sum_{i=1}^{W} y_i \log\big(\frac{\exp(u_i^\top v_c)}{\sum_{w=1}^{W}\exp(u_w^\top v_c)}\big)$$

We note that y is one hot encoded, so only one entry would be 1. Therefore,

$$J = -y_o\Big[u_i^\top v_c - \log\big(\sum_{w=1}^{W}\exp(u_w^\top v_c)\big)\Big]$$

Where o is the ground truth index.
Now we take the partial derivative w.r.t to v_c

$$\frac{\partial J}{\partial v_c} = -\Big[u_o - \frac{\sum_{w=1}^{W}\exp(u_w^\top v_c)u_w}{\sum_{x=1}^{W}\exp(u_x^\top v_c)}\Big]$$

$$= \sum_{w=1}^{W}\Big(\frac{\exp(u_w^\top v_c)}{\sum_{x=1}^{W}\exp(u_x^\top v_c)}u_w\Big) - u_o$$

Note that u_o is simply the encoding of the ground truth label, so $u_o = Uy$, and the minuend is simply $\sum_{w=1}^{W}(\hat{y_w}u_w)$, which is equal to $U\hat{y}$ Therefore

$$\frac{\partial J}{\partial v_c} = U(\hat{y} - y)$$

(c) This is similar to question (b). When w = o, i.e. u_w is a true outside vector

$$\frac{\partial J}{\partial u_w} = (\hat{y} - y)v_c$$

When $w \neq o$, the subtrahend is zero.

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

(d)

$$\begin{aligned}\frac{d}{dx}\sigma(x) &= \frac{d}{dx}\frac{1}{1+e^{-x}} \\ &= \frac{e^{-x}}{(1+e^{-x})^2} \\ &= \frac{1+e^{-x}-1}{(1+e^{-x})^2} \\ &= \frac{1}{1+e^{-x}}(1-\frac{1}{1+e^{-x}}) \\ &= \sigma(x)(1-\sigma(x))\end{aligned}$$

(e) K refers to the set of negative samples

$$J_{neg} = -\log(\sigma(u_o^\top v_c)) - \sum_{k=1}^{K} log(\sigma(-u_k^\top v_c))$$

The partial derivative w.r.t to v_c:

$$\frac{\partial J_{neg}}{\partial v_c} = -\sigma(u_o^\top v_c)u_o + \sum_{k=1}^{W}\sigma(u_k^\top v_c)u_k$$

The partial derivative w.r.t to the u_o and u_k, the vector that corresponds to the groundtruth label and negative samples, respectively:

$$\frac{\partial J_{neg}}{\partial u_o} = -\sigma(u_o^\top v_c)v_c$$

$$\frac{\partial J_{neg}}{\partial u_k} = \sum_{k=1}^{K}\sigma(u_k^\top v_c)v_c$$

(f) This is straightforward since the derivative is simply the sum of the derivative for each word.

$$\frac{\partial J_{skip-gram}}{\partial U} = \sum_j \frac{\partial J_{skip-gram}(v_c, w_{t+j}, U)}{\partial U}$$

$$\frac{\partial J_{skip-gram}}{\partial v_c} = \sum_j \frac{\partial J_{skip-gram}(v_c, w_{t+j}, U)}{\partial v_c}$$

$$\frac{\partial J_{skip-gram}}{\partial v_w} = 0$$