Stanford CS224n.
Assignment 1

Neural Networks:

2a. $\sigma'(n) = \sigma(n)(1-\sigma(n))$

2b. $\hat{y} = \text{softmax}(\theta)$ and $\frac{dCE}{d\theta}$. $CE = -\sum_i y_i \log(\hat{y}_i)$

assume $y_k = 1$

$$\frac{dCE}{d\theta} = \frac{dCE}{d\hat{y}} \cdot \frac{d\hat{y}}{d\theta}$$

$$\frac{dCE}{d\hat{y}} = \begin{bmatrix} \frac{dCE}{d\hat{y}_1} & \frac{dCE}{d\hat{y}_2} & \cdots & \frac{dCE}{d\hat{y}_n} \end{bmatrix}$$

$$\frac{d\hat{y}}{d\theta} = \begin{bmatrix} \frac{d\hat{y}_1}{d\theta_1} & \frac{d\hat{y}_1}{d\theta_2} & \cdots & \\ \frac{d\hat{y}_2}{d\theta_1} & \ddots & & \\ \vdots & & & \frac{d\hat{y}_n}{d\theta_n} \end{bmatrix}$$

Simplifying: $\frac{dCE}{d\hat{y}_i}$ when $i == k$: $CE = -\log(\hat{y}_k)$, $\frac{dCE}{d\hat{y}_k} = \frac{-1}{\hat{y}_k}$; else $\frac{dCE}{d\hat{y}_i} = 0$

$$\Rightarrow \frac{dCE}{d\hat{y}} = \begin{bmatrix} 0 & \cdots & \frac{-1}{\hat{y}_k} & \cdots & 0 \end{bmatrix}$$

$$\therefore \boxed{\frac{dCE}{d\theta} = \frac{dCE}{d\hat{y}} \cdot \frac{d\hat{y}}{d\theta} = \begin{bmatrix} \frac{d\hat{y}_k}{d\theta_1} & \frac{d\hat{y}_k}{d\theta_2} & \cdots & \frac{d\hat{y}_k}{d\theta_n} \end{bmatrix} \times \frac{-1}{\hat{y}_k}}$$

$$\hat{y}_k = \frac{e^{\theta_k}}{\sum_i e^{\theta_i}} \qquad \frac{d\hat{y}_k}{d\theta_k} = \frac{\sum_i e^{\theta_i} e^{\theta_k} - (e^{\theta_k})^2}{\left(\sum_i e^{\theta_i}\right)^2}$$

$$= \hat{y}_k(1-\hat{y}_k)$$

$$\frac{d\hat{y}}{d\theta_{i(i\neq k)}} = \frac{-e^{\theta_i} e^{\theta_k}}{\sum_j e^{\theta_j}} = -\hat{y}_i \hat{y}_k$$

$$\Rightarrow \frac{dCE}{d\theta_j} = \begin{bmatrix} \hat{y}_k - 1 & \text{if } k == j \\ \hat{y}_j & \text{else} \end{bmatrix}$$

2c. find $\frac{dCE}{dX} = \begin{bmatrix} \frac{dCE}{dx_1} & \cdots \\ \frac{dCE}{dx_2} \\ \end{bmatrix}$

$h = \text{sigmoid}(xW_1 + b_1)$

$\hat{y} = \text{softmax}(hW_2 + b_2)$

for a single sample x:

from (2b). $\frac{dCE}{d(hW_2+b_2)} = \begin{bmatrix} \hat{y}_k - 1 & \text{if } k == j \\ \hat{y}_j & \text{else} \end{bmatrix} = \hat{y} - y$

$$\frac{dCE}{dx} = \frac{dCE}{d(hW_2+b_2)} \cdot \frac{d(hW_2+b_2)}{dh} \cdot \frac{dh}{d(xW_1+b_1)} \cdot \frac{d(xW_1+b_1)}{dx}$$

$$\Rightarrow \frac{d(hW_2+b_2)}{dh} = W_2^T \Rightarrow \frac{d(xW_1+b_1)}{dx} = W_1^T$$

$$\frac{dh}{d(xW_1+b_1)} = \begin{bmatrix} \frac{dh}{dxW_{1C_1}+b_{11}} & \frac{dh}{dxW_{1C_2}+b_{12}} & \cdots \end{bmatrix}$$

$$= \begin{bmatrix} \frac{dh_1}{dxW_{1C_1}+b_{11}} & \frac{dh_1}{dxW_{1C_2}+b_{12}} \\ \frac{dh_2}{dxW_{1C_1}+b_{11}} & \\ \vdots & \end{bmatrix}$$

$h_i = \text{sigmoid}(xW_{1C_i} + b_{1i})$

$$\frac{dh_i}{dxW_{1Cj}+b_{ij}} = \begin{cases} h_i(1-h_i) & (i == j) \\ 0 & \text{else} \end{cases}$$

$$\frac{dh}{d(xW_1+b_1)} = \begin{bmatrix} h_1(1-h_1) & 0 & \cdots \\ 0 & h_2(1-h_2) & \\ \vdots & & \ddots \end{bmatrix}$$

$$\therefore \frac{dCE}{dx_i} = \left(\left(\frac{dCE}{do_i} \cdot w_2^T\right) * \left[h_i(1-h_i) \rightarrow\right]\right) \cdot w_1^T$$

$$\frac{dCE}{dX} = \left[\left(\frac{dCE}{d\textcircled{H}} \cdot w_2^T\right) * \left(H * (1-H)\right)\right] \cdot w_1^T$$

$*$ → element wise multiplication.

$\textcircled{H} = \begin{bmatrix} o_1 \\ o_2 \\ \vdots \end{bmatrix}$

29. Computations: find $\frac{dCE}{dw_1}, \frac{dCE}{dw_2}, \frac{dCE}{db_1}, \frac{dCE}{db_2}$

[Gradients for back prop]

1. $\frac{dCE}{dw_2} = \frac{dCE}{d(hw_2+b_2)} \cdot \frac{d(hw_2+b_2)}{dw_2}$

note that → hw_2+b_2 is a vector

→ w_2 is a matrix

$\Rightarrow \frac{d(hw_2+b_2)}{dw_2}$ => 3 dimensional.

$\therefore$ assume $w_2 = [c_1 \; c_2 \cdots]$ $c_i \Rightarrow i^{th}$ column of w_2

$$\frac{d(hw_2+b_1)}{dc_i} = \frac{d(hw_2)}{dc_i} = \left[\frac{dhw_2}{dc_{i1}} \quad \frac{dhw_2}{dc_{i2}} \cdots\right]$$

$$= \begin{bmatrix} \frac{dhc_1}{dc_{i1}} & \frac{dhc_1}{dc_{i2}} \cdots \\ \frac{dhc_2}{dc_{i1}} & \ddots \\ \vdots & \end{bmatrix}$$

(i^{th} row = h, rest = 0) $\Rightarrow$

$$= \begin{bmatrix} 0 \cdots \\ \leftarrow h \rightarrow \\ 0 \cdots \\ \vdots \end{bmatrix}$$

$$\therefore \frac{dCE}{dc_i} = \frac{dCE}{d(hw_2+b_2)} \cdot \frac{d(hw_2+b_2)}{dc_i} = (\hat{y}_i - y_i) * h. \quad \cdots (a)$$

$$\frac{dCE}{dw_2} = \left[\left(\frac{dCE}{dc_1}\right)^T \left(\frac{dCE}{dc_2}\right)^T \cdots \right]$$

$$= \left[(\hat{y}_1 - y_1) h^T \quad (\hat{y}_2 - y_2) h^T \cdots \right]$$

$$= h^T \cdot (\hat{y} - y)$$

$$\frac{dCE}{db_2} = \frac{dCE}{d(hw_2+b_2)} \cdot \frac{d(hw_2+b_2)}{db_2}$$

$$= \frac{dCE}{d(hw_2+b_2)} \cdot I$$

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

$$\frac{dCE}{dw_1} = \frac{dCE}{d(xw_1+b_1)} \cdot \frac{d(xw_1+b_1)}{dw_1}$$

again, $w_1 = [c_1 \cdots c_n]$

$$\frac{d(xw_1+b_1)}{dw_1} = \left[\left(\frac{d(xw_1+b_1)}{dc_1}\right)^T \left(\frac{d(xw_1+b_1)}{dc_2}\right)^T \cdots \right]$$

$$\frac{d(xw_1+b_1)}{dc_i} = \left[\frac{d(xw_1+b_1)}{dc_{i1}} \quad \frac{d(xw_1+b_1)}{dc_{i2}} \cdots \right]$$

$$= \begin{bmatrix} \frac{dxc_1}{dc_{i1}} & \frac{dxc_1}{dc_{i2}} \cdots \\ \frac{dxc_2}{dc_{i1}} & \ddots \\ \vdots & \end{bmatrix}$$

$$\left(\begin{matrix} i^{th} \text{ row} = x \\ \text{rest} = 0 \end{matrix} \right) \Rightarrow \begin{bmatrix} \vdots & \\ 0 \cdots & \\ \leftarrow x \rightarrow & \\ 0 \cdots & \\ \vdots & \end{bmatrix}$$

from (2c) $\frac{dCE}{d(xw_1+b_1)} = \left(\frac{dCE}{do} \cdot w_2^T\right) * \left[h_i(1-h_i)\right]$

Note that this is very similar to case (a),

$$\frac{dCE}{dw_1} = x^T \cdot \left[\left(\frac{dCE}{do} \cdot w_2^T\right) * \left[h_i(1-h_i)\right]\right]$$

$$\frac{dCE}{db_1} = \frac{dCE}{d(xw_1+b_1)} \cdot \frac{d(xw_1+b_1)}{db_1}$$

$$= \frac{dCE}{d(xw_1+b_1)} \; I = \left(\frac{dCE}{do} \cdot w_2^T\right) * \left[h_i(1-h_i)\right]$$

Note that these are derivatives wrt a single sample.

for X (all samples):

$$\frac{dCE}{dw_2} = \left[\frac{dCE}{d(h_1w_2+b_2)} \quad \frac{dCE}{d(h_2w_2+b_2)} \cdots\right] \begin{bmatrix} \frac{d(h_1w_2+b_2)}{dw_2} \\ \frac{d(h_2w_2+b_2}{dw_2} \\ \vdots \end{bmatrix}$$

$= \sum_i h_i^T (\hat{y}_i - y_i)$ (simply summation of all matrices)

$= H^T \cdot (\hat{y} - y)$

similarly:

$$\frac{dCE}{dw_1} = X^T \cdot \left[\left(\frac{dCE}{d\Theta} \cdot w_2^T\right) * \left[H(1-H)\right]\right]$$

$$\frac{dCE}{db_2} = \sum_i (\hat{y} - y)_i$$

$$\frac{dCE}{db_1} = \sum_i \left[\left(\frac{dCE}{d\theta_i} \cdot w_2^T\right) * \left[h_i(1-h_i)\right]\right]$$

3. word2vec. (Softmax)

o → output
c → current

3a. $\hat{y}_o = p(o|c) = \frac{\exp(u_o^T v_c)}{\sum_{w=1}^{V} \exp(u_w^T v_c)}$

$U = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}$

$CE = -\log \hat{y}_o \quad | \; y \rightarrow$ one-hot vector

$y, \hat{y} \rightarrow$ row. vec.

$$\frac{dCE}{dv_c} = \frac{dCE}{d\hat{y}_o} \cdot \frac{d\hat{y}_o}{dv_c}$$

$$\frac{dCE}{d\hat{y}_o} = \frac{-1}{\hat{y}_o}$$

$$\frac{d\hat{y}_o}{dv_c} = \left[\frac{d\hat{y}_o}{dv_{c_1}} \quad \frac{d\hat{y}_o}{dv_{c_2}} \quad \cdots \right]$$

$$\frac{d\hat{y}_o}{dv_{ci}} = \frac{d \dfrac{e^{u_o^T v_c}}{\sum_{w=1}^{V} e^{u_w^T v_c}}}{dv_{ci}}$$

$$= \frac{\left(\sum_{w=1}^{V} e^{u_w^T v_c}\right) \cdot e^{u_o^T v_c} \cdot u_{oi} - e^{u_o^T v_c} \cdot \sum_{w=1}^{V} u_{wi} e^{u_w^T v_c}}{\left(\sum_{w=1}^{V} e^{u_w^T v_c}\right)^2}$$

$$= \hat{y}_o \cdot u_{oi} - \hat{y}_o \cdot \sum_{w=1}^{V} u_{wi} \hat{y}_i$$

$$= \hat{y}_o \left[u_{oi} - \hat{y} \cdot U_{ci} \right]$$

$$= \hat{y}_o \left[y \cdot U_{ci} - \hat{y} \cdot U_{ci} \right]$$

$$\frac{d\hat{y}_0}{dv_{ci}} = \hat{y}_0\left[y \cdot v_{ci} - \hat{y} \cdot v_{ci}\right] = \hat{y}_0 \cdot [y - \hat{y}]\, v_{ci}$$

$$\frac{d\hat{y}_0}{dv_c} = [y - \hat{y}] \cdot U \cdot \hat{y}_0$$

$$\frac{dCE}{dv_c} = \frac{d\hat{y}_0}{dv_c} \times -\frac{1}{\hat{y}_0} = [\hat{y} - y] \cdot U$$

3b. $$\frac{dCE}{dU} = v_c^T \cdot (\hat{y} - y)$$

$$\frac{d\hat{y}_0}{dU_{0i}} = \left(\sum_{w=1}^{V} e^{u_w^T v_c}\right) \cdot e^{u_0^T v_c} \cdot v_{ci} -$$

$$\frac{e^{u_0^T v_c} \cdot e^{u_0^T v_c} \cdot v_{ci}}{\left(\sum_{w=1}^{V} e^{u_w^T v_c}\right)^2}$$

$$= \hat{y}_0 \cdot v_{ci} - (\hat{y}_0)^2 v_{ci}$$

$$= \hat{y}_0 (v_{ci} - \hat{y}_0 v_{ci})$$

$$= \hat{y}_0 (1 - \hat{y}_0) \cdot v_{ci}$$

$$\frac{dCE}{dU_{0i}} = (\hat{y}_0 - 1) v_{ci}$$

$$\boxed{\frac{dCE}{dU_0} = (\hat{y}_0 - 1) v_c} = (\hat{y}_0 - y_0) v_c$$

$$\frac{dCE}{dU_{wi}} \quad w \neq 0 = \frac{\left(\sum_{w=1}^{V} e^{u_w^T v_c}\right) \cdot 0 - e^{u_0^T v_c} \cdot e^{u_w^T v_c} \cdot v_{ci}}{(\sum \cdot)^2 \times -\hat{y}_0}$$

$$= \hat{y}_w \cdot v_{ci}$$

$$\boxed{\frac{dCE}{dU_w} = \hat{y}_w v_c} \Rightarrow (\hat{y}_w - y_w) v_c$$

3c (negative sampling).

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

find: $\frac{\partial J}{\partial v_c}$ $\frac{\partial J}{\partial u_o}$ $\frac{\partial J}{\partial u_i}$ $i \neq o$

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

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

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

$$\frac{\partial J}{\partial u_{i \neq o}} = -(\sigma(-u_o^T v_c) - 1) v_c$$