

# Stanford CS224n

## Assignment 1

Neural Networks:

$$2a. \hat{q}'(n) = \sigma(n)(1-\sigma(n))$$

$$2b. \hat{q} = \text{softmax}(o) \quad \text{and} \quad \frac{dCE}{d\theta}, \quad CE = -\sum_i q_i \log(\hat{q}_i)$$

assume  $q_k = 1$

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

$$\frac{dCE}{d\hat{q}} = \left[ \frac{dCE}{d\hat{q}_1}, \frac{dCE}{d\hat{q}_2}, \dots, \frac{dCE}{d\hat{q}_n} \right]$$

$$\frac{d\hat{q}}{d\theta} = \begin{bmatrix} \frac{d\hat{q}_1}{d\theta_1} & \frac{d\hat{q}_1}{d\theta_2} & \dots \\ \vdots & \vdots & \vdots \\ \frac{d\hat{q}_n}{d\theta_1} & \frac{d\hat{q}_n}{d\theta_2} & \dots \\ \frac{d\hat{q}_n}{d\theta_n} & & \end{bmatrix}$$

Simplifying  $\frac{dCE}{d\hat{q}_i}$  when  $i=k$  else

$$CE = -\log(\hat{q}_k) \quad \frac{dCE}{d\hat{q}_i} = 0$$

$$\frac{dCE}{d\hat{q}_k} = -\frac{1}{\hat{q}_k}$$

$$\Rightarrow \frac{dCE}{d\hat{q}} = \left[ 0 \dots -\frac{1}{\hat{q}_k} \dots 0 \right]$$

$$\therefore \frac{dCE}{d\theta} = \frac{dCE}{d\hat{q}} \cdot \frac{d\hat{q}}{d\theta} = \left[ \frac{d\hat{q}_k}{d\theta_1}, \frac{d\hat{q}_k}{d\theta_2}, \dots, \frac{d\hat{q}_k}{d\theta_n} \right] \times -\frac{1}{\hat{q}_k}$$

$$\hat{q}_k = \frac{e^{\theta_k}}{\sum_i e^{\theta_i}}$$

$$\frac{d\hat{q}_k}{d\theta_k} = \frac{\sum_i e^{\theta_i} e^{\theta_k} - (e^{\theta_k})^2}{(\sum_i e^{\theta_i})^2}$$

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

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

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

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

$$h = \text{sigmoid}(xw_1 + b_1)$$

$$\hat{y} = \text{softmax}(hw_2 + b_2)$$

for a single sample  $x$ :

from ②b.  $\frac{dCE}{d(hw_2 + b_2)} = \begin{bmatrix} \hat{y}_{k-1} & \text{if } k=j \\ \hat{y}_j & \text{else} \end{bmatrix} = \hat{y} - y$

$$\frac{dCE}{dx} = \frac{dCE}{d(hw_2 + b_2)} \cdot \frac{d(hw_2 + b_2)}{dh} \cdot \frac{dh}{d(xw_1 + b_1)} \cdot \frac{d(xw_1 + b_1)}{dx}$$

$$\Rightarrow \frac{d(hw_2 + b_2)}{dh} = w_2^T \Rightarrow \frac{d(xw_1 + b_1)}{dx} = w_1^T$$

$$\frac{dh}{d(xw_1 + b_1)} = \begin{bmatrix} \frac{dh}{dxw_{11} + b_{11}} & \frac{dh}{dxw_{12} + b_{12}} & \dots \end{bmatrix}$$

$$= \begin{bmatrix} \frac{dh_1}{dxw_{11} + b_{11}} & \frac{dh_1}{dxw_{12} + b_{12}} \\ \frac{dh_2}{dxw_{11} + b_{11}} & \ddots \\ \vdots & \ddots \end{bmatrix}$$

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

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

$$\frac{dh}{d(xw_1 + b_1)} = \begin{bmatrix} h_1(1-h_1) & 0 & \dots \\ 0 & h_2(1-h_2) & \dots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

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

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

$\rightarrow$  element wise multiplication.

$$\Theta = \begin{bmatrix} \Theta_1 \\ \Theta_2 \\ \vdots \end{bmatrix}$$

29. Computations: find  $\frac{dCE}{dw_1}, \frac{dCE}{dw_2}, \frac{dCE}{db_1}, \frac{dCE}{db_2}$   
[gradient 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  $\rightarrow hw_2 + b_2$  is a vector

$\rightarrow w_2$  is a matrix

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

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

$$\frac{d(hw_2 + b_2)}{dc_i} = \frac{d(hw_2)}{dc_i} = \left[ \frac{dhw_2}{dc_{i1}} \quad \frac{dhw_2}{dc_{i2}} \dots \right]$$

$$= \left[ \frac{dhc_1}{dc_{i1}} \quad \frac{dhc_1}{dc_{i2}} \dots \right]$$

$$\left( \begin{array}{l} i^{\text{th}} \text{ row} = h \\ \text{rest} = 0 \end{array} \right) \Rightarrow \left[ \begin{array}{c} \frac{dhc_1}{dc_{i1}} \quad \frac{dhc_1}{dc_{i2}} \dots \\ \vdots \\ \frac{dhc_2}{dc_{i1}} \quad \ddots \\ \vdots \\ \vdots \end{array} \right]$$

$$= \left[ \begin{array}{c} \vdots \dots \\ \leftarrow h \rightarrow \\ \vdots \dots \end{array} \right]$$

$$\therefore \frac{dCE}{dc_i} = \frac{dCE}{d(hw_2 + b_2)} \cdot \frac{d(hw_2 + b_2)}{dc_i} = (\hat{y}_i - y_i) * h_i. \dots @$$

$$\begin{aligned}\frac{dCE}{dw_2} &= \left[ \left( \frac{dCE}{dc_1} \right)^T \left( \frac{dCE}{dc_2} \right)^T \dots \right] \\ &= \left[ (\hat{y}_1 - y)^T (\hat{y}_2 - y)^T \dots \right] \\ &= h^T \cdot (\hat{y} - y)\end{aligned}$$

$$\begin{aligned}\frac{dCE}{db_2} &= \frac{dCE}{d(hw_2 + b_2)} \cdot \frac{d(hw_2 + b_2)}{db_2} \\ &= \frac{dCE}{d(hw_2 + b_2)} \cdot I \\ &= (\hat{y} - y)\end{aligned}$$

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

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

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

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

$$= \left[ \begin{array}{cc} \frac{d(xc_1)}{dc_{i1}} & \frac{d(xc_1)}{dc_{i2}} \dots \\ \frac{d(xc_2)}{dc_{i1}} & \ddots \\ \vdots & \end{array} \right]$$

$$= \left[ \begin{array}{c} \vdots \\ \ddots \\ \xrightarrow{x} \\ \ddots \\ \vdots \end{array} \right]$$

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

$$\text{From (24)}: \frac{dCE}{d(xw_1 + b_1)} = \left( \frac{dCE}{d\theta} \cdot w_2^T \right) * [h_i(1-h_i)]$$

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

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

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

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

Note that these are derivatives wrt a single sample.

for  $X$  (all samples):

$$\frac{dCE}{dw_2} = \begin{bmatrix} \frac{dCE}{d(h_1 w_2 + b_2)} & \frac{dCE}{d(h_2 w_2 + b_2)} & \dots \end{bmatrix} \begin{bmatrix} \frac{dh_1 w_2 + b_2}{dw_2} \\ \frac{dh_2 w_2 + b_2}{dw_2} \\ \vdots \end{bmatrix}$$

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

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

similarly:

$$\frac{dCE}{db_1} = X^T \cdot \left[ \left( \frac{dCE}{d\theta} \cdot w_2^T \right) * [H(1-H)] \right]$$

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

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

### 3. Word2Vec. (Softmax)

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

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

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

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

$$\frac{d\hat{q}_o}{dv_c} = \left[ \frac{d\hat{q}_o}{dv_{c1}} \quad \frac{d\hat{q}_o}{dv_{c2}} \quad \dots \right]$$

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

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

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

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

$$= \hat{q}_o \left[ q \cdot v_{ci} - \hat{q}_o \cdot v_{ci} \right]$$

$\circ \rightarrow \text{output}$   
 $c \rightarrow \text{current}$

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

$4, 9 \rightarrow \text{row, Vec}$

$$\frac{d\hat{q}_0}{dv_{ci}} = \hat{q}_0 [q \cdot v_{ci} - \hat{q} \cdot v_{ci}] = \hat{q}_0 [q - \hat{q}] \cdot v_{ci}$$

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

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

$$3b. \quad \frac{dCE}{dU} = v_c \cdot (\hat{q} - q)$$

$$\frac{d\hat{q}_0}{dU_{ci}} = \left( \sum_{w=1}^V e^{U_w v_c} \right) \cdot e^{U_0 v_c} \cdot v_{ci} - \frac{e^{U_0 v_c} - e^{U_0 v_c} \cdot v_{ci}}{\left( \sum_{w=1}^V e^{U_w v_c} \right)^2}$$

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

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

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

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

$$\boxed{\frac{dCE}{dU_0} = (\hat{q}_0 - 1) v_c} = (\hat{q}_0 - q_0) v_c$$

$$\frac{dCE}{dU_{wi}} \stackrel{w \neq 0}{=} \frac{\left( \sum_{w=1}^V e^{U_w v_c} \right) \cdot 0 - e^{U_0 v_c} \cdot e^{U_w v_c} \cdot v_{ci}}{(\sum \cdot)^2 \times -\hat{q}_0}$$

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

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

3c (negative sampling).

$$d(\sigma, v_c, u) = -\log(\sigma(u_0^T v_c)) - \sum_{k=1}^K \log(\sigma(-u_k^T v_c))$$

find:  $\frac{dd}{dv_c}$   $\frac{dd}{du_0}$   $\frac{dd}{du_i \neq 0}$

$$\frac{dd}{dv_c} = \frac{-1 \cdot \sigma'(1-\sigma)}{\cancel{\sigma(u_0^T v_c)}} \cdot u_0 - \sum_{k=1}^K \frac{-1 \cdot \sigma'(1-\sigma)}{\cancel{\sigma(-u_k^T v_c)}} \cdot u_k$$

$$\boxed{\frac{dd}{dv_c} = (\sigma(u_0^T v_c) - 1) u_0 - \sum_{k=1}^K (\sigma(u_k^T v_c) - 1) u_k}$$

$$\frac{dd}{du_0} = (\sigma(u_0^T v_c) - 1) v_c$$

$$\frac{dd}{du_i \neq 0} = -(\sigma(-u_0^T v_c) - 1) v_c$$

