Assignment #1

1. Softmax

(a) for all dimensions $i \in [1, \dim(x)]$

$$\text{softmax}(x+c)_i = \frac{e^{(x_i+c)}}{\sum_j e^{(x_j+c)}}$$

$$= \frac{e^{x_i} \cdot e^{c}}{e^{c} \sum_j e^{x_j}}$$

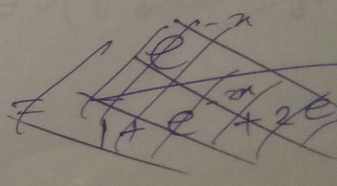
$$= \frac{e^{x_i}}{\sum_j e^{x_j}}$$

$$= \text{softmax}(x)_i$$

2. Neural Network Basics

(a) $$\sigma(x) = \frac{1}{1+e^{-x}}$$

$$\sigma'(x) = \frac{-(-e^{-x})}{(1+e^{-x})^2}$$



$$= \frac{e^{-x}}{(1+e^{-x})} \cdot \frac{1}{(1+e^{-x})}$$

$$= \sigma(x)(1-\sigma(x))$$

(b)

$$CE(y,\hat{y}) = -\sum_i y_i \log(\hat{y}_i)$$

$$= -\sum_i y_i \log\left(\frac{e^{\theta_i}}{\sum_j e^{\theta_j}}\right)$$

Let k^{th} element of $y = 1$ & all others are zero

$$\frac{\partial CE(y,\hat{y})}{\partial \theta_l} = -y_l \frac{\partial}{\partial \theta_l}\left(\log(e^{\theta_l}) - \log\left(\sum_i e^{\theta_i}\right)\right) - \frac{\partial}{\partial \theta_l}\sum_{i \neq l} y_i \left(\log(e^{\theta_i}) - \log\left(\sum_j e^{\theta_j}\right)\right)$$

$$= -y_l\left(1 - \frac{1}{\sum_i e^{\theta_i}} e^{\theta_l}\right) + \sum_{i \neq l} \frac{y_i \cdot e^{\theta_l}}{\sum_j e^{\theta_j}}$$

$$= -y_l(1-\hat{y}_l) + \hat{y}_l \sum_{i \neq l} y_i$$

$$= y_l(\hat{y}_l - 1) = \begin{cases} \hat{y}_l - y_l & \text{if } l = * \\ \hat{y}_l - \end{cases}$$

$$\nabla_\theta CE(y,\hat{y}) = \cup$$

if $l = k$,

$$\frac{\partial CE(y,\hat{y})}{\partial \theta_l} = -(1-\hat{y}_l) = \hat{y}_l - 1$$

if $l \neq k$,

$$\frac{\partial}{\partial \theta_l} CE(y,\hat{y}) = \hat{y}_l$$

So,

$$\boxed{\frac{\partial CE(y,\hat{y})}{\partial \theta} = \hat{y} - y}$$

(c) $$J = -\sum_i y_i \log(\hat{y}_i)$$

Let $z_2 = h W_2 + b_2$

$z_1 = x W_1 + b_1$

then

$h = \text{sigmoid}(z_1) = \sigma(z_1)$

$\hat{y} = \text{softmax}(z_2)$

$$\frac{\partial J}{\partial x_j} = \frac{\partial J}{\partial z_2} \cdot \frac{\partial z_2}{\partial x_j}$$

$$= (\hat{y} - y) \cdot \frac{\partial z_2}{\partial x_j}$$

$$\delta_1 = \hat{y} - y$$

$$\frac{\partial z_2}{\partial x_j} = \sum_k \frac{\partial z_2}{\partial h_k} \cdot \frac{\partial h_k}{\partial x_j}$$

$$= \sum_k W_{2_{\cdot k}} \cdot \sigma'(z_{1_k}) \cdot W_{1_{jk}}$$

$$W_2 \cdot \left(\sigma'(z_1) \circ W_{1_{j\cdot}}^T\right)$$

$$\frac{\partial J}{\partial x_j} = \sum_k \delta_1 \cdot W_{2_{\cdot k}}^T \sigma'(z_{1_k}) W_{1_{jk}}$$

$$= \sum_k \delta_1 W_{2_{\cdot k}}^T \sigma'(z_{1_k}) W_{1_{jk}}$$

$$\delta_{2_k}$$

$$\delta_2 = \delta_1 W_2^T$$

$$\frac{\partial J}{\partial x_j} = \sum_k \delta_{2_k} \sigma'(z_{1_k}) W_{1_{jk}} = \sum_k \delta_{3_k} W_{1_{jk}}$$

$$\delta_{3_k}$$

$$\frac{\partial J}{\partial x} = \delta_3 W_1^T$$

(d) $x \in \mathbb{R}^{D_x}$

$\hat{y} \in \mathbb{R}^{D_y}$

H hidden units.

Parameters $\Rightarrow$ W_1, W_2, b_1, b_2

$W_1 \in \mathbb{R}^{H \times D_x}$

$b_1 \in \mathbb{R}^{H}$

$W_2 \in \mathbb{R}^{H \times D_y}$

$b_2 \in \mathbb{R}^{D_y}$

$$\# \text{ Parameters} = H(D_x + 1) + D_y(H + 1)$$

3. word2vec

(a) $$J(\theta) = CE(y, \hat{y})$$

$$= -\sum_{w=1}^{W} y_w \log(\hat{y}_w)$$

$\because$ o is the expected word.

$\therefore$ $y_o = 1$ & $y_i = 0 \quad \forall\, i \neq o$

$$J(\theta) = -\log(\hat{y}_o)$$

$$= -\log\left(\frac{\exp(u_o^T v_c)}{\sum_{w=1}^{W}\exp(u_w^T v_c)}\right)$$

$$= -u_o^T v_c + \log\sum_{w=1}^{W}\exp(u_w^T v_c)$$

$$\frac{\partial J}{\partial v_c} = -u_o + \frac{\sum_{w=1}^{W}\exp(u_w^T v_c)\,u_w}{\sum_{i=1}^{W}\exp(u_i^T v_c)}$$

$$\boxed{\frac{\partial J}{\partial v_c} = -u_o + \sum_{w=1}^{W}\hat{y}_w u_w}$$

Let

$$U = \begin{bmatrix} | & | & & | \\ u_1 & u_2 & \cdots & u_W \\ | & | & & | \end{bmatrix}$$

then

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

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

(b) from (a)

$$J(\theta) = -u_o^T v_c + \log \sum_{w=1}^{W} \exp(u_w^T v_c)$$

$$\frac{\partial J}{\partial u_i} = \mathbb{1}_{\{i=o\}}(-v_c) + \frac{\exp(u_o^T v_c)\, v_c}{\sum_{j=1}^{W} \exp(u_j^T v_o)}$$

$$= \mathbb{1}_{\{i=o\}}(-v_c) + \hat{y}_i v_c$$

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

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

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

(c)

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

$$\frac{\partial J}{\partial v_c} = -\frac{\sigma'(u_o^T v_c)\, u_o}{\sigma(u_o^T v_c)} - \sum_{k=1}^{K} \frac{d}{\partial v_c} \log(\sigma(-u_k^T v_c))$$

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

$\because \sigma(-x) = 1 - \sigma(x)$

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

$$\frac{\sigma(u_o^T v_c)\,(1-\sigma(u_o^T v_c))}{\sigma(u_o^T v_c)}\, u_o$$

$$= \frac{\sigma(u_o^T v_c)\,(1-\sigma(u_o^T v_c))}{\log\left(\frac{1}{1+\exp(-u_o^T v_c)}\right)}\, u_o$$

$$= \frac{\sigma(u_o^T v_c)(1-\sigma(u_o^T v_c))}{-\log(1+\exp(-u_o^T v_c))}\, u_o$$

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

$$\frac{\partial J}{\partial u_i} = \mathbb{1}_{\{i=o\}} \left[-\frac{\sigma(u_o^T v_c)\,(1-\sigma(u_o^T v_c))}{\sigma(u_o^T v_c)}\, v_c \right]$$

$$- \sum_{k=1}^{K} \mathbb{1}_{\{k=i\}} \left[\frac{\sigma(-u_k^T v_c)\,(1-\sigma(-u_k^T v_c))}{\sigma(-u_k^T v_c)} (-v_c) \right]$$

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

$$\frac{\partial J}{\partial u_k} = -\left(\sigma(-u_k^T v_c) - 1\right) v_c \qquad \forall\, k = 1, 2, \ldots, K$$

$$\frac{\partial J}{\partial u_i} = 0 \qquad \forall\, i \notin \{1, 2, \ldots, k\} \cup \{o\}$$

The negative sampling loss function requires only K (≈ 100) training examples to compute while softmax-CE loss function requires all training examples (words) which is of the order of millions. i.e.

$$\text{speed-up ratio} \cong \frac{W}{K}$$

(d) for skip-gram

$$J_{\text{skip-gram}}(\text{word}_{c-m \ldots c+m}) = \sum_{\substack{-m \le j \le m, \\ j \neq 0}} f(w_{c+j}, v_c)$$

$$= \sum_{\substack{-m \le j \le m, \\ j \neq 0}} J_{\text{softmax-CE}}(o, v_c, U)$$

From previous parts, we know to compute the gradients $\frac{\partial F(w_i, v)}{\partial U}$ and $\frac{\partial f(w_i, v)}{\partial v}$.

$$\frac{\partial J_{skip\text{-}gram}(word_{c-m,\dots c+m})}{\partial U} = \sum_{\substack{-m \le j \le m, \\ j \neq 0}} \frac{\partial f(w_{c+j}, v_c)}{\partial U}$$

$$\frac{\partial J_{skip\text{-}gram}(word_{c-m,\dots c+m})}{\partial v_c} = \sum_{\substack{-m \le j \le m, \\ j \neq 0}} \frac{\partial f(w_{c+j}, v_c)}{\partial v_c}$$

$$\frac{\partial J_{skip\text{-}gram}(word_{c-m,\dots c+m})}{\partial v_j} = 0 \quad \forall\ j \neq c$$

For CBOW

$$\frac{\partial J_{CBOW}(word_{c-m,\dots c+m})}{\partial U} = \frac{\partial f(w_c, \hat{v})}{\partial U}$$

$$\frac{\partial J_{CBOW}(word_{c-m,\dots c+m})}{\partial v_j} = \frac{\partial f(w_c, \hat{v})}{\partial \hat{v}}, \quad \forall\ j \in \{c-m, \dots, c-1, c+1, \dots, c+m\}$$

$$\frac{\partial J_{CBOW}(word_{c-m,\dots c+m})}{\partial v_j} = 0, \quad \forall\ j \notin \{c-m, \dots, c-1, c+1, \dots, c+m\}$$

4. Sentiment Analysis

(b) To avoid overfitting to the training data and make the model generalize well on unseen examples.

(c) To chose best regularization parameter, we tried the regularization parameter values from 0.0, 0.00001, 0.00003, 0.0001, 0.0003, 0.001, 0.003, 0.01 and choose the one that maximizes the dev set accuracy.

This strategy gives the best regularization parameter with value of 1e-4 with train, dev and test accuracies of 29.12, 31.43 and 27.56.