Assignment I

Part I, Softmax:

a) $\text{softmax}(x_i+c) = \dfrac{e^{x_i+c}}{\sum_j e^{x_j+c}} = \dfrac{e^{x_i}e^c}{\sum_j e^{x_j}e^c} = \dfrac{\cancel{e^c}}{\cancel{e^c}} \dfrac{e^{x_i}}{\sum_j e^{x_j}}$, so

$\text{softmax}(x_i+c) = \text{softmax}(x_i)$ ✓

b)

Part II, NN Basics

a) $\sigma(x) = \dfrac{1}{1+e^{-x}}$ $\quad \sigma'(x) = \dfrac{0-(1+e^{-x})'}{(1+e^{-x})^2} = \dfrac{e^{-x}}{(1+e^{-x})^2} = \dfrac{(1+e^{-x})-1}{(1+e^{-x})^2} =$

$= \sigma(x) - \sigma^2(x) = \sigma(x)(1-\sigma(x))$ ✓

b) $CE(y,\hat{y}) = -\sum_i y_i \log(\hat{y}_i)$ $\quad y_k = 1$, $y_{-k} = 0$, so

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

$CE(y,\hat{y}) = -\log(\hat{y}_k)$, $\hat{y}_k = \dfrac{e^{\theta_k}}{\sum_i e^{\theta_i}}$

$\dfrac{\partial CE(y,\hat{y})}{\partial\theta} = \dfrac{\partial}{\partial\theta}\left(-[\log(e^{\theta_k}) - \log(\sum_i e^{\theta_i})]\right) =$

$= \dfrac{\partial}{\partial\theta}\left(-\theta_k + \log(\sum_i e^{\theta_i})\right) = \dfrac{-\partial\theta_k}{\partial\theta} + \dfrac{\partial}{\partial\theta}\left(\log \sum_i e^{\theta_i}\right) =$

$= \dfrac{-\partial\theta_k}{\partial\theta} + \dfrac{1}{\sum_i e^{\theta_i}} \sum_i \dfrac{\partial e^{\theta_i}}{\partial\theta}$

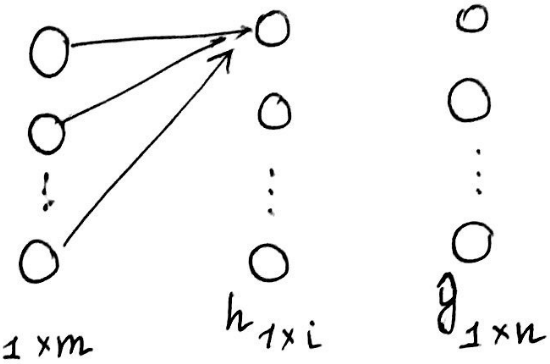
1) $i = k$ $\quad \dfrac{-\partial(\theta_k)}{\partial\theta_k} = -1$ and $\dfrac{1}{\sum_i e^{\theta_i}} \sum_i \dfrac{\partial e^{\theta_i}}{\partial\theta_k} = \dfrac{e^{\theta_k}}{\sum_i e^{\theta_i}} = \hat{y}_k$, so

$\dfrac{\partial CE(y,\hat{y})}{\partial\theta} = \hat{y}_i - 1.$

2) $i \neq k$, $i = j$ $\quad \dfrac{\partial\theta_k}{\partial_j} = 0$, $\dfrac{1}{\sum_i e^{\theta_i}} \sum_i \dfrac{\partial e^{\theta_i}}{\partial\theta_j} = \dfrac{e^{\theta_j}}{\sum_i e^{\theta_i}} = \hat{y}_j$

$\dfrac{\partial CE(y,\hat{y})}{\partial\theta} = \hat{y}_i$

c



$W_1\,(m\times i)$ $\quad$ $W_2\,(i\times n)$

$b_1\,(1\times i)$

$z_1 = xW_1 + b_1 = \quad z_1 \in \mathbb{R}^{1\times i}$ (2)

$= [x_1 w_{11} + x_2 w_{21} + \dots x_m w_{m1} \;—\; \dots x_1 w_{1i} + x_2 w_{2i} + \dots + x_m w_{mi}]_{1\times i}$

$a_1 = \sigma(z_1) = [\sigma(z_{11}) \;\; \sigma(z_{12}) \dots \sigma(z_{1i})]$

$a_1 \in \mathbb{R}^{1\times i}$

$z_2 = a_1 W_2 + b_2 \quad z_2 \in \mathbb{R}^{1\times n}$

$z_2 = \cancel{[a_1 w_{11} + a_2 w_{21} + \dots a_k w_{k1} \;—\; a_1 w_{1n} + a_2 w_{2n} + \dots a_k w_{kn}}$

$z_2 = [a_1 w_{11} + \dots a_i w_{i1} \;—\; a_1 w_{1n} + \dots a_i w_{in}]$

$a_2 = \sigma(z_2) \qquad a_2 \in \mathbb{R}^{1\times n}$

$J = -\log(a_k^{(2)}) \qquad \frac{\partial J}{\partial x} = \frac{\partial J}{\partial z^{(2)}} \cdot \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial x}$

$\delta_1 = \frac{\partial J}{\partial z^{(2)}} = \begin{cases} a_i^{(2)} - 1 & i = k \\ a_i^{(2)} - 0 & i \neq k \end{cases} = \underline{\underline{a^{(2)} - y}} \in \mathbb{R}^{1\times n}$

$\delta_2 = \frac{\partial J}{\partial a^{(1)}} = \underset{1\times n}{\delta_1} \cdot \underset{n\times i}{W_2^T} \quad \in \mathbb{R}^{1\times i}$

$\delta_3 = \frac{\partial J}{\partial z^{(1)}} = \delta_2 \circ \sigma'(z^{(1)}) \quad \in \mathbb{R}^{1\times i}$

$\delta_4 = \underset{1\times i}{\delta_3} \cdot \underset{\substack{\cancel{m\times i} \\ i\times m}}{W_1^T} \qquad \in \mathbb{R}^{1\times m}$

—

Scanned with CamScanner

d)

$X =$	$1 \times m$	$1 \times D_x$
$W_1 \in$	$m \times i$	$D_x \cdot H$
$b_1 \in$	$1 \times i$	$1 \times H$
$W_2 \in$	$i \times n$	$H \times D_y$
$b_2 \in$	$1 \times n$	$1 \times D_y$

$+ D_x \cdot H + H + H \cdot D_y + D_y$ ③

$\underline{(D_x + 1) \cdot H + (H+1) D_y}$ //

e), f) g) progr. $\qquad$ m # samples

$\frac{\partial J}{\partial W_1} \quad \frac{\partial J}{\partial W_2} \quad \frac{\partial J}{\partial b_1} \quad \frac{\partial J}{\partial b_2}$

① $\frac{\partial J}{\partial W_1} = \frac{\partial J}{\partial z^{(2)}} \frac{\partial z^{(2)}}{\partial a^{(1)}} \frac{\partial a^{(1)}}{\partial z^{(1)}} \frac{\partial z^{(1)}}{\partial W_1}$

$\delta_1 = a^{(2)} - y \qquad \delta_2 = \delta_1 \cdot W_2^T \in R^{1 \times i} \qquad \underbrace{\delta_3 = \delta_2 \, \sigma'(z^{(1)})}_{1 \times i} \qquad \delta_4 = \cancel{\underset{1\times i}{\delta_3} \cdot \underset{1 \times m}{X}} = \frac{\partial J}{\partial W_1}$

$= \underset{m \times 1}{X^T} \underset{1 \times i}{\delta_3} = m \times i \in R^{m \times i}$

② $\frac{\partial J}{\partial b_1} = \sum \delta_3 \quad \frac{1}{m} \sum_{i=1}^{m} \delta_3^{\,i} \in R^{1 \times i}$

③ $\frac{\partial J}{\partial W_2} = \frac{\partial J}{\partial z^{(2)}} \frac{\partial z^{(2)}}{W_2} = \cancel{\underset{1 \times n}{\delta_1} \cdot \underset{1 \times i}{a_1}}$ // $\underset{i \times 1}{a_1^T} \underset{1 \times n}{\delta_1} \quad \underline{i \times n}$ // $\in R^{i \times n}$

④ $\frac{\partial J}{\partial b_2} = \sum \delta_{1 i} \quad \frac{1}{m} \sum_{i=1}^{m} \delta_1^{\,i} \quad \in R^{1 \times n}$

Word2vec.

a)

$\hat{y}_o = p(o|c) \quad J = -\sum_i y_i \log(\hat{y}_i) = -\log \hat{y}_K$

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

$$+ \frac{\partial}{\partial v_c} \log\left(\sum_{w=1}^{W} \exp(u_w^T v_c)\right)^{\mathrm{II}}$$

I $\quad -u_o$

II $$\frac{1}{\sum_{w=1}^{W} \exp(u_w^T v_c)} \cdot \sum_{x=1}^{V} \frac{\partial}{\partial v_c} \exp(u_x^T v_c) =$$

$$= \frac{1}{\sum_{w=1}^{W} \exp(u_w^T v_c)} \cdot \sum_{x=1}^{V} \exp(u_x^T v_c) \cdot u_x =$$

$$= \sum_{x=1}^{V} \frac{\exp(u_x^T v_c) \cdot u_x}{\sum_{w=1}^{W} \exp(u_w^T v_c)} = \sum_{x=1}^{V} p(x|c)\, u_x \text{ , so}$$

$$\frac{\partial J}{\partial v_c} = -u_o + \sum_{x=1}^{V} p(x|c)\, u_x = -u_o + \sum_{x=1}^{V} \hat{y}_x u_{x_i} \quad //$$

b) $$\frac{\partial J}{\partial u} = -\frac{\partial}{\partial u_w} \log\left(\frac{\exp(u_o^T v_c)}{\sum_{w=1}^{W} \exp(u_w^T v_c)}\right) = -\frac{\partial}{\partial u_w}\left(\log(\exp(u_o^T v_c))\right)^{\mathrm{I}} +$$

$$+ \frac{\partial}{\partial v_c} \log\left(\sum_{w=1}^{W} \exp(u_w^T v_c)\right)^{\mathrm{II}}$$

I $$\begin{cases} 0 & \text{if } w \neq \text{'o'} \\ -v_c & \text{if } w = \text{'o'} \end{cases}$$

II $$\frac{1}{\sum_{w=1}^{W} \exp(u_w^T v_c)} \cdot \sum_{x=1}^{V} \exp(u_w^T v_c) \cdot v_c = \sum_{x=1}^{V} p(x|c) v_c =$$

$$= \sum_{x=1}^{V} \hat{y}_x v_c \text{ , so} \quad \frac{\partial J}{\partial u_w} = \begin{cases} \sum_{x=1}^{V} \hat{y}_x v_c & , \ w \neq \text{'o'} \\ \sum_{x=1}^{V} \hat{y}_x v_c - v_c & , \ w = 0 \end{cases}$$

(c) 5

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

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

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

$$\frac{\partial J}{\partial u_o} = -\frac{1}{\sigma(u_o^T v_c)} \cdot \sigma'(u_o^T v_c)\, v_c = -(1 - \sigma(u_o^T v_c))\, v_c =$$

$$= (\sigma(u_o^T v_c) - 1)\, v_c$$

$$\frac{\partial J}{\partial u_k} = + \sum_{k=1}^{K} \frac{1}{\sigma(-u_k^T v_c)} \cdot \sigma'(-u_k^T v_c) \cdot v_c =$$

$$= \cancel{\sum_{k=1}^{K} (1 - \sigma(u_k^T v_c))\, v_c} = \sum_{k=1}^{K} (1 - \sigma(-u_k^T v_c))\, v_c \quad \text{for all } k$$

d) $\dfrac{\partial F(w_i, \hat{v})}{\partial U}$ $\quad \dfrac{\partial F(w_i, \hat{v})}{\partial V}$

$$\frac{\partial J_{skip\text{-}gram}(w_{t-m \ldots t+m})}{\partial U} = \sum_{-m \le j \le m, j \ne 0} \frac{\partial F(w_{t+j}, v_c)}{\partial U}$$

$$\frac{\partial J_{skip\text{-}gram}(w_{t-m \ldots t+m})}{\partial v_c} = \sum_{-m \le j \le m, j \ne 0} \frac{\partial F(w_{t+j}, v_c)}{\partial v_c}$$

$$\frac{\partial J_{skip\text{-}gram}(w_{t-m \ldots t+m})}{\partial v_{w_{t+j}}} = 0 \quad , j \ne c$$

$$\frac{\partial J_{CBOW}(w_{t-m \dots t+m})}{\partial u} = \frac{\partial F(w_t, \hat{v})}{\partial u}$$

$$\frac{\partial J_{CBOW}(w_{t-m \dots t+m})}{\partial v_{w_{t+j}}} = \frac{\partial F(w_t, \hat{v})}{\partial \hat{v}} \quad j \in (-m, \dots, -1, +1, \dots, +m)$$

$$\frac{\partial J_{CBOW}(w_{t-m \dots t+m})}{\partial v_{w_{t+j}}} = 0 \quad , j \notin \{-m, \dots, -1, +1, \dots, +m\}$$

Scanned with CamScanner