NLP

NLP Stanford - Assigment 01

1 Softmax

(a) Probar que $softmax(x) = softmax(x+c)$

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

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

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

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

$$= softmax(x)_{(i)}$$

2 Neural Networks Basics

(a) Derive the sigmoid function

$$\sigma(x) = \frac{1}{1+e^{-x}} \qquad " f(x) = \frac{g(x)}{h(x)} \quad f'(x) = \frac{g'(x)h(x) - g(x)h'(x)}{h^2(x)}$$

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

$$g'(x) = \frac{\partial}{\partial x} 1 = 0 \qquad h'(x) = \frac{\partial}{\partial x}(1+e^{-x}) = \frac{\partial}{\partial} 1 + \frac{\partial}{\partial x} e^{-x}$$

$$= 0 + e^{-x} = -e^{-x}$$

$$\sigma'(x) = \frac{0(1+e^{-x}) - 1(e^{-x})}{(1+e^{-x})(1+e^{-x})} = \frac{e^{-x}}{(1+e^{-x})(1+e^{-x})} = \left(\frac{1}{1+e^{-x}}\right)\left(\frac{e^{-x}}{1+e^{-x}}\right)$$

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

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

(b) Derivate the crossentropy function:

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

$$\hat{y}_i = softmax(\theta_i) = \frac{e^{\theta_i}}{\sum_j e^{\theta_j}}$$

chain rule: $y = f(u), u = g(x)$ $\frac{dy}{dx} = \frac{dy}{du} \frac{du}{dx} = \frac{\partial f(u)}{\partial u} \frac{\partial g(x)}{\partial x}$

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

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

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

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

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

$$= \underbrace{-\sum_i y_i \theta_i}_{(1)} + \underbrace{\sum_i \log \sum_j e^{\theta_j}}_{(2)}$$

(1) $$\frac{\partial}{\partial \theta_i} - \sum_i y_i \theta_i = -y_i$$

(2) $$\frac{\partial}{\partial \theta_i} \sum_i \log \sum_j e^{\theta_j} = \frac{\partial}{\partial \theta_i} \underbrace{\log}_{f(u)} \underbrace{\sum_j e^{\theta_j}}_{g(\theta_i)} \qquad f(u) = \log u \quad u = \sum_j e^{\theta_j}$$

$$\frac{\partial}{\partial \theta_i} \log \sum_j e^{\theta_j} = \frac{\partial}{\partial u} f(u) \frac{\partial}{\partial \theta} g(\theta_i)$$

$$= \frac{\partial}{\partial u} \log_e(u) \frac{\partial}{\partial \theta} \sum_j e^{\theta_j}$$

$$= \frac{1}{u \log_e e} \frac{\partial}{\partial \theta_j} \sum_j e^{\theta_j}$$

$$= \frac{1}{u} \frac{\partial}{\partial \theta_j} \sum_j e^{\theta_j}$$

$$= \frac{1}{\sum_j e^{\theta_j}} e^{\theta_i} = \frac{e^{\theta_i}}{\sum_j e^{\theta_j}}$$

Juntando (1) y (2):

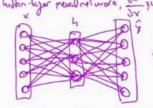
$$\frac{\partial}{\partial \theta_i} CE(y, \hat{y}) = \frac{\partial}{\partial \theta_i} - \sum_i y_i \theta_i + \frac{\partial}{\partial \theta_i} \log \sum_j e^{\theta_j}$$

$$= -y_i + \frac{e^{\theta_i}}{\sum_j e^{\theta_j}} = -y_i + softmax(\theta_i) = -y_i + \hat{y}_i$$

$$= \hat{y}_i - y_i$$

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

(c) Derivate the gradients with respect to the inputs x to an hidden-layer neural network, $\frac{\partial J}{\partial x}$, where $J = CE(y, \hat{y})$



$h: sigmoid(xw_1 + b_1)$

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

$$h = sigmoid(xw_1 + b_1) = sigmoid(z_1) \;/\; z_1 = xw_1 + b_1$$

$$\hat{y} = softmax(hw_2 + b_2) = softmax(z_2) \;/\; z_2 = hw_2 + b_2$$

$$\frac{\partial CE(y, \hat{y})}{\partial x} = \frac{\partial CE(y, \hat{y})}{\partial z_2} \cdot \frac{\partial z_2}{\partial h} \cdot \frac{\partial h}{\partial z_1} \frac{\partial z_1}{\partial x} \quad \text{by chain rule.}$$

Just do for "i",

- $$\frac{\partial CE(y_i, \hat{y}_i)}{\partial z_2} = \hat{y}_i - y_i$$

- Sabemos que $z_2 : g(h) \;/\; z_2 : \mathbb{R}^n \to \mathbb{R}^m$

$$\text{Jacobian} = \frac{\partial z_2}{\partial h} = \begin{bmatrix} \frac{\partial z_1^{(2)}}{\partial h_1} & \cdots & \frac{\partial z_1^{(2)}}{\partial h_m} \\ \frac{\partial z_2^{(2)}}{\partial h_1} & & \frac{\partial z_2^{(2)}}{\partial h_m} \\ \vdots & & \vdots \end{bmatrix}$$

for conveniance $z_2 = z^{(2)}$

$$\begin{bmatrix} \vdots & & \\ \frac{\partial z_m^{(1)}}{\partial h_1} & & \frac{\partial z_m^{(1)}}{\partial h_m} \end{bmatrix}$$

for convenience: $W_2 = W^{(2)}$; $W^{(2)}_{ij}$ weigh of neuron hidden i to output j
$b_2 = b^{(2)}$

$$z_2 = [\underbrace{W^{(2)}_{11}h_1 + W^{(2)}_{21}h_2 + \ldots + W^{(2)}_{m1}h_m + b^{(2)}_1}_{z_{2_1}}, \ldots, \underbrace{W^{(2)}_{1m}h_1 + W^{(2)}_{2m}h_2 + \ldots + W^{(2)}_{mm}h_m + b_m}_{z_{2m}}]$$

$$\frac{\partial z_2}{\partial h} = \begin{bmatrix} w_{11} & w_{21} & \cdots & w_{m1} \\ w_{12} & w_{22} & & w_{m2} \\ \vdots & & & \vdots \\ w_{1m} & w_{2m} & & w_{mm} \end{bmatrix} = W_2^T$$

- $\text{sigmoid}(z) = \sigma(z) = \frac{1}{1+e^{-z}}$; $\frac{\partial \sigma(z)}{\partial z} = \sigma(z)(1-\sigma(z))$

for convenience $z_1 = z^{(1)}$

$$\frac{\partial h_i}{\partial z_i^{(1)}} = \sigma(z_i^{(1)})(1-\sigma(z_i^{(1)}))$$

- $z_1 = xW_1 + b_1$; for convenience $z_1 = z^{(1)}$; $z^{(1)}: \mathbb{R}^{l} \to \mathbb{R}^{m}$

$w_1 = w^{(1)}$

$W^{(1)}_{ij}$ = "for input i to hidden neuron j"

$b_1 = b^{(1)}$

$$\text{Jacobian } \frac{\partial z^{(1)}}{\partial x} = \begin{bmatrix} \frac{\partial z_1^{(1)}}{\partial x_1} & \cdots & \frac{\partial z_1^{(1)}}{\partial x_l} \\ \frac{\partial z_2^{(1)}}{\partial x_1} & & \frac{\partial z_2^{(1)}}{\partial x_l} \\ \vdots & & \vdots \\ \frac{\partial z_m^{(1)}}{\partial x_1} & & \frac{\partial z_m^{(1)}}{\partial x_l} \end{bmatrix}$$

$$z^{(1)} = [\underbrace{x_1w_{11} + x_2w_{21} + \ldots + x_l w_{l1} + b^{(1)}_1}_{z^{(1)}_1}, \ldots, \underbrace{x_1 w_{1m} + x_2 w_{2m} + \ldots + x_l w_{lm} + b^{(1)}_m}_{z^{(1)}_m}]$$

$$\frac{\partial z^{(1)}}{\partial x} = \begin{bmatrix} w^{(1)}_{11} & w^{(1)}_{21} & \cdots & w^{(1)}_{l1} \\ w^{(1)}_{12} & w_{22} & & w_{l2} \\ \vdots & \vdots & & \vdots \\ w^{(1)}_{1m} & w^{(1)}_{2m} & & w^{(1)}_{lm} \end{bmatrix} = W^{(1)T} = W_1^T$$

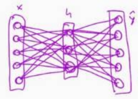
$$\Rightarrow \frac{\partial CE(y,\hat{y})}{\partial x} = \frac{\partial CE(y,\hat{y})}{\partial z_2} \cdot \frac{\partial z_2}{\partial h} \cdot \frac{\partial h}{\partial z_1} \frac{\partial z_1}{\partial x}$$

$$= ((\hat{y}-y) W_2^T \circ (\sigma(x)(1-\sigma(x)))\, W_1^T$$

(d) How many parameters are there in this neural network assuming the input D_x-dimensional, the output is D_y-dimensional and there are H hidden units?

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

(Plus 1) for a neural network with a one hidden layer. Calculate the derivate $\frac{\partial CE}{\partial w_1}, \frac{\partial CE}{\partial b_1}, \frac{\partial CE}{\partial w_2}, \frac{\partial CE}{\partial b_2}$



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

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

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

- $\frac{\partial CE(y,\hat{y})}{\partial w_2} = \frac{\partial CE(y,\hat{y})}{\partial z_2} \cdot \frac{\partial z_2}{\partial w_2} = (\hat{y}-y)\frac{\partial z_2}{\partial w_2}$

$z_2 = z^{(2)} / z^{(2)}: \mathbb{R}^{m} \to \mathbb{R}^{m}$

for convenience: $W_2 = W^{(2)}$; $W^{(2)}_{ij}$ weigh of neuron hidden i to output j
$b_2 = b^{(2)}$

$$z_2 = [\underbrace{W^{(2)}_{11}h_1 + W^{(2)}_{21}h_2 + \ldots + W^{(2)}_{m1}h_m + b^{(2)}_1}_{z_{2_1}}, \ldots, \underbrace{W^{(2)}_{1m}h_1 + W^{(2)}_{2m}h_2 + \ldots + W^{(2)}_{mm}h_m + b_m}_{z_{2m}}]$$

$$W^{(2)} = \begin{bmatrix} w^{(2)}_{11} & w^{(2)}_{12} & \cdots & w^{(2)}_{1m} \\ w^{(2)}_{21} & w^{(2)}_{22} & & w^{(2)}_{2m} \\ \vdots & & & \\ w^{(2)}_{m1} & w^{(2)}_{m2} & & w^{(2)}_{mm} \end{bmatrix}$$

$z_i^{(2)}: \mathbb{R}^{m} \to \mathbb{R}$ $\quad \frac{\partial z_i^{(2)}}{\partial w^{(2)}_{\cdot i}} \in \mathbb{R}^{1\times m}$

$h \in \mathbb{R}^{1\times m}$

$\frac{\partial z^{(2)}}{\partial w^{(2)}} = h^T$ $\quad z^{(2)}: \mathbb{R}^{m\times m} \to \mathbb{R}^{m}$

intuición: $\frac{\partial z_i^{(2)}}{\partial w^{(2)}} = \begin{bmatrix} 0 \\ 0 \\ h_i \\ \vdots \\ 0 \end{bmatrix} \Rightarrow \frac{\partial z^{(2)}}{\partial w^{(2)}} = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_m \end{bmatrix}$

- $\frac{\partial CE(y,\hat{y})}{\partial w_2} = h^T(\hat{y}-y)$

- $\frac{\partial CE(y,\hat{y})}{\partial b_2} = \frac{\partial CE(y,\hat{y})}{\partial z_2} \frac{\partial z_2}{\partial b_2} = (\hat{y}-y)$

$\frac{\partial z_2}{\partial b_2}$; $z_2: \mathbb{R}^{m} \to \mathbb{R}^{m}$ $\quad J = \frac{\partial z_2}{\partial b_2} \in \mathbb{R}^{m\times m}$; $b_2 = b^{(2)}$, $z_2 = z^{(2)}$

$$z_2 = [\underbrace{W^{(2)}_{11}h_1 + W^{(2)}_{21}h_2 + \ldots + W^{(2)}_{m1}h_m + b^{(2)}_1}_{z_{2_1}}, \ldots, \underbrace{W^{(2)}_{1m}h_1 + W^{(2)}_{2m}h_2 + \ldots + W^{(2)}_{mm}h_m + b_m}_{z_{2m}}]$$

$$J = \begin{bmatrix} \frac{\partial z_1^{(2)}}{\partial b_1^{(2)}} & \frac{\partial z_1^{(2)}}{\partial b_2^{(2)}} & \cdots & \frac{\partial z_1^{(2)}}{\partial b_m^{(2)}} \\ \frac{\partial z_2^{(2)}}{\partial b_1^{(2)}} & \frac{\partial z_2^{(2)}}{\partial b_2^{(2)}} & & \frac{\partial z_2^{(2)}}{\partial b_m^{(2)}} \\ \vdots & & & \\ \frac{\partial z_m^{(2)}}{\partial b_1^{(2)}} & \frac{\partial z_m^{(2)}}{\partial b_2^{(2)}} & & \frac{\partial z_m^{(2)}}{\partial b_m^{(2)}} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = I$$

$$\frac{\partial CE(\hat{y},y)}{\partial b_2} = \frac{\partial CE(\hat{y},y)}{\partial z_2} \frac{\partial z_2}{\partial b_2} = (\hat{y}-y)I = \hat{y}-y$$

- $\frac{\partial CE(\hat{y},y)}{\partial w_1}$; $w_1 = w^{(1)}$

$$\frac{\partial CE(\hat{y},y)}{\partial w_1} = \frac{\partial CE(\hat{y},y)}{\partial z_1} \frac{\partial z_1}{\partial w_1} = (\hat{y}-y) w_2^T \circ (\sigma(z_1)(1-\sigma(z_1))) \frac{\partial z_1}{\partial w_1} \quad \ldots (1)$$

$z_1: \mathbb{R}^{l\times m} \to \mathbb{R}^{m}$, tal como $\frac{\partial z_2}{\partial w_2} = h^T \Rightarrow \frac{\partial z_1}{\partial w_1} = x^T \quad \ldots (2)$

Juntado (1) y (2):

$$\frac{\partial CE(\hat{y},y)}{\partial w_1} = x^T ((\hat{y}-y) w_2^T \circ (\sigma(z_1)(1-\sigma(z_1)))) \quad w_1 \in \mathbb{R}^{l\times m}$$

$$= x^T ((\hat{y}-y)_{1\times m}\, w^T_{2\,m\times m} \circ \ldots)$$

$$= x^T_{l\times 1}((\hat{y}-y) w_2^T \circ \ldots)_{1\times m}$$

$$= x^T((\hat{y}-y) w_2^T \circ \ldots)_{l\times m}$$

- $\frac{\partial CE(\hat{y},y)}{\partial b_1} = \frac{\partial CE(y,\hat{y})}{\partial z_1} \frac{\partial z_1}{\partial b_1}$

$z_1: \mathbb{R}^{l} \to \mathbb{R}^{m}$ $\quad \frac{\partial z_1}{\partial x} \in \mathbb{R}^{m\times l} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & & 1 \end{bmatrix} = I$

$$\Rightarrow \frac{\partial CE(\hat{y},y)}{\partial b_1} = ((\hat{y}-y) w_2^T \circ (\sigma(z_1)(1-\sigma(z_1)))) I$$

$$= (\hat{y}-y) w_2^T \circ (\sigma(z)(1-\sigma(z)))$$