CS224N Assignment 1

Adriano Carmezim

September 2017

1 Softmax

(a)

$$
\begin{aligned}
softmax(x+c) &= \frac{e^{x+c}}{\sum_j e^{x+c}} \\
&= \frac{e^x e^c}{\sum_j e^x e^c} \\
&= \frac{e^x e^c}{e^c \sum_j e^x} \\
&= \frac{e^x}{\sum_j e^x} \\
&= softmax(x)
\end{aligned}
$$

2 Neural Networks Basics

(a)

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

$$
\begin{aligned}
\nabla\sigma = \frac{\partial\sigma}{\partial x} &= \frac{d}{dx}\frac{1}{1+e^{-x}} \\
&= \frac{d}{dx}(1+e^{-x})^{-1} \\
&= (1+e^{-x})^{-2}(-e^{-x}) \\
&= \frac{e^{-x}}{(e^{-x}+1)^2} \\
&= \sigma(x)(1-\sigma(x))
\end{aligned}
$$

(b)

$$CE(y,\hat{y}) = -\sum_i y_i log(\hat{y_i}), \text{ where } \hat{y} = softmax(\Theta_i)$$

$$\nabla_\Theta CE(y,\hat{y}) = \frac{\partial CE(y,\hat{y})}{\partial \Theta_i}$$
$$= \sum_i y_i \frac{\partial log(softmax(\Theta_i))}{\partial \Theta}$$
$$= \frac{\partial log(\frac{e^{\Theta_k}}{\sum_j e^{\Theta_j}})}{\partial \Theta}$$
$$= \frac{\partial log(e^{\Theta_k})}{\partial \Theta} - \frac{\partial log(\sum_j e^{\Theta_j})}{\partial \Theta_i}$$
$$= \frac{\partial \Theta_k}{\partial \Theta} - \frac{1}{\sum_j e^{\Theta_j}} e^{\Theta_i}$$
$$= y - \hat{y}_i$$

(c)

$$\frac{\partial J}{\partial x}, \text{ where } J = CE(y,\hat{y}),$$
$$h = sigmoid(xW_1 + b_1),$$
$$\hat{y} = softmax(hW_2 + b_2)$$

$$\frac{\partial CE(y,\hat{y})}{\partial x} = \frac{\partial CE(y,\hat{y})}{\partial (hW_2+b_2)} \cdot \frac{\partial (hW_2+b_2)}{\partial h} \cdot \frac{\partial h}{\partial (xW_1+b_1)} \cdot \frac{\partial xW_1+b_1}{\partial x}$$
$$= (y-\hat{y}) \cdot W_2 \cdot ((xW_1+b_1) - (xW_1+b_1)^2) \cdot W_1$$

(d)

$$P = \text{total number of parameters},$$
$$P = H(D_x+1) + D_y(H+1)$$

3 word2vec

(a)

$$J_{softmax-CE}(o, v_c, U) = CE(y,\hat{y})$$
$$J = -\sum_{i=1}^{W} y_i log \frac{e^{u_i^T v_c}}{\sum_{w=1}^{W} e^{u_w^T}}$$
$$\frac{\partial J}{\partial v_c} = \frac{\partial CE}{\partial U^T v_c} \cdot \frac{\partial U^T v_c}{\partial v_c}$$
$$= U^T \cdot (\hat{y} - y)$$

(b)

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

(c)

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

$$\frac{\partial J}{\partial u_o} = -\frac{1}{\sigma(u_o^T v_c} \cdot \sigma(u_o^T v_c)(10\sigma(u_o^T v_c)) \cdot v_c$$
$$= (\sigma(u_o^T v_c) - 1) \cdot v_c$$

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

(d)

skip-gram:

$$\sum_{-m \leq j \leq m \neq 0} \frac{\partial F(w_{c+j}, v_c)}{\partial v_j} = 0, \forall j \neq c$$

CBOW:

$$\frac{\partial F(w_c, \hat{v})}{\partial v_j} = 0, \forall j \notin \{c - m, ..., c + m\}$$
$$\frac{\partial F(w_c, \hat{v})}{\partial v_j} = \frac{\partial F(w_c, \hat{v})}{\partial \hat{v}}, \forall j \in \{c - m, ..., c + m\}$$

4 Sentiment Analysis

(d)

GloVe was trained on a considerable larger corpus and as higher dimensional word vectors proportionally encode more information it yielded a better accuracy in the results