Assignment 1

Chen Luo

March 21, 2018

1 Softmax

1.1 a

$$softmax(\boldsymbol{x}+c) = \frac{e^{\boldsymbol{x}+\boldsymbol{c}}}{\sum_{j=1}^{d} e^{x_j+c}} \tag{1}$$

$$= \frac{e^{\boldsymbol{x}}e^{c}}{e^{c}\sum_{j=1}^{d} e^{x_j}} = softmax(\boldsymbol{x}) \tag{2}$$

2 NN Basic

2.1 a

$$\frac{\partial\sigma(x)}{\partial x} = -1(1+e^{-x})^{-2}\frac{\partial(1+e^{-x})}{\partial x} \tag{3}$$

$$= \frac{e^{-x}}{(1+e^{-x})^2} \tag{4}$$

$$= \frac{e^{-x}+1-1}{(1+e^{-x})^2} \tag{5}$$

$$= \sigma(x) - \sigma(x)^2 \tag{6}$$

2.2 b

$$CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = -\sum_i y_i \log(\hat{y}_i) \tag{7}$$

$$\hat{\boldsymbol{y}} = softmax(\boldsymbol{\theta}) = \frac{e^{\boldsymbol{\theta}}}{\sum_j e^{\theta_j}} \tag{8}$$

$$\frac{\partial CE(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\partial \boldsymbol{\theta}} = -\frac{\partial \sum_i 1_{[y_i==1]} \left\{ \log e^{\theta_i} - \log \sum_j e^{\theta_j} \right\}}{\partial \boldsymbol{\theta}} \tag{9}$$

$$= -\frac{\partial \boldsymbol{y}^T \boldsymbol{\theta} - log \sum_j e^{\theta_j}}{\partial \boldsymbol{\theta}} \tag{10}$$

$$= -\boldsymbol{y} + \frac{\log \mathbf{1}^T e^{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}} \tag{11}$$

$$= -\boldsymbol{y} + \frac{e^{\boldsymbol{\theta}}}{\mathbf{1}^T e^{\boldsymbol{\theta}}} \tag{12}$$

$$= -\boldsymbol{y} + softmax(\boldsymbol{\theta}) \tag{13}$$

$$= \hat{\boldsymbol{y}} - \boldsymbol{y} \tag{14}$$

2.3 c

$$J = CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = CE(\boldsymbol{y}, softmax(\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{b}_2)) \tag{15}$$

$$= CE(\boldsymbol{y}, softmax(\sigma(\boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b}_1)\boldsymbol{W}_2 + \boldsymbol{b}_2)) \tag{16}$$

Actually,

$$J = -\boldsymbol{y}_{1\times D_y} \log(\boldsymbol{h}_{1\times H}\boldsymbol{W}_{H\times D_y} + \boldsymbol{b}_{1\times D_y})^T \tag{17}$$

$$\frac{\partial J}{\partial \boldsymbol{x}} = \frac{\partial J}{\partial(\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{b}_2)} \frac{\partial(\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{b}_2)}{\partial \boldsymbol{h}} \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}} \tag{18}$$

$$= (\hat{\boldsymbol{y}} - \boldsymbol{y})\boldsymbol{W}_2^T \sigma(\boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b}_1)\{\mathbf{1} - \sigma(\boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b}_1)\}^T \boldsymbol{W}_1^T \tag{19}$$

, where $\frac{\partial(\boldsymbol{h}\boldsymbol{W}_2+\boldsymbol{b}_2)}{\partial \boldsymbol{h}}$ is Jaccobian matrix.

Or, [TODO]

$$d(\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{b}_2) = d\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{h}d\boldsymbol{W}_2 + d\boldsymbol{b}_2 \tag{20}$$

$$vec(d(\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{b}_2)) = vec(d\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{h}d\boldsymbol{W}_2 + d\boldsymbol{b}_2) \tag{21}$$

$$= vec(\mathbf{1}d\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{h}d\boldsymbol{W}_2\boldsymbol{I} + \mathbf{1}d\boldsymbol{b}_2\boldsymbol{I}) \rightarrow preparation for \tag{22}$$

$$vec(AXB) = (B^T \otimes A)vec(X) \tag{23}$$

$$= \boldsymbol{W}_2^T \otimes 1 vec(d\boldsymbol{h}) + \boldsymbol{I} \otimes \boldsymbol{h} vec(d\boldsymbol{W}_2) + \boldsymbol{I} \otimes 1 vec(d\boldsymbol{b}_2) \tag{24}$$

2.4 d

$$\boldsymbol{W}_1 : D_x \times H \tag{25}$$
$$\boldsymbol{b}_1 : 1 \times H \tag{26}$$
$$\boldsymbol{W}_2 : H \times D_y \tag{27}$$
$$\boldsymbol{b}_2 : 1 \times D_y \tag{28}$$
$$H(D_x + D_y + 1) + D_y \tag{29}$$

2.5 g

$$dJ = (\hat{\boldsymbol{y}} - \boldsymbol{y})^T d(\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{b}_2) = (\hat{\boldsymbol{y}} - \boldsymbol{y})^T (d\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{h}d\boldsymbol{W}_2 + d\boldsymbol{b}_2) \tag{30}$$
$$= (\hat{\boldsymbol{y}} - \boldsymbol{y})^T d\boldsymbol{h}\boldsymbol{W}_2 + (\hat{\boldsymbol{y}} - \boldsymbol{y})^T \boldsymbol{h}d\boldsymbol{W}_2 + (\hat{\boldsymbol{y}} - \boldsymbol{y})^T d\boldsymbol{b}_2) \tag{31}$$
$$= trace((\hat{\boldsymbol{y}} - \boldsymbol{y})^T d\boldsymbol{h}\boldsymbol{W}_2 + (\hat{\boldsymbol{y}} - \boldsymbol{y})^T \boldsymbol{h}d\boldsymbol{W}_2 + (\hat{\boldsymbol{y}} - \boldsymbol{y})^T d\boldsymbol{b}_2)) \tag{32}$$
$$\frac{\partial J}{\partial \boldsymbol{W}_2} = (\boldsymbol{W}_2(\hat{\boldsymbol{y}} - \boldsymbol{y})^T)^T \tag{33}$$
$$= \boldsymbol{h}^T(\hat{\boldsymbol{y}} - \boldsymbol{y}) \tag{34}$$
$$\frac{\partial J}{\partial \boldsymbol{b}_2} = \hat{\boldsymbol{y}} - \boldsymbol{y} \tag{35}$$
$$d\boldsymbol{h} = d(\sigma(\boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b}_1)) = \sigma'(\cdot) \odot (\boldsymbol{x}d\boldsymbol{W}_1 + d\boldsymbol{b}_1) \tag{36}$$
$$dJ = trace((\hat{\boldsymbol{y}} - \boldsymbol{y})^T \sigma'(\cdot) \odot (\boldsymbol{x}d\boldsymbol{W}_1 + d\boldsymbol{b}_1)\boldsymbol{W}_2) \tag{37}$$
$$= trace(\boldsymbol{W}_2(\hat{\boldsymbol{y}} - \boldsymbol{y})^T \sigma'(\cdot) \odot (\boldsymbol{x}d\boldsymbol{W}_1 + d\boldsymbol{b}_1)) \tag{38}$$
$$Apply\ law\ of\ trace\ and\ element - wise\ product: \tag{39}$$
$$trace(A^T(B \odot C)) = trace((A \odot B)^T C) \tag{40}$$
$$= trace([(\hat{\boldsymbol{y}} - \boldsymbol{y})\boldsymbol{W}_2^T \odot \sigma'(\cdot)]^T (\boldsymbol{x}d\boldsymbol{W}_1 + d\boldsymbol{b}_1)) \tag{41}$$
$$\frac{\partial J}{\partial \boldsymbol{W}_1} = \{[(\hat{\boldsymbol{y}} - \boldsymbol{y})\boldsymbol{W}_2^T \odot \sigma'(\cdot)]^T \boldsymbol{x}\}^T \tag{42}$$
$$= \boldsymbol{x}^T[(\hat{\boldsymbol{y}} - \boldsymbol{y})\boldsymbol{W}_2^T \odot \sigma'(\cdot)] \tag{43}$$
$$\frac{\partial J}{\partial \boldsymbol{b}_1} = [(\hat{\boldsymbol{y}} - \boldsymbol{y})\boldsymbol{W}_2^T \odot \sigma'(\cdot)] \tag{44}$$

3 word2vec

Notice: $\boldsymbol{\nu}$ and $\boldsymbol{\mu}$ are column vectors. Previously, all vectors are row vectors.

3.1 a

$$J_{softmax-CE}(o, \boldsymbol{\nu}_c, \boldsymbol{U}) = CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) \tag{45}$$

$$\hat{\boldsymbol{y}} = softmax(\{\boldsymbol{U}^T_{\{d\times V\}} \times \boldsymbol{\nu}_{c\{d\times 1\}}\}^T) \tag{46}$$

$$dCE(\cdot) = (\hat{\boldsymbol{y}} - \boldsymbol{y}) d\boldsymbol{\nu}_c^T \boldsymbol{U} \tag{47}$$

$$= trace[(\hat{\boldsymbol{y}} - \boldsymbol{y}) d\boldsymbol{\nu}_c^T \boldsymbol{U}] \tag{48}$$

$$= trace[\boldsymbol{U}(\hat{\boldsymbol{y}} - \boldsymbol{y}) d\boldsymbol{\nu}_c^T] \tag{49}$$

$$\frac{\partial CE(\cdot)}{\partial \boldsymbol{\nu}_c^T} = (\hat{\boldsymbol{y}} - \boldsymbol{y})^T \boldsymbol{U}^T \tag{50}$$

$$\frac{\partial CE(\cdot)}{\partial \boldsymbol{\nu}_c} = \boldsymbol{U}(\hat{\boldsymbol{y}} - \boldsymbol{y}) \tag{51}$$

3.2 b

$$dCE(\cdot) = (\hat{\boldsymbol{y}} - \boldsymbol{y}) \boldsymbol{\nu}_c^T d\boldsymbol{U} \tag{52}$$

$$= trace[(\hat{\boldsymbol{y}} - \boldsymbol{y}) \boldsymbol{\nu}_c^T d\boldsymbol{U}] \tag{53}$$

$$\frac{\partial CE}{\partial \boldsymbol{U}} = \boldsymbol{\nu}_c (\hat{\boldsymbol{y}} - \boldsymbol{y})^T \tag{54}$$

3.3 c

Assume that for $\boldsymbol{\nu}_c$, $\boldsymbol{k}_{neg}$ is a row vector of K-hot (ensure $\boldsymbol{k}_{neg\{o\}} \neq 1$) and $\boldsymbol{k}$ is one-hot vector that $\boldsymbol{k}_o = 1$. Rewrite $J_{neg-sample}(\cdot)$ as follow,

$$J_{neg-sample}(\cdot) = -\log(\sigma(\boldsymbol{\nu}_c^T \boldsymbol{U}))\boldsymbol{k}^T - \log(\sigma(-\boldsymbol{\nu}_c^T \boldsymbol{U}))\boldsymbol{k}_{neg}^T \tag{55}$$

$$= -\log(\sigma(\boldsymbol{\nu}_c^T \boldsymbol{U}\boldsymbol{k}^T)) - \log(\sigma(-\boldsymbol{\nu}_c^T \boldsymbol{U}\boldsymbol{k}_{neg}^T)) \tag{56}$$

In the last equation, we put all vectors into element-wise operator.

$$dJ_{neg-sample}(\cdot) = \frac{\sigma(\boldsymbol{\nu}_c^T \boldsymbol{U}\boldsymbol{k}^T)(1 - \sigma(\boldsymbol{\nu}_c^T \boldsymbol{U}\boldsymbol{k}^T))}{\sigma(\boldsymbol{\nu}_c^T \boldsymbol{U}\boldsymbol{k}^T)} \tag{57}$$

$$\frac{\partial J_{neg-sample}(\cdot)}{\partial \boldsymbol{\nu}_c} = \tag{58}$$

TODO

$$\frac{\partial J}{\partial \boldsymbol{\nu}_c} = -\boldsymbol{\mu}_o \frac{\sigma(\boldsymbol{\mu}_o^T \boldsymbol{\nu}_c)(1-\sigma(\boldsymbol{\mu}_o^T \boldsymbol{\nu}_c))}{\sigma(\boldsymbol{\mu}_o^T \boldsymbol{\nu}_c)} + \sum_{k=1}^{K} \boldsymbol{\mu}_k (1-\sigma(-\boldsymbol{\mu}_k^T \boldsymbol{\nu}_c)) \tag{59}$$

$$= -\boldsymbol{\mu}_o (1-\sigma(\boldsymbol{\mu}_o^T \boldsymbol{\nu}_c)) + \sum_{k=1}^{K} \boldsymbol{\mu}_k (1-(1-\sigma(\boldsymbol{\mu}_k^T \boldsymbol{\nu}_c)) \tag{60}$$

$$= -\boldsymbol{\mu}_o (1-\sigma(\boldsymbol{\mu}_o^T \boldsymbol{\nu}_c)) + \sum_{k=1}^{K} \boldsymbol{\mu}_k (\sigma(\boldsymbol{\mu}_k^T \boldsymbol{\nu}_c)) \tag{61}$$

$$dJ(\cdot) = -\frac{d(\sigma(\cdot))}{\sigma(\cdot)} + \sum_{k=1}^{K} (1-\sigma(-\boldsymbol{\mu}_k^T \boldsymbol{\nu}_c)) d\boldsymbol{\mu}_k^T \boldsymbol{\nu}_c \tag{62}$$

$$= -trace((1-\sigma(\boldsymbol{\mu}_o^T \boldsymbol{\nu}_c)) d\boldsymbol{\mu}_o^T \boldsymbol{\nu}_c) \tag{63}$$

$$= trace(\boldsymbol{\nu}_c (\sigma(\boldsymbol{\mu}_o^T \boldsymbol{\nu}_c) - 1) d\boldsymbol{\mu}_o^T) \tag{64}$$

$$\frac{\partial J}{\partial \boldsymbol{\mu}_o} = \boldsymbol{\nu}_c (\sigma(\boldsymbol{\mu}_o^T \boldsymbol{\nu}_c) - 1) \tag{65}$$

$$\tag{66}$$

$$\frac{\partial J}{\partial \boldsymbol{\mu}_k} = \boldsymbol{\nu}_c \sigma(\boldsymbol{\mu}_k^T \boldsymbol{\nu}_c), for\ k = [1, 2, ..., K] \tag{67}$$

Computation efficiency of negtive-sampling loss, speed up ratio is approximately

$$\frac{V}{K+1}$$

3.4 d

For skip-gram:

As the definition of word-vector, there are no items of $\boldsymbol{\nu}_k$ where $k \neq c$ in cost function $J(o, \boldsymbol{\nu}_c, \boldsymbol{U})$ or $F(\cdot)$.

$$\frac{\partial J_{skip-gram}(\omega_{t-m}, ...\omega_{t+m})}{\partial \boldsymbol{\nu}_k} = \mathbf{0}, for\ any\ k \neq c \tag{68}$$

$$\tag{69}$$

For the remaining parts is only applying the gradients derived by previous questions into the sum operator. The inner parts depend on the specified definition of $F(o, c)$.

For CBOW:
If softmax-cross-entropy cost,

$$p(\omega_t, \hat{\boldsymbol{\nu}}) = \frac{\exp(\boldsymbol{\mu}_{\omega_t}^T \sum_{-m \le j \le m, j \neq 0} \boldsymbol{\nu}_{\omega_j})}{\sum_{v=1}^{V} \exp(\boldsymbol{\mu}_v^T \sum_{-m \le j \le m, j \neq 0} \boldsymbol{\nu}_{\omega_j})} \tag{70}$$

$$J_{CBOW} = F(\omega_t, \hat{\boldsymbol{\nu}}) = CE(\boldsymbol{y}, \hat{\boldsymbol{y}} = softmax(\boldsymbol{U}_{\{d \times V\}}^T \times \hat{\boldsymbol{\nu}}_{\{d \times 1\}}^T)) \tag{71}$$

$$dJ_{CBOW} = \boldsymbol{U}(\hat{\boldsymbol{y}} - \boldsymbol{y}) d\hat{\boldsymbol{\nu}} \tag{72}$$

$$= \sum_{-m \le j \le m, j \neq 0} \boldsymbol{U}(\hat{\boldsymbol{y}} - \boldsymbol{y}) d\boldsymbol{\nu}_{\omega_j} \tag{73}$$