Assignment 1

Chen Luo

April 8, 2018

1 Softmax

1.1 a

$$\begin{aligned} softmax(\boldsymbol{x} + c) &= \frac{e^{\boldsymbol{x+c}}}{\sum_{j=1}^{d} e^{x_j + c}} && (1) \\ &= \frac{e^{\boldsymbol{x}} e^{c}}{e^{c} \sum_{j=1}^{d} e^{x_j}} = softmax(\boldsymbol{x}) && (2) \end{aligned}$$

2 NN Basic

2.1 a

$$\begin{aligned} \frac{\partial \sigma(x)}{\partial x} &= -1(1+e^{-x})^{-2} \frac{\partial (1+e^{-x})}{\partial x} && (3) \\ &= \frac{e^{-x}}{(1+e^{-x})^2} && (4) \\ &= \frac{e^{-x}+1-1}{(1+e^{-x})^2} && (5) \\ &= \sigma(x) - \sigma(x)^2 && (6) \end{aligned}$$

2.2 b

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

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

$$\begin{aligned}
\frac{\partial CE(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\partial \boldsymbol{\theta}} &= -\frac{\partial \sum_i 1_{[y_i==1]} \left\{ \log e^{\theta_i} - \log \sum_j e^{\theta_j} \right\}}{\partial \boldsymbol{\theta}} &(9)\\
&= -\frac{\partial \boldsymbol{y}^T \boldsymbol{\theta} - log \sum_j e^{\theta_j}}{\partial \boldsymbol{\theta}} &(10)\\
&= -\boldsymbol{y} + \frac{\log \mathbf{1}^T e^{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}} &(11)\\
&= -\boldsymbol{y} + \frac{e^{\boldsymbol{\theta}}}{\mathbf{1}^T e^{\boldsymbol{\theta}}} &(12)\\
&= -\boldsymbol{y} + softmax(\boldsymbol{\theta}) &(13)\\
&= \hat{\boldsymbol{y}} - \boldsymbol{y} &(14)
\end{aligned}$$

2.3 c

$$\begin{aligned}
J = CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) &= CE(\boldsymbol{y}, softmax(\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{b}_2)) &(15)\\
&= CE(\boldsymbol{y}, softmax(\sigma(\boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b}_1)\boldsymbol{W}_2 + \boldsymbol{b}_2)) &(16)
\end{aligned}$$

Actually,

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

$$\begin{aligned}
\frac{\partial J}{\partial \boldsymbol{x}} &= \frac{\partial J}{\partial(\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{b}_2)} \frac{\partial(\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{b}_2)}{\partial \boldsymbol{h}} \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{x}} &(18)\\
&= (\hat{\boldsymbol{y}} - \boldsymbol{y})\boldsymbol{W}_2^T \sigma(\boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b}_1)\{\mathbf{1} - \sigma(\boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b}_1)\}^T \boldsymbol{W}_1^T &(19)
\end{aligned}$$

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

Or, [TODO]

$$\begin{aligned}
d(\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{b}_2) &= d\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{h}d\boldsymbol{W}_2 + d\boldsymbol{b}_2 &(20)\\
vec(d(\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{b}_2)) &= vec(d\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{h}d\boldsymbol{W}_2 + d\boldsymbol{b}_2) &(21)\\
&= vec(\mathbf{1}d\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{h}d\boldsymbol{W}_2\boldsymbol{I} + \mathbf{1}d\boldsymbol{b}_2\boldsymbol{I}) \rightarrow preparation for &(22)\\
vec(AXB) &= (B^T \otimes A)vec(X) &(23)\\
&= \boldsymbol{W}_2^T \otimes 1 vec(d\boldsymbol{h}) + \boldsymbol{I} \otimes \boldsymbol{h} vec(d\boldsymbol{W}_2) + \boldsymbol{I} \otimes 1 vec(d\boldsymbol{b}_2) &(24)
\end{aligned}$$

2.4 d

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

2.5 g

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

3 word2vec

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

3.1 a

$$
\begin{aligned}
J_{softmax-CE}(o, \boldsymbol{\nu}_c, \boldsymbol{U}) &= CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) && (45)\\
\hat{\boldsymbol{y}} &= softmax(\{\boldsymbol{U}^T_{\{d\times V\}} \times \boldsymbol{\nu}_{c\{d\times 1\}}\}^T) && (46)\\
dCE(\cdot) &= (\hat{\boldsymbol{y}} - \boldsymbol{y})d\boldsymbol{\nu}_c^T\boldsymbol{U} && (47)\\
&= trace[(\hat{\boldsymbol{y}} - \boldsymbol{y})d\boldsymbol{\nu}_c^T\boldsymbol{U}] && (48)\\
&= trace[\boldsymbol{U}(\hat{\boldsymbol{y}} - \boldsymbol{y})d\boldsymbol{\nu}_c^T] && (49)\\
\frac{\partial CE(\cdot)}{\partial \boldsymbol{\nu}_c^T} &= (\hat{\boldsymbol{y}} - \boldsymbol{y})^T\boldsymbol{U}^T && (50)\\
\frac{\partial CE(\cdot)}{\partial \boldsymbol{\nu}_c} &= \boldsymbol{U}(\hat{\boldsymbol{y}} - \boldsymbol{y}) && (51)
\end{aligned}
$$

3.2 b

$$
\begin{aligned}
dCE(\cdot) &= (\hat{\boldsymbol{y}} - \boldsymbol{y})\boldsymbol{\nu}_c^T d\boldsymbol{U} && (52)\\
&= trace[(\hat{\boldsymbol{y}} - \boldsymbol{y})\boldsymbol{\nu}_c^T d\boldsymbol{U}] && (53)\\
\frac{\partial CE}{\partial \boldsymbol{U}} &= \boldsymbol{\nu}_c(\hat{\boldsymbol{y}} - \boldsymbol{y})^T && (54)
\end{aligned}
$$

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,

$$
\begin{aligned}
J_{neg-sample}(\cdot) &= -\log(\sigma(\boldsymbol{\nu}_c^T\boldsymbol{U}))\boldsymbol{k}^T - \log(\sigma(-\boldsymbol{\nu}_c^T\boldsymbol{U}))\boldsymbol{k}_{neg}^T && (55)\\
&= -\log(\sigma(\boldsymbol{\nu}_c^T\boldsymbol{U}\boldsymbol{k}^T)) - \log(\sigma(-\boldsymbol{\nu}_c^T\boldsymbol{U}\boldsymbol{k}_{neg}^T)) && (56)
\end{aligned}
$$

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

$$
\begin{aligned}
dJ_{neg-sample}(\cdot) &= \frac{\sigma(\boldsymbol{\nu}_c^T\boldsymbol{U}\boldsymbol{k}^T)(1-\sigma(\boldsymbol{\nu}_c^T\boldsymbol{U}\boldsymbol{k}^T))}{\sigma(\boldsymbol{\nu}_c^T\boldsymbol{U}\boldsymbol{k}^T)} && (57)\\
\frac{\partial J_{neg-sample}(\cdot)}{\partial \boldsymbol{\nu}_c} &= && (58)
\end{aligned}
$$

TODO

$$
\begin{aligned}
\frac{\partial J}{\partial \boldsymbol{\nu}_c} &= -\boldsymbol{\mu}_o \frac{\sigma(\boldsymbol{\mu}_o^T \boldsymbol{\nu}_c)(1-\sigma(\boldsymbol{\mu}_o^T \boldsymbol{\nu}_c))}{\sigma(\boldsymbol{\mu}_o^T \boldsymbol{\nu}_c)} + \sum_{k=1}^{K} \boldsymbol{\mu}_k (1-\sigma(-\boldsymbol{\mu}_k^T \boldsymbol{\nu}_c)) \qquad (59)\\
&= -\boldsymbol{\mu}_o (1-\sigma(\boldsymbol{\mu}_o^T \boldsymbol{\nu}_c)) + \sum_{k=1}^{K} \boldsymbol{\mu}_k (1-(1-\sigma(\boldsymbol{\mu}_k^T \boldsymbol{\nu}_c)) \qquad (60)\\
&= -\boldsymbol{\mu}_o (1-\sigma(\boldsymbol{\mu}_o^T \boldsymbol{\nu}_c)) + \sum_{k=1}^{K} \boldsymbol{\mu}_k (\sigma(\boldsymbol{\mu}_k^T \boldsymbol{\nu}_c)) \qquad (61)
\end{aligned}
$$

$$
\begin{aligned}
dJ(\cdot) &= -\frac{d(\sigma(\cdot))}{\sigma(\cdot)} + \sum_{k=1}^{K} (1-\sigma(-\boldsymbol{\mu}_k^T \boldsymbol{\nu}_c)) d\boldsymbol{\mu}_k^T \boldsymbol{\nu}_c \qquad (62)\\
&= -trace((1-\sigma(\boldsymbol{\mu}_o^T \boldsymbol{\nu}_c)) d\boldsymbol{\mu}_o^T \boldsymbol{\nu}_c) \qquad (63)\\
&= trace(\boldsymbol{\nu}_c (\sigma(\boldsymbol{\mu}_o^T \boldsymbol{\nu}_c) - 1) d\boldsymbol{\mu}_o^T) \qquad (64)\\
\frac{\partial J}{\partial \boldsymbol{\mu}_o} &= \boldsymbol{\nu}_c (\sigma(\boldsymbol{\mu}_o^T \boldsymbol{\nu}_c) - 1) \qquad (65)\\
& \qquad (66)
\end{aligned}
$$

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

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 $F(\cdot)$.

$$
\begin{aligned}
\frac{\partial J_{skip-gram}(\omega_{t-m}, ...\omega_{t+m})}{\partial \boldsymbol{\nu}_k} &= \mathbf{0}, for\ any\ k \neq c \qquad (68)\\
& \qquad (69)
\end{aligned}
$$

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,

$$\begin{aligned} p(\omega_t, \hat{\boldsymbol{\nu}}) &= \frac{\exp(\boldsymbol{\mu}_{\omega_t}^T \sum_{-m\leq j\leq m, j\neq 0} \boldsymbol{\nu}_{\omega_j})}{\sum_{v=1}^{V} \exp(\boldsymbol{\mu}_{v}^T \sum_{-m\leq j\leq m, j\neq 0} \boldsymbol{\nu}_{\omega_j})} \tag{70} \\ J_{CBOW} = F(\omega_t, \hat{\boldsymbol{\nu}}) &= CE(\boldsymbol{y}, \hat{\boldsymbol{y}} = softmax(\boldsymbol{U}_{\{d\times V\}}^T \times \hat{\boldsymbol{\nu}}_{\{d\times 1\}}^T)) \tag{71} \\ dJ_{CBOW} &= \boldsymbol{U}(\hat{\boldsymbol{y}} - \boldsymbol{y})d\hat{\boldsymbol{\nu}} \tag{72} \\ &= \sum_{-m\leq j\leq m, j\neq 0} \boldsymbol{U}(\hat{\boldsymbol{y}} - \boldsymbol{y})d\boldsymbol{\nu}_{\omega_j} \tag{73} \end{aligned}$$

In the context window, $\boldsymbol{\nu}_{\omega_j}$ may be duplicated. All of the gradients of $\boldsymbol{\nu}_{\omega_j}$ is $\boldsymbol{U}(\hat{\boldsymbol{y}} - \boldsymbol{y})$ times *occurence* of ω_j

4 Remark of word2vec

Classification task to classify whether two words would be occur together (with different representations)

5 sentiment classification

5.1 b

Prevent parameters grow too large. −¿ More stable model −¿ Prevent overfitting.