PSet Assignment #1

1 Softmax

a) Trivial

b) Code

2 Neural Network Basics

a) $\forall x \in \mathbb{R}$, $\sigma(x) = \frac{1}{1+e^{-x}}$
$\forall x \in \mathbb{R}$:

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

b) $\boldsymbol{y} \in \mathbb{R}^n$, $\exists k \in [\![1, n]\!] \quad \boldsymbol{y} = \boldsymbol{e}_k$. Therefore

$$\forall \theta \in \mathbb{R}^n \quad CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = -\log(\hat{y_k}) = -\theta_k + \log(\sum_{i=1}^{n} e^{\theta_i})$$

Then $\forall \theta \in \mathbb{R}^n$:

$$\nabla_\theta CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \hat{\boldsymbol{y}} - \boldsymbol{y}$$

c)

$$\boldsymbol{h} = \sigma(\boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b}_1)$$

$$\hat{\boldsymbol{y}} = \text{softmax}(\boldsymbol{h}\boldsymbol{W}_2 + \boldsymbol{b}_2)$$

$\forall \boldsymbol{x} \in \mathbb{R}^n$:

$$
\begin{aligned}
\nabla_x CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) &= \sum_{i=1}^{n} \frac{\partial CE(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\partial x_i} \boldsymbol{e}_i \\
&= \nabla_\theta CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) \cdot (\frac{\partial \theta}{\partial \boldsymbol{x}})^\top \\
\forall i \in [\![1, D_x]\!], \forall j \in [\![1, D_y]\!] \quad (\frac{\partial \theta_j}{\partial x_i}) &= (\frac{\partial(\sum_{l=1}^{h} h_l (W_2)_{lj} + (b_2)_j)}{\partial x_i}) \\
&= (\frac{\partial(\sum_{l=1}^{h} h_l (W_2)_{lj} + (b_2)_j)}{\partial x_i}) \\
&= (\frac{\partial(\sum_{l=1}^{h} \sigma(\sum_{m=1}^{D_x} x_m (W_1)_{ml} + (b1)_l)(W_2)_{lj})}{\partial x_i}) \\
&= (\sum_{l=1}^{h} (W_1)_{il} \sigma'(\sum_{m=1}^{D_x} x_m (W_1)_{ml} + (b1)_l)(W_2)_{lj})) \\
&= (\boldsymbol{W}_1)_{i.} \cdot (\boldsymbol{h'} * (\boldsymbol{W}_2)_{.j}) \\
\text{Then } \nabla_x CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) &= (\hat{\boldsymbol{y}} - \boldsymbol{y}) \boldsymbol{W}_2^\top \boldsymbol{D} \boldsymbol{W}_1^\top \text{ where } \boldsymbol{D} = \text{diag}(\sigma'(\boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b}_1))
\end{aligned}
$$

d) There are $H(1 + D_x) + D_y(1 + H)$ parameters.

e) Code

f) Code

g) Code

3 word2vec

a) $\hat{\boldsymbol{y}}_o = p(\boldsymbol{o}|\boldsymbol{c}) = \frac{\exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}{\sum_{w=1}^{W} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}$.
We can rewrite $\hat{\boldsymbol{y}}_o$ as follows :

$$\hat{\boldsymbol{y}}_o = \text{softmax}(\boldsymbol{U}^\top \boldsymbol{v}_c)_o \quad \text{where } \boldsymbol{U} = [\boldsymbol{u}_1, ..., \boldsymbol{u}_W]$$

Therefore, if $\boldsymbol{y} = \boldsymbol{e}_o$, then $\forall \boldsymbol{v}_c \in \mathbb{R}^n$

$$
\begin{aligned}
\nabla_{\boldsymbol{v}_c} CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) &= \boldsymbol{U} \nabla_{\boldsymbol{U}^\top \boldsymbol{v}_c} CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) \\
&= \boldsymbol{U}(\hat{\boldsymbol{y}} - \boldsymbol{y})
\end{aligned}
$$

b)
- If $w \neq o$:

$$
\begin{aligned}
\nabla_{\boldsymbol{u}_w} CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) &= (\frac{\partial(\boldsymbol{U}^\top \boldsymbol{v}_c)}{\partial \boldsymbol{u}_w})^\top (\hat{\boldsymbol{y}} - \boldsymbol{y}) \\
&= (\sum_{i=1}^{W} \frac{\partial((\boldsymbol{u}_i^\top \boldsymbol{v}_c)\boldsymbol{e}_i)}{\partial \boldsymbol{u}_w})^\top (\hat{\boldsymbol{y}} - \boldsymbol{y}) \\
&= \hat{\boldsymbol{y}}_w \cdot \boldsymbol{v}_c
\end{aligned}
$$

- If $w = o$:

$$\nabla_{\boldsymbol{u}_o} CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = (\hat{\boldsymbol{y}}_o - 1) \cdot \boldsymbol{v}_c$$

c) Now $J_{\text{neg-sample}}(\boldsymbol{o}, \boldsymbol{v}_c, \boldsymbol{U}) = -\log((\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)) - \sum_{k=1}^K \log(\sigma(-\boldsymbol{u}_k^\top \boldsymbol{v}_c))$

- $\forall \boldsymbol{v}_c \in \mathbb{R}^n$

$$\begin{aligned}\nabla_{\boldsymbol{v}_c} CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) &= -\frac{\sigma(-\boldsymbol{u}_o^\top \boldsymbol{v}_c)\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}{\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)} \boldsymbol{u}_o + \sum_{k=1}^K \frac{\sigma(-\boldsymbol{u}_k^\top \boldsymbol{v}_c))\sigma(\boldsymbol{u}_k^\top \boldsymbol{v}_c)}{\sigma(-\boldsymbol{u}_k^\top \boldsymbol{v}_c)} \boldsymbol{u}_k \\ &= -\sigma(-\boldsymbol{u}_o^\top \boldsymbol{v}_c)\boldsymbol{u}_o + \sum_{k=1}^K \sigma(\boldsymbol{u}_k^\top \boldsymbol{v}_c)\boldsymbol{u}_k\end{aligned}$$

- If $w \neq o$:

$$\nabla_{\boldsymbol{u}_w} CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \sigma(\boldsymbol{u}_w^\top \boldsymbol{v}_c)\boldsymbol{v}_c$$

- If $w = o$:

$$\nabla_{\boldsymbol{u}_o} CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = -\sigma(-\boldsymbol{u}_o^\top \boldsymbol{v}_c)\boldsymbol{v}_c$$

d) $J_{\text{skip-gram}}(\text{word}_{c-m \dots c+m}) = \sum_{-m \leq j \leq m, j \neq 0} F(\boldsymbol{w}_{c+j}, \boldsymbol{v}_c)$

- $\forall \boldsymbol{v}_c \in \mathbb{R}^n$

$$\nabla_{\boldsymbol{v}_c} J = \sum_{-m \leq j \leq m, j \neq 0} \frac{\partial F(\boldsymbol{w}_{c+j}, \boldsymbol{v}_c)}{\partial \boldsymbol{v}_c}$$

-

$$\nabla_{\boldsymbol{u}_w} J = \sum_{-m \leq j \leq m, j \neq 0} \frac{\partial F(\boldsymbol{w}_{c+j}, \boldsymbol{v}_c)}{\partial \boldsymbol{u}_w}$$

$$J_{\text{CBOW}}(\text{word}_{c-m \dots c+m}) = F(\boldsymbol{w}_c, \hat{\boldsymbol{v}}) \text{ with } \hat{\boldsymbol{v}} = \sum_{-m \leq j \leq m, j \neq 0} \boldsymbol{v}_{c+j}$$

-

$$\nabla_{\boldsymbol{v}_c} J = \vec{0}$$

-

$$\nabla_{\boldsymbol{u}_w} J = \frac{\partial F(\boldsymbol{w}_c, \hat{\boldsymbol{v}})}{\partial \boldsymbol{u}_w}$$

e) Code

f) Code

g) Code

h) Code

4 Sentiment Analysis

a) Code

b) Blah blah blah ... OverFitting ... Blah blah blah more Bias for less Variance ... Blah blah blah

c) Code

d) Code