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

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

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

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

c)

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

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

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

$$\begin{aligned}
\nabla_{\boldsymbol{x}} CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) &= \nabla_{\boldsymbol{\theta}} CE(\boldsymbol{y}, \hat{\boldsymbol{y}})(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{x}}) \\
&\text{where } \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{x}} \text{ is the Jacobian Matrix :}
\end{aligned}$$

$$\begin{aligned}
\forall i \in [\![1, D_y]\!], \forall j \in [\![1, D_x]\!] \quad (\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{x}})_{ij} &= (\frac{\partial \theta_i}{\partial x_j}) \\
&= (\frac{\partial(\sum_{l=1}^{h} h_l (W_2)_{li} + (b_2)_i)}{\partial x_j}) \\
&= (\frac{\partial(\sum_{l=1}^{h} \sigma(\sum_{m=1}^{D_x} x_m (W_1)_{ml} + (b_1)_l)(W_2)_{li})}{\partial x_j}) \\
&= (\sum_{l=1}^{h} (W_1)_{jl} \sigma'(\sum_{m=1}^{D_x} x_m (W_1)_{ml} + (b_1)_l)(W_2)_{li})) \\
&= (\boldsymbol{W}_1)_{j.} \cdot (\boldsymbol{h'}^\top * (\boldsymbol{W}_2)_{.i}) \\
&= (\boldsymbol{W}_1)_{j.} \cdot (\boldsymbol{D}\boldsymbol{W}_2)_{.i} \text{ where } \boldsymbol{D} = \text{diag}(\sigma'(\boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b}_1)) \\
&= ((\boldsymbol{W}_1 \boldsymbol{D} \boldsymbol{W}_2)^\top)_{ij}
\end{aligned}$$

$$\text{Then } \nabla_{\boldsymbol{x}} CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = (\hat{\boldsymbol{y}} - \boldsymbol{y}) \boldsymbol{W}_2^\top \boldsymbol{D} \boldsymbol{W}_1^\top \text{ where } \boldsymbol{D} = \text{diag}(\sigma'(\boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b}_1))$$

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

e) Code

f) Code

g)

$$\begin{aligned}
\nabla_{\boldsymbol{W}_2} CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) &= \nabla_{\boldsymbol{\theta}} CE(\boldsymbol{y}, \hat{\boldsymbol{y}})(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{W}_2}) \\
&= \boldsymbol{h}^\top (\hat{\boldsymbol{y}} - \boldsymbol{y})
\end{aligned}$$

$$\begin{aligned}
\nabla_{\boldsymbol{b}_2} CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) &= \nabla_{\boldsymbol{\theta}} CE(\boldsymbol{y}, \hat{\boldsymbol{y}})(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{b}_2}) \\
&= (\hat{\boldsymbol{y}} - \boldsymbol{y})
\end{aligned}$$

$$\begin{aligned}
\nabla_{\boldsymbol{W}_1} CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) &= \nabla_{\boldsymbol{\theta}} CE(\boldsymbol{y}, \hat{\boldsymbol{y}})(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{W}_1}) \\
&= \nabla_{\boldsymbol{\theta}} CE(\boldsymbol{y}, \hat{\boldsymbol{y}})(\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{h}})(\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{W_1}}) \\
&= \nabla_{\boldsymbol{\theta}} CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) \boldsymbol{W}_2^\top (\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{W_1}}) \\
&= \boldsymbol{x}^\top \nabla_{\boldsymbol{\theta}} CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) \boldsymbol{W}_2^\top \boldsymbol{D}
\end{aligned}$$

$$\nabla_{\boldsymbol{b}_1} CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \nabla_{\boldsymbol{\theta}} CE(\boldsymbol{y}, \hat{\boldsymbol{y}}) \boldsymbol{W}_2^\top \boldsymbol{D}$$

3 word2vec

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

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

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

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

b) $\forall w \in [\![1, W]\!]$:

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

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

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

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

- If $w \neq o$:

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

- If $w = o$:

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

  Conclusion : $\forall w \in [\![1, W]\!]$:

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

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

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

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

-

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

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

-

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

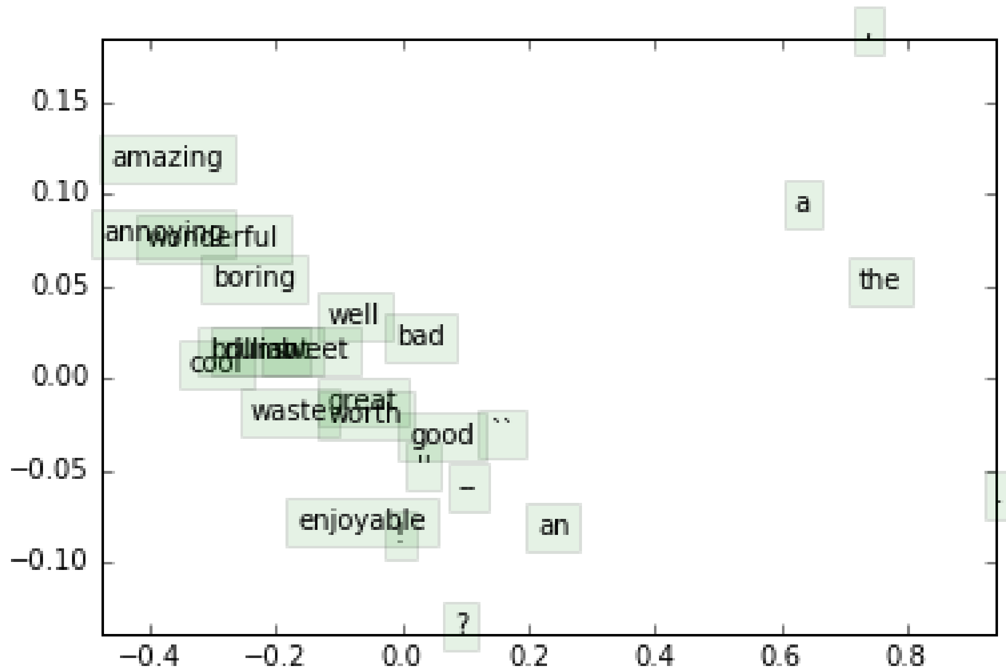
-

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

e) Code

f) Code

g) Code



h)

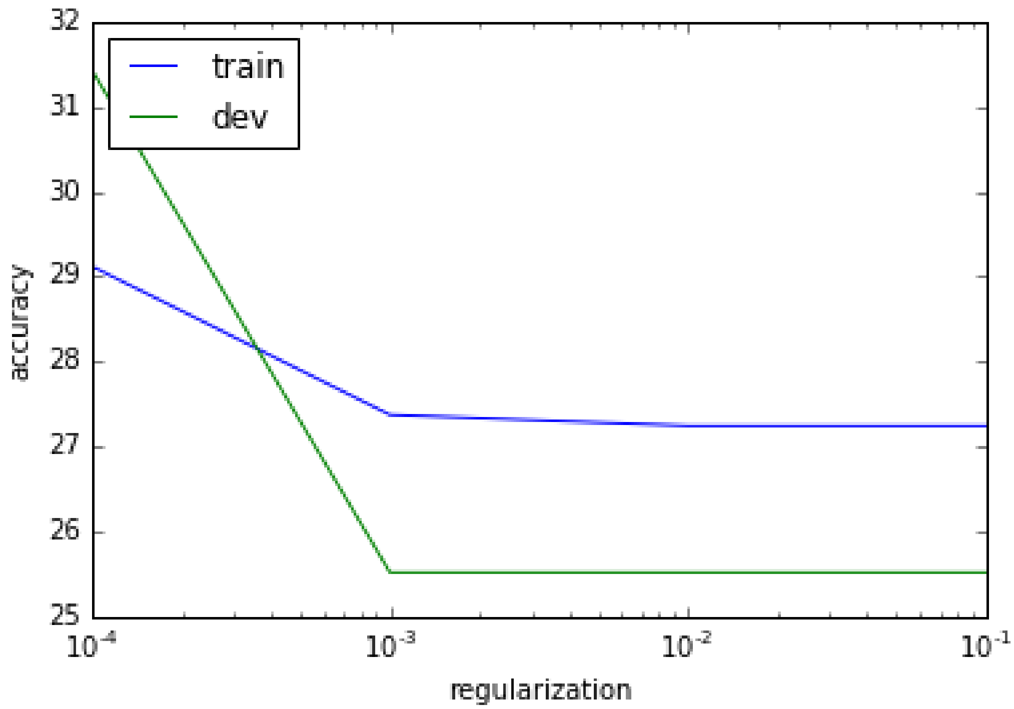
4 Sentiment Analysis

a) $J = \frac{1}{N}\sum_{i=1}^{N} CE(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_i) + \frac{\lambda}{2}\|\boldsymbol{W}\|^2$ with $\hat{\boldsymbol{y}}_i = \text{softmax}(\boldsymbol{x}_i \boldsymbol{W})$
$\forall\ \boldsymbol{W}$:

$$\begin{aligned}
\nabla_{\boldsymbol{W}} J &= \frac{1}{N}\sum_{i=1}^{N} \nabla_{\boldsymbol{W}} CE(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_i) + \frac{\lambda}{2}\nabla_{\boldsymbol{W}}\|\boldsymbol{W}\|^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i^\top (\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i) + \lambda \boldsymbol{W}
\end{aligned}$$

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

c) I selected $\lambda \in \{0.1, 0.01, 0.01, 0.001, 0.0001\}$



32
31
30
29
28
27
26
25
accuracy
train
dev
10^{-4}
10^{-3}
10^{-2}
10^{-1}
regularization

d)