Stanford CS 224n Assignment 2

Kuei-Da Liao (lkueida@eng.ucsd.edu)

October 3, 2020

1. Written: Understanding word2vec

(a)

$$-\sum_{w\in Vocab} y_w \log(\hat{y}_w) = -\sum_{w\in \text{Vocab}} \mathbb{1}(w=o)\log(\hat{y}_w) = -\log(\hat{y}_o)$$

(b)

$$-\log P(O=o|C=c) = -\boldsymbol{u}_o^\top \boldsymbol{v}_c + \log \sum_{w\in \text{Vocab}} \exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c)$$

$$\begin{aligned}
\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}}{\partial \boldsymbol{v}_c} &= -\boldsymbol{u}_o^\top + \frac{1}{\sum_{w'\in \text{Vocab}} \exp(\boldsymbol{u}_{w'}^\top \boldsymbol{v}_c)} \frac{\partial \sum_{w\in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}{\partial \boldsymbol{v}_c} \\
&= -\boldsymbol{u}_o^\top + \sum_{w\in \text{Vocab}} \boldsymbol{u}_w^\top \frac{\exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}{\sum_{w'\in \text{Vocab}} \exp(\boldsymbol{u}_{w'}^\top \boldsymbol{v}_c)} \\
&= -\boldsymbol{u}_o^\top + \sum_{w\in \text{Vocab}} \boldsymbol{u}_w^\top \hat{y}_w \\
&= \sum_{w\in \text{Vocab}} (\hat{y}_w - \mathbb{1}(w=o))\boldsymbol{u}_w^\top
\end{aligned}$$

(c)

$$\begin{aligned}
\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{u}_o} &= -\boldsymbol{v}_c^\top + \frac{1}{\sum_{w'\in \text{Vocab}} \exp(\boldsymbol{u}_{w'}^\top \boldsymbol{v}_c)} \frac{\partial \sum_{w\in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}{\partial \boldsymbol{u}_o} \\
&= -\boldsymbol{v}_c^\top + \sum_{w\in \text{Vocab}} \boldsymbol{v}_c^\top \hat{y}_w \mathbb{1}(w=o) \\
&= -\boldsymbol{v}_c^\top + \boldsymbol{v}_c^\top \hat{y}_o \\
&= (\hat{y}_o - 1)\boldsymbol{v}_c^\top
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{u}_{w,w\neq o}} &= \frac{1}{\sum_{w'\in \text{Vocab}} \exp(\boldsymbol{u}_{w'}^\top \boldsymbol{v}_c)} \frac{\partial \sum_{w\in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}{\partial \boldsymbol{u}_{w,w\neq o}} \\
&= \sum_{w\in \text{Vocab}, w\neq o} \boldsymbol{v}_c^\top \frac{\exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}{\sum_{w'\in \text{Vocab}} \exp(\boldsymbol{u}_{w'}^\top \boldsymbol{v}_c)} \\
&= \sum_{w\in \text{Vocab}, w\neq o} \hat{y}_w \boldsymbol{v}_c^\top
\end{aligned}$$

$$\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{U}} = (\hat{y} - e_o^\top)\boldsymbol{v}_c^\top$$

(d)

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

(e)

$$\begin{aligned}\frac{\partial \boldsymbol{J}_{\text{neg-sample}}}{\partial \boldsymbol{v}_c} &= -\frac{1}{\cancel{\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}}\cancel{\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}\sigma(-\boldsymbol{u}_o^\top \boldsymbol{v}_c)\boldsymbol{u}_o^\top - \sum_{k=1}^{K} -\frac{1}{\cancel{\sigma(-\boldsymbol{u}_k^\top \boldsymbol{v}_c)}}\cancel{\sigma(-\boldsymbol{u}_k^\top \boldsymbol{v}_c)}\sigma(\boldsymbol{u}_k^\top \boldsymbol{v}_c)\boldsymbol{u}_k^\top \\ &= (\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c) - 1)\boldsymbol{u}_o^\top + \sum_{k=1}^{K}\sigma(\boldsymbol{u}_k^\top \boldsymbol{v}_c)\boldsymbol{u}_k^\top\end{aligned}$$

$$\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{u}_o} = -\frac{1}{\cancel{\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}}\cancel{\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}\sigma(-\boldsymbol{u}_o^\top \boldsymbol{v}_c)\boldsymbol{v}_c = (\sigma(\boldsymbol{u}_o^\top \boldsymbol{v}_c) - 1)\boldsymbol{v}_c$$

$$\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{u}_k} = -\frac{1}{\cancel{\sigma(-\boldsymbol{u}_k^\top \boldsymbol{v}_c)}}\cancel{\sigma(-\boldsymbol{u}_k^\top \boldsymbol{v}_c)}\sigma(\boldsymbol{u}_k^\top \boldsymbol{v}_c)(-\boldsymbol{v}_c) = \sigma(\boldsymbol{u}_k^\top \boldsymbol{v}_c)\boldsymbol{v}_c$$

(f) (i)

$$\frac{\partial \boldsymbol{J}_{\text{skip-gram}}(\boldsymbol{v}_c, w_{t-m}, ..., w_{t+m}, \boldsymbol{U})}{\partial \boldsymbol{U}} = \sum_{\substack{-m \le j < m \\ j \ne 0}} \frac{\partial \boldsymbol{J}(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U})}{\partial \boldsymbol{U}}$$

(ii)

$$\frac{\partial \boldsymbol{J}_{\text{skip-gram}}(\boldsymbol{v}_c, w_{t-m}, ..., w_{t+m}, \boldsymbol{U})}{\partial \boldsymbol{v}_c} = \sum_{\substack{-m \le j < m \\ j \ne 0}} \frac{\partial \boldsymbol{J}(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U})}{\partial \boldsymbol{v}_c}$$

(iii)

$$\frac{\partial \boldsymbol{J}_{\text{skip-gram}}(\boldsymbol{v}_c, w_{t-m}, ..., w_{t+m}, \boldsymbol{U})}{\partial \boldsymbol{v}_{w, w \ne c}} = 0$$