Word2vec Assignment 2

Congfeng Yin

2022-05-30

1 符号说明

- d 词向量的维度
- n 词汇数量
- $\mathbf{U} \in \mathbb{R}^{d \times n}$ 每一列是一个词向量，$\mathbf{u}_w \in \mathbb{R}^{d \times 1}$
- $\mathbf{V} \in \mathbb{R}^{d \times n}$ 每一列是一个词向量，$\mathbf{v}_w \in \mathbb{R}^{d \times 1}$
- $\mathbf{y} \in \mathbb{R}^{n \times 1}$ 真实值，one-hot 向量
- $\hat{\mathbf{y}} \in \mathbb{R}^{n \times 1}$ 预测值，表示属于某个词的概率

2 Question (a)

Defination of $\mathbf{y}$:

$$y_w = \begin{cases} 1 & w = o \\ 0 & w \neq o \end{cases}$$

Hence,

$$-\sum_{w=1}^{V} y_w \log(\hat{y}_w) = -y_o \log(\hat{y}_o) = -\log(\hat{y}_o) = -\log P(O = o|C = c)$$

3 Question (b)

$$
\begin{aligned}
\frac{\partial \mathcal{J}_{naive-softmax}(\mathbf{v}_c, o, \mathbf{U})}{\partial \mathbf{v}_c} &= \frac{\partial}{\partial \mathbf{v}_c}[-\mathbf{u}_o^T\mathbf{v}_c + \log \sum_{w \in Vocab} \exp(\mathbf{u}_w^T\mathbf{v}_c)] \\
&= \frac{\partial}{\partial \mathbf{v}_c}[\log \sum_{w \in Vocab} \exp(\mathbf{u}_w^T\mathbf{v}_c)] - \mathbf{u}_o \\
&= \frac{\sum_{x \in Vocab} \exp(\mathbf{u}_x^T\mathbf{v}_c)\mathbf{u}_x}{\sum_{w \in Vocab} \exp(\mathbf{u}_w^T\mathbf{v}_c)} - \mathbf{u}_o \quad {}^1 \\
&= \sum_{x \in Vocab} \frac{\exp(\mathbf{u}_x^T\mathbf{v}_c)}{\sum_{w \in Vocab} \exp(\mathbf{u}_w^T\mathbf{v}_c)} \mathbf{u}_x - \mathbf{u}_o \\
&= \sum_{x \in Vocab} [\frac{\exp(\mathbf{u}_x^T\mathbf{v}_c)}{\sum_{w \in Vocab} \exp(\mathbf{u}_w^T\mathbf{v}_c)} \mathbf{u}_x - y_x\mathbf{u}_x] \\
&= \sum_{x \in Vocab} \mathbf{u}_x(\hat{y}_x - y_x) \\
&= \sum_{w \in Vocab} \mathbf{u}_w(\hat{y}_w - y_w) \\
&= \mathbf{U}(\hat{\mathbf{y}} - \mathbf{y}) \in \mathbb{R}^{d \times 1}
\end{aligned}
$$

推导结果是 outside vector $\mathbf{u}_w$ 的期望减去 $\mathbf{u}_w$ 的实际值，所以可以用 $\mathbf{U}(\hat{\mathbf{y}} - \mathbf{y})$ 表示。

4 Question (c)

$$
\frac{\partial \mathcal{J}_{naive-softmax}(\mathbf{v}_c, o, \mathbf{U})}{\partial \mathbf{u}_w} = \frac{\partial}{\partial \mathbf{u}_w}[-\mathbf{u}_o^T\mathbf{v}_c + \log \sum_{w \in Vocab} \exp(\mathbf{u}_w^T\mathbf{v}_c)]
$$

[1] 对数函数的复合函数求导

If $w = o$,

$$\begin{aligned}
\frac{\partial \mathcal{J}_{naive-softmax}(\mathbf{v}_c, o, \mathbf{U})}{\partial \mathbf{u}_w} &= -\mathbf{v}_c + \frac{1}{\sum_{w \in Vocab} \exp(\mathbf{u}_w^T \mathbf{v}_c)} \frac{\partial}{\partial \mathbf{u}_o} \sum_{w \in Vocab} \exp(\mathbf{u}_w^T \mathbf{v}_c) \\
&= -\mathbf{v}_c + \frac{1}{\sum_{w \in Vocab} \exp(\mathbf{u}_w^T \mathbf{v}_c)} \frac{\partial}{\partial \mathbf{u}_o} \exp(\mathbf{u}_o^T \mathbf{v}_c) \\
&= -\mathbf{v}_c + \frac{\exp(\mathbf{u}_o^T \mathbf{v}_c)}{\sum_{w \in Vocab} \exp(\mathbf{u}_w^T \mathbf{v}_c)} \mathbf{v}_c \\
&= [P(O = o | C = c) - 1] \mathbf{v}_c \\
&= (\hat{y}_w - y_w) \mathbf{v}_c \quad (w = o)
\end{aligned}$$

If $w \neq o$,

$$\begin{aligned}
\frac{\partial \mathcal{J}_{naive-softmax}(\mathbf{v}_c, o, \mathbf{U})}{\partial \mathbf{u}_w} &= \frac{\exp(\mathbf{u}_w^T \mathbf{v}_c)}{\sum_{w \in Vocab} \exp(\mathbf{u}_w^T \mathbf{v}_c)} \mathbf{v}_c \\
&= P(O = w | C = c) \mathbf{v}_c \\
&= (\hat{y}_w - y_w) \mathbf{v}_c \quad (w \neq o)
\end{aligned}$$

i.e.

$$\frac{\partial \mathcal{J}_{naive-softmax}(\mathbf{v}_c, o, \mathbf{U})}{\partial \mathbf{u}_w} = (\hat{y}_w - y_w) \mathbf{v}_c \in \mathbb{R}^{d \times 1}$$

5 Question (d)

$$\begin{aligned}
\frac{\partial \mathcal{J}_{naive-softmax}(\mathbf{v}_c, o, \mathbf{U})}{\partial \mathbf{U}} &= [\frac{\partial}{\partial \mathbf{u}_1}, \cdots, \frac{\partial}{\partial \mathbf{u}_n}] \mathcal{J}_{naive-softmax}(\mathbf{v}_c, o, \mathbf{U}) \\
&= [(\hat{y}_1 - y_1), \cdots, (\hat{y}_n - y_n)] \mathbf{v}_c \\
&= \mathbf{v}_c (\hat{\mathbf{y}} - \mathbf{y})^T \in \mathbb{R}^{d \times n}
\end{aligned}$$

6 Question (e)

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

7 Question (f)

$$\frac{\partial\mathcal{J}_{neg-sample}(\mathbf{v}_c,o,\mathbf{U})}{\partial\mathbf{v}_c} = -[1-\sigma(\mathbf{u}_o^T\mathbf{v}_c)]\mathbf{u}_o + \sum_{k=1}^{K}[1-\sigma(-\mathbf{u}_k^T\mathbf{v}_c)]\mathbf{u}_k$$

$$\frac{\partial\mathcal{J}_{neg-sample}(\mathbf{v}_c,o,\mathbf{U})}{\partial\mathbf{u}_o} = -[1-\sigma(\mathbf{u}_o^T\mathbf{v}_c)]\mathbf{v}_c$$

$$\frac{\partial\mathcal{J}_{neg-sample}(\mathbf{v}_c,o,\mathbf{U})}{\partial\mathbf{u}_k} = [1-\sigma(-\mathbf{u}_k^T\mathbf{v}_c)]\mathbf{v}_c$$

$\mathcal{J}_{naive-softmax}$ 里边包含 $\hat{\mathbf{y}}$，需要计算 $\mathbf{v}_c$ 和语料库中所有词向量的内积，$\mathcal{J}_{neg-sample}$ 只用计算相关的几个词向量，所以计算效率更高。

8 Question (g)

Assume $\mathbf{u}_i = \mathbf{u}_k$ when $i \in \{1,\cdots,m\}$, $\mathbf{u}_i \neq \mathbf{u}_k$ when $i \in \{m+1,\cdots,K\}, m \leq K$,

$$\begin{aligned}\frac{\partial\mathcal{J}_{neg-sample}(\mathbf{v}_c,o,\mathbf{U})}{\partial\mathbf{u}_k} &= \frac{\partial}{\partial\mathbf{u}_k}[-\sum_{k=1}^{m}\log(\sigma(-\mathbf{u}_k^T\mathbf{v}_c)) - \sum_{k=m+1}^{K}\log(\sigma(-\mathbf{u}_k^T\mathbf{v}_c))]\\ &= -\sum_{k=1}^{m}\frac{\sigma(-\mathbf{u}_k^T\mathbf{v}_c)[1-\sigma(-\mathbf{u}_k^T\mathbf{v}_c)]}{\sigma(-\mathbf{u}_k^T\mathbf{v}_c)}(-\mathbf{v}_c)\\ &= \sum_{k=1}^{m}[1-\sigma(-\mathbf{u}_k^T\mathbf{v}_c)]\mathbf{v}_c\end{aligned}$$

[1] 分式求导公式 $(\frac{u}{v})' = \frac{u'v-uv'}{v^2}$

9 Question (h)

9.1 (i)

$$\frac{\partial \mathcal{J}_{skip-gram}(\mathbf{v}_c, w_{t-m}, \cdots, w_{t+m}, \mathbf{U})}{\partial \mathbf{U}} = \sum_{-m \leq j \leq m, j \neq 0} \frac{\partial \mathcal{J}(\mathbf{v}_c, w_{t+j}, \mathbf{U})}{\partial \mathbf{U}}$$

9.2 (ii)

$$\frac{\partial \mathcal{J}_{skip-gram}(\mathbf{v}_c, w_{t-m}, \cdots, w_{t+m}, \mathbf{U})}{\partial \mathbf{v}_c} = \sum_{-m \leq j \leq m, j \neq 0} \frac{\partial \mathcal{J}(\mathbf{v}_c, w_{t+j}, \mathbf{U})}{\partial \mathbf{v}_c}$$

9.3 (iii)

$$\frac{\partial \mathcal{J}_{skip-gram}(\mathbf{v}_c, w_{t-m}, \cdots, w_{t+m}, \mathbf{U})}{\partial \mathbf{v}_w} = 0$$