CS224N-2021 Assignment 2:Word2vec

Jingwei Huang

2022-06-01

1 Written: Understanding word2vec

1.1 参数说明

在给定中心词 $C = c$，语料库（Vocab）和上下文窗口大小（context window）时：

- d：词向量的维度。
- n：词汇数量，$n = |Vocab|$。
- o：代表中心词 $C = c$ 上下文窗口中的词汇。
- w：代表预料库的任意单词，$w \in \text{Vocab}$。
- $\boldsymbol{y}$：$\boldsymbol{y} \in R^{|Vocab|}$，one-hot 向量，$y_w$ 代表单词 $W = w$ 和中心词 $C = c$ 的真实相关性，若该词出现在中心词的上下文中，则 $y_w = 1$，反之为 $y_w = 0$。
- $\hat{\boldsymbol{y}}$：$\hat{\boldsymbol{y}} \in R^{|Vocab|}$，$\hat{y}_w$ 代表模型对单词 $W = w$ 和中心词 $C = c$ 的相关性预测概率，$\hat{y_o} = P(O = o \mid C = c)$。
- $\boldsymbol{V}$：中心词矩阵，v_w 代表单词 w 作为中心词时的向量表示。
- $\boldsymbol{U}$：上下文矩阵，u_w 代表单词 w 作为上下文词时的向量表示。

1.2 Question (a)

当给定中心词 $C=c$ 和和上下文窗口大小（context window）时，那么就可确定该词的上下文词汇 O。由 $\boldsymbol{y}$ 的定义可知：

1）若 $w \notin O$，$y_w=0$ 进而使得 $y_w \log p(\hat{y}_w)=0$。

2）若 $w \in O$，$y_w=1$ 进而使得 $y_w \log p(\hat{y}_w)=\log p(\hat{y}_w)$。

因此

$$-\sum_{w \in Vocab} y_w \log\left(\hat{y}_w\right)=-\sum_{o \in O} \log\left(\hat{y}_o\right)=-\log\left(\hat{y}_o\right)$$

特别注意，$(\hat{y}_o)$ 代表中心词 $C=c$ 和上下文词汇 $O=O$ 成对计算的结果。

1.3 Question (b)

因为

$$\begin{aligned}
\boldsymbol{J}_{\text{naive-softmax}}\left(\boldsymbol{v}_c, o, \boldsymbol{U}\right) &=-\log P\left(O=o \mid C=c\right) \\
&=-\log \frac{\exp\left(\boldsymbol{u}_o^{\top} \boldsymbol{v}_c\right)}{\sum_{w \in \text{ Vocab }} \exp\left(\boldsymbol{u}_w^{\top} \boldsymbol{v}_c\right)} \\
&=-\boldsymbol{u}_o^{\top} \boldsymbol{v}_c+\log \sum_{w \in \text{ Vocab }} \exp\left(\boldsymbol{u}_w^{\top} \boldsymbol{v}_c\right)
\end{aligned}$$

所以根据链式求导法则，得

$$\begin{aligned}
\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}\left(\boldsymbol{v}_c, o, \boldsymbol{U}\right)}{\partial \boldsymbol{v}_c} &=-\boldsymbol{u}_o+\frac{\sum_{w \in \text{ Vocab }} \boldsymbol{u}_w \exp\left(\boldsymbol{u}_w^{\top} \boldsymbol{v}_c\right)}{\sum_{w \in \text{ Vocab }} \exp\left(\boldsymbol{u}_w^{\top} \boldsymbol{v}_c\right)} \\
&=-\boldsymbol{u}_o+\sum_{w \in \text{ Vocab }} \boldsymbol{u}_w \frac{\exp\left(\boldsymbol{u}_w^{\top} \boldsymbol{v}_c\right)}{\sum_{w \in \text{ Vocab }} \exp\left(\boldsymbol{u}_w^{\top} \boldsymbol{v}_c\right)} \\
&=-\boldsymbol{u}_o+\sum_{w \in \text{ Vocab }} \boldsymbol{u}_w P(O=w \mid C=c) \\
&=-\boldsymbol{u}_o+\sum_{w \in \text{ Vocab }} \boldsymbol{u}_w \hat{y}_w \\
&=\sum_{w \in \text{ Vocab }} \boldsymbol{u}_w \hat{y}_w-\boldsymbol{u}_w y_w \\
&=\sum_{w \in \text{ Vocab }} \boldsymbol{u}_w(\hat{y}_w-y_w) \\
&=\mathbf{U}(\hat{\mathbf{y}}-\mathbf{y}) \in \mathbb{R}^{d \times 1}
\end{aligned}$$

推导结果是 outside vectoru_w 的期望减去 u_w 的实际值，所以可以用 $\mathbf{U}(\hat{\mathbf{y}}-\mathbf{y}) \in \mathbb{R}^{d\times 1}$ 表示。

1.4 Question (c)

由上面推导过程知

$$\boldsymbol{J}_{\text{naive-softmax}}\left(\boldsymbol{v}_c, o, \boldsymbol{U}\right) = -\boldsymbol{u}_o^\top \boldsymbol{v}_c + \log \sum_{w\in \text{ Vocab}} \exp\left(\boldsymbol{u}_w^\top \boldsymbol{v}_c\right)$$

（1）当 $w=o$ 时，$\boldsymbol{u}_w=\boldsymbol{u}_o$ 和 $y_w=1$

$$\begin{aligned}
\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}\left(\boldsymbol{v}_c, o, \boldsymbol{U}\right)}{\partial \boldsymbol{u}_w} &= -\boldsymbol{v}_c + \frac{\boldsymbol{v}_c \exp\left(\boldsymbol{u}_o^\top \boldsymbol{v}_c\right)}{\sum_{w\in \text{ Vocab}} \exp\left(\boldsymbol{u}_w^\top \boldsymbol{v}_c\right)} \\
&= -\boldsymbol{v}_c + \boldsymbol{v}_c P(O=o \mid C=c) \\
&= -\boldsymbol{v}_c + \boldsymbol{v}_c \hat{y}_w \\
&= \boldsymbol{v}_c(\hat{y}_w - 1) \\
&= \boldsymbol{v}_c(\hat{y}_w - y_w)
\end{aligned}$$

（2）当 $w\neq o$ 时，$y_w=0$

$$\begin{aligned}
\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}\left(\boldsymbol{v}_c, o, \boldsymbol{U}\right)}{\partial \boldsymbol{u}_w} &= \frac{\boldsymbol{v}_c \exp\left(\boldsymbol{u}_w^\top \boldsymbol{v}_c\right)}{\sum_{w\in \text{ Vocab}} \exp\left(\boldsymbol{u}_w^\top \boldsymbol{v}_c\right)} \\
&= \boldsymbol{v}_c P(O=w \mid C=c) \\
&= \boldsymbol{v}_c \hat{y}_w \\
&= \boldsymbol{v}_c \hat{y}_w - \boldsymbol{v}_c y_w \\
&= \boldsymbol{v}_c(\hat{y}_w - y_w)
\end{aligned}$$

综上所述

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

1.5 Question (d)

因为 $\boldsymbol{U}=[\boldsymbol{u_1},...,\boldsymbol{u_w},...\boldsymbol{u_n}]$，其中 $n=|Vocab|$。根据上述推导过程，可得下面结论：

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

1.6 Question (e)

因为

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

所以

$$
\begin{aligned}
\frac{\mathrm{d} \sigma(x)}{\mathrm{d} x} &= \frac{e^x}{\left(e^x+1\right)^2}=\frac{e^x+1-1}{\left(e^x+1\right)^2}=\frac{1}{e^x+1}-\frac{1}{\left(e^x+1\right)^2} \\
&= \sigma(x)-(\sigma(x))^2=\sigma(x) \times(1-\sigma(x))
\end{aligned}
$$

1.7 Question (f)

由函数定义知

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

由链式求导法则可得，$\boldsymbol{J}_{\text{neg-sample }}\left(\boldsymbol{v}_c, o, \boldsymbol{U}\right)$ 对 $\boldsymbol{v}_c$ 的偏导数为

$$
\frac{\partial \boldsymbol{J}_{\text{neg-sample }}\left(\boldsymbol{v}_c, o, \boldsymbol{U}\right)}{\partial \boldsymbol{v}_c}=-\boldsymbol{u}_o\left(1-\sigma\left(\boldsymbol{u}_o^{\top} \boldsymbol{v}_c\right)\right)+\sum_{k=1}^K \boldsymbol{u}_k\left(1-\sigma\left(-\boldsymbol{u}_k^{\top} \boldsymbol{v}_c\right)\right)
$$

因为 u_k 是随机负采样词的向量表示，而 u_o 是 outside 词的向量表示，这两者之间没有联系。因此 $\boldsymbol{J}_{\text{neg-sample }}\left(\boldsymbol{v}_c, o, \boldsymbol{U}\right)$ 对 $\boldsymbol{u}_w$ 的偏导数需要进行分情况讨论，具体如下所示

（1）当 $w=o$ 时，$\boldsymbol{u}_k=\boldsymbol{u}_o$ 和

$$
\begin{aligned}
\frac{\partial \boldsymbol{J}_{\text{neg-sample }}\left(\boldsymbol{v}_c, o, \boldsymbol{U}\right)}{\partial \boldsymbol{u}_o} &= -\boldsymbol{v}_c\left(1-\sigma\left(\boldsymbol{u}_o^{\top} \boldsymbol{v}_c\right)\right) \\
&= -\left[1-\sigma\left(\mathbf{u}_o^T \mathbf{v}_c\right)\right] \mathbf{v}_c
\end{aligned}
$$

(2) 当 $w=k$ 时

$$\begin{aligned}
\frac{\partial \boldsymbol{J}_{\text{neg-sample }}\left(\boldsymbol{u}_{k}, o, \boldsymbol{U}\right)}{\partial \boldsymbol{u}_{k}} &=\sum_{k=1}^{K} \boldsymbol{v}_{c}\left(1-\sigma\left(-\boldsymbol{u}_{k}^{\top} \boldsymbol{v}_{c}\right)\right) \\
&=\left[1-\sigma\left(-\mathbf{u}_{k}^{T} \mathbf{v}_{c}\right)\right] \mathbf{v}_{c}
\end{aligned}$$

该方法使用随机负采样的方法，在语料库随机抽取 K 个单词来替代整个语料库。那么在计算中心词 $C=c$ 与其他单词的相关性概率时，改动前需要进行 $|Vocab|$ 次计算，改动后仅仅进行 K 次计算。

1.8 Question (g)

因为 $\boldsymbol{U}=[\boldsymbol{u_1},...,\boldsymbol{u_w},...\boldsymbol{u_K}]$，其中 $w\in|1,K|$。根据上述推导过程，可得下面结论：

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

1.9 Question (h)

(i) 对于中心词的上下文词汇

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

(ii) 损失函数对中心词向量的偏导数为

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

(iii) 因为 $\mathbf{v}_w$ 是，不参与以单词 $C=c$ 的损失函数。因此该损失函数对 $\mathbf{v}_w$ 的偏导数为 0。

$$\frac{\partial \mathcal{J}_{\text{skip-gram }}\left(\mathbf{v}_{c}, w_{t-m}, \cdots, w_{t+m}, \mathbf{U}\right)}{\partial \mathbf{v}_{w}}=0$$

2 Coding: Implementing word2vec

2.1 Question (a)

to do!

2.2 Question (b)

to do!

2.3 Question (c)

to do!