HW 2

Omer Cohen 308428127
Nimrod Admoni 203860721

November 2022

1

Integrating eq. 3 into eq. 4 gives:

$$\begin{aligned}\theta^* &= \arg\max_{\theta} \sum_{(x,y)\in D} \log \frac{e^{x_0 \cdot y_c}}{\sum_{z\in V} e^{x_0 \cdot z_c}} \\ &= \arg\max_{\theta} \sum_{(x,y)\in D} \left[x_0 \cdot y_c - \log \sum_{z\in V} e^{x_0 \cdot z_c} \right]\end{aligned}$$

2

Denote the objective as $L(x_0)$ we have:

$$\frac{\partial}{\partial x_0} L(x_0) = \sum_{(x,y)\in D} y_c - \frac{\sum_{z\in V} z_c \cdot e^{x_0 \cdot z_c}}{\sum_{z\in V} e^{x_0 \cdot z_c}}$$

3

3.1

Because every word is associated with two vectors in $\mathbb{R}^{500}$ and we have 500K words the number of parameters is:

$$N = 2 \cdot 500 \cdot 500k = 500,000,000$$

3.2

The model is not computationally feasible. We have 500M parameters and in each epoch we need to iterate over all parameters and update them all.

4

We want to model to see also negative examples - words that have low probability to be one after the other. The corpus only contain evidence for positive examples. Sampling words uniformly from the vocabulary results in two words that have low probability to be one after each other and this is why it is a good solution for negative examples.

5

Words that are highly similar are more likely to appear in a similar context, therefore they are more likely to have closer embedding. For example it is more likely that we can replace the word "car" with the word "vehicle" in a sentence and it will remain valid. Therefore "car" and "vehicle" are more likely the have closer vectors.