CS 224n Assignment 2: word2vec

Chrysa Dikonimaki

August 2020

1 Written: Understanding word2vec

1.a

Since y_w is 1 only for one word (the word that we represent as o):

$$-\sum_{w \in Vocab} y_w \log(\hat{y_w}) = -y_o \log(\hat{y_o}) = -\log(\hat{y_o})$$

1.b

$$\hat{y_o} = P(O = o|C = c) = \frac{exp(u_o^T v_c)}{\sum_{w \in Vocab} exp(u_w^T v_c)}$$

so :

$$\begin{aligned}
J_{naive-softmax} &= -\log(\hat{y_o}) \\
&= -\log \frac{exp(u_o^T v_c)}{\sum_{w \in Vocab} exp(u_w^T v_c)} \\
&= -u_o^T v_c + \log \sum_{w \in Vocab} exp(u_w^T v_c)
\end{aligned}$$

Thus:

$$\begin{aligned}
\frac{\partial J_{naive-softmax}}{\partial v_c} &= -u_o^T + \frac{\sum_{w \in Vocab} exp(u_w^T v_c) * u_w^T}{\sum_{w \in Vocab} exp(u_w^T v_c)} \\
&= -u_o + \sum_{w \in Vocab} P(O = w|C = c) * u_w \\
&= -u_o + U[P(O = w_1|C = c), P(O = w_2|C = c), ...] \\
&= -U * y + U * \hat{y} \\
&= U(\hat{y} - y)
\end{aligned}$$

1.c

1st case: w=o

$$\begin{aligned}\frac{\partial J_{naive-softmax}}{\partial u_o} &= -v_c + \frac{exp(u_o^T v_c) * v_c}{\sum_{w \in Vocab} exp(u_w^T v_c)} \\ &= -v_c + y^T \hat{y} v_c \\ &= v_c(y^T \hat{y} - 1)\end{aligned}$$

2nd case: $w \neq o$

$$\begin{aligned}\frac{\partial J_{naive-softmax}}{\partial u_w} &= -0 + \frac{exp(u_w^T v_c) * v_c}{\sum_{w \in Vocab} exp(u_w^T v_c)} \\ &= \hat{y}_w v_c\end{aligned}$$

1.d

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

1.e

$$\begin{aligned}
\frac{\partial J_{neg-sample}}{\partial v_c} &= -\frac{\sigma(-u_o^T v_c)\sigma(u_o^T v_c) * u_o^T}{\sigma(u_o^T v_c)} - \sum_{k=1}^{K} \frac{\sigma(u_k^T v_c)\sigma(-u_k^T v_c)(-u_k^T)}{\sigma(-u_k^T v_c)} \\
&= -\sigma(-u_o^T v_c) * u_o^T - \sum_{k=1}^{K} \sigma(u_k^T v_c)(-u_k^T) \\
&= -\sigma(-u_o^T v_c) * u_o^T + \sum_{k=1}^{K} \sigma(u_k^T v_c) u_k^T \\
&= -\sigma(-u_o^T v_c) * u_o + \sum_{k=1}^{K} \sigma(u_k^T v_c) u_k
\end{aligned}$$

$$\begin{aligned}
\frac{\partial J_{neg-sample}}{\partial u_o} &= -\frac{\sigma(-u_o^T v_c)\sigma(u_o^T v_c) * v_c}{\sigma(u_o^T v_c)} - 0 \\
&= -\sigma(-u_o^T v_c) v_c
\end{aligned}$$

$$\begin{aligned}
\frac{\partial J_{neg-sample}}{\partial u_k} &= -0 - \frac{\sigma(u_k^T v_c)\sigma(-u_k^T v_c)(-v_c)}{\sigma(-u_k^T v_c)} \\
&= \sigma(u_k^T v_c) v_c
\end{aligned}$$

1.f

$$\frac{\partial J_{skip-gram}}{\partial U} = \sum_{-m \le j \le m, j \ne 0} \frac{\partial J}{\partial U}$$

$$\frac{\partial J_{skip-gram}}{\partial v_c} = \sum_{-m \le j \le m, j \ne 0} \frac{\partial J}{\partial v_c}$$

$$\frac{\partial J_{skip-gram}}{\partial v_c} = 0$$

2 Coding: Implementing word2vec