Assignment 2

Nikolaos Smyrnioudis

August 26, 2020

1 Written

1.1 a

Since the true distibution y_w is a one hot encoding of the desired outside word, $y_w = 1$ iff $w = o$

1.2 b

Compute the partial derivative of $J_{naive-softmax}(v_c, o, U)$ with respect to v_c . Please write your answer in terms of y, ŷ, and U

$$
\begin{aligned}
J_{naive\_softmax}(v_c, o, U) &= -log(P(O = o|C = c)) \\
J_{naive\_softmax}(v_c, o, U) &= -(u_o^T v_c) + log(\sum_{w \in Vocab} exp(u_w^T v_c)) \\
\frac{\partial J}{\partial v_c} &= -u_o + \frac{1}{\sum_{w \in Vocab} exp(u_w^T v_c)} * \sum_{w \in Vocab} (exp(u_w^T v_c) * u_w)) \\
&= -u_o + \sum_{w \in Vocab} \frac{(exp(u_w^T v_c) * u_w))}{\sum_{w \in Vocab} exp(u_w^T v_c)} \\
&= -u_o + \sum_{w \in Vocab} (P(O = w|C = c) * u_w) \\
&= -U * y + U * \hat{y} \\
&= U * (\hat{y} - y) \\
P(O = o|C = c) &= \frac{exp(u_o^T v_c)}{\sum_{w \in Vocab} exp(u_w^T v_c)}
\end{aligned}
$$

1.3 c

Compute the partial derivative of $J_{naive-softmax}(v_c, o, U)$ with respect to each of the 'outside' word vectors, u_w 's. There will be two cases: when $w = o$, the true 'outside' word vector, and $w \neq o$, for all other words. Please write you answer in terms of y, ŷ, and v_c .

$$J_{naive\_softmax}(v_c, o, U) = -log(P(O = o|C = c))$$
$$J_{naive\_softmax}(v_c, o, U) = -(u_o^T v_c) + log(\sum_{w \in Vocab} exp(u_w^T v_c))$$

Case when $w = o$

$$\frac{\partial J}{\partial u_w} = -v_c + \frac{1}{\sum_{w \in Vocab} exp(u_w^T v_c)} * \exp(u_o^T v_c) * v_c$$
$$\frac{\partial J}{\partial u_w} = v_c * (\frac{1}{\sum_{w \in Vocab} exp(u_w^T v_c)} * \exp(u_o^T v_c) - 1)$$
$$\frac{\partial J}{\partial u_w} = v_c * (P(O = o|C = c) - 1)$$
$$\frac{\partial J}{\partial u_w} = v_c * (y^T \hat{y} - 1)$$

Case when $w \neq o$

$$\frac{\partial J}{\partial u_w} = \frac{1}{\sum_{w \in Vocab} exp(u_w^T v_c)} * \exp(u_w^T v_c) * v_c$$
$$\frac{\partial J}{\partial u_w} = P(O = w|C = c) * v_c$$
$$\frac{\partial J}{\partial u_w} = \hat{y_w} * v_c$$

1.4 d

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

1.5 e

$$
\begin{aligned}
J_{neg\_sample}(v_c, o, U) &= -log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} log(\sigma(-u_k^T v_c)) \\
\frac{\partial J}{\partial v_c} &= -\frac{1}{\sigma(u_o^T v_c)} * \sigma(u_o^T v_c)\sigma(-u_o^T v_c) * u_o \\
&\quad + \sum_{k=1}^{K} \frac{1}{\sigma(-u_k^T v_c)} * \sigma(-u_k^T v_c) * \sigma(u_k^T v_c) u_k \\
&= -\sigma(-u_o^T v_c) * u_o + \sum_{k=1}^{K} \sigma(u_k^T v_c) u_k \\
\frac{\partial J}{\partial u_o} &= -\frac{1}{\sigma(u_o^T v_c)} * \sigma(u_o^T v_c)\sigma(-u_o^T v_c) * v_c \\
&= -\sigma(-u_o^T v_c) * v_c \\
\frac{\partial J}{\partial u_k} &= \sigma(u_k^T v_c) v_c
\end{aligned}
$$

1.6 f

$$
J_{skip-gram}(u_c, w_{t-m}, \ldots, w_{t+m}, U) = \sum_{-m \le j \le m; j \ne 0} J(v_c, w_{t+j}, U)
$$

1.6.1 i

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

1.6.2 ii

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

1.6.3 iii

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

2 Results

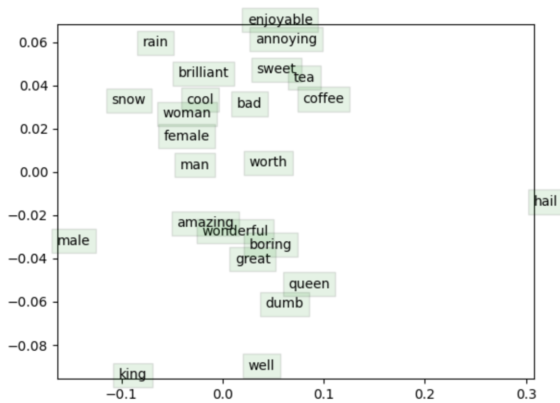

Figure 1: A plot of the resulting word vectors

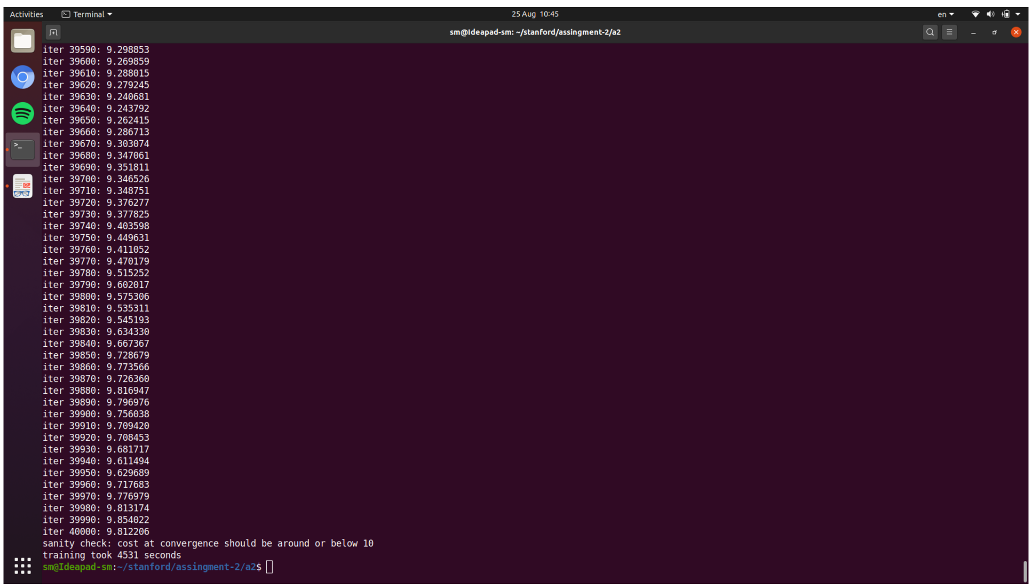

Figure 2: The time needed to train and output loss