CS 224n Assignment #2: word2vec

Jai Gupta

January 21, 2019

Q (1.a): Show that the naive-softmax loss given in Equation (2) is the same as the cross-entropy loss between y and $\hat{y}$; i.e., show that

$$-\sum_{w \epsilon Vocab} y_w log(\hat{y}_w) = -log((\hat{y}_o) \tag{0.1}$$

Solution:

$$-\sum_{w \epsilon Vocab} y_w log(\hat{y}_w) = -(1 * log(\hat{y}_o) + \sum_{w \epsilon Vocab - o} 0 * y_w log(\hat{y}_w)) = -log(\hat{y}_o) \tag{0.2}$$

Q (1.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, $\hat{y}$, and U.
Solution:

$$J_{naive-softmax}(v_c, o, U) = -log(\hat{y}_o)$$

where

$$\begin{aligned} \hat{y}_o &= P(O = o|C = c) \\ &= \frac{exp(u_o^T v_c)}{\sum_{w \epsilon Vocab} exp(u_w^T v_c)} \end{aligned}$$

Taking derivative on both sides, we get:

$$\frac{\partial J}{\partial v_c} = -\frac{1}{\hat{y}_o} * \frac{\partial \hat{y}_o}{\partial v_c}$$

$$
\begin{aligned}
\frac{\partial \hat{y}_o}{\partial v_c} &= \frac{\partial P(O=o|C=c)}{\partial v_c} \\
&= \frac{\partial exp(u_o^T v_c)}{\partial v_c} * \frac{1}{\sum_{w\epsilon Vocab} exp(u_w^T v_c)} + exp(u_o^T v_c) * \frac{\partial (\sum_{w\epsilon Vocab} exp(u_w^T v_c))^{-1}}{\partial v_c} \\
&= exp(u_o^T v_c) * u_o * \frac{1}{\sum_{w\epsilon Vocab} exp(u_w^T v_c)} + exp(u_o^T v_c) * \frac{-\sum_{w\epsilon Vocab} u_w * exp(u_w^T v_c)}{(\sum_{w\epsilon Vocab} exp(u_w^T v_c))^2} \\
&= \hat{y}_o u_o - \hat{y}_o \frac{\sum_{w\epsilon Vocab} u_w * exp(u_w^T v_c)}{\sum_{w\epsilon Vocab} exp(u_w^T v_c)} \\
&= \hat{y}_o u_o - \hat{y}_o \frac{\sum_{w\epsilon Vocab} u_w * exp(u_w^T v_c)}{\sum_{w\epsilon Vocab} exp(u_w^T v_c)} \\
&= \hat{y}_o U^T y - \hat{y}_o U^T * \hat{y} \\
&= \hat{y}_o U^T (y - \hat{y}) \\
\frac{\partial J}{\partial v_c} &= -\frac{1}{\hat{y}_o} * \frac{\partial \hat{y}_o}{\partial v_c} \\
&= -\frac{1}{\hat{y}_o} * \hat{y}_o U^T (y - \hat{y}) \\
&= U^T (\hat{y} - y)
\end{aligned}
$$

Q (1.c): Compute the partial derivatives 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, $\hat{y}$, and v_c.
Solution:

$$J_{naive-softmax}(v_c, o, U) = -log(\hat{y}_o)$$

where

$$
\begin{aligned}
\hat{y}_o &= P(O=o|C=c) \\
&= \frac{exp(u_o^T v_c)}{\sum_{w\epsilon Vocab} exp(u_w^T v_c)}
\end{aligned}
$$

Taking derivative on both sides with respect to u_w, we get:

$$
\begin{aligned}
\frac{\partial \hat{y}_o}{\partial u_w} &= \frac{\partial P(O=o|C=c)}{\partial u_w} \\
&= \frac{1}{\sum_{x\epsilon Vocab} exp(u_x^T v_c)} * \frac{\partial exp(u_o^T v_c)}{\partial u_w} + exp(u_o^T v_c) * \frac{\partial (\sum_{x\epsilon Vocab} exp(u_x^T v_c))^{-1}}{\partial u_w}
\end{aligned}
$$

Case #1: when $w \neq o$

$$
\begin{aligned}
\frac{\partial \hat{y}_o}{\partial u_w} &= 0 + exp(u_o^T v_c) * \frac{\partial(\sum_{x\epsilon Vocab} exp(u_x^T v_c))^{-1}}{\partial u_w} \\
&= -\frac{exp(u_o^T v_c)}{\sum_{x\epsilon Vocab} exp(u_x^T v_c))^2} exp(u_w^T v_c) v_c \\
&= -v_c \frac{exp(u_o^T v_c)}{\sum_{x\epsilon Vocab} exp(u_x^T v_c)} \frac{exp(u_w^T v_c)}{\sum_{x\epsilon Vocab} exp(u_x^T v_c)} \\
&= -v_c \hat{y}_o \hat{y}_w \\
\frac{\partial J}{\partial u_w} &= \frac{-1}{\hat{y}_o} \frac{\partial \hat{y}_o}{\partial u_w} \\
&= \frac{v_c \hat{y}_o \hat{y}_w}{\hat{y}_o} \\
&= v_c \hat{y}_w
\end{aligned}
$$

Case #2: when $w = o$

$$
\begin{aligned}
\frac{\partial \hat{y}_o}{\partial u_o} &= \frac{1}{\sum_{x\epsilon Vocab} exp(u_x^T v_c)} * \frac{\partial exp(u_o^T v_c)}{\partial u_o} + exp(u_o^T v_c) * \frac{\partial(\sum_{x\epsilon Vocab} exp(u_x^T v_c))^{-1}}{\partial u_o} \\
&= v_c \frac{exp(u_o^T v_c)}{\sum_{x\epsilon Vocab} exp(u_x^T v_c)} - \frac{exp(u_o^T v_c)}{(\sum_{x\epsilon Vocab} exp(u_x^T v_c))^2} exp(u_o^T v_c) v_c \\
&= v_c \hat{y}_o - v_c \hat{y}_o^2 \\
&= v_c \hat{y}_o (1 - \hat{y}_o) \\
\frac{\partial J}{\partial u_o} &= \frac{-1}{\hat{y}_o} \frac{\partial \hat{y}_o}{\partial u_o} \\
&= \frac{-v_c \hat{y}_o (1 - \hat{y}_o)}{\hat{y}_o} \\
&= v_c (\hat{y}_o - 1)
\end{aligned}
$$

From the above two cases,

$$
\frac{\partial J}{\partial u_w} = \begin{cases} v_c(\hat{y}_w - 1), & \text{if } w = o \\ v_c \hat{y}_w, & \text{otherwise} \end{cases}
$$

This can in short be written as:

$$
\frac{\partial J}{\partial u_w} = v_c(\hat{y}_w - y_w)
$$

Hence,

$$
\frac{\partial J}{\partial U} = (\hat{y} - y)^T v_c
$$

Q (1.d): Derivative of Sigmoid function.
Solution:
$\sigma(x)$ is a element wise function if x is a vector. Differentiating w.r.t x, we get a Jacobian(J). J(i, j) would be $\frac{\partial \sigma(x_i)}{\partial x_j}$ where i represents the row index and j represents the column index.

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

Normally, in such cases, components of the given vector are independent to each other. In such cases:

$$
\frac{\partial x_i}{\partial x_j} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases}
$$

So for such vectors, the Jacobian will be a diagonal matrix where ith diagonal element will be $\sigma(x_i)(1-\sigma(x_i))$

Q (1.e): Negative Sampling Softmax Loss.
Solution:

$$
J_{neg-sample}(v_c, o, U) = -log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} log(\sigma(-u_k^T v_c))
$$

Part (b) Derivative with respect v_c

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

Part (b) Derivative with respect u_w
Case #1: $w = o$

$$
\begin{aligned}
\frac{\partial J_{neg-sample}}{\partial u_o} &= -\frac{\sigma(u_o^T v_c)(1-\sigma(u_o^T v_c))v_c}{\sigma(u_o^T v_c)} + 0 \\
&= -(1-\sigma(u_o^T v_c))v_c \\
&= (1-\sigma(u_o^T v_c))v_c \\
&= \sigma(-u_o^T v_c)v_c
\end{aligned}
$$

Case #1: $w \neq o$

$$
\begin{aligned}
\frac{\partial J_{neg-sample}}{\partial u_w} &= 0 - \frac{\sigma(-u_w^T v_c)(1-\sigma(-u_w^T v_c))(-v_c)}{\sigma(-u_w^T v_c)} \\
&= -(1-\sigma(-u_w^T v_c))(-v_c) \\
&= (1-\sigma(-u_w^T v_c))\, v_c \\
&= \sigma(u_w^T v_c)\, v_c
\end{aligned}
$$

Q (1.f): Derivatives for skip-gram model.
(i):

$$
\frac{\partial J_{skip-gram} m(v_c, w_{t-m}, \ldots w_{t+m}, U)}{\partial U} = - \sum_{\substack{-m<=j<=m \\ j\neq 0}} \frac{\partial J(v_c, w_{t+j}, U)}{\partial U}
$$

(ii):

$$
\frac{\partial J_{skip-gram} m(v_c, w_{t-m}, \ldots w_{t+m}, U)}{\partial v_c} = - \sum_{\substack{-m<=j<=m \\ j\neq 0}} \frac{\partial J(v_c, w_{t+j}, U)}{\partial v_c}
$$

(iii):
Since the loss function depends on only v_c and no other center word vector, derivative with respect to v_w, where $w \neq c$, will be 0.

$$
\frac{\partial J_{skip-gram} m(v_c, w_{t-m}, \ldots w_{t+m}, U)}{\partial v_w} = 0
$$

Q (2.c):

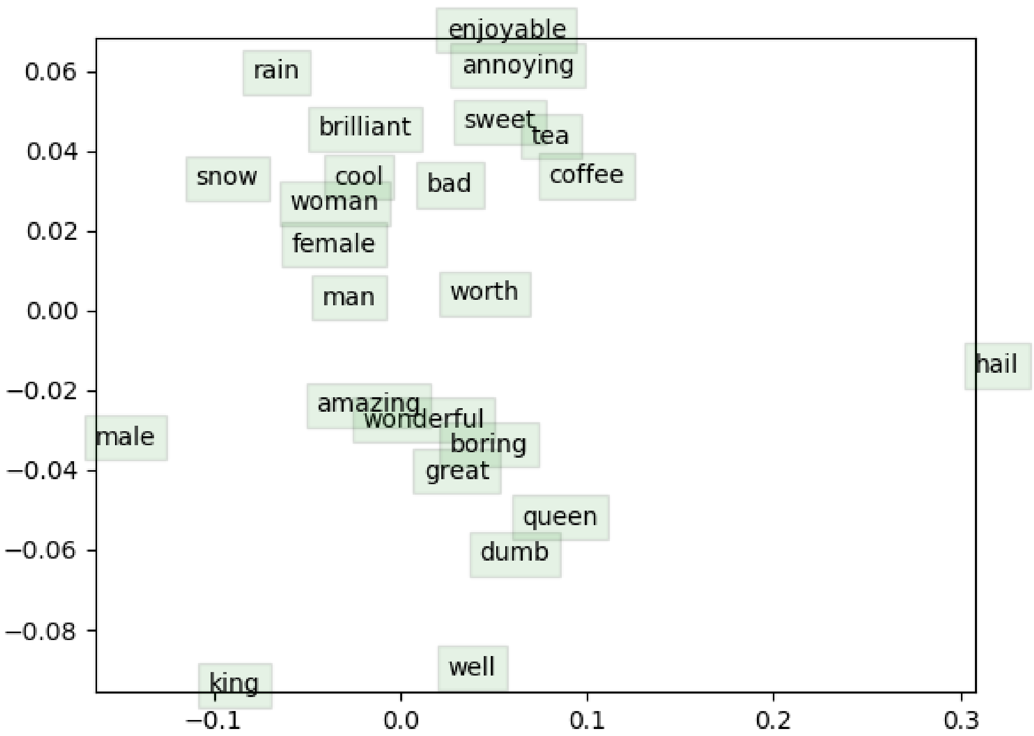

Figure 0.1: Word Vector Visualization