cs224n-2019

Falco Winkler

June 2020

These are the answers to the second assignment of the course with very detailed steps.

(a)

We know that y_w is a one hot vector that is 1 iff $w = o$ (it is 1 for the outside word).

$$y = \begin{cases} 1 & w = o \\ 0 & otherwise \end{cases}$$

$$-\sum_{w \in Vocab}^{V} y_w log(\hat{y_w}) = -(y_{w1} log(\hat{y}_{w1}) + ... + y_{wo} log(\hat{y}_{wo}) + ... + y_{wv} log(\hat{y}_{wv})) =$$

$$-(0 log(\hat{y}_{w1}) + ... + 1 log(\hat{y}_{wo}) + ... + 0 log(\hat{y}_{wv})) =$$

$$-log(\hat{y}_{wo})$$

(b)

We know from the lecture that

$$\frac{\delta}{\delta v_c} J(v_c, o, U) = -(u_o - \sum_{x=1}^{V} p(x|c) u_x)$$

$\hat{y}$ is the vector of probabilities of x to be an outside word of a given c. So we can rewrite:

$$-(u_o - \sum_{x=1}^{V} \hat{y}_x u_x) =$$

$$-u_o + \hat{y}_1 u_1 + ... + \hat{y}_o u_o + ... + \hat{y}_V u_V =$$

$$\hat{y}_1 u_1 + ... + \hat{y}_o u_o - u_o + ... + \hat{y}_V u_V =$$

$$\hat{y}_1 u_1 + ... + (\hat{y}_o - 1) u_o + ... + \hat{y}_V u_V =$$

$$(\hat{y}_1 - 0) u_1 + ... + (\hat{y}_o - 1) u_o + ... + (\hat{y}_V - 0) u_V =$$

$$(\hat{y} - y) U$$

(c)

In the case of $w = o$

$$\frac{\delta}{\delta u_o} - logP(x|o) =$$

$$-(\frac{\delta}{\delta u_o} logP(x|o)) =$$

$$-(\frac{\delta}{\delta u_o} log \frac{e^{u_o^T v_c}}{\sum_{w \in Vocab}^{V} e^{u_w^T v_c}}) =$$

$$-(\frac{\delta}{\delta u_o} log(e^{u_o^T v_c}) - \frac{\delta}{\delta u_o} log(\sum_{w \in Vocab}^{V} e^{u_w^T v_c})) =$$

$$-(v_c - \frac{1}{\sum_{w \in Vocab}^{V} e^{u_w^T v_c}} \frac{\delta}{\delta u_o} (\sum_{w \in Vocab}^{V} e^{u_w^T v_c})) =$$

$$-(v_c - \frac{1}{\sum_{w \in Vocab}^{V} e^{u_w^T v_c}} e^{u_o^T v_c} v_c)) =$$

$$-(v_c - \frac{e^{u_o^T v_c}}{\sum_{w \in Vocab}^{V} e^{u_w^T v_c}} v_c)) =$$

$$-(v_c - P(O = o|C = c)v_c) =$$

$$P(O = o|C = c)v_c - v_c =$$

$$(P(O = o|C = c) - 1)v_c$$

If $w = x$ for any word $x \in Vocab$ and $x \neq o$:

$$-(\frac{\delta}{\delta u_x} log \frac{e^{u_o^T v_c}}{\sum_{w \in Vocab}^{V} e^{u_w^T v_c}}) =$$

$$-(\frac{\delta}{\delta u_x} log(e^{u_o^T v_c}) - \frac{\delta}{\delta u_x} log(\sum_{w \in Vocab}^{V} e^{u_w^T v_c})) =$$

$$-0 + \frac{\delta}{\delta u_x} log(\sum_{w \in Vocab}^{V} e^{u_w^T v_c}) =$$

$$\frac{1}{\sum_{w \in Vocab}^{V} e^{u_w^T v_c}} \frac{\delta}{\delta u_x} (\sum_{w \in Vocab}^{V} e^{u_w^T v_c}) =$$

$$\frac{e^{u_x^T v_c}}{\sum_{w \in Vocab}^{V} e^{u_w^T v_c}} v_c =$$

$$P(O = x|C = c)v_c$$

So, if we derive with respect to the outside-word vector of the actual outside word o we are looking at, the gradient is the (probability - 1) multiplied with every vector entry of the word vector, and in the other case it is simply the probability multiplied with every vector entry. That means, if we want the gradient for U, we take each of those probabilities and multiply it element-wise with the representation of the current center word v_c. Written in vector form, that is

$$\frac{\delta}{\delta U} J_{naive-softmax(v_c,o,U)} = (\hat{y} - y)^T v_c$$

(d)

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

$$\frac{\delta}{\delta x}\frac{e^x}{1+e^x} = \frac{e^x(e^x+1) - e^x e^x}{(e^x+1)^2} =$$

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

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

$$\frac{e^{-x}e^x + e^{-x} - e^x e^{-x}}{e^{-x}e^x + e^{-x}} = \tag{1}$$

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

(e) (1.)

$$\frac{\delta}{\delta v_c} - log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} log(\sigma(-u_k^T v_c)) =$$

$$-\frac{\delta}{\delta v_c} log(\sigma(u_o^T v_c)) - \frac{\delta}{\delta v_c}\sum_{k=1}^{K} log(\sigma(-u_k^T v_c)) \overset{2}{=}$$

$$-\frac{(1-\sigma(u_o^T v_c))\sigma(u_o^T v_c)}{\sigma(u_o^T v_c)}\frac{\delta u_o^T v_c}{\delta v_c} - \sum_{k=1}^{K}\frac{(1-\sigma(-u_k^T v_c))\sigma(-u_k^T v_c)}{\sigma(-u_k^T v_c)}\frac{\delta - u_k^T v_c}{\delta v_c} =$$

$$-(1-\sigma(u_o^T v_c))u_o - \sum_{k=1}^{K}(1-\sigma(-u_k^T v_c))(-u_k) =$$

$$-(1-\sigma(u_o^T v_c))u_o + \sum_{k=1}^{K}(1-\sigma(-u_k^T v_c))u_k$$

$$k(x) = h(f(g(x))) \iff k'(x) = h'(f(g(x)))f'(g(x))g'(x) \tag{2}$$

(e) (2.)

$$\frac{\delta}{\delta u_o} - log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} log(\sigma(-u_k^T v_c)) =$$

$$\frac{\delta}{\delta u_o} - log(\sigma(u_o^T v_c)) =$$

$$-\frac{(1-\sigma(u_o^T v_c))\sigma(u_o^T v_c)}{\sigma(u_o^T v_c)} \frac{\delta u_o^T v_c}{\delta u_o} =$$

$$-(1-\sigma(u_o^T v_c))v_)$$

(e) (3.)

$$\frac{\delta}{\delta u_k} - log(\sigma(u_o^T v_c)) - \sum_{k=1}^{K} log(\sigma(-u_k^T v_c)) =$$

$$-\frac{\delta}{\delta u_k} log(\sigma(-u_k^T v_c)) =$$

$$-(1-\sigma(-u_k^T v_c))(-v_c)$$

$$(1-\sigma(-u_k^T v_c))v_c$$

The loss function is easier to compute, because the sum only computed on a subsample of size K (negative words) and not the whole vocabulary.

(f)

This is indeed very simple: We simply have to sum the derivatives of J for the w that are in the window size.

(f) (1.)

$$\frac{\delta J_{skip-gram}(v_c, w_{t-m}, ..., w_{t+m}, U)}{\delta U} =$$

$$\sum_{-m<=j<=m, j!=0} \frac{\delta}{\delta U} J(v_c, w_t + j, U)$$

(f) (2.)

$$\frac{\delta J_{skip-gram}(v_c, w_{t-m}, ..., w_{t+m}, U)}{\delta v_c} =$$

$$\sum_{-m<=j<=m, j!=0} \frac{\delta}{\delta v_c} J(v_c, w_t + j, U)$$

(f) (2.) $x! = c$

$$\frac{\delta J_{skip-gram}(v_c, w_{t-m}, ..., w_{t+m}, U)}{\delta v_x} = 0$$

If we derive with respect to a center vector v_x where x is not the center word c, we can observe that this vector is not contained in the expression of J, therefore the derivative is zero.