A2 Solution

Alok Regmi*

June 11, 2021

Contents

Q.1.a

Show that naive-softmax loss is the same as the cross-entropy loss between y and $\hat{y}$.

Answer

We know that y is one hot vector with only one 1 for true outside word. Thus,

$$-\sum_{w \in vocab} y_w log(\hat{y}_w)$$

becomes :

$$-[0log(\hat{y}_1) + ... + 1log(\hat{y}_0), ..., 0log(\hat{y}_{|V|})] = -log(\hat{y}_0)$$

Q.1.b

Compute the partial derivative of $J_{naive-softmax}(uc, o, U) = -logP(O = o|C = c)$ w.r.t. v_c.

*sagar.r.alok@gmail.com

Answer

Given : $-\log P(O=o \mid C=c) = -\log \frac{e^{u_o^T v_c}}{\sum_{w \in Vocab} e^{u_w^T v_c}}$, then we have :

$$\begin{aligned}
\frac{\partial J}{\partial v_c} &= -u_o + \frac{1}{\sum_{w \in vocab} e^{u_w^T v_c}} \sum_{w \in Vocab} \left(e^{u_w^T v_c} u_w \right) \\
&= -u_o + \sum_{w \in Vocab} \left(\frac{e^{u_w^T v_c}}{\sum_{w \in vocab} e^{u_w^T v_c}} u_w \right) \\
&= -u_o + \sum_{w \in vocab} (P(u_w|v_c) u_w) \\
&= -u_o + \textstyle\sum_{w \in vocab} (\hat{y}_w u_w)
\end{aligned}$$

Q.1.c

Compute partial derivative of $J_{naive-softmax}$ with repect to each of outside word vectors u_w's. Two cases when w = 0 and w $\neq$ 0.

Answer

1. When w = 0,

$$\begin{aligned}
\frac{\partial J}{\partial u_{w=o}} &= -v_c + \frac{1}{\sum_{w \in vocab} e^{u_w^T v_c}} e^{u_o^T v_c} v_c \\
&= v_c(\hat{y_o} - 1)
\end{aligned}$$

2. When w $\neq$ 0,

$$\begin{aligned}
\frac{\partial J}{\partial u_{w \neq o}} &= \frac{1}{\sum_{w \in vocab} e^{u_w^T v_c}} e^{u_o^T v_c} v_c \\
&= v_c(\hat{y_{w \neq 0}})
\end{aligned}$$

Q.1.d

Compute the partial derivative of $J_{naive-softmax}$ with respect to U.

answer

Since, the vector is one-hot encoded vector, we get this simple representation:

$$\frac{\partial J}{\partial U} = \left[\frac{\partial J}{\partial U_1}, \frac{\partial J}{\partial U_2}, ..., \frac{\partial J}{\partial U_{|vocab|}} \right]$$

Q.1.e

Partial Derivative of sigmoid function.

answer

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

Then, derivative of sigmoid is:

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

Q.1.f

Compute the derivative of $J_{neg-sample}$ with respect to v_c, u_o, u_k.

Answer

Given:

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

Now,

$$\begin{aligned}\frac{\partial J}{\partial v_c} &= \frac{-1}{\sigma(u_o^T v_c)}\sigma^{'}(u_o^T v_c)u_o - \sum_k \frac{\sigma^{'}(-u_k^T v_c)}{\sigma(-u_k^T v_c)}(-u_k) \\ &= (\sigma(u_o^T v_c) - 1)u_o + \sum_k \big(1 - \sigma(-u_k^T v_c))u_k\big)\end{aligned}$$

$$\begin{aligned}\frac{\partial J}{\partial u_o} &= \frac{-1}{\sigma(u_o^T v_c)}\sigma^{'}(u_o^T v_c)v_c \\ &= (\sigma(u_o^T v_c) - 1)v_c\end{aligned}$$

$$\begin{aligned}\frac{\partial J}{\partial u_k} &= -\sum_k \frac{\sigma^{'}(-u_k^T v_c)}{\sigma(-u_k^T v_c)}(-v_c) \\ &= \big(1 - \sigma(-u_k^T v_c))v_c\big)\end{aligned}$$

Q.1.g

Without the assumption that K negative samples are distinct, find derivative of $J_{neg-sample}$ w.r.t. u_k.

Answer

In our previous example, when derivative w.r.t. u_k , we had the sum term gone since each sample was independent of another thus derivative w.r.t other samples will be 0 for each sample. Now, in this case, we can't assume that is the case.

$$
\begin{aligned}
\frac{\partial J}{\partial u_k} &= -\sum_{k} \frac{\sigma^{'}(-u_k^T v_c)}{\sigma(-u_k^T v_c)}(-v_c) \\
&= \sum_{j=k} \big(1 - \sigma(-u_j{}^T v_c))v_c\big)
\end{aligned}
$$

Q.1.h

Now, for skip gram with context window, find three partial derivatives.

Answer

Given:

$$
J_{\text{skip-gram}}\ (\boldsymbol{v}_c, w_{t-m}, \ldots w_{t+m}, \boldsymbol{U}) = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \boldsymbol{J}\,(\boldsymbol{v}_c, w_{t+j}, \boldsymbol{U})
$$

Now, iii. When w $\neq$ c: In this case, the derivative will be equal to 0.