CS 224n Assignment #2: word2vec

Intae Jun

1 Written: Understanding word2vec (31 points)

(a) (2 points) Prove that the naive-softmax loss (Eq. 2) is the same as the cross-entropy loss between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$.

Since the true label $\boldsymbol{y}_w$ is a one-hot vector, all the terms where $w \neq o$ are reduced and only the term where $w = o$ remains.

(b) (7 points)

(i) Compute the partial derivative of $\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})$ with respect to $\boldsymbol{v}_c$.

Please write your answer in terms of $\boldsymbol{y}, \hat{\boldsymbol{y}}, \boldsymbol{U}$, and show your work toreceive full credit.

$$
\begin{aligned}
\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{v}_c} &= -\frac{\partial \log P(O = o|C = c)}{\partial \boldsymbol{v}_c} \\
&= -\frac{1}{P(O = o|C = c)} \frac{\partial P(O = o|C = c)}{\partial \boldsymbol{v}_c} \\
&= -\frac{\boldsymbol{u}_o \exp(\boldsymbol{u}_o^T \boldsymbol{v}_c) \sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^T \boldsymbol{v}_c) - \exp(\boldsymbol{u}_o^T \boldsymbol{v}_c) \sum_{w \in \text{Vocab}} \boldsymbol{u}_w \exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)}{P(O = o|C = c)(\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^T \boldsymbol{v}_c))^2} \\
&= -\frac{\boldsymbol{u}_o \exp(\boldsymbol{u}_o^T \boldsymbol{v}_c) \sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^T \boldsymbol{v}_c) - \exp(\boldsymbol{u}_o^T \boldsymbol{v}_c) \sum_{w \in \text{Vocab}} \boldsymbol{u}_w \exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)}{\exp(\boldsymbol{u}_o^T \boldsymbol{v}_c)(\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^T \boldsymbol{v}_c))} \\
&= -\frac{\boldsymbol{u}_o \sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^T \boldsymbol{v}_c) - \sum_{w \in \text{Vocab}} \boldsymbol{u}_w \exp(\boldsymbol{u}_w^T \boldsymbol{v}_c))}{\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)} \\
&= -\boldsymbol{u}_o + \sum_{w \in \text{Vocab}} P(O = w|C = c)\boldsymbol{u}_w \\
&= \boldsymbol{U}^T(\hat{\boldsymbol{y}} - \boldsymbol{y}).
\end{aligned}
$$

(ii) When $\hat{\boldsymbol{y}} - \boldsymbol{y} = 0$, the gradient becomes zero.

(iii) As the gradient is subtracted from the word vector $\boldsymbol{v}_c$, the vector is updated as follows:

$$\boldsymbol{v}_c^{(i+1)} \leftarrow \boldsymbol{v}_c^{(i)} \eta \boldsymbol{U}^T(\hat{\boldsymbol{y}} - \boldsymbol{y}),$$

where i is the index for the iteration number and eta is the learning rate for the gradient descent method. By subtracting this gradient, we are moving the center word vector $\boldsymbol{v}_c$ towards the correct output vector $\boldsymbol{u}_o$ and away from the incorrect predictoins $\boldsymbol{U}^T\hat{\boldsymbol{y}}$.

(iv) For two different words $x \neq y$, if, for some scalar α, they have a relationship such as $\boldsymbol{u}_x = \alpha \boldsymbol{u}_y$, the L2 normalized vectors $\boldsymbol{u}'_x$ and $\boldsymbol{u}'_y$ are not discriminated since L2 normalization makes every vector have the same magnitude. However the primary information contained in the vectors, such as representation of each word does not disappear. Also, keeping consistency in vector scale are recommended for some models that are sensitive to the scale of input features.

(c) (5 points) Compute the partial derivatives of $\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})$ with respect to each of the 'outside' word vectors, $\boldsymbol{u}_w$'s. There will be two cases: when $w = o$, the true 'outside' word vector, and $w \neq o$, for all other words. Please write your answer in terms of $\boldsymbol{y}, \hat{\boldsymbol{y}}$, and $\boldsymbol{v}_c$. In this subpart, you may use specific elements within these terms as well (such as $\boldsymbol{y}_1, \boldsymbol{y}_2, ...$).

$$\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) = -\log P(O = o|C = c) = -\log \frac{\exp(\boldsymbol{u}_o^T \boldsymbol{v}_c)}{\sum_w \exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)} = -\boldsymbol{u}_o^T \boldsymbol{v}_c + \log \sum_w \exp(\boldsymbol{u}_w^T \boldsymbol{v}_c).$$

If $w = o$ (i.e. the outside word is the answer):

$$\begin{aligned}
\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{u}_w} &= -\boldsymbol{v}_c + \frac{\partial}{\partial \boldsymbol{u}_w}(\log \sum_w \exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)) \\
&= -\boldsymbol{v}_c + \frac{\boldsymbol{v}_c \sum_w \exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)}{\log \sum_w \exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)} \\
&= -\boldsymbol{v}_c + P(O = o|C = c)\boldsymbol{v}_c \\
&= (\hat{\boldsymbol{y}} - \boldsymbol{y})^T \boldsymbol{v}_c.
\end{aligned}$$

If $w \neq o$ (i.e. the outside word is not the answer):

$$\begin{aligned}
\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{u}_w} &= -\frac{\partial}{\partial \boldsymbol{u}_w}(\boldsymbol{u}_o^T \boldsymbol{v}_c) + \frac{\partial}{\partial \boldsymbol{u}_w}(\log \sum_w \exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)) \\
&= 0 + \frac{\boldsymbol{v}_c \sum_w \exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)}{\log \sum_w \exp(\boldsymbol{u}_w^T \boldsymbol{v}_c)} \\
&= P(O = w|C = c)\boldsymbol{v}_c \\
&= \hat{\boldsymbol{y}}^T \boldsymbol{v}_c \\
&= (\hat{\boldsymbol{y}} - \boldsymbol{y})^T \boldsymbol{v}_c.
\end{aligned}$$

(d) (1 point) Write down the partial derivative of $\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})$ with respect to $\boldsymbol{U}$.

.

(e) (2 points) Compute the derivate of $f(x)$ with respect to x.

If $x > 0$, $f(x) = x$ and $f'(x) = 1$. If $x < 0$, then $f(x) = \alpha x$ and $f'(x) = \alpha$.

In summary, we have that

$$\frac{d}{dx}f(x) = \begin{cases} x & (x > 0), \\ \alpha & (x < 0). \end{cases}$$

(f) (3 points) Compute the derivative of $\sigma(x)$ with respect to x, where x is a scalar.

$$\begin{aligned} \frac{d}{dx}\left(\frac{e^x}{e^x+1}\right) &= \frac{e^x(e^x+1) - e^x(e^x)}{(e^x+1)^2} \\ &= \frac{e^x}{(e^x+1)^2} \\ &= \sigma(x)(1-\sigma(x)). \end{aligned}$$

(g) (6 points)

(i) .

(h) (2 points)

(i) .