

CS224n Winter 2019 Homework 2: word2vec

SUNet ID: 05794739

Name: Luis Perez

Collaborators:

By turning in this assignment, I agree by the Stanford honor code and declare that all of this is my own work.

Problem 1

- (a) The key insight for this equality is that the vector of the true distribution \mathbf{y} is one-hot encoded vector with 1 for the true, outside word o and 0 everywhere else. We therefore have:

$$\begin{aligned} \text{CrossEntropy}(\mathbf{y}, \hat{\mathbf{y}}) &= - \sum_{w \in \text{Vocab}} y_w \log(\hat{y}_w) \\ &= -1 \cdot \log(\hat{y}_o) - \sum_{w \in \text{Vocab}, w \neq o} 0 \cdot \log(\hat{y}_w) \\ &= -\log(\hat{y}_o) \\ &= -\log P(O = o \mid C = c) = \mathbf{J}_{\text{naive-softmax}}(\mathbf{v}_c, o, \mathbf{U}) \end{aligned}$$

- (b) We compute the partial derivate of the cross-entropy loss with respect to v_c .

$$\begin{aligned} \frac{\partial J}{\partial \mathbf{v}_c} &= -\frac{\partial}{\partial \mathbf{v}_c} \left[\sum_{w \in \text{Vocab}} \mathbf{y}_w \log \hat{\mathbf{y}}_w \right] && \text{(Results from 1a)} \\ &= - \sum_{w \in \text{Vocab}} \mathbf{y}_w \frac{\partial}{\partial \mathbf{v}_c} \log \frac{\exp \mathbf{u}_w^T \mathbf{v}_c}{\sum_{k \in \text{Vocab}} \exp \mathbf{u}_k^T \mathbf{v}_c} && \text{(Definition of } \hat{\mathbf{y}}_w) \\ &= - \sum_{w \in \text{Vocab}} \frac{\mathbf{y}_w}{\hat{\mathbf{y}}_w} \frac{\partial}{\partial \mathbf{v}_c} \frac{\exp \mathbf{u}_w^T \mathbf{v}_c}{\sum_{k \in \text{Vocab}} \exp \mathbf{u}_k^T \mathbf{v}_c} && \text{(Derivative of } \log x \text{ and Chain rule)} \\ &= - \sum_{w \in \text{Vocab}} \frac{\mathbf{y}_w}{\hat{\mathbf{y}}_w} \left[\frac{\mathbf{u}_w \exp(\mathbf{u}_w^T \mathbf{v}_c) \sum_{k \in \text{Vocab}} \exp \mathbf{u}_k^T \mathbf{v}_c}{\left(\sum_{k \in \text{Vocab}} \exp \mathbf{u}_k^T \mathbf{v}_c \right)^2} - \frac{\exp(\mathbf{u}_w^T \mathbf{v}_c) \sum_{k \in \text{Vocab}} \mathbf{u}_k \exp \mathbf{u}_k^T \mathbf{v}_c}{\left(\sum_{k \in \text{Vocab}} \exp \mathbf{u}_k^T \mathbf{v}_c \right)^2} \right] && \text{(Quotient Rule of Derivates)} \\ &= - \sum_{w \in \text{Vocab}} \frac{\mathbf{y}_w}{\hat{\mathbf{y}}_w} \left[\mathbf{u}_w \hat{\mathbf{y}}_w - \hat{\mathbf{y}}_w \sum_{k \in \text{Vocab}} \frac{\mathbf{u}_k \exp \mathbf{u}_k^T \mathbf{v}_c}{\sum_{\ell \in \text{Vocab}} \exp \mathbf{u}_\ell^T \mathbf{v}_c} \right] && \text{(Simplification using definition of } \hat{\mathbf{y}}_w, \text{ reindex sum)} \\ &= - \sum_{w \in \text{Vocab}} \mathbf{y}_w \mathbf{u}_w + \left(\sum_{w \in \text{Vocab}} \mathbf{y}_w \right) \left(\sum_{k \in \text{Vocab}} \mathbf{u}_k \hat{\mathbf{y}}_k \right) && \text{(Defintion of } \hat{\mathbf{y}}_k, \text{ distribute sum, simplify)} \\ &= \mathbf{U}[\hat{\mathbf{y}} - \mathbf{y}] && \text{(Convert to matrix form.)} \end{aligned}$$

(c) We compute the partial derivate of the cross-entry loss with respect to \mathbf{u}_w . We have:

$$\begin{aligned}
\frac{\partial J}{\partial \mathbf{u}_w} &= -\frac{\partial}{\partial \mathbf{u}_w} \left[\sum_{k \in Vocab} \mathbf{y}_k \log \hat{\mathbf{y}}_k \right] && \text{(Results from 1a)} \\
&= -\sum_{k \in Vocab} \mathbf{y}_k \frac{\partial}{\partial \mathbf{u}_w} \log \frac{\exp \mathbf{u}_w^T \mathbf{v}_c}{\sum_{k \in Vocab} \exp \mathbf{u}_k^T \mathbf{v}_c} && \text{(Definition of } \hat{\mathbf{y}}_k \text{)} \\
&= -\sum_{k \in Vocab} \frac{\mathbf{y}_k}{\hat{\mathbf{y}}_k} \frac{\partial}{\partial \mathbf{u}_w} \frac{\exp \mathbf{u}_k^T \mathbf{v}_c}{\sum_{\ell \in Vocab} \exp \mathbf{u}_\ell^T \mathbf{v}_c} && \text{(Derivative of } \log x \text{ and Chain rule)} \\
&= -\frac{\mathbf{y}_w \mathbf{v}_c \exp(\mathbf{u}_w^T \mathbf{v}_c) \sum_{\ell \in Vocab} \exp \mathbf{u}_\ell^T \mathbf{v}_c - \mathbf{v}_c (\exp \mathbf{u}_w^T \mathbf{v}_c)^2}{\hat{\mathbf{y}}_w (\sum_{\ell \in Vocab} \exp \mathbf{u}_\ell^T \mathbf{v}_c)^2} \\
&\quad + \sum_{k \in Vocab, k \neq w} \frac{\mathbf{y}_k \mathbf{v}_c \exp(\mathbf{u}_w^T \mathbf{v}_c) \exp(\mathbf{u}_k^T \mathbf{v}_c)}{\hat{\mathbf{y}}_k (\sum_{\ell \in Vocab} \exp \mathbf{u}_\ell^T \mathbf{v}_c)^2} && \text{(Quotient Rule and split into cases)} \\
&= -\mathbf{v}_c \mathbf{y}_w [1 - \hat{\mathbf{y}}_w] + \mathbf{v}_c \sum_{k \in Vocab, k \neq w} \mathbf{y}_k \hat{\mathbf{y}}_w && \text{(Simplify)} \\
&= [-\mathbf{y}_w + \hat{\mathbf{y}}_w (\mathbf{y}_w + \sum_{k \in Vocab, k \neq w} \mathbf{y}_k)] \mathbf{v}_c && \text{(Refactoring)} \\
&= [\hat{\mathbf{y}}_w - \mathbf{y}_w] \mathbf{v}_c
\end{aligned}$$

We can, in fact, write the above for the entire matrix \mathbf{U} as follows:

$$\frac{\partial J}{\partial \mathbf{U}} = \mathbf{v}_c [\hat{\mathbf{y}} - \mathbf{y}]^T$$

(d) We compute the derivative element by element. We have:

$$\begin{aligned}
\frac{d}{d\mathbf{x}_i} \sigma(\mathbf{x}_i) &= \frac{d}{d\mathbf{x}_i} \left[\frac{e^{\mathbf{x}_i}}{1 + e^{\mathbf{x}_i}} \right] \\
&= \frac{e^{\mathbf{x}_i} (1 + e^{\mathbf{x}_i}) - e^{\mathbf{x}_i} \cdot e^{\mathbf{x}_i}}{(1 + e^{\mathbf{x}_i})^2} \\
&= \left(\frac{e^{\mathbf{x}_i}}{1 + e^{\mathbf{x}_i}} \right) \left(1 - \frac{e^{\mathbf{x}_i}}{1 + e^{\mathbf{x}_i}} \right) \\
&= \sigma(\mathbf{x}_i) [1 - \sigma(\mathbf{x}_i)]
\end{aligned}$$

As such, we have the vector derivate as:

$$\frac{d}{d\mathbf{x}} \sigma(\mathbf{x}) = \sigma(\mathbf{x}) \circ [\mathbf{1} - \sigma(\mathbf{x})]$$

where \circ represents element-wise vector product.

- (e) We compute the requested derivatives for the negative sampling loss function. First, with respect to \mathbf{u}_o .

$$\begin{aligned}
\frac{\partial J}{\partial \mathbf{u}_o} &= -\frac{\partial}{\partial \mathbf{u}_o} \log \sigma(\mathbf{u}_o^T \mathbf{v}_c) && \text{(Only the first term depends on } \mathbf{u}_o) \\
&= -\frac{\sigma(\mathbf{u}_o^T \mathbf{v}_c)(1 - \sigma(\mathbf{u}_o^T \mathbf{v}_c))}{\sigma(\mathbf{u}_o^T \mathbf{v}_c)} \mathbf{v}_c && \text{(Derivative of } \log x \text{ and results from 1d)} \\
&= -(1 - \sigma(\mathbf{u}_o^T \mathbf{v}_c)) \mathbf{v}_c
\end{aligned}$$

Next, with respect to \mathbf{u}_k .

$$\begin{aligned}
\frac{\partial J}{\partial \mathbf{u}_k} &= -\frac{\partial}{\partial \mathbf{u}_k} \log \sigma(-\mathbf{u}_k^T \mathbf{v}_c) && \text{(Only the } k+1\text{-th term depends on } \mathbf{u}_k) \\
&= \frac{\sigma(-\mathbf{u}_k^T \mathbf{v}_c)(1 - \sigma(-\mathbf{u}_k^T \mathbf{v}_c))}{\sigma(-\mathbf{u}_k^T \mathbf{v}_c)} \mathbf{v}_c && \text{(Derivative of } \log x \text{ and results from 1d)} \\
&= (1 - \sigma(-\mathbf{u}_k^T \mathbf{v}_c)) \mathbf{v}_c
\end{aligned}$$

And finally, with respect to \mathbf{v}_c .

$$\begin{aligned}
\frac{\partial J}{\partial \mathbf{v}_c} &= -\frac{\partial}{\partial \mathbf{v}_c} \log \sigma(\mathbf{u}_o^T \mathbf{v}_c) - \sum_{k=1}^K \frac{\partial}{\partial \mathbf{v}_c} \log \sigma(-\mathbf{u}_k^T \mathbf{v}_c) \\
&= -(1 - \sigma(\mathbf{u}_o^T \mathbf{v}_c)) \mathbf{u}_o + \sum_{k=1}^K (1 - \sigma(-\mathbf{u}_k^T \mathbf{v}_c)) \mathbf{u}_k && \text{(Previous results)}
\end{aligned}$$

Computing the naive-softmax requires iterating over the entire vocabulary, which can be extremely large, while the negative sampling loss requires considering only $K + 1$ samples.

- (f) We now compute the skip-gram loss function. We have:

$$\text{i} \quad \frac{\partial \mathbf{J}_{\text{skip-gram}}(\mathbf{v}_c, w_{t-m}, \dots, w_{t+m}, \mathbf{U})}{\partial \mathbf{U}} = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \frac{\partial \mathbf{J}(\mathbf{v}_c, w_{t+m}, \mathbf{U})}{\partial \mathbf{U}}$$

$$\text{ii} \quad \frac{\partial \mathbf{J}_{\text{skip-gram}}(\mathbf{v}_c, w_{t-m}, \dots, w_{t+m}, \mathbf{v}_c)}{\partial \mathbf{v}_c} = \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \frac{\partial \mathbf{J}(\mathbf{v}_c, w_{t+m}, \mathbf{U})}{\partial \mathbf{v}_c}$$

$$\text{iii} \quad \frac{\partial \mathbf{J}_{\text{skip-gram}}(\mathbf{v}_c, w_{t-m}, \dots, w_{t+m}, \mathbf{v}_c)}{\partial \mathbf{v}_w} = 0 \quad (w \neq c)$$