Author: Benjamin Smidt
Created: October 3rd, 2022
Last Updated: October 3rd, 2022

CS 224N A2: Word Vectors

Note to the reader. This is my work for assignment one of Stanford's course CS 224N: Natural Language Processing with Deep Learning. You can find the lecture Winter 2021 lectures series on YouTube here. This document is meant to be used as a reference, explanation, and resource for the assignment, not necessarily a comprehensive overview of Word Vectors. If there's a typo or a correction needs to be made, feel free to email me at benjamin.smidt@utexas.edu so I can fix it. Thank you! I hope you find this document helpful :).

Contents

1 Written Understanding

1.1 Problem 1-A

Instructions
Prove that the naive-softmax loss is the same as the cross-entropy loss between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$, i.e. (note that $\boldsymbol{y}, \hat{\boldsymbol{y}}$ are vectors and $\hat{\boldsymbol{y}}_o$ is a scalar).

Solution
To start with, our naive-softmax loss is defined as

$$\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) = -\log P(O = o | C = c)$$

where

$$P(O = o \mid C = c) = \frac{\exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}{\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}$$

Let's define our variables. $\hat{\boldsymbol{y}}$ is our score vector. Note that the numerator is a vector of length $\boldsymbol{V}$ while the denominator is a scalar (this notation is a bit abusive but I think it actually makes things clearer). Thus, each index in the vector can be interpeted as the probability that the corresponding word (using the the index and one hot vector) is the center word.

$$\hat{\boldsymbol{y}} = \frac{\exp(\boldsymbol{u}^\top \boldsymbol{v}_c)}{\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}$$

$\boldsymbol{y}$ is the one hot vector of the true center word. Then

$$-\sum_{w \in \text{Vocab}} \boldsymbol{y}_w \log(\hat{\boldsymbol{y}}_w) = \boldsymbol{-y}_1 \log(\hat{\boldsymbol{y}}_1) + ... + -\boldsymbol{y}_o \log(\hat{\boldsymbol{y}}_o) + ... + -\boldsymbol{y}_w \log(\hat{\boldsymbol{y}}_w)$$

Where the index $\boldsymbol{o}$ indicates the index containing the only 1 within $\hat{\boldsymbol{y}}$ (since it is a one-hot vector).

$$-\sum_{w \in \text{Vocab}} \boldsymbol{y}_w \log(\hat{\boldsymbol{y}}_w) = \boldsymbol{-}(0) \log(\hat{\boldsymbol{y}}_1) + ... + -\boldsymbol{(}1) \log(\hat{\boldsymbol{y}}_o) + ... + -(0) \log(\hat{\boldsymbol{y}}_w)$$

$$-\sum_{w \in \text{Vocab}} \boldsymbol{y}_w \log(\hat{\boldsymbol{y}}_w) = -\log(\hat{\boldsymbol{y}}_o)$$

$$-\sum_{w \in \text{Vocab}} \boldsymbol{y}_w \log(\hat{\boldsymbol{y}}_w) = -\log(\frac{\exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}{\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)})$$

$$-\sum_{w \in \text{Vocab}} \boldsymbol{y}_w \log(\hat{\boldsymbol{y}}_w) = -\log P(O = o | C = c)$$

$$-\sum_{w \in \text{Vocab}} \boldsymbol{y}_w \log(\hat{\boldsymbol{y}}_w) = \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})$$

I know the instructions said one line but I was going for clarity here. Obviously, I could not define the variables (leave it to you to figure out what they mean) and just write some one liner that connects the dots. However, I wanted to make this as clear to understand as possible. Hopefully this leaves no room for ambiguity.

1.2 Problem 1-B

Instructions

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 $\boldsymbol{y}$, $\hat{\boldsymbol{y}}$, and $\boldsymbol{U}$. Additionally, answer the following two questions with one sentence each: (1) When is the gradient zero? (2) Why does subtracting this gradient, in the general case when it is nonzero, make $\boldsymbol{v}_c$ a more desirable vector (namely, a vector closer to outside word vectors in its window)?

Solution

We start with our definition of $\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})$.

$$\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) = -log\frac{\exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}{\sum_{w\in\text{Vocab}}\exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}$$

$$\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{v}_c} = \frac{\partial}{\partial \boldsymbol{v}_c} - log\frac{\exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}{\sum_{w\in\text{Vocab}}\exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}$$

$$\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{v}_c} = \frac{\partial}{\partial \boldsymbol{v}_c} - \boldsymbol{u}_o^\top \boldsymbol{v}_c + \frac{\partial}{\partial \boldsymbol{v}_c}\ log \sum_{w\in\text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)$$

Everything up to this point should be easy to follow if you remember calculus (although see the first lecture where he goes through these exact steps in detail if you are confused). I then do a quick change of variables to ensure I know what I'm taking my partial derivative with respect to.

$$\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{v}_c} = -\boldsymbol{u}_o + \frac{1}{\sum_{w\in\text{Vocab}}\exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}\frac{\partial}{\partial \boldsymbol{v}_c}\sum_{j\in\text{Vocab}} \exp(\boldsymbol{u}_j^\top \boldsymbol{v}_c)$$

$$\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{v}_c} = -\boldsymbol{u}_o + \frac{\sum_{j\in\text{Vocab}}\boldsymbol{u}_j^\top\exp(\boldsymbol{u}_j^\top \boldsymbol{v}_c)}{\sum_{w\in\text{Vocab}}\exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}$$

From here I hope you can see that if you remove the $\boldsymbol{u}_j^\top$ form the second sum term you're left with $\hat{\boldsymbol{y}}$. Simplifying with this in mind we get the following.

$$\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{v}_c} = -\boldsymbol{U}^\top \boldsymbol{y} + \boldsymbol{U}^\top \hat{\boldsymbol{y}}_{\boldsymbol{w}}$$

$$\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{v}_c} = \boldsymbol{U}^\top(\hat{\boldsymbol{y}} - \boldsymbol{y})$$

1. The gradient is zero when $\hat{\boldsymbol{y}} = \boldsymbol{y}$. Obviously if our predicted and correct vectors are equivalent then our accuracy is perfect and there's no update that could improve the loss.
2. Because we're doing gradient descent (as opposed to ascent), the update adds some portion ($\propto \alpha$) of $\boldsymbol{U}^\top \boldsymbol{y}$ and subtracts some portion of $\boldsymbol{U}^\top \hat{\boldsymbol{y}}$. This makes intuitive sense because adding the correct vector $\boldsymbol{u}_o^\top$ makes it more like that vector (which is what we want) and subtracting by $\boldsymbol{U}^\top \hat{\boldsymbol{y}}$, the weighted average of our incorrect vectors that are producing the loss, makes it less like those vectors (again, what we want).

1.3 Problem 1-C

<u>Instructions</u>
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 $\boldsymbol{y}_1$, $\boldsymbol{y}_2$, $\ldots$). Note that $\boldsymbol{u}_w$ is a vector while $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots$ are scalars.

<u>Solution</u>
Case 1: $\boldsymbol{u}_w = u_o$
We start with the case that $\boldsymbol{u}_w = u_o$. That is, the gradient with respect to the correct embedding vector.

$$\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) = -log \frac{\exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}{\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}$$

$$\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_o} = \frac{\partial}{\partial \boldsymbol{u}_o} - log \frac{\exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}{\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}$$

$$\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_o} = \frac{\partial}{\partial \boldsymbol{u}_o} - log \frac{\exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}{\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}$$

$$\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_o} = \frac{\partial}{\partial \boldsymbol{u}_o} - \boldsymbol{u}_o^\top \boldsymbol{v}_c + \frac{\partial}{\partial \boldsymbol{u}_o} \; log \sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)$$

$$\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_o} = -\boldsymbol{v}_c + \frac{1}{\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)} \frac{\partial}{\partial \boldsymbol{u}_o} \sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)$$

$$\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_o} = -\boldsymbol{v}_c + \frac{\boldsymbol{v}_c \exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}{\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}$$

$$\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_o} = -\boldsymbol{v}_c + \boldsymbol{v}_c(\hat{\boldsymbol{y}} \cdot \boldsymbol{y})$$

$$\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_o} = \boldsymbol{v}_c(\hat{\boldsymbol{y}}_{\mathbf{0}} - 1)$$

Note that $\hat{\boldsymbol{y}} \cdot \boldsymbol{y}$ is the dot product of the two vectors.

Case 2: $\boldsymbol{u}_w = \boldsymbol{u}_o$
Now we move onto the case that $\boldsymbol{u}_w \neq \boldsymbol{u}_o$. That is, the gradient with respect to any vector $\boldsymbol{u}_w$ that isn't $\boldsymbol{u}_o$. I'll use the notation $\boldsymbol{u}_j$ to indicate a specific $\boldsymbol{u}_w$ and (hopefully) prevent any confusion.

$$\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) = -log \frac{\exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}{\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}$$

$$\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_j} = \frac{\partial}{\partial \boldsymbol{u}_j} - log \frac{\exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}{\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}$$

$$\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_j} = \frac{\partial}{\partial \boldsymbol{u}_j} - log \frac{\exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}{\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}$$

$$\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_j} = \frac{\partial}{\partial \boldsymbol{u}_j} - \boldsymbol{u}_o^\top \boldsymbol{v}_c + \frac{\partial}{\partial \boldsymbol{u}_j} \; log \sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)$$

$$\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_j} = \frac{1}{\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)} \frac{\partial}{\partial \boldsymbol{u}_j} \sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)$$

$$\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_j} = \frac{\boldsymbol{v}_c \exp(\boldsymbol{u}_j^\top \boldsymbol{v}_c)}{\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}$$

$$\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_j} = \boldsymbol{v}_c \; \hat{\boldsymbol{y}}_{\boldsymbol{j}}$$

1.4 Problem 1-D

Instructions

Write down the partial derivative of $\boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})$ with respect to $\boldsymbol{U}$. Please break down your answer in terms of $\frac{\partial \boldsymbol{J}(\boldsymbol{v}_c,o,\boldsymbol{U})}{\partial \boldsymbol{u}_1}, \frac{\partial \boldsymbol{J}(\boldsymbol{v}_c,o,\boldsymbol{U})}{\partial \boldsymbol{u}_2}, \cdots, \frac{\partial \boldsymbol{J}(\boldsymbol{v}_c,o,\boldsymbol{U})}{\partial \boldsymbol{u}_{|\text{Vocab}|}}$. The solution should be one or two lines long.

Solution

We already know

$$\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_j} = \boldsymbol{v}_c\ \boldsymbol{\hat{y}_j} \text{ and } \frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_0} = \boldsymbol{v}_c(\boldsymbol{\hat{y}_0} - 1)$$

Since $\boldsymbol{y}$ is a one hot vector, then

$$\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_j} = \boldsymbol{v}_c(\boldsymbol{\hat{y}} - \boldsymbol{y})$$

With that we can write

$$\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{U}} = \frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_1}, ..., \frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{u}_V}$$

where V is the index of the last word vector in $\boldsymbol{U}$. It follows then that

$$\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{U}} = \boldsymbol{v}_c\ \boldsymbol{\hat{y}_1}, \boldsymbol{v}_c\ \boldsymbol{\hat{y}_2}, ..., \boldsymbol{v}_c\ \boldsymbol{\hat{y}_V}$$

$$\frac{\partial \boldsymbol{J}_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U})}{\partial \boldsymbol{U}} = \boldsymbol{v}_c(\boldsymbol{\hat{y}} - \boldsymbol{y})^\top$$

where we use the outer product to get the proper shape for $\boldsymbol{U}$.

1.5 Problem 1-E

Instructions

The ReLU (Rectified Linear Unit) activation function is given by the Equation:

$$f(x) = \max(0, x) \tag{1}$$

Please compute the derivative of $f(x)$ with respect to x, where x is a scalar. You may ignore the case that the derivative is not defined at 0.

Solution

We can break ReLU into two cases.

$$f(x) = \begin{cases} 0 & \text{if } \ x < 0 \\ x & \text{if } \ x > 0 \end{cases}$$

Then the derivation becomes trivial.

$$f'(x) = \begin{cases} 0 & \text{if } \ x < 0 \\ 1 & \text{if } \ x > 0 \end{cases}$$

1.6 Problem 1-F

Instructions

The sigmoid function is given by:

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

Please compute the derivative of $\sigma(x)$ with respect to x, where x is a scalar. Hint: you may want to write your answer in terms of $\sigma(x)$.

Solution

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

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

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