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).

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

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

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

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

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

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