Word2Vec

Jaime Sancho Molero

December 22, 2022

1 Introduction

In this document I am going to derive the backpropagation equations for a Word2Vec using the **skip-gram** algorithm. The characteristics of this model are:

- **Loss function**: Negative sample
- **Window size**: m
- **Negative Sampling words**: K

2 Backpropagation Negative Sampling

First of all, let's define our cost function for a single pair of words:

$$J = -\log \sigma(\mathbf{u}_o^T \mathbf{v}_c) - \sum_{s=1}^{s=K} \log(1 - \sigma(\mathbf{u}_s^T \mathbf{v}_c)) \tag{1}$$

where the underscripts o, c stands for *outside* and *center* respectively.

Computing the derivatives of this function is straightforward with a little knowloedge of vector calculus. Let's start with the center word and the outside word:

$$\nabla_{\mathbf{v}_c} J = -\frac{1}{\sigma}\sigma(1-\sigma)\mathbf{u}_o + \sum_{s=1}^{s=K} \frac{1}{1-\sigma}\sigma(1-\sigma)\mathbf{u}_s \implies$$

$$\nabla_{\mathbf{v}_c} J = -\big(1 - \sigma(\mathbf{u}_o^T \mathbf{v}_c)\big)\mathbf{u}_o + \sum_{s=1}^{s=K} \sigma(\mathbf{u}_s^T \mathbf{v}_c)\mathbf{u}_s \tag{2}$$

Doing the same but respect with the outside words:

$$\nabla_{\mathbf{u}_o} J = -\big(1 - \sigma(\mathbf{u}_o^T \mathbf{v}_c)\big)\mathbf{u}_o \tag{3}$$

since the negative sampling vectors $\mathbf{u_s}$ are, in general given the large size of the corpus, are not equal to $\mathbf{u_o}$.

The last gradient we need to compute is with respect the sampling words $\mathbf{u}_s$. We have to do the same as with the two previous gradients:

$$\nabla_{\mathbf{u}_s} J = \sigma(\mathbf{u}_s^T \mathbf{v}_c)\mathbf{v}_c \qquad s = 1, 2, ..., K \tag{4}$$

3 Backpropagation Skip - gram

The cost function for the Skip-gram model for a single window:

$$J_{\textbf{skip-gram}} = \sum_{-m \leq j \leq m, j \neq 0} J(\mathbf{v_c}, u_{w_{t+j}}) \tag{5}$$

where J can be any cost function related to word vectors (it can be Naive Bayes but in this case we will use Negative sampling 1). The subscript w_{t+j} stands for the word in the window at position $t + j$. For example, in the phrase:

"Machine learning uses a series of mathematical tools such as ..."

the first window, if we consider window size of 2, is:

"Machine learning uses a series"

The center word is: "uses". The word at position w_{-2} is "Machine".

3.1 Center word

$$\nabla_{\mathbf{v}_c} J_{\textbf{skip-gram}} = \sum_{-m \leq j \leq m, j \neq 0} \nabla_{\mathbf{v}_c} J \tag{6}$$

Since we already computed the quantity $\nabla_{\mathbf{v}_c} J$ (equation (2))this equation is done.

3.2 Window word

$$\nabla_{\mathbf{u}_{w_k}} J_{\textbf{skip-gram}} = \sum_{-m \leq j \leq m, j \neq 0} \nabla_{\mathbf{u}_{w_k}} J \ \ \delta_{w_k, w_{t+j}} = \nabla_{\mathbf{u}_{w_k}} J \tag{7}$$

3.3 Negative Sampling Word

Since now the gradient we are going to compute is respect a sampling word $\mathbf{u}_s$ we have to remember that $\mathbf{u}_s \neq \mathbf{u}_{w_{t+j}}$. Having these mind:

$$\nabla_{\mathbf{u}_s} J_{\textbf{skip-gram}} = \sum_{-m \leq j \leq m, j \neq 0} \nabla_{\mathbf{u}_s} J = 2m \, \nabla_{\mathbf{u}_s} J \tag{8}$$