Kate Lassiter

Problem 1:

CBOW objective:

$$L\ (A,\ B)\ =\ -\ log\ p(wc|w_{o-m},...,w_{o-1},w_{o+1},...,w_{o+m})$$

$$L\ (A,\ B)\ =\ -\ log\ p(wc|a_{avg})$$

$$L\ (A,\ B)\ =\ -\ log\ \frac{exp\ b_{wc}^{T}a_{avg}}{\sum_{w\in v} exp\ b_{w}^{T}a_{avg}}$$

Use the following *negative sampling* approximation:

$$\sum_{w\in v} exp\ b_{w}^{T}a_{avg} \sim \left(\frac{1}{1+exp-\ b_{wc}^{T}a_{avg}}\right)\ E_{w_k\sim\ Psample(w)}\left(\prod_{k=1}^{K}\left(\frac{1}{1+exp b_{wk}^{T}a_{avg}}\right)\right)$$

$$log\sum_{w\in v} exp\ b_{w}^{T}a_{avg} \sim\ log\left(\frac{1}{1+exp-\ b_{wc}^{T}a_{avg}}\right)+\ E_{w_k\sim\ Psample(w)}\left(\sum_{k=1}^{K} log\left(\frac{1}{1+exp b_{wk}^{T}a_{avg}}\right)\right)$$

$$-\ log\sum_{w\in v} exp\ b_{w}^{T}a_{avg} \sim -\ \ log(\sigma(\ b_{wc}^{T}a_{avg}))\ \ -\ E_{w_k\sim\ Psample(w)}\left(\sum_{k=1}^{K} log(\sigma(\ -\ \ b_{wk}^{T}a_{avg}))\right)$$

Objective using the approximation:

$$L\ (A,\ B)\ =\ -\ log\ \frac{exp\ b_{wc}^{T}a_{avg}}{\sum_{w\in v} exp\ b_{w}^{T}a_{avg}}$$

$$L\ (A,\ B)\ =\ -\ b_{wc}^{T}a_{avg}\ +\ log\sum_{w\in v} exp\ b_{w}^{T}a_{avg}$$

$$L\ (A,\ B)\ =\ -\ b_{wc}^{T}a_{avg}\ -\ log(\sigma(\ b_{wc}^{T}a_{avg}))\ -\ E_{w_k\sim\ Psample(w)}\left(\sum_{k=1}^{K} log(\sigma(\ -\ \ b_{wk}^{T}a_{avg}))\right)$$

$$L\ (A,\ B)\ =\ -\ b_{wc}^{T}a_{avg}\ -\ log\left(\frac{1}{1+exp-\ b_{wc}^{T}a_{avg}}\right)\ -\ E_{w_k\sim\ Psample(w)}\left(\sum_{k=1}^{K} log\left(\frac{1}{1+exp\ b_{wk}^{T}a_{avg}}\right)\right)$$

$$L\ (A,\ B)\ =\ -\ b_{wc}^{T}a_{avg}\ -\ log(1\ )\ +\ log(1\ +\ exp\ -\ \ b_{wc}^{T}a_{avg})\ -\ \left(\sum_{k=1}^{K} log(1)\right)\ +\ \sum_{k=1}^{K} log(1\ +\ exp\ b_{wk}^{T}a_{avg}))$$

$$\boxed{L\ (A,\ B)\ =\ -\ b_{wc}^{T}a_{avg}+\ log(1\ +\ exp\ -\ \ b_{wc}^{T}a_{avg})\ \ +\ \sum_{k=1,\ w_k\sim\ Psample(w)}^{K} log(1\ +\ exp\ b_{wc}^{T}a_{avg}))}$$

Associated vectors for validation words:

Kate Lassiter

```
32496it [09:54, 60.75it/s]| epoch  10 | 32500/32580 batches | loss    1.331
32509it [09:55, 34.93it/s]money: work, much, them, it, even, result, use, support, what, god
lion: convention, measured, consists, statement, navy, euro, succeeded, hills, beer, punishment
africa: europe, america, india, china, germany, france, asia, east, north, south
musician: writer, singer, actor, author, actress, poet, march, kingdom, january, addition
dance: music, history, able, country, whole, same, view, idea, style, subject
```

Some vectors make more sense than others. Clearly, the word "africa" returns a list of all locationally related words, like countries and directions like "north" and "south". Likewise, "musician" returns similar words like "writer", "singer", "actor" and other performance/artistic words. However, the vectors for "lion" and "money" don't really return logical words in relation.

Mathematical Problems

Problem 1 Let w be some word in the vocabulary $\mathcal{V}$ and let e_w be it's one-hot encoding (pretend the word is actually integer w, we might have $itos[w]$ = "cat" for example depending on how we set up the hash map between words and integers). Explain why $B^{\intercal}e_w = b_w \in \mathbb{R}^d$ and why this multiplication selects the w^{th} column of $B^{\intercal}$. Remember, if $B \in \mathbb{R}^{|\mathcal{V}| \times d}$ then $B^{\intercal} \in \mathbb{R}^{d \times |\mathcal{V}|}$.

A one hot encoded vector is a representation where only one value is one and the rest are zero.
For example:
V={"cat","dog","frog"}
Represented by a VxV one hot encoding matrix e

	Dog	Cat	Frog
Dog	1	0	0
Cat	0	1	0
Frog	0	0	1

Picking out the word "cat" from e would result in the column $e_w \in \mathbb{R}^{|V| x1}$ = e_{cat} = [0,1,0]. Crucially, only one dimension will be non-zero. This dimension represents the word "cat".
Take for example matrix B $\in \mathbb{R}^{|V| \times d}$

	1	2	3	4
Dog	7	5.2	2	3
Cat	4.1	1	8	5
Frog	3	5	0	2.1

Multiplying matrix $B^{\intercal} \in \mathbb{R}^{dx|V|}$ by vector e_{cat} results in a new vector of dimension d x 1 (Recall: d x V ⊗V x 1 = d x 1). This will just be the full column of B (with all its dimensions) , b_{cat}, for only the word "cat" because all the other multiplications will be zero due to the other zero values in e_{cat}.

	$B^{\intercal}$				e_{cat}							$B^{\intercal}$ e_{cat}
	Dog	**Cat**	Frog		Cat							
1	7	**4.1**	3		0	Dog		7x0	4.1x1	3x0		**[4.1,**
2	5.2	**1**	5	**x**	1	Cat	**=**	5.2x0	1x1	5x0	**=**	**1,**
3	2	**8**	0		0	Frog		2x0	8x1	0x0		**8,**
4	3	**5**	2.1					3x0	5x1	2.1x0		**5]**
											=	b_{cat}

Kate Lassiter

Problem 2 Assume you do CBOW and Skip-Gram with negative sampling. Assume $m = 1$. Which method, on average, will get more training samples? Suppose there are 3 sentences with 7, 8, and 11 tokens. How many training sampling (positive training samples), will each method get. Draw a picture of a sentence with token counts and think about the number of samples each method gives. This is why Skip-Gram is used more often. It is more "sample efficient": you get more training data per Corpus.

On average there will be more training samples for Skip –Gram.

Skip –Gram:

Training data: Get pairs of the center word and all other words in a fixed window

sentence1 = (token1,token2, token3, token4, token5, token6, token7)

data1=(token2,token1),(token2,token3),(token3,token2),(token3,token4),(token4,token3),(token4,token5),(token5,token4),(token5,token6),(token6,token5),(token6,token7)

= 10 training samples

sentence2 = (token1,token2, token3, token4, token5, token6, token7,token8)

data2=(token2,token1),(token2,token3),(token3,token2),(token3,token4),(token4,token3),(token4,token5),(token5,token4),(token5,token6),(token6,token5),(token6,token7),(token7,token6),(token7,token8)

= 12 training samples

sentence3 = (token1,token2, token3, token4, token5, token6, token7,token8,token9, token10, token11)

data3=(token2,token1),(token2,token3),(token3,token2),(token3,token4),(token4,token3),(token4,token5),(token5,token4),(token5,token6),(token6,token5),(token6,token7),(token7,token6),(token7,token8),(token8,token7),(token8,toke9),(token9,token8),(token9,token10),(token10,token9),(token10,token11)

=18 training samples

Positive samples are drawn:

$p(w_o|w_c) \sim \sigma(b^T_{wo} a_{wc}) = \frac{1}{1+exp - b^T_{wo} a_{wc}}$

Sample K words w_k that are not in context and we know these have a negative label.

They are predicted with probability: $p(w_k|w_c) \sim 1 - \sigma(b^T_{wk} a_{wc}) = \sigma(-b^T_{wk} a_{wc}) = \frac{1}{1+exp b^T_{wk} a_{wc}}$

Denominator becomes:

Kate Lassiter

$$- \ log \sum_{w \in v} exp b_{w}^{T} a_{wc} \sim - \ log(\sigma(b_{wo}^{T} a_{wc})) \ - \ E_{w_k \sim Psample(w)} (\sum_{k=1}^{K} log(\sigma(\ - \ b_{wk}^{T} a_{wc})))$$

<u>CBOW:</u>

Training data: Get pairs of the center word and all other words in a fixed window

sentence1 = (token1,token2, token3, token4, token5, token6, token7)

data1=(token1,token3,token2),(token2,token4,token3),(token3,token5, token4),(token4,token6, token5),(token5,token7, token6)

=5 training samples

sentence2 = (token1,token2, token3, token4, token5, token6, token7,token8)

data2=(token1,token3,token2),(token2,token4,token3),(token3,token5, token4),(token4,token6, token5),(token5,token7, token6),(token6,token8, token7)

=6 training samples

sentence3 = (token1,token2, token3, token4, token5, token6, token7,token8,token9, token10, token11)

data3=(token1,token3,token2),(token2,token4,token3),(token3,token5, token4),(token4,token6, token5),(token5,token7, token6),(token6,token8, token7),(token7,token9, token8),(token8,token10, token9),(token9,token11, token10)

=9 training samples

Positive samples are drawn:

$$p(w_o | w_c) \sim \sigma(b_{wo}^{T} a_{avg}) = \frac{1}{1+exp-b_{wo}^{T} a_{avg}}$$

Negative samples are drawn:

$$p(w_k | w_c) \sim \ 1 \ - \ \sigma(b_{wk}^{T} a_{avg}) \ = \ \sigma(\ - \ b_{wk}^{T} a_{avg}) \ = \frac{1}{1+expb_{wk}^{T} a_{avg}}$$

Denominator becomes:

$$- \ log \sum_{w \in v} exp \ b_{w}^{T} a_{avg} \sim - \ log(\sigma(b_{wo}^{T} a_{avg})) \ - \ E_{w_k \sim Psample(w)} (\sum_{k=1}^{K} log(\sigma(\ - \ b_{wk}^{T} a_{avg})))$$

Kate Lassiter

Problem 3 In class we looked at the formula for the Skip-Gram for 1 sample (w_c, w_o) and got

$$\mathcal{L}(A, B) = -\log p(b_{w_o}|a_{w_c})) = -b_{w_o}^{\intercal} a_{w_c} + \log \sum_{w \in \mathcal{V}} \exp b_w^{\intercal} a_{w_c}$$

Then, we said that the gradients were as below. Prove this. Also, explain why $\frac{\partial \mathcal{L}}{\partial a_{w_c}}$ can be be interpreted as a difference between a hard guess and an expected value.

***Skip–Gram* Objective:**

$$L\ (A,\ B)\ =\ -\ log\ p(b_{wo}|a_{wc})$$

$$L\ (A,\ B)\ =\ -\ log\ \frac{exp\ b^T_{wo}a_{wc}}{\sum_{w \in v} exp\ b^T_w a_{wc}}$$

$$L\ (A,\ B)\ =\ -\ b^T_{wo}a_{wc}\ +\ log\ \sum_{w \in v} exp\ b^T_w a_{wc}$$

b_{w0}:

Recall:

$$\frac{d}{dx}[f(g(x))] = f'(g(x))g'(x)$$

$$\frac{exp\ b^T_w a_{wc}}{\partial b_{wo}} = 0\ for\ all\ b_w\ except\ for\ b_{wo}$$

Thus:

$$\frac{\partial L\ (A,B)}{\partial b_{wo}}\ =\ -\ a_{wc}\ +\ (\frac{1}{\sum_{w \in v} exp\ b^T_w a_{wc}})\ (a_{wc} exp\ b^T_{w0} a_{wc})$$

$$\boxed{\frac{\partial L\ (A,B)}{\partial b_{wo}}\ =\ -\ a_{wc}\ +\ \frac{a_{wc} exp\ b^T_{w0} a_{wc}}{\sum_{w \in v} exp\ b^T_w a_{wc}}}$$

a_{wc} :

$$\frac{L\ (A,B)}{\partial a_{wc}}\ =\ -\ b_{wo}\ +\ (\frac{1}{\sum_{u \in v} exp\ b^T_u a_{wc}})\ (\sum_{w \in v}\ b_w exp\ b^T_w a_{wc})$$

$$\frac{L\ (A,B)}{\partial a_{wc}}\ =\ -\ b_{wo}\ +\ \frac{\sum_{w \in v} b_w exp\ b^T_w a_{wc}}{\sum_{u \in v} exp\ b^T_u a_{wc}}$$

Kate Lassiter

$$\boxed{\frac{L\ \ (A,B)}{\partial a_{wc}} = -\ b_{wo} + \sum_{w\in v} b_w \frac{exp\ b_w^T a_{wc}}{\sum_{u\in v} exp\ b_u^T a_{wc}}}$$

This is $E[b_w]-b_{wo}$:

$\frac{L\ \ (A,B)}{\partial a_{wc}}$ can be interpreted as the difference between a hard guess and an expected value. This is because the expected value of a random variable is equal to the value of that variable times the probability of observing that variable. In this case the probability p(w) = $\frac{exp\ b_w^T a_{wc}}{\sum_{u\in v} exp\ b_u^T a_{wc}}$ and this is multiplied by the actual observed value b_w, for every possible value of w, resulting in the expectation of b_w, $E[b_w]$. This is the value that would be theoretically achieved on average given a large number of trials. Subtract from that the actual value for b_{w0} and the gradient equals the expected value of b_w minus a hard guess of the value b_w, it's actual value.