COMP 472 - Artificial Intelligence

Word Embeddings

Solutions

Question 1 Consider the following sentence:

"the cat drinks the milk"

We will use this sentence to train a CBOW Word2Vec model. Assume that:

- you want to produce word embeddings of dimension 2,
- you use a context window of size 2 (1 word before and 1 word after the target word), and
- your vocabulary only contains the words in the sentence above

(a) Using only the sentence above, how many instances will be generated as training set?

3 instances

Instance	Context Word-1	Context Word+1	To Predict
1	the	drinks	cat
2	cat	the	drinks
3	drinks	milk	the

(b) List the one-hot vectors that correspond to each word in the vocabulary. (Assume alphabetical ordering)

Word	Hot Vector			
cat	1	0	0	0
drinks	0	1	0	0
milk	0	0	1	0
the	0	0	0	1

(c) List the one-hot vectors that correspond to each training instance in the input layer.

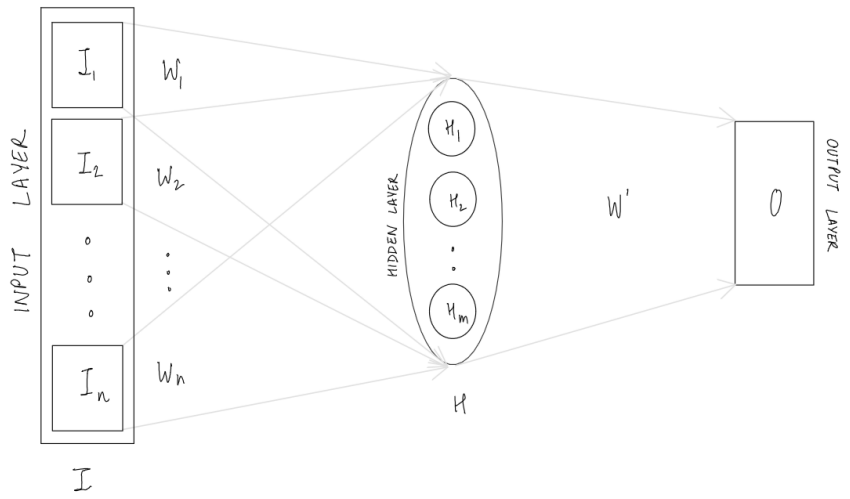
Instance	Context	Word	Hot Vector			
1	Context Word-1	the	0	0	0	1
	Context Word+1	drinks	0	1	0	0
2	Context Word-1	cat	1	0	0	0
	Context Word+1	the	0	0	0	1
3	Context Word-1	drinks	0	1	0	0
	Context Word+1	milk	0	0	1	0

(d) How many nodes will the hidden layer contain?
Number of nodes in hidden layer = 2

(e) What is the target hot vector for each training instance?

Instance	To Predict	Hot Vector			
1	cat	1	0	0	0
2	drinks	0	1	0	0
3	the	0	0	0	1

(f) Assume that the Word2Vec model is trained with the standard network depicted below:



i. What will be the values of n and m?
$n = 2$, $m = 2$

ii. What will be the sizes of I, W_i (for each $1 \leq i \leq n$), W' and O?
Size of $I_1 = I_2 = 1 \times 4$, $I = 2 \times 4$
Size of $W_1 = W_2 = 4 \times 2$
Size of $W' = 2 \times 4$
Size of $O = 1 \times 4$

(g) Assume that we have these weight vectors:

$$W = \begin{bmatrix} 2 & 6 \\ 4 & 3 \\ 1 & 4 \\ 5 & 2 \end{bmatrix}$$

$$W' = \begin{bmatrix} 6 & 2 & 8 & 3 \\ 4 & 5 & 9 & 7 \end{bmatrix}$$

To compute the final probabilities at the output layer, we use the softmax function as shown in class. Recall that for a given vector of size k, the softmax function is defined as:

$$p_i = \frac{e^{x_i}}{\sum_{i=1}^{k} e^{x_i}}, \text{ where } 1 \leq i \leq k$$

i. Trace the first feed forward pass in the network and show the values propagated all the way to the output layer.

Instance 1 -

Instance	Context	Word	Hot Vector				To Predict	Target			
1	Context Word-1 Context Word+1	the drinks	0 0	0 1	0 0	1 0	cat	1	0	0	0

Calculate the output of each hidden node for each context word

$$H = I \times W = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 2 & 6 \\ 4 & 3 \\ 1 & 4 \\ 5 & 2 \end{bmatrix} = \begin{bmatrix} 5 & 2 \\ 4 & 3 \end{bmatrix}$$

Take the average

$$H_{AVG} = \begin{bmatrix} 4.5 & 2.5 \end{bmatrix}$$

Calculate output

$$O = H_{AVG} \times W' = \begin{bmatrix} 4.5 & 2.5 \end{bmatrix} \times \begin{bmatrix} 6 & 2 & 8 & 3 \\ 4 & 5 & 9 & 7 \end{bmatrix} = \begin{bmatrix} 37 & 21.5 & 58.5 & 31 \end{bmatrix}$$

Calculate softmax probabilities for the output

$$softmax(O) = softmax(\begin{bmatrix} 37 & 21.5 & 58.5 & 31 \end{bmatrix})$$
$$= \begin{bmatrix} 4.6 \times 10^{-10} & 8.53 \times 10^{-17} & 0.99 & 1.14 \times 10^{-12} \end{bmatrix}$$

ii. What is the error after the first pass?

Calculate error

$$E = O - T = \begin{bmatrix} 4.6 \times 10^{-10} & 8.53 \times 10^{-17} & 0.99 & 1.14 \times 10^{-12} \end{bmatrix} - \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} \approx -1 & 8.53 \times 10^{-17} & 0.99 & 1.14 \times 10^{-12} \end{bmatrix}$$