

# Assignment 2: Language Models

Apoorv Sharma

1/24/2021

## Question 1

Consider 3 words:  $i, j$  and  $k$  that have index values  $m, m+1$  and  $m+2$ , respectively  
From the language we know:

$$P(w_m = i) = \frac{1}{|V|}$$

$$P(w_{m+1} = j) = \frac{1}{|V|}$$

$$P(w_{m+2} = k) = \frac{1}{|V|}$$

Thus:

$$P(w_m \neq i \text{ OR } w_{m+1} \neq j) = 1 - \frac{1}{V^2}$$

$$P(w_{m+1} \neq i \text{ OR } w_{m+2} \neq j \mid w_{m+1} \neq j) = 1 - \frac{1}{V} \frac{1}{V-1} = 1 - \frac{1}{V^2}$$

Thus for a corpus size of  $M$  with vocabulary size  $V$  we have

$$(1 - \frac{1}{V^2}) \cdot (1 - \frac{1}{V^2})^{M-1} = (1 - \frac{1}{V^2})^M$$

## Question 2

Following are the inputs and desired output of the neural network:

| x1 | x2 | f(x1, x2) |
|----|----|-----------|
| 1  | 1  | -1        |
| 1  | 0  | 1         |
| 0  | 1  | 1         |
| 0  | 0  | -1        |

Since we have 1 hidden layer, the model can be represented by the following function:

$$f(\vec{x}; \vec{W}, \vec{c}, \vec{w}, \vec{b}) = \text{sign}(\vec{w}^T \cdot \max(0, \vec{W}^T \cdot x + c) + b)$$

Assign the following weights:

$$W = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

$$c = \begin{pmatrix} 0 \\ -1 \end{pmatrix}$$

$$w = \begin{pmatrix} 1 \\ -2 \end{pmatrix}$$

$$b = 0$$

Proving the weights

$$X \cdot W = \begin{pmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 1 & 1 \\ 2 & 2 \end{pmatrix}$$

$$\max(X \cdot W + c) = \max\left(\begin{pmatrix} 0 & 0 \\ 1 & 1 \\ 1 & 1 \\ 2 & 2 \end{pmatrix} + \begin{pmatrix} 0 & -1 \\ 0 & -1 \\ 0 & -1 \\ 0 & -1 \end{pmatrix}\right) = \max\left(\begin{pmatrix} 0 & -1 \\ 1 & 0 \\ 1 & 0 \\ 2 & 1 \end{pmatrix}\right) = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 2 & 1 \end{pmatrix}$$

$$\text{sign}((X \cdot W + c) \cdot w) = \text{sign}\left(\begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 2 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ -2 \end{pmatrix}\right) = \text{sign}\left(\begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}\right) = \begin{pmatrix} -1 \\ 1 \\ 1 \\ -1 \end{pmatrix}$$

### Question 3

If all the weights are initialized to 0, the derivative with respect to the loss function will be the same for every weight in the weight matrix. Therefore all the weights will have the same value in the subsequent iterations and the weights will never change. As a result the model will never be trained.