

Step 2 (linear)

- The first step is multiplying

The input by a matrix filled with model parameters learned

in training

since the multiplications are multiplying

each row in the matrix by the whole vector so think of

- each row as asking a question
- and this dot product multiplication is
- ≈ 1 if the vector answers the question with a yes and ≤ 0 if not and the

$$\begin{matrix} \text{matrix} & \text{vector} & \text{result} \\ \begin{bmatrix} 1 & 2 & 5 & \dots \\ 7 & \dots & \dots & \dots \\ 3 & \dots & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} & \begin{bmatrix} 5 \\ 7 \\ 3 \\ 0 \end{bmatrix} & = \begin{bmatrix} 2 \\ 5 \\ 9 \\ 0 \end{bmatrix} \end{matrix}$$

$$\begin{bmatrix} 5 \\ 7 \\ 3 \\ 0 \end{bmatrix} \xrightarrow{\text{linear}} \begin{bmatrix} 2 \\ 5 \\ 9 \\ 0 \end{bmatrix}$$

$$\begin{matrix} \text{Q: First name?} \\ \text{Michael} \end{matrix} \rightarrow R_0 \quad \begin{matrix} \approx 1 = \text{yes} \\ \leq 0 = \text{no} \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 5 & \dots \\ 7 & \dots & \dots & \dots \\ 3 & \dots & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} 5 \\ 7 \\ 3 \\ 0 \end{bmatrix} = \begin{bmatrix} R_0 \cdot E \\ \vdots \end{bmatrix}$$

ans the question

Questions can be more complicated Like asking 2 questions in 2A case for both Q's to be yes the $R_0 \cdot E \approx 2 = \text{yes}$ and $\leq 1 = \text{NO}$

- Very often this step requires adding another vector to the output called the Bias which also has model parameters this Bias normalizes the dot product so in the case of $R_0 \cdot E \approx 2$ for yes the bias would learn to have -1 as first element so that the result is $\approx 1 = \text{yes}$ $\leq 0 = \text{NO}$

$$\begin{matrix} \text{Q: First name Michael?} \\ \text{R0 last name Jordan} \end{matrix} \rightarrow \text{Bias}$$

$$\begin{bmatrix} 3 & 6 & 5 & \dots \\ 7 & \dots & \dots & \dots \\ 3 & \dots & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} 5 \\ 7 \\ 3 \\ 0 \end{bmatrix} + \begin{bmatrix} -1 \\ -5 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 5 \\ 9 \\ 0 \end{bmatrix}$$

result of linear step

≈ 1 if E encodes "Michael"
 ≤ 0 if otherwise Jordan

- in our case E does encode MJ
- Bias makes sure $\leq 0 = \text{NO}$

- The number of rows in the matrix \approx num of question asked is 50K in GPT-3 which is $4 \times$ the dimension space of embedding
- * choosing a clean multiple for the dimension is hardware friendly

★ You can think of R_0 in the last Ex as having vectors for M and J hence the dot product is largest if E also has directionality in the M and J direction (which it does) so think of the operation $R_0 \cdot E = (M+J) \cdot E$ 65 once again Matrix has billions