Mathematical details

20 June 2023   09:06

Bigram LM

Assumption

$$P\left(y^{<1>}, y^{<2>}, \dots y^{<T>}\right)$$
$$= P\left(y^{<1>}\right) P\left(y^{<2>} \mid y^{<1>}\right) \dots$$
$$\sim P\left(y^{<T>} \mid y^{<T-1>}\right)$$

Training strategy

$$\underline{x} = \begin{pmatrix} \text{vocab.size}, \\ 1 \end{pmatrix} \qquad \underline{y} = \begin{pmatrix} \text{vocab.size} \\ 1 \end{pmatrix}$$
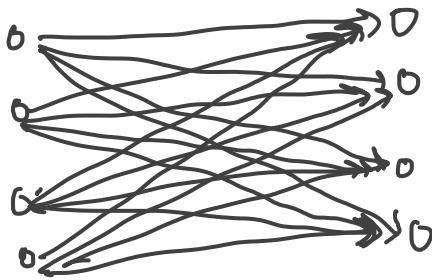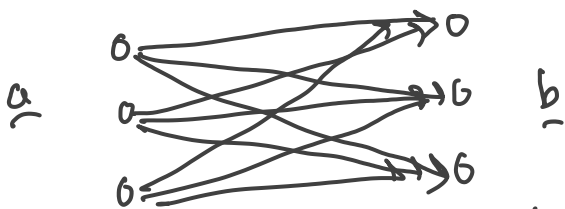
lets say you have
vocab = [a, b, c]
Now if want to get
$P(b|a)$ where one-hot $(a) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$

one-hot $(b) = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$

$\underline{a}$ $\underline{b}$

W = weight matrix can be defined
as

$$-\log \begin{pmatrix} P(a) & P(b|a) & P(c|a) \\ P(b) & P(a|b) & P(c|b) \\ P(c) & P(a|c) & P(c) \end{pmatrix}$$