15-780 HW2

Linear Hypothesis for Language Modeling

Part a. Assume that we consider a language modeling setting where our input x is a one-hot vector encoding a single previous word

$$x \in [0,1]^k = [\text{One-Hot}(\text{word}_i)] \tag{1}$$

(where One-Hot represents the operator we dicussed in class, a vector with all zeros except a one in the position corresponding to the value of word_i) and the target y is a discrete encoding of the next word

$$y \in 1, \ldots, k \tag{2}$$

where k is the number of possible classes (which is also the vocabulary size). Suppose we have a set of target probabilities of next word given the previous word:

$$p(\texttt{word}_{i+1} = j \mid \texttt{word}_i = r)$$

(i.e., we have a target probability value for each next work given the previous word). Show that a linear hypothesis function, followed by a softmax operation to convert this to probabilities, can encode *any* such probabilities. In other words, show that there exists some $\theta \in \mathbb{R}^{k \times k}$ such that the target probabilities are given by

$$\text{softmax}(\theta^T x) \tag{3}$$

Part b. Now suppose the input $x \in [0,1]^{2k}$ contains a concatenation of the previous *two* words

$$x = \begin{bmatrix} \text{One-Hot}(\text{word}_{i-1}) \\ \text{One-Hot}(\text{word}_i) \end{bmatrix}. \tag{4}$$

Show that a linear hypothesis function *cannot* express all possible two-word probabilities. I.e., show that there exist target probabilitty distributions

$$P(\texttt{word}_{i+1} = j \mid \texttt{word}_i = r,\ \texttt{word}_{i-1} = s)$$

such that one *cannot* encode this probability distribution using the probabilities given by

$$\text{softmax}(\theta^T x) \tag{5}$$

for any $\theta \in \mathbb{R}^{2k \times k}$.

Part c. In part (b), is there an alternative representation of the inputs (i.e., not just as the concatenation of two one-hot vectors of each word, but via some other representation), that *does* make it possible to present arbitrary distributions conditioned on the past two words?