README

Jeff Zhang

July 15, 2024

Language models are probability distributions over sequences of tokens. More formally, we define the sample space Ω^d to be the d-dimensional cartesian product of a set of tokens, the event space (and σ-algebra) $\mathcal{F}$ to be the set of all possible token sentences, and probability measure $\mathbb{P} : \mathcal{F} \mapsto [0, 1]$.

Letting the random vector $\mathbf{X} : \Omega^d \mapsto \mathbb{R}^d$, we can encode sentences into a more computable measurable space, $\mathbb{R}^d$. (denotational vs distributional semantics?). To evaluate the probability of an encoded sentence, we compute (joint pdf ==¿ conditional ==¿ chain rule??):

$$\begin{aligned} p_{\mathbf{X}}(\mathbf{x}) &= \mathbb{P}[\mathbf{X} = \mathbf{x}] \\ &= \mathbb{P}[\mathbf{X^{-1}}(\mathbf{x})] \\ &= \mathbb{P}[\{\omega \in \Omega^d : \mathbf{X}(\omega) = \mathbf{x}\}] \end{aligned} \tag{1}$$

The goal is to approximate $p_{\mathbf{X}}(\mathbf{x})$ given $\mathcal{D} = \{(x_1, y_1), ..., (x_n, y_n)\}$ (assuming $(x_i, y_i) \sim p$), with some hypothesis $h : \mathcal{X} \mapsto \mathcal{Y}$. With parametric models, this is done by posing parameter estimation as an optimization problem (ERM?) $\operatorname{argmin}_\theta \mathcal{L}(\theta)$.

In this document we will cover the major models $h \in \mathcal{H} = \{n{-}gram, mlp, rnn, transformer, ...\}$ that have been used for natural language processing and understanding.

(PAC?, theoretical optimal?)