Hidden Markov Models

Justin T. Chiu

February 6, 2020

Abstract

TODO

1 Problem Setup

We apply hidden markov models (HMMs) to language modeling, where we would like to model sentences $x_{1:T}$. The generative process of an HMM is as follows:

1. Choose an initial state $z_0 \sim \text{Cat}()$
2. For each time $t \in \{1, \ldots, T\}$ choose a state $z_t \mid z_{t-1} \sim \text{Cat}()$
3. For each time $t \in \{0, \ldots, T\}$ choose a word $x_t \mid z_t \sim \text{Cat}()$.

This gives the following joint distribution:

$$\log p_\theta(x_{0:T}, z_{0:T}) = \log p_\theta(x_0, z_0) + \sum_{t=1}^{T} \log p_\theta(x_t, z_t \mid z_{t-1})$$

2 Parameter estimation

We maximize the evidence of the observed sentences $\log p(x_{0:T} = \log \sum_{z_{0:T}} p(x_{1:T}, z_{0:T})$.

2.1 Gradient of evidence

Let $\psi_0(z_0, z_1) = \log p(x_{0:1}, z_{0:1})$ and $\psi_t(z_t, z_{t+1}) = \log p(x_{t+1}, z_{t+1} \mid z_t)$ for $t \in \{1, \ldots, T-1\}$. After conditioning on the observed $x_{0:T}$, we can express the evidence as the following: $Z_x = \log p(x_{0:T}) = \sum_{t=0}^{T-1} \psi_t(z_t, z_{t+1})$ where Z_x is the clamped partition function.

2.2

2.2.1 Very high training loss

Surrogate loss is a loose bound, but that is ok. We proved gradient estimator is correct.