Hidden Markov Models

Justin T. Chiu

February 26, 2020

Abstract

TODO

1 Introduction

TODO

2 Problem Setup

We apply hidden markov models (HMMs) to language modeling, where we would like to model sentences $\mathbf{x}_{0:T}$. The generative process of an HMM is as follows:

1. Choose an initial state $z_0 \sim \text{Cat}(\nu), \nu \in \mathbb{R}^K$
2. For each time $t \in \{1, \ldots, T\}$ choose a state $z_t \mid z_{t-1} \sim \text{Cat}(\theta_{z_{t-1}}), \theta \in \mathbb{R}^{K \times K}$
3. For each time $t \in \{0, \ldots, T\}$ choose a word $x_t \mid z_t \sim \text{Cat}(\phi_{z_t}), \phi \in \mathbb{R}^{K \times V}$.

This gives the following joint distribution:

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

3 Parameter estimation

We maximize the evidence of the observed sentences $\log p(\mathbf{x}_{0:T}) = \log \sum_{\mathbf{z}_{0:T}} p(\mathbf{x}_{1:T}, \mathbf{z}_{0:T})$ via gradient ascent.

3.1 Gradient of the evidence

Let $\psi_0(z_0, z_1) = \log p(\mathbf{x}_{0:1}, \mathbf{z}_{0:1})$ and $\psi_t(z_t, z_{t+1}) = \log p(x_{t+1}, z_{t+1} \mid z_t)$ for $t \in \{1, \dots, T-1\}$. Additionally, let $\oplus$ and $\otimes$ be addition and multiplication in the log semiring. After conditioning on the observed $\mathbf{x}_{0:T}$, we can express the evidence as the following:

$$A_x = \log p(\mathbf{x}_{0:T}) = \bigoplus_{\mathbf{z}_{0:T}} \bigotimes_{t=0}^{T-1} \psi_t(z_t, z_{t+1})$$

where $A_{\mathbf{x}}$ is the clamped log partition function.

We show that the gradient of the log partition function with respect to the $\psi_t(z_t, z_{t+1})$ is the first moment (a general result for the cumulant generating function of exponential family distributions). Given the value of the derivative $\frac{\partial A_{\mathbf{x}}}{\partial \psi_t(z_t, z_{t+1})}$, we can then apply the chain rule to compute the total derivative with respect to downstream parameters. For example, the gradient with respect to the transition matrix of the HMM is given by

$$\frac{\partial A_{\mathbf{x}}}{\partial \theta_{ij}} = \sum_t \frac{\partial A_{\mathbf{x}}}{\partial \psi_t(i,j)} \frac{\partial \psi_t(i,j)}{\partial \theta_{ij}}$$

Recall that the derivative of logaddexp ($\oplus$ in the log semiring) is

$$\frac{\partial}{\partial x} x \oplus y = \frac{\partial}{\partial x} \log e^x + e^y = \frac{e^x}{e^x + e^y} = \exp(x - (x \oplus y)),$$

while the derivative of addition ($\otimes$ in the log semiring) is

$$\frac{\partial}{\partial x} x \otimes y = 1.$$

The derivative of the clamped log partition function $A_{\mathbf{x}}$ is given by

$$\begin{aligned}
\frac{\partial A_{\mathbf{x}}}{\partial \psi_t(i,j)} &= \frac{\partial}{\partial \psi_t(i,j)} \bigoplus_{\mathbf{z}_{0:T}} \bigotimes_{t'} \psi_{t'}(z_{t'}, z_{t'+1}) \\
&= \frac{\partial}{\partial \psi_t(i,j)} \left(\left(\bigoplus_{\mathbf{z}_{0:T}: z_t = i, z_{t+1} = j} \bigotimes_{t'} \psi_{t'}(z_{t'}, z_{t'+1}) \right) \oplus \left(\bigoplus_{\mathbf{z}_{0:T}: z_t \neq i, z_{t+1} \neq j} \bigotimes_{t'} \psi_{t'}(z_{t'}, z_{t'+1}) \right) \right) \\
&= \exp\left(\left(\bigoplus_{\mathbf{z}_{0:T}: z_t = i, z_{t+1} = j} \bigotimes_{t'} \psi_{t'}(z_{t'}, z_{t'+1}) \right) \frac{\partial}{\partial \psi_t(i,j)} \left(\bigotimes_{t'} \psi_{t'}(z_{t'}, z_{t'+1}) \right) - A_{\mathbf{x}} \right) \\
&= \exp\left(\left(\bigoplus_{\mathbf{z}_{0:T}: z_t = i, z_{t+1} = j} \bigotimes_{t'} \psi_{t'}(z_{t'}, z_{t'+1}) \right) - A_{\mathbf{x}} \right)
\end{aligned}$$

which is the edge marginal for z_t, z_{t+1} obtained via the forward-backward algorithm. The second equality is a result of the associativity of $\oplus$, while the third and fourth equalities are applications of the derivatives derived above.

3.1.1 Very high training loss

It is okay if the surrogate loss is a loose bound. We proved gradient estimator is correct. This just means that the ELBO under a pairwise approximation is loose.

4 Cloned HMM (CHMM)

Inspired by Dedieu et al. (2019), we introduce a sparse emission constraint that allows us to efficiently compute derivatives in large state spaces. We constrain each word x to be emit only by $z \in \mathcal{C}_x \subset \mathcal{Z}$:

$$p(x \mid z) \propto \begin{cases} \phi_{z,x} & z \in \mathcal{C}_x \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

When conditioning on

References

Antoine Dedieu, Nishad Gothoskar, Scott Swingle, Wolfgang Lehrach, Miguel Lázaro-Gredilla, and Dileep George. Learning higher-order sequential structure with cloned hmms, 2019.