Notes on Hidden Markovs Models

Jeff Miller

March 5, 2016

Contents

Hidden Markov models (HMMs) are a surprisingly powerful tool for modeling a wide range of sequential data, including speech, written text, genomic data, weather patterns, financial data, animal behaviors, and many more applications. Dynamic programming enables tractable inference in HMMs, including finding the most probable sequence of hidden states using the Viterbi algorithm, probabilistic inference using the forward-backward algorithm, and parameter estimation using the Baum–Welch algorithm.

1 Setup

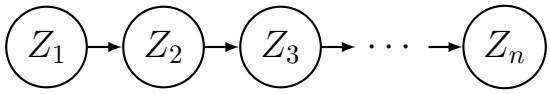
1.1 Refresher on Markov chains

- Recall that $(Z_1, \ldots, Z_n)$ is a Markov chain if

$$Z_{t+1} \perp (Z_1, \ldots, Z_{t-1}) \mid Z_t$$

for each t, in other words, "the future is conditionally independent of the past given the present."

- This is equivalent to saying that the distribution respects the following directed graph:

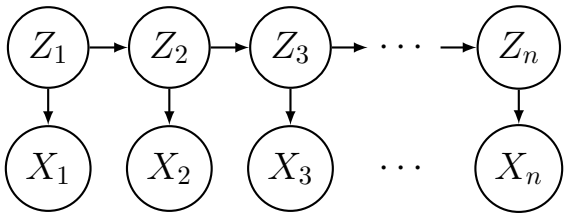


- A Markov chain is a natural model to use for sequential data when the present state Z_t contains all of the information about the future that could be gleaned from $Z_1, \ldots, Z_t$. In other words, when Z_t is the "complete state" of the system.

- If Z_t is sufficiently rich, then this may be the case, but oftentimes we only get to observe an incomplete or noisy version of Z_t. In such cases, a hidden Markov model is preferable.

1.2 Hidden Markov model

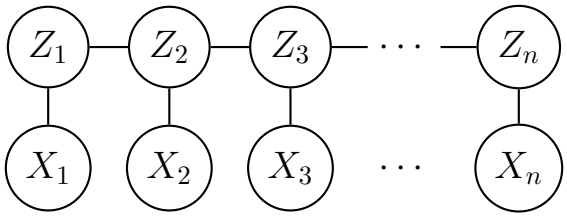
- A hidden Markov model is a distribution $p(x_1, \ldots, x_n, z_1, \ldots, z_n)$ that respects the following directed graph:



In other words, it factors as

$$p(x_{1:n}, z_{1:n}) = p(z_1)p(x_1|z_1)\prod_{t=2}^{n} p(z_t|z_{t-1})p(x_t|z_t).$$

- It turns out that in this case, it is equivalent to say that the distribution respects the following undirected graph:



- $Z_1, \ldots, Z_n$ represent the "hidden states", and $X_1, \ldots, X_n$ represent the sequence of observations.
- Assume that $Z_1, \ldots, Z_n$ are discrete random variables taking finitely many possible values. For simplicity, let's denote these possible values as $1, \ldots, m$. In other words, $Z_t \in \{1, \ldots, m\}$.
- Assume that the "transition probabilities" $T(i,j) = \mathbb{P}(Z_{t+1} = j \mid Z_t = i)$ do not depend on the time index t. This assumption is referred to as "time-homogeneity." The $m \times m$ matrix T in which entry (i,j) is $T(i,j)$ is referred to as the "transition matrix." Note that every row of T must sum to 1. (A nonnegative matrix with this property is referred to as a "stochastic matrix").
- Assume that the "emission distributions" $\varepsilon_i(x_t) = p(x_t \mid Z_t = i)$ do not depend on the time index t. While we assume the Z's are discrete, the X's may be either discrete or continuous, and may also be multivariate.
- The "initial distribution" π is the distribution of Z_1, that is, $\pi(i) = \mathbb{P}(Z_1 = i)$.

1.3 Example

- $m = 2$ hidden states, i.e., $Z_t \in \{1, 2\}$
- Initial distribution: $\pi = (0.5, 0.5)$
- Transition matrix:
$$T = \begin{bmatrix} .9 & .1 \\ .2 & .8 \end{bmatrix}$$
- Emission distributions:
$$X_t \mid Z_t = i \sim \mathcal{N}(\mu_i, \sigma_i^2)$$
where $\mu = (-1, 1)$ and $\sigma = (1, 1)$.

2 Overview of dynamic programming for HMMs

- There are three main algorithms used for inference in HMMs: the Viterbi algorithm, the forward-backward algorithm, and the Baum–Welch algorithm.

- In the Viterbi algorithm and the forward-backward algorithm, it is assumed that all of the parameters are known—in other words, the initial distribution π, transition matrix T, and emission distributions ε_i are all known.
- The Viterbi algorithm is an efficient method of computing the sequence $z_1^*, \ldots, z_n^*$ with the highest probability given $x_1, \ldots, x_n$, that is, computing

$$z_{1:n}^* = \operatorname*{argmax}_{z_{1:n}} p(z_{1:n} | x_{1:n}).$$

  Naively maximizing over all sequences would take order nm^n time, whereas the Viterbi algorithm only takes nm^2 time.
- The forward-backward algorithm enables one to efficiently compute a wide range of conditional probabilities given $x_{1:n}$, for example,
  - $\mathbb{P}(Z_t = i \mid x_{1:n})$ for each i and each t,
  - $\mathbb{P}(Z_t = i, Z_{t+1} = j \mid x_{1:n})$ for each i, j and each t,
  - $\mathbb{P}(Z_t \neq Z_{t+1} \mid x_{1:n})$ for each t,
  - etc.
- The Baum–Welch algorithm is a method of estimating the parameters of an HMM (the initial distribution, transition matrix, and emission distributions), using expectation-maximization and the forward-backward algorithm.
- Historical fun facts:
  - The term "dynamic programming" was coined by Richard Bellman in the 1940s, to describe his research on certain optimization problems that can be efficiently solved with recursions.
  - How does it involve "programming"? In this context, "programming" means optimization. As I understand it, this terminology comes from the 1940s during which there was a lot of work on how to optimize military plans or "programs", in the field of operations research. So, what is "dynamic" about it? There's a funny story on Wikipedia about why he called it "dynamic" programming.

3 Viterbi algorithm

(to do)

4 Forward-backward algorithm

(to do)