Hidden Markov Models

David Larson

November 13, 2001

1 Introduction

This paper will present a definition and some of the mathematics behind Hidden Markov Models (HMMs). It will also discuss some of the usefulness and applications of these models. For a more detailed description, see Durbin et. al.[1] or Rabiner[2].

2 Hidden Markov Models

2.1 Definition

Hidden Markov models find their use in categorizing sequences of data, in our case, DNA.[2, 1] The idea behind them is simple: a HMM is a model for generating a data sequence by following a stochastic procedure. The model contains a finite, usually small number of different states; the sequence is generated by moving from state to state and at each state, producing a piece of data.

In a regular (not hidden) Markov Model, the data produced at each state is predetermined (for example, you have states for the bases A, T, G, and C). The history of states is given explicitly in the data. See figure 1 for a diagram of a regular Markov model. In a HMM, the history of states the model took cannot generally be determined from the data sequence.

Rabiner and Durbin et. al.[2, 1] use notation similar to the following. If there are N states, then each state is represented by S_i, where $i = 1...N$. The probability of moving from state S_i to state S_j is given by the matrix element a_{ij}. The probability of producing or emitting the data O_k in a state S_i is $e_i(O_k)$. In our example, $O_k \in \{A, T, G, C\}$. See figure 2 for a diagram of a three-state HMM. Let $O = O_1O_2O_3...O_T$ be a sequence of T observations (data) and let $Q = q_1q_2q_3...q_T$ be the sequence of T states the model went through to produce those observations. The values π_i are the probabilities of starting in the states S_i.

2.2 Generic Use

HMMs are commonly used to categorize data sequences. For example, HMMs could be used to distinguish between coding and non-coding regions of DNA.[3] HMMs can do this, with significant accuracy, by the following steps.

1. Assume the sequences could have been generated by HMMs.

2. Determine (from intuition) the topology of the HMMs. The topology refers to the number of states and how they are connected, but not to probabilities.

3. For each category of sequences, using a number of sequences known to be in that category, use an iterative training process to find the best parameters π_i, a_{ij}, and $e_i(O_k)$ for a HMM to model that category.

4. For an unknown sequence, determine which HMM models it with the highest probability.

2.3 Probabilities

Rabiner[2] gives three main problems that must be solved to understand HMMs. This paper will only look at the first (the easiest).

1. Given a data sequence and a HMM, how does one calculate the probability that the sequence was generated by the HMM?

2. Given a data sequence and a HMM, how does one find the most probable sequence of states for generating the given data?

3. Given a data sequence and topology for a HMM, how does one calculate the parameters $e_i(b)$ and a_{ij} so that the HMM best models the data?

The solution of the first problem is used to classify sequences from a HMM, and the solution of the third problem is used to train the HMM.

The first problem is solved recursively. The important quantity used in the calculation is $\alpha_t(i)$. This is the probability of the model generating the first t data and ending in the state q_i. As Rabiner puts it,

$$\alpha_t(i) = P(O_1 O_2 O_3 ... O_t, q_t = S_i)$$

The important point is that each $\alpha_t(i)$ can be calculated from all of the $\alpha_{t-1}(i)$ values for $i = 1...N$. Specifically,

$$\alpha_t(i) = e_i(O_t) \sum_{j=1}^{N} \alpha_{t-1}(j) a_{ji}$$

The probability of ending in a state S_i and producing the correct sequence up to that point is: the probability of producing O_t times the probability of transferring to the state S_i from the previous distribution of states $\alpha_{t-1}(j)$.

The first distribution is clearly given by

$$\alpha_1(i) = \pi_i e_i(O_1)$$

from the probability distribution π_i of initial states of the HMM. To calculate the final probability of the HMM producing the given sequence, simply sum the final probabilities $\alpha_T(j)$ over all states $j = 1...N$.

$$P(O) = \sum_{j=1}^{N} \alpha_T(j)$$

Note that throughout these calculations, we have assumed that the HMM is completely known.

3 Applications

Markov models have a large range of applications, both inside and outside of biology.

When looking at DNA as the data sequence, they can be used to tell the difference between coding and non-coding sections of DNA. Borodovsky and McIninch use the Markov model shown in figure 3 in their paper describing the GENMARK algorithm.[3] That model has a built-in period of 3 to match the genetic code.

Eddy has written software using HMMs to search large databases for specific proteins.[4]

To extend the theory of HMMs, look at how they can be applied to wavelet transforms[5], and look at coupled hidden Markov models[6].

HMMs can deal with much more than DNA or proteins, though. An object's x,y coordinates as a function of time (from videotape, perhaps) provide two sequences of data a HMM could use. Coupled HMMs provide an extension of HMMs to deal with multiple streams of data[6].

HMMs have been used to recognize sign language[7]. Oliver has used it to recognize driver behavior for applications in SmartCars[8, 9], and try to recognize and classify human interactions from videotape[10]. HMMs are also commonly used in speech recognition[2]. In general, as long as a system can be reduced to multiple streams of data, HMMs can be applied to analyze those sequences.

References

[1] Richard Durbin, Sean Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge, 1998.

[2] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

[3] Mark Borodovsky and James McIninch. Genmark: Parallel gene recognition for both dna strands. *Computers Chem.*, 17(2):123–133, 1993.

[4] Sean R. Eddy. Hmmer: Profile hidden markov models for biological sequence analysis. http://hmmer.wustl.edu/, 2001.

[5] Matthew Crouse, Robert Nowak, and Richard Baraniuk. Wavelet-based statistical signal processing using hidden markov models. *IEEE Transactions on Signal Processing*, 1997. Web address is citeseer.nj.nec.com/crouse98waveletbased.html.

[6] M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In Proceedings of IEEE CVPR97, 1996. Web address is citeseer.nj.nec.com/article/brand96coupled.html.

[7] Christopher Lee and Yangsheng Xu. Online, interactive learning of gestures for human/robot interfaces. *IEEE International Conference on Robotics and Automation, Minneapolis, MN*, 4:2982–2987, 1996.

[8] Nuria Oliver and Alex Pentland. Driver behavior recognition and prediction in a smartcar. In *Proceedings of SPIE Aerosense2000 'Enhanced and Synthetic Vision' Orlando, Florida*, April 2000.

[9] Nuria Oliver and Alex Pentland. Graphical models for driver behavior recognition in a smartcar. In *Proceedings of IEEE Intl. Conference on Intelligent Vehicles 2000 Detroit, Michigan*, October 2000.

[10] Nuria Oliver, Barbara Rosario, and Alex Pentland. A bayesian computer vision system for modeling human interaction. In *ICVS*, pages 255–272, 1999. Web address is citeseer.nj.nec.com/article/oliver99bayesian.html.

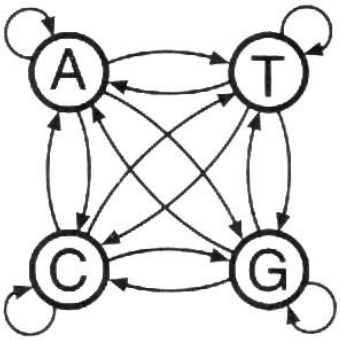

Figure 1: Durbin et. al.'s figure of a simple Markov model for generating a DNA sequence.

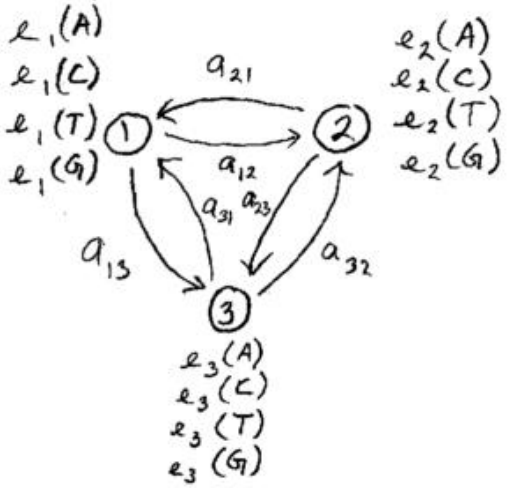

Figure 2: A three-state hidden Markov model.

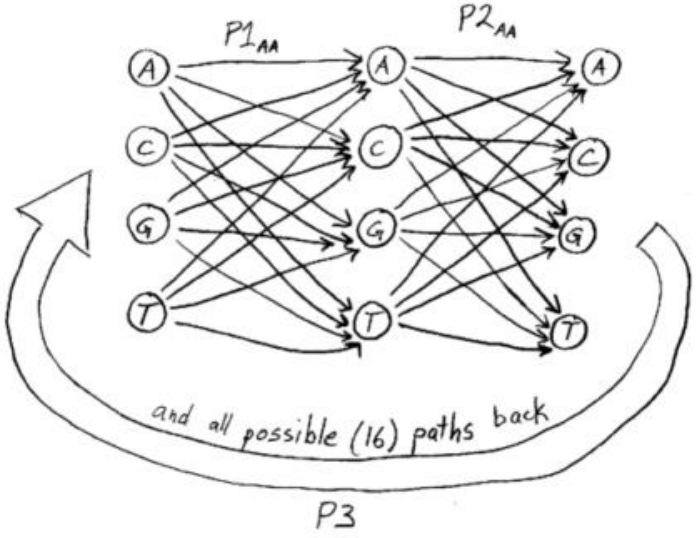

Figure 3: The Markov model for the GENMARK program.