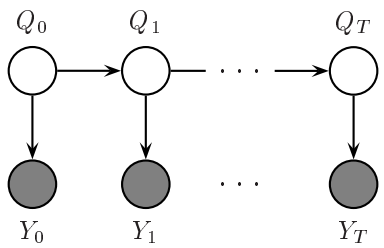
CS281A/Stat241A: Statistical Learning Theory

Hidden Markov Models

Lecturer: Michael I. Jordan *Scribes: Peter Broadwell and Stephen Sorkin*

1 Formal Structure

Recall the formal structure of a Hidden Markov Model (HMM). In the system illustrated below, q_i is the state of the system at time i, represented by a multinomial random variable. At time i, observation y_i is made. The ranom variable embodied by each y can be as complex as desired, since the inference calculation only requires the probability of an observation given the state at a particular time step.



This model can be made more complex by adding extra arrows from the observations to other state nodes. This, however, does not change the inference procedure that will be described. It will just add extra parameters to the system. Our goal is to consider an arbitrary system where the Y_i are high-dimensional and the Q_i are low-dimensional.

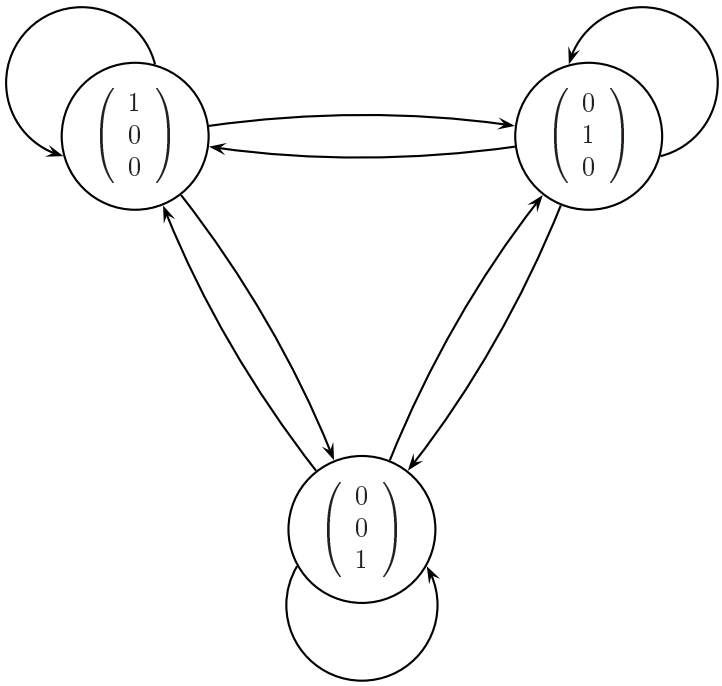
Consider the following parameterization of the model:

$$\begin{aligned} A &: a_{ij} = \Pr(Q_{t+i}^j = 1 | Q_t^i = 1) \\ \pi &: \pi_i = \Pr(Q_0^i = 1) \\ &: \Pr(Y_t | Q_t, \eta) \end{aligned}$$

We call the last parameter B if the observations are multinomial as well. Specifically, A characterizes the probability of one state given the previous and π characterizes the probability distribution over the first state.

2 Applications

Before considering the inference problem on HMMs, we look at a few of their applications. First we note that some of the applications are motiviated by alternate representations of HMMs. The structure of the HMM-derived stochastic automaton, for example, provides a more intuitive representation of the state-changing characteristics of an HMM:



This is a stochastic automaton with three states. The edges indicate the transition probabilities to the other states. At any time step q_t, the system is in one of the states.

2.1 Gene Finding

One appliation of HMMs is in the field of molecular biology. Consider the following segment of a chromosome:

promoters	**UTR**	**exon 1**	**intron 1**	**...**	**exon N**	**intron N**
GTACAACCA	ACA	GCATC	TTACCAT	...	GCAAA	TCCCAGT

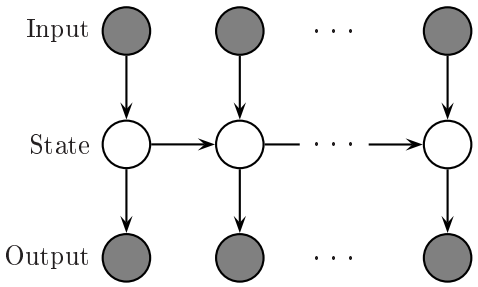
Based on our knowledge of the process of RNA transcription, it's possible to identify the different regions of a chromosome by progressing down the length of the chromosome and using a properly-parameterized HMM to determine the state (region) we're most likely to be in at any given point.

2.2 Part-of-Speech classification

HMMs are useful for language modeling because they can represent a simple parser. The current state at a given point in a sentence describes the part of speech of that word, and the state transition probabilities define the probability of the next word's part of speech, given the current state and the input.

2.3 Stochastic Transducer

Another language-based application is translation: the figure below illustrates the use of an HMM as a "stochastic transducer", which generates a mappping from one string to another.



2.4 Classification

Finally, HMMs can be applied to classification problems such voice recognition — identifying spoken words from a finite possible set. This can be accomplished by "training" a set of HMMs with pre-recorded voices, then applying live input to the entire set and choosing the most likely match (for example, the word that was most likely spoken out of the set of digits "one" through "ten").

3 Parameter Estimation via EM

Given the structure of an HMM, we would like to estimate the unknown parameters of the model. This is a simple graphical model, so EM is the algorithm of choice. First consider the complete likelihood of the graphical model where the parameter vector is $\theta = (\pi, A, \eta)$:

$$\begin{aligned} p(q, y|\theta) &= \pi_{q_0} \prod_{t=0}^{T-1} a_{q_t, q_{t-1}} \prod_{t=0}^{T} p(y_t|q_t, \eta) \\ \pi_{q_0} &\triangleq \prod_i \pi_i^{q_0^i} \\ a_{q_t, q_{t+1}} &\triangleq \prod_{i,j} a_{ij}^{q_t^i q_{t+1}^j} \end{aligned}$$

This yields the complete log likelihood, assuming that the observations are of a multinomial random variable ($\eta_{ij} \triangleq P(Y_t^i = 1|Q_t^i = 1)$):

$$\begin{aligned} \ell_C(\theta|q, y) &= \log\{p(q, y|\theta)\} \\ &= \log\left\{ \prod_i \pi_i^{q_0^i} \prod_{t=0}^{T-1} \prod_{i,j} a_{ij}^{q_t^i q_{t+1}^j} \prod_{t=0}^{T} \prod_{i,j} \eta_{ij}^{q_t^i y_t^j} \right\} \\ &= \sum_i q_0^i \log \pi_i + \sum_{i,j} \sum_{t=0}^{T-1} q_t^i q_{t+1}^j \log a_{ij} + \sum_{i,j} \sum_{t=0}^{T} q_t^i y_t^j \log \eta_{ij} \end{aligned}$$

It can then be observed that this probability distribution is in the exponential family and the following relationships hold:

- q_0^i is sufficient for π_i
- $\sum_{t=0}^{T-1} q_t^i q_{t+1}^j$ is sufficient for a_{ij}
- $\sum_{t=0}^{T} q_t^i y_y^j$ is sufficient for η_{ij}

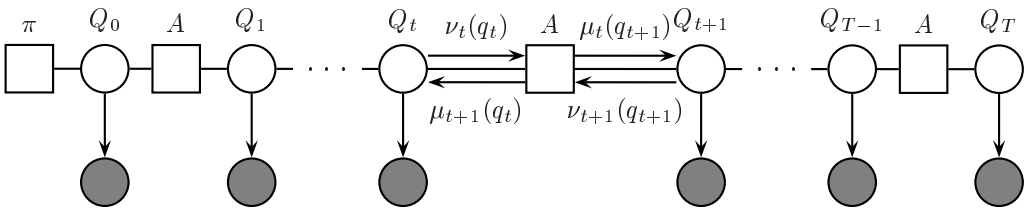
We then take the expected value of the above expression, conditioning on $y, \theta^{(p)}$:

$$\langle \ell_C(\theta|q,y)\rangle_{y,\theta^{(p)}} = \sum_i \langle q_0^i\rangle \log \pi_i + \sum_{i,j}\sum_{t=0}^{T-1} \langle q_t^i q_{t+1}^j\rangle \log a_{ij} + \sum_{i,j}\sum_{t=0}^{T} \langle q_t^i\rangle y_t^j \log \eta_{ij}$$

Hence it suffices to compute $\langle q_t^i\rangle_{y,\theta^{(p)}}$ and $\langle q_t^i q_t^j\rangle_{y,\theta^{(p)}}$. These are just marginal distributions, however:

$$\begin{aligned}\langle q_t^i\rangle_{y,\theta^{(p)}} &\triangleq E[Q_t^i|Y,\theta^{(p)}] \\ &= P(Q_t^i = 1|Y,\theta^{(p)})\end{aligned}$$

These values can be computed with the sum-product algorithm. However, to remain consistent with the HMM literature, we will show how to derive these values from a similar message-passing algorithm on the corresponding factor graph. Consider the corresponding factor graph for an HMM (note that this is not a true factor graph since the Q_t, Y_t are adjacent; their factor nodes are implicit):



Here we will call $\alpha(q_t) \triangleq \nu_t(q_t)$ the message passed from Q_t towards Q_{t+1}, and $\beta(q_t) \triangleq \mu_{t+1}(q_t)$ the message passed in the opposite direction on the same edge. Using the definitions of μ, ν from the sum-product algorithm on factor graphs, we have:

$$\begin{aligned}\beta(q_t) &= \sum_{q_{t+1}} \overbrace{p(y_{t+1}|q_{t+1})\beta(q_{t+1})}^{\text{incoming messages}} \overbrace{a_{q_t,q_{t+1}}}^{\text{factor}} \\ \alpha(q_t) &= p(y_t|q_t)\sum_{q_{t-1}} \alpha(q_{t-1}) a_{q_{t-1},q_t}\end{aligned}$$

In the two special cases we have $\alpha(q_1) = p(y_1|q_1)\pi$ and $\beta(q_T) = 1$. We can then define:

$$\gamma(q_t) \triangleq p\left(q_t|y,\theta^{(p-1)}\right)$$

$$\gamma(q_t) \propto \alpha(q_t)\beta(q_t)$$

Hence this algoirthm correctly computes $\langle q_t^i \rangle$. What about the other required expectation, $\langle q_t^i q_{t+1}^j \rangle$? Note that:

$$\begin{aligned} \alpha(q_t) &= p(y_0, \ldots, y_t, q_t) \\ \beta(q_t) &= p(y_{t+1}, \ldots, y_T | q_t) \end{aligned}$$

We then define:

$$\begin{aligned} \xi(q_t, q_{t+1}) &\triangleq p(q_t, q_{t+1} | y) \\ &= \frac{p(y|q_t, q_{t+1}) p(q_t, q_{t+1})}{p(y)} \\ &= \frac{p(y_0, \ldots, y_t | q_t) p(y_{t+1} | q_{t+1}) p(y_{t+2}, \ldots, y_T | q_{t+1}) p(q_{t+1} | q_t) p(q_t)}{p(y)} \\ &= \frac{\alpha(q_t) p(y_{t+1} | q_{t+1}) \beta(q_{t+1}) a_{q_t, q_{t+1}}}{p(y)} \end{aligned}$$

3.1 E Step

We now have the definitions necessary to compute the E step of the EM algorithm:

1. Compute $\alpha(q_t)$ and $\beta(q_t)$ for all t.
2. Compute $\gamma(q_t)$.
3. Compute $\xi(q_t, q_{t+1})$.

3.2 M Step

For the M step, we take the derivative of the log likelihood with the estimated sufficient statistics from the E step plugged in, set equal to 0, and solve. See the text for more details.