CS281A/Stat241A: Statistical Learning Theory

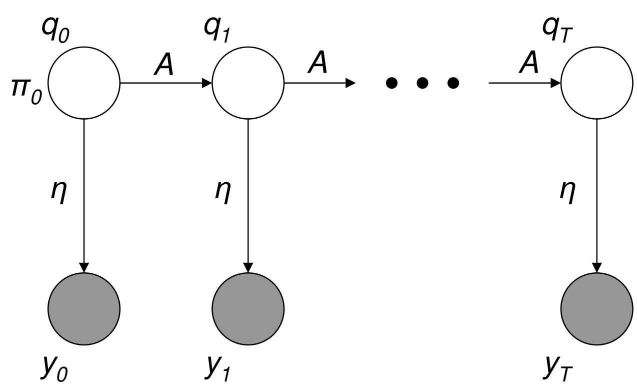
Hidden Markov Models & The Multivariate Gaussian (10/26/04)

Lecturer: Michael I. Jordan *Scribes: Jonathan W. Hui*

1 Hidden Markov Models

As a brief review, hidden Markov models (HMMs) are appropriate for modeling sequential data. Thus, HMMs have been applied to speech recognition, gene finding, and other applications which may involve, but are not restricted to, parsing or segmenting.

The formal structure of an HMM is shown below, where the representation can be viewed as a chain of mixture models. The "hidden" states are denoted by q_t and the observed values are denoted by y_t where t is a specific point in time.



2 HMM Parameter Estimation

Recall from last lecture the complete log likelihood for the HMM:

$$l_c(\theta) \quad = \quad \sum_{i=1}^{M} q_0^i \log \pi_i + \sum_{t=0}^{T-1} \sum_{i,j=1}^{M} q_t^i q_{t+1}^j \log a_{ij} + \sum_{t=0}^{T} \log p(y_t|q_t, \eta) \tag{1}$$

Where $p(y_t|q_t, \eta)$ is the probability distribution of each output node.

We observe that the complete probability distribution is in the exponential family where q_0^i is the sufficient statistic for π_i, and $\sum_{t=0}^{T-1} q_t^i q_{t+1}^j$ is the sufficient statistic for a_{ij} (the sufficient statistic for η depends on the distribution chosen for $p(y_t|q_t, \eta)$).

Note: In this discussion, we have left the distribution on the output values arbitrary and thus ignore the $\langle \log p(y_t|q_t, y) \rangle$ term. Refer to chapter 12 of the text for an example where the outputs y_t are multinomial variables.

2.1 E Step

For the E step, we take the expected value of the complete log likelihood, conditioning on the parameters at iteration p, $\theta^{(p)}$.

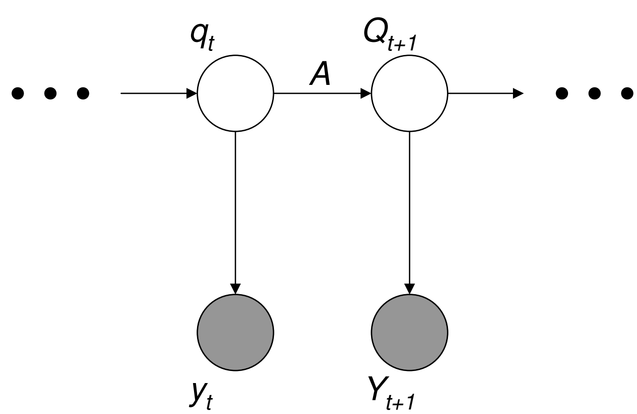
$$\begin{aligned}
\langle l_c(\theta|q,y)\rangle_{y,\theta^{(p)}} &= \langle\sum_{i=1}^{M} q_0^i \log \pi_i + \sum_{t=0}^{T-1}\sum_{i,j=1}^{M} q_t^i q_{t+1}^j \log a_{ij} + \sum_{t=0}^{T} \log p(y_t|q_t,\eta)\rangle \qquad (2)\\
&= \sum_{i=1}^{M}\langle q_0^i\rangle \log \pi_i + \sum_{t=0}^{T-1}\sum_{i,j=1}^{M}\langle q_t^i q_{t+1}^j\rangle \log a_{ij} + \sum_{t=0}^{T}\langle \log p(y_t|q_t,\eta)\rangle \qquad (3)
\end{aligned}$$

Thus, we must compute $\langle q_0^i\rangle_{y,\theta^{(p)}}$ and $\langle q_t^i q_{t+1}^j\rangle_{y,\theta^{(p)}}$. However, these are just marginal distributions:

$$\begin{aligned}
\langle q_0^i\rangle_{y,\theta^{(p)}} &= E(q_0^i|y,\theta^{(p)}) \qquad (4)\\
&= p(q_0^i = 1|y,\theta^{(p)}) \qquad (5)\\
\langle q_t^i q_{t+1}^j\rangle_{y,\theta^{(p)}} &= \sum_{t=0}^{T-1} E(q_t^i q_{t+1}^j|y,\theta^{(p)}) \qquad (6)\\
&= \sum_{t=0}^{T-1} p(q_t^i q_{t+1}^j|y,\theta^{(p)}) \qquad (7)
\end{aligned}$$

We can compute these values via the SUM-PRODUCT algorithm.

While the marginals can be computed via the SUM-PRODUCT algorithm, we will remain consistent with the HMM literature and show how to calculate these values via the *alpha-beta* algorithm (also called *forward-backward*). We first show that the calculation of the β's in the alpha-beta algorithm is identical to the SUM-PRODUCT algorithm. Consider the fragment of the graphical model representation of the HMM below:



The β's of the alpha-beta algorithm is given by:

$$\beta(q_t) = \sum_{q_{t+1}} p(y_{t+1}|q_{t+1})\beta(q_{t+1})a_{q_t,q_{t+1}} \qquad (8)$$

By the SUM-PRODUCT algorithm, the message sent from q_{t+1} to q_t is given by:

$$m_{q_{t+1}}(q_t) = \sum_{q_{t+1}} m_{q_{t+2}}(q_{t+1})p(y_{t+1}|q_{t+1})a_{q_t,q_{t+1}} \qquad (9)$$

We see that this is exactly the same as the calculation of the β's where $\beta(q_t) = m_{q_{t+1}}(q_t)$. Note that we drop the q_{t+1} notation in the β's since the chain structure of the HMM already implies that the message is sent by q_{t+1}. We can also write $\beta(q_t)$ as follows:

$$\beta(q_t) \equiv p(y_{t+1}, \ldots, y_T | q_t) \tag{10}$$

which is the probability of emitting a partial sequence of outputs $y_{t+1}, \ldots, y_T$ given that the system starts in state q_t.

In the alpha-beta algorithm, the α's are defined to be:

$$\alpha(q_t) \equiv p(y_0, \ldots, y_t, q_t) \tag{11}$$

which is the probability of emitting a partial sequence of outputs $y_0, \ldots, y_t$ and ending up in state q_t.

2.2 M Step

We define γ_t^i to be equal to $\langle q_t^i \rangle$, and $\xi_{t,t+1}^{ij}$ to be equal to $\langle q_t^i q_{t+1}^j \rangle$. We can write the expected complete log likelihood as:

$$\langle l_c(\theta) \rangle = \sum_{i=1}^{M} \gamma_t^i \log \pi_i + \sum_{t=0}^{T-1} \sum_{i,j=1}^{M} \xi_{t,t+1}^{ij} \log a_{ij} + \sum_{t=0}^{T} \langle \log p(y_t | q_t, y) \rangle \tag{12}$$

In deriving the M step for the HMM, we use a lemma that is useful throughout this class, and thus we present it here. Given $J(\pi)$ as follows:

$$J(\pi) = \sum_i a_i \log \pi_i \tag{13}$$

we would like to maximize $J(\pi)$ such that $\sum_i \pi_i = 1$ and $\pi_i > 0$. The solution is $\hat{\pi}_i = \frac{a_i}{\sum_j a_j}$. To see that this is true, we simply take the derivative with respect to π and set to zero. We use a Lagrange multiplier to represent the constraint that the π_i's must sum to one.

$$\widetilde{J}(\pi) = \sum_i a_i \log \pi_i + \lambda(1 - \sum_i \pi_i) \tag{14}$$

$$\frac{\partial J}{\partial \pi_i} = \frac{a_i}{\pi_i} - \lambda \tag{15}$$

$$\lambda = \frac{a_i}{\pi_i} \tag{16}$$

$$\frac{a_i}{\lambda} = \pi_i \tag{17}$$

$$\Rightarrow \hat{\pi}_i = \frac{a_i}{\sum_j a_j} \tag{18}$$

Using this lemma, we derive the following equations for the M step.

$$\hat{\pi}_i^{(p+1)} = \frac{\gamma_0^i}{\sum_j \gamma_0^j} \tag{19}$$

$$\hat{\alpha}_{ij}^{(p+1)} = \frac{\sum_{t=0}^{T-1} \xi_{t,t+1}^{i,j}}{\sum_{k=1}^{M} \sum_{t=1}^{T-1} \xi_{t,t+1}^{i,k}} \tag{20}$$

$$= \frac{\sum_{t=0}^{T-1} \xi_{t,t+1}^{T-1}}{\sum_{t=0}^{T-1} \gamma_t^i} \tag{21}$$

In the case of HMMs, these equations are also known as the "Baum-Welch updates".

Note: In some cases, we would like to calculate the configuration of states on the HMM that has the highest probability given observed values for y_t. We can solve this by using the well-known Verterbi algorithm, which essentially is the MAX-PRODUCT algorithm.

2.3 Concrete Example

To give a concrete example, we compute the expected complete log likelihood when the probability distribution on the output values is Gaussian.

$$
\begin{aligned}
\langle l_c(\theta)\rangle &= \cdots + \sum_{t=0}^{T} \langle \log p(y_t|q_t, \eta)\rangle && (22)\\
&= \cdots + \sum_{t=0}^{T} \langle \log \prod_i p(y_t|q_t^i = 1, \eta)^{q_t^i}\rangle && (23)\\
&= \cdots + \sum_{t=0}^{T} \langle \sum_i q_t^i \log p(y_t|q_t^i = 1, \eta)\rangle && (24)\\
&= \cdots + \sum_{t=0}^{T} \sum_i \langle q_t^i\rangle \left(-\frac{1}{2\sigma^2}(y_t - \mu_i)^2\right) && (25)\\
&= \cdots - \sum_{t=0}^{T} \sum_i \gamma_t^i \left(\frac{1}{2\sigma^2}(y_t - \mu_i)^2\right) && (26)
\end{aligned}
$$

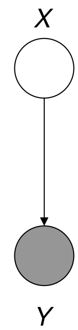
Note that this is a weighted least squares problem.

2.4 Numerical Issues

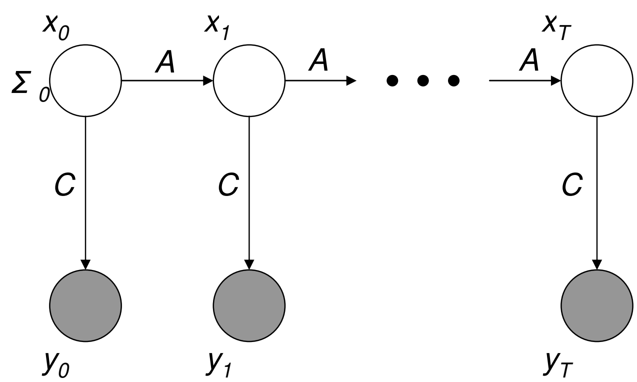
When implementing an HMM on the computer, special care has to be taken with regard to numerical issues. Specifically, the forward and backward recursions involve repeated multiplications of probabilities (i.e. numbers less than one). These repeated multiplications of small numbers generally lead to underflow. However, we can simply get around this problem by normalizing after each recursion step. See chapter 12 for more details on how to derive these normalization update equations.

3 Factor Analysis Models and HMMs

So far, we've viewed HMMs as a chain of mixture models where the states q_t are based on a discrete latent variable. We can also consider *factor analysis models* which are based on continuous latent variables. The underlying graphs of these models are identical. Roughly, factor analysis can be considered a probabilistic form of principle component analysis (PCA).



In the dynamical generalization of factor analysis, called the Kalman Filter, the hidden states are represented by x_t and the observed values as y_t. To represent the transition between nodes, we allow the mean of the state at time $t+1$ be a linear function of the state at time t plus Gaussian error ϵ_t, with mean 0. The initial state, x_0, is endowed with a Gaussian distribution with mean 0 and covariance Σ_0. The state space model is shown below.



This dynamical generalization of factor analysis yields time series analysis methods known as the *Kalman filter* and the *Rauch-Tung-Striebel* smoother.

4 Multivariate Gaussians

We often express the multivariate Gaussian distribution using the parameters μ and Σ, where μ is a $d \times 1$ vector and Σ is a $d \times d$, symmetric matrix. Using these parameters, we have the following form for the density function:

$$p(x|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \texttt{exp} \left\{ -\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right\} \tag{27}$$

where x is a vector in R^d.

Alternatively, we can use a different parameterization, the canonical parameterization. We define the canonical parameters as follows:

$$\Lambda = \Sigma^{-1} \tag{28}$$

$$\eta = \Sigma^{-1}\mu \tag{29}$$

Note that these are invertible and that we can calculate the moment parameters as follows:

$$\mu = \Lambda^{-1}\eta \tag{30}$$
$$\Sigma = \Lambda^{-1} \tag{31}$$

Using the canonical parameterization, we obtain the following density function:

$$p(x|\eta,\Lambda) = \texttt{exp}\left\{\eta^T x - \frac{1}{2}x^T\Lambda x + a(\eta,\Lambda)\right\} \tag{32}$$

We now introduce the trace trick. We define the trace of a square matrix A to be the sum of the diagonal elements a_{ii} of A:

$$\texttt{tr}(A) \equiv \sum_i a_{ii} \tag{33}$$

An important property is its invariance to cyclical permutations of matrix products:

$$\texttt{tr}(ABC) = \texttt{tr}(CAB) = \texttt{tr}(BCA) \tag{34}$$

Using the trace trick, we can have the following:

$$x^T\Lambda x = \texttt{tr}(x^T\Lambda x) = \texttt{tr}(\Lambda x x^T) \tag{35}$$

Thus, we can rewrite the density function as follows:

$$p(x|\eta,\Lambda) = \texttt{exp}\left\{\eta^T x - \frac{1}{2}\texttt{tr}(\Lambda x x^T) + a(\eta,\Lambda)\right\} \tag{36}$$

We now see that sufficient statistic is (x, xx^T) and the canonical parameter is $(\eta, -\frac{1}{2}\Lambda)$.

We can partition the $d \times 1$ parameter vector x into a $p \times 1$ sub-vector x_1 and $q \times 1$ sub-vector x_2, where $n = p + q$. The corresponding partitions of the μ and Σ parameters are:

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \tag{37}$$
$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \tag{38}$$

In the next class, we will talk about inverses of partition matrices. Given a block matrix:

$$M = \begin{pmatrix} E & F \\ G & H \end{pmatrix} \tag{39}$$

The Schur complement of the matrix M with respect to H, denoted $M\backslash H$, is defined to be $E - FH^{-1}G$. We will also obtain an important result involving the determinant of M:

$$|M| = |H||M\backslash H| \tag{40}$$