Prof. Dahleh, Prof. Mitter
Scribed by Rajiv Menjoge

Lecture 15
Th 4/5

Last lecture, we began to describe a model which incorporates sample dependence called the Hidden Markov Model. The model assumes the sequence to be analyzed $Y^\ell = (Y_1, ..., Y_\ell)$ has a corresponding sequence of random variables $X^\ell = (X_1, ...X_\ell)$ such that each X_i takes values in a finite set. In addition Y^ℓ and X^ℓ have the following properties:

$$P(X_{t+1}|X^t, Y^t) = P(X_{t+1}|X_t)$$
$$P(Y_t|X^t, Y^{t-1}) = P(Y_t|X_t)$$

Let $P(X_{t+1} = j|X_t = i)$ be denoted a_{ij} and let $P(Y_t = \nu|X_t = i)$ be denoted $b_i(\nu)$. Hence, we have two sets of parameters. The random variable X^ℓ can be thought of as a typical Markov Process, and the random variable Y^ℓ can be thought of as being derived in some stochastic way from this process.

An alternate description of a Hidden Markov Model is to denote $P(Y_t = \nu, X_t = j|X_{t-1} = i)$ as $[M(\nu)]_{ij}$ so that $P(Y_t = \nu, X_t|X_{t-1}) = M(\nu)$. In this lecture, we will discuss how to compute the following three quantities of interest:

1. $P(Y^\ell)$
2. $P(X_t = i|Y^\ell)$
3. $\text{argmax}_{a_{ij}, b_i(\nu)} \log(P(Y^\ell))$

1 Computing $P(Y^\ell)$

Note that the event that $Y^\ell = \nu$ is the event that $Y^\ell = (Y_1 = \nu_1, ..., Y_\ell = \nu_l)$. Now first consider the quantity $P(Y^\ell = \nu, X^\ell = q)$. Assume for now that we are given all parameter values. We then have:

$$\begin{aligned}P(Y^\ell = \nu, X^\ell = q) &= P(Y_\ell = \nu_l, X_\ell = q_l|Y^{\ell-1} = \nu^{\ell-1}, X^{\ell-1} = q^{\ell-1})P(Y^{\ell-1} = \nu^{\ell-1}, X^{\ell-1} = q^{\ell-1}) \\ &= P(Y_\ell = \nu_l, X_\ell = q_l|X_{\ell-1} = v_{\ell-1})P(Y^{\ell-1} = \nu^{\ell-1}, X^{\ell-1} = q^{\ell-1})\end{aligned}$$

We are given the value of the first expression and have a recursion here. We can repeat this procedure now on $P(Y^{\ell-1}, X^{\ell-1})$ and etc. till we get the following expression:

$$\begin{aligned}P(Y^\ell = \nu, X^\ell = q) &= P(x_0 = q_0) \times \textstyle\prod_{i=1}^{\ell} P(Y_i = \nu_i, X_i = q_i|X_{i-1} = q_{i-1}) \\ &= P(x_0 = q_0) \times \textstyle\prod_{i=1}^{\ell} a_{q_{i-1}q_i} b_{q_i}(v_i)\end{aligned}$$

It follows that, where e is a vector with all entries equal to 1 and π is the distribution of the initial state:

$$
\begin{aligned}
P(Y^\ell = \nu) &= \sum_q P(Y^\ell = \nu, X^\ell = q) \\
&= \sum_q P(x_0 = q_0) \times (\prod_{i=1}^{\ell} P(Y_i = \nu_i, X_i = q_i | X_{i-1} = q_{i-1})) \\
&= \pi^T \times (\prod_{i=1}^{\ell} M(v_i))e
\end{aligned}
$$

2 Computing $P(X_t = i|Y^\ell = \nu)$

We use a forward recursion and a backward recursion in order to compute this. Note that this computation is a filtering problem if $t = \ell$, a smoothing problem if $t < \ell$ and a prediction problem if $t > \ell$. For the forward recursion, denote $\alpha_t(i) = P(Y^t = v^t, X_t = i)$. For the backward recursion, denote $\beta_t(i) = P(Y^\ell_{t+1} = v^\ell_t, X_t = i)$, where $Y^\ell_{t+1} = \{Y_{t+1}, Y_{t+2}, ...Y_\ell\}$.

Note that in this notation, $\sum_{i=1}^{\ell} \alpha_l(i) = P(Y^\ell = \nu)$. Also, we have that:

$$
\begin{aligned}
\sum_{i=1}^{\ell} \alpha_t(i)\beta_t(i) &= \sum_{i=1}^{l} P(Y^t = v^t, X_t = i) \times P(Y^\ell_{t+1} = v^\ell_t, X_t = i | X_t = i, Y^t = v^t) \\
&= \sum_{i=1}^{l} P(Y^t = v^t, X_t = i, Y^\ell_{t+1} = v^\ell_t, X_t = i) \\
&= P(Y^\ell = \nu)
\end{aligned}
$$

Using similar arguments, we can derive that $P(X_t = i|Y^\ell = v) = \frac{\alpha_t(i)\beta_t(i)}{\sum_i \alpha_t(i)\beta_t(i)}$. It turns out that we can set up a recursion for $\alpha_t(i)$ and $\beta_t(i)$ and we can find that $\alpha_t(i) = \sum_j a_{ji} b_i(\nu_t)\alpha_{t-1}(j)$ and that $\beta_t(i) = \sum_j a_{ji} b_i(v_t)\beta_{t+1}(j)$. Now the value of $\alpha_1(j) = P(Y_1 = \nu_1, X_1 = i) = b_i(\nu_1)\pi_i$ and the value of $\beta_{\ell-1}(i) = P(Y^\ell = \nu_l | X_{\ell-1} = i) = b_i(\nu_l)$. Since both of these values are known, we can compute $\alpha_t(i)$ and $\beta_t(i)$ for any t. From here, we can calculate $P(X_t = i|Y^\ell = v) = \frac{\alpha_t(i)\beta_t(i)}{\sum_i \alpha_t(i)\beta_t(i)}$. The proof that $\alpha_t(i) = \sum_j a_{ji} b_i(\nu_t)\alpha_{t-1}(j)$ is shown below. The proof that $\beta_t(i) = \sum_j a_{ji} b_i(v_t)\beta_{t+1}(j)$ is similar.

$$
\begin{aligned}
\alpha_t(i) &= P(Y^t = \nu^t, X_t = i) \\
&= P(Y^{t-1} = \nu^{t-1}, Y_t = \nu_t, X_t = i) \\
&= \sum_j P(Y^{t-1} = \nu^{t-1}, Y_t = \nu_t, X_t = i, X_{t-1} = j) \\
&= \sum_j P(Y_t = \nu_t, X_t = i | X_{t-1} = j, Y^{t-1} = \nu^{t-1}) P(X_{t-1} = j, Y^{t-1} = \nu^{t-1}) \\
&= \sum_j a_{ji} b_i(\nu_t)\alpha_{t-1}(j)
\end{aligned}
$$

3 Computing the maximum likelihood estimates of π, a_{ij}, and $b_i(\nu)$

The problem this section concerns itself with is finding the values π, a_{ij}, and $b_i(\nu)$ which maximize the log-likelihood, $\log(P(Y^n))$. Let us assume that $a_{ij}, b_i(\nu)$ are unknown and our class is $C = a_{ij}, b_i(\nu)$, where α denotes an instance of these parameters. Here, we have that $R^\ell_{emp}(\alpha) =$

$\frac{1}{\ell}\log P_\alpha(Y^\ell = \nu)$. There is a result which says that if Y^ℓ is a stationary process, then $R^\ell_{emp}(\alpha) \longrightarrow \lim_{\ell\to\infty} \frac{1}{\ell}E(\log(P_\alpha(Y^\ell)))$. Also, where $\alpha_l = \text{argmax}(R^\ell_{emp}(\alpha))$, we have that

$$\alpha_l \longrightarrow \text{argmin}(\lim_{\ell\to\infty} \frac{1}{\ell} D(P(Y^\ell), P_\alpha(Y^\ell)))$$

By the results of section 2, the log-likelihood is given by $\frac{1}{\ell}\log\sum_q \pi(q_0)\prod_{i=1}^{\ell} a_{q_{i-1}q_i} b_{q_i}(\nu)$. Though this appears to look like the corresponding problem for Markov Models, it is nontrivial. The main issue is the fact that the variables q are unknown. It would be much easier if we did know q, because then we wouldn't have to deal with a log over a sum. The standard method for finding solutions to this problem is an iterative algorithm known as the Expectation Maximization (EM) Algorithm, which is often used to solve maximum likelihood problems which involve missing data. Intuitively, here, q is treated as though it were missing data. The algorithm attempts to find parameters that maximize the expected value of the likelihood over the estimated distribution of q. Given the new parameters, it can update the estimated distribution over q. The algorithm is written below.

1. Pick α
2. Expectation Step:
   Set $J(\tilde{\alpha}, \alpha) = \sum_q \log P_{\tilde{\alpha}}(Y^\ell = \nu, X^\ell = q) P_\alpha(Y^\ell = \nu, X^\ell = q)$
3. Maximization Step: Find $\alpha^\star = \text{argmax}_{\tilde{\alpha}}\, J(\tilde{\alpha}, \alpha)$
4. Set α to be $\alpha^\star$ and return to step 2.

The justification for this algorithm is in the following proposition. Here, for shorthand, $P(Y^\ell = \nu)$ is denoted $P(\nu)$ and $P(Y^\ell = \nu, X^\ell = q)$ is denoted $P(\nu, q)$

Theorem 1. *If $J(\alpha^\star, \alpha) > J(\alpha, \alpha)$, then $P_{\alpha^\star}(Y^\ell = \nu) > P_\alpha(Y^\ell = \nu)$*

Proof.

$$\begin{aligned}
\log \frac{P_{\alpha^\star}(\nu)}{P_\alpha(\nu)} &= \log \frac{\sum_q P_{\alpha^\star}(\nu, q)}{P_\alpha(\nu)} \\
&= \log \frac{1}{P_\alpha(\nu)} \sum_q P_\alpha(\nu, q) \frac{P^\star_\alpha(\nu, q)}{P_\alpha(\nu, q)} \\
&\geq \frac{1}{P_\alpha(\nu)} \sum_q P_\alpha(\nu, q) \log \frac{P^\star_\alpha(\nu, q)}{P_\alpha(\nu, q)} \\
&= \frac{1}{P_\alpha(\nu)} \sum_q (J(\alpha^\star, \alpha) - J(\alpha, \alpha)) \\
&> 0
\end{aligned}$$

□

The third line follows from Jensen's Inequality and the fifth line follows from the given. It turns out that the solution to this maximization problem in this case is as follows:

$$
\begin{aligned}
\alpha^\star &= (\pi^\star, a^\star, b^\star), \\
\text{Such that :} & \\
\pi^\star(i) &= P_\alpha(Y^\ell = \nu, X_0 = i)/P_\alpha(Y^\ell = \nu), \\
a^\star_{ij} &= \sum_{t=1}^{\ell} P_\alpha(Y^\ell = \nu, X_{t-1} = i, X_t = j) / \sum_{t=1}^{\ell} P_\alpha(Y^\ell = \nu, X_{t-1} = i), \\
b^\star_i(v) &= \sum_{\{t \,:\, y_t = v\}} P_\alpha(Y^\ell = \nu, X_t = i) / \sum_{t=1}^{\ell} P_\alpha(Y^\ell = \nu, X_t = i).
\end{aligned}
$$